New Folder Name The Physics of LIGO

Lecture Notes & Exercises, Two volumes

THE PHYSICS OF LIGO

I. Lecture Notes & Exercises

Materials from a Course taught at Caltech
by members of the LIGO team and others
in Spring 1994

organized and edited by Kip S. Thorne

California Institute of Technology

1994

PREFACE

In the spring term of 1994, I organized a course at Caltech on *The Physics of LIGO* (i.e., the physics of the Laser Interferometer Gravitational Wave Observatory). The course consisted of eighteen 1.5-hour-long tutorial lectures, delivered by members of the LIGO team and others, and it was aimed at advanced undergraduates and graduate students in physics, in applied physics, and in engineering and applied sciences, and also at interested postdoctoral fellows, research staff, and faculty.

In my mind the course had several purposes: (i) It used LIGO as a vehicle for teaching students about the physics and technology of high-precision physical experiments. (ii) It served as a tutorial on the physics of LIGO for scientists and engineers, who had joined the LIGO team in the preceding year in preparation for the beginning of LIGO's construction. (iii) It served as an introduction to the science and technology of LIGO for other members of the Caltech community: In spring 1994, LIGO was just beginning to emerge from two years of controversy on the Caltech campus, and a number of faculty and staff wanted to learn in detail about the LIGO team's interferometer R&D, so they could form opinions of their own about whether the Project was well conceived and its interferometer development was being well executed. (It is my impression, in retrospect, that most and perhaps all of the faculty and staff who attended the course regularly emerged with a positive view LIGO.)

The lectures were delivered in Room 107 Downs on Wednesdays from 1:00 to 2:30 PM and Fridays from 10:30AM to noon. The audience typically consisted of about 5 undergraduates, 10 graduate students, 5 postdoctoral fellows, 8 professors, and 15 members of the LIGO team—and, for some lectures, rather more than this, especially more professors. The audience was mostly from Physics and Engineering, but a smattering of other disciplines was represented (including even an occasional social scientist). The undergraduates and some of the graduate students took the course for credit under the rubric of Physics 103.

These two Volumes contain the materials distributed at the lectures, augmented occasionally in Volume I by lecture notes that Malik Rakhmanov (the grader) or I have written, describing the lecture. More specifically:

Volume I contains (i) copies of the transparencies used in each lecture, or—in the case of lectures not based on transparencies—notes on the lecture prepared by Rakhmanov or me; and (ii) lists of references and sets of exercises prepared by me and/or the lecturers.

Volume II contains copies of the most important of the references that the lecturers chose to accompany their lectures. Some references are extracted from textbooks or technical monographs, others are from the original scientific literature, and a few are preprints of papers not yet published. Because we have not sought, from the publishers of these references, permission for widespread duplication and distribution, only a few copies of

Volume II are being made; and Volume II carries an admonition on its cover page that it should not be reproduced.

For these volumes I have given sequential capital-letter labels (A, B, C, ... Z, AA, BB, ... YY) to all the readings that appear in Volume II, and have revised the reference lists in Volumes I and II to reflect this labeling. References not included as readings are now labeled with lower-case letters (a, b, c, ... z).

These Volumes will be of value not only as a historical record, but also as a reference source for members of the LIGO team and others, and as an aid for people who did not attend the lectures and who want to begin learning about LIGO. For example, people who join the LIGO team during the next several years may find these volumes helpful in getting oriented. (To those who joined the team during the summer or early autumn of 1994, I apologize that I have been so slow in putting these volumes together.)

I thank the lecturers for the extensive time, energy, and enthusiasm that they put into this course. No single person could possibly have delivered this set of lectures, especially not I! I also thank Robbie Vogt and Stan Whitcomb who, as Director and Deputy Director of LIGO, encouraged me in 1993 to organize this course and encouraged the members of the LIGO team to help me make it a reality. Finally, for their enthusiastic backing of this effort, I thank the entire LIGO team, Barry Barish (LIGO PI), Tom Everhart (the Caltech President), Paul Jennings (the Provost), Charles Peck (the Chair of Physics, Mathematics, and Astronomy), and a number of Caltech faculty members.

Kip S. Thorne
Caltech
20 October 1994

CONTENTS OF VOLUME I

Note: The contents of each lecture are described below in outline form (though the lecture typically does not follow the outline sequentially). For each lecture, this volume contains: (i) a list of references prepared by the lecturer and broken down into "Assigned Reading" and "Supplementary Reading," with comments about the relevance of each reference;* (ii) a set of exercises that the reader is invited to try, as a tool to better understanding;† and (iii) the transparencies from which the lecture was delivered and/or notes on the lecture prepared by Kip, or by Malik Rakhmanov.

* Those references labeled by capital letters (A, B, C, ... Z, AA, BB, ... YY) are reproduced in Volume II; those labeled by lower-case letters are not

† The students who took this course for credit were required to work many of these exercises or do supplementary reading and write essays about it.

4. **Idealized theory of interferometers—I** *by Kip S. Thorne* [8 April]
   a. Gaussian beams and their manipulation
   b. beam splitters and mirrors
   c. simple delay-line interferometers
   d. Fabry-Perot cavities
   e. simple Fabry-Perot interferometers

5. **Idealized theory of interferometers—II** *by Ronald W. P. Drever* [13 April]
   a. Power recycled interferometers
   b. Resonantly recycled interferometers
   c. Dual (or signal) recycled interferometers
   d. Doubly resonant signal recycled interferometer
   e. Resonant sideband extraction

6. **Overview of a real interferometer** *by Stanley E. Whitcomb* [15 April]
   a. what is in a real interferometer
   b. survey of potential noise sources
   c. scaling of noise sources
   d. introduction to control systems
   e. diagnostic techniques for real interferometers

7. **Lasers and input optics—I** *by Robert E. Spero* [20 April]
   a. Fabry-Perot cavities as displacement sensors
   b. shot noise in photodetection, signal-to-noise ratio
   c. effect of mirror losses; equivalence of active and passive cavities
   d. phase modulation to sense optical phase and eliminate sensitivity to laser intensity fluctuations; sideband analysis; reflection ("Pound-Drever") locking
   e. experimental demonstration of shot noise limited sensitivity.
   f. non-recombined and recombined optical configurations; sensitivity vs. storage time, visibility, modulation waveform
   g. optimization of optical and modulation parameters

8. **Lasers and input optics—II** *by Alex Abramovici* [22 April]
   a. general requirements on light for LIGO interferometers
   b. configuration of the light source (laser, mode cleaner, other components and subsystems)
   c. Argon ion laser; its single-frequency operation and frequency prestabilization
   d. beam jitter in terms of mode superposition
   e. mode cleaner and mode matching
   f. Nd:YAG laser and frequency doubling

9. **Optical elements** *by Rick L. Savage* [27 April]
   a. overview of LIGO's requirements for mirrors, beam splitters, phase modulators, photodiodes, pick offs, etc.
   b. detailed requirements for test-mass optics:
   general requirements—total losses, reflectivity, radius of curvature, etc.
   mirror surface imperfections and how they influence the interferometer;
   contamination-induced mirror heating.
   c. the LIGO core optics pathfinder program:
   mirror substrates (mechanical quality factors, polishing, Zernike polynomials, measurement of polished surfaces);
   mirror coatings.

10. **Control systems for test-mass position and orientation** *by Seiji Kawamura* [29 April]
   a. test-mass suspension systems
   b. test-mass position and orientation damping
   c. transfer function of a pendulum
   d. sensors and actuators
   e. test mass orientation noise in a Fabry-Perot interferometer
   f. noise from the control system

11. **Optical topology for the locking and control of an interferometer, and signal extraction** *by Martin W. Regehr* [4 May]
   a. overview and explanation of the modulation methods used to extract the gravitational wave signal
   b. methods of extracting the auxiliary signals necessary for locking an interferometer
   c. analysis of multivariable control systems, with examples

12. **Seismic isolation** *by Lisa A. Sievers* [6 May]
   a. seismic background: its origin and spectrum
   b. isolation stacks: basic theory, design issues, chosen design and performance
   c. isolation via pendulum suspension; compound pendulum
   d. active isolation systems

13&14. **Test masses and suspensions and their thermal noise** *by Aaron Gillespie* [11 May and 13 May]
   a. key issues in elasticity theory
   b. fluctuation-dissipation theorm
   c. causes of losses in materials
   d. frequency dependence of noise
   e. suspension noise
   f. violin-mode noise
   g. internal-mode noise
   h. "excess" (non-Gaussian) noise
   i. choices of materials

15. **Light scattering and its control** *by Kip S. Thorne* [18 May, 1st half]
    a. how scattered light can imitate a gravitational wave; the magnitude of the danger
    b. control of scattered light by baffles and by choice of materials
    c. the chosen LIGO baffle design

16. **Squeezed light and its potential use in LIGO** *by H. Jeff Kimble* [18 May 2nd half, and 20 May]
    a. theory of squeezing
    b. practical methods of squeezing
    c. present state of the art
    d. use of squeezed vacuum state in interferometers
    e. methods to beat the standard quantum limit

17. **The physics of vacuum systems, and the LIGO vacuum system** *by Jordan Camp* [25 May]
    a. basic physics and engineering of vacuum systems
    b. noise in an interferometer due to residual gas
    c. LIGO vacuum specifications
    d. LIGO's special low-hydrogen steel
    e. outgasing and pumping strategy
    f. construction of the vacuum system

18. **The 40 meter prototype interferometer as an example of many of the issues studied in this course** *by Robert E. Spero* [27 May]
    The following abstract gives the flavor of how Spero approached this topic:
    Building gravity wave detectors like the 40 m interferometer or LIGO proceeds in two steps: constructing an array of test masses that is free from external disturbances and other sources of displacement noise, and devising a sensitive readout of the relative positions of these masses. The noise sources that constrain sensitivity can be classified as *fundamental*, meaning they were expected, ultimately, to limit the detector's sensitivity and their limits were estimated (around 1971) before the first detectors were built, and *technical*, meaning that, whether they initially were thought of or not, they are unlikely to place ultimate limits on sensitivity. Since the required sensitivity is many orders of magnitude greater than anything previously achieved, one might worry about a third class of noise sources: unanticipated fundamental phenomena revealed in the course of the R&D, which will limit the ultimate sensitivities. Luckily, no such phenomena have been discovered. The 40 m interferometer has been invaluable at sorting out which sources of noise (both fundamental and technical) are the most important in the short run and the long, and in guiding the design of LIGO. Our current understanding of the effects of imperfections in phenomena such as phase modulation, intensity stabilization, mechanical servos, and lock acquisition is due largely to investigations conducted on the 40 m and similar experimental interferometers.

Note: A tour of the 40 meter prototype interferometer was taken twice outside lecture hours: once early in the term, largely for impressionistic purposes; once at the end of the course, following up on the last lecture. A tour of the LIGO Optics laboratory in the basement of West Bridge was also taken twice: once in the middle of the term focusing on issues discussed in the term's first half; once at the end of the term, focusing on issues from the term's second half.

CONTENTS OF VOLUME II

Note: For each lecture, this volume contains the list of references prepared by the lecturer with comments about the relevance of each reference, followed by a copy of each of the most important references. In the contents below, we list the copied references for each lecture. The copied references are labeled sequentially by capital letters (A, B, C, ... , Z, AA, BB, ... YY). The other references are labeled sequentially by lower-case letters.

J. B. J. Meers, “Recycling in laser-interferometric gravitational-wave detectors,” *Phys. Rev. D*, **38**, 2317–2326.

K. B. J. Meers, *Physics Letters A*, “The frequency response of interferometric gravitational wave detectors,” *Physics Letters A*, **142**, 465 (1989).

L. B. J. Meers and R. W. P. Drever, “Doubly-resonant signal recycling for interferometric gravitational-wave detectors.” (preprint)

M. J. Mizuno, K. A. Strain, P. G. Nelson, J. M. Chen, R. Schilling, A. Rudiger, W. Winkler and K. Danzman, “Resonant sideband extraction: a new configuration for interferometric gravitational wave detectors,” *Phys. Lett. A*, **175**, 273–276 (1993).

N. R. W. P. Drever, “Interferometric Detectors of Gravitational Radiation,” in *Gravitational Radiation*, N. Deruelle and T. Piran, eds. (North Holland, 1983).

6. **Overview of a real interferometer** *by Stanley E. Whitcomb* [15 April]

O. D. Shoemaker, R. Schilling, L. Schnupp, W. Winkler, K. Maischberger, A. Rudiger, “Noise behavior of the Garching 30-meter prototype gravitational-wave interferometer,” *Physical Review D,* **38**, 423-432 (1988).

P. Benjamin C. Kuo, *Automatic Control Systems* (Prentice-Hall), “Chapter 1. Introduction.”

7. **Lasers and input optics—I** *by Robert E. Spero* [20 April]

Q. A. Rüdiger, R. Schilling, L. Schnupp, W. Winkler, H. Billing and K. Maischberger, “A mode selector to suppress fluctuations in laser beam geometry,” *Optica Acta,* **28**, 641–658 (1981).

R. A. Yariv, *Optical Electronics* (Saunders College Publishing, 1991), “Chapter 10. Noise in Optical Detection and Generation.”

S. A. Abramovici and Z. Vager, “Comparison between active- and passive-cavity interferometers,” *Phys. Rev. A*, **33**, 3181 (1986).

T. J. Gea-Banacloche, “Passive versus active interferometers: Why cavity losses make them equivalent,” *Phys. Rev. A*, **35**, 2518 (1987).

U. T.M. Niebauer, R. Schilling, K. Danzmann, A. Rudiger, W. Winkler, “Nonstationary Shot Noise and its Effect on the Sensitivity of Interferometers *Phys. Rev A* **43**, 5022–5029 (1991).

V. P.H. Roll, R. Krotkov, and R.H. Dicke, “The equivalence of inertial and passive gravitational mass,” *Ann. Phys* **26**, 442 (1964); pages 466–470.

8. **Lasers and input optics—II** *by Alex Abramovici* [22 April]

W. A. Rüdiger, R. Schilling, L. Schnupp, W. Winkler, H. Billing and K. Maischberger, “A mode selector to suppress fluctuations in laser beam geometry,” *Optica Acta,* **28**, 641–658 (1981).

X. W. Koechner, *Solid-State Laser Engineering* (Springer Verlag, Berlin, 1988), “Chapter 1. Introduction.”

9. **Optical elements** *by Rick L. Savage* [27 April]

Y. W. Winkler, K. Danzmann, A. Rüdiger and R. Schilling, “Optical Problems in Interfereometric Gravitational Wave Antennas,” in *The Sixth Marcel Grossmann*

Meeting, eds. H. Sato and T. Nakamura (World Scientific, Singapore, 1991), pp. 176–191.

Z. D. Malacara, *Optical Shop Testing* (John Wiley and Sons, New York, 1978), section 1.2, "Fizeau Interferometer," pp. 19–37.

AA. H. A. Macleod, *Thin-Film Optical Filters,* 2nd edition (Adam Hilger Ltd., Bristol, 1986), "Introduction," pp. 1–10.

10. **Control systems for test-mass position and orientation** *by Seiji Kawamura* [29 April]

BB. S. Kawamura and M. E. Zucker, "Mirror orientation noise in a Fabray-Perot interferometer gravitational wave detector," *Applied Optics*, in press.

CC. M. Stephens, P. Saulson, and J. Kovalik, "A double pendulum vibration isolation system for a laser interferometric gravitational wave antenna," *Rev. Sci. Instrum.*, **62**, 924–932 (1991).

DD. R. C. Dorf, *Modern Control Systems* 5th editon (Addison-Wesley, 1989): §§7.1 and 7.2 of Chapter 7, "Frequency Response Methods;" and §§8.1–8.4 of Chapter 8, "Stability in the Frequency Domain."

11. **Optical topology for the locking and control of an interferometer, and signal extraction** *by Martin W. Regehr* [4 May]

EE. P. W. Milonni and J. H. Eberly, *Lasers* (Wiley, New York, 1988): §§12.9 "AM Locking" and 12.10 "FM Locking," pp. 385–390.

FF. C. N. Man, D. Shoemaker, M. Pham Tu and D. Dewey, "External modulation technique for sensitive interferometric detection of displacements," *Physics Letters A*, **148**, 8–16.

GG. John H. Moore, Christopher C. Davis, and Michael A. Coplan, *Building Scientific Apparatus: A Practical Guide to Design and Construction* (Addison-Wesley, 1983), Sec. 6.8.3 "The lock-in amplifier and gated integrator or boxcar," (pp. 435–437).

HH. Paul Horowitz and Winfield Hill, *The Art of Electronics* (Cambridge University Press, Cambridge, 1980), Sec. 14.15 "Lock-in detection" (pp. 628–631) and an earlier section to which it refers, Sec. 9.29 "PLL components, Phase detector" (pp. 429–430).

12. **Seismic isolation** *by Lisa A. Sievers* [6 May]

II. Leonard Meirovitch, *Elements of Vibration Analysis* (McGraw-Hill, 1986), pp. 39–57.

JJ. R. del Fabbro, A. di Virgilio, A. Giazotto, H. Kautzky, V. Montelatici, and D. Passuello, "Three-dimensional seismic super-attenuator for low frequency gravitational wave detection," *Physics Letters A*, **124**, 253–257 (1987).

KK. C. A. Cantley, J. Hough, and N. A. Robertson, "Vibration isolation stacks for gravitational wave detectors—Finite element analysis," *Rev. Sci. Instrum.*, **63**, 2210–2219 (1992).

LL. M. Stephens, P. Saulson, and J. Kovalik, "A double pendulum vibration isolation system for a laser interferometric gravitational wave antenna," *Rev. Sci. Instrum.*, **62**, 924–932 (1991).

MM. L. Ju, D. G. Blair, H. Peng, and F. van Kann, "High dynamic range measurements of an all metal isolator using a sapphire transducer," *Mass. Sci. Technol.*, **3**, 463–470 (1992).

13&14. **Test masses and suspensions and their thermal noise** *by Aaron Gillespie* [11 May and 13 May]

NN. H. B. Callen and T. A. Welton, "Irreversibility and generalized noise," *Phys. Rev.*, **83**, 34–40 (1951).

OO. Peter R. Saulson, "Thermal noise in mechanical experiments," *Phys. Rev. D*, **42**, 2437–2445 (1990).

PP. Aaron Gillespie and Frederick Raab, "Thermal noise in mechanical experiments," *Phys. Rev. D*, **42**, 2437–2445 (1990).

QQ. Aaron Gillespie and Frederick Raab, "Thermally excited vibrations of the mirrors of a laser interferometer gravitational wave detector," unpublished (1994).

RR. Aaron Gillespie and Frederick Raab, "Suspension losses in the pendula of laser interferometer gravitational wave detectors," *Phys. Lett. A*, in press (1994).

15. **Light scattering and its control** *by Kip S. Thorne* [18 May, 1st half]

SS. J. M. Elson, H. E. Bennett, and J. M. Bennett, "Scattering from Optical Surfaces," in *Applied Optical Engineering*, Vol. VII (Academic Press 1979), Chapter 7, pp. 191–243.

16. **Squeezed light and its potential use in LIGO** *by H. Jeff Kimble* [18 May 2nd half, and 20 May]

TT. C. M. Caves, "Quantum mechanical noise in an interferometer," *Phys. Rev. D*, **23**, 1693–1708 (1981).

UU. D. F. Walls, "Squeezed states of light," *Nature*, **306**, 141–146 (1983).

VV. M. Xiao, L. A. Wu, and H. J. Kimble, "Precision measurement beyond the shot-noise limit," *Phys. Rev. Lett.*, **59**, 278–281 (1987).

17. **The physics of vacuum systems, and the LIGO vacuum system** *by Jordan Camp* [25 May]

WW. J. Moore, C. Davis, M. Coplan, *Building Scientific Apparatus* (Addison-Wesley, 1983), Chapter 3. "Vacuum technology."

18. **The 40 meter prototype interferometer as an example of many of the issues studied in this course** *by Robert E. Spero* [27 May]

XX. Rainer Weiss, "Electromagnetically coupled broadband gravitational antenna," *Quart. Prog. Rep. Res. Lab. Electron. M.I.T.* **105**, 54 (1972).

YY. Robert L. Forward, "Wideband laser-interferometer gravitational-radiation experiment," *Phys. Rev. D* **17**(2), 379–390 (1977).

BATCH
START

1. overview

STAPLE
OR
DIVIDER

LECTURE 1: OVERVIEW

Lecture by Kip S. Thorne

Assigned Reading:

A. "Gravitational Radiation" by Kip S. Thorne, in *300 Years of Gravitation*, eds. S. W. Hawking and W. Israel (Cambridge University Press, 1987), pages 330–350; 378—383; 414–420. [This article is a somewhat out of date review of the entire field of gravitational radiation. The assigned portions deal with (i) those aspects of the theory of gravitational waves that we will need in this course, (ii) the waves emitted by coalescing compact binaries, and (iii) the basic idea of an interferometric gravitational-wave detector (called a "beam" detector in this article. Note that in this article Newton's gravitation constant G and the speed of light c are set equal to unity.]

B. "LIGO: The Laser Interferometer Gravitational-Wave Observatory" by Alex Abramovici, William E. Althouse, Ronald W. P. Drever, Yekta Gürsel, Seiji Kawamura, Frederick J. Raab, David Shoemaker, Lisa Sievers, Robert E. Spero, Kip S. Thorne, Rochus E. Vogt, Rainer Weiss, Stanley E. Whitcomb, and Michael E. Zucker, *Science*, **256**, 325–333 (1992). [This article, written in 1992 by members of the LIGO Science Team, is an overview of the LIGO Project.]

Suggested Supplementary Reading:

A. "Gravitational Radiation" (see above):

- a. pp. 351–364, on methods by which the generation of gravitational waves is computed and on various effects that occur in the propagation of gravitational waves from their sources to Earth.
- b. pp. 364–400, on astrophysical sources of gravitational waves.
- c. pp. 400–415, on bar detectors.
- d. *NOT* pp. 415–435, on interferometric detectors; we will be studying this material later in the course
- d. pp. 435–445, on other methods of detecting gravitational waves.

C. "Gravitational Radiation: An Introductory Review" by Kip S. Thorne, in *Gravitational Radiation*, eds. Nathalie Deruelle and Tsvi Piran (North-Holland, Amsterdam, 1983), pp. 1–58. [This is an introduction to the theory of gravitational waves aimed at people who know the basic concepts and formalism of general relativity.]

A Few Suggested Problems

1. *Behavior of gravitational-wave fields under a rotation of axes.* Derive Eq. (7d) of Ref. [A]; show that it can be rewritten in the form

$$h_+^{\rm new} + ih_\times^{\rm new} = \left(h_+^{\rm old} + ih_\times^{\rm old}\right) e^{-i2\Delta\Psi}$$

2. *Beam pattern of an interferometer.* Derive Eqs. (103) and (104) of Ref. [A] for the response of an interferometric detector to gravitational waves that come from an

arbitrary direction and have an arbitrary polarization. Note that these equations are the precise version of the interferometer response described in Eq. (2) of Ref. [B].

3. *Factors of G and c.* Equations (12)–(19) of Ref. [A], which embody the quadrupole-moment formalism for computing the generation of gravitational waves, are written in so-called "geometrized units" with $G = c = 1$. Restore the factors of G and c so these equations are all dimensionally correct in cgs or mks units.
4. *Gravitational waves from an inspiraling compact binary: Order-of-magnitude analysis.*
   a. Consider a binary system made of two black holes or neutron stars in a circular orbit that gradually shrinks due to gravitational radiation reaction. Use an order-of-magnitude Newtonian (Keplerian) analysis and the quadrupole-moment formulas to compute the binary's rate of loss of energy dE/dt to gravitational waves, and thence the rate at which the binary's orbital radius shrinks, and thence the rate at which the gravitational-wave frequency f increases. Your final answer should be an order-of-magnitude version of Eq. (42d) of Ref. [A].

   b. Restore the factors of G and c to Eq. (42d), and then insert numbers to infer the gravitational-wave frequency in Hertz as a function of time to final coalescence in seconds. Compare your answer with the time markings for NS/NS inspiral in Figure 10 of Ref. [B]. (Both neutron stars there are assumed to have masses of 1.4 suns.)

   c. The waves' "characteristic amplitude" h_c (which is rigorously defined in terms of "optimal signal processing"—a topic to be studied in the next two lectures) is approximately the waves' amplitude h times the square root of the number $n = f^2/(df/dt)$ of cycles that the wave spends near a given frequency, $h_c \simeq h\sqrt{n}$. Compute h, n, and h_c in order of magnitude from the quadrupole-moment formalism. Your answers should agree with Eqs. (42) and (46b) of Ref. [A].
5. *Memory of a Gravitational Wave* Give an example of a source of gravitational waves for which one or both of the fields h_+ and $h_\times$ begin with zero value, then oscillate in some manner, and then instead of returning to zero they settle down into a finite final value. The net change in the wave field is called the wave's "memory". Discuss the prospects for LIGO to measure such a memory.

Lecture 1
Overview of Gravitational Waves

by Kip S. Thorne, 30 March 1994

This lecture actually consumed one and a half days of the course; lecture 2, just half a day.

The record of this lecture consists of three parts:

- Prose notes by Malik Rakhmanov, based on the first half of Kip's lecture (which was presented at the blackboard).
- Copies of transparencies for the second half of the lecture.
- Appendix: Kip's original handwritten notes for the first half, from which he lectured.

Lecture 1. Overview: [Notes by Malik Rakhmanov with annotations by Kip]

1. Gravity as Space-Time Curvature

a) Newtonian Description of Gravity.

In Newtonian theory the gravitational potential Φ is generated by the mass distribution (mass density ρ)

$$\nabla^2 \Phi = 4\pi G \rho \quad , \quad G - \text{Newton const.}$$

The gravitational force per unit mass due to the Newtonian potential is

$$\vec{g} = -\nabla \Phi .$$

The equivalence principle: the effect of gravity is locally equivalent to inertia forces and thus disappears for a freely falling observer. However, for any two freely falling observers their relative acceleration does not vanish. This is a true effect of gravity. It is given by

$$\partial_k g_j = -\partial_k \partial_j \Phi = -R_{k0j0} ,$$

where $R_{\mu\nu\rho\sigma}$ is Riemann curvature tensor. The components R_{k0j0} are analogons to the components of the electric field, because this part of the gravitational force does not depend on the velocity of the body.

b) Other parts of gravitational forces.

In general relativity, There are other contributions to the grav. force. These are linear and quadratic in the velocity of the body:

$$R_{k0jl} u^l \quad , \quad R_{kmjl} u^m u^l . \quad \text{where } u^l \text{ is the velocity.}$$

The first is the analog of the Lorentz force with R_{k0jl} playing a role of magnetic field. These velocity dependent contributions to the grav. force become important when the velocity of the body is comparable with the speed of light and thus will not concern zo us for LIGO.

2. Gravitational Waves.

a) Transverse traceless tensor. In the weak field approximation the grav. waves are described by symmetric transverse traceless tensor of a second rank $h_{\mu\nu}$:

$$R_{j0k0} = -\frac{1}{2}\ddot{h}_{jk},$$

h_{jk} are components ~~polarizations~~ of the grav. wave propagating with the speed of light. If the wave is propagating in z-direction then the only nonzero components of $h_{\mu\nu}$ are h_{xx}, h_{xy}, h_{yx} and h_{yy} with the following identities:

$$h_{xx} = -h_{yy} \equiv h_+ ,$$
$$h_{xy} = h_{yx} \equiv h_\times .$$

Therefore there are only two independent degrees of freedom (polarizations) denoted by h_+ (h-plus) and $h_\times$ (h-cross).

,) Field Lines, Quadrupolar Forces. Let the test mass be in the $z=0$ plane and have coordinates (x,y). Suppose the grav. wave is passing through the lab. Then the acceleration of the test mass relative to the origin is

$$g_j = \frac{1}{2}\sum_{k=1,2}\ddot{h}_{jk}x^k,$$

or in terms of independent polarizations

$$g_x = \frac{1}{2}(\ddot{h}_+ x + \ddot{h}_\times y),$$
$$g_y = \frac{1}{2}(-\ddot{h}_+ y + \ddot{h}_\times x).$$

Here we assumed.. that the wave is coming from directly above or underneath us. Note that $\vec{g}$ is divergence free. To picture the field lines of $\vec{g}$ let us assume first that $h_\times = 0$. The field lines are shown on Fig. 1.

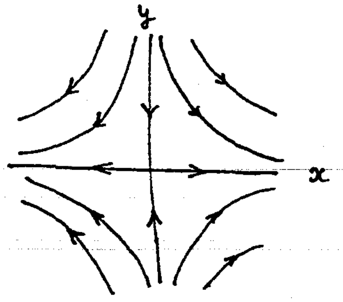


Fig. 1

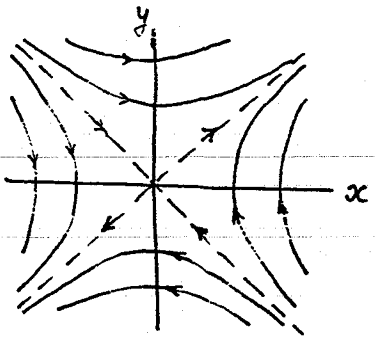
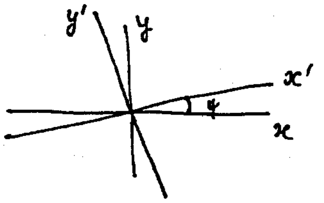


Fig. 2

In the case when $h_+ = 0$ the picture is rotated by 45° (Fig. 2) The functions $h_+ (t - \frac{z}{c})$ and $h_\times (t - \frac{z}{c})$ are called „+ - waveform" and „ x - waveform".

:) <u>Behavior of GW under rotations.</u> In new coordinate frame x' and y', rotated by the angle ψ with respect to the original coordinate frame (x, y) the polarizations are given by

$$h'_+ = h_+ \cos 2\psi + h_\times \sin 2\psi ,$$
$$h'_\times = h_\times \cos 2\psi - h_+ \sin 2\psi .$$



The fact that the angle enters as 2ψ is due to spin 2 of the gravitons that carry the grav. wave. Note that in general:

$$\text{spin} = \frac{2\pi}{\text{Return Angle}} .$$

For photons (vector polarizations) the return angle is 2π and $s=1$. For the waveforms above one can see that the return angle is π, and $s=2$.

d) Newton Potential for GW.

Over regions of space small compared to a wavelength, The grav. waves can be described by Newtonian potential as well. The part of the potential responsible for the gravitational waves is

$$\Phi = -\frac{1}{2}\ddot{h}_+ (x^2-y^2) - \ddot{h}_\times xy .$$

e) Energy Density and Flux

The stress-energy tensor of the GW is obtained by averaging the squared gradient of the wave field over several wavelengths. For the wave propagating in z-direction

$$T_{00} = -T_{0z} = T_{zz} = \frac{c^3}{16\pi G} \left\langle (\dot{h}_+^2 + \dot{h}_\times^2) \right\rangle =$$

$$= 320 \ \frac{\text{erg}}{\text{cm}^2 \ \text{sec}} \left(\frac{f}{1 \ \text{KHz}} \right)^2 \left\langle \frac{h_+^2 + h_\times^2}{(10^{-21})^2} \right\rangle .$$

The numerical factor can also be written as 0.32×10^{26} Jansky. This expression gives us an order of magnitude of $\sim 10^{-9}$ erg/cm² sec estimate for energy flux from an axially symmetric super nova burst at the Virgo Cluster.

3. Laser Interferometer

a) Approximations. The version of the theory of Laser interferometer detectors that we shall use ~~design~~ relies heavily on the following assumptions:

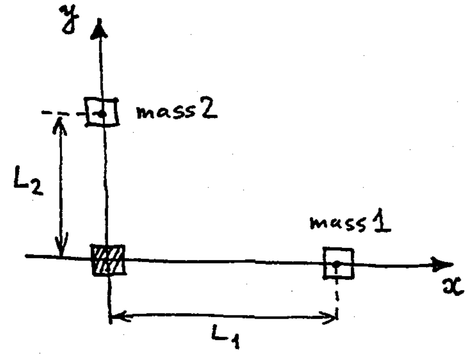
$L \ll \lambda$. That is the size of the detector L should be much smaller than the wavelength of the grav. wave. Typical wavelengths are

$$\lambda = 300 \div 30{,}000 \ \text{km} .$$

For LIGO with the arm length $L = 4$ km this assumption is well satisfied.

$v \ll c$. The velocities of the test masses should be much smaller than the speed of light in order to neglect the velocity dependent terms in the grav. force ($\sim R_{ijk0}, R_{ijkl}$). ~~With this assumption the detector output is simply the wavefront~~.

b) Schematic Picture of the Laser Interferometer.



Suppose the coordinate system is chosen so that one of the test masses (shown in dark) is at the origin and two other masses are placed on the axes. In the absence of GW both arm lengths L_1 and L_2 are equal to L. Assume also that the test masses move freely in the horizontal direction. ~~are freely falling~~, Then for the masses #1 and #2 the equations of motion are

$$\ddot{x}_j = \frac{1}{2} \sum_k \ddot{h}_{jk} x^k .$$

With above approximations integration is easy and yields

$$\Delta x_j = \frac{1}{2} \sum_k h_{jk} x^k .$$

For the special orientation of the interferometer arms (shown above), we obtain

$$\begin{cases} \Delta L_1 = \frac{1}{2} h_+ L_1 , \\ \Delta L_2 = -\frac{1}{2} h_+ L_2 . \end{cases}$$

(One arm is streatched the other is squeezed and vice-versa). The output in LIGO is

$$\Delta L_1 - \Delta L_2 = h_+ L \quad , \quad \text{or}$$

$$\frac{\Delta L}{L} = h_+ \quad , \quad \text{where} \quad \Delta L = \Delta L_1 - \Delta L_2 \, .$$

Thus the output is simply the waveform $h_+(t)$. Of course, the gravity wave may come from an arbitrary direction in the sky (with coordinates θ and φ) and can be polarized along the axes rotated by the angle ψ with respect to the arms. In this case the output is

$$h(t) \equiv \frac{\Delta L(t)}{L} = F_+(\theta,\varphi,\psi)\, h_+(t) + F_\times(\theta,\varphi,\psi)\, h_\times(t) \, ,$$

where
$$F_+ = \frac{1}{2}\left(1 + \cos^2\theta\right)\cos 2\varphi \cos 2\psi - \cos\theta \sin 2\varphi \sin 2\psi \, ,$$

$$F_\times = \frac{1}{2}\left(1 + \cos^2\theta\right)\cos 2\varphi \sin 2\psi + \cos\theta \sin 2\varphi \cos 2\psi \, .$$

Averaging over the whole sky gives $\sqrt{\langle F_+^2 \rangle} = 1/\sqrt{5}$. For unpolarized waves the pattern is proportional to

$$\sqrt{F_+^2 + F_\times^2} = \sqrt{\frac{1}{2}\sin^4\theta \cos^2 2\varphi + 2\cos^2\theta} \, .$$

This antenna beam pattern is shown on the next page.

In reality the test masses are not free but are hung on wires. However, since the swinging frequency of the suspension is of the order of 1 Hz, which is much less than the frequency of the GW (~ 10 Hz ÷ 1 KHz) the restoring force is unimportant and the masses move in the gravitational wave as free masses.

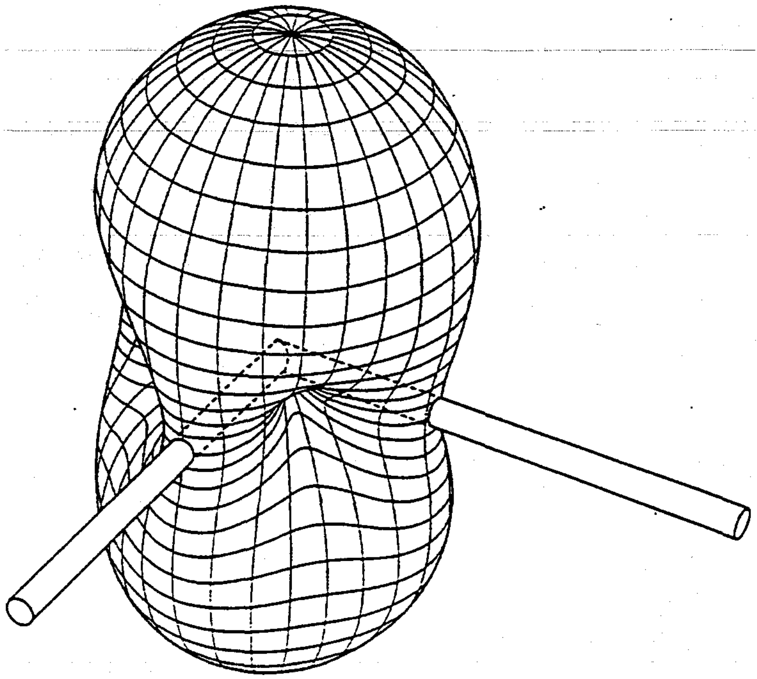


Figure B–2 The angular response pattern of an interferometer with orthogonal arms to unpolarized gravitational radiation. The tubes penetrating the response pattern surface represent the interferometer arms.

[From the 1989 LIGO Construction Proposal]

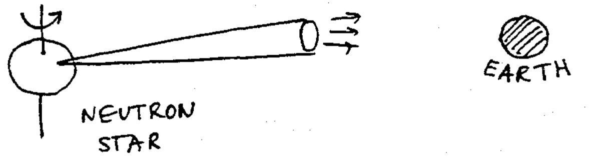
4. Other Kinds of Gravitational Detectors

a) Bar Detectors (pioneered by J. Weber). Unlike the interferometers operating in a broad band of wavelengths the bar detectors operate at the fundamental mode of the bar ($f \sim 1000$ Hz). Coupling of a transducer to the bar determines the bandwidth Δf, typically $\Delta f \ll f$. This is one of the disadvantages of the bar detectors. Another disadvantage of the bars compared to the interferometers is poorer prospects for future sensitivity improvements. ~~sensitivity~~. This is because $\Delta L \sim hL$ and the bars are short.

b) Other possible ways of detecting GW are Earth Normal modes (period of order of 10's of minutes) and Doppler Tracking (period ~ few minutes to hours; in this case the assumption $L \ll \lambda$ is not satisfied).

c) Interferometers in Space. Such detectors (which might fly in ~2010), can operate at the frequencies $10^{-2} \div 10^{-4}$ Hz. For even lower frequencies one has to take into account the R_{oijk} and R_{ijkl} -terms in the gravitational force.

d) Down at still lower frequencies there are two techniques that are impressive: Pulsar Timing (Taylor at Princeton)



~~Passing by~~ Gravitational waves passing through the pulsar cause slowing down and accelerating the NS clock (same for the Earth). The accuracy is 10^{-8} Hz.

e) Anisotropy of CMB (COBE group). Some portion of the anisotropy in the CMB observed by COBE might be due to gravitational waves. ~~due to gravitational waves caused by the primordial distribution of matter was observed~~. The frequency range here is $10^{-16} \div 10^{-18}$ Hz. These are due to ~~the~~ Early Universe Sources (parametric amplification of vacuum fluctuations that emerge from the Planck era, or phase transitions in the vacuum of the early universe) not the black holes and neutron stars.

5. Propagation of the Gravitational Waves

Gravitational waves are described by the symmetric transverse traceless tensor

$$h_{jk} = h_{jk}\left(t - \frac{z}{c}\right).$$

The tensor satisfies the wave equation

$$\Box h_{jk} + \underbrace{2R^{(0)}_{ijmk}\, h^{im}}_{\uparrow\ \text{Negligible when wavelength } \lambda \ll R} = 0$$

where $\Box$ is curved D'Alembertian with respect to background metric $g^{(0)}_{\mu\nu}$ and $R^{(0)}_{\mu\nu\alpha\beta}$ is Riemann curvature tensor associated with $g^{(0)}_{\mu\nu}$.

In the limit of $\lambda \ll R$ (here R is the ~~an instantaneous~~ radius of curvature of the space-time through which the waves propagate), the gravity waves behave like electromagnetic waves, i.e. they show redshifts, deflection (lensing) and etc; and the curvature coupling term is negligible. However the propagation of gravity waves through matter is different from electromagnetic. The gravity waves have negligible sattering, absoption and dispersion. For example, in order to slow down the gravitational wave by one just wavelength one has to create an enormous density of oscillators, e.g. made out of neutron stars, a ~~Such~~ density large ~~would be~~ enough to close the universe.

∂) Wave generation.

Consider the waves from a source, and decompose those waves into contributions from the source's mass multipole moments. A general multipole moment has dimensions $\mathcal{M}_\ell \sim ML^\ell$, where M is the mass and L is the size of ~~an~~ the ~~object~~ source. h_{jk} is dimensionless. Then the dimensional analysis, ~~lead to~~ plus the fact that (by energy conservation) h_{jk} must die out as $1/r$, gives

$$h \sim \sum_\ell \frac{G}{rc^{\ell+2}} \frac{\partial^\ell \mathcal{M}_\ell}{\partial t^\ell} .$$

(distance from source)

For $\ell=0$ $\mathcal{M}_0 \sim M$, but M cannot oscillate and thus there are no gravity waves produces by $\mathcal{M}_0$. Similarly, there are no gravity waves generated by dipole moment, Since ~~Indeed~~, $\dot{\mathcal{M}}_1$ = (momentum of source) = constant. ~~= const~~. Only for $\ell \geqq 2$ there are gravity waves. [In general it is not possible to have radiation in ~~of~~ the field with spin s ~~by~~ from a source with multipole moments $\mathcal{M}_\ell$ with $\ell < s$. This fact is a theorem in canonical field theory.] ~~From dimensio~~

From dimensional analysis we conclude that for quadrupole moment

$$h \sim \frac{G}{rc^4} \frac{ML^2}{T^2} \sim \frac{G}{c^2} \frac{E_{kin}/c^2}{r} ,$$

where E_{kin} is kinetic energy of nonspherical motion. The virial theorem says that E_{kin} is of the same order as the gravitational potential energy. That means that to radiate strongly, the source has to be highly compact. Strong sources of gravitational radiation are black holes and neutron stars. For two neutron stars or black holes inspiraling toward ~~around~~ each other and colliding with E_{kin}/c^2 of the order of a solar mass

$$h \sim \begin{cases} 10^{-23} & \text{at} \quad r \sim 10^{10} \text{ lyrs (Hubble distance)} \\ 10^{-20.5} & \text{at} \quad r \sim 50 \text{ Mlyrs (Virgo Cluster)} \\ 10^{-17.5} & \text{at} \quad r \sim 30 \text{ Klyrs (our galaxy)} \end{cases}$$

Such ~~The~~ events are very-very rare in our galaxy. So ~~the~~ such strong gravitational radiation has to come from somewhere between Virgo and the edge of the Universe. In this case

$$h \lesssim 10^{-21}.$$

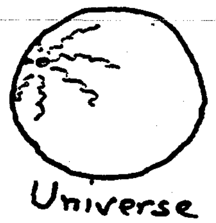
The actual formula for the gravitational radiation from the quadrupole oscillations is the following

$$h_{jk} = \frac{2}{r}\frac{G}{c^4}\overset{\text{mass quadrupole moment}}{\ddot{Q}_{jk}}\left(t - \frac{r}{c}\right), \quad \left(\text{where } Q_{jk} = \int(\rho x^j x^k - \tfrac{1}{3} r^2 \delta^{jk}) d^3x \text{ in Newtonian limit}\right)$$

Then for the radiation energy we obtain:

$$\frac{dE}{dt} = \frac{1}{5}\frac{G}{c^5}\sum_{jk}\left[\dddot{Q}_{jk}\right]^2.$$

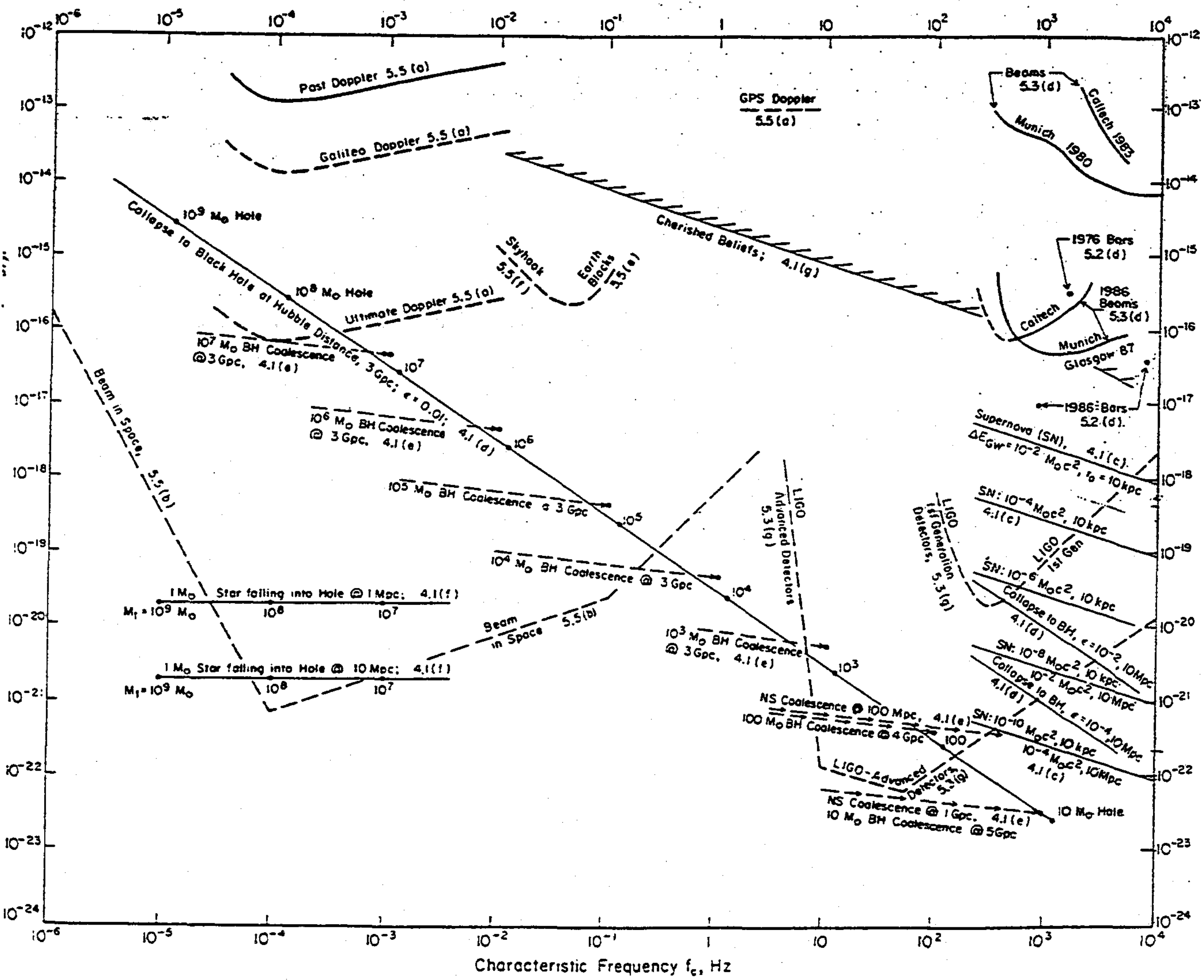
Overview of Grav'l Waves



- Prediction: Poincaré (1905), Einstein (1915)
- Experimental Confirmation:
  Hulse & Taylor
- The most mature detection methods:

f	λ	Method	Sources
$\sim 10^{-16}$ Hz	$\sim 10^{9}$ lt.yrs.	Anisotropy of Cosmic μWave Background Rad'n	Primordial
$\sim 10^{-9}$ Hz	~ 10 lt.yrs.	Timing of millisecond pulsars	Primordial Cosmic strings
$\sim 10^{-4}$ to $\sim 10^{-1}$ Hz	~ 0.01 AU to ~ 10 AU	Doppler tracking of Spacecraft Laser interferometers in space [LISAG ... ESA]	Binary Stars Supermassive BH's (10^{3}-$10^{7} M_{\odot}$) - formation - coalescence - inspiral into
~ 10 to $\sim 10^{3}$ Hz	~ 300 km to $\sim 30{,}000$ km	Laser interferometers on Earth [LIGO/VIRGO]	Inspiral of NS/NS NS/BH & BH/BH binaries (1 to 1000 $M_{\odot}$)
$\sim 10^{3}$ Hz	~ 300 km	Resonant bars	NS & BH coalescence Supernovae Rotating NS's

Fig. 9.4. The characteristic amplitudes h_c (equation (31b)) and frequencies f_c (equation (31a)) of gravitational waves from several postulated *burst sources* (thin curves), and the sensitivities $h_{3/\mathrm{yr}}$ of several existing and planned detectors (thick curves and circles) ($h_{3/\mathrm{yr}}$ is the amplitude h_c of the weakest source that can be detected three times per year with 90% confidence by two identical detectors operating in coincidence). The abbreviations BH, NS and SN are used for black hole, neutron star and supernova. The sources are discussed in detail in the indicated subsections of Section 9.4.1, and the detectors in the indicated subsections of Section 9.5.



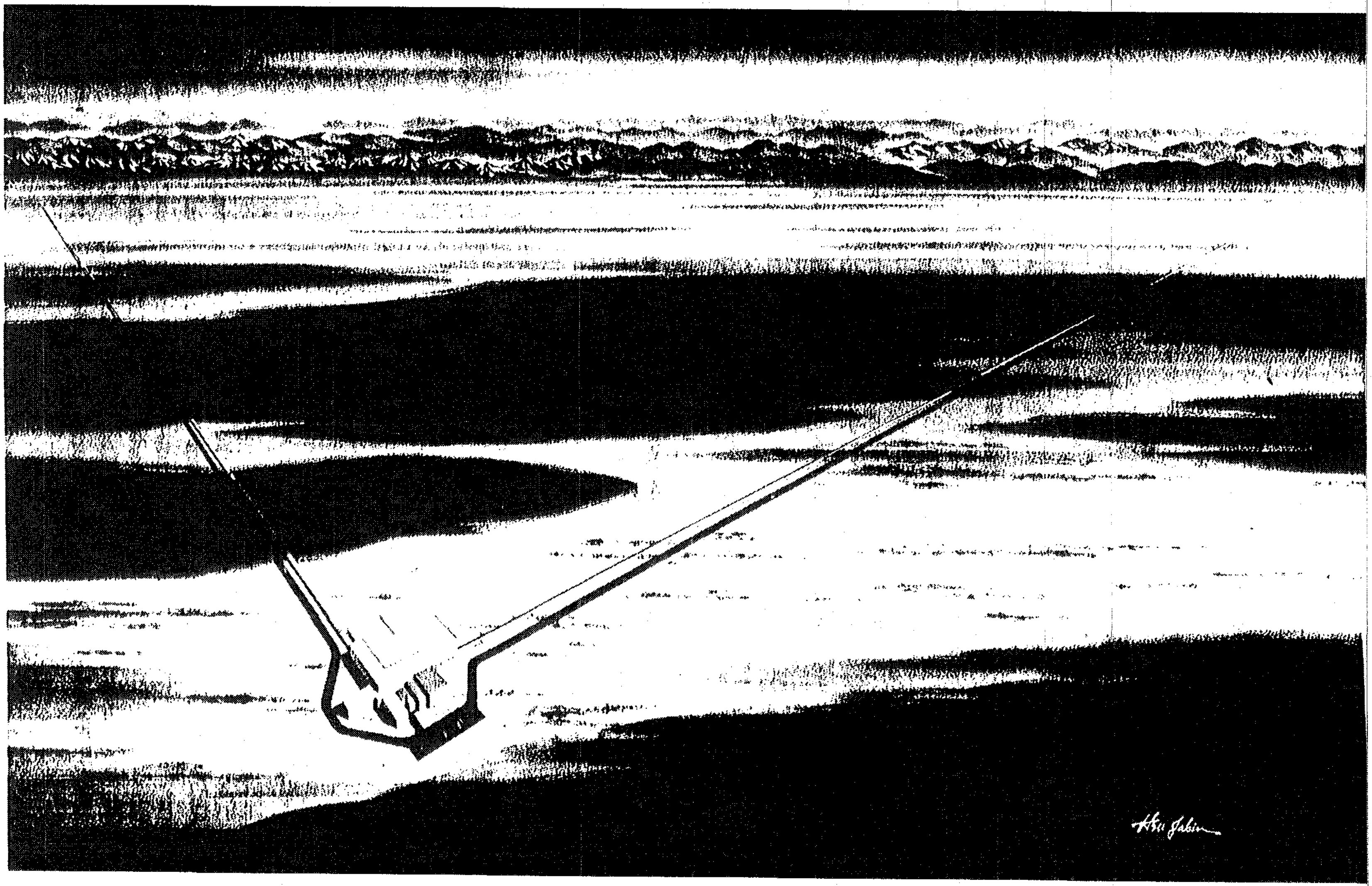
From Ref. A [K.S. Thorne, "Gravitational Radiation", in 300 Years of Gravitation.]

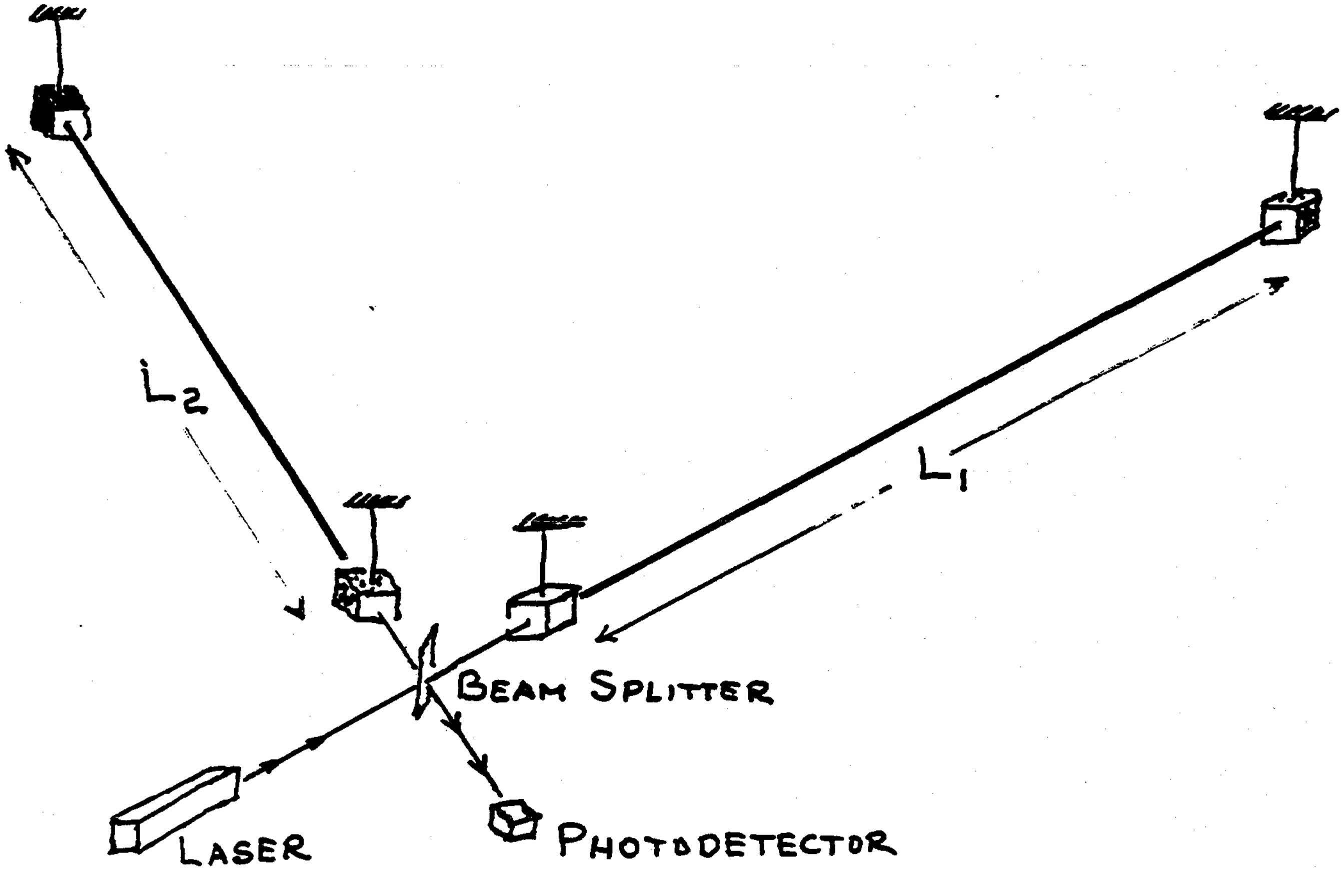
INTERFEROMETER NETWORK [based on 25 years of research & development]

- Facilities to house many successive generations of interferometric detectors
- LIGO [LASER INTERFEROMETER GRAVITATIONAL-WAVE OBSERVATORY] U.S.A. (N.S.F. - Physics)
  - Caltech & MIT [Barish, Vogt, Whitcomb, Weiss, Raab ....~70 scientists & engineers & staff
    - JILA [Faller], Moscow [Braginsky], ... Stanford [Byers], Syracuse [Saulson], ...
    - Many industrial contractors
  - Two Facilities: Livingston, Louisiana
    Hanford, Washington
  - Schedule:
    Begin Construction: 2 weeks ago
    End Construction: ≈1998
    First Interferometers Installed: ≈1999
    First Gravitational-Wave Searches: ≈2000
- VIRGO FRANCE [CNRS], ITALY [INFN]
  - Brillet & Giazotto
  - One Facility: Pisa, Italy
  - Schedule:
    About one year behind LIGO

Angular resolution ~10's of arc min to few degrees; Both waveforms

- Britain & Germany [Hough, Danzman; Hanover]
- Japan
- Australia
- 17/17 COSTS: ≈ $250M LIGO facilities; several $M/interferometer





$$\frac{\Delta L}{L} \equiv \frac{L_1 - L_2}{L} = h(t) = \underbrace{h_+(t)} F_+(\theta,\varphi) + \underbrace{h_\times(t)} F_\times(\theta,\varphi)$$

Gravitational Waveforms

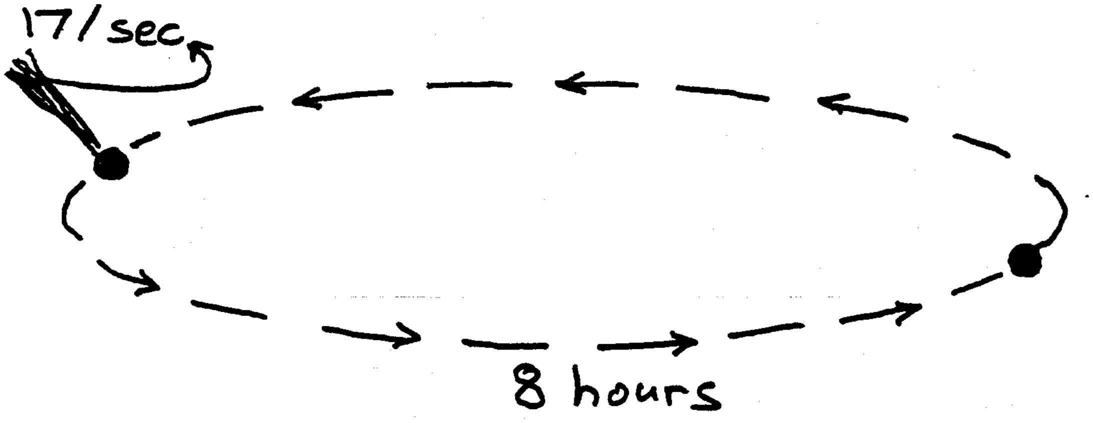
INFORMATION TO BE EXTRACTED FROM WAVES

- Background Issues:
  - Enormous difference between EM & Grav'l Waves
    - emission mechanisms
    - interaction with matter
  - ⇒ Most grav'l sources not seen EM'ly
    - and conversely
    - Potential for great surprises
  - Strongest sources:
    - Extragalactic ... near cosmological
  - No imaging
- Coalescing Binaries [NS/NS, NS/BH, BH/BH]

  The "bread & butter (rice & potato)" sources
  - Map spacetime geometry of black holes
  - Explore nonlinear dynamics of spacetime curvature
  - Measure neutron star masses & radii
    → nuclear equation of state
  - Probe large-scale structure of Universe
- Stochastic Background?
- Pulsars?
- Supernovae?
- The Unexpected

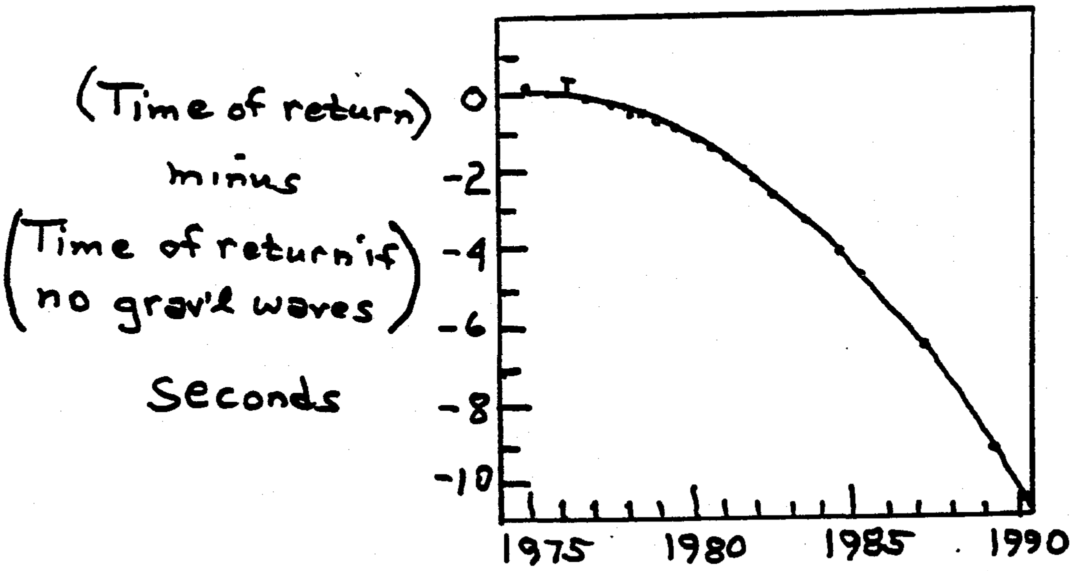
NYC

BINARY NEUTRON STARS



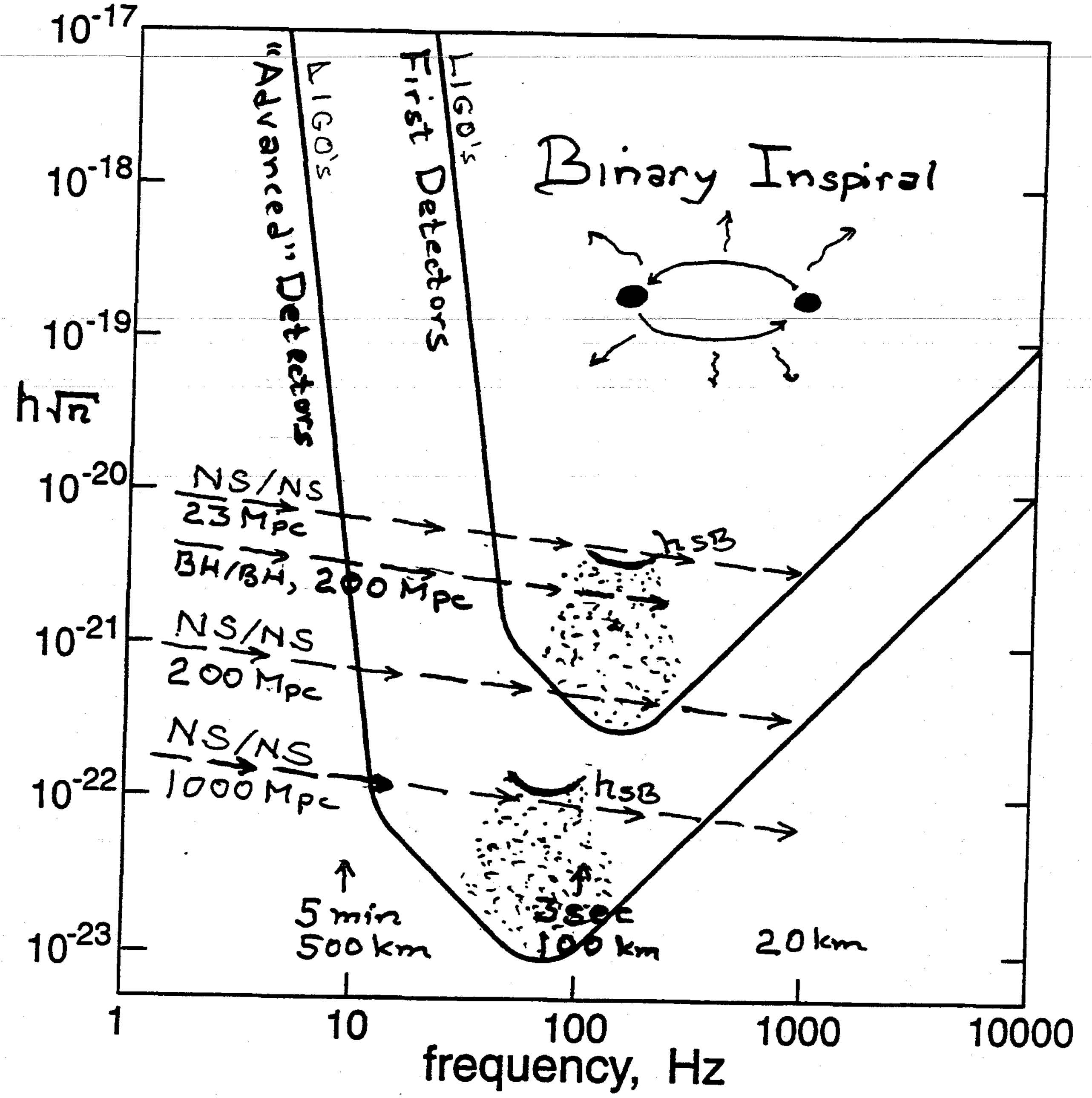
PSR 1913+16 .... RUSSELL HULSE (U. Mass)
JOSEPH TAYLOR (Princeton)

(Indirect but firm observational evidence for grav'l waves)



Time until final coalescence:
~100 million years

LIGO/VIRGO can measure:
Last 15 minutes

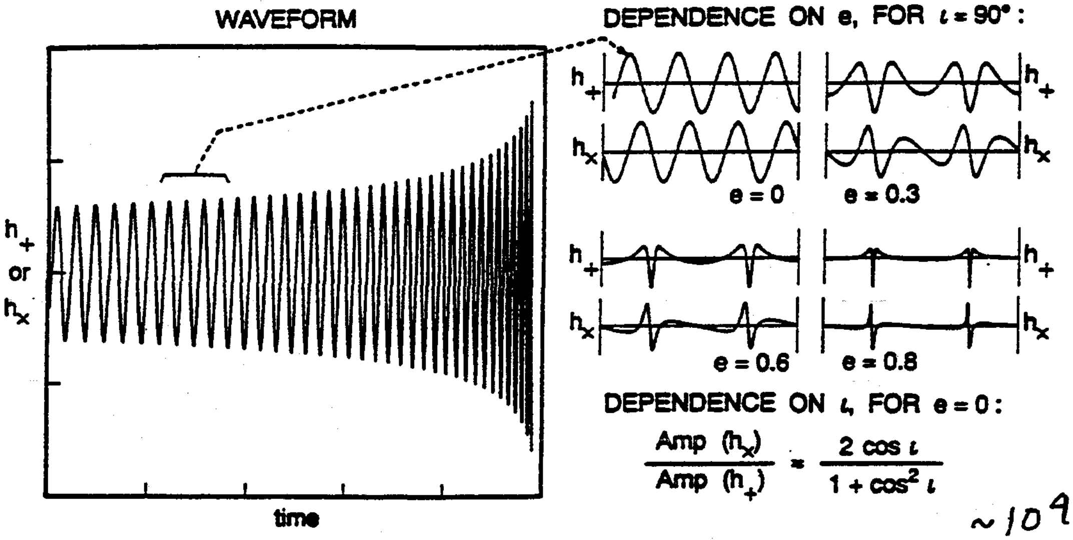


- Rate ∝ $1/h_{SB}^3$
- If rate is 3/yr out to 200 Mpc

Then: "Advanced detectors" should see:

NS/NS: ~1/day out to 1000 Mpc

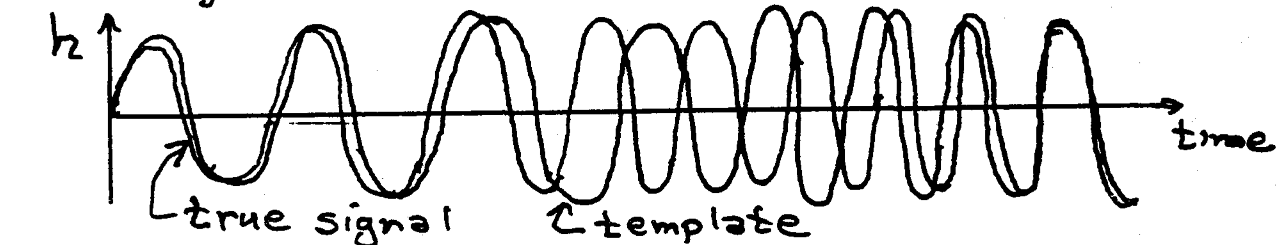
NS/BH & BH/BH: ~ten/day out to cosmological distances

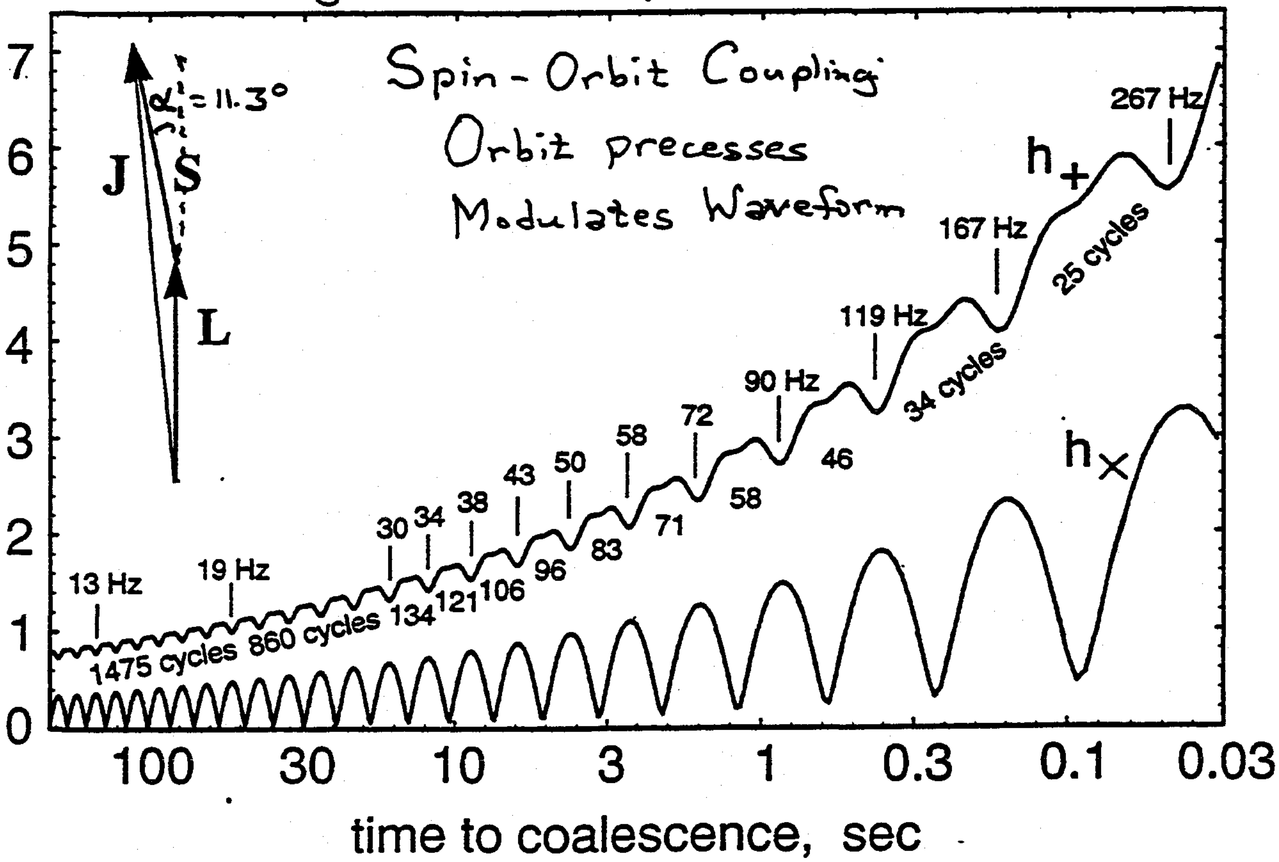


$$\frac{\text{Amp }(h_\times)}{\text{Amp }(h_+)} = \frac{2\cos\iota}{1+\cos^2\iota}$$

~10^4 cycles
~10^5 radians

Data analysis: Matched Filter





APPENDIX: Kip's Personal Lecture Notes for ~~first~~ half of lecture

I. Overview

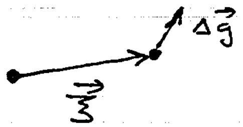
A. Gravity as Spacetime Curvature

1. Throw eraser; jump; earth & balls; nonwalking of inertial frames

2. Newtonian description: $\nabla^2\Phi = 4\pi G\rho$

$$\underset{\sim}{g} = -\underset{\sim}{\nabla}\Phi \quad .. \quad g_j = -\frac{\partial\Phi}{\partial x^j}$$

$$\Delta g_j = -\underbrace{\frac{\partial^2\Phi}{\partial x^j\partial x^k}}_{+R_{j0k0}}\xi^k$$



... "Riemann curvature"

— analog of electric field

~~B.~~ 3. Other pieces: $\left\{\begin{array}{l}\text{gravitomagnetic}\\ \text{spatial curvature}\end{array}\right\}$ — will not concern us — important @ high speeds

B. Gravitational Waves Passing Through Lab

1. Assumptions — under which this is full story:

a. $v/c << 1$

b. $L << \lambda$

2. $$\boxed{\Delta g_j = -R_{j0k0}\xi^k = +\frac{1}{2}\ddot{h}^{TT}_{jk}\xi^k}$$ – No interaction @ light

↑ propagates @ light speed

$$\boxed{\delta\xi_j = \frac{1}{2}h^{TT}_{jk}\xi^k \text{ for free masses}}$$

3. Cartesian coordinates; propagate in Z-direction

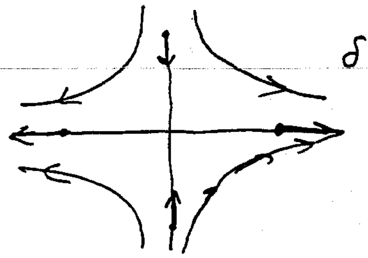
~~a. Lines of force ... relative to some origin (center of coordinates)~~;

~~Δg_x~~ =

~~a. $h^{TT}_{xx} = -h^{TT}_{yy}$; h~~

a. $$\boxed{\begin{array}{l} h_+ \equiv h^{TT}_{xx} = -h^{TT}_{yy} \\ h_\times \equiv h^{TT}_{xy} = h^{TT}_{yx} \end{array}}$$

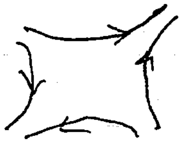
b. $$\boxed{\begin{aligned}\Delta g_x &= \tfrac{1}{2}\ddot{h}_+ \xi_x \\ \Delta g_y &= -\tfrac{1}{2}\ddot{h}_+ \xi_y\end{aligned}}$$



$$\delta\xi_x^{\rm free} = \tfrac{1}{2} h_+ \xi_x$$
$$\delta\xi_y^{\rm free} = -\tfrac{1}{2} h_+ \xi_y$$

Quadrupolar; transverse

c. $$\boxed{\begin{aligned}\Delta g_x &= h_\times \xi_y \\ \Delta g_y &= h_\times \xi_x\end{aligned}}$$

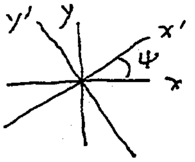


d. $$\left.\begin{aligned} h_+ &= h_+(t - z/c) \\ h_\times &= h_\times(t - z/c)\end{aligned}\right\} \text{waveforms}$$

~~4/ Graviton Spin & Rest Mass~~

5. Behavior under

e. Behavior under Rotation of Axes



$$h_+' + i h_\times' = (h_+ + i h_\times)\, e^{-i2\psi}$$

$$h_+' = h_+ \cos 2\psi + h_\times \sin 2\psi$$

$$h_\times' = h_\times \cos 2\psi - h_+ \sin 2\psi$$

4. Graviton Spin & Rest Mass

5. Newton Potential for GW's

$$\boxed{\begin{aligned}\Phi &= -h^{TT}_{jk} x^j x^k \\ &= -h_+(x^2 - y^2) - h_\times \cdot 2xy\end{aligned}}$$

6. Energy Density & Flux

a. $$T_{00} = \frac{c^3}{G}\frac{1}{16\pi}\langle \dot{h}_+^2 + \dot{h}_\times^2\rangle = 320\,\frac{\rm erg}{\rm cm^2 s}\left\langle \frac{h_+^2 + h_\times^2}{(10^{-21})^2}\right\rangle \left(\frac{f}{1\,\rm kHz}\right)^2$$

$= 0.32\,{\rm W/m^2} = 0.32 \times 10^{26}$ Jansky [SN @ Virgo @ peak of outburst: $\sim 10^{-9}\,{\rm erg/cm^2 s}$]

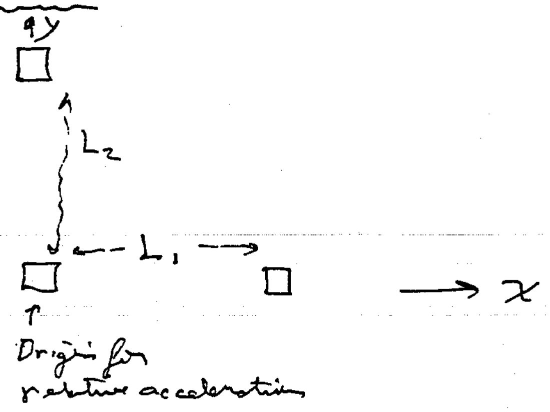
~~C. Interferometric Detectors~~

~~C. Influence of Waves on GW Detectors~~

C. Influence of Waves on Detectors

1. Laser Interferometer:

a. Free test masses



b. Waves from overhead

$$\frac{\Delta L_1}{L} = \tfrac{1}{2}h_+ \quad , \quad \frac{\Delta L_2}{L} = -\tfrac{1}{2}h_+$$

$$\frac{\Delta L}{L} \equiv \frac{\Delta(L_1 - L_2)}{L} = h_+$$

c. Hang masses in earth's field

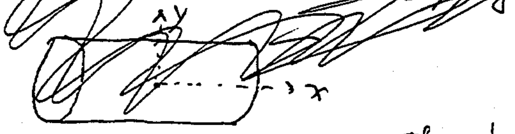
– now have restoring forces

$$\boxed{\begin{array}{l} \text{If} \quad f \gg 1\,\text{Hz} \\ \text{Then} \quad \frac{\Delta L}{L} = h_+ \end{array}}$$

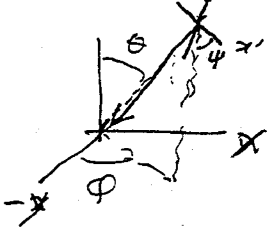
d. Note GW's will have $f < 10^4\,\text{Hz} \Rightarrow \lambda > 30\,\text{km}$; $L = 4\,\text{km}$

so $L < \lambda \Rightarrow$ OK.

~~2. Bar Detector:~~



e. For waves from other direction:



$$\boxed{h(t) \equiv \frac{\Delta L}{L}(t) = F_+ h_+ + F_\times h_\times}$$

$$F_+ = \tfrac{1}{2}(1+\cos^2\theta)\cos 2\phi \cos 2\psi - \cos\theta \sin 2\phi \sin 2\psi$$

$$F_\times = \tfrac{1}{2}(1+\cos^2\theta)\cos 2\phi \sin 2\psi + \cos\theta \sin 2\phi \cos 2\psi$$

"Quadrupolar Beam Pattern"; see page 4a

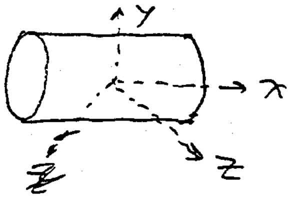
$$\langle F_+^2 \rangle^{1/2} = \frac{1}{\sqrt{5}}$$

f. Show pattern $\sqrt{F_+^2 + F_\times^2}$ for unpolarized waves

g. $f \sim 10$ to 1000 Hz

2. Bar Detectors

a.



b. $\rho \ddot{\xi}_j =$ (elastic restoring force) $+ \frac{1}{2} \ddot{h}^{TT}_{jk} x^k$

c. Drive normal modes

... monitor

d. Narrow band in practice (not in principle)

e. $f \sim 1000$ Hz

~~3? Spacecraft Tracking~~

3. Earth Normal Modes

a. $f \sim$ 10's of minutes

b. poor sensitivity

4. Doppler Tracking — L not $\ll \lambda$

a. f's ~ minutes to hours — noise

5. Intfb in space

a. Ibid

4a

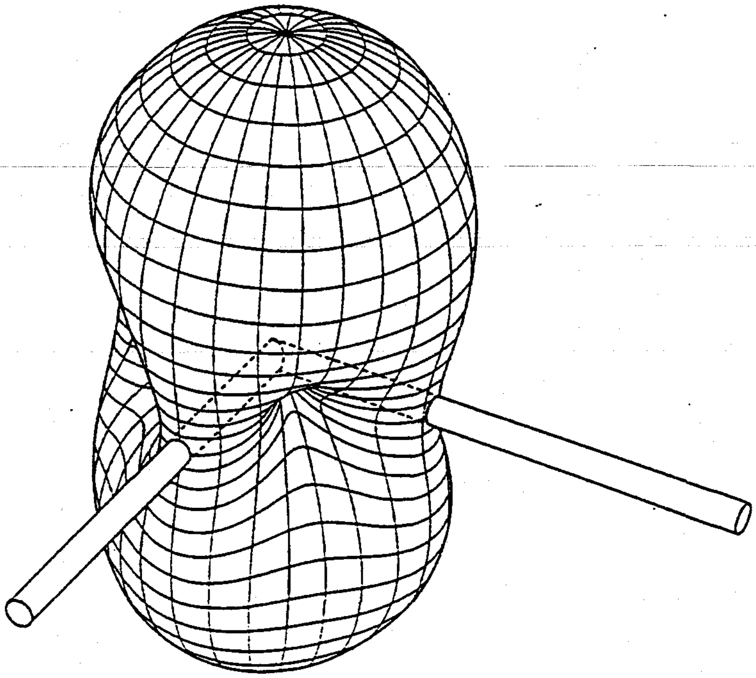


Figure B–2 The angular response pattern of an interferometer with orthogonal arms to unpolarized gravitational radiation. The tubes penetrating the response pattern surface represent the interferometer arms.

The time dependence of the input light at the beam splitter is given by

$$E = E_0 e^{-i\omega t}. \tag{B.18}$$

The two waves leaving the beam splitter are

$$\begin{aligned} E_{10} &= r_s E \qquad \text{the wave launched into arm 1,} \\ E_{20} &= t_s E \qquad \text{the wave launched into arm 2,} \end{aligned} \tag{B.19}$$

where t_s and r_s are the transmission and reflection coefficients of the beam splitter. The two waves next pass through optical phase modulators which are crystals with the property that their optical index of refraction is linearly proportional to an applied modulating field (Pockels effect). The modulating field is chosen to be at a radio frequency (RF), ω_m, sufficiently high that the laser amplitude noise and the noise in the photodetection circuitry at this frequency are close to fundamental limits. The phase modulation adds a time dependent phase to the optical beams given by

$$\phi(t) = \pm\Gamma\sin(\omega_m t), \tag{B.20}$$

6. Pulsar Timing

a. $f \sim 1/10\,\text{yrs}$ to $1/\text{yr}$ $(\sim 10^{-8}\,\text{Hz})$

7. Microwave Anisotropies

a. $f \sim 10^{-16}\,\text{Hz}$ to $10^{-18}\,\text{Hz}$

~~D. Grav'l Wave Generation~~ End of Wed Lecture

cap: Wave ~~field~~ $h^{TT}_{jk} = h^{TT}_{kj}$ … $h^{TT}_{xx} = -h^{TT}_{yy} \equiv h_+(t - z/c)$; $h^{TT}_{xy} = h^{TT}_{yx} \equiv h_\times(t - z/c)$;
not ~~affected~~ by boosts ~~[illegible]~~ L

D. Propagation of Grav'l Waves

1. In brief: (geometric optics)

Same as EM except negligible absorption, scattering, dispersion

2. Example — a universe filled @ NS's with oscillate

~~ℓ = distance for phase shift of $\pi/2$~~

~~R = radius of curvature~~

~~$\frac{\ell}{R} = \begin{pmatrix} 1 \text{ off resonance} \\ \frac{1}{Q} \text{ on resonance} \end{pmatrix} \frac{1}{(n R^3)^{1/2} \left(\frac{GM}{Rc^2}\right)^{1/2} (\omega R)}$~~

3. Grav'l redshift

Grav'l lensing

E. Generation of Grav'l Waves

1. Multipole Expansion:

a. h dimensionless, constructed from $\mathcal{M}_\ell \sim mL^\ell$, G, c, time derivs;

$$\Rightarrow h \sim \overset{1/t}{\frac{\partial^\ell}{\partial t^\ell}} \frac{\overbrace{(G/c^2)\mathcal{M}_\ell}^{L^{\ell}/t^\ell}}{r} \frac{1}{c^\ell} \;\ldots\; h \sim \frac{G}{c^{2+\ell}} \frac{\partial^\ell \mathcal{M}_\ell / \partial t^\ell}{r}$$

b. Monopole — mass — can't oscillate

c. Dipole: $\dot{\vec{D}}$ — momentum — can't oscillate

d. Lowest order: Quadrupole $\left[\begin{array}{c}\text{general thm:}\\ \ell \leq s = 2\end{array}\right]$

$$h \sim \frac{G}{c^4}\,\frac{\ddot{Q}}{r} \sim \frac{G}{c^4}\,\frac{ML^2/T^2}{r}$$

$$\sim \frac{G(E_{ns}^{kin}/c^2)}{r} \cdot \frac{1}{c^2}$$ ... Newtonian potential of nonspherical kin. energy

e. Numbers: ~~fo~~

$$h \sim 10^{-23}\left(\frac{E_{kin}^{ns}/c^2}{M_\odot}\right)\left(\frac{10^{10}\,lyrs}{r}\right)$$

$$\sim 10^{-20.5} \text{ @ Virgo Cluster}$$

$$\sim 10^{-17.5} \text{ @ galactic center}$$

f. Implications:

i. Strongest sources:

- Highly compact
- Strong self gravity
- Short lived

~~S/N~~;
NS's, BH's .... S/N ... binary coalescence

2. Precise Quadrupolar Formulae:

~~a. $\frac{dE_{GW}}{dt}$~~ =

a. $$h_{jk}^{TT} = \frac{2}{r}\,\frac{G}{c^4}\,\frac{\partial^2}{\partial t^2}\left(\mathcal{I}_{jk}\right)^{TT}_{ret}$$

[EM analog: electric dipole radn

$$E_j = \frac{+1}{c^2}\,\frac{\partial^2}{\partial t^2}\left(\frac{D_j^T}{r}\right)_{ret}$$

$$\mathcal{I}_{jk} = \mathcal{J}_{jk} = Q_{jk} = \int \rho\,(x_j x_k - \tfrac{1}{3} r^3 \delta_{jk})\,d^3x$$

TT ≡ "transverse-traceless"

b. $$\frac{dE_{GW}}{dt} = \frac{G}{c^5}\frac{1}{5}\sum_{j,k}\left(\dddot{\mathcal{I}}_{jk}\right)^2$$

nonlinearities

3. For sources of interest, relativistic corrections, can be VIP

— e.g. black-hole collisions ... ρ ~~not~~ even defined !!

— ∃ sophisticated techniques to compute
— PN Expansions
— Numerical relativity,

To VG's —

F. Sources

1. Overview by wave band [VG]

a. High-f [1 to 10^4 Hz] — earth based bars, interf's

~~compact~~

i. 10^4 Hz as upper limit on stellar sources

$$\ldots \quad f \simeq \frac{1}{\pi}\left(\frac{GM}{R^3}\right)^{1/3} \lesssim \frac{1}{\pi}\left(\frac{GM}{[2GM/c^2]^3}\right)^{1/3}$$

↑ BH

@ $M \gtrsim 2M_\odot$ → $f \lesssim 10$ kHz

already done in Wed

ii. Coalescing ~~bin~~ compact binaries @ $M \lesssim$ few 100 $M_\odot$

S/N

Pulsars

Early Universe — { BB, cosmic strings, phase transitions

b. Middle-f [1 to 10^{-5} Hz] — S/c; LISA

i. Supermassive bh's

ii. Ordinary binaries

iii. early universe

c. Low-f [10^{-5} to 10^{-9} Hz] — Pulsar timing

early universe

d. VLF [10^{-9} to 10^{-18} Hz] — Microwave anisotropies

early universe [very impressive]

2. Source / Sensitivity Comparisons [VG's]

~~3. General Remarks About Sources: Compare @ EM~~

~~a. Emission Mechanisms: incoherent / coherent~~

~~b. Absorption & Scattering strong / negligible~~

~~c. Conditions for strong radiation:~~

~~diffuse gas / highly compact, large energy~~

TO VG's

~~d. Implications —~~

- ~~most em not seen grav~~
- ~~most grav not seen em~~
- ~~big surprises [Radio Ay]~~
- ~~poor present knowledge about sources.~~

~~e. Never Image: $\lambda \sim cL/\dot{L} \sim \frac{c}{v} L \gtrsim L$~~

G. Overview of LIGO/VIRGO/network [VG's]

BATCH
START

(2.) Random Processes

STAPLE
OR
DIVIDER

Ph 103c: The Physics of LIGO

1 April 1994

LECTURE 2: RANDOM PROCESSES

Lecture by Kip S. Thorne

Assigned Reading:

D. Pages 5-1 through 5-24 of "Chapter 5. Random Processes" from the textbook manuscript *Applications of Classical Physics* by Roger Blandford and Kip Thorne.

Suggested Supplementary Reading:

a. L. A. Wainstein and V. D. Zubakov, *Extraction of Signals from Noise* (Prentice Hall, London, 1962; Dover, New York, 1970). [This wonderful book—a sort of biblical primer on the subject—is long since out of print. Kip will put his personal xerox copy on reserve in Millikan Library for a few weeks, along with the library's only copy.]

Two Suggested Problems from Blandford and Thorne's "Chapter 5, Random Processes":

5.1 *Bandwidths of a finite-Fourier-transform filter and an averaging filter* [page 5-21]

5.2 *Wiener's Optimal Filter* [page 5-22]. This is an especially important exercise, since the optimal filter underlies much of the data analysis to be done in LIGO.

Lecture 2
Random Processes

by Kip S. Thorne, 1 April 1994

'his lecture actually consumed only half of the 90 minutes on 1 April; the completion of ecture 1 consumed the other half.

'his lecture was largely just a blackboard presentation of the key issues in Reference D ›ages 5-1 through 5-24 of "Chapter 5. Random Processes" from the textbook manuscript .*pplications of Classical Physics* by Roger Blandford and Kip Thorne]. Since that reference . included in Volume II, we here present, as a record of Lecture 2, only the scrawled notes 'om which Kip lectured.

Lecture 2 Random Processes Tutorial · 1/4 ~~3/31/94~~ 4/1/94 1

1. Examples of RP ~~[illegible]~~ $I_{PD}(t)$; $h(t) = C \cdot I_{PD}(t)$; $x(t)$ test mass

~~$\Delta L(t) = L\,h(t)$~~ [Mean Removed]

2. Sp Noise spectrum — $\boxed{\text{VG show}}$ — what means?

~~Show~~ ↑ $h(f) \equiv \sqrt{S_h(f)}$

Like FT: $\tilde{h}'(f) = \int_{-\infty}^{+\infty} h(t)\, e^{i2\pi f t}\, dt$

... no ... diverged; complex

Try $\lim_{T\to\infty} \frac{1}{T}\left|\int_{-T/2}^{+T/2} h(t)\, e^{i2\pi f t}\, dt\right|^2$ Same f & -f

→ fold over

→ $G_h(f) \equiv S_h(f) \equiv [\tilde{h}(f)]^2 = \lim_{T\to\infty} \frac{2}{T}\left|\int_{-T/2}^{T/2} h(t)\, e^{i2\pi f t} dt\right|$

[For Em wave $G_E(f) = \frac{4\pi}{c}\frac{dE}{dA\,dt\,df}$] Units: strain/$\sqrt{Hz}$

3. Correlation function

$$C_h(\tau) \equiv \overline{h(t)\,h(t+\tau)} = \lim_{T\to\infty} \frac{1}{T}\int_{T/2}^{T/2} h(t)\, h(t+\tau)\, d\tau$$

4. Wiener-Khintchine Thm

$$C_h(\tau) = \int_0^\infty G_h(f)\cos 2\pi f\tau\, df$$

$$G_h(f) = 4\int_0^\infty C_h(\tau)\cos 2\pi f\tau\, df$$

5. Variance:

$$C_h(0) = \int_0^\infty G_h(f)\, df = \overline{[h(t)]^2} = \sigma_h^2$$

— but might not converge @ low f.

6. Filtering

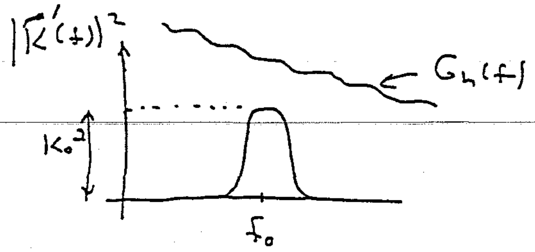
a. $H(t) \equiv \int_{-\infty}^{+\infty} K(t-t')\, h(t')\, dt'$

b. If were ~~not~~ confined to finite time: $\tilde{H}'(f) = \tilde{K}'(f)\, h'(f)$

c For RP $G_H(f) = \underbrace{|\tilde{K}'(f)|^2}_{\text{converges}} G_h(f)$

d.

Band Pass Filter



~~$G_H(f)$~~

$$\sigma_H^2(f) = \int ~~K_0~~ |\tilde{K}'(f)|^2 \underbrace{G_h(f)}_{\text{cnst}}\, df = G_h(f_0) \underbrace{\int |\tilde{K}'(f)|^2\, df}_{K_0^2 \Delta f}$$

$$\sigma_H^2(f) = \underbrace{[G_h(f_0) \cdot \Delta f]}_{\text{rms fluctuations of } h \text{ in bandwidth } f_0} \cdot K_0^2$$

a. Simple example of Bandpass filter:

$$H(t) = \int_{t-\Delta t}^{t} \cos[2\pi f_0 (t-t')]\, y(t')\, dt'$$

$$\boxed{\Delta f = 1/\Delta t}$$

e.

xample: A line spike in spectrum ... large rms fluctuations near that f.

Wiener Optimal Filter:

a. $h(t) = \overset{}{\underset{\uparrow \text{unknown}}{A}} S(t) + n(t)$ .... want to find ~~S(t)~~ out if $S(t)$ there and if so how strong, A

Best way should be to cross correlate: $\int h(t) \cdot S(t)\, dt$

– Better: Suppress ~~comp~~ frequencies where detector is noisy .... $\tilde{S}_F^*(f) = \dfrac{\tilde{S}'(f)}{G_h(f)}$ .. then cross correlate:

$$W = \int h(t)\, S_F(t)\, dt = \int \frac{\tilde{h}'(f)\, \tilde{S}'^*(f)\, df}{G_h(f)}$$

b. Signal contribution to W is S; noise N is random @ some mean $\bar{N}$: $W = S+N$.

$$\boxed{\frac{S}{\bar{N}} = 4\int_0^\infty \frac{|\tilde{S}(f)|^2}{G_h(f)}\,df}$$

11. Knowing spectrum does not tell us $P(N)$, the probability distn of N. But if we know it is Gaussian, then

~~$P(N) = \frac{1}{\sqrt{2\pi\bar{N}}}\exp\left[\frac{(N-\bar{N})^2}{2\bar{N}}\right]$~~

$$\boxed{P(N) = \frac{1}{\sqrt{2\pi\bar{N}^2}}\exp\left[-\frac{N^2}{2\bar{N}^2}\right]}$$

~~Key issue in LIGO~~ is

12. If instrument clean, ~~then~~ enough [not likely], noise is superposition of tiny influences of many different things

Central limit thm ⇒ Gaussian

13. Central issue will be: Gaussian or not?

BATCH START

Signal Processing

STAPLE
OR
DIVIDER

LECTURE 3: SIGNAL PROCESSING

Lecture by Eanna E. Flanagan

Assigned Reading:

A. "Gravitational Radiation" by Kip S. Thorne, in *300 Years of Gravitation*, eds. S. W. Hawking and W. Israel (Cambridge University Press, 1987), pages 366–371; 385–386; 393–395. [This is the review article handed out last Friday. The assigned sections outline the data processing methods for detecting burst, periodic and stochastic gravitational waves.]

E. "Data Processing, Analysis and Storage for Interferometric Antennas", B.F. Schutz, in "The Detection of Gravitational radiation", edited by D. Blair, (Cambridge 1989) pp. 406–416; 420–422; 428-429; 445–447. [To be handed out on Friday]

Suggested Supplementary Reading:

A. The remainder of Ref. [A] above.

F. "The Last Three Minutes: Issues in Gravitational Wave Measurements of Coalescing Compact Binaries", C. Cutler *et al*, Phys. Rev. Lett. **70**, 2984 (1993). [This is a overview of what is understood to date about the potential for extracting useful information from detected binary inspiral waveforms.]

b. E. S. Phinney, Astrophys. J. **380**, L17 (1991). [This article gives the most up-to-date estimates of the rate of binary neutron star inspirals in the Universe, and discusses in detail the astronomical observations that underlie these estimates.]

c. "Near optimal solution to the inverse problem for gravitational wave bursts", Y. Gursel and M. Tinto, Phys. Rev. D 40, 3884 (1989). [This article describes how best to reconstruct the gravitational waveforms $h_+(t)$ and $h_\times(t)$ for a detected burst of unknown form, from the (noisy) outputs of 3 interferometers.]

d. S. Smith, PhD thesis, Caltech (1987). [A description of the last real search for gravitational waves using data from the 40m prototype interferometer.]

A Few Suggested Problems

1. *The detectability of neutron star – neutron star inspirals at 1000 Mpc by LIGO.*

   a. Suppose that the burst of gravitational waves produced by a neutron star – neutron star inspiral at 1000 Mpc passes through the Earth. Calculate from the following foundations the signal-to-noise ratio obtained after optimal signal processing by one of the two LIGO interferometers: Assume that the "advanced detector" sensitivity benchmark given in Ref. [B] of lecture 1 has been achieved. Approximate this noise curve by the formula

$$S_h(f) = \begin{cases} (h_m^2/f_m)(f/f_m)^2 & f \geq f_m \\ (h_m^2/f_m)(f/f_m)^{-4} & f < f_m, \end{cases}$$

where $h_m = 1.0 \times 10^{-23}$ and $f_m = 70$Hz. Use the waveform $h(t)$ given in Eqs. (26), (42) and (104) of Ref. [A], also Eq. (29) (with RHS multiplied by a correction factor of 2), and assume that the waves' polarization and the relative orientation of the binary and of the interferometer are such that the signal-to-noise is maximized. Assume both neutron stars have masses of $1.4M_\odot$. [*Hint:* Use the stationary phase approximation to evaluate the Fourier transform].

b. The current best estimate of the neutron star – neutron star merger rate, inferred from the statistics of observed neutron star binaries in our own galaxy, is that there should be 3 per year within a distance of 200 Mpc (uncertain to within a factor ~ 2 in the distance). Assuming this merger rate, estimate to within a factor ~ 2 the number of inspiral events per year that will produce a signal-to-noise ratio ≥ 6 in each of LIGO's two interferometers (which is roughly the criterion for successful detection), if the advanced detector sensitivity levels are achieved.

2. *The shapes of Wiener optimal filters in the time domain.*

   a. For the waveform $h(t)$ and noise spectrum $S_h(f)$ of question 1, numerically calculate and plot the optimal filter $K(t)$. Compare it's shape to that of the original waveform.

   b. A chance near-collision of two neutron stars in which their gravitational attraction substantially alters their velocities will produce "gravitational bremsstrahlung" or braking radiation. These waves will have a *memory*: the test masses in a nearby detector would be left with a permanent relative displacement following the waves passage, corresponding to a nonzero final value of $h(t)$. Approximate the memory part of the waveform by a step function, and numerically calculate and plot the optimal filter for the memory $K(t)$. Compare it's shape to that of the original waveform.

3. *Prospects for observing the violent final stages of black-hole – black hole mergers.* One of the aims of LIGO is to measure waves produced by the highly nonlinear dynamics in the final stages of black hole – black hole mergers. Such measurements, if they agree with supercomputer simulations, would convincingly demonstrate the existence of black holes and for the first time experimentally probe general relativity in the highly nonlinear regime. A major effort (called the Grand Challenge project) is currently underway in the numerical relativity community to calculate the waveforms; the task is expected to take several years.

   a. Suppose that a pair of rapidly spinning, $15M_\odot$ black holes coalesce at a cosmological redshift of $z\Delta\lambda/\lambda = 1$. Assume that the final plunge and coalescence of the holes (after the gradual inspiral) radiates 5% of the total mass-energy of the system into gravitational waves, and that this energy is uniformly distributed in frequency between ~ 150 Hz and ~ 350 Hz (the latter being frequency of the lowest quasinormal mode of the final $\sim 30M_\odot$ black hole). Use Eq. (35) of Ref. [A] to estimate the signal-to-noise ratio in one of the LIGO interferometers after optimal filtering, assuming the noise spectrum of question 1. Note that the effect of redshift on energy flux is exactly the same for gravitational waves as for electromagnetic waves, and also that the energy will be distributed between ~ 75

Hz and $\sim$ 175 Hz as measured at the detector. Assume a Hubble constant of 75 $\mathrm{km\,s^{-1}\,Mpc^{-1}}$ so that the luminosity distance is 4.7 Gpc.

b. What is the signal-to-noise squared per cycle if the waveform contains 10 cycles? What are the prospects for measuring the detailed shape of the waveform?

4. *Phase incoherence effects in searches for a gravitational wave stochastic background* The method outlined in the lecture for searching for a stochastic background by cross correlating the outputs of two interferometers assumed that the both interferometers respond to the same gravitational wave signal $h(t)$. The two LIGO detectors will be approximately parallel, but will be separated by $\sim$ 3000 km. For what range of gravitational wave frequencies will the phase lag between the detectors be small compared to unity for all propagation directions? Will these phase lags necessitate a modification of the search algorithm outlined in the lecture?

Lecture 3
Signal Processing in LIGO and in Prototype Interferometers

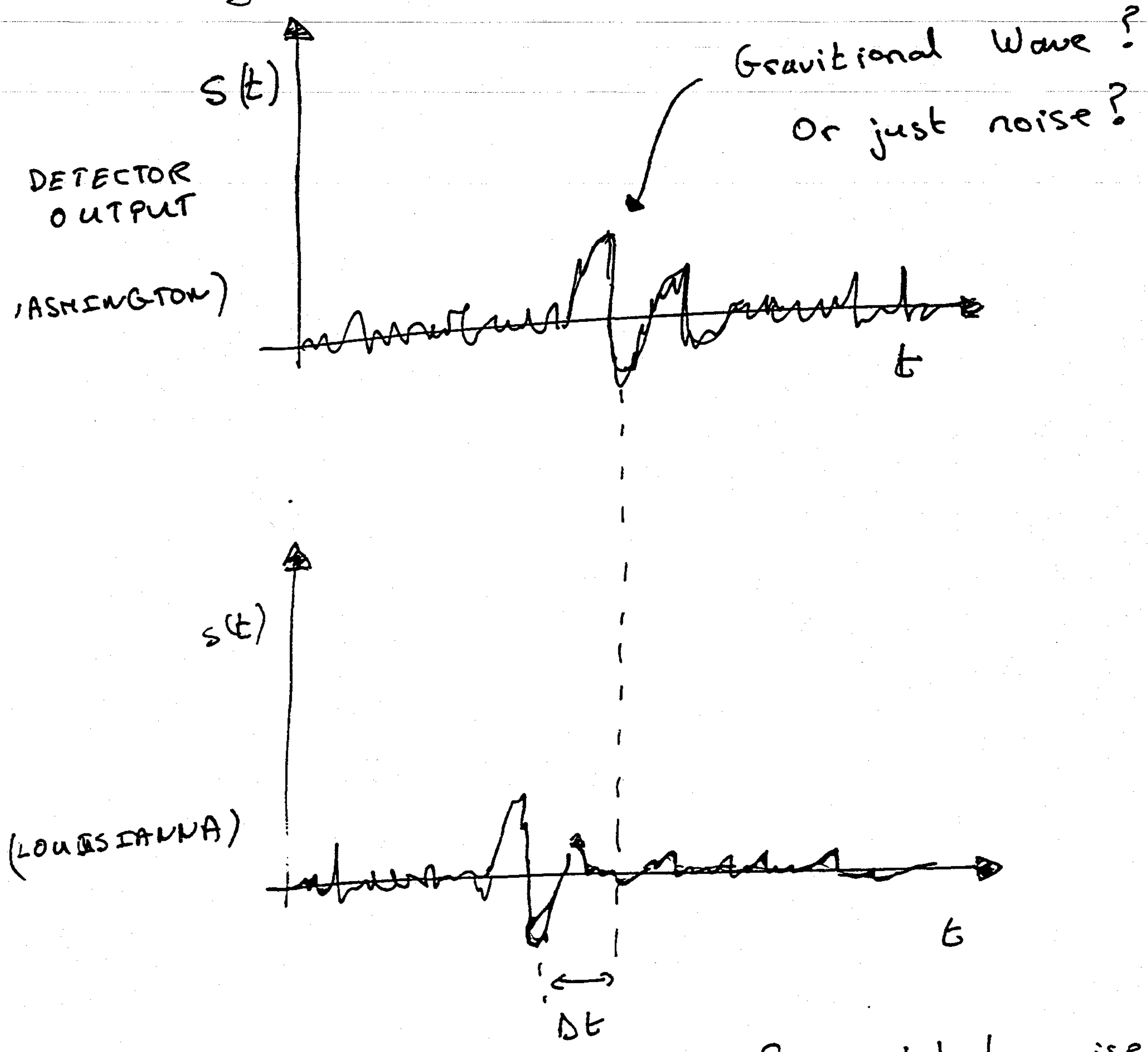
by Eanna E. Flanagan, 6 April 1994

Flanagan lectured from the transparencies that follow...

Ph 103c

LECTURE 3
SIGNAL PROCESSING IN GRAVITATIONAL WAVE DETECTORS

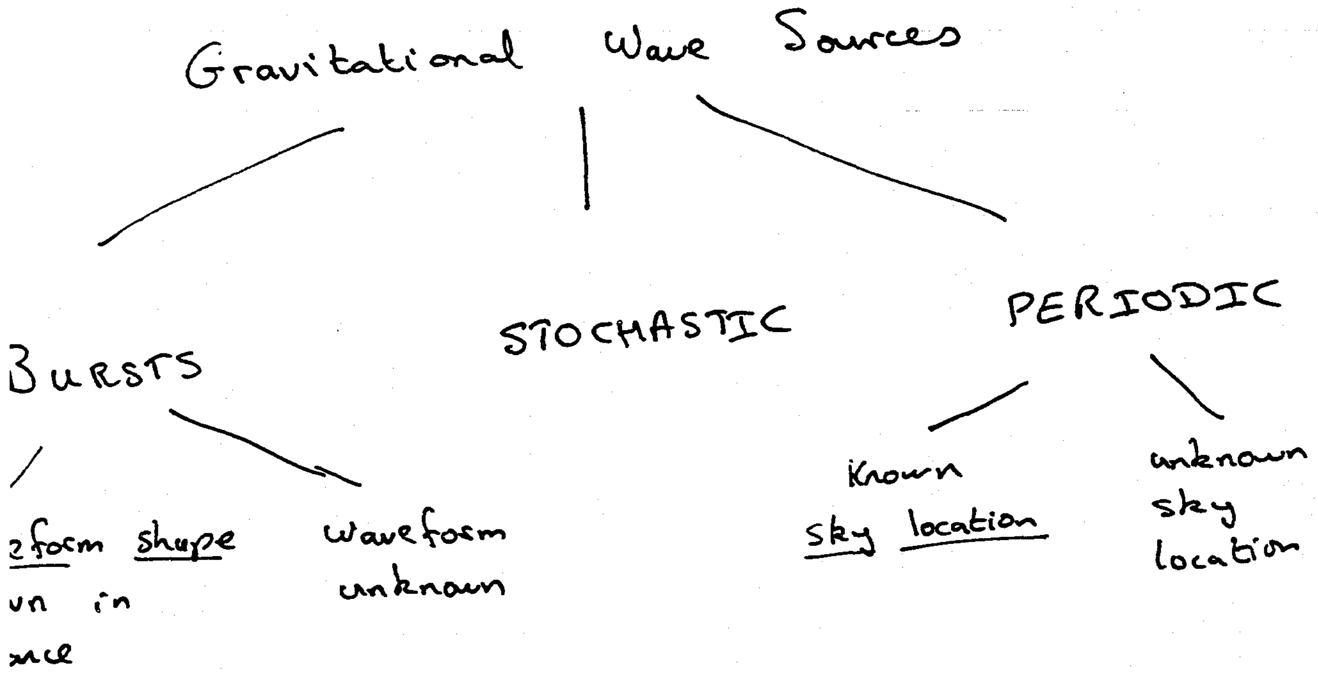
by EANNA E. FLANAGAN, 6 April 1994



Gravitional Wave? Or just noise?

- How often would this occur from detector noise?
- How likely is it to be a real signal?
- What threshold criteria should be used?

OVERVIEW OF LECTURE :



For each type of wave

- How detect ?
- Prospects for detection ?

- What useful information (physics & astrophysics) do the waves bring ?
- How do we extract the information ?
- How well can we hope to do ?

INFORMATION CARRIED BY GRAVITATIONAL WAVES

- Limits in a world without detector noise??

eminders:

EM astronomy: Basic observable is

$$I_\nu = \frac{dE}{dA\, dt\, df}$$

← Photons typically arrive phase incoherently.

GW astronomy:

Gravitons phase coherent.

Get $I_\nu \propto |\tilde{h}_+(f)|^2 + |\tilde{h}_\times(f)|^2$

but <u>also</u> $h_+(t)$, $h_\times(t)$.

Extraction of waveforms

$$h_1(t) = F_{1+}(\theta,\varphi)\, h_+(t) + F_{1\times}(\theta,\varphi)\, h_\times(t)$$

$$h_2(t) = F_{2+}(\theta,\varphi)\, h_+(t-\tau_{12}) + F_{2\times}(\theta,\varphi)\, h_\times(t-\tau_{12})$$

$$h_3(t) = F_{3+}(\theta,\varphi)\, h_+(t-\tau_{13}) + F_{3\times}(\theta,\varphi)\, h_\times(t-\tau_{13})$$

- Solve for $\tau_{12}(\theta,\varphi)$, $\tau_{13}(\theta,\varphi)$ and thus h_+, $h_\times$.

- Cannot "aim" at portion of sky ... must respond to waves from all directions.

- Ultimately limited by <u>background gravitational wave noise</u> ("Stochastic background"; analog of microwave background).

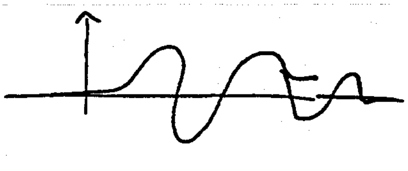
Expected to be several orders of magnitude smaller than detector noise for LIGO (although is limiting factor for space-based detector systems).

Some Reminders:

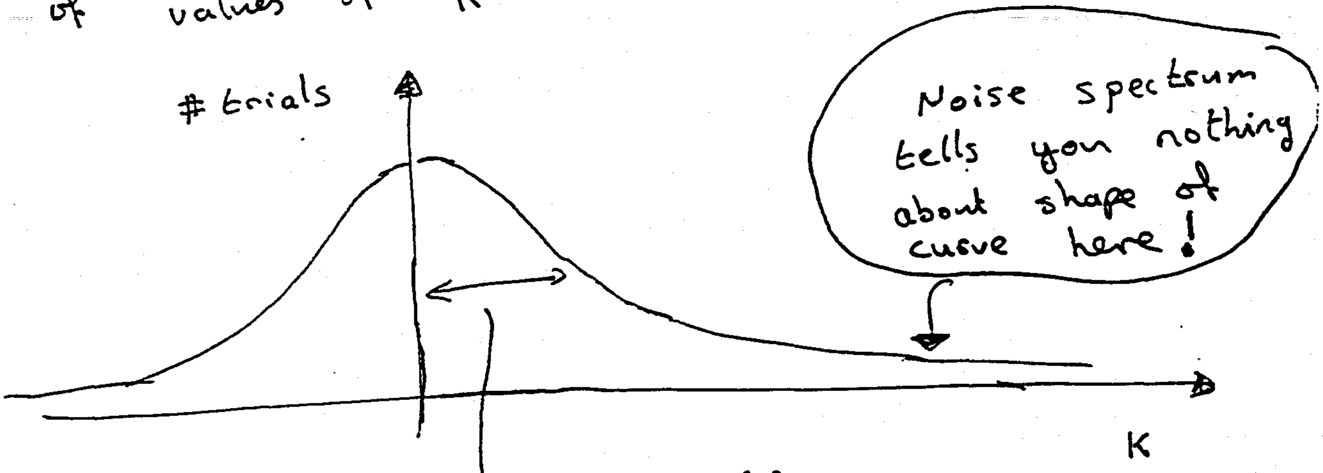
- Detector output: $h(t) = \underbrace{n(t)}_{\text{noise}} + \underbrace{s(t)}_{\text{G.W. signal}}$

Meaning of noise spectrum $G_h(f) = \tilde{h}(f)^2$:

- Pick any function $K(t)$, eg.



- Work out $K = \int dt\, K(t)\, n(t)$
- Repeat many times and plot distribution of values of K



$(\text{rms value of } K)^2$

$= (\text{width})^2$

$$= \overline{K^2} = \int_0^\infty G_h(f)\, |\tilde{K}(f)|^2\, df$$

Reminders ...

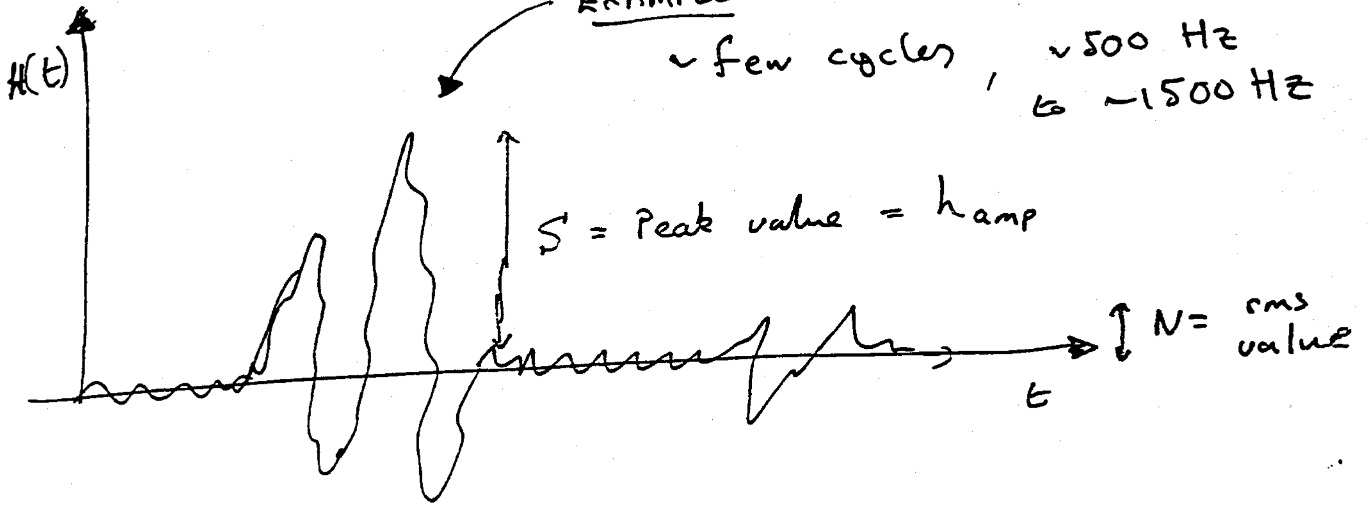
Linear filtering:

$$\underbrace{H(t)}\ =\ \int_{-\infty}^{t} dt'\ K(t-t')\ h(t')$$

New, filtered output has spectrum $G_H(f) = G_n(f)\,|\check{K}(f)|^2$

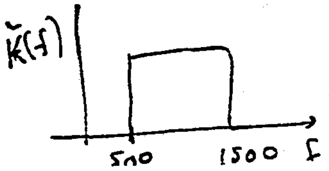
How does filtering help?

Case 1: Signals of unknown shape.

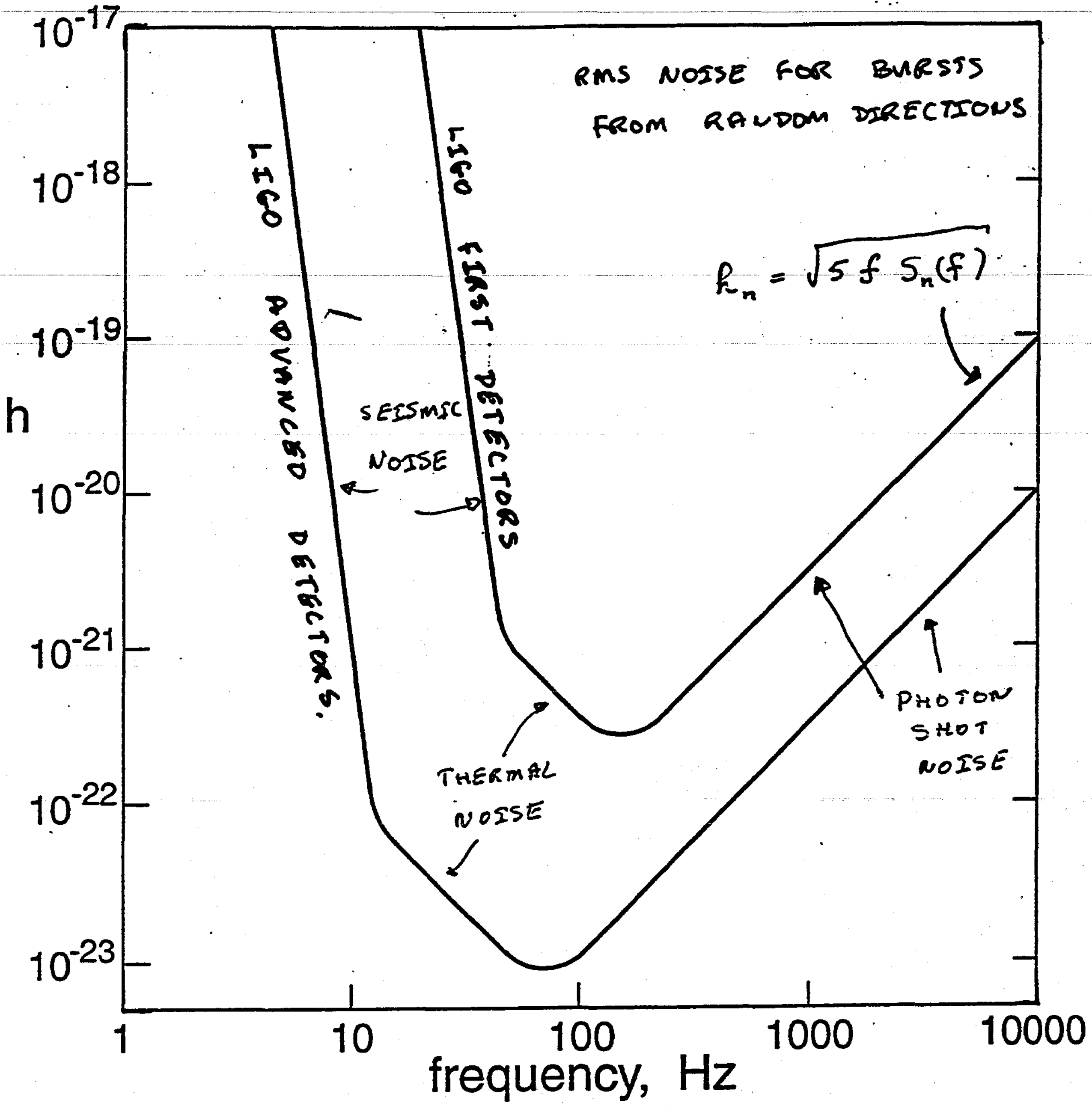


EXAMPLE: Waves from supernovae, ~ few cycles, ~500 Hz to ~1500 Hz

- After filtering with bandpass filter with $\Delta f \sim f$, find $\boxed{\dfrac{S}{N} \simeq \dfrac{h_{amp}}{\sqrt{G_n(f)\ f}}}$



LIGO NOISE SPECTRUM.

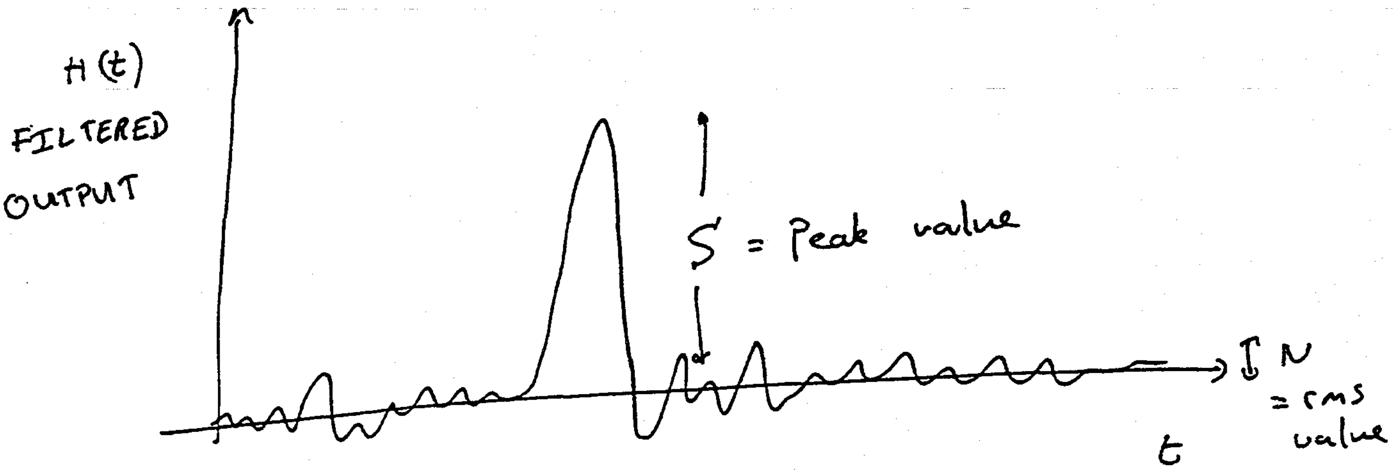


$$h_{out}(t) = h_{true}(t) + n(t)$$

$$\Delta n(f_0, \Delta f)_{rms} = \sqrt{S_n(f_0)\,\Delta f}$$

Case 2: Waveform Shape known:

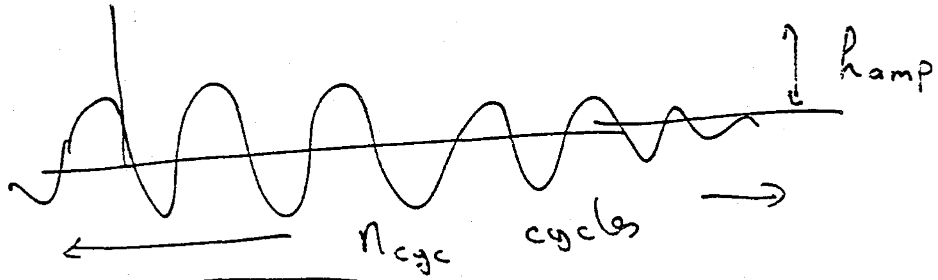
- Suppose a signal is present $h(t) = n(t) + s(t)$
  Apply some filter $\tilde{H}(f) = \tilde{K}(f)\,\tilde{h}(f)$



$\frac{S}{N}$ depends on shape of filter $K(t)$.

Optimum is $\underbrace{\boxed{\tilde{K}(f) = \frac{\tilde{s}(f)}{G_n(f)}}}_{\text{WIENER OPTIMAL FILTER.}}$, then $\underbrace{\boxed{\frac{S^2}{N^2} = 4\int_0^{\infty} \frac{|\tilde{s}(f)|^2}{G_n(f)}\,df}}_{\text{SIGNAL-TO-NOISE RATIO}}$

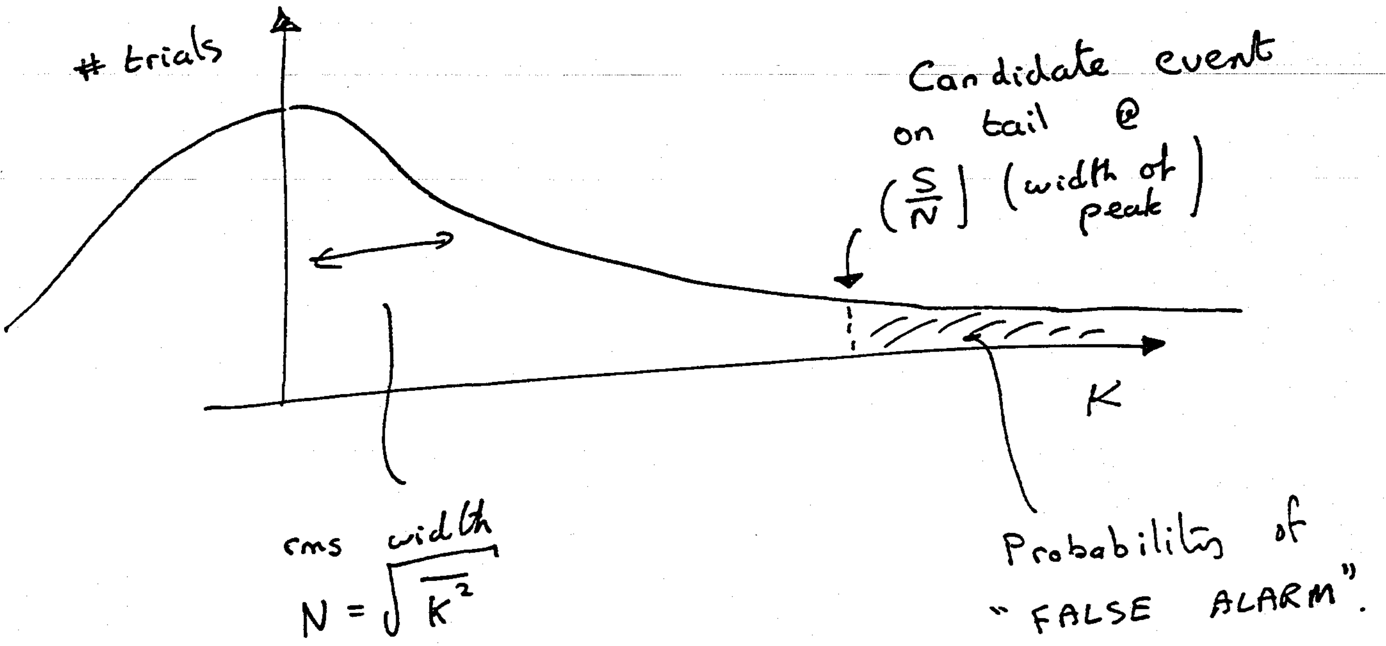
- If $s(t) =$



Then $\frac{S}{N} \simeq \sqrt{n_{cyc}}\ \frac{h_{amp}}{}$

Meaning of $\frac{S}{N}$, and thresholds.

Recall.. $$K = \int K(t)\, n(t)\, dt$$

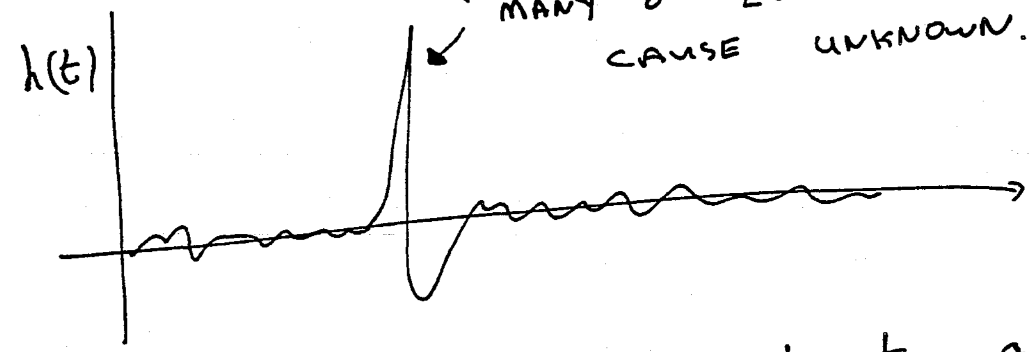


- If noise is Gaussian, then probability $= \sqrt{\frac{2}{\pi}} \int_{\frac{S}{N}}^{\infty} e^{-\frac{1}{2}x^2}\, dx \simeq \sqrt{\frac{2}{\pi}} \left(\frac{S}{N}\right)^{-1} e^{-\frac{1}{2}(S/N)^2}$

- Example : 1 False alarm / year @ 1 kHz sampling rate $\Rightarrow \left(\frac{S}{N}\right)_{\text{THRESHOLD}} \simeq 6.6$

- Key Issue : How non-Gaussian is the noise?

NON GAUSSIAN NOISE



"MANY σ" EVENTS FREQUENTLY SEEN, CAUSE UNKNOWN.

- LIGO has been designed to get around this problem.

Combating Non Gaussian Noise:

① COINCIDENCE TESTING

Example: 100 Hz sampling
1 Glitch / minute in each of 3 detectors

$\Rightarrow P_{\text{GLITCH, 1 DETECTOR}} \sim 3\times10^{-4}$

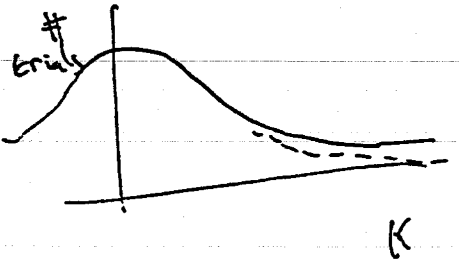
$P_{\text{GLITCH, 3 DETECTORS}} \sim 3\times10^{-10}$

$\Rightarrow$ FALSE ALARM RATE ~ 3×10^{7} s
~ 1 YEAR.

- TWO WIDELY SEPARATED DETECTORS
- MIDSTATIONS
- PARALLEL DETECTORS.

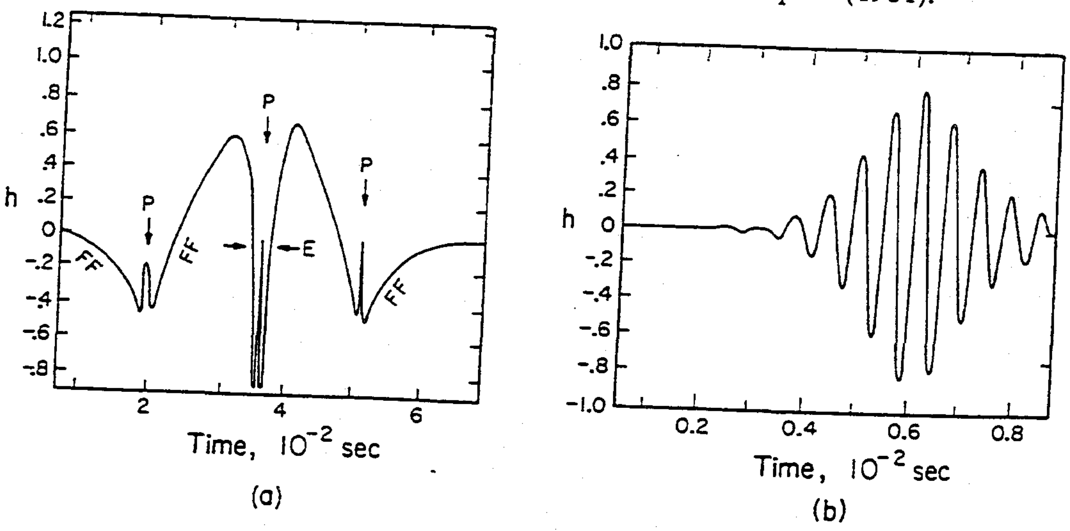
② Filtering.

- Amount of non-Gaussian tail in distribution of $K = \int K(t)\, n(t)\, dt$ depends on shape of K.



- Worst for $K(t) \simeq$ pulse shape
- Smaller for $K(t) =$ Optimal filter for binaries.

Fig. 9.3. Wave forms produced by two very different scenarios for the collapse of a normal star to form a neutron star. Wave form (*a*) is from Saenz and Shapiro (1978); (*b*) is from Saenz and Shapiro (1981).



COALESCING COMPACT BINARIES

- Neutron star / neutron star inspirals most promising & reliable source for LIGO / VIRGO because

(i) Wave form is well understood; source is very "clean". [Improves detectability $\frac{S}{N} \propto \sqrt{n_{cyc}}$]

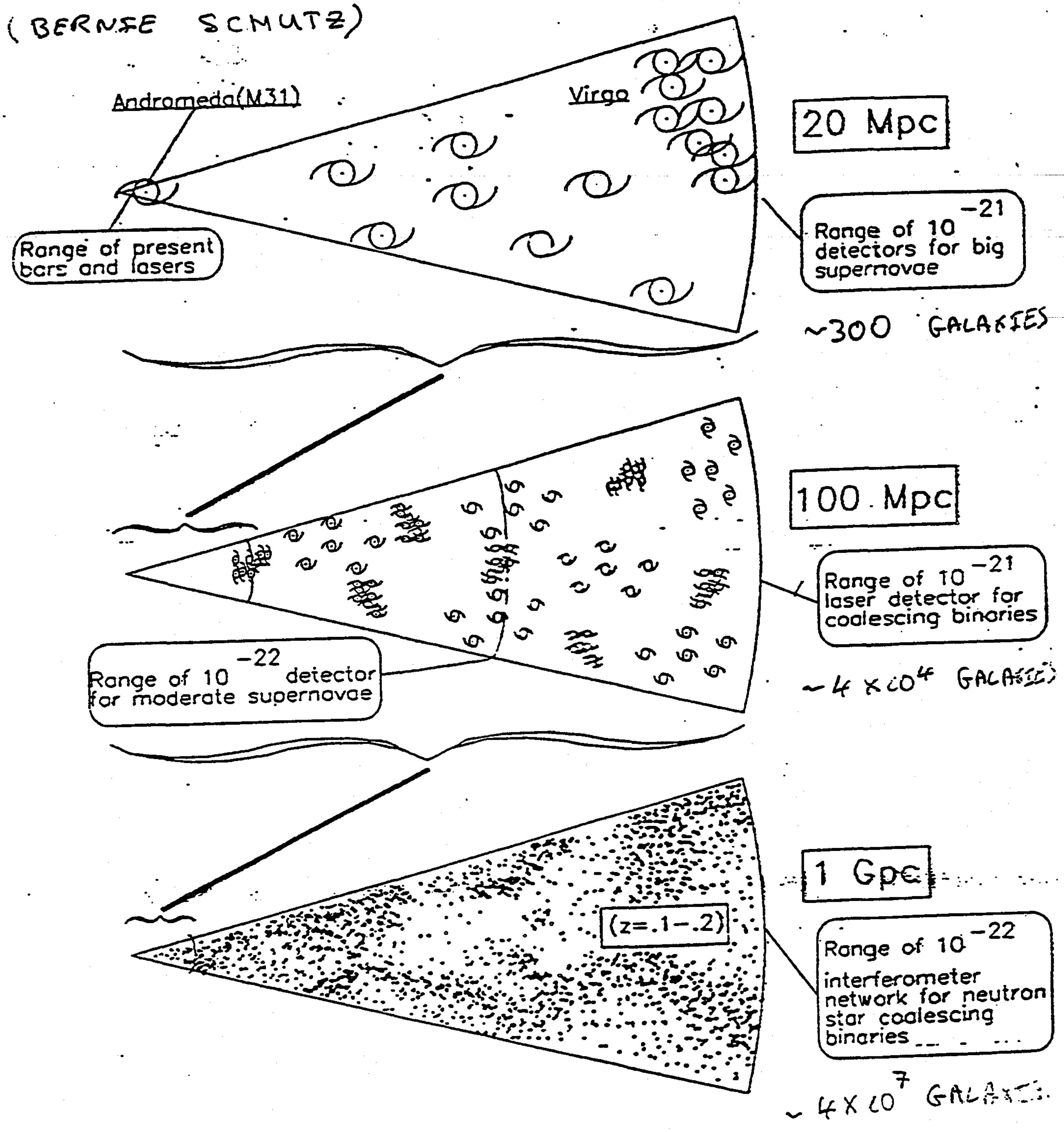
(ii) Event rates are known with relatively high confidence. 3/yr within 200 Mpc, Phinney, Narayan et al.

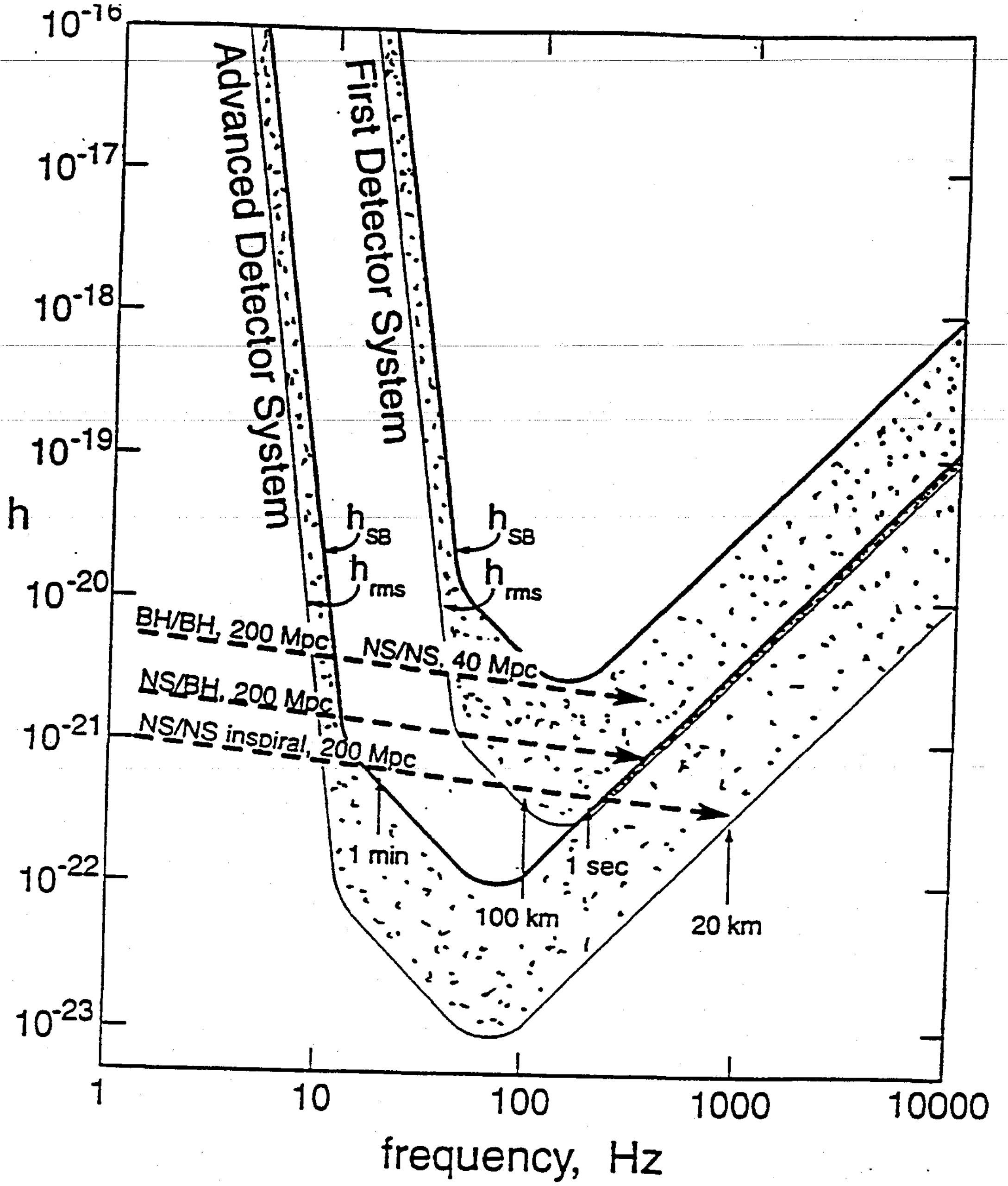
- Event rates imply, for LIGO / VIRGO with "advanced detectors"

	Range		Detection Rate	
NS/NS	~1 Gpc	$\times 2^{\pm 1}$	~0.5 day	$\times 2^{\pm 6}$
NS/BH	~3 Gpc		several/day	
BH/BH	~6 Gpc			

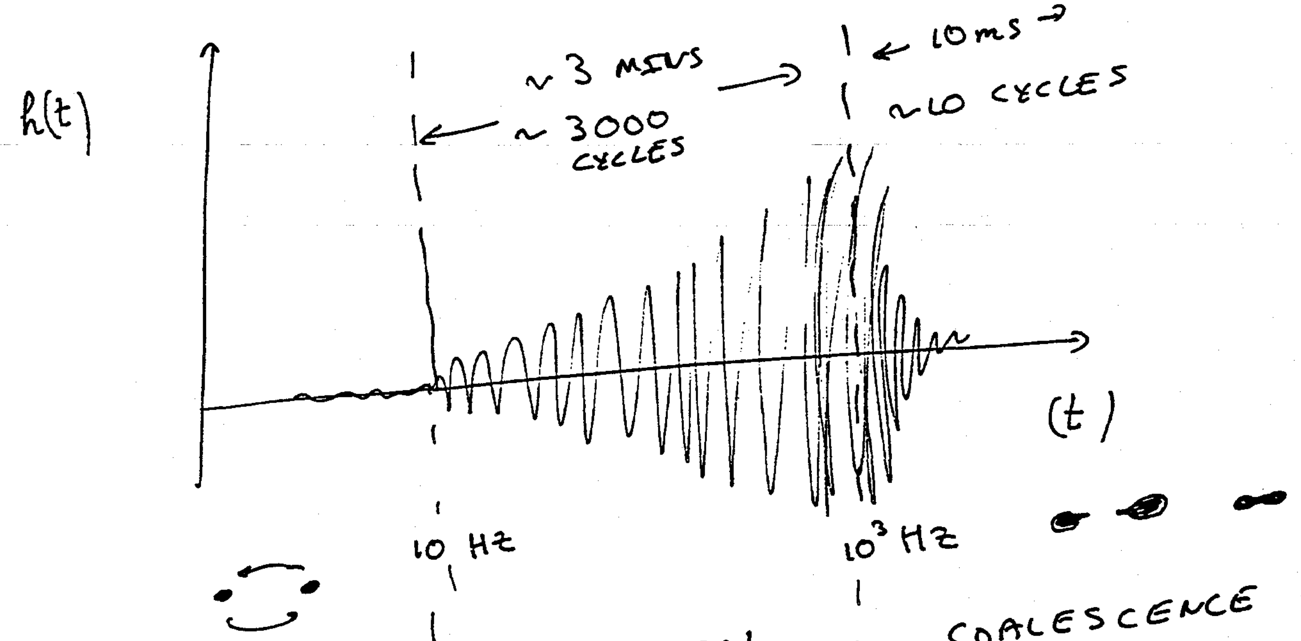
ILLUSTRATION OF DISTANCE SCALES:

(BERNIE SCHUTZ)


Andromeda(M31)
Virgo
20 Mpc
Range of present bars and lasers
Range of 10^{-21} detectors for big supernovae
~300 GALAXIES
100 Mpc
Range of 10^{-21} laser detector for coalescing binaries
Range of 10^{-22} detector for moderate supernovae
~4×10^4 GALAXIES
1 Gpc
(z=.1−.2)
Range of 10^{-22} interferometer network for neutron star coalescing binaries
~4×10^7 GALAXIES


10^{-16}
10^{-17}
10^{-18}
10^{-19}
10^{-20}
10^{-21}
10^{-22}
10^{-23}
h
1
10
100
1000
10000
frequency, Hz
Advanced Detector System
First Detector System
h_{SB}
h_{rms}
h_{SB}
h_{rms}
BH/BH, 200 Mpc
NS/NS, 40 Mpc
NS/BH, 200 Mpc
NS/NS inspiral, 200 Mpc
1 min
100 km
1 sec
20 km

THE WAVEFORM



INSPIRAL STAGE

$\sim 10^{53}$ ergs

$\sim 10^{76}$ gravitons

INFORMATION:

- Distance
- Sky location
- Masses x 2
- Orbital elements
- Spins (?)

COALESCENCE STAGE

$\sim 10^{34}$ ergs.

INFORMATION:

- Neutron star radius
- Black hole nonlinear dynamics.

DETECTION

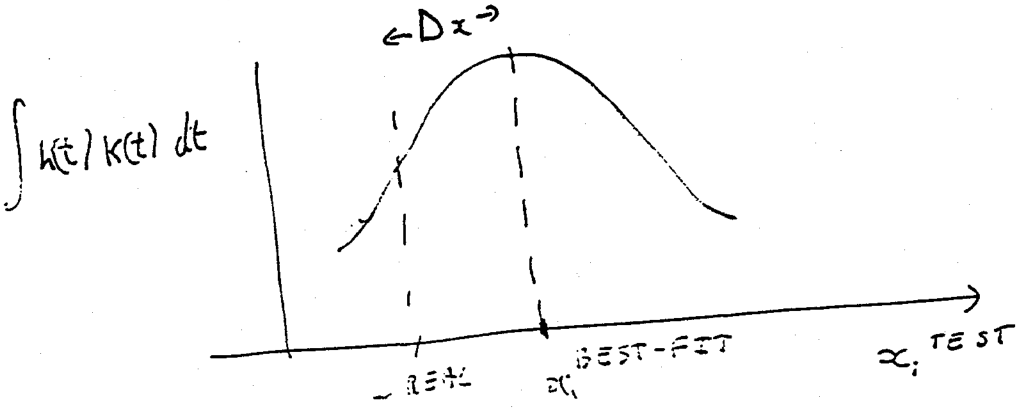
- Optimal filter separately at each detector with $\sim 10^5$ template shapes.

- $10^5 \times \left(10^{10} \text{ "start times" per year}\right) = \dfrac{10^{15}}{\text{year}}$

$\Rightarrow \left(\frac{S}{N}\right)_{\text{threshold}} \simeq 6.0$ [SETS RANGE OF DETECTORS].

- Computation rate $\sim$ 10 Giga Flops.

- Parameter Extraction $\quad h(t) = n(t) + s(t; x_i^{REAL})$
  - Family of templates $\quad K(t, x_i^{TEST})$



GRAVITATIONAL WAVEFORM:

- CALCULATED BY EXPANDING IN

$$\Phi \sim \frac{GM}{c^2 r} \sim \left(\frac{v}{c}\right)^2 \sim \left(\frac{GMf}{c^3}\right)^{2/3} \sim 0.1$$

$$h(t) \approx \underbrace{m(t)}_{\text{AMPLITUDE MODULATION}} Q(\text{angles}) \underbrace{\left(\frac{M_c}{D}\right)(M_c f)^{2/3}}_{\text{"STANDARD CANDLE"}} \cos\Big[\overset{2\times \text{ ORBITAL PHASE}}{\underbrace{2\pi \int f(t)\,dt}_{\text{PHASE EVOLUTION INFORMATION}}}\Big]$$

$$M_c = \mu^{3/5} M^{2/5}$$

$$= \text{"Chirp mass"}$$

$$\frac{d}{dt} f(t) \propto M_c^{5/3} f^{11/3} \left[1 - \left(\frac{743}{336} + \frac{11}{4}\frac{\mu}{M}\right)(\pi M f)^{2/3} + \cdots \right]$$

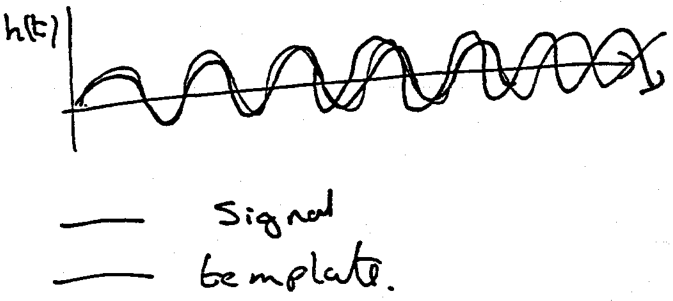
INFORMATION EXTRACTION.

- Waveform very "clean"

- Relativistic effects large

- Accurate measurements of phase evolution

$$\Delta\Phi \sim \frac{1}{S/N}\,\frac{1}{N_{cycles}}$$



- Parameter extraction accuracies:

Distance Inclination of orbit	$\sim 20-30\,\%$
Sky Location	$\sim 1^\circ \times 1^\circ$
$\mathcal{M} = \mu^{3/5} M^{2/5}$	$\sim 0.1\,\%$
μ	$\sim 10\,\%$
time	~ 1 ms.

- Distribution of mass on large ($\gtrsim 200$ Mpc) scales
- Cosmological information
- Mass spectra of NS's and BH's.

PROBLEM | THEORETICAL CHALLENGE

- Today we have only approximate templates (filters) whose phasing is not quite right "POST 1.5 NEWTONIAN".

- Effect on Detection:
  - Crude templates adequate for detecting $\gtrsim 90\%$ of signals.

- Effect on parameter extraction:
  - Introduces systematic errors into measurement.

To ensure that systematic errors $\lesssim$ statistical errors

need phasing accurate to $\sim 10^{-5}$.

$\boxed{\text{TASK}}$ - Compute the templates $\left[\text{up to } \left(\frac{v}{c}\right)^{11}\right]$

- Will occupy experts for several years.

PERIODIC SOURCES

[SKY LOCATION KNOWN]

- Detection

- Take long stretch of data, correct for Doppler shifts due to Earths motion, and Fourier transform.

- Bandwidth $\Delta f \sim \frac{1}{T_{obs}}$

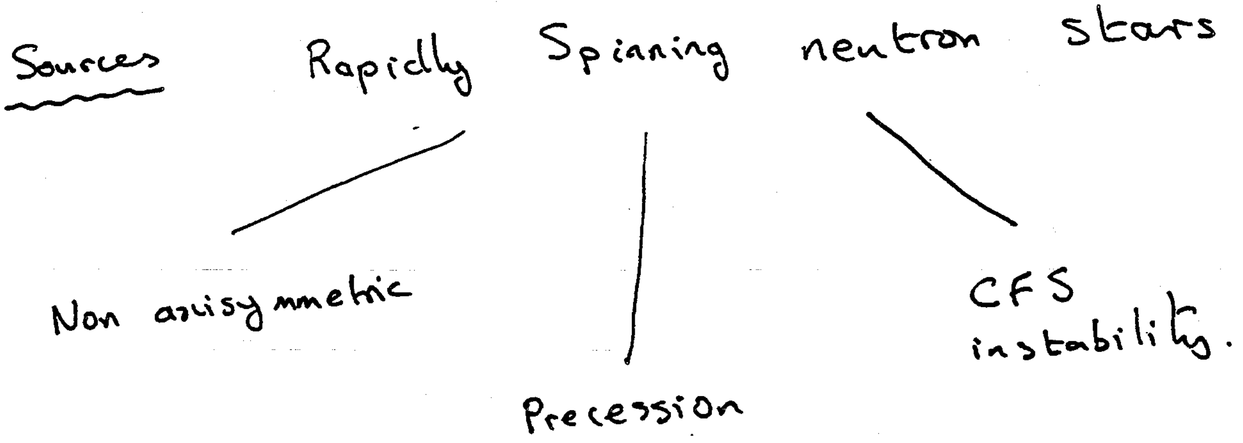
$$\Rightarrow \quad \frac{S}{N} \sim \frac{h_{amp}}{\sqrt{G_n(f)\,\Delta f}} \sim \frac{h_{amp}}{\sqrt{G_n(f)/T_{obs}}}.$$

- If frequency unknown require

$$e^{-\frac{1}{2}(S/N)^2} \lessapprox \frac{\Delta f}{f},$$

$$\Rightarrow \text{CAN SEE} \quad h_{amp} \sim \sqrt{2\log(f/\Delta f)}\,\sqrt{S_n(f)/T_{obs}}$$

$$\sim 10^{-27} \quad \text{if} \quad T_{obs} = 1\ \text{YR}$$



Sources Rapidly Spinning neutron stars

Non axisymmetric

Precession

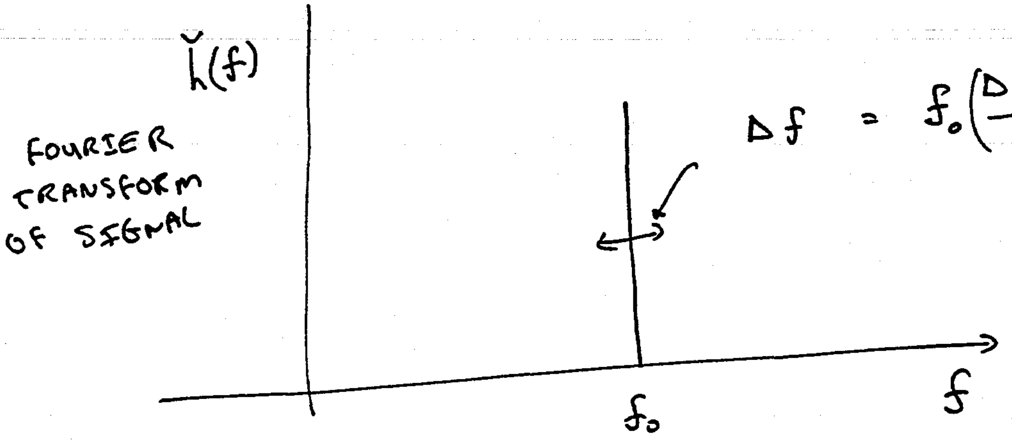
CFS instability.

- Asphericity ε gives

$$h_{amp} \sim 10^{-27}\left(\frac{\varepsilon}{10^{-6}}\right)\left(\frac{f}{1\ \mathrm{kHz}}\right)\left(\frac{10\ \mathrm{kpc}}{r}\right).$$

SKY LOCATION UNKNOWN --- ALL SKY SEARCHES.

- Time varying Doppler shifting of Earths motion spreads signal in frequency space.



$$\Delta f = f_0\left(\frac{\Delta V_{EARTH}}{c}\right)$$

- Correction must be separately applied to each source location on sky.

$$N_{patches} \sim 10^{13}\left(\frac{f}{1\ kHz}\right)^2\left(\frac{T_{obs}}{10^7\ s}\right)^4$$

- All sky search is computation limited.

$$T_{obs} \lesssim (10^6\ s)\left(\frac{f}{1\ kHz}\right)^{-\frac{3}{5}}\left(\frac{T_{analysis}}{1\ year}\right)^{\frac{1}{5}}\left(\frac{Speed}{1\ Teraflop}\right)^{\frac{1}{5}}.$$

- h_{limit} (1 Teraflop) $\simeq 0.1\ h_{limit}$ (infinite computing power)

- Clever new ideas needed!

STOCHASTIC WAVES

- Phase incoherent superposition of gravitons travelling from all directions, "background noise".

- Quantified by

$$\Omega_{gw}(f) = \frac{dE}{d^3x \; d\ln f} \times \left(\frac{1}{\rho_{crit}} \right)$$

$$\left[\text{Analogous} \quad \Omega_{em} \sim 10^{-3} \right].$$

- $$h_{rms} \sim 10^{-24} \left(\frac{\Omega_{gw}}{10^{-10}} \right)^{\frac{1}{2}} \left(\frac{f}{100\,\text{Hz}} \right)^{-1}$$ Visible above noise only if $\Omega \gtrsim 10^{-6}$

- Detection :

$$\left.\begin{aligned} h_1(t) &= n_1(t) + n_{SB}(t) \\ h_2(t) &= \underbrace{n_2(t)}_{\substack{\text{detector}\\\text{noise}\\\uparrow\\\text{Uncorrelated}}} + \underbrace{n_{SB}(t)}_{\substack{\text{Stochastic}\\\text{Background}\\\text{noise}\\ \\\text{correlated.}}} \end{aligned}\right\} \begin{array}{l}\text{2}\\\text{independent}\\\text{detectors.}\end{array}$$

- Bandpass filter detector outputs to chosen frequency band, then compute

$$\int dt \; h_1(t)\, h_2(t) = \underbrace{\int n_1 n_2 + \int n_1 n_{SB} + \int n_2 n_{SB}}_{\text{all these terms average to zero}}$$

$$+ \underbrace{\int n_{SB}(t)^2 \, dt}_{\text{grows } \propto T_{obs}}$$

- Using 2 LIGO detectors could see signal if

$$\Omega_{gw} \gtrsim 10^{-10} \left(\frac{T_{obs}}{1 \text{ year}} \right)^{-1}$$

in $30 \text{ Hz} \lesssim f \lesssim 100 \text{ Hz}$

- Conventional wisdom is that $\Omega_{gw} \ll 10^{-10}$ but theorists can accomodate $\Omega_{gw} \gtrsim 10^{-10}$.

Sources [Highly speculative]

- Parametric amplification of gravitational wave vacuum fluctuations during inflation.

  $\Omega_{expected} \sim 10^{-13}$ [Grischuk]

- Phase transitions in early Universe

- Cosmic strings vibrating

  $\Omega_{expected} \sim 10^{-7}$

BATCH
START

4. & 5 Idealized Interferometer

STAPLE
OR
DIVIDER

Ph 103c: The Physics of LIGO 8 April 1994

LECTURE 4.
IDEALIZED THEORY OF INTERFEROMETRIC DETECTORS — I.

Lecture by Kip S. Thorne

Assigned Reading:

A. "Gravitational Radiation" by Kip S. Thorne, in *300 Years of Gravitation*, eds. S. W. Hawking and W. Israel (Cambridge University Press, 1987), pages 414–425; ending at beginning of first full paragraph on 425. [This material uses the phrase *beam detector* for an *interferometric gravitational-wave detector*. The principal results quoted in this lecture are derived in the exercises below.]

G. The following portions of "Chapter 7. Diffraction" from the textbook manuscript *Applications of Classical Physics* by Roger Blandford and Kip Thorne: Section 7.2 (pages 7-2 to 7-7), and Section 7.5 (pages 7-20 to 7-27). [This material develops the foundations of the theory of diffraction (Green's theorem and the Helmholtz-Kirchoff formula), explores semi-quantitatively the spreading of a transversely collimated beam of light, develops the formalism of *paraxial Fourier optics* for analyzing quantitatively the propagation of collimated light beams, and uses that formalism to derive the evolution of the cross sectional shape of a Gaussian beam, of the sort used in LIGO.]

Suggested Supplementary Reading:

H. A. E. Siegman, *Lasers* (University Science Books, Mill Valley CA, 1986), chapter 17, "Physical Properties of Gaussian Beams." [This chapter develops in full detail the paraxial-Fourier-optics theory of the manipulation of Gaussian beams by a system of lenses and mirrors, and the shapes of the Gaussian modes of an optical resonator (Fabry-Perot cavity).]

A Few Suggested Problems

1. *Shot Noise.* Reread the discussion of shot noise on pages 5-20 and 5-21 of Blandford and Thorne, *Random Processes* (which was passed out last week). In that discussion let the random process $y(t)$ be the intensity $I(t) = d(\text{energy})/dt$ of a laser beam, and let $F(t)$ be the intensity carried by an individual photon, which has frequency ω.
   (a) Explain why $\tilde{F}(0)$, the Fourier transform of F at zero frequency, is the photon energy $\hbar\omega$.
   (b) Show that the spectral density of I (the "shot-noise spectrum") is

$$G_I(f) = 2\bar{I}\hbar\omega\ , \tag{1}$$

   where $\bar{I}$ is the beam's mean intensity.
   (b) Let $N(t)$ be the number of photons that the beam carries into a photodiode between time t and time $t + \hat{\tau}$ (so $\hat{\tau}$ is the averaging time): $N(t) = \int_t^{t+\hat{\tau}} I(t')dt'$.

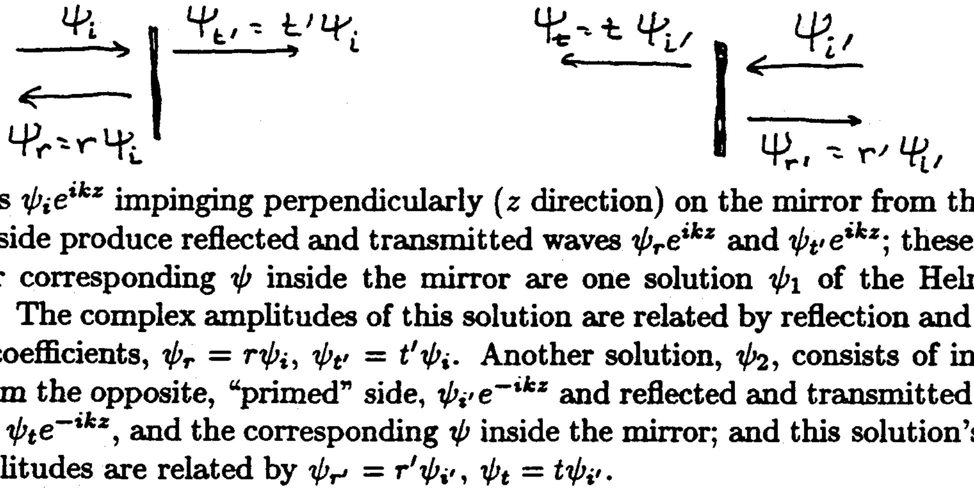
This $N(t)$ is a linear functional of $I(t')$. Use the theory of linear signal processing to derive the spectral density $G_N(f)$ of $N(t)$, and then compute the mean square fluctuations of N: $(\sigma_N)^2 = \int_0^\infty G_N(f)df$. Your result should be $\sigma_N = \sqrt{\bar{N}}$, where $\bar{N}$ is the mean number of photons that arrive in the averaging time $\hat{\tau}$. This is the standard "square-root-of-N" fluctuation in photon arrival for a laser beam.

2. *Reciprocity Relations for a Mirror and a Beam Splitter.* Modern mirrors, beam splitters, and other optical devices are generally made of glass or fused silica (quartz), with dielectric coatings on their surfaces. The coatings consist of alternating layers of materials with different dielectric constants, so the index of refraction n varies periodically. If, for example, the period of n's variations is half a wavelength of the radiation that impinges on the device, then waves reflected from successive dielectric layers build up coherently, producing a large net reflection coefficient. In this exercise we shall derive the reciprocity relations for a mirror of this type, with normally incident radiation. The generalization to radiation incident from other directions, and to other dielectric optical devices is straightforward.

   The foundation for the analysis is the wave equation,

$$\left(-\frac{\partial^2}{\partial t^2} + \frac{c^2}{n^2(\mathbf{x})}\nabla^2\right)\psi = 0$$

   satisfied by any Cartesian component ψ of the electric field, and the assumption that ψ is precisely monochromatic with angular frequency ω. These imply that the spatial dependence of ψ is governed by the Helmholtz equation with spatially variable wave number $k(\mathbf{x}) = n(\mathbf{x})\omega/c$: $\nabla^2\psi + k^2\psi = 0$.



   Let waves $\psi_i e^{ikz}$ impinging perpendicularly (z direction) on the mirror from the "unprimed" side produce reflected and transmitted waves $\psi_r e^{ikz}$ and $\psi_{t'} e^{ikz}$; these waves and their corresponding ψ inside the mirror are one solution ψ_1 of the Helmholtz equation. The complex amplitudes of this solution are related by reflection and transmission coefficients, $\psi_r = r\psi_i$, $\psi_{t'} = t'\psi_i$. Another solution, ψ_2, consists of incident waves from the opposite, "primed" side, $\psi_{i'} e^{-ikz}$ and reflected and transmitted waves $\psi_{r'} e^{+ikz}$, $\psi_t e^{-ikz}$, and the corresponding ψ inside the mirror; and this solution's complex amplitudes are related by $\psi_{r'} = r'\psi_{i'}$, $\psi_t = t\psi_{i'}$.

   (a) Show that ψ obeys Green's theorem [Equation (7.3) of Blandford and Thorne] throughout the mirror. Apply Green's theorem, with ψ and ψ_0 chosen to be various pairs of ψ_1, ψ_2, ψ_1^*, ψ_2^* (where the star denotes complex conjugation). Thereby obtain four relationships between r, r', t, and t'.

   (b) Show that these relationships can be written in the form

$$r = \sqrt{\mathcal{R}}e^{2i\beta}, \quad r' = -\sqrt{\mathcal{R}}e^{2i\beta'}, \quad t = t' = \sqrt{\mathcal{T}}e^{i(\beta+\beta')},$$

where β and β' are unconstrained phases, and $\mathcal{R}$ and $\mathcal{T}$, the power reflection and transmission coefficients are related by

$$\mathcal{R} + \mathcal{T} = 1\,, \tag{2}$$

which is just energy conservation.

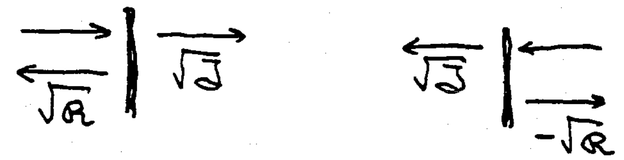
(c) Show that, if one moves the origin of coordinates as seen from the unprimed side by $\delta z = -k\beta$, and moves the origin as seen from the primed side by $\delta z = +k\beta'$, one thereby will make all the reflection and transmission coefficients real:

$$t = t' = \sqrt{\mathcal{T}}, \quad r = -r' = \sqrt{\mathcal{R}}\,. \tag{3}$$

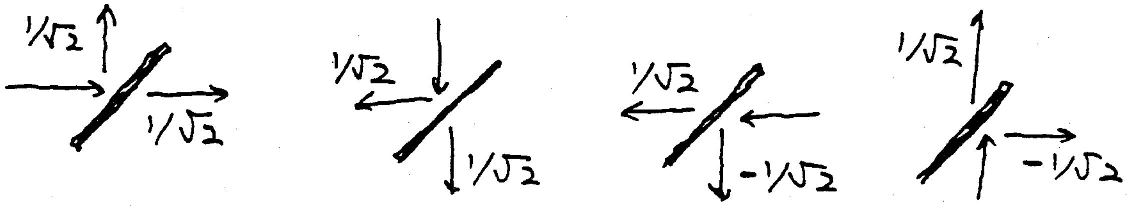
Thus, with an appropriate choice of origin on each side of the mirror, the coefficients can always be made real.

The same is true for the reflection and transmission coefficients of any other optical device made of a lossless, spatially variable dielectric. In particular, for a perfect, 50/50 beam splitter, the transmission coefficient becomes, with appropriate choice of origins, $1/\sqrt{2}$ from each and every one of the four input ports, and the reflection coefficient becomes $+1/\sqrt{2}$ from the input ports on one side of the beam splitter and $-1/\sqrt{2}$ from the input ports on the other side of the beam splitter. These results are summarized by the following figures:

mirror :



beam-splitter



2. *Transfer Function and Photon Shot Noise for a Delay-Line Interferometer* In class, Kip derived the "tranfer function" for a delay-line interferometer in the limiting regime where the waveform $h(t)$ is nearly constant during the time $2BL/c$ that the light is stored in the interferometer arms (during B round trips in an arm whose length is L). His result was

$$I_{PD}(t) = I_1(t) + 2\sqrt{\bar{I}_1 I_0}\, BkLh(t) \tag{4}$$

where I_0 is the mean laser input power entering the beamsplitter, $I_1(t)$ is the (slightly fluctuating because of shot noise) intensity of the light falling onto the photodiode in the absence of a gravitational-wave signal, $\bar{I}_1$ is the mean intensity onto the photodiode, B is the number of round trips in the arms of the interferometer, $k = \omega/c = 2\pi/\lambda_e$ is the light's wave number, L is the arm length, and $h(t)$ is the gravitational waveform. Kip used this and the shot-noise spectral density [Eq. (1) above] to derive the following expression for the shot-noise contribution to the interferometer's gravitational-wave noise output:

$$G_h(f) = \frac{\hbar\omega}{2I_0(BkL)^2} \, . \tag{5}$$

(a) Use the same method of analysis as Kip did in class to derive the transfer function when the gravitational wave is sinusoidal in time with angular frequency $\Omega = 2\pi f$, i.e. when $h(t) = h_o \cos(\Omega t) = h_o \mathrm{Real}(e^{-i\Omega t})$, with a frequency f high enough (gravitational wavelength short enough) that the waveform *can* vary significantly while the light is stored in the arms. Your result should be the same as Eq. (4), with B replaced by

$$B_{\text{eff}} = B\frac{\sin(f/f_0)}{f/f_0} \, , \quad f_0 \equiv \frac{c}{2\pi BL} = \frac{119\text{Hz}}{(B/100)(L/4\text{km})} \, . \tag{6}$$

(b) Show that the shot-noise contribution to $G_h(f)$ has the form (5) with B replaced by B_{eff}.

3. *Transfer Function and Photon Shot Noise for a Fabry-Perot Interferometer.* In class, Kip showed that for a Fabry-Perot interferometer in the regime of slow variations of $h(t)$ the transfer function and photon shot noise have the forms (4) and (5), with B replaced by

$$B_{\text{eff}} = \frac{4}{(1-\mathcal{R})} \tag{7}$$

where $\mathcal{R}$ is the power reflectivity of the interferometer's corner mirrors and where it is assumed that the end mirrors are perfectly reflecting. Show that, if the variations of $h(t)$ are not assumed to be slow, then the transfer function (for monochromatic gravitational waves) has the form (4) and the shot noise contribution to $G_h(f)$ has the form (5), with B replaced by

$$B_{\text{eff}} = \frac{B}{\sqrt{1+(f/f_0)^2}} \, , \tag{8}$$

where f_0 is as in Eq. (6) above.

Lecture 4
Idealized Theory of Interferometers — I.

by Kip S. Thorne, 8 April 1994

Thorne lectured at the blackboard. The following are the notes from which he lectured, cleaned up a bit to make them more understandable.

Kp Thorne
Ph 106

Letwe 4 — Idealized Theory of Interferometers

4/8/94

1. Overview of How ~~[illegible]~~ an Interferometer Works & Orders of Magnitude

a. **Goal:** $h \sim 10^{-22}$; $\frac{\Delta L}{L} = h$; $L = 4\text{km} \sim 10^{6}\text{cm}$

$\Rightarrow \Delta L \sim 10^{-16}\text{cm}$; $\lambda_e \simeq 0.5\mu m \approx 10^{-4}\text{cm}$

$\Rightarrow$ measure 10^{-12} of λ_e ! seems outrageous

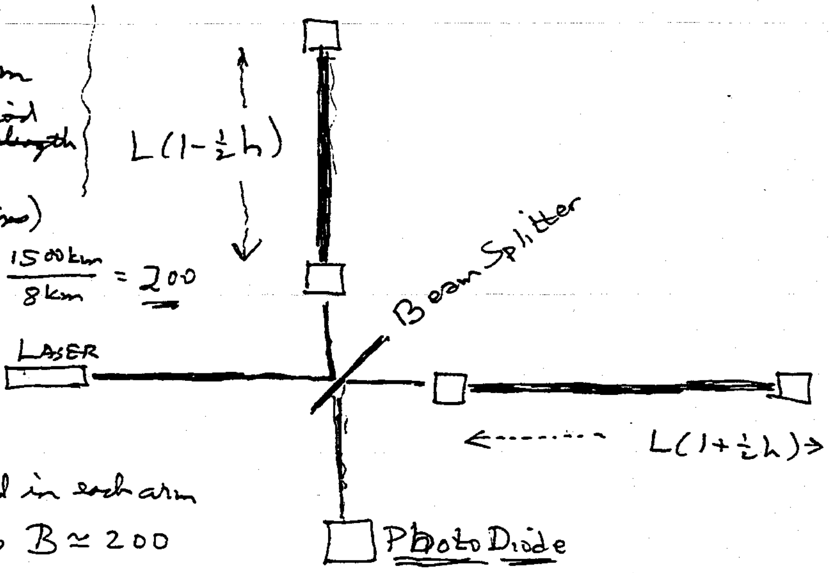
b. GW $f \sim 100\,Hz$

$\lambda \sim 3000\,km$

store light for $\frac{1}{2}$ ~~wavelength~~ period

$\Rightarrow B = (\#\text{ of round trips})$

$= \frac{\lambda/2}{2L} = \frac{1500\text{km}}{8\text{km}} = 200$



(So light is stored in each arm for, on average, $B \simeq 200$ round trips)

c. Phase shift: $\Delta\Phi = \cancel{\frac{\lambda_e}{2\pi}} \overset{k}{\left(\frac{2\pi}{\lambda_e}\right)} 2B\Delta L = 200 \times \frac{2\pi}{5\times10^{-5}\text{cm}} 2\times10^{-16}\text{cm}$

$\simeq 10^{-9}$

d. How accurately can this phase shift be measured?

If clever & good: $\Delta\Phi \simeq \frac{1}{\sqrt{N_\gamma}}$ [Proof: $\overset{\hbar\omega\Delta N_\gamma}{\Delta E}\, \Delta t \gtrsim \hbar \Rightarrow$
$\underbrace{\Delta N}_{\sqrt{N_\gamma}} \underbrace{\omega\Delta t}_{\Delta\Phi} \gtrsim 1$]

number of photons from laser in time $1/f = 0.01\,sec$

[Photon shot noise]

$\Rightarrow$ need 10^{18} photons in $0.01\,sec$

$\Rightarrow I = \frac{10^{18} \times \overset{=10^{-27}}{\hbar\omega} = 5\times10^{15}}{10^{-2}\text{sec}} \simeq 5\times10^{15-27+18+2} \simeq 5\times10^{8}\,\frac{\text{erg}}{\text{s}}$

$\simeq$ 50 Watts

[Can be achieved @ 5 Watt laser ~~and~~ a 10-fold recycling of used light.]

e. Won't vibrations of atoms in mirror prevent measurement of such tiny motions? No –

i. Individual atoms vibrate at $f \sim 10^{13}\,Hz$, far above interferometers' gravity-wave band

ii. Only concern is lowest frequency normal modes which have thermal amplitudes

$$\sqrt{\frac{kT}{m\Omega^2}} \simeq \sqrt{\frac{(1.38\times 10^{-16} erg/K)(300K)}{10^4 g\ (10^5/s)^2}} \simeq \sqrt{4\times 10^{-16+2-4-10}}$$

$$\simeq 2\times 10^{-14}\,cm.$$

The interferometer averages over many periods and sees only <u>changes</u> of amplitude – which are made small by giving mirrors high mechanical Q's.

This <u>thermal</u> noise will be discussed in later lectures.

2. Ways to Produce Multiple Bounces in Interferometer Arms

a. Delay Line [many ~~two~~ discrete spots on each mirror]

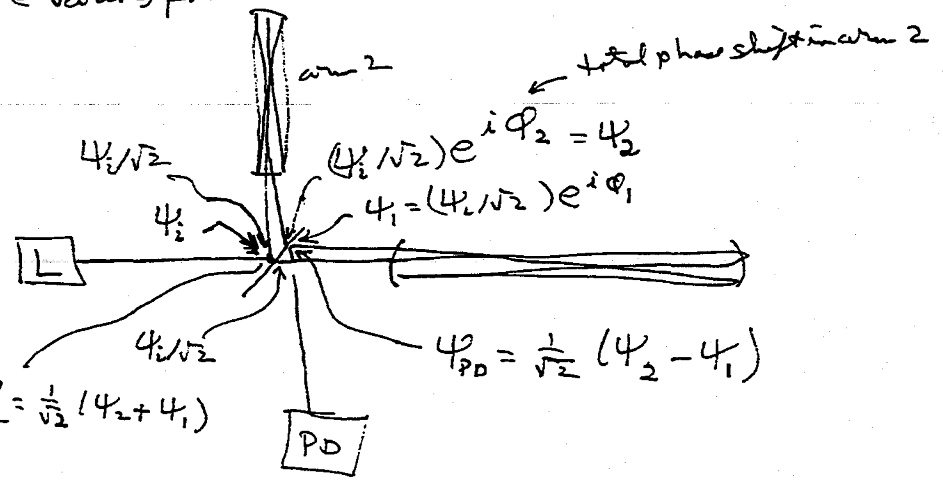
b. Fabry-Perot [one spot on each mirror]

3. Delay Line — @ h constant during storage time: $2BL \ll \lambda_{GW}/2$

a. Describe light by $\psi = \frac{E_x}{\sqrt{4\pi}}$ ... so $\boxed{I = |\psi|^2}$

$\psi = \psi e^{-i\omega t}$; $\psi = e^{ikx}$ if propagate in x direction; $\boxed{k = \frac{\omega}{c} = \frac{2\pi}{\lambda_e}}$

b. Field @ various points


arm 2
total phase shift in arm 2
$(\psi_i/\sqrt{2})e^{i\Phi_2} = \psi_2$
$\psi_i/\sqrt{2}$
$\psi_1 = (\psi_i/\sqrt{2})e^{i\Phi_1}$
ψ_i
L
$\psi_i/\sqrt{2}$
$\psi_{PD} = \frac{1}{\sqrt{2}}(\psi_2 - \psi_1)$
$\psi_L = \frac{1}{\sqrt{2}}(\psi_2 + \psi_1)$
PD

c. A bit of algebra!

$\psi_L = \frac{1}{2}\psi_i\left(e^{i\Phi_2} + e^{i\Phi_1}\right)$ [is field going back toward laser from interferometer]

$\Rightarrow \boxed{|\psi_L|^2 = \underbrace{|\psi_i|^2}_{I_0} \cos^2\left(\frac{\Phi_2 - \Phi_1}{2}\right)}$ intensity toward laser

$\psi_{PD} = \frac{1}{2}\psi_i\left(e^{i\Phi_2} - e^{i\Phi_1}\right) \leftarrow$ [Field going toward photodiode]

$\Rightarrow \boxed{|\psi_{PD}|^2 = \underbrace{|\psi_i|^2}_{I_0} \sin^2\left(\frac{\Phi_2 - \Phi_1}{2}\right)}$ intensity toward PD

d. Operate PD on "dark fringe" ["dark port"] $\boxed{I_0 = \text{Laser Power}}$

So $\boxed{\Phi_2 - \Phi_1 \equiv \Phi_0 \ll 1 \text{ before wave arrives}}$

e. Effect of wave:

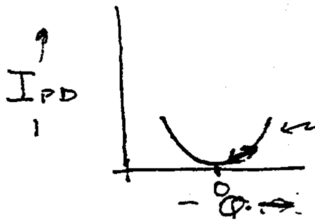
$$\boxed{\delta\varphi_2 = -\delta\varphi_1 = k\cdot \underbrace{\tfrac{h}{2}L}_{\delta L \text{ in each arm}} \cdot \underbrace{2B}_{\#\text{ of one way trips}}}$$

$\boxed{B = \#\text{ of round trips in each arm}}$

$$\delta\varphi = \tfrac{1}{2}(\delta\varphi_2 - \delta\varphi_1)$$

$$\boxed{\delta\varphi = B\cdot 2k\delta L = B\cdot kLh}$$

f. $$\boxed{I_{PD}(t) = I_0\left(\varphi_0^2 + 2\varphi_0 BkL\,h(t)\right) \quad \text{If } 2BL \ll \lambda_{GW}/2}$$ ← Intensity of light into photodiode



as h oscillates, $\delta\varphi$ oscillates, and I_{PD} oscillates up and down on side of parabola

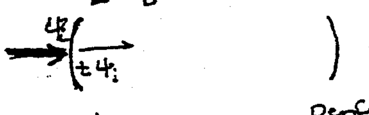
g. WHY dark port toward PD?

- Keep power on PD low
- Send light toward laser, so it can be recycled back into interferometer with new laser light

4. Fabry-Perot: ~~Single Cavity~~

a. Look at a single arm [cavity]

b. How it gets excited:



power transmission $= 1-R \sim 10^{-2}$
t = amp transmission $\sim 1/10$

perfect mirror (losses few $\times 10^{-5}$)

i. Turn on light suddenly, on resonance so $2KL$ = multiple of 2π

ii. First pass of light down arm

$\psi = t\psi_i$ [t = amplitude transmission]

returns in phase ⇒ next pass ~~$t\,r\,\psi_i$~~ $r(t\psi_i)$

returns in phase again ⇒ next: ~~t~~ $r^2(t\psi_i)$

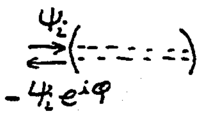
⋮

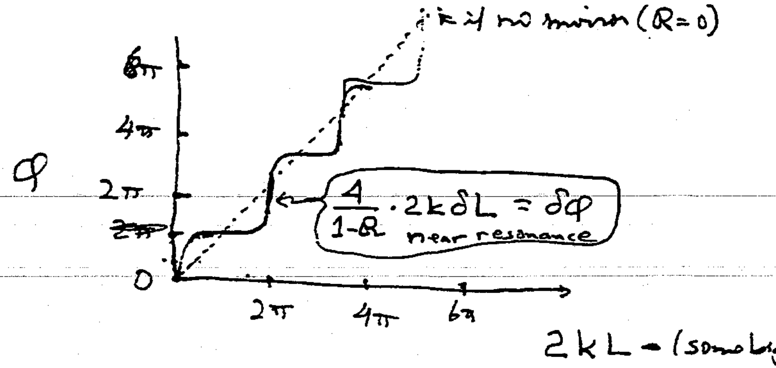
$$\psi_{inside} = t\psi_i(1+r+r^2+\dots) = \frac{t\psi_i}{1-r} \simeq \frac{\sqrt{1-r^2}\,\psi_i}{\frac{1}{2}(1-r^2)}$$

$\psi_{inside} = 2\psi_i$ / [illegible] ; $\boxed{I_{inside} = \frac{4}{1-R} I_0}$

c. If off resonance: cannot excite cavity with laser light

d. Phase shift as function of length of cavity (arm)





$\frac{4}{1-R} \cdot 2k\delta L = \delta\varphi$ near resonance

... i.e. $\frac{4}{1-R}$ plays role of the "Q"

e. Compare with delay line. — Same, with

$$\boxed{B = \frac{4}{1-R}}$$

--- effective number of bounces in a Fabry-Perot interferometer.

5. Shot Noise [cf. Random Processes Chapter, Ref. D]

a. The ~~beam~~ $I_{PD} = I_0 \Phi_0^2$ — in absence of GW's — (intensity toward photodiode) consists of photons which arrive randomly at photodiode. Each photon carries energy $\hbar\omega$, so average rate of arrival is $R = \frac{I_{PD}}{\hbar\omega} \simeq (10^{18}/\text{sec}) \frac{I}{1W}$

Since duration τ_p of each photon pulse is $> 10^{-15}$ sec, $\boxed{R\tau_p \gg 1}$

b. This random arrival means I_{PD} fluctuates randomly,

$I_{PD}(t)$, with some spectral density

$G_{I_{PD}}(f)$.

c. GW Frequencies of interest to us, $f \sim 0.01$ sec, are $\ll \frac{1}{\tau_p}$.

At these frequencies, the shapes of the pulses cannot influence $G_{I_{PD}}(f)$ — low f limit!! ⇒ $\boxed{G_I(f) \simeq G_I(0) = \text{const}}$

d. Exercise: From requirement that if $\bar{N} = \frac{I}{\hbar\omega}\hat{\tau}$ is mean # that arrive during time $\hat{\tau}$, then $\sigma_N = \sqrt{\bar{N}}$, we get

$$\boxed{G_I(f) = 2\bar{I}\hbar\omega} \quad \text{— Check units : } \frac{I^2}{Hz}$$

6. Shot noise ~~limit~~ in $h(t)$: (Translate this photon shot noise into an equivalent noise in grav. wave signal)

a. $I_{PD}(t) = I_0[\Phi_0^2 + 2\Phi_0 BkL\,h(t)]$

b. Rewrite formula on page 4 with $I_0\Phi_0^2 = I_1(t)$; $I_0\Phi_0 = \sqrt{I_0\bar{I}_1}$

$$\boxed{I_{PD} = I_1(t) + 2\sqrt{\bar{I}_1 I_0}\;BkL\,h(t)}$$

Then $G_{I_1}(f) = 4\bar{I}_1 I_0 (BkL)^2 G_h(f)$

$$\Rightarrow \boxed{G_h(f) = \frac{\overset{\leftarrow 2\bar{I}_1\hbar\omega}{G_{I_1}(f)}}{4\bar{I}_1 I_0 (BkL)^2}}$$

c. $$\boxed{G_h(f) = \frac{\hbar\omega}{2I_0(BkL)^2} \quad \text{--- white noise spectrum}}$$

d. $$\boxed{h_{rms} = \sqrt{f\,G_h(f)} = \underbrace{\frac{1}{2BkL}}_{\frac{\lambda_e}{2BL}} \cdot \frac{1}{\sqrt{\underbrace{(I_0/\hbar\omega)\underbrace{(1/2f)}_{P/2\,=}}_{\text{\# of photons available in 1/2 GW period}}}}}$$

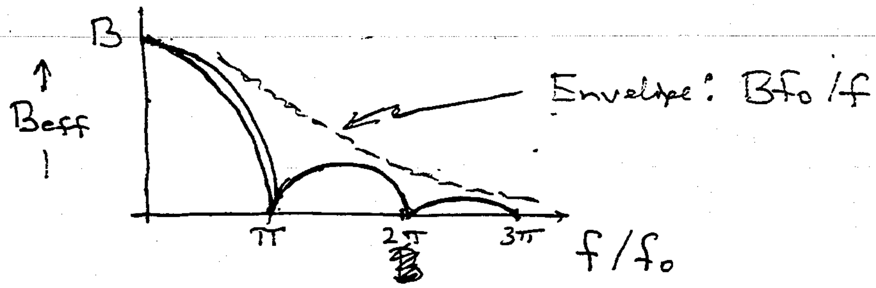
These are the shot noise limits on our interferometer — valid both for Delay Line and Fabry Perot.

7. What if $\lambda_{GW} \lesssim 2BL$?

a. Delay Line:

put shift onto light, then remove, then put on again...

$$\Rightarrow \boxed{B_{eff} = B \cdot \left| \frac{\sin\left(B \cdot \overset{\leftarrow 2\pi L/\lambda_{GW}}{(2\pi f L/c)}\right)}{B \cdot 2\pi f L/c} \right| = B \frac{\sin(f/f_0)}{f/f_0}}$$



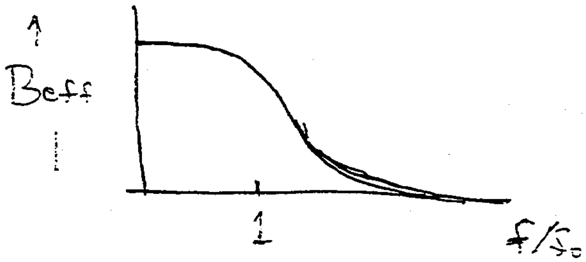
$$\boxed{f_0 = \frac{c}{2\pi BL} = \frac{119\,Hz}{(B/100)(L/4km)}}$$

$$\boxed{\frac{Bf_0}{f} = \frac{c}{2\pi L f} = \frac{\lambda_{GW}}{L}}$$

b. Fabry-Perot:

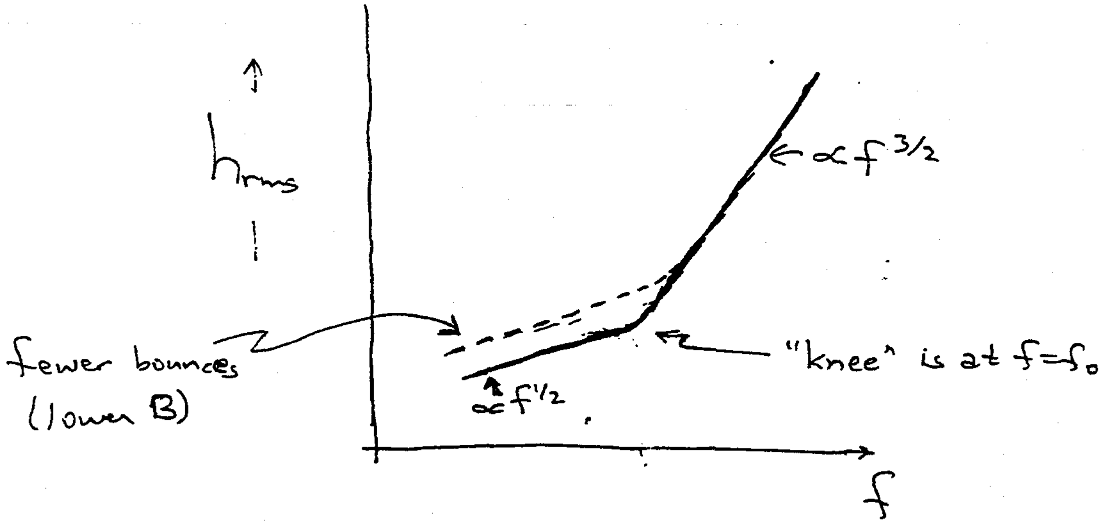
photons stored for a statistically varying length of time $\Rightarrow$ smooth die out of B:

$$\boxed{B_{eff} = \frac{B}{\sqrt{1+(f/f_0)^2}}} \simeq \begin{cases} B & f \ll f_0 \\ \frac{Bf_0}{f} = \frac{2\pi L}{cf} = \frac{L}{\lambda_{GW}} & f \gg f_0 \end{cases}$$



8. Bottom Line

$$\boxed{h_{rms} = \sqrt{f\, G_h(f)} = \frac{\bar{\lambda}_e}{2BL}\,\frac{1}{\sqrt{(I_0/\hbar\omega)(1/2f)}}\cdot\sqrt{1+(f/f_0)^2}}$$

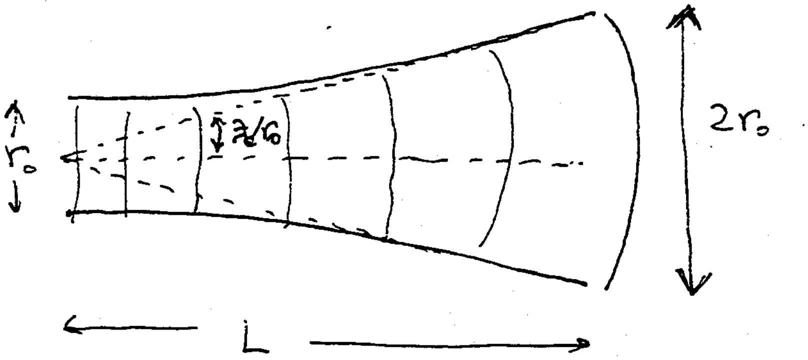


9. Gaussian Beams & Paraxial Optics [§7.5 of Ref. G]

a. Basic idea of Wave-spreading

$$\Delta k_y \, r_0 \sim 1$$

$$\Delta k_y \sim \frac{1}{r_0} \qquad \frac{\Delta k_y}{k} \sim \frac{\bar{\lambda}_e}{r_0} = \text{opening angle}$$



↑ distance for factor 2 spreading

$$\frac{2r_0}{L} \sim \frac{\bar{\lambda}_e}{r_0} \Rightarrow \boxed{r_0 \sim \sqrt{L\bar{\lambda}_e}}$$

b. This fixes size of beam in interferometer arms:

$$r_0 \sim \sqrt{L\bar{\lambda}_e} \sim \sqrt{(4\times10^5\,\text{cm})(4\times10^{-5}\,\text{cm})} \sim 4\,\text{cm}$$

c. Transverse profile is Gaussian

$$\boxed{\psi \approx \exp\left(-\frac{r^2}{r_0^2[1+z^2/z_0^2]}\right)} \quad \boxed{z_0 = r_0^2/\bar{\lambda}_e}$$

10

d. Beam is matched into cavity using lenses that manipulate its radius and its radius of curvature of phase fronts.

LECTURE 5.
IDEALIZED THEORY OF INTERFEROMETRIC DETECTORS—II.

Lecture by Ronald W. P. Drever

Assigned Reading:

I. R. W. P. Drever, "Fabry-Perot cavity gravity-wave detectors" by R. W. P. Drever, in *The Detection of Gravitational Waves*, edited by D. G. Blair (Cambridge University Press, 1991), pages 306–317. [This is a qualitative overview of Fabry-Perot gravitational-wave detectors, with emphasis on recycling in the later part (pages 312—317).]

J. B. J. Meers, "Recycling in laser-interferometric gravitational-wave detectors," *Phys. Rev. D*, **38**, 2317–2326. [This is the paper in which Meers introduced his idea of dual recycling and sketched out its features. You are not expected to master all the equations in this paper—which Meers just gives without derivation—but you might try deriving some of the equations as a homework exercise.]

Suggested Supplementary Reading:

K. B. J. Meers, *Physics Letters A*, "The frequency response of interferometric gravitational wave detectors," *Physics Letters A*, **142**, 465 (1989). [In this paper Meers discusses in some detail the frequency responses and sensitivities of various configurations of recycled interferometers.]

L. B. J. Meers and R. W. P. Drever, "Doubly-resonant signal recycling for interferometric gravitational-wave detectors." (preprint) [This paper introduces a new recycling configuration, not considered in previous papers.]

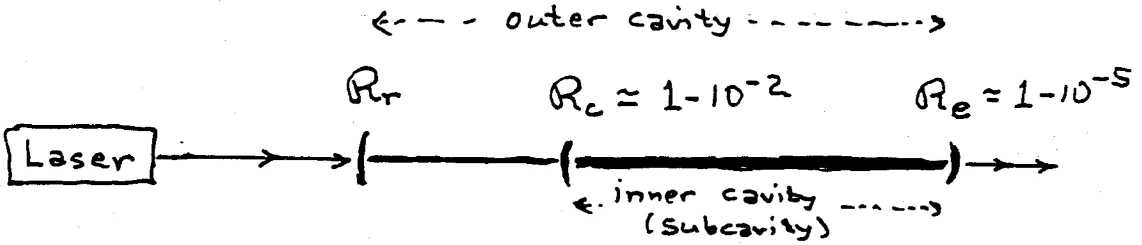
M. J. Mizuno, K. A. Strain, P. G. Nelson, J. M. Chen, R. Schilling, A. Rudiger, W. Winkler and K. Danzman, "Resonant sideband extraction: a new configuration for interferometric gravitational wave detectors," *Phys. Lett. A*, **175**, 273–276 (1993). [This is yet another recycling configuration]

N. R. W. P. Drever, "Interferometric Detectors of Gravitational Radiation," in *Gravitational Radiation*, N. Deruelle and T. Piran, eds. (North Holland, 1983); section 8 (pages 331-335). [This is the article in which Drever first presented in detail his ideas of power recycling and resonant recycling.]

A Few Suggested Problems

Note: Of all configurations for a recycled interferometer, the only one that is reasonably easy to analyze is power recycling. For this reason, and because this is the type of recycling planned for the first LIGO interferometers, I have chosen to focus solely on power recycling in the following exercises. — Kip.

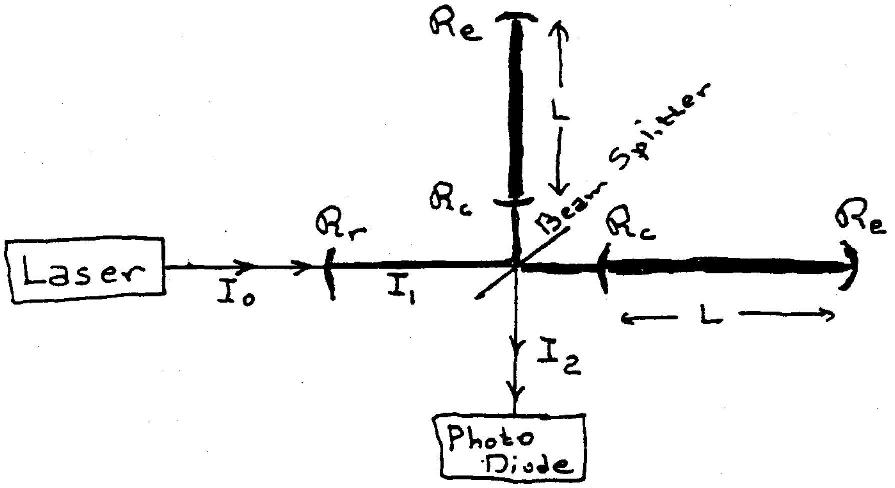
1. *Simplified Configuration of Nested Cavities that Illustrates Power Recycling:* Consider the configuration of two nested optical cavities shown below:



outer cavity
$\mathcal{R}_r$
$\mathcal{R}_c \simeq 1-10^{-2}$
$\mathcal{R}_e \simeq 1-10^{-5}$
Laser
inner cavity
(subcavity)

All three mirrors are assumed ideal in the sense that they do not scatter or absorb any light; therefore each of them satisfies the reciprocity relations of Assignment 4, Eq. (3). Assume that the power reflectivities of the subcavity are fixed: $\mathcal{R}_e$ is the highest reflectivity the experimenter has available; $\mathcal{R}_c$ is a much lower reflectivity, carefully designed to store the light in the subcavity for a chosen length of time. What reflectivity $\mathcal{R}_r$ should the recycling mirror have in order to maximize the light intensity in the subcavity, when both cavities are operating on resonance? Use physical reasoning to guess the answer before doing the calculation.

2. *Optimization of a Power Recycled Interferometer.* Consider the power-recycled interferometer shown below.



a. Suppose the interferometer is operated with the photodiode very near a dark fringe, so the light power I_2 is many orders of magnitude less than I_1. As in exercise 1, let $\mathcal{R}_e$ and $\mathcal{R}_c$ be fixed. How should $\mathcal{R}_r$ be chosen to maximize the power in the interferometers' two arms? Guess the answer on physical grounds before doing the calculation.

b. Again, suppose that I_2 is many orders of magnitude less than I_1. Let a low-frequency gravitational wave (one with $2\pi f BL/c \ll 1$ where $B = 4/(1-\mathcal{R}_c)$ is the effective number of round trips in the arms) impinge on the interferometer. How should $\mathcal{R}_r$ be chosen so as to maximize the gravitational-wave signal to noise ratio in the interferometer? Guess the answer on physical grounds before doing the calculation.

c. Suppose that the mirrors in the two arms are slightly imperfect, and their imperfections cause a mismatching of the phase fronts of the light from the two arms at the beam splitter. As a result, the ratio $I_2/I_1 \equiv \alpha$ has some modest value (e.g. 0.01) instead of being arbitrarily small. In this case, how should $\mathcal{R}_r$ be chosen so as to maximize the signal to noise ratio? Guess the answer on physical grounds before doing the calculation.

3. *Scaling of Photon Shot Noise with Arm Length.* We saw in Kip's lecture that, if one has mirrors of sufficiently high reflectivity and one uses a simple (nonrecycled) interferometer, then the photon shot noise $h_{\rm rms} = \sqrt{f G_h(f)}$ is independent of the interferometer's arm length.

   Suppose, instead, that (i) the highest achievable power reflectivity is $\mathcal{R} = 1 - 10^{-5}$, (ii) one can do as good a job of phase-front matching at the interferometer as one wishes, so in the above drawing $I_2/I_1 = \alpha$ can be made as small as one wishes, (iii) one has a fixed laser power I_0 (say, 10 Watts) available, (iv) one operates the interferometer in a power-recycled mode, as in the above figure. *Show* that in this case the photon shot noise $h_{\rm rms}$ scales as $1/\sqrt{L}$ in the full LIGO frequency band (a result quoted on page 314 of Ref. I).

   Note: Another example of arm-length scaling is described on page 316 of Ref. I: A resonant-recycled or dual-recycled interferometer looking for periodic gravitational waves, e.g. from a pulsar, has photon shot noise $h_{\rm rms} \propto 1/L$.

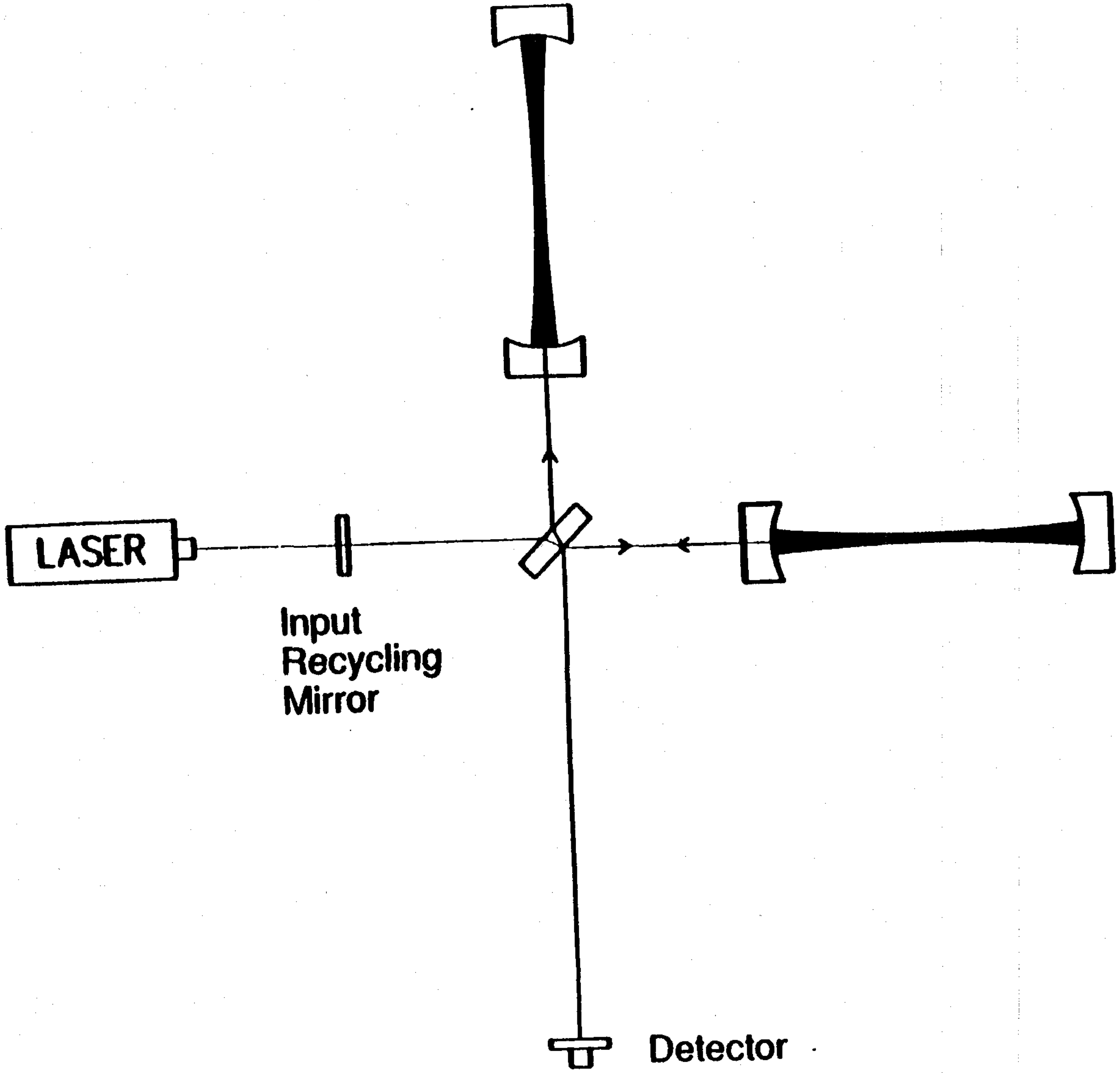
Lecture 5
Idealized Theory of Interferometers — II.

by Ronald W. P. Drever, 13 April 1994

Drever lectured from the attached transparencies. His lecture focussed on optical configurations for interferometers that are more complex than the simple interferometers of Thorne's lecture:

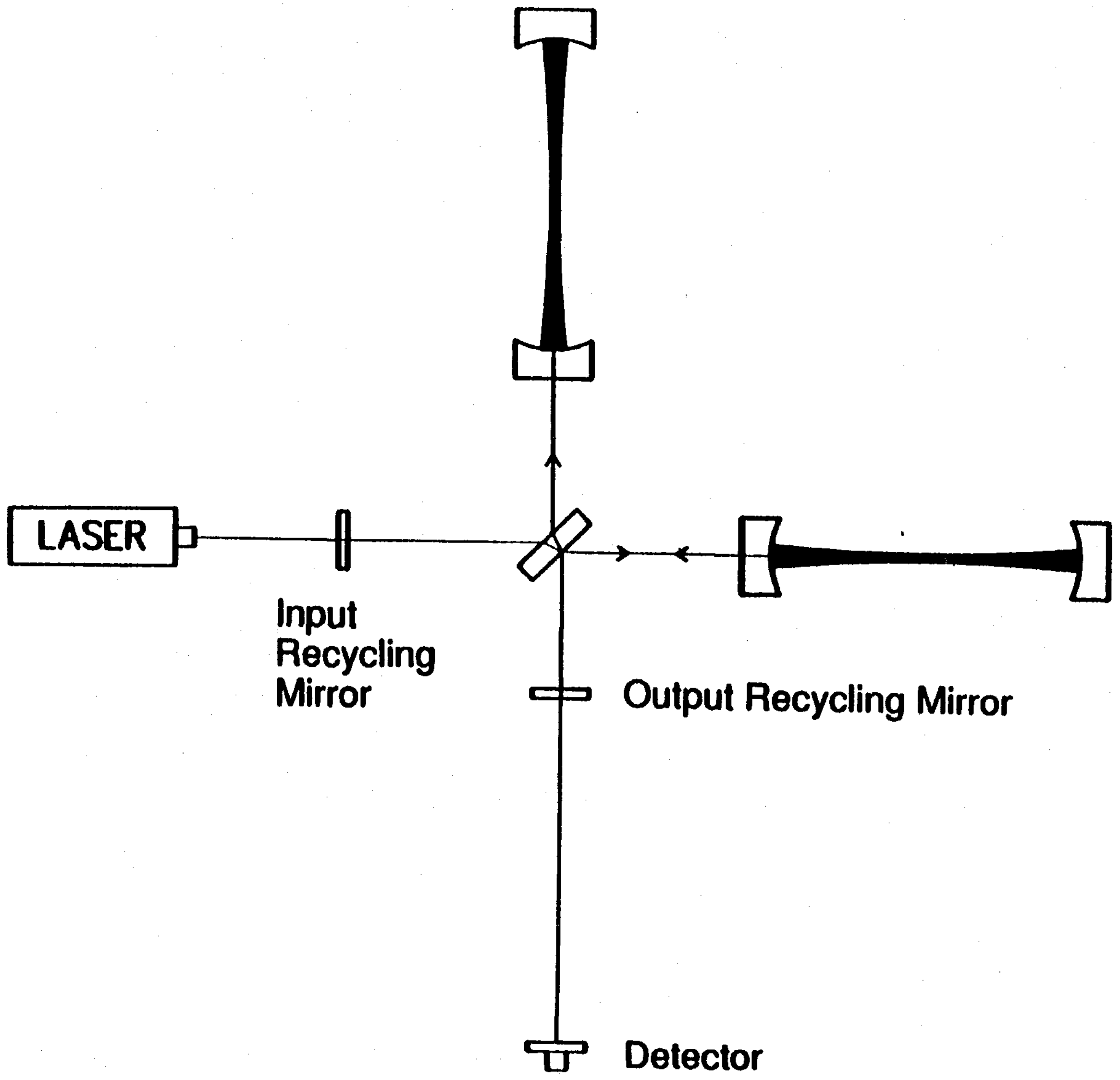
- A *recycling interferometer* (more normally called *power recycling interferometer* these days), in which the light going back from the interferometer toward the laser gets recycled back into the interferometer together with and in phase with new laser light. Such power recycling (invented in the early 1980s by Drever) will be used in LIGO's first interferometers.
- A *dual recycling interferometer* (also called *signal recycling*) in which light power is recycled at the laser's port of the interferometer, and the gravity-wave signal is recycled (and thereby enchanced) at the photodetector's port. Such signal recycling can be used to enhance the interferometer's performance in the vicinity of (most) any desired gravity-wave frequency and over (most) any desired bandwidth around that frequency. It is likely to find application, for example, in deep searches for the gravitational waves from pulsars. A dual recycled interferometers has the added benefit of less sensitivity to irregularities in the mirrors and beam splitter than an ordinary interferometer.
- A *resonant recycling interferometer*, a configuration that accomplishes the same thing as signal recycling but in a more complicated and less practical way. (This configuration was invented by Drever in the early 1980s; signal recycling is an improvement on it, devised by Brian Meers in the late 1980s.)
- A *doubly resonant signal recycling interferometer*. This configuration, invented by Drever and Meers, recycles (and enhances) both signal sidebands of the light's carrier frequency; ordinary signal recycling recycles and enhances only one signal sideband.
- A *resonant sideband extraction interferometer*, whose configuration looks just like that of a dual recycling interferometer but performs quite differently because of a different fine tuning of the location and reflectivity of the recycling mirrors. This configuration, invented recently by M.J. Mizuno (a Japanese graduate student working with the Garching, Germany group) stores the carrier-frequency light in the arms for a long time, while resonantly extracting the signal sideband from the arms after about a half cycle of the gravitational wave. It thereby achieves a broad-band sensitivity comparable to that of a power recycled interferometer, but with far less light power passing through the beam splitter and hence with less problem from high-power heating of the splitter.

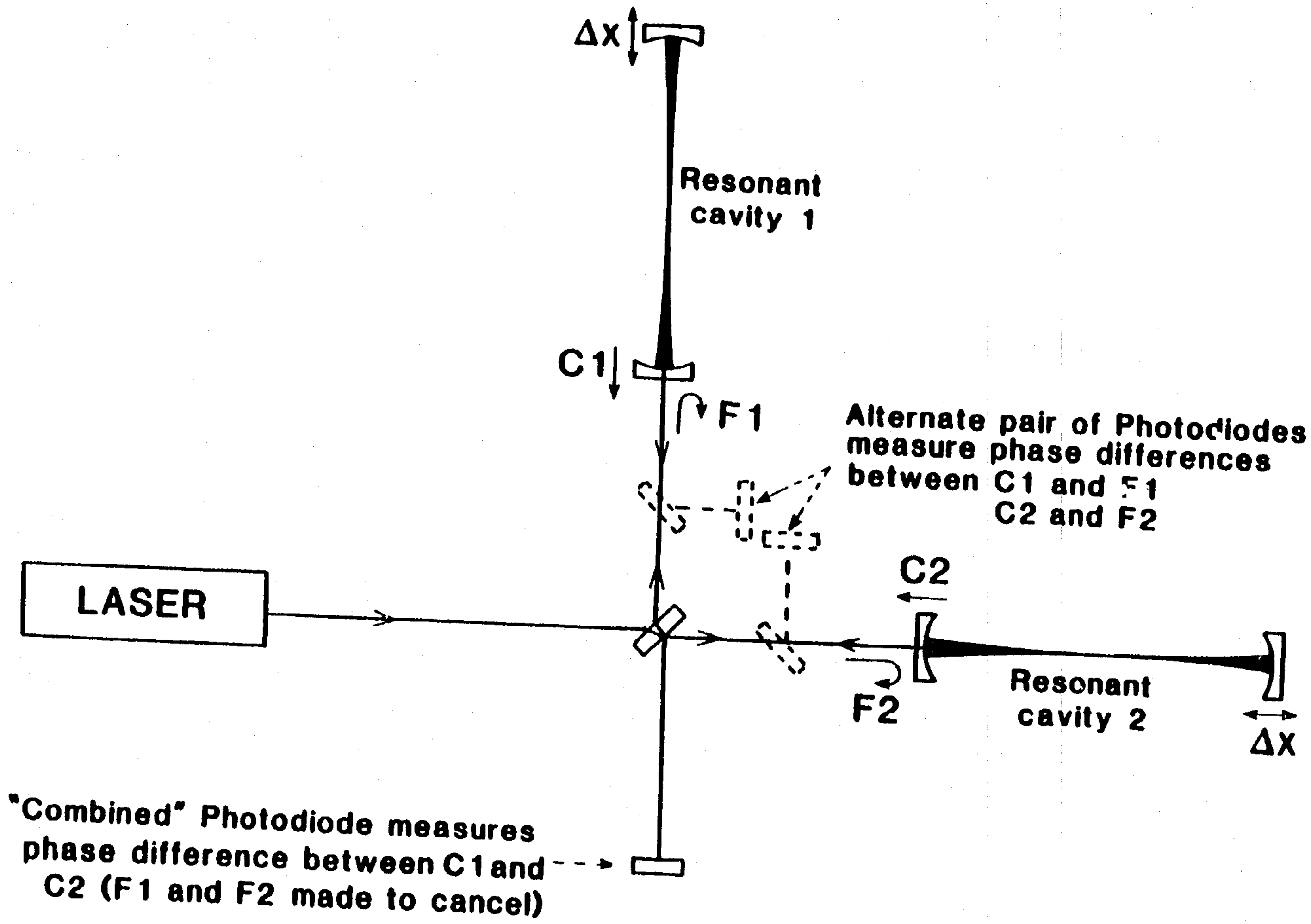
RECYCLING INTERFEROMETER



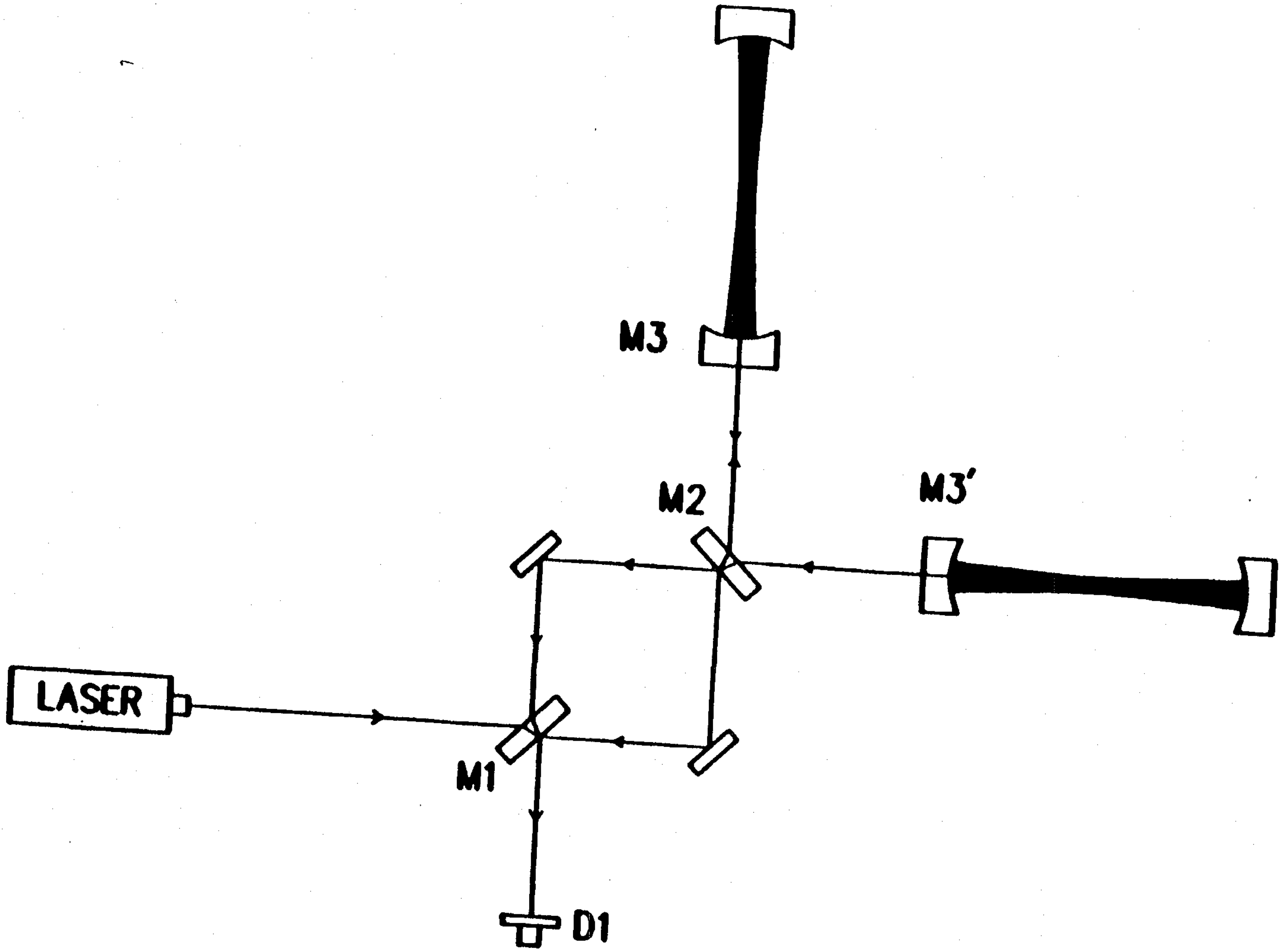
DUAL RECYCLING INTERFEROMETER

(Also Resonant Side Band Extraction Interferometer)

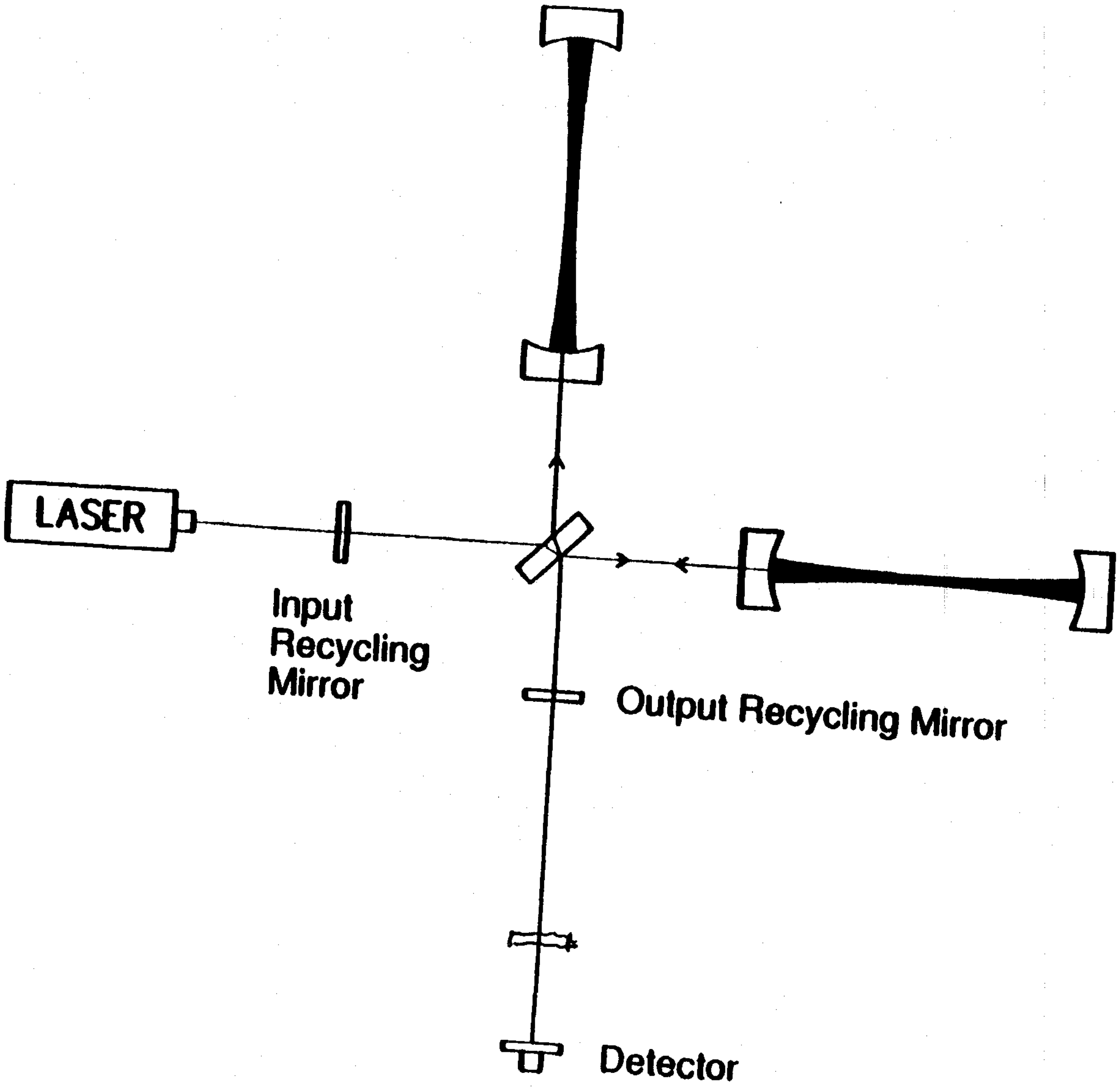


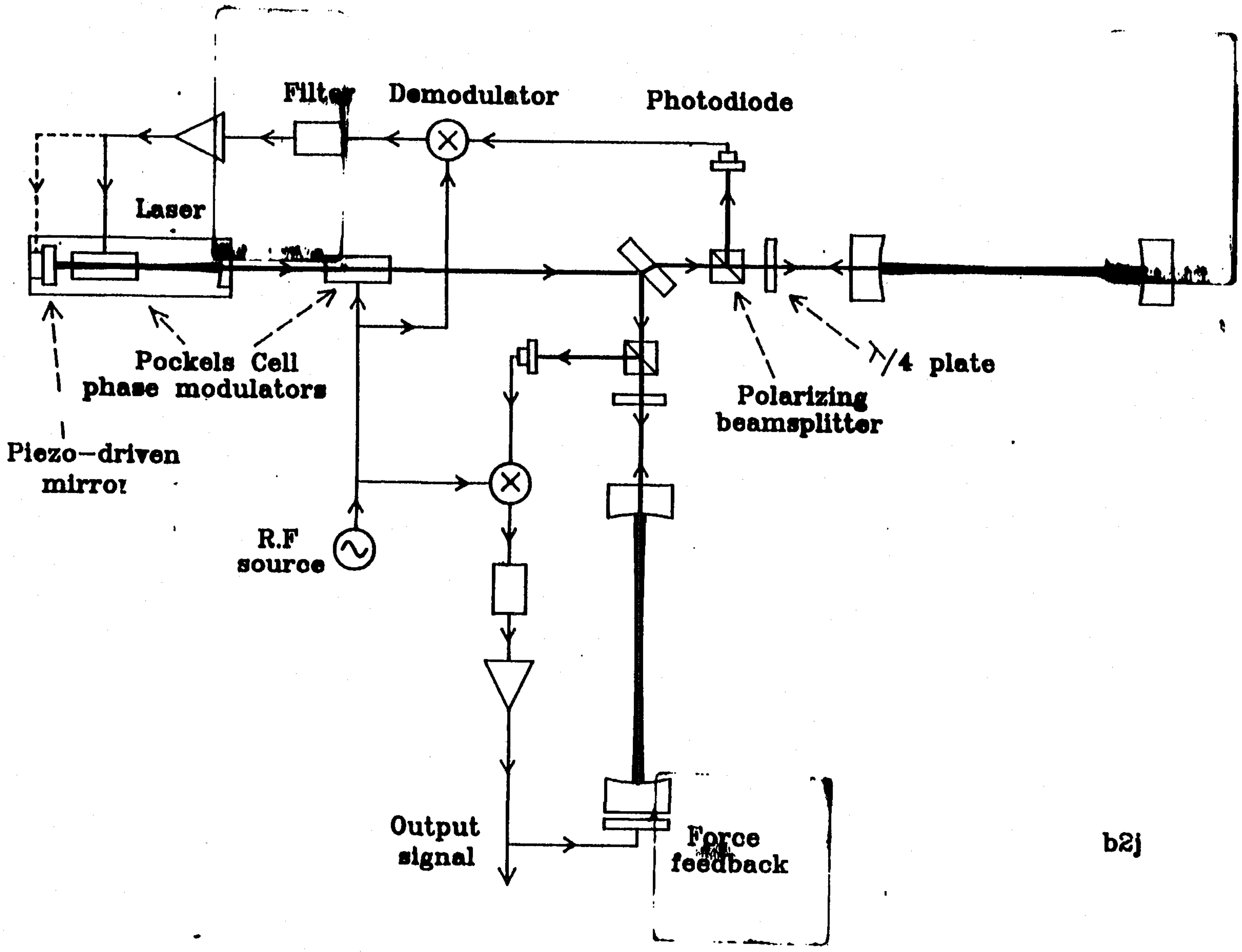

Δx
Resonant
cavity 1
C1
F1
Alternate pair of Photodiodes
measure phase differences
between C1 and F1
C2 and F2
C2
LASER
F2
Resonant
cavity 2
Δx
"Combined" Photodiode measures
phase difference between C1 and
C2 (F1 and F2 made to cancel)

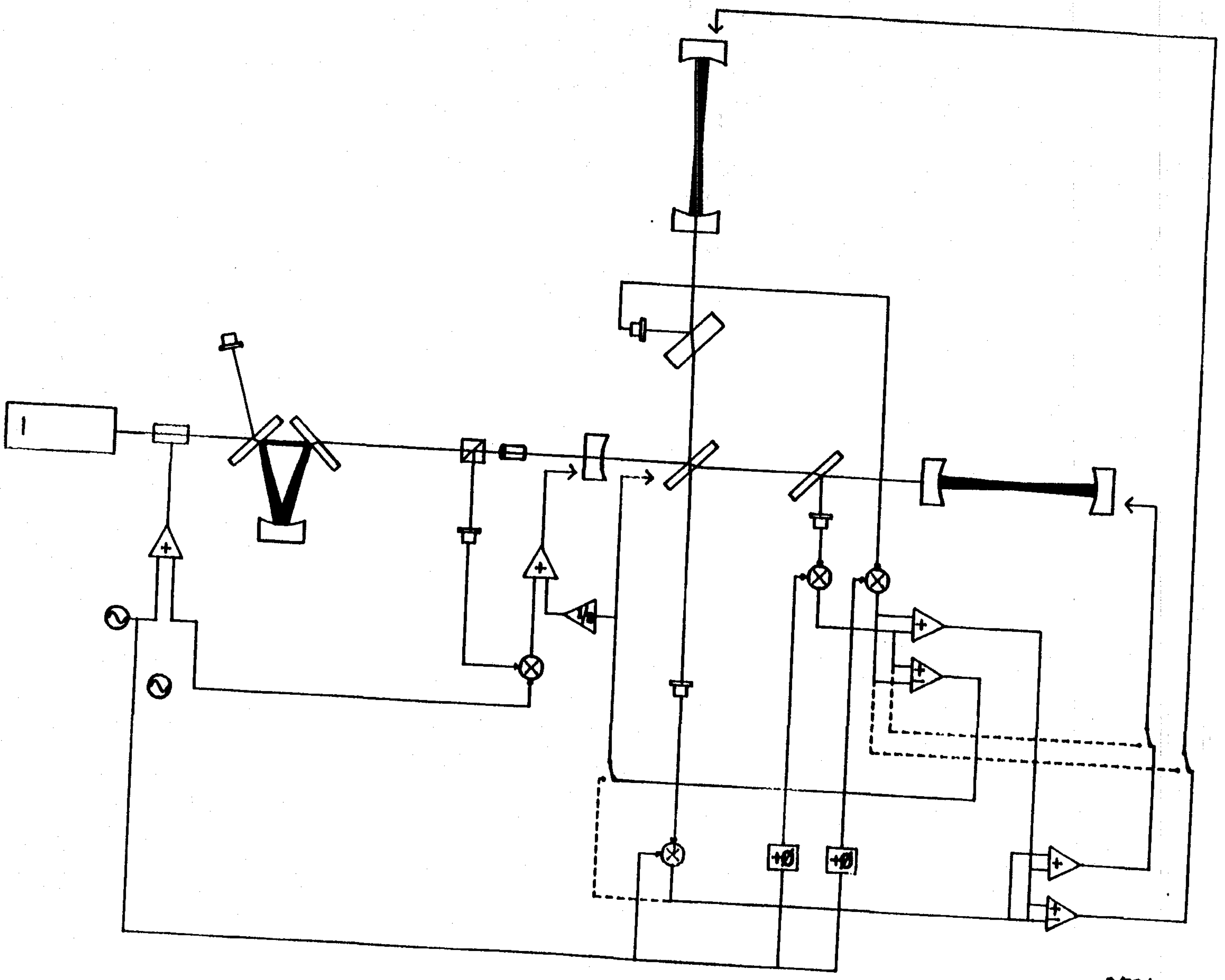
RESONANT RECYCLING INTERFEROMETER



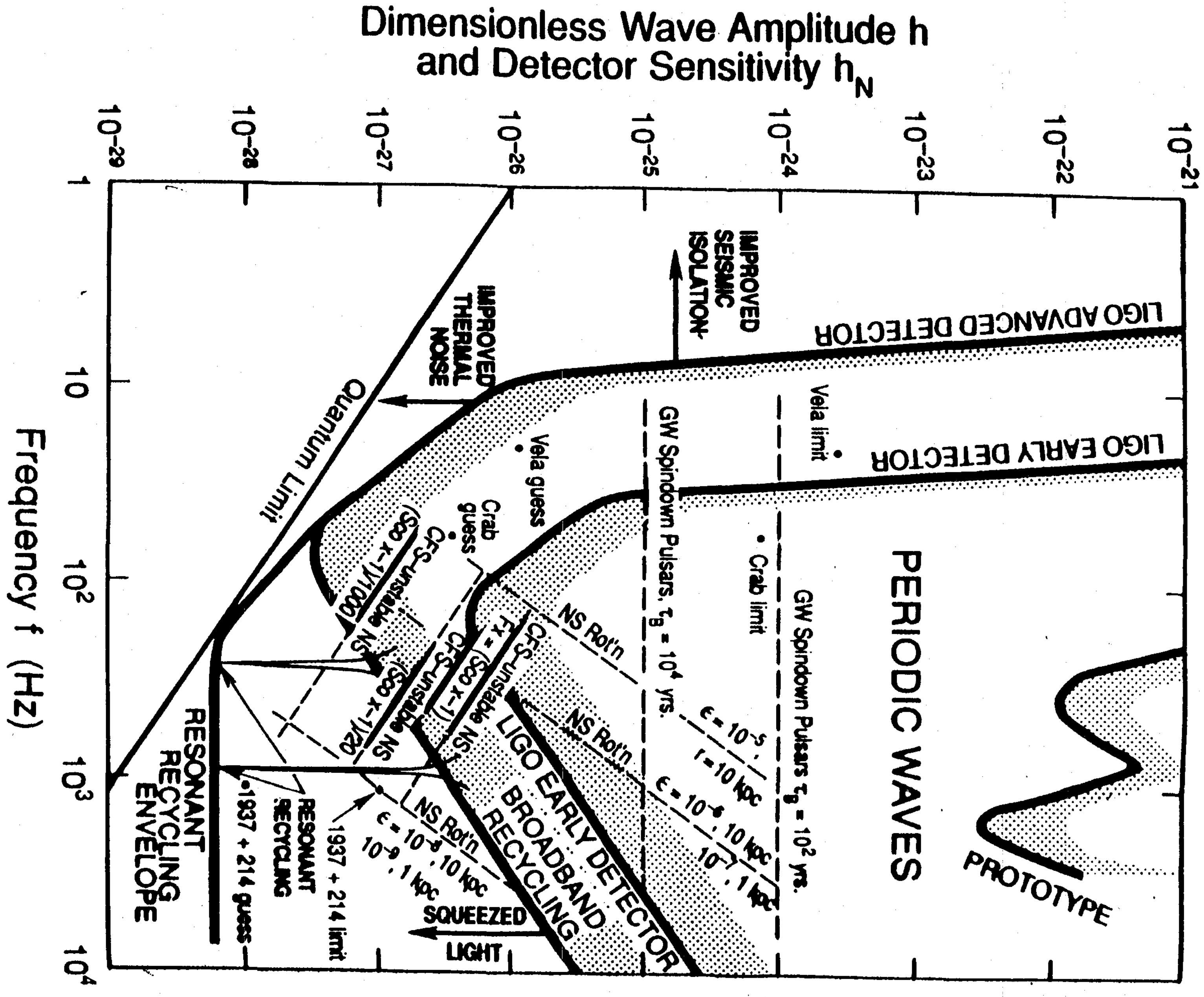
DOUBLY-RESONANT SIGNAL RECYCLING INTERFEROMETER

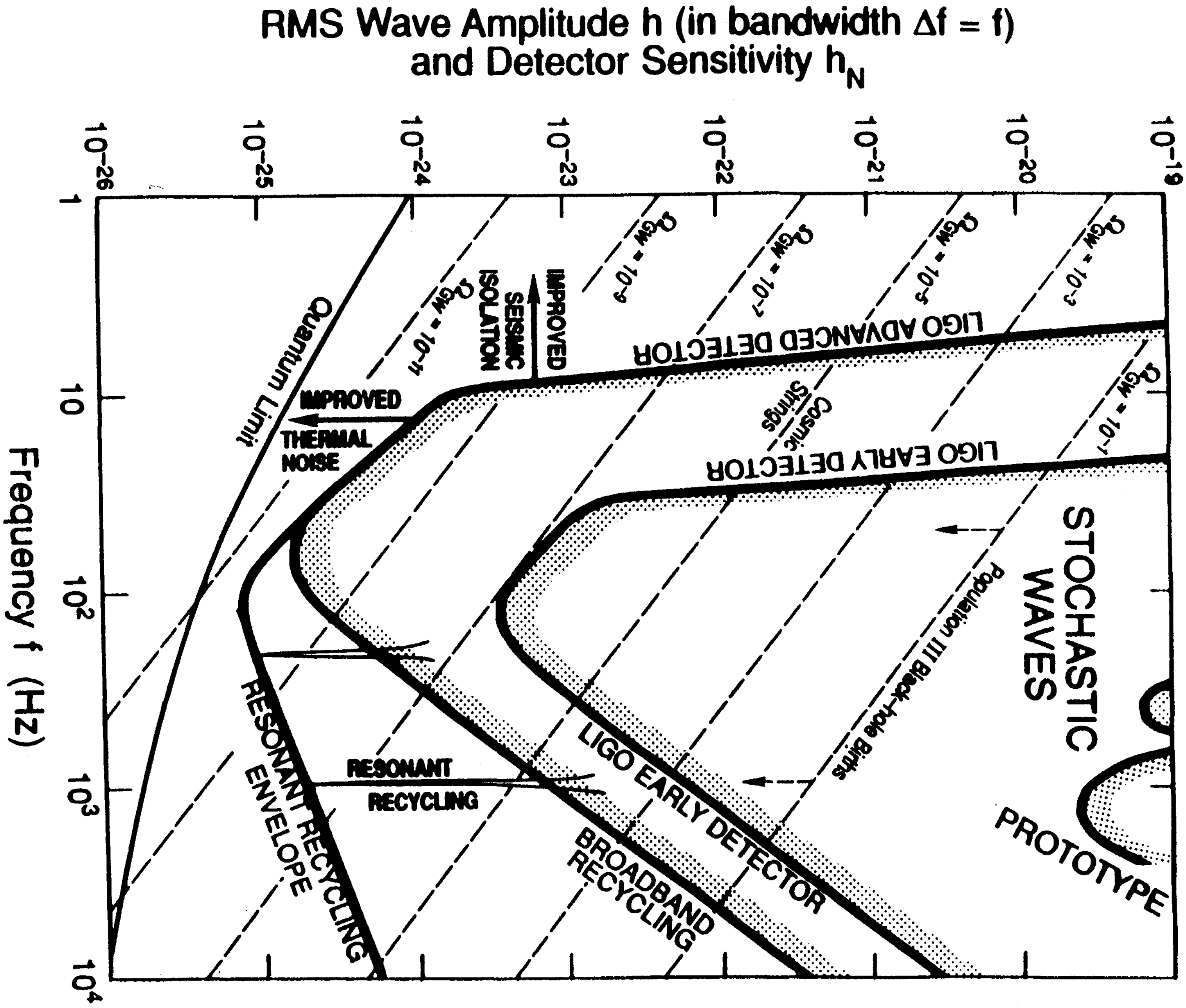



Filter
Demodulator
Photodiode
Laser
Pockels Cell
phase modulators
Piezo–driven
mirror
R.F
source
Polarizing
beamsplitter
λ/4 plate
Output
signal
Force
feedback



9/22/92
[illegible]


Dimensionless Wave Amplitude h
and Detector Sensitivity h_N
10^{-29}
10^{-28}
10^{-27}
10^{-26}
10^{-25}
10^{-24}
10^{-23}
10^{-22}
10^{-21}
Frequency f (Hz)
1
10
10^2
10^3
10^4
PERIODIC WAVES
PROTOTYPE
LIGO ADVANCED DETECTOR
LIGO EARLY DETECTOR
IMPROVED SEISMIC ISOLATION
IMPROVED THERMAL NOISE
Quantum Limit
Vela limit
Vela guess
Crab limit
Crab guess
GW Spindown Pulsars $\tau_g = 10^2$ yrs.
GW Spindown Pulsars, $\tau_g = 10^4$ yrs.
NS Rot'n
$\epsilon = 10^{-5}$, $r = 10$ kpc
NS Rot'n
$\epsilon = 10^{-6}$, 10 kpc
10^{-7}, 1 kpc
NS Rot'n
$\epsilon = 10^{-8}$, 10 kpc
10^{-9}, 1 kpc
CFS-unstable NS (Sco x-1)/1000
CFS-unstable NS (Sco x-1)/20
CFS-unstable NS, F.x = (Sco x-1)
LIGO EARLY DETECTOR BROADBAND RECYCLING
SQUEEZED LIGHT
1937 + 214 limit
1937 + 214 guess
RESONANT RECYCLING
RESONANT RECYCLING ENVELOPE


RMS Wave Amplitude h (in bandwidth Δf = f)
and Detector Sensitivity h_N
10^{-26}
10^{-25}
10^{-24}
10^{-23}
10^{-22}
10^{-21}
10^{-20}
10^{-19}
1
10
10^2
10^3
10^4
Frequency f (Hz)
$\Omega_{GW} = 10^{-3}$
$\Omega_{GW} = 10^{-5}$
$\Omega_{GW} = 10^{-7}$
$\Omega_{GW} = 10^{-9}$
$\Omega_{GW} = 10^{-11}$
$\Omega_{GW} = 10^{-1}$
Quantum Limit
IMPROVED SEISMIC ISOLATION
IMPROVED THERMAL NOISE
LIGO ADVANCED DETECTOR
LIGO EARLY DETECTOR
Cosmic Strings
Population III Black-hole Births
STOCHASTIC WAVES
PROTOTYPE
LIGO EARLY DETECTOR
BROADBAND RECYCLING
RESONANT RECYCLING
RESONANT RECYCLING ENVELOPE

BATCH
START

⑥ Interferometer

STAPLE
OR
DIVIDER

LECTURE 6.
OVERVIEW OF A REAL INTERFEROMETER

Lecture by Stanley E. Whitcomb

Assigned Reading:

O. D. Shoemaker, R. Schilling, L. Schnupp, W. Winkler, K. Maischberger, A. Rüdiger, "Noise behavior of the Garching 30-meter prototype gravitational-wave interferometer," *Physical Review D*, **38**, 423-432 (1988). [This is a fairly complete description of a prototype delay-line interferometer, with all the complications of a real device.]

P. The first one or one and a half chapters of any introductory text on servosystems (also called closed loop control systems or servo loops). Don't labor slavishly over the mathematical details, but do try to get a "feel" for how servo loops work. One possibility for this is the first chapter of Benjamin C. Kuo, *Automatic Control Systems* (Prentice-Hall), which is being passed out to the class. Note that this chapter is very qualitative; you might want to dig into other books for more quantitative detail.

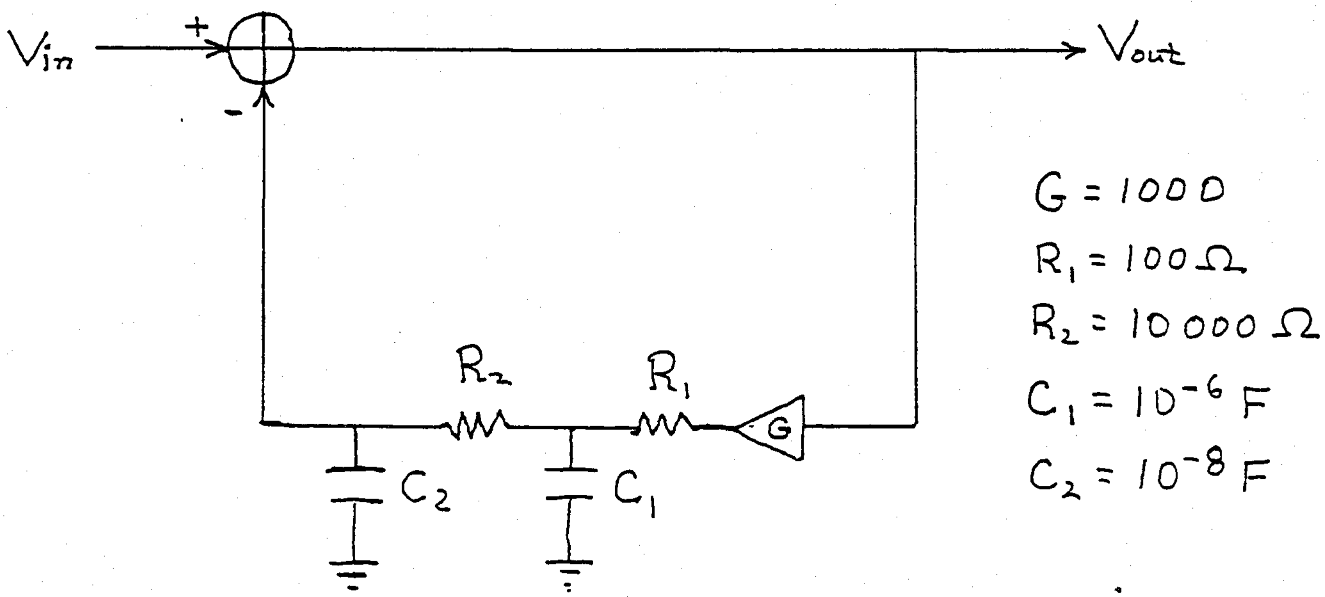
Suggested Supplementary Reading:

Read more deeply into your favorite servo text.

A Few Suggested Problems

1. *Garching Prototype.* There was a discrepancy between the observed and predicted noise spectrum in the Garching prototype (Figure 4 of Reference 1 above). The spectrum's shape is about right, suggesting that maybe the Garching group identified the right noise sources but made a calibration error that produced the numerical disagreement. On the other hand, the disagreement is approximately a factor 3, which seems a large error for a group that is generally regarded as very careful. What do you think about this? What information in the paper might lead you to one conclusion or another about the discrepancy?

2. *Example of a Servo loop.* Consider the following servo loop in an electronic circuit. It is designed to strongly suppress the input voltage V_{in} at frequencies well below some critical frequency ω_o, and pass the voltage signal more or less unchanged at frequencies well above ω_o. For the values of the servo amplifier gain G and the resistances and capacitances shown in the figure, what is the frequency ω_o? Is there any frequency region in which the servo is unstable, in the sense that it strongly amplifies the input voltage signal (i.e., it oscillates with large amplitude when a small amplitude stimulus is applied)? If so, how might you change the circuit to get rid of the unwanted amplification, while maintaining the original goals of voltage suppression well below ω_o and passing the signal unscathed well above ω_o? [*Note* (for theorists who might not know such things): The device symbolized is a voltage amplifier with gain G and it can be regarded as having infinite input impedance and zero output impedance; the device symbolized produces an output that is the difference of its two inputs, and it can be regarded as having infinite input impedances.]

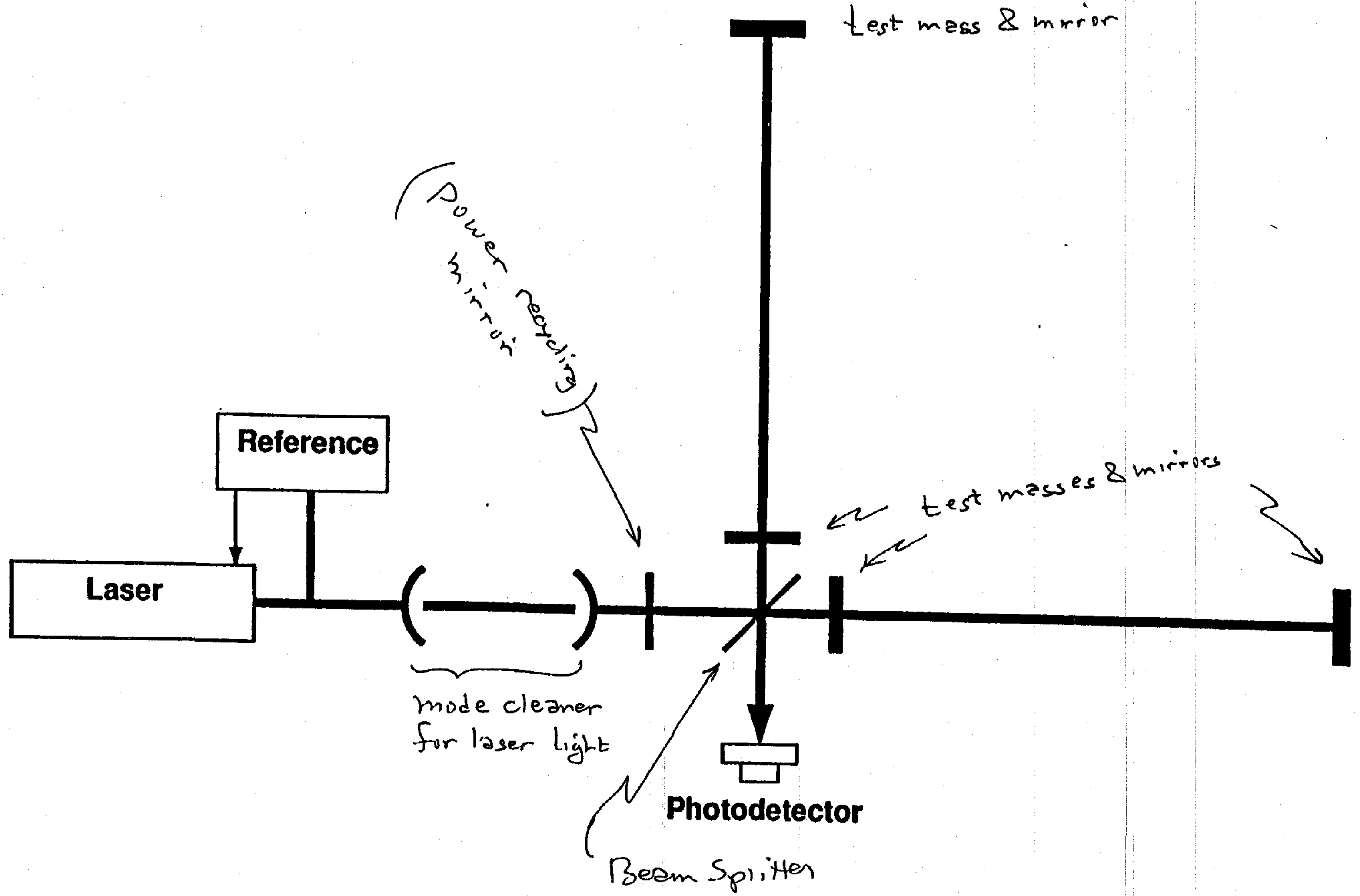


Lecture 6
Overview of a Real Interferometer

by Stanley E. Whitcomb, 15 April 1994

Whitcomb lectured from the following transparencies. Kip has annotated them, based on Whitcomb's lecture.

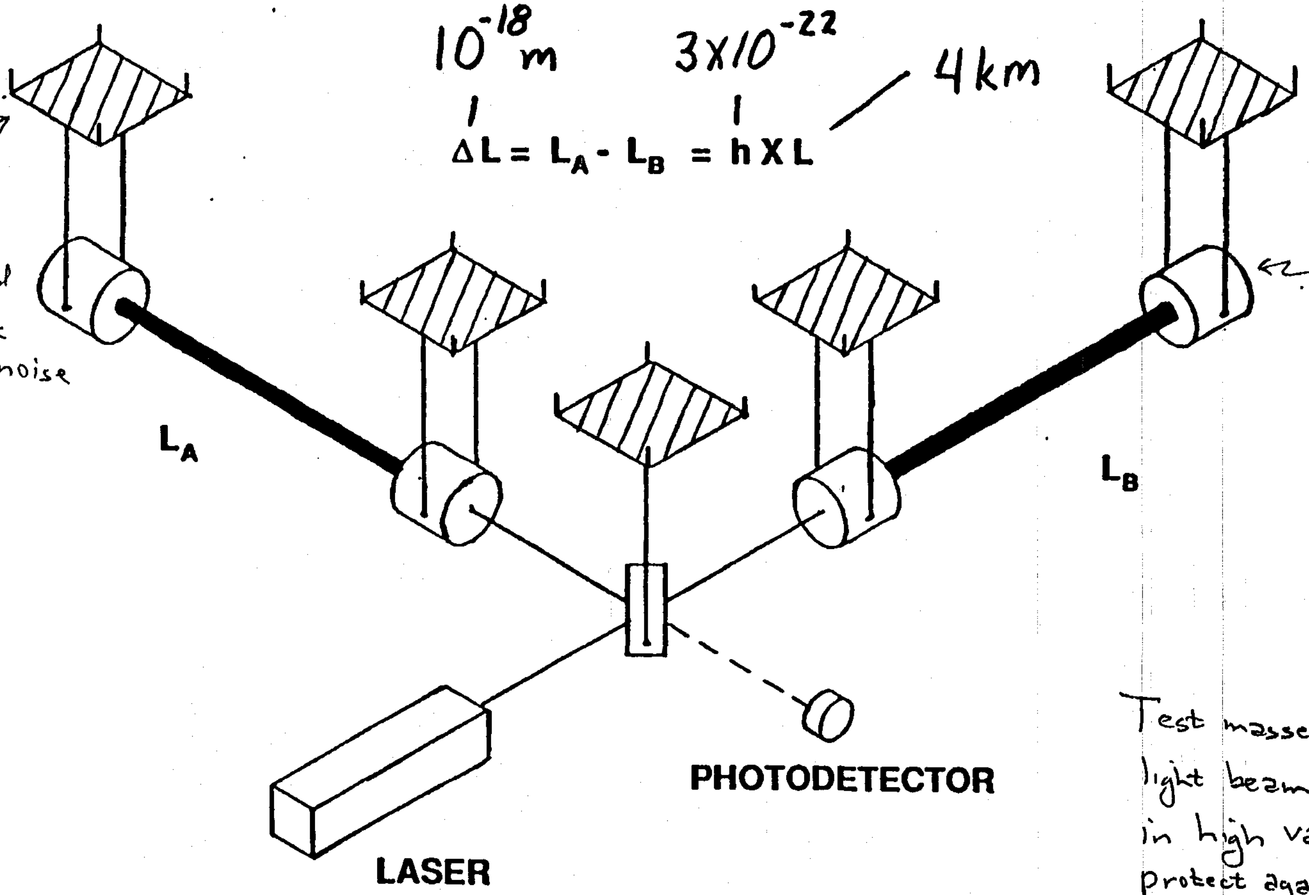
Optical Layout and Operation



WHAT'S IN A REAL INTERFEROMETER?

	______'s GUESS	CIT 40-m
# PHOTODIODES		52
# LENSES		20
# SERVOCONTROL LOOPS		
MASS OF IFO (EXCLUDING VAC. SYSTEM)		6000kg

SCHEMATIC INTERFEROMETRIC DETECTOR



Suspension points must be ~~pr~~ isolated from seismic & acoustic noise

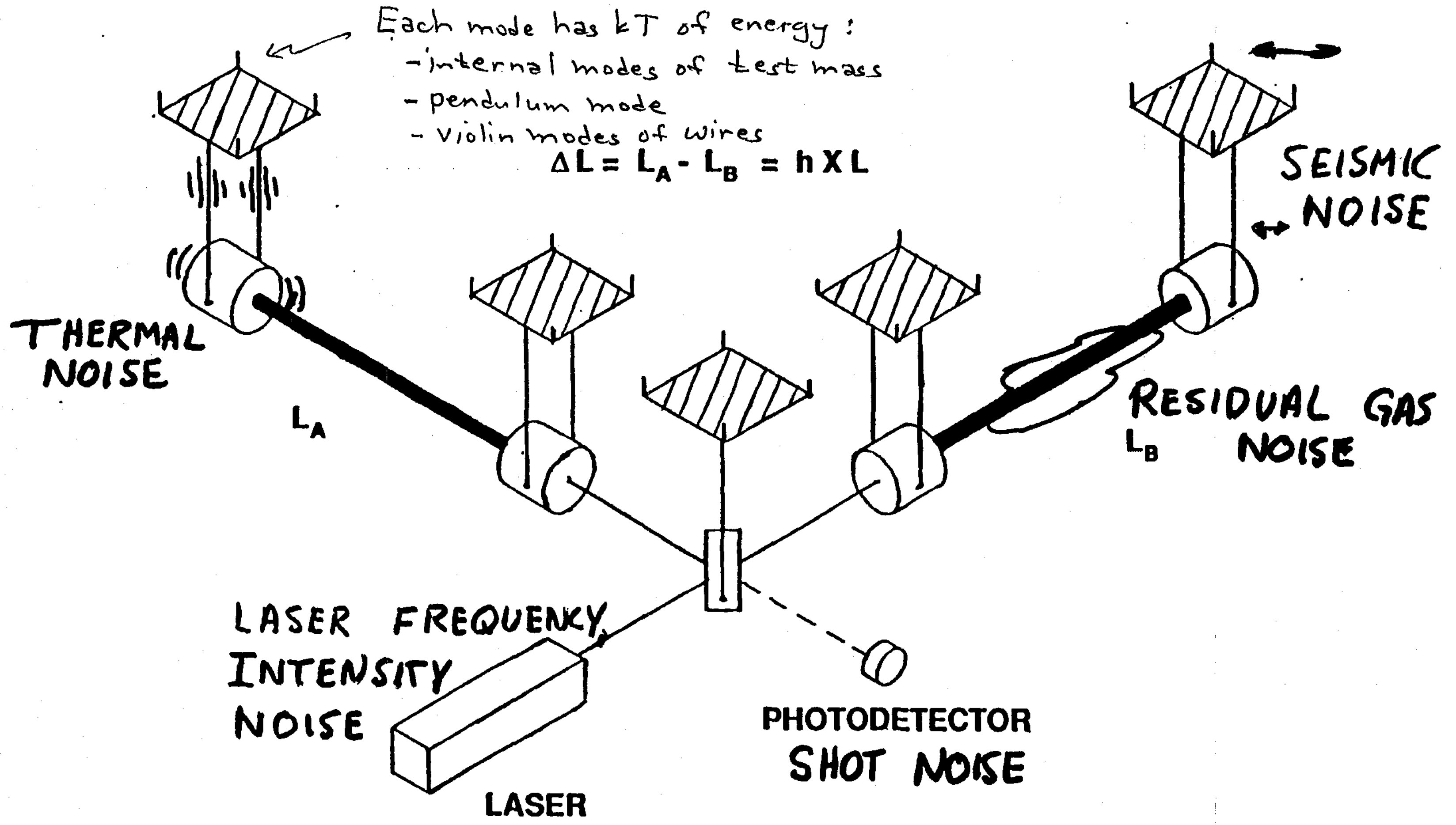
mirrors must be held in position to accuracy $\simeq 10^{-12}$ m

frequency stability must be improved by ~10^{10}; amplitude stability, by ~100 to 1000

Test masses and light beam must be in high vacuum to protect against

- buffeting of test masses by air molecules
- fluctuations of index of refraction. causing fluctuations of phase shift

SCHEMATIC INTERFEROMETRIC DETECTOR



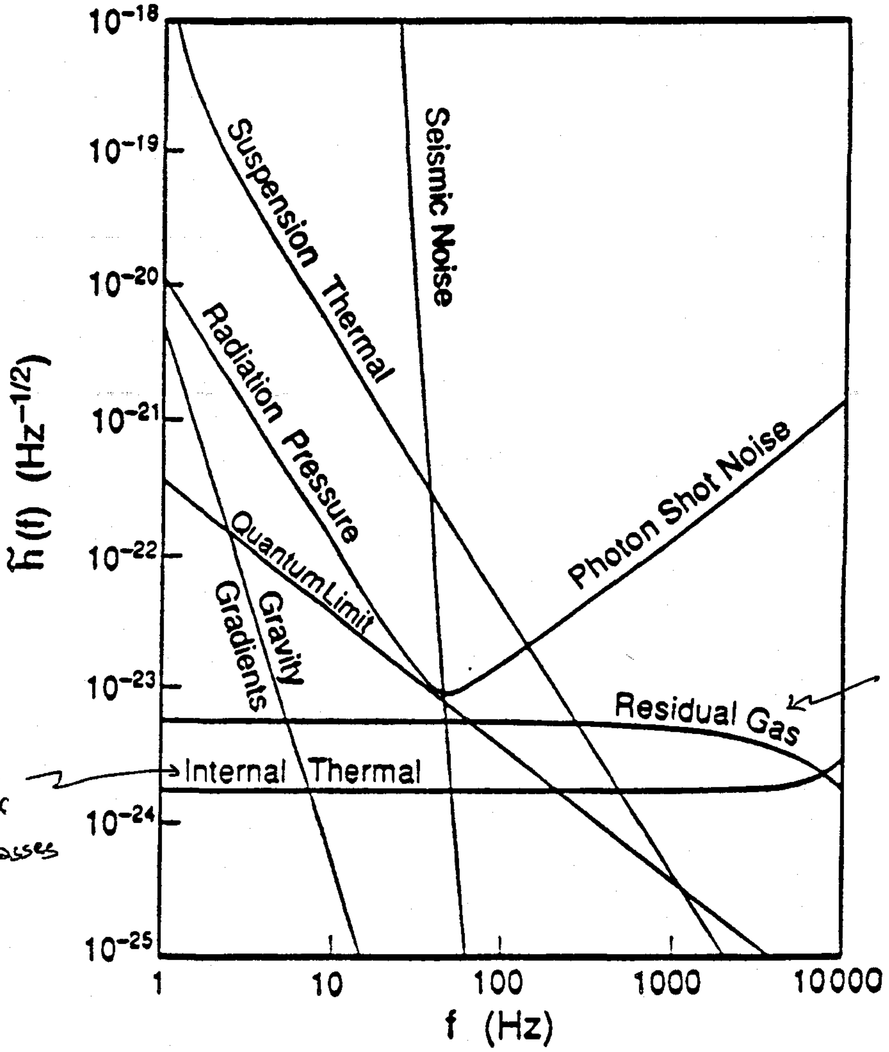
Each mode has kT of energy :
- internal modes of test mass
- pendulum mode
- violin modes of wires

$$\Delta L = L_A - L_B = h \times L$$

Noise Budget For First LIGO Detectors

- **5 Watt Laser**
- **Mirror Losses 50 ppm**
- **Recycling Factor of 30**
- **10 kg Test Masses**
- **Suspension Q=10^7**

One makes a design, then computes the noise of each source, then adds the noises in quadrature, then iterates the design...



Internal modes of test masses

Index of refractin fluctuations

LIGO

A USEFUL CLASSIFICATION SCHEME ← of noise sources

DISPLACEMENT NOISE

REAL MOTION OF MIRROR SURFACE

EXAMPLES: SEISMIC NOISE
THERMAL NOISE
ELECTRONIC NOISE WHICH DRIVES CONTROL OF T.M.

SENSING NOISE

NOISE WHICH APPEARS IN THE READ-OUT SYSTEM (OPTICAL INTERFEROMETER + PHOTODETECTORS) AND IS INTERPRETED AS APPARENT MOTIONS

EXAMPLES: SHOT NOISE
RESIDUAL GAS NOISE
ELECTRONIC NOISE IN PHOTODETECTOR

A (NOT VERY USEFUL) CLASSIFICATION ← of noise Sources

FUNDAMENTAL VS. TECHNICAL

FUNDAMENTAL

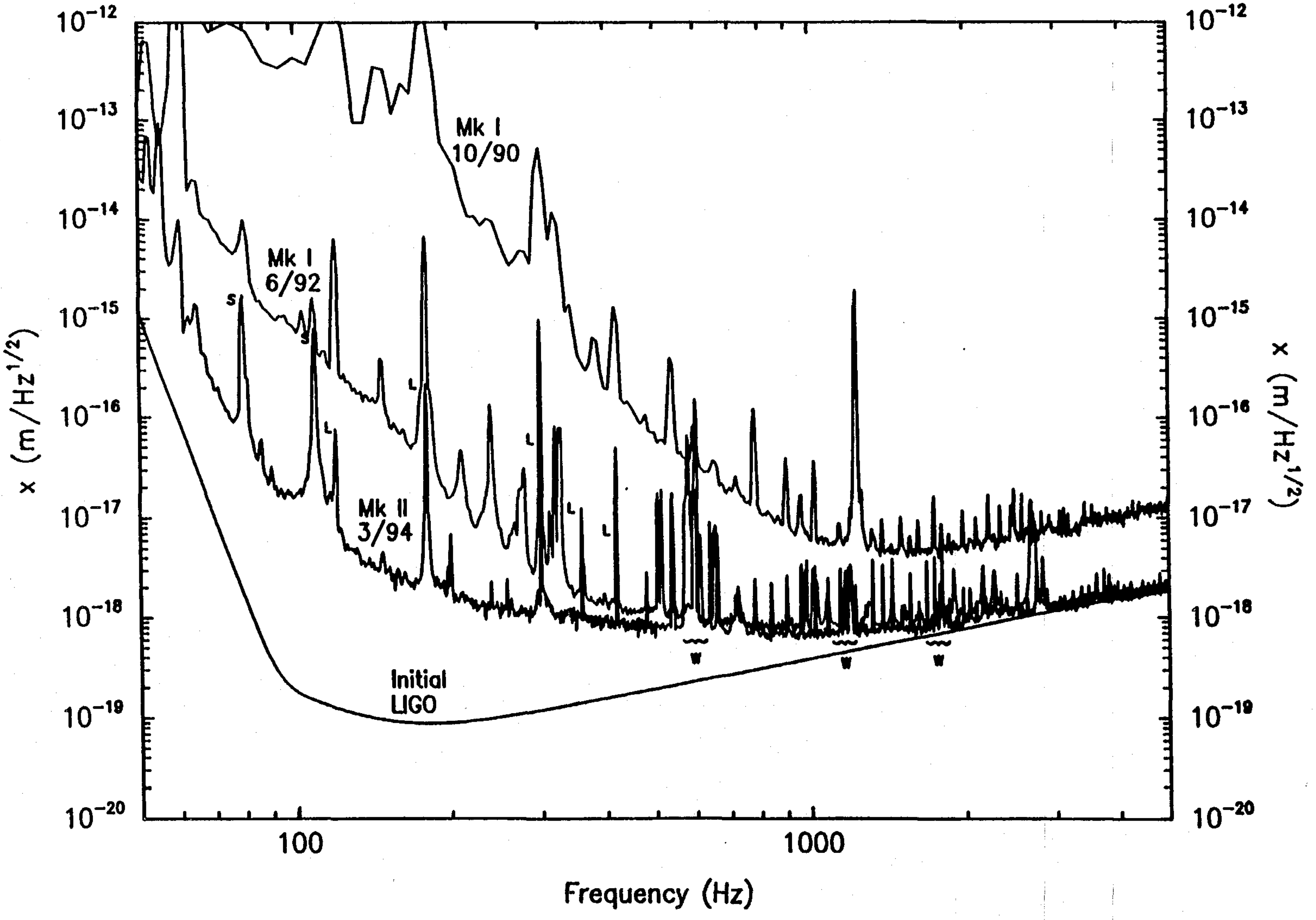
INTERESTING NOISE SOURCES, USUALLY WORKED ON BY PHYSICISTS, INVOLVE $\hbar$, kT, Q, ...

TECHNICAL

NOISE SOURCES SOMEONE ELSE SHOULD WORK ON, INVOLVE

$$\underbrace{e^2(f)\,,\; i^2(f)}_{\text{Voltage \& current noise in amplifiers}}\,,\; \underbrace{R}_{\text{Johnson noise in resistors}}\,,\; \left.\frac{dV}{dt}\right|_{max}\,,\; \ldots$$

Displacement Sensitivity of Caltech 40 m Interferometer



SCALING

(ONE OF THE) MOST COMMON QUESTIONS:

- HOW DOES NOISE SCALE WITH _X_?
- _X_ TYPICALLY LENGTH (OR LASER POWER, OR TEST MASS Q, OR ...)

WRONG ANSWER:

"THE NOISE SCALES AS $\frac{1}{L}$"

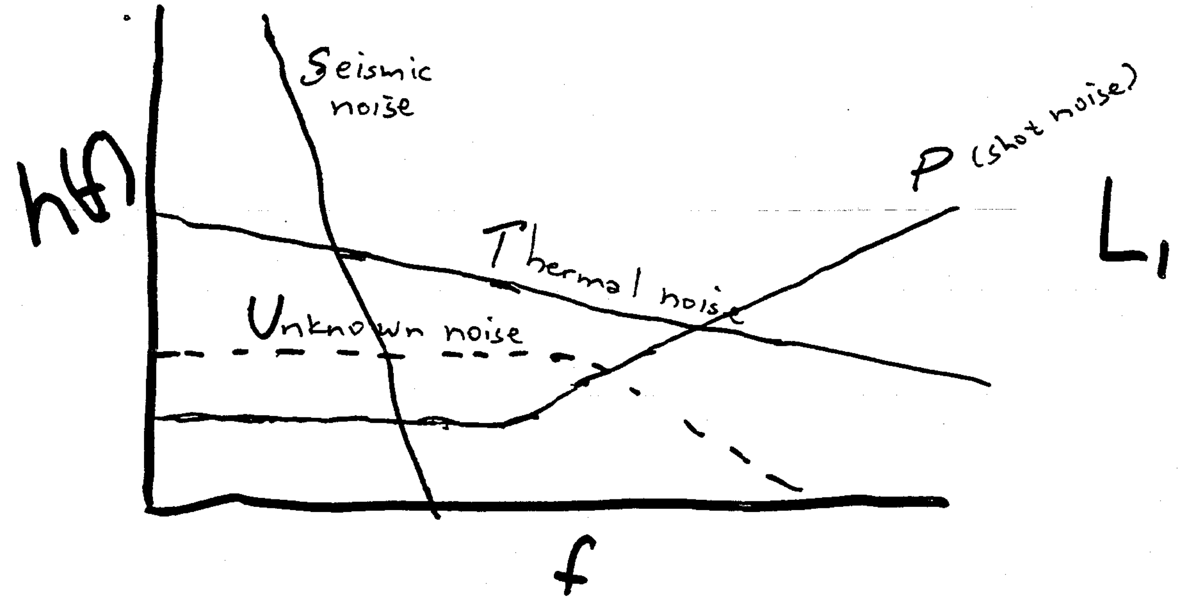
OR $1/L^{1/2}$ OR L^0 ...

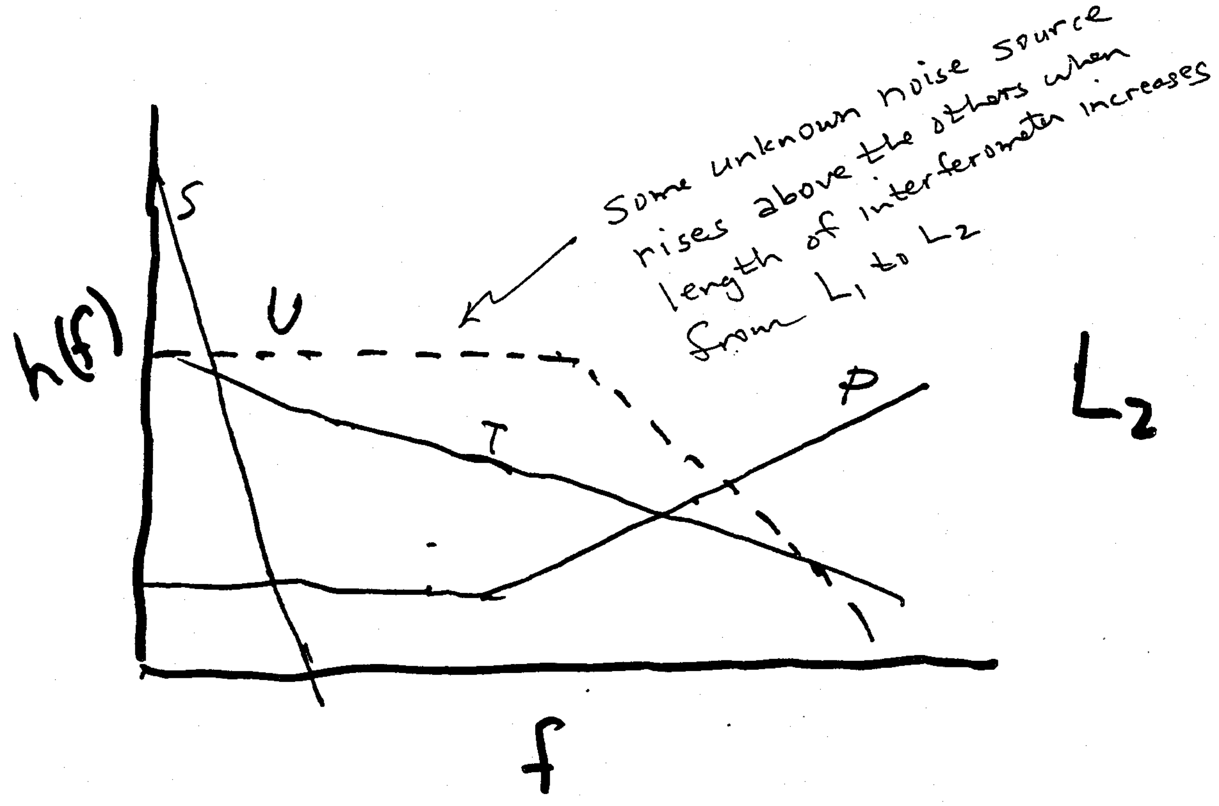
OR $P^{-1/2}$ OR ...

RIGHT ANSWER:

NOISE SPECTRUM IS A COMPOSITE OF SEVERAL COMPONENT CURVES (CORRESPONDING TO DIFFERENT SOURCES). EACH SOURCE HAS ITS OWN CHARACTERISTIC SCALING WITH THE INTERFEROMETER PARAMETERS (**L** OR **P** OR ...). TO PREDICT THE SCALING FROM L_1 TO L_2 (OR P_1 TO P_2), ONE MUST DECOMPOSE L_1 SPECTRUM INTO COMPONENT PIECES, THEN SCALE EACH TO L_2, THEN COMBINE.

EXAMPLE





Some unknown noise source rises above the others when length of interferometer increases from L_1 to L_2

IMPLICIT VS. EXPLICIT SCALING

ALL NOISE SOURCES HAVE EXPLICIT DEPENDENCE ON L (OR P OR Q OR...) AND AN IMPLICIT DEPENDENCE

EXAMPLE: INTERNAL MODE THERMAL NOISE

$$\tilde{h}_{IT}(f) = \frac{\tilde{x}_{IT}(f)}{L}$$

EXPLICIT DEPENDENCE L^{-1}

IMPLICIT DEPENDENCE

SIZE OF T.M. SCALES AS $\sqrt{L}$

ω_{MODE} SCALES AS $1/\text{SIZE}$

Q_{MODE} MAY CHANGE WITH $\frac{\text{VOLUME}}{\text{AREA}}$

NUMBER OF CONTRIBUTING MODES CHANGES WITH BEAM SIZE

⋮

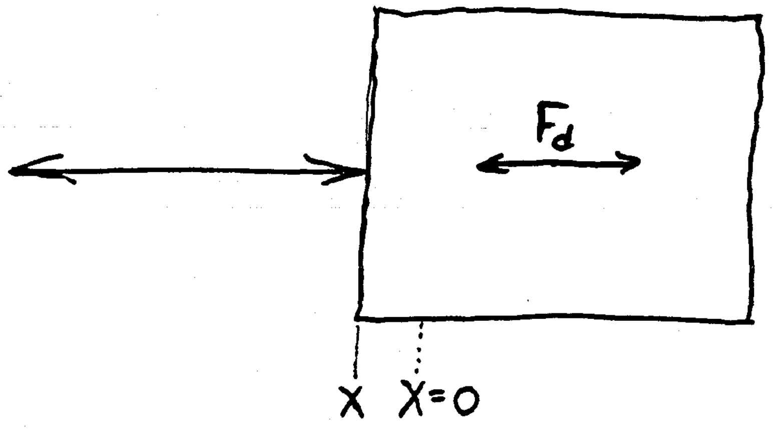
EXAMPLE 2: SHOT NOISE

EXPLICIT: $\tilde{h}_{SN}(f) \propto P^{-1/2}$

IMPLICIT: ? ← [Homework Problem]

WHY DO WE NEED SERVO'S ?

EXAMPLE



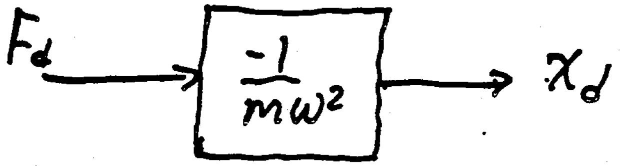
DISTURBING FORCE $F_d = F_{do} e^{i\omega t}$

CAUSE DISPLACEMENT OF MASS FROM DESIRED POSITION

$$\ddot{X} = F/m$$

$$\Rightarrow \quad X = X_{do}\, e^{i\omega t}$$

$$X_{do} = -\frac{F_d}{m\omega^2}$$

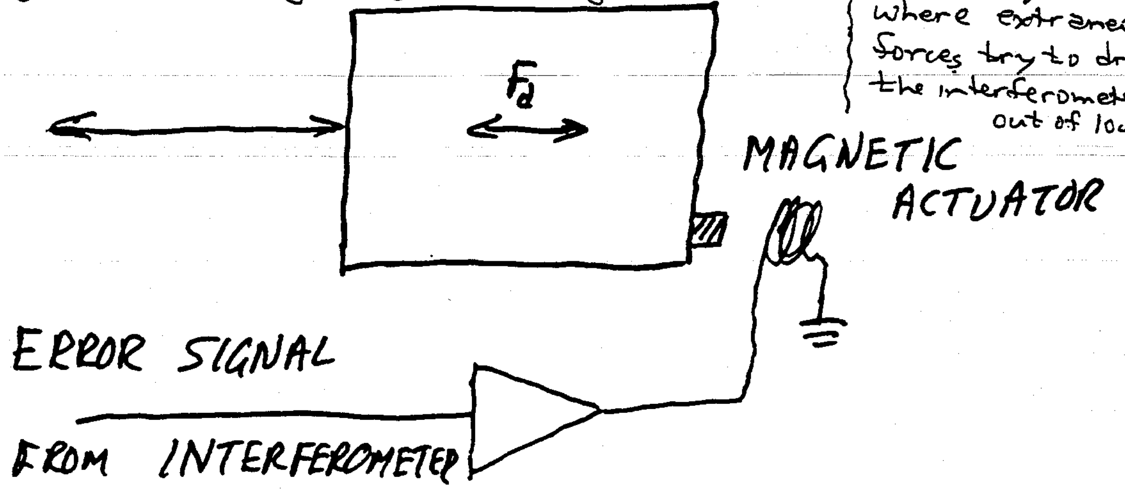


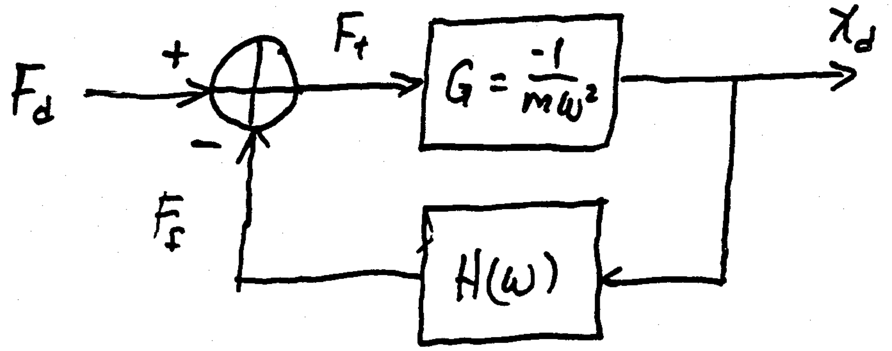
$G(\omega) \equiv$ TRANSFER FUNCTION

SERVOLOOP – FIRST ATTEMPT

If the test-mass position is servoed as shown at the gravity-wave frequency, then the servo signal is our gravity wave signal.

But this servoing is mostly done at lower frequencies where extraneous forces try to drive the interferometer out of lock.





FIRST GUESS

$$H(\omega) = k\,X$$

$$F_t = F_d - F_f$$

$$= F_d + \frac{k}{m\omega^2} F_t$$

$$\Rightarrow \quad F_t = \frac{F_d}{1 - \frac{k}{m\omega^2}} = \frac{F_d}{1 - \frac{\omega_0^2}{\omega^2}}$$

THE <u>GOOD</u>

$$X_{d,} = -\frac{1}{m\omega^2} F_t$$

$$= \frac{F_d}{k - m\omega^2}$$

FOR $\omega << \sqrt{\frac{k}{m}} \equiv \omega_0$

$$X_d = \frac{F_d}{k}$$

cf. $X_d = \frac{F_d}{m\omega^2}$ WITHOUT LOOP

SINCE $k >> m\omega^2$

EFFECT OF F_d HAS BEEN REDUCED at low ω

NOTE: THE LARGER k IS THE MORE X_d IS REDUCED.

"HIGH GAIN IS GOOD" (SOMETIMES)

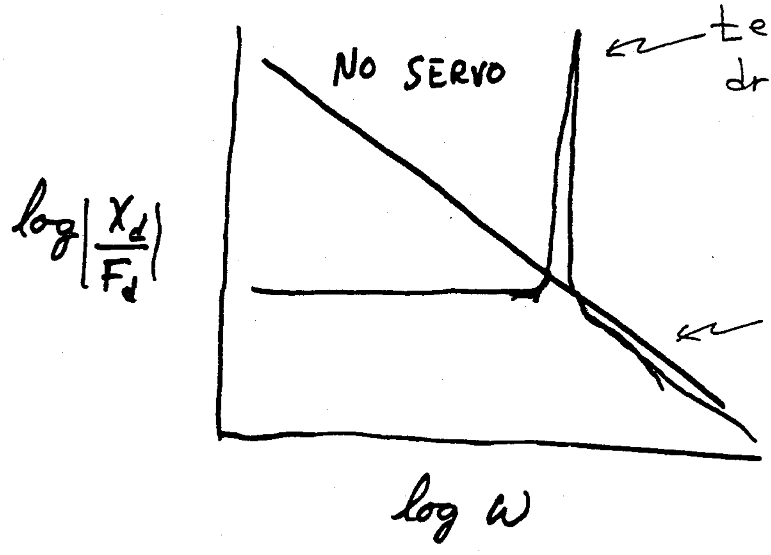
THE BAD

$$X_d = \frac{F_d}{K - m\omega^2}$$

AS $\omega \rightarrow \sqrt{\frac{K}{m}}$

$$\left|\frac{X_d}{F_d}\right| \rightarrow \infty$$

OSCILLATION



test mass oscillates, driven by servo system

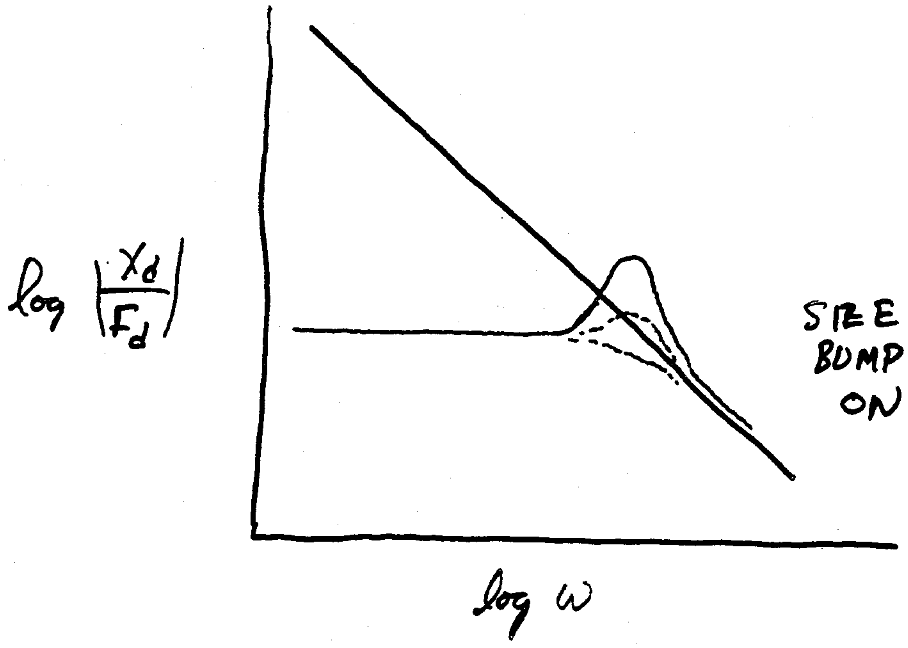
response at high f is like that of an unservoed mass

THE FIX

WE HAVE EFFECTIVELY MADE A MASS ON A SPRING. TO SUPPRESS OSCILLATIONS WE ADD DAMPING

$$H(\omega) = kx + b\dot{x}$$
$$= (k + i\omega b)\,x$$

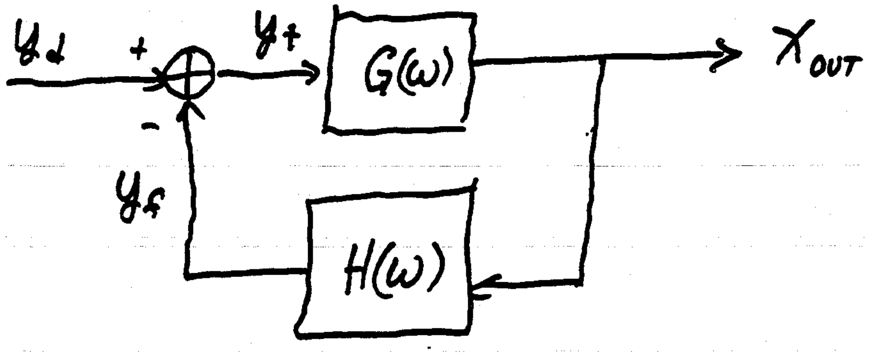
$$\Rightarrow \quad x_d = \frac{F_d}{k - m\omega^2 + i\omega b}$$



SIZE OF BUMP DEPENDS ON b

By choosing the damping b correctly, we can make the peak almost go away

GENERAL CASE



$$X_{OUT} = \frac{G\, y_d}{1 + GH}$$

$$y_f = \frac{GH\, y_d}{1 + GH} \rightarrow y_d \quad \text{IF } |GH| >> 1$$

GH "OPEN LOOP GAIN"

(the gain you would have if you look at y_f with no feedback applied)

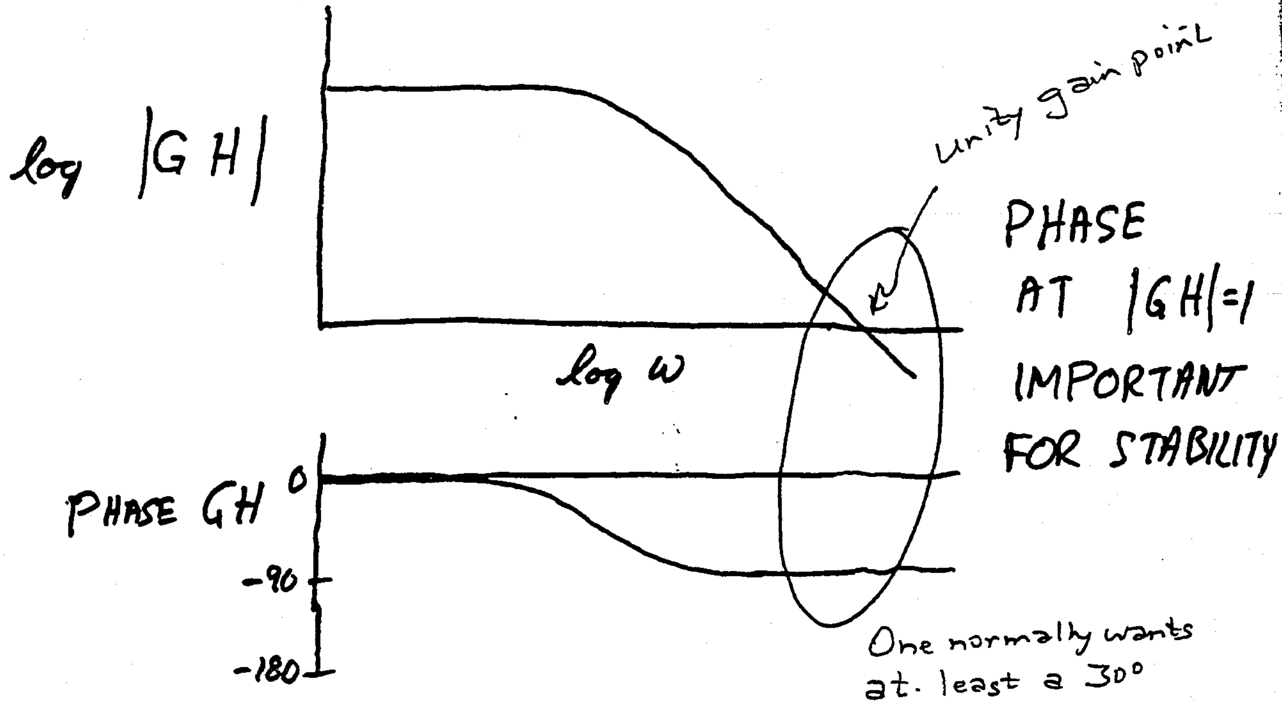
NOTE! PHASE OF GH AT FREQUENCY WHERE $|GH| = 1$ IS CRITICAL TO STABILITY

"UNITY GAIN POINT"

If phase ≈ 180° at unity gain point, then loop oscilates

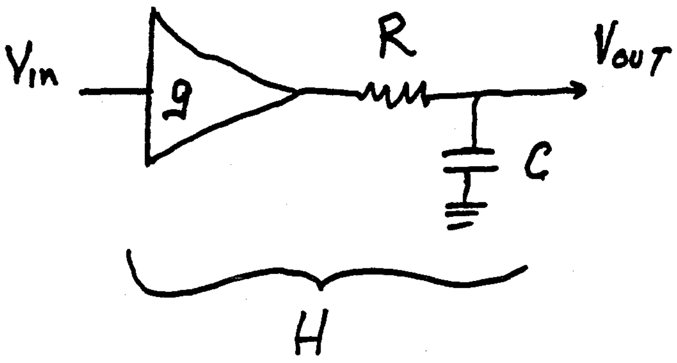
EARLY EXAMPLE in this lecture WAS A CASE OF "UNITY GAIN OSCILLATION"

BODE DIAGRAMS



One normally wants at least a 30° phase margin at unity gain point (phase > -180°)

TYPICAL FEED BACK ELEMENT



$$H(\omega) = \frac{g\, Z_c}{Z_R + Z_c}$$

$$= \frac{g \frac{1}{i\omega C}}{R + \frac{1}{i\omega C}} = \frac{g}{1 + i\frac{\omega}{\omega_0}}$$

This is well behaved

How do you do experiments with an interferometer? How do you test it and work with it?

DIAGNOSTIC TECHNIQUES

PARAMETER VARIATION

- USED TO VERIFY UNDERSTANDING OF NOISE SOURCES WITH EASY-TO-VARY PARAMETERS
- (USUALLY) REQUIRES NOISE SOURCE UNDER STUDY TO BE DOMINANT OVER SOME FREQ. RANGE
- MOST USEFUL FOR COMPARISON WITH DETAILED PHYSICAL MODELS

STIMULUS - RESPONSE

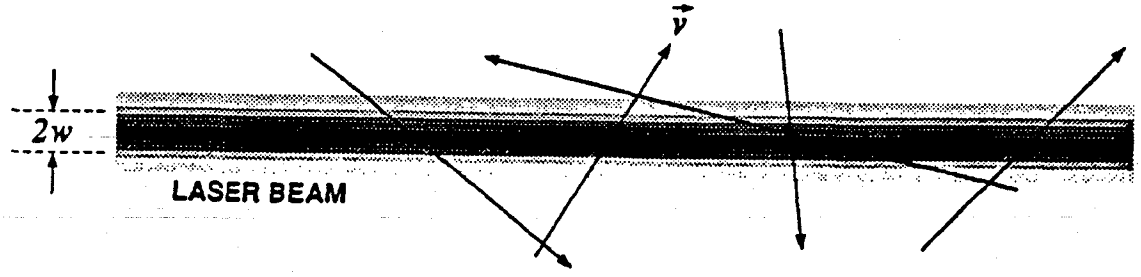
- MOST COMMONLY USED TECHNIQUE
- USED TO MAKE PREDICTIONS OF INDIVIDUAL -NOISE CONTRIBUTIONS FOR COMPARISON WITH OBSERVED NOISE LEVEL
- DOES <u>NOT</u> REQUIRE DETAILED NOISE MODEL FOR NOISE SOURCE OR FOR IT TO BE DOMINANT

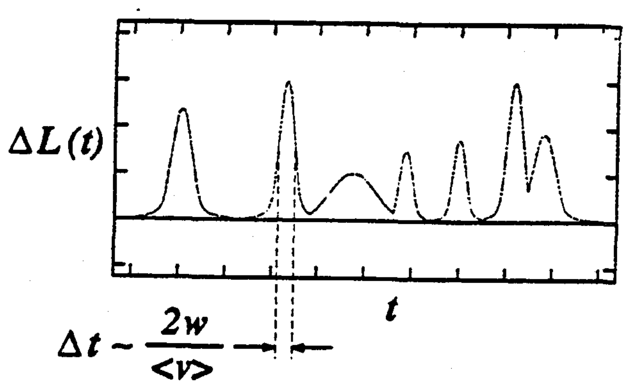
CONFIGURATION CHANGE

- EXTREME VARIANT OF "PARAMETER VARIATION"
- DIFFICULT AND LENGTHY TO IMPLEMENT
- POSSIBILITY OF CHANGING MANY PARAMETERS AT ONCE
- IN SPITE OF THESE LIMITATIONS, STILL THE BEST (ONLY) METHOD FOR SOME STUDIES

Residual Gas Index Fluctuation Noise

(Example of parameter variation).





$$\widetilde{\Delta L}(f) \propto P^{1/2}\, m^{1/4}\, \alpha \left(\frac{L}{w}\right)^{1/2}$$

where P is the partial pressure of the gas,
α is its molecular polarizability, and
m is its molecular weight.

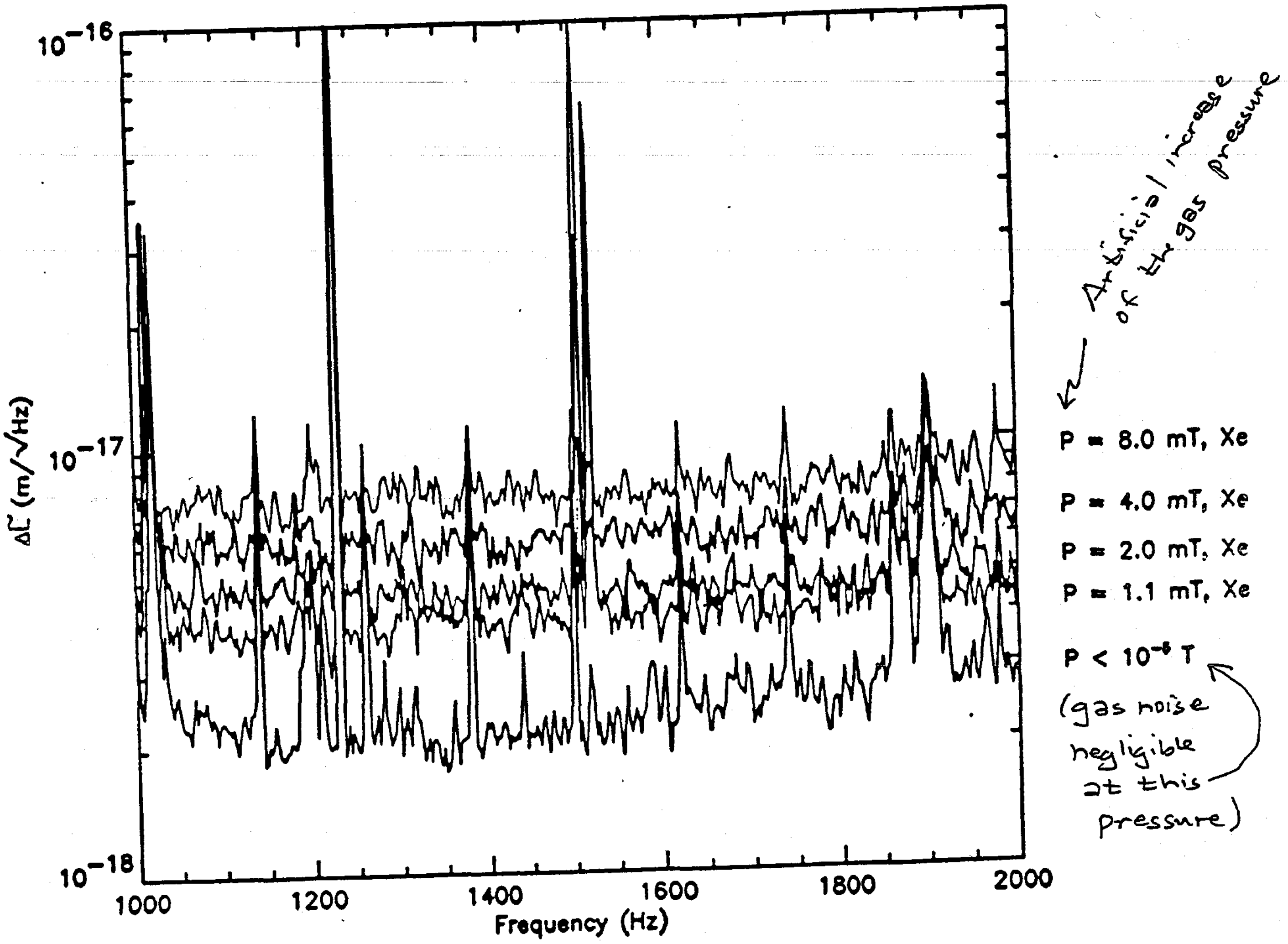
} One can change these parameters to see how they affect the noise

For diffraction-limited optical systems $w \propto L^{1/2}$ so strain noise

$$\tilde{h}(f) = \widetilde{\Delta L}(f)/L \propto L^{3/4}.$$

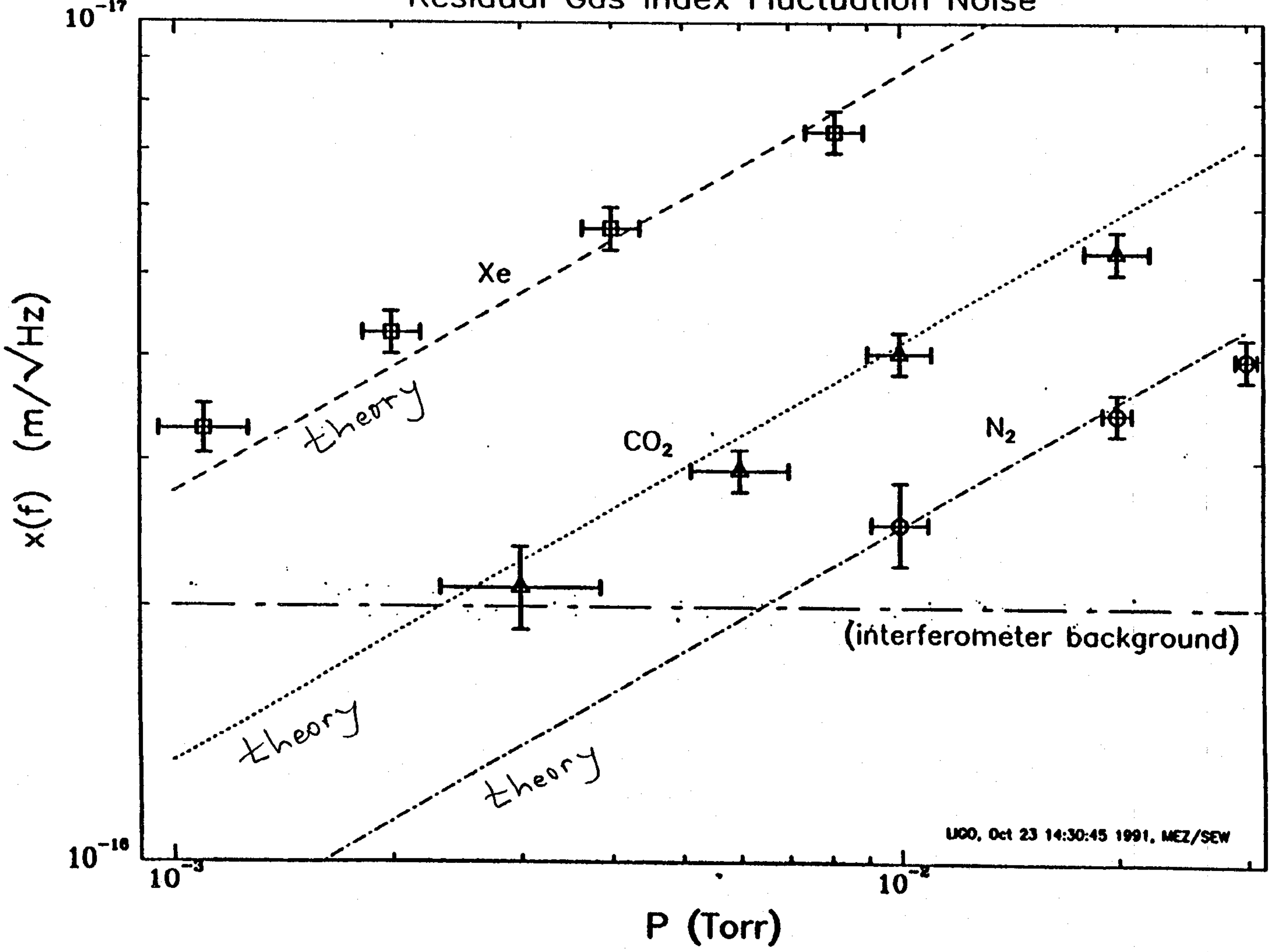
LIGO

Residual Gas Index Fluctuation Noise



note: $1\ T = 1\ \text{torr} = 1\ \text{mm Hg} = \frac{1}{760}$ atmosphere

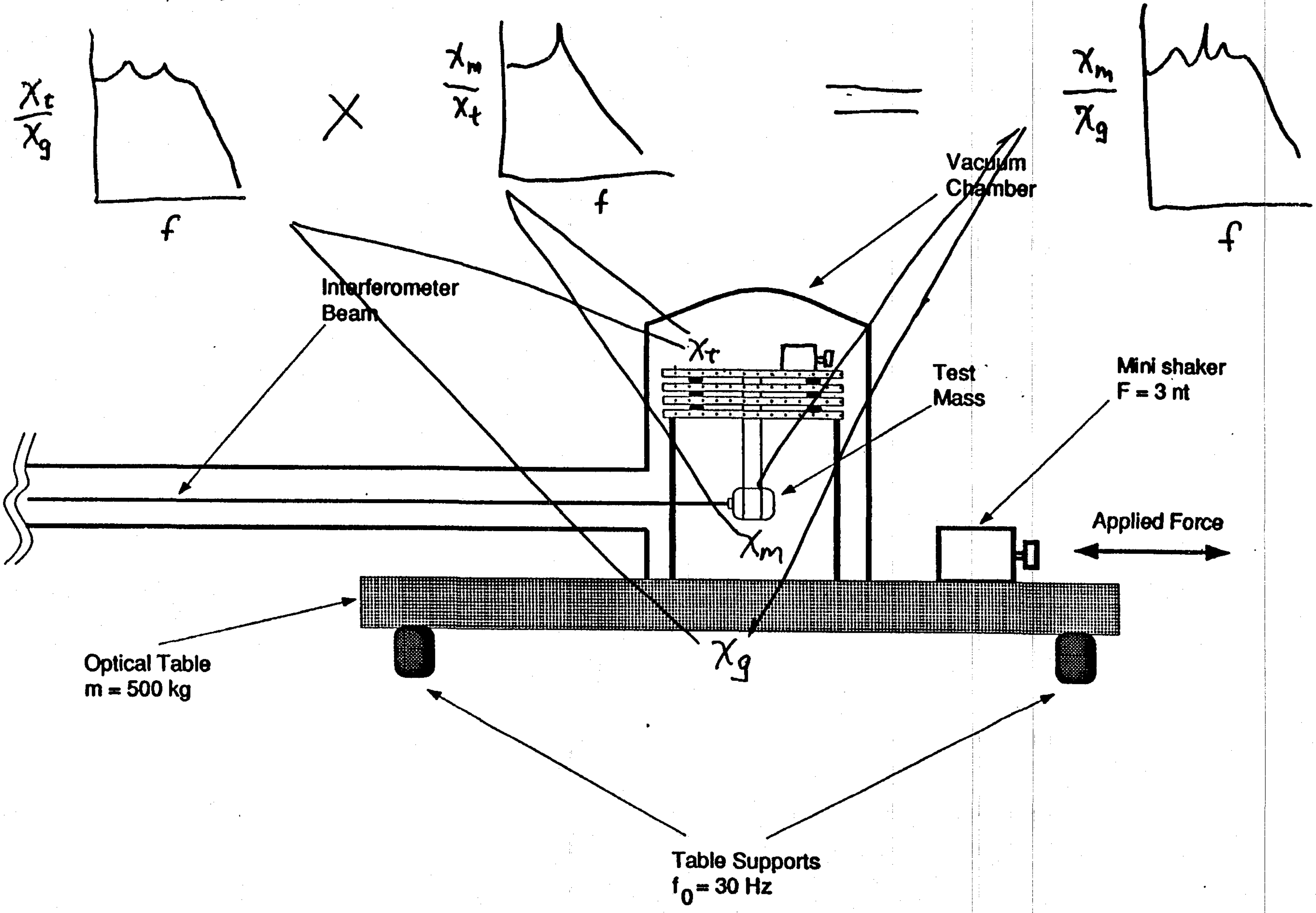
Residual Gas Index Fluctuation Noise



Experiment and theory agree well

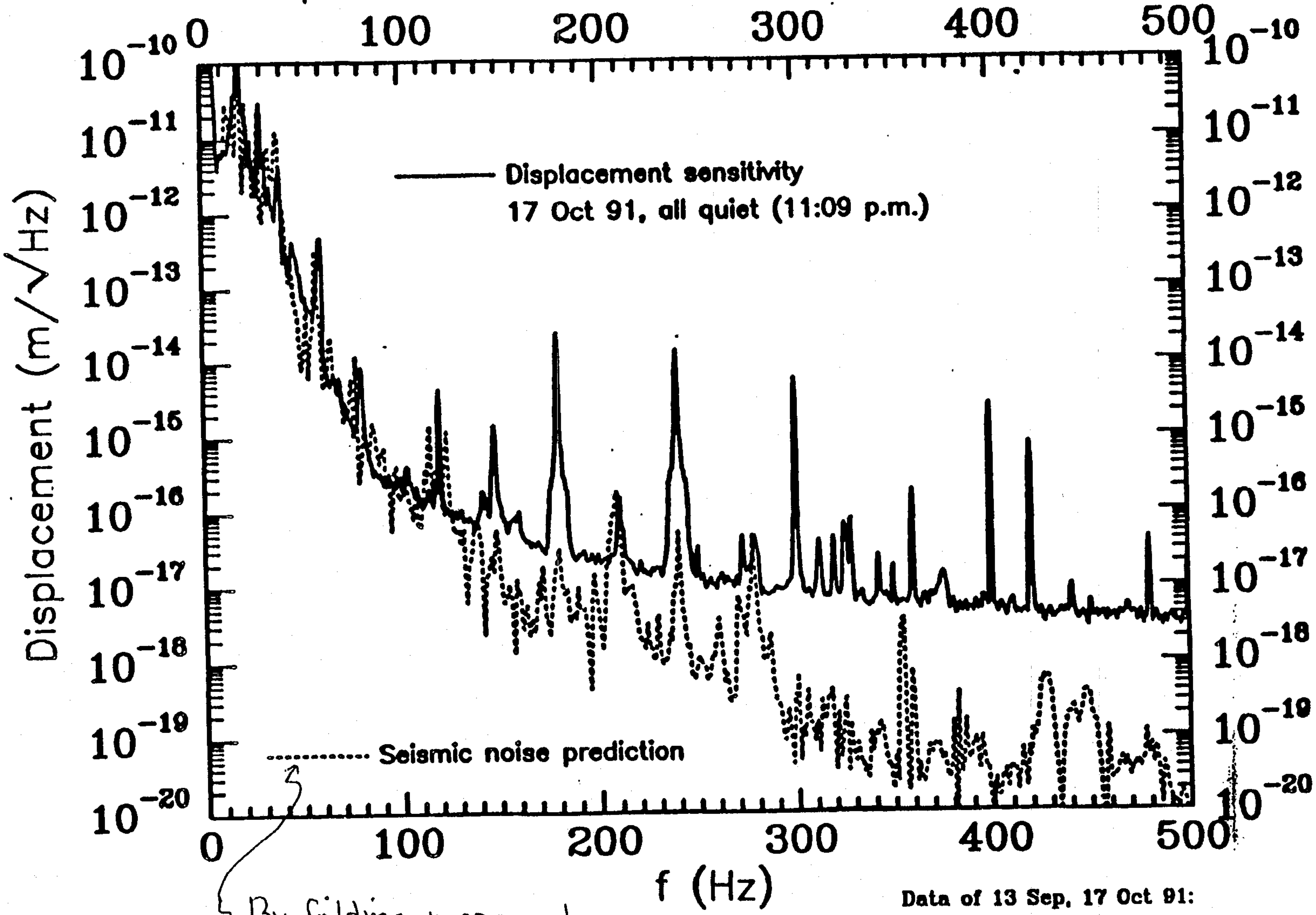
Example of Stimulus - Response:

Measurement of mechanical transfer functions for seismic & acoustic noise



40 meter interferometer

SEISMIC NOISE CONTRIBUTION---VERTICAL (H)
Input excitation measured by accelerometer



By folding measured
transfer function
into spectrum of ground motions

Data of 13 Sep, 17 Oct 91:
shv1, shv4, shv2, shv8, late, late1; sepspv.sm

BATCH
START

⑦ & ⑧ Lasers and Input Optics

STAPLE
OR
DIVIDER

Ph 103c: The Physics of LIGO 20 April 1994

LECTURE 7
LASERS AND INPUT OPTICS—I

Lecture by Robert Spero

Assigned Reading:

Q. A. Rüdiger, R. Schilling, L. Schnupp, W. Winkler, H. Billing and K. Maischberger, "A mode selector to suppress fluctuations in laser beam geometry," *Optica Acta,* **28**, 641–658 (1981).

R. "Noise in Optical Detection and Generation," Chapter 10 of A. Yariv, *Optical Electronics* (Saunders College Publishing, 1991).

Suggested Supplementary Reading:

The following two articles explain how frequency noise originating in vacuum fluctuations is fundamentally independent of the properties of atoms and depends only on the mirror properties and other cavity losses, and that the sensitivity of interferometric gravitational wave detectors is the same whether the arms are empty ("passive") cavities, as in LIGO, or idealized ("active" cavity) lasers.

S. "Comparison Between Active-cavity and Passive-cavity Interferometers," Abramovici A., Vager Z, *Phys. Rev.* **A33** (5), 3181-3184 (1986).

T. "Passive Versus Active Interferometers–Why Cavity Losses Make them Equivalent", J. Geabanacloche, *Phys. Rev. A* **35**(6), 2518–2522 (1987).

U. T.M. Niebauer, R. Schilling, K. Danzmann, A. Rudiger, and W. Winkler, "Nonstationary Shot Noise and its Effect on the Sensitivity of Interferometers" *Phys. Rev A* **43**(9), 5022–5029(1991). This paper resolves a long-standing 15% discrepancy between the calculated and observed shot noise in the German 30 m interferometer, having to do with the shape of the waveforms used for modulation and demodulation.

V. P.H. Roll, R. Krotkov, and R.H. Dicke, *Ann. Phys* **26**, 442 (1964). This paper, though long, is fun to read. It describes a classic experiment to measure the equivalence of inertial and gravitational mass, and is an excellent example of how experiments are designed, operated, and analyzed. The pages excerpted for the handout show how an optical lever reads the torsion balance deflection. Dicke's clever design, using a vibrating wire that casts a shadow on a photdetector, is a prototype for the use of modulation to reduce noise.

A Few Suggested Problems:

The 40 m interferometer operates in an "unrecombined" configuration: the reflected beams from the two arms' input mirrors do not interfere, and are separately detected. The shot noise levels from the two photodetectors add in quadrature; in the case of identical arms the total shot noise equivalent displacement is

$$\Delta\tilde{L}(f) \equiv \tilde{x}(f) = \frac{l}{4\pi\tau_E}\sqrt{3\mathcal{F}}\left[\frac{\lambda \mathrm{h}}{cP}\left(1+[f/f_k]^2\right)\right]^{1/2}$$

where $\mathcal{F}$ contains terms that depend on the depth of modulation Γ. ($\Gamma = 1$ corresponds to phase modulation of amplitude 1 radian.)

$$\mathcal{F} = \frac{1}{3}\left[\frac{M^{-1} + A^2J_0^2 - 2AJ_0^2 + 2AJ_0J_2}{MA^2J_0^2J_1^2}\right]$$

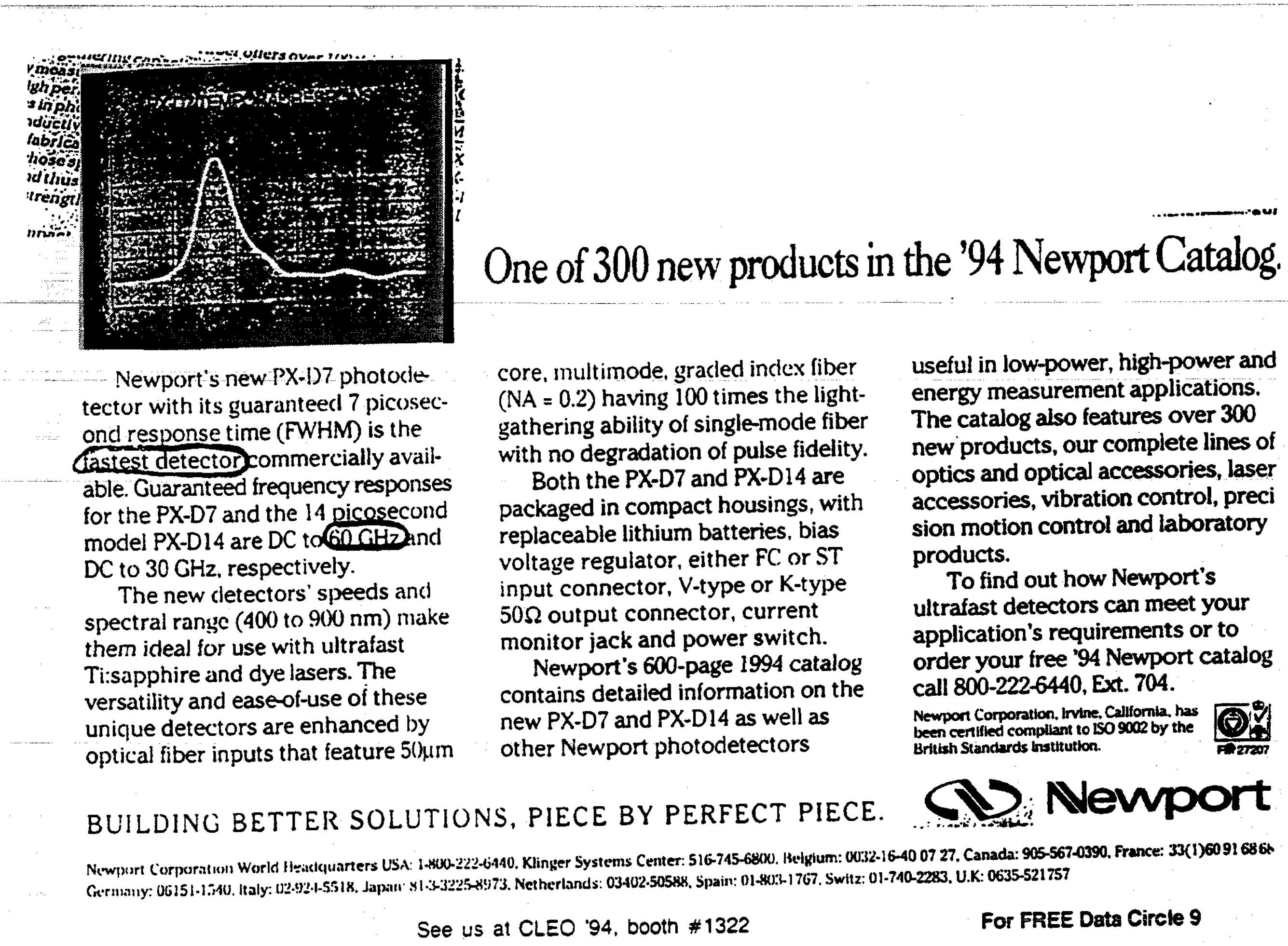
M is the (energy) mode matching fraction; $0 < M < 1$, $M = 1$ being the case of perfect alignment of the mirrors and proper matching of the laser gaussian beam parameters to the cavity mirror curvatures and separation. The mismatched fraction of the laser beam $(M-1)$ does not participate in the interfence, but does add to the shot noise. The 40 m interferometer operates with $M \approx 0.9$. J_0, J_1, J_2 are Bessel functions evaluated at Γ. Each cavity has input mirror transmission T and the sum of other losses L. τ_E, the cavity energy storage time, is the time it takes the intensity of the light "leaking" out of one of the arm cavities to drop from its starting level by a factor of e, after the input light is turned off; $\tau_E = \tau_t/(L+T)$, with $\tau_t = 2l/c$ the round-trip transit time and l the length of each arm. The cavity knee frequency is $f_k = 1/(4\pi\tau_E)$. $A = 2T/(L+T)$ is the amplitude of the cavity field leaking back out through the input mirror on resonance, in the absence of modulation. It is normalized to the input amplitude, and is constrained by $0 < A < 2$. λ is the optical wavelength, P is the total power (corrected for inefficiency in the photodiodes and other losses outside the arm cavities) incident on the beamsplitter, and f is the signal frequency.

1. Suppose the beamsplitter is not symmetric: that is, if P_1 and P_2 represent the power incident on the two arms, $P_1 = P\alpha$, $P_2 = P(1-\alpha)$, $\alpha \neq 0.5$. How does the sensitivity change from the symmetric case? How much asymmetry is required to degrade the sensitivity by 10%?
2. Verify that the modulation function $\mathcal{F}$ has a minimum value of 1. What parameters are required to approach this value? Optimization of interferometer sensitivity requires minimization of $\mathcal{F}$. Explain how the optimum value of Γ depends on the mode matching M and the mirror transmission and loss, T and L.
3. Even with l as short as 40 m, it is possible–using readily available very low-loss mirrors–to make f_k lower than the lowest expected detectable signal frequency f of approximately 100 Hz. Verify that for $f > f_k$, the shot-noise limited strain sensitivity $\tilde{h}(f) = \tilde{x}(f)l$ is independent of l, and make a plot sketching $\tilde{h}(f)$ for various values of l, all other parameters held fixed. The currently achieved shot-noise limited displacement sensitivity of the 40 m interferometer is approximately the same as the requirement for initial LIGO ($l = 4$ km) detectors. What are the implications for the design of LIGO detectors? For R&D on the 40 m interferometer?
4. Compare the shot noise sensitivity above to the "recombined" but not recycled calculation of Lecture 4. Explain why the sensitivity is worse for the unrecombined configuration.

Lecture 7
Lasers and Input Optics — I. Mechanics of Signal and Noise

by Robert Spero, 20 April 1995

Spero lectured from the following transparencies.



One of 300 new products in the '94 Newport Catalog.

Newport's new PX-D7 photodetector with its guaranteed 7 picosecond response time (FWHM) is the fastest detector commercially available. Guaranteed frequency responses for the PX-D7 and the 14 picosecond model PX-D14 are DC to 60 GHz and DC to 30 GHz, respectively.

The new detectors' speeds and spectral range (400 to 900 nm) make them ideal for use with ultrafast Ti:sapphire and dye lasers. The versatility and ease-of-use of these unique detectors are enhanced by optical fiber inputs that feature 50µm core, multimode, graded index fiber (NA = 0.2) having 100 times the light-gathering ability of single-mode fiber with no degradation of pulse fidelity.

Both the PX-D7 and PX-D14 are packaged in compact housings, with replaceable lithium batteries, bias voltage regulator, either FC or ST input connector, V-type or K-type 50Ω output connector, current monitor jack and power switch.

Newport's 600-page 1994 catalog contains detailed information on the new PX-D7 and PX-D14 as well as other Newport photodetectors useful in low-power, high-power and energy measurement applications. The catalog also features over 300 new products, our complete lines of optics and optical accessories, laser accessories, vibration control, precision motion control and laboratory products.

To find out how Newport's ultrafast detectors can meet your application's requirements or to order your free '94 Newport catalog call 800-222-6440, Ext. 704.

Newport Corporation, Irvine, California, has been certified compliant to ISO 9002 by the British Standards Institution. FM 27207

BUILDING BETTER SOLUTIONS, PIECE BY PERFECT PIECE. Newport

Newport Corporation World Headquarters USA: 1-800-222-6440, Klinger Systems Center: 516-745-6800, Belgium: 0032-16-40 07 27, Canada: 905-567-0390, France: 33(1)60 91 68 68
Germany: 06151-1540, Italy: 02-924-5518, Japan: 81-3-3225-8973, Netherlands: 03402-50588, Spain: 01-803-1767, Switz: 01-740-2283, U.K: 0635-521757

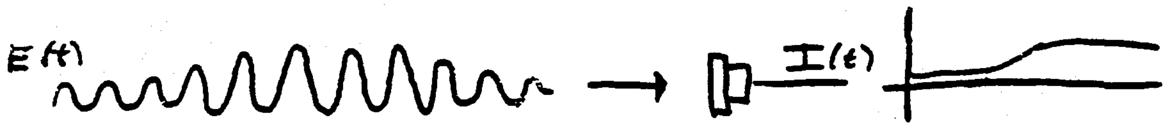
See us at CLEO '94, booth #1322

For FREE Data Circle 9

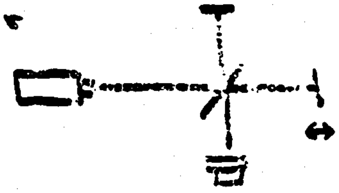
"FAST" PHOTODETECTOR RESPONDS TO $f \lesssim 10^{11}$ Hz

cf $\nu_{\text{OPTICAL}} = 6 \cdot 10^{14}$ Hz

Photodetection senses average of $(\text{field})^2$ incident on sensor:

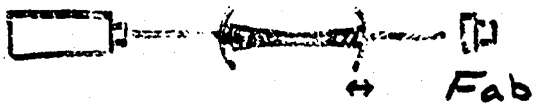


PHASE SENSITIVITY requires phase-to-intensity converter



Michelson Interferometer

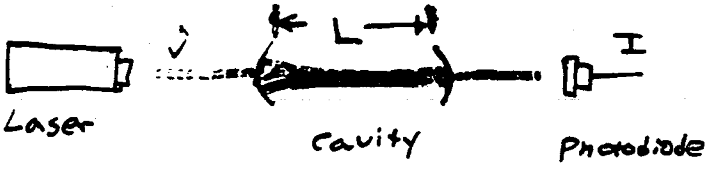
or

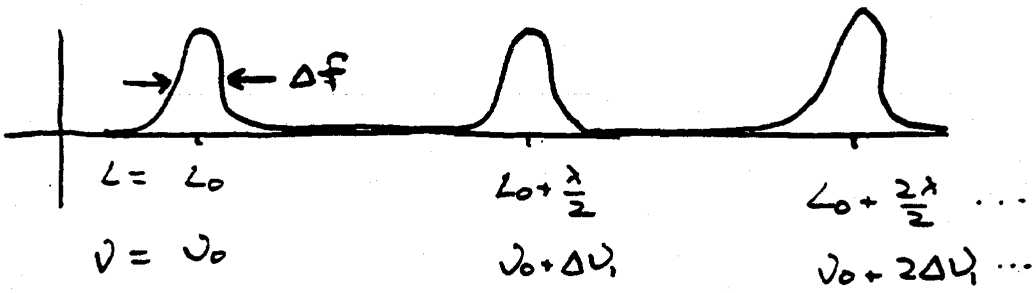


Fabry-Perot cavity

CAVITY AS DISPLACEMENT SENSOR

1) TRANSMITTED LIGHT





RESONANCE CONDITION: $L_0 = \frac{n\lambda}{2}$

To maintain resonance when ν or L changes:

$$\Delta L \propto \Delta\lambda \qquad \lambda = \frac{c}{\nu}$$

or $$\frac{\Delta L}{L} = -\frac{\Delta\nu}{\nu}$$

TERMINOLOGY

"FREE SPECTRAL RANGE" $= \Delta\nu_1$

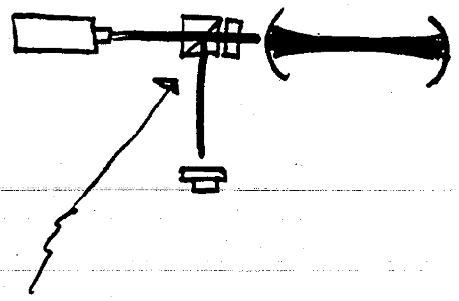
$$\Delta\nu_1 = \nu_0 \frac{\lambda}{2} \frac{1}{L_0} = \frac{c}{2L_0}$$

"Q" $= \frac{\nu_0}{\Delta f}$ $(3 \cdot 10^{12}$ in 40 m interferometer$)$

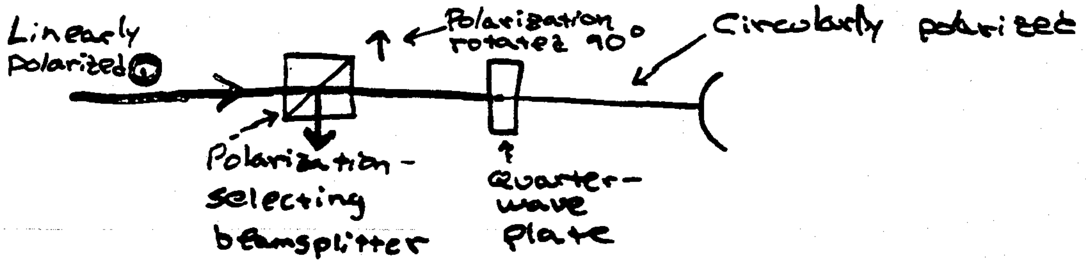
"Finesse" $= \frac{\Delta\nu_1}{\Delta f}$ $(6 \cdot 10^{3})$

"Bandwidth" $= \Delta f$ (300 Hz)

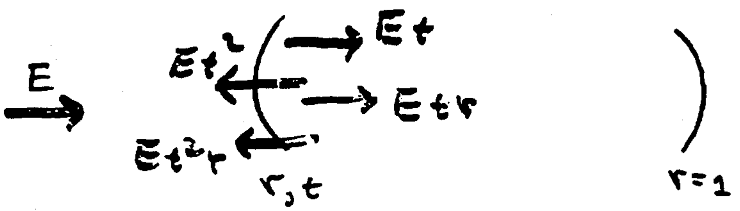
2) REFLECTED LIGHT



"Circulator" selects light from cavity by polarization trick



FIELDS ON RESONANCE, STEADY STATE



$$A = Et^2 + Et^2r + Et^2r^2 + \cdots$$

$$= ET(1 + r + r^2 + \cdots)$$

$$= \frac{ET}{1-r} = \frac{ET}{1-\sqrt{R}}$$

$t^2 = T$

$r^2 = R$

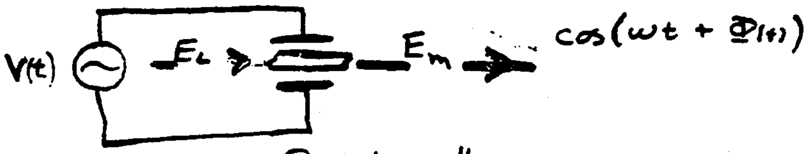
$R + T + L = 1$

$(L = \text{Loss})$

$$1-\sqrt{R} = 1-\sqrt{1-\epsilon} = \tfrac{1}{2}\epsilon = \tfrac{1}{2}(1-R) = \tfrac{1}{2}(T+L)$$

$$A = \frac{2ET}{T+L}, \quad A_0 \equiv \frac{A}{E} = \frac{2T}{T+L} \qquad 0 < A_0 < 2$$

PHASE MODULATION



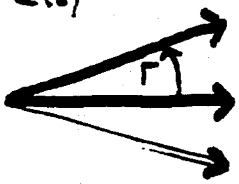
Pockels cell
Electrooptic crystal

$\Phi(t) \propto V(t)$; the index of refraction change, and the optical pathlength change and $\Phi(t)$, are proportional to the electric field across the crystal.

$E_L(t) = \cos\omega t$

$E_m(t) = \cos(\omega t + \Phi(t)) = \cos\omega t \cos\Phi(t) - \sin\omega t \sin\Phi(t)$

$\Phi(t) = \Gamma \sin\omega_m t$

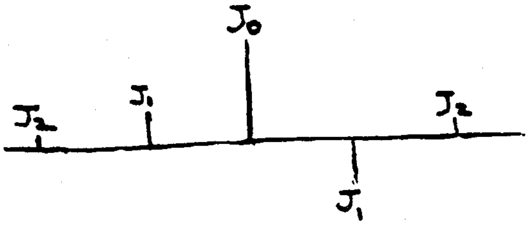


2Γ = peak-to-p variation in E_m phase

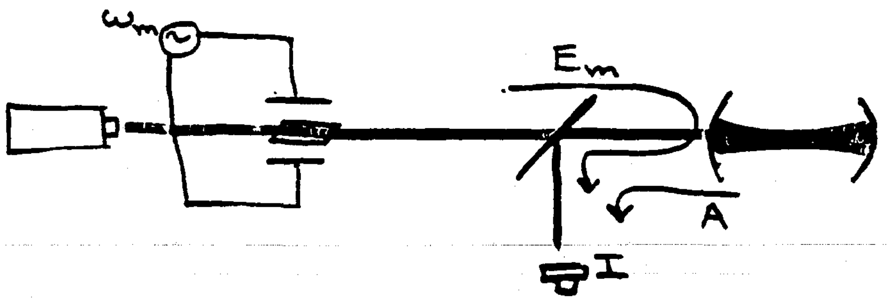
$$\left[\begin{array}{l} \cos(\Gamma\sin\theta) = J_0(\Gamma) + 2J_2(\Gamma)\cos 2\theta + \dots \\ \sin(\Gamma\sin\theta) = 2J_1(\Gamma)\sin\theta + 2J_3(\Gamma)\sin 3\theta + \dots \\ 2\cos a\cos b = \cos(a+b) + \cos(a-b) \\ 2\sin a\sin b = \cos(a+b) - \cos(a-b) \end{array}\right]$$

$$\begin{aligned} E(t) = {} & J_0\cos\omega t - J_1[\cos(\omega+\omega_m)t - \cos(\omega-\omega_m)t] \\ & + J_2[\cos(\omega+2\omega_m)t + \cos(\omega-2\omega_m)t] \\ & - J_3[\cdots] \end{aligned}$$

Sideband
Spectrum



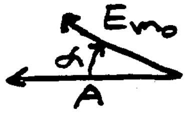
INTERFERENCE AND MODULATION TOGETHER



$$E_m = E_{m0}\left(J_0(\Gamma) + 2J_2(\Gamma)\cos 2\omega_m t + \cdots\right)$$

$\omega_m >> \Delta f$; no sidebands on A

I is intensity resulting from sum of fields E_m and A:



$$I = |A - E_{m0}|^2 = A^2 + E_{m0}^2 - 2AE_{m0}\cos\alpha$$

$$A = E_{m0}J_0(\Gamma)\frac{2T}{T+L} = E_{m0}J_0(\Gamma)A_0$$

↑ coefficient for resonant sideband

Sideband spectrum of I : $\alpha = \underset{\text{Signal, } \ll 1}{\alpha_0} + \Gamma\cos\omega_m t$ ← Modulation

$$\cos\alpha = \cos\alpha_0 \cos\left[\Gamma \sin\omega_m t\right] - \sin\alpha_0 \sin\left[\Gamma\sin\omega_m t\right]$$

$$\left[\ \sin(\Gamma\cos\theta) = 2J_1(\Gamma)\sin\theta + 2J_3(\Gamma)\sin 3\theta + \cdots\ \right]$$

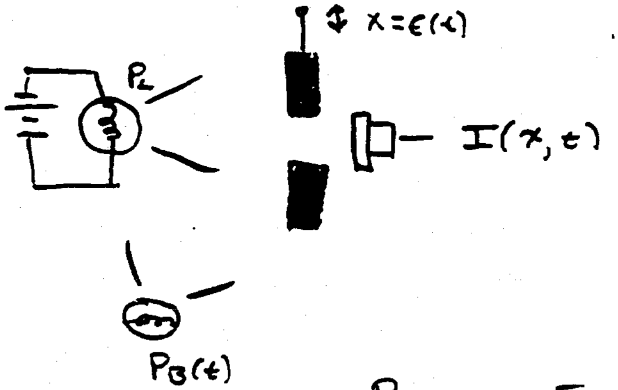
$$\frac{I_{\text{o.c.}}}{E_{m0}^2} = 1 + J_0^2 A_0^2 - 2A_0 J_0^2 \quad \text{Background light}$$

$$\frac{I_\omega}{E_{m0}^2} = -4J_1(\Gamma)\sin\omega_m t\cdot\sin\alpha_0 \sin\omega_m t$$

(Component at ω_m) ≈ linear in α_0

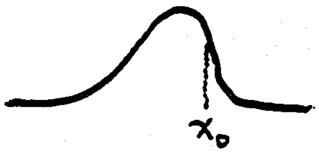
DIGRESSION ON MODULATION AND SIGNAL EXTRACTION

Experiment: Measure the position of an aperture in the presence of fluctuating background illumination.



Response Function $T(x)$:

$$I = (P_L + P_B)\,T(x)$$

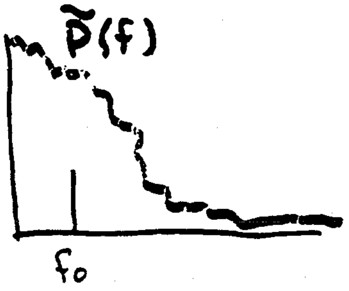


offset photodiode position to x_0 for linear response

$X = X_0 - \epsilon$ $\qquad x_0 \gtrsim 1$, $\epsilon \ll 1$, $T(\epsilon) \simeq 1 - X_0^2 + 2\epsilon X_0$

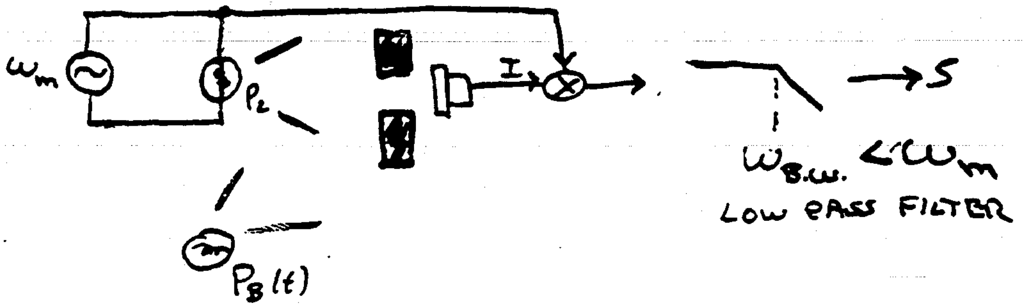
THE CATCH: (#1) Fluctuation of P_B changes I; indistinguishable from change due to ϵ.

Problem if $\tilde{P}(f)$ is strong at frequency of interest, f_0



Poor Signal-to-noise ratio

THE SOLUTION: (#1) Modulate P_L at frequency where P_B is quiet. Demodulate I synchronously



$$P_L = P_{L0}\,(1 + \underbrace{\alpha}_{\text{Depth of modulation}} \cos\omega_m t)$$

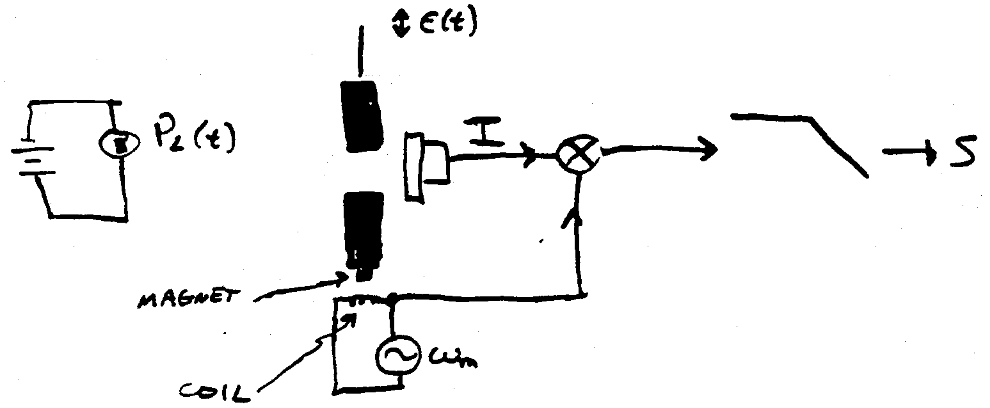
$$I = \left[P_L\,(1+\alpha\cos\omega_m t) + P_B(t)\right]T(x)$$

$$\begin{aligned} I\cos\omega_m t = {} & \tfrac{1}{2}\alpha P_L T(x) \\ & + \{\cos\omega_m t \ \ \text{terms}\} \\ & + \{\cos 2\omega_m t \ \ \text{terms}\} \end{aligned}$$

$S = \frac{1}{2}\alpha P_L T(x)$, independent of P_B

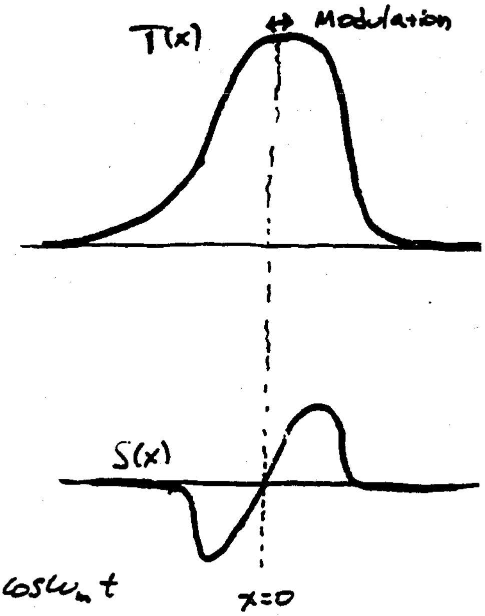
CATCH #2: P_L fluctuates

SOLUTION #2: Modulate aperture position (at frequency where P_L is quiet)



Photodiode is nominally centered on aperture

Component of I at ω_m is zero until aperture moves off center:



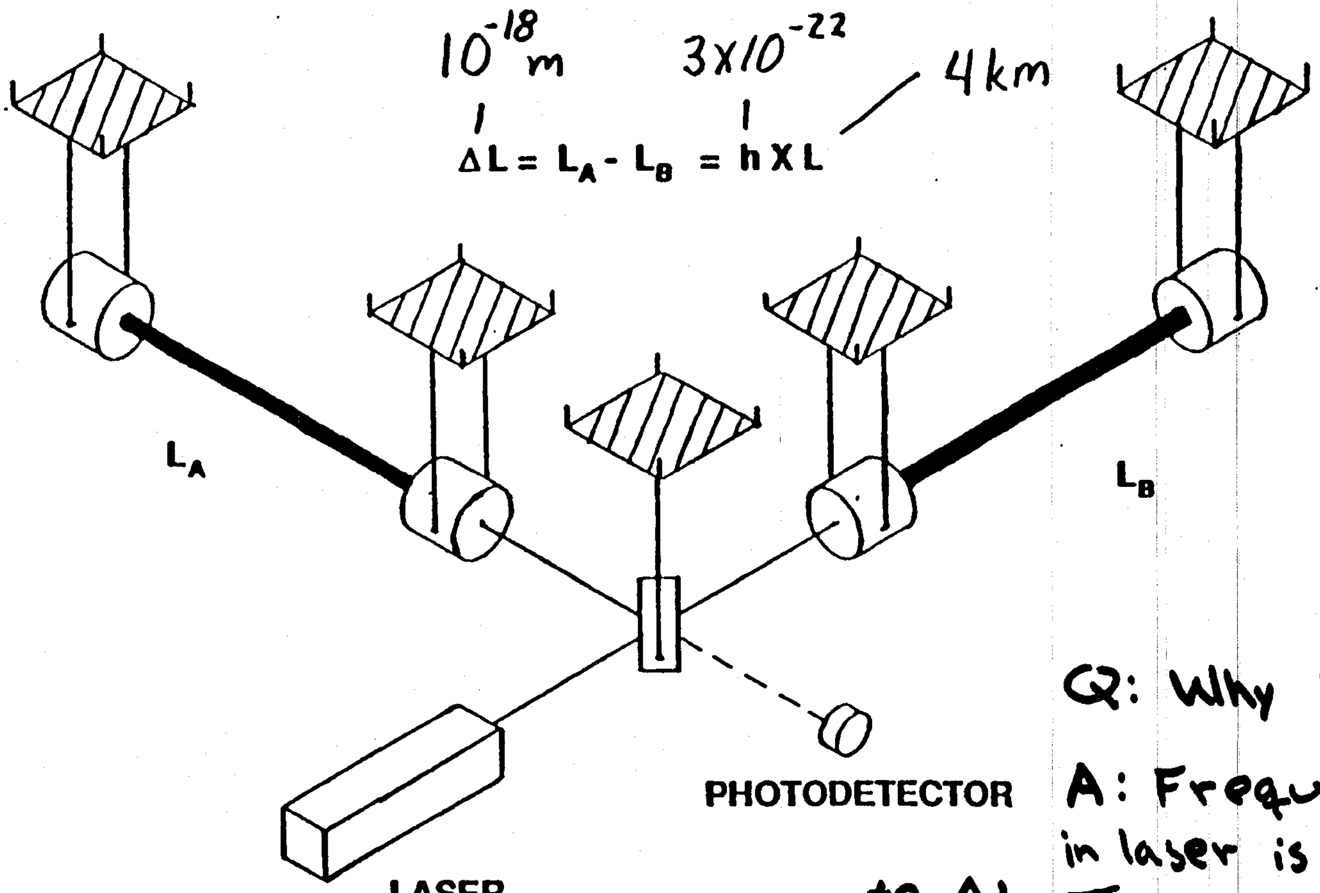
$$X = \epsilon(t) + \beta\cos\omega_m t$$

$$T(x) = 1 - X^2 = 1 - \epsilon^2 - \beta^2\cos^2\omega_m t - 2\epsilon\beta\cos\omega_m t$$

↑

Term selected by demodulation

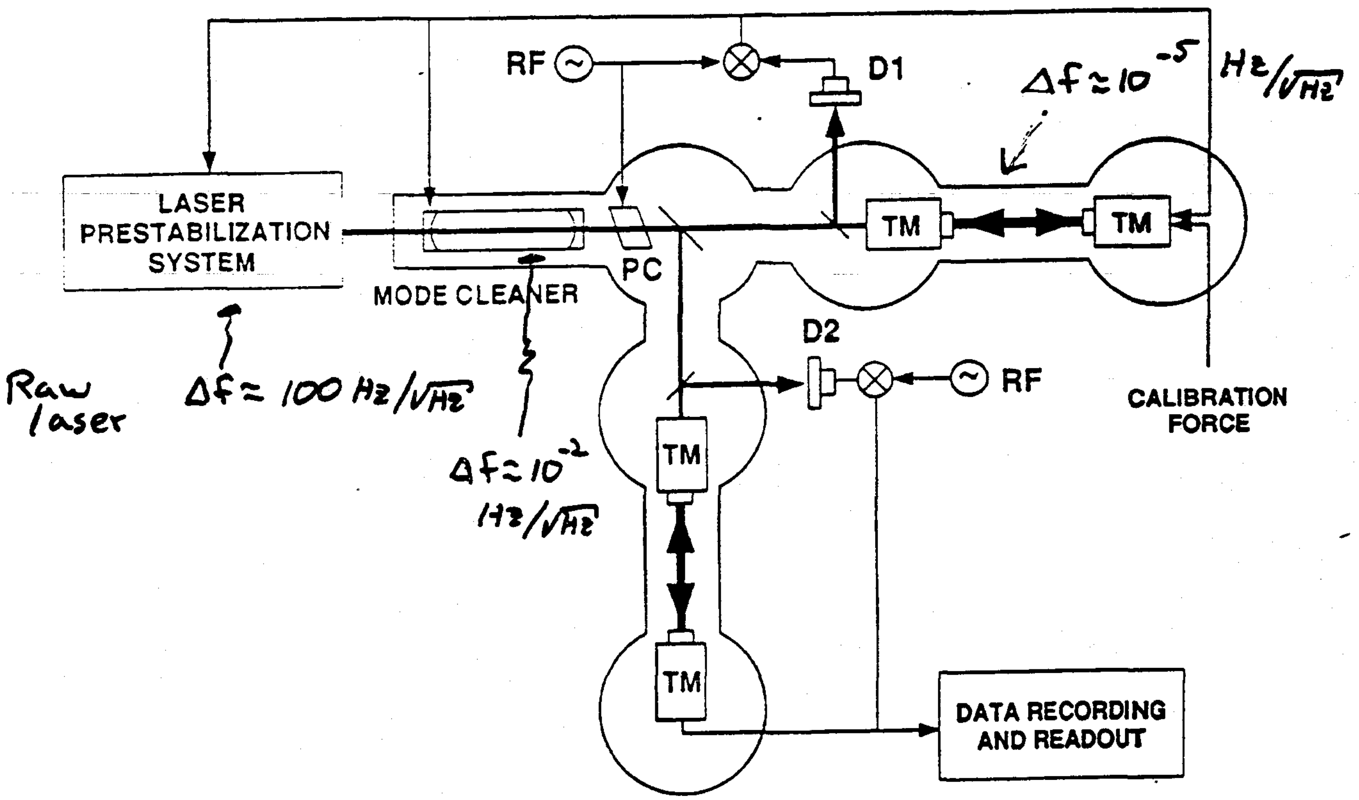
SCHEMATIC INTERFEROMETRIC DETECTOR



Q: Why 2 arms?

A: Frequency noise in laser is equivalent to ΔL. Two arms can be configured to eliminate Δf (by frequency-stabilizing servos) or to reduce sensitivity to Δf (by making arms identical and subtracting).

$$\Delta\tilde{f} = \nu_0 \frac{\tilde{x}(f)}{L_0}$$



$$\frac{\tilde{x}(f)}{L_0} \approx \frac{1}{4\pi\tau_E}\sqrt{\frac{3\lambda h}{cP}}$$ (low frequency limit)

$$\tau_E = \frac{\tau_r}{L+T}$$

τ_r ← Round trip light travel time $= \frac{2L_0}{c}$

L ← Mirror loss

T ← Mirror transmission

Response of Arm n to $\Delta f \propto \tau_{E,n} \cdot P_n$ ← Input power

Arm mismatch: 40 m ~ 30%

LIGO ~ 3%

VERIFICATION OF SHOT NOISE SENSITIVITY

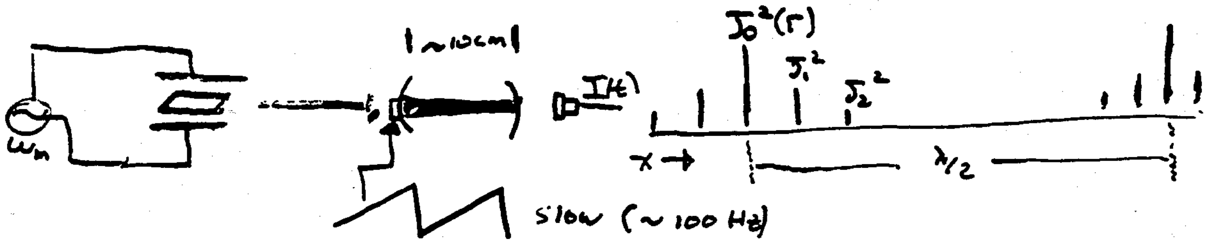
1) Add pure shot noise to photodiodes with flashlight, and observe increase in $\tilde{x}(f)$

2) [Also verification of calculation] Measure relevant parameters needed to calculate $\tilde{x}(f)$, and compare with calibrated measurement

2a "Visibility" $V = 1 - \frac{I_{D.C.}}{E_{mo}^2} = A_0 J_0^2(\Gamma)(2-A_0)$

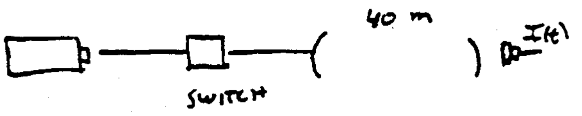
$A_0 = \frac{2T}{T+L}$ V=1 for DARK FRING

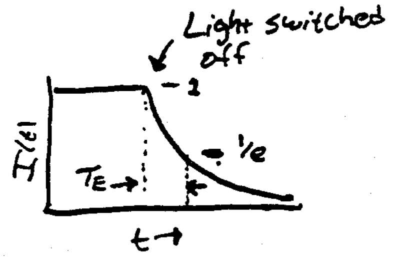
2b Γ with optical spectrum analyzer:



slow (~100 Hz) ramp sweeps one mirror position with amplitude ~ λ

2c T+L by ringdown

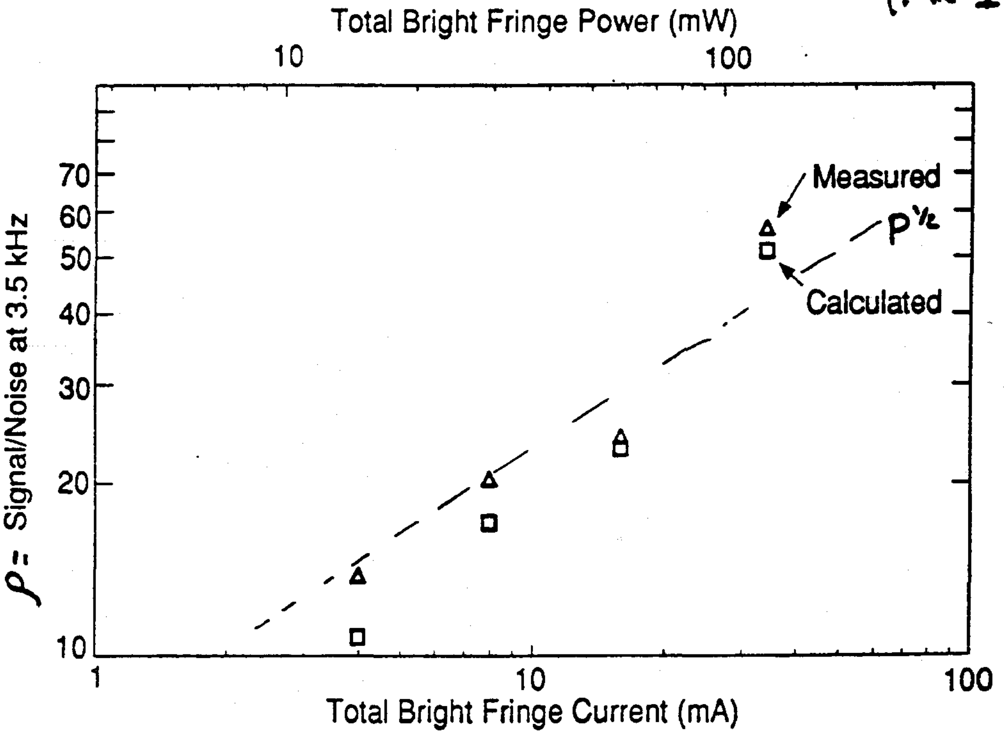




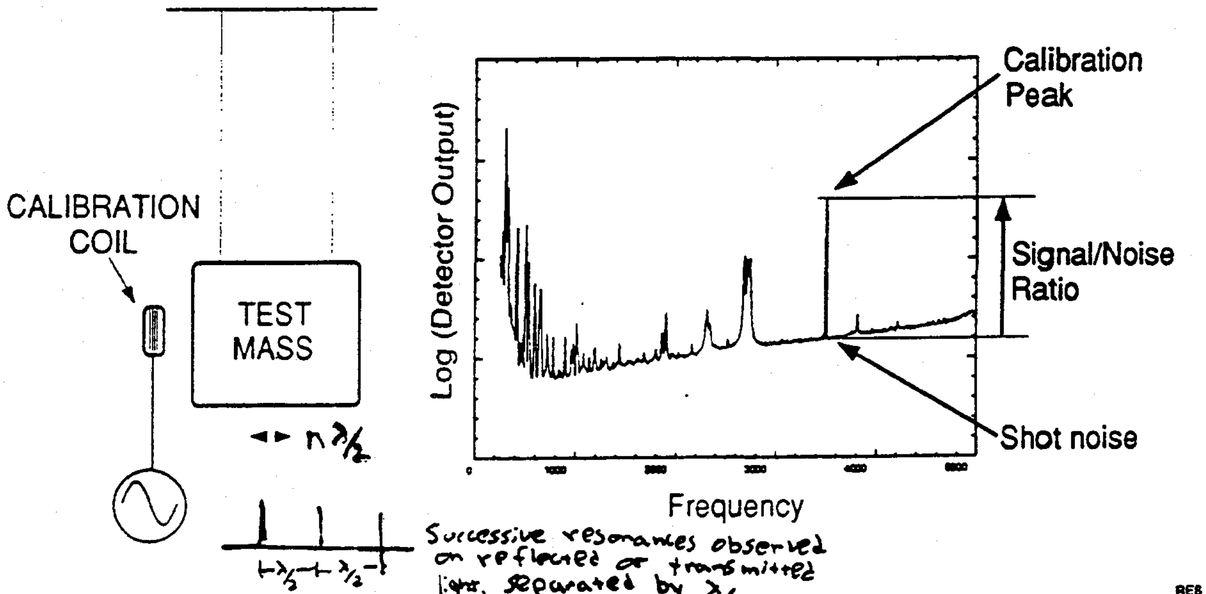
$\tau_E = \frac{\tau_t}{L+T}$

POWER DEPENDENCE OF SHOT NOISE

("Mk I")



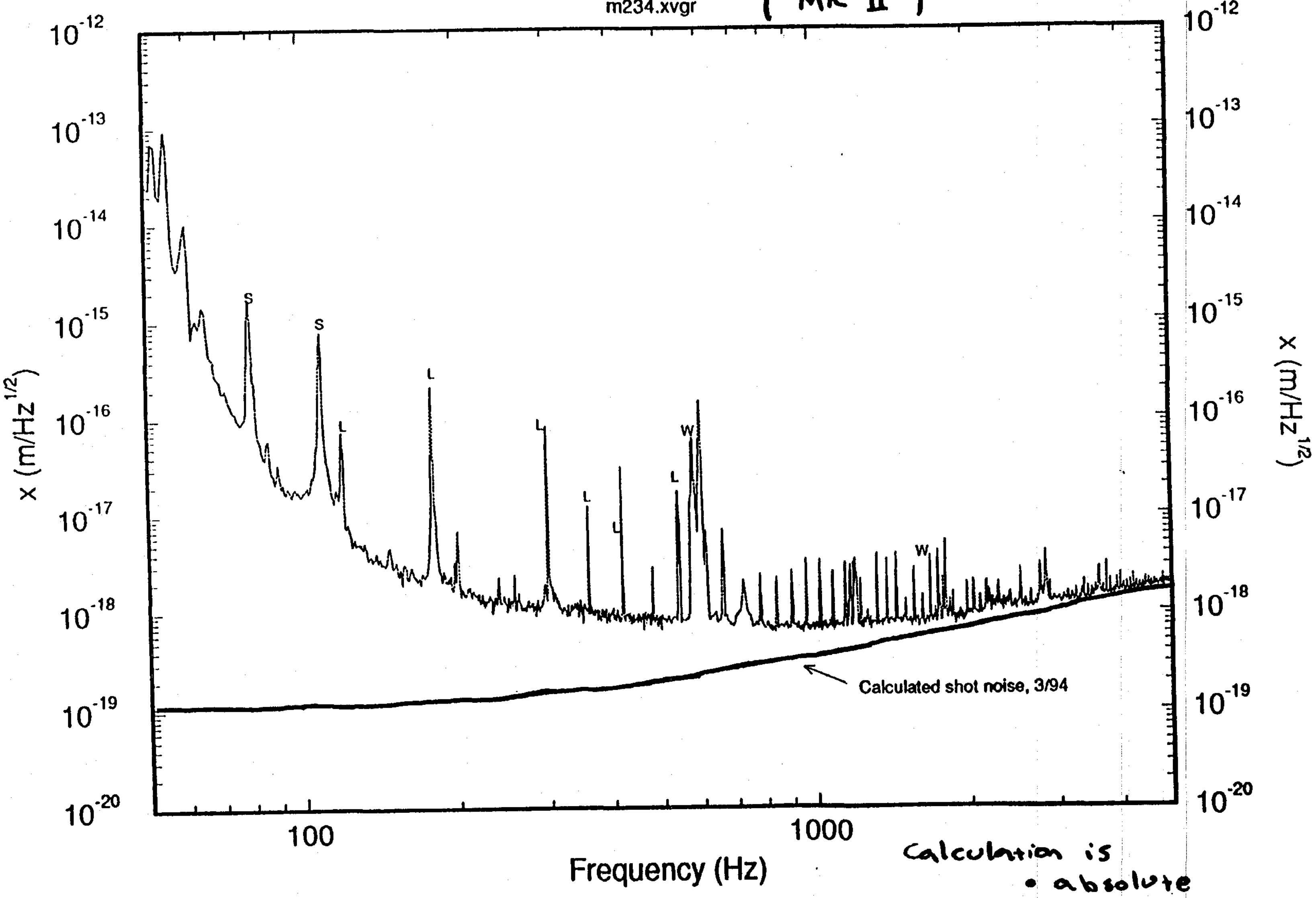
$$\rho_{CALCULATED} = \sqrt{\rho_{SHOT}^2 + \rho_{DARK}^2}$$



Successive resonances observed on reflected or transmitted light, separated by λ/2

Displacement Sensitivity of Caltech 40 m Interferometer

m234.xvgr ("Mk II")



Calculation is
- absolute
- not compensated for noise other than shot noise.

LECTURE 8.
LASERS AND INPUT OPTICS — II

Lecture by Alex Abramovici

Assigned Reading:

W. A. Rüdiger, R. Schilling, L. Schnupp, W. Winkler, H. Billing and K. Maischberger, "A mode selector to suppress fluctuations in laser beam geometry," *Optica Acta*, **28**, 641–658 (1981). [This paper describes, first in simple terms and then in terms of a mode decomposition, the use of a *mode cleaning Fabry-Perot cavity* to precondition the light that is injected into an interferometer. The preconditioning includes suppression of beam wiggle, suppression of beam-diameter pulsations, and suppression of other unwanted spatial modes of the laser light. Also described is the use of lenses to adjust the radius of curvature of the beam's phase fronts so as to match the desired eigenmodes of the mode cleaner and of each arm of the interferometer. Note that, at the time this paper was written, "supermirrors" with losses far far less than 0.01 were not yet available, and the degree to which one can control mirror heating by keeping the mirrors extremely clean was not yet understood.]

X. Those students who are not familiar with the physics of lasers should also read the introductory chapter of a good text on laser physics; for example, Chapter 1, "Introduction" of W. Koechner, *Solid-State Laser Engineering* (Springer Verlag, Berlin, 1988), which is being passed out.

Suggested Supplementary Reading:

H. Read more deeply into your favorite laser physics text. Most especially, read the material dealing with Gaussian beams and their manipulation, e.g. the material already suggested in Lecture 4: Reference H — chapter 17, "Physical Properties of Gaussian Beams," of A. E. Siegman, *Lasers* (University Science Books, Mill Valley CA, 1986).

A Few Suggested Problems

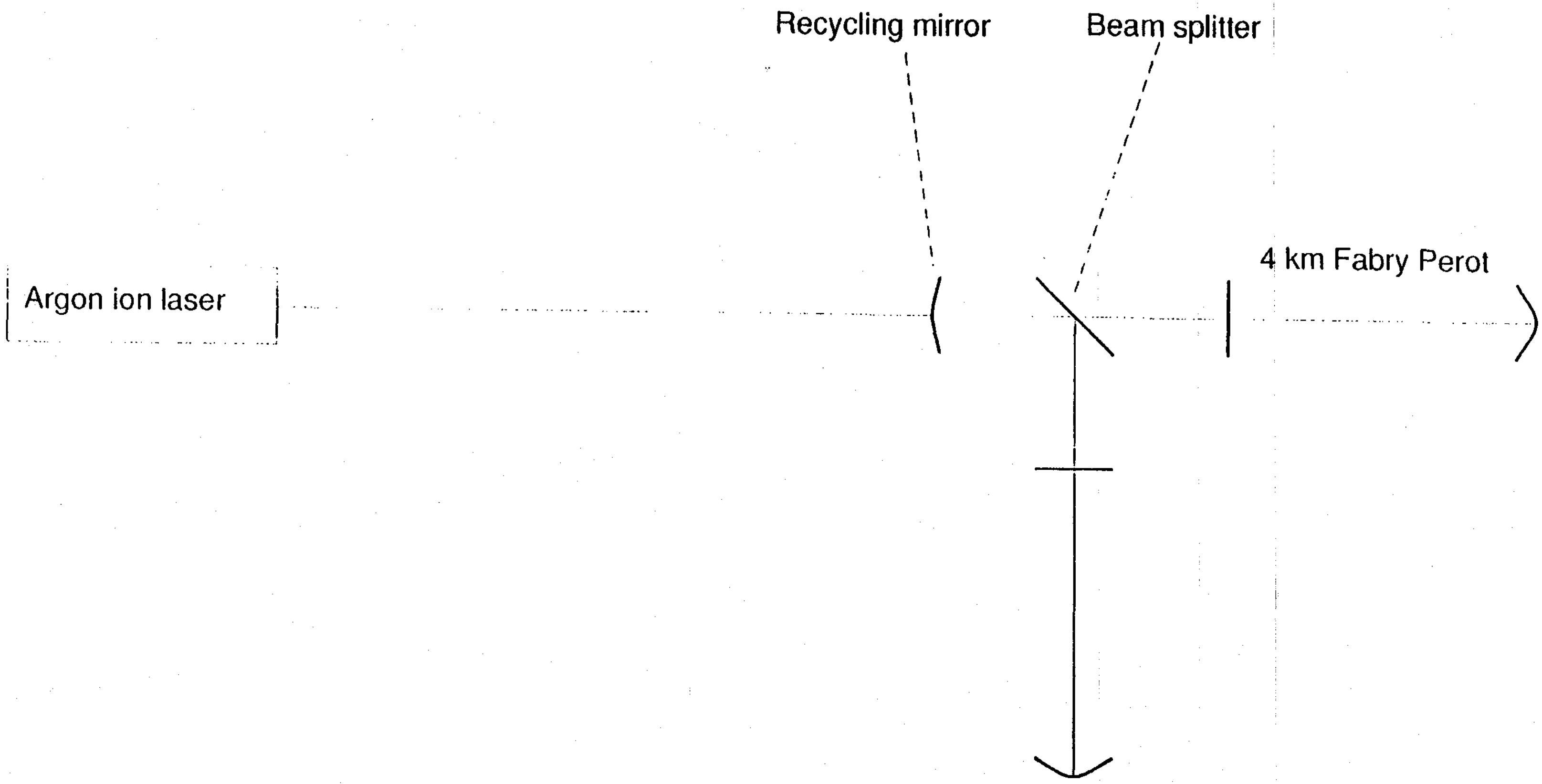
1. *Laser Stabilization by Locking to a Cavity.* The frequency of a laser is stabilized by locking it to an eigenmode of an optical cavity using a feedback loop. Suppose that, in the absence of the feedback loop, the laser's frequency differs from the cavity's eigenfrequency by an amount $\Delta\nu \equiv \nu_o$ (the "initial detuning"). When the feedback system is turned on, the residual detuning is $\Delta\nu \equiv \nu_1 = \nu_o/(1+G)$, where $G \gg 1$ is the gain of the feedback system, which is proportional to the power I of the light beam. What is the rms frequency fluctuation σ_ν induced by an rms fluctuation σ_I of the optical power?
2. *Mode Cleaning Cavity—I.* The Fabry-Perot cavity that will make up each arm of LIGO's standard, broad-band interferometric gravitational-wave detector will have a corner mirror with modest power transmisivity $\mathcal{T}_c \equiv 1-\mathcal{R}_c \sim 10^{-2}$ and an end mirror with tiny transmissivity $\mathcal{T}_e \sim 10^{-5}$. With this huge difference of transmissivities, almost all the light injected into the cavity through the corner mirror ultimately leaves back through the corner mirror; hardly any leaks out the end mirror. In a mode cleaning Fabry-Perot cavity, by contrast, the two mirrors are chosen to have identical transmisivities $\mathcal{T}$. In this case show that, if the cavity (which is idealized as having no absorption or scattering) is driven on resonance through the left mirror, all the light leaves the cavity through the right mirror. If the cavity is driven off resonance, what fraction of the light goes out each end?
3. *Mode Cleaning Cavity—II.* Consider a mode-cleaning cavity consisting of two identical concave mirrors with radii of curvature $R = 1$m and power transmissivities $\mathcal{T} = 0.005$, separated by 1.5m along the optic axis. The cavity is driven by laser light that is primarily in the TEM$_{00}$ mode, with a small admixture of TEM$_{01}$; and it is driven on a TEM$_{00}$ resonance so all the light in that mode passes through the cavity from one side to the other. What fraction of the light in the TEM$_{01}$ mode passes through?
4. *Changes in wavefront curvature on reflection from a curved mirror.* The phase variation in the transverse plane (i.e. at constant z) for a diverging Gaussian beam propagating in the z direction (so $\psi \propto e^{+i(kz-\omega t)}$), with a phase-front radius of curvature R, is $\psi \propto e^{ikr^2/2R}$ [cf. Eq. (3.5) in Reference W above, or Eq. (7.35) of Reference G: chapter 7, "Diffraction," of Blandford and Thorne, *Applications of Classical Physics.*] What will be the transverse variation of this same wave (a) after reflecting off a planar mirror set up normal to the z axis? (b) after reflecting off a concave spherical mirror with radius of curvature $R_m = R$? (c) after passing through a converging lens with focal length f?

Lecture 8
Lasers and Input Optics — II.

by Alex Abarmovici, 22 April 1995

Abramovici lectured from the following transparencies. Kip has annotated them, based on Abramovici's lecture.

RECOMBINED, POWER RECYCLED FABRY-PEROT INTERFEROMETER



FREQUENCY NOISE SPECIFICATION

the ΔL we want to measure

$$\boxed{\Delta\nu} = \frac{\boxed{\Delta L}}{L}\cdot\nu\cdot\frac{B}{\Delta B}$$

the maximum frequency fluctuations we can live with

- ΔL: **displacement noise.**
- L: **interferometer arm length.**
- ν: **frequency of light.**
- B: **arm cavity bandwidth.**
- $\Delta B/B$: **Degree of matching between interferometer arms. Working assumption: 0.001.**

POWER STABILITY SPECIFICATION

(fluctuations of laser power I cause fluctuations of phase shift and thence of the measured ΔL)

$$\boxed{\Delta I} = I \cdot \frac{\Delta L}{x_0}$$

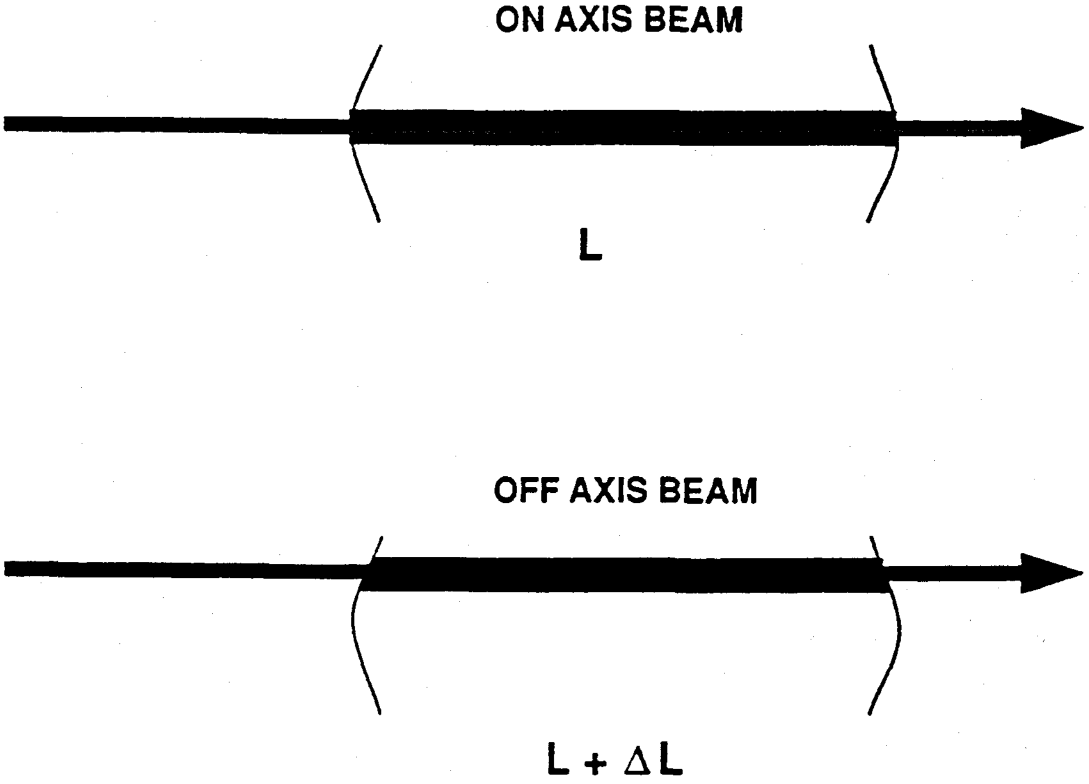
↑ maximum power fluctuations we can live with

- I: **optical power.**
- x_0: **interferometer arm detuning from resonance, expressed in length units.** $x_0 = 10^{-13}$ **m was measured at the 40 m lab.**

$$X_0 = L \frac{\Delta \nu}{\nu}$$ ← detuning

LIGO

FREQUENCY NOISE CAUSED BY BEAM JITTER



$$\Delta\nu = \nu \frac{\Delta L}{L}$$

UPPER LIMIT ON BEAM JITTER

frequency noise due to fluctuations Δε in amplitudes of modes that were not supposed to be excited

$$\boxed{\Delta\nu} = \frac{\pi c}{2L\mathcal{F}^2} \cdot \frac{|\,\epsilon_0 \Delta\epsilon\,|\,\sin\phi_N}{1 + R_1R_2 - 2\sqrt{R_1R_2}\cos\phi_N}$$

- $\Delta\epsilon$: **amplitude of fluctuating higher transverse mode, responsible for beam jitter. Examples:**

  A.

  $$\Delta\epsilon = \frac{\pi w_0}{\lambda} \cdot \Delta\theta$$

  for angular beam jitter by $\Delta\theta$. w_0: beam waist parameter, λ: wavelength.

  B.

  $$\Delta\epsilon = \frac{\Delta y}{w_0}$$

  for lateral beam displacement by Δy.

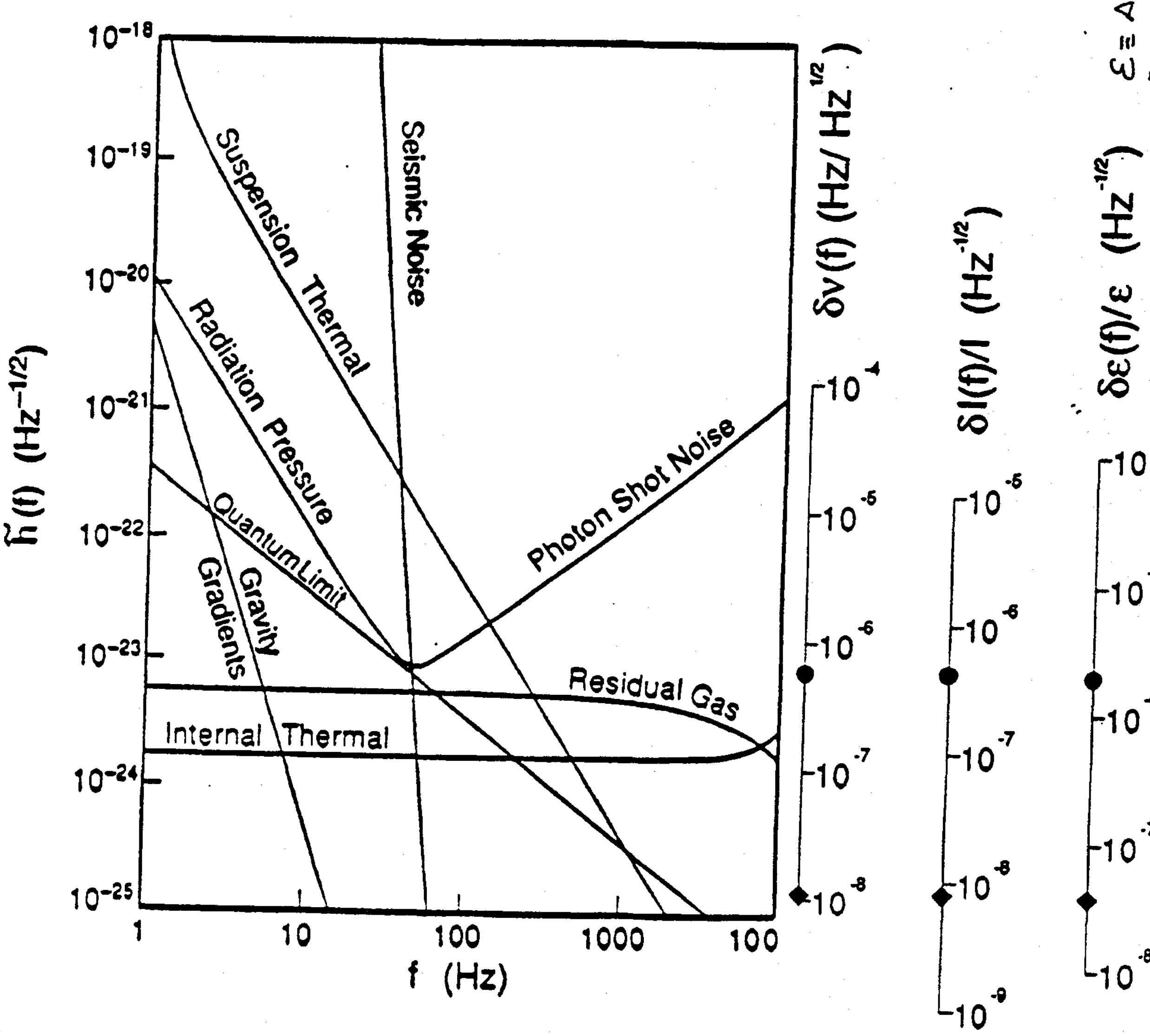
- ϵ_0: **amplitude of higher transverse mode, corresponding to static misalignment. $\epsilon_0 = 0.3$ is assumed.**
- $\mathcal{F}$: **resonator finesse.**
- $R_{1,2}$: **mirror reflectivities (power).**
- ϕ_N: **phase depending on resonator geometry and transverse mode index.**

Noise Budget For First LIGO Detectors

- **5 Watt Laser**
- **Mirror Losses 50 ppm**
- **Recycling Factor of 30**
- **10 kg Test Masses**
- **Suspension Q=10^7**

● Initial interferometers

◆ Subsequent, advanced interferometers

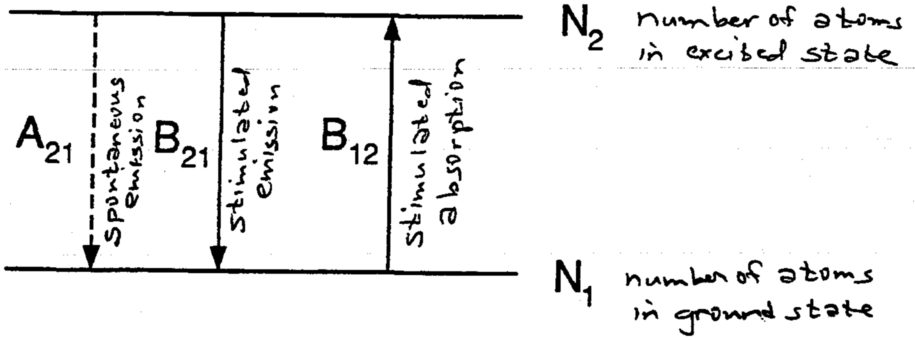


$\mathcal{E} \equiv$ Amplitude of desired TEM_{00} mode
[$\delta\mathcal{E}(f)/\mathcal{E}$ is $\approx 10^{-4}$ before stabilization]

LIGO

How a Laser Works —

Einstein Coefficients



Steady state: (e.g., if immersed in black body radiation)

$$\frac{dN_1}{dt} = A_{21}N_2 + \rho(\nu)B_{21}N_2 - \rho(\nu)B_{12}N_1 = 0$$

Boltzmann distribution:

$$\frac{N_2}{N_1} = \frac{g_2}{g_1}exp(-h\nu/kT)$$ << 1 at optical frequencies

Spectral density of energy:

$$\rho(\nu) = \frac{\frac{A_{21}}{B_{21}}}{\frac{B_{12}}{B_{21}}\frac{g_1}{g_2}exp(h\nu/kT) - 1}$$

Compare with Planck formula:

$$\rho(\nu) = \frac{8\pi\nu^2}{c^3}\frac{h\nu}{exp(h\nu/kT) - 1}$$

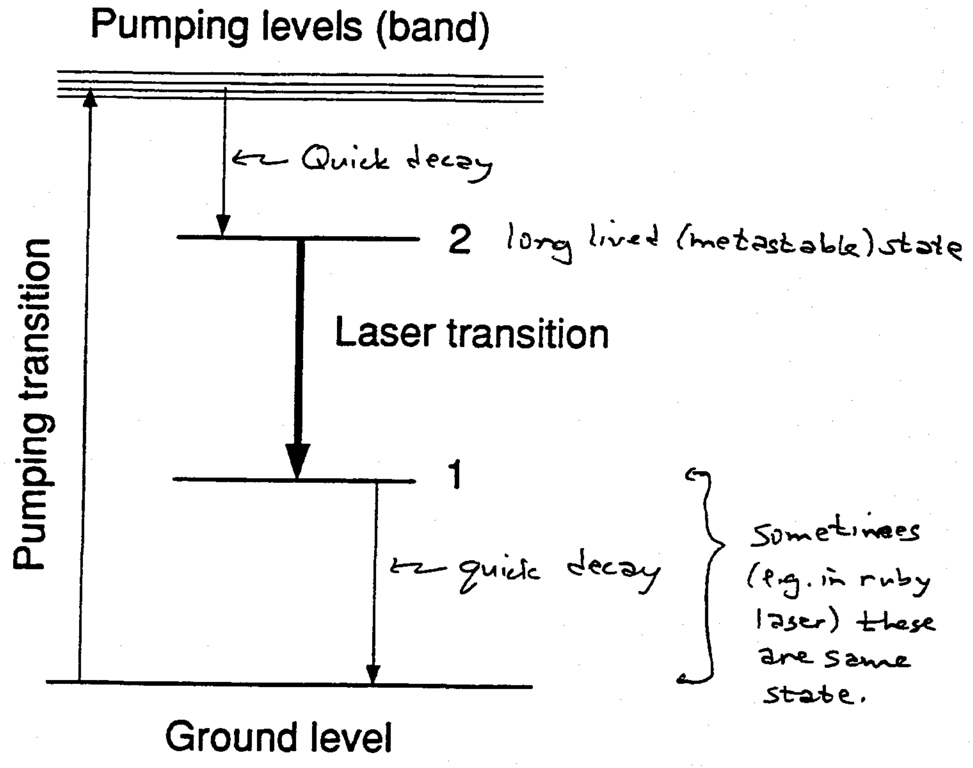
to derive relationships between Enstein coefficients:

$$B_{12} = \frac{g_2}{g_1}B_{21} \quad ; \quad \frac{A_{21}}{B_{21}} = \frac{8\pi\nu^2 h\nu}{c^3}$$

To get lasing, need $N_1 << N_2$
(population inversion)

FOUR-LEVEL LASER OPERATION

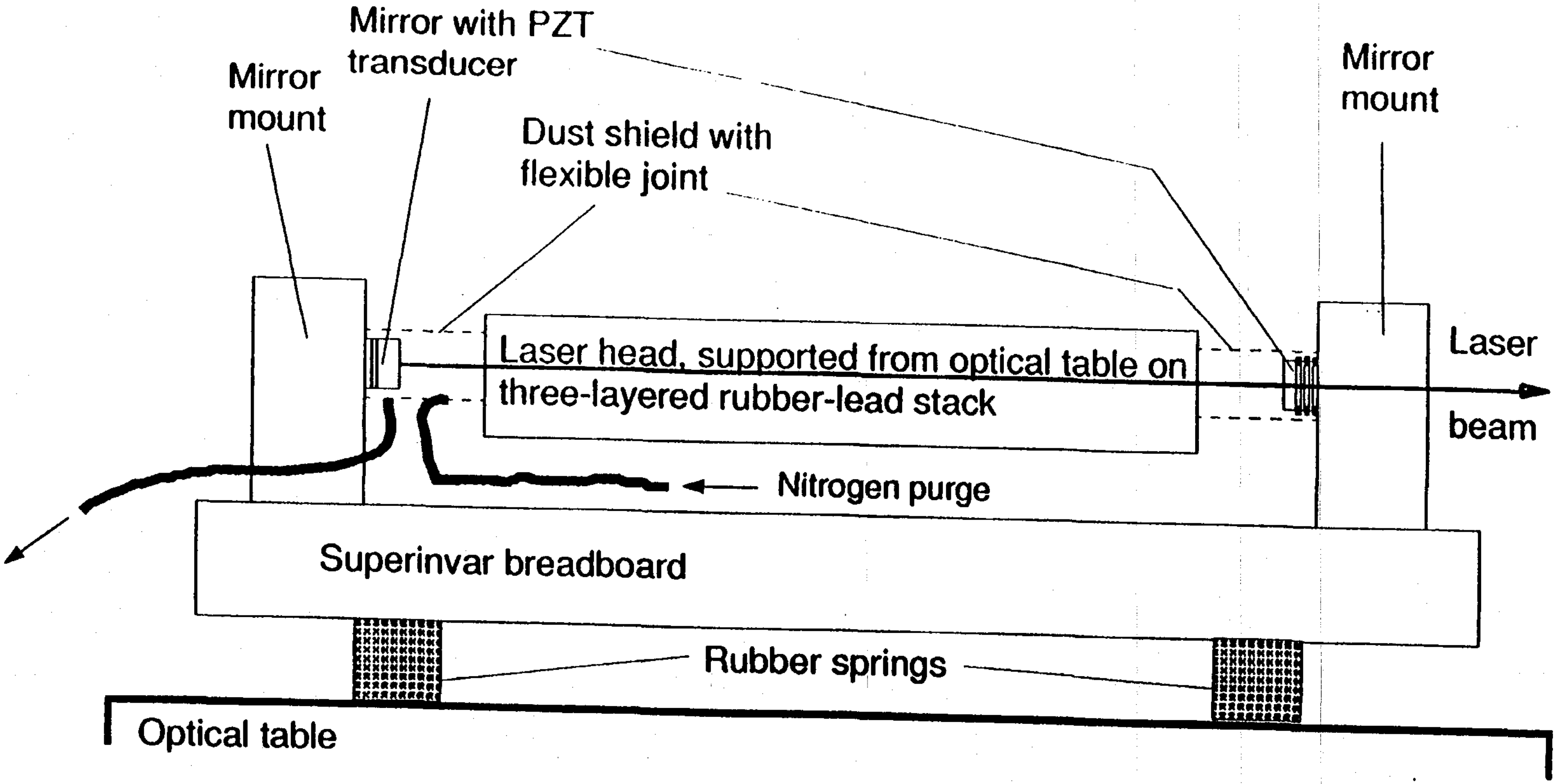
– a very common way to get population inversion



Argon Ion Lasers are currently used in the 40 m prototype and planned for use in LIGO's first interferometers. Nd:YAG lasers are under development at Stanford & elsewhere for future use in LIGO

REBUILDING OF COMMERCIAL ARGON ION LASER

(take out mirrors and attach them to a new mount that holds the cavity length more rigid, in the presence of vibrations due to water cooling system, etc.)



Mirror with PZT transducer
Mirror mount
Dust shield with flexible joint
Mirror mount
Laser head, supported from optical table on three-layered rubber-lead stack
Laser beam
Nitrogen purge
Superinvar breadboard
Rubber springs
Optical table

NEEDED NOISE SUPPRESSION FACTORS

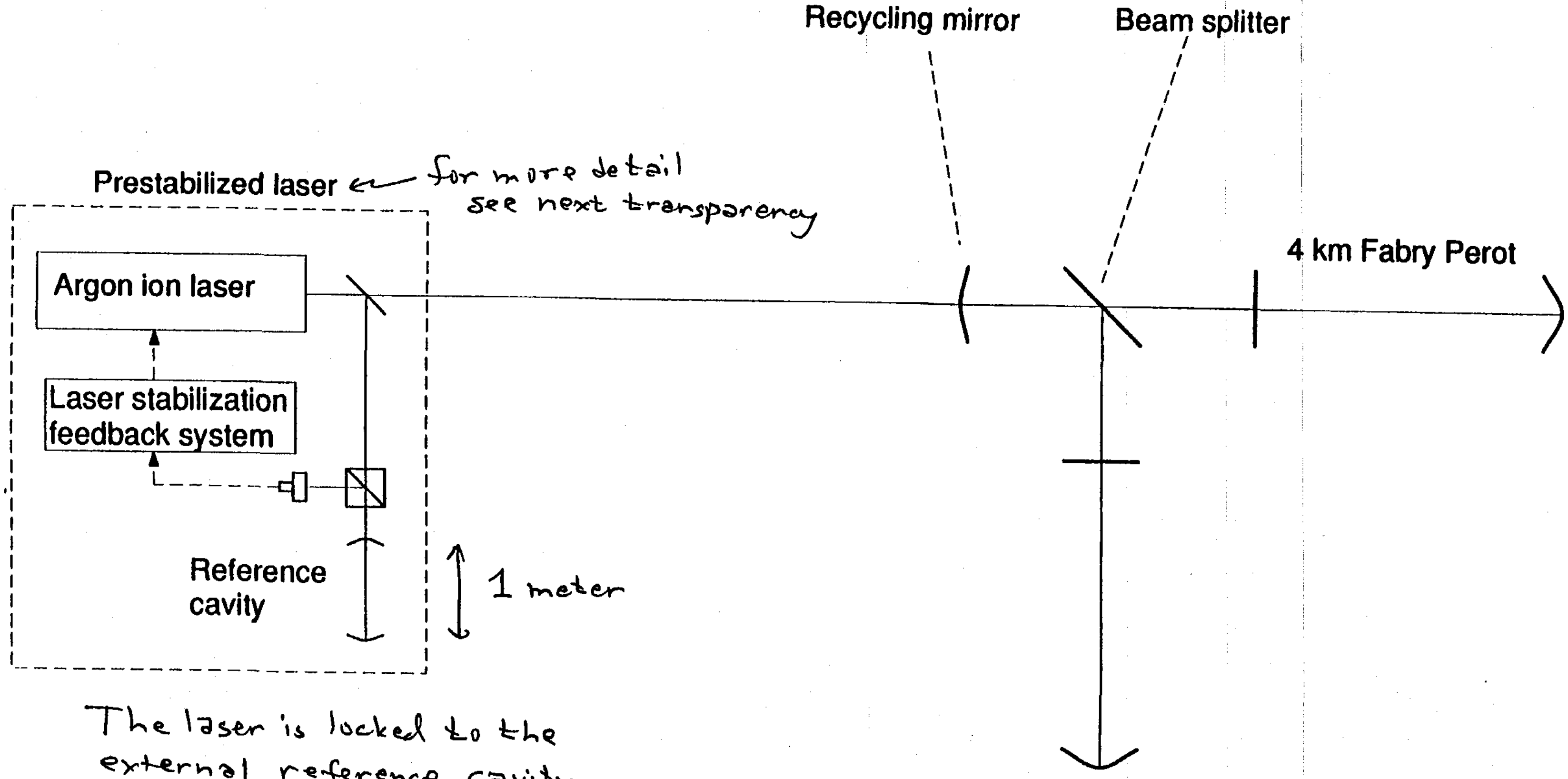
after remounting the laser's mirrors

(at 100 Hz)

Noise type	Suppression factor
Frequency noise	$5 \cdot 10^{10}$
Intensity noise	50
Beam jitter	75

LIGO

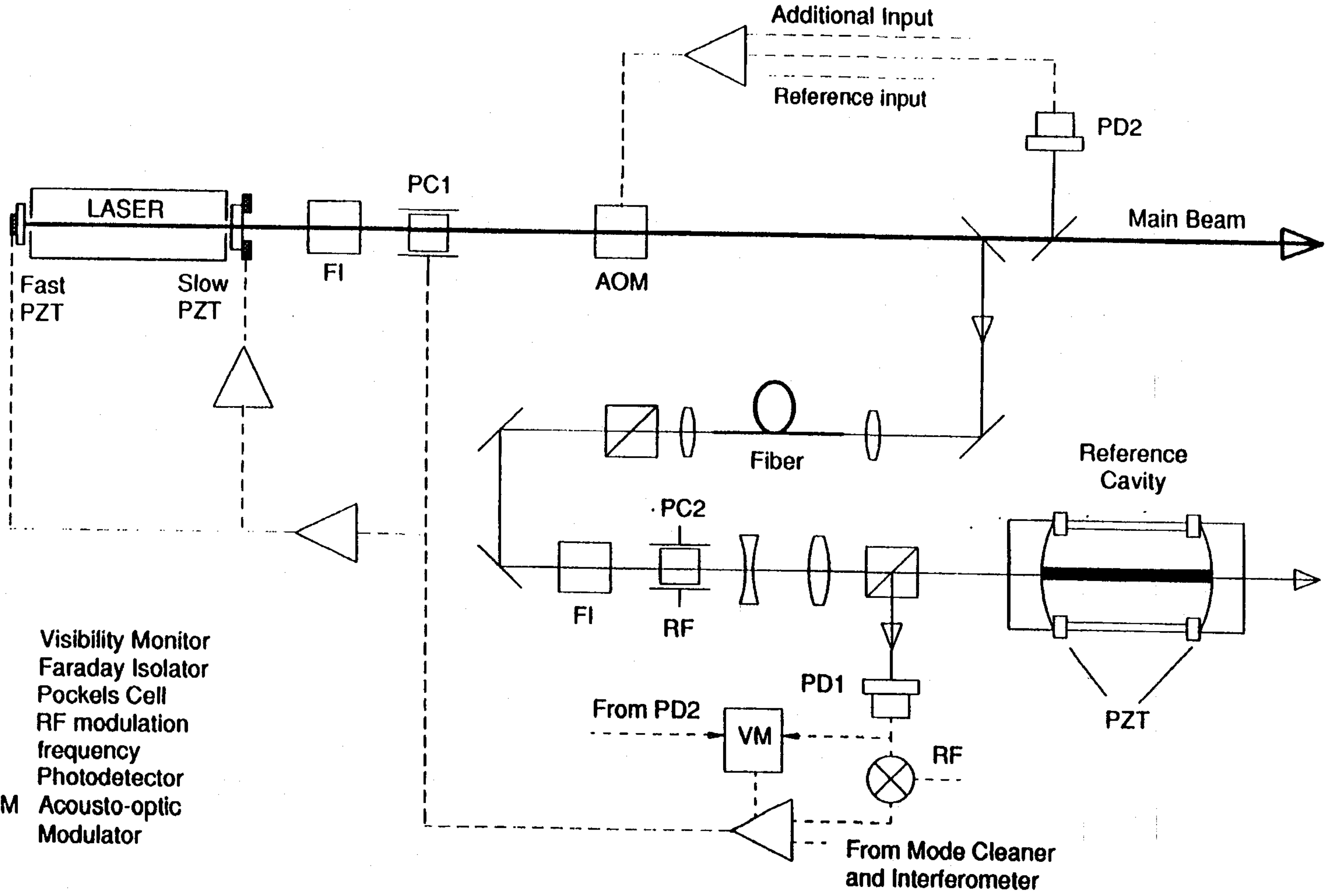
RECOMBINED, POWER RECYCLED FABRY-PEROT INTERFEROMETER



for more detail see next transparency

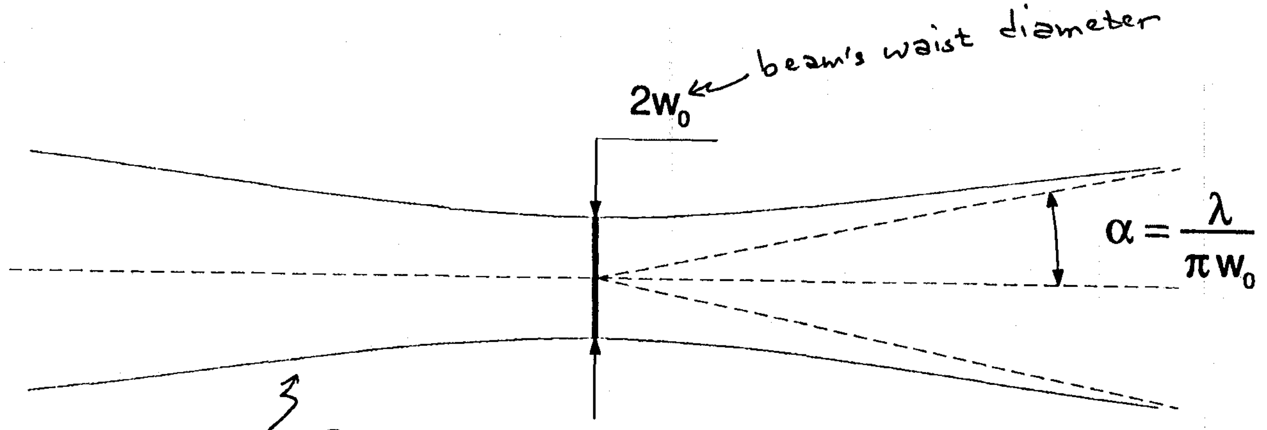
The laser is locked to the external reference cavity. This improves its frequency stability by ~106.

PRESTABILIZED LASER

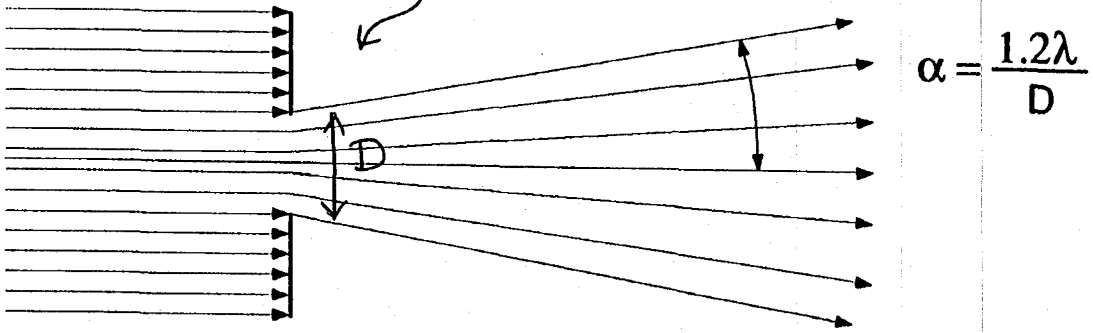


VM	Visibility Monitor
FI	Faraday Isolator
PC	Pockels Cell
RF	RF modulation frequency
PD	Photodetector
AOM	Acousto-optic Modulator

GAUSSIAN BEAMS AND DIFFRACTION

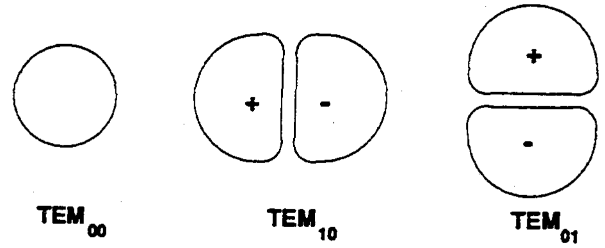


Gaussian-beam diffraction is similar to the diffraction of light passing through a circular aperature

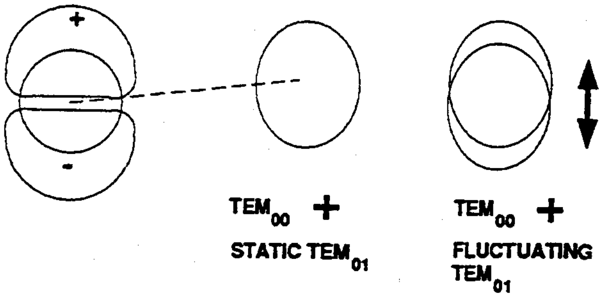


CONNECTION BETWEEN HIGHER TRANSVERSE MODES AND BEAM JITTER

ELECTROMAGNETIC FIELD MODES



EFFECT OF PRESENCE OF HIGHER MODE

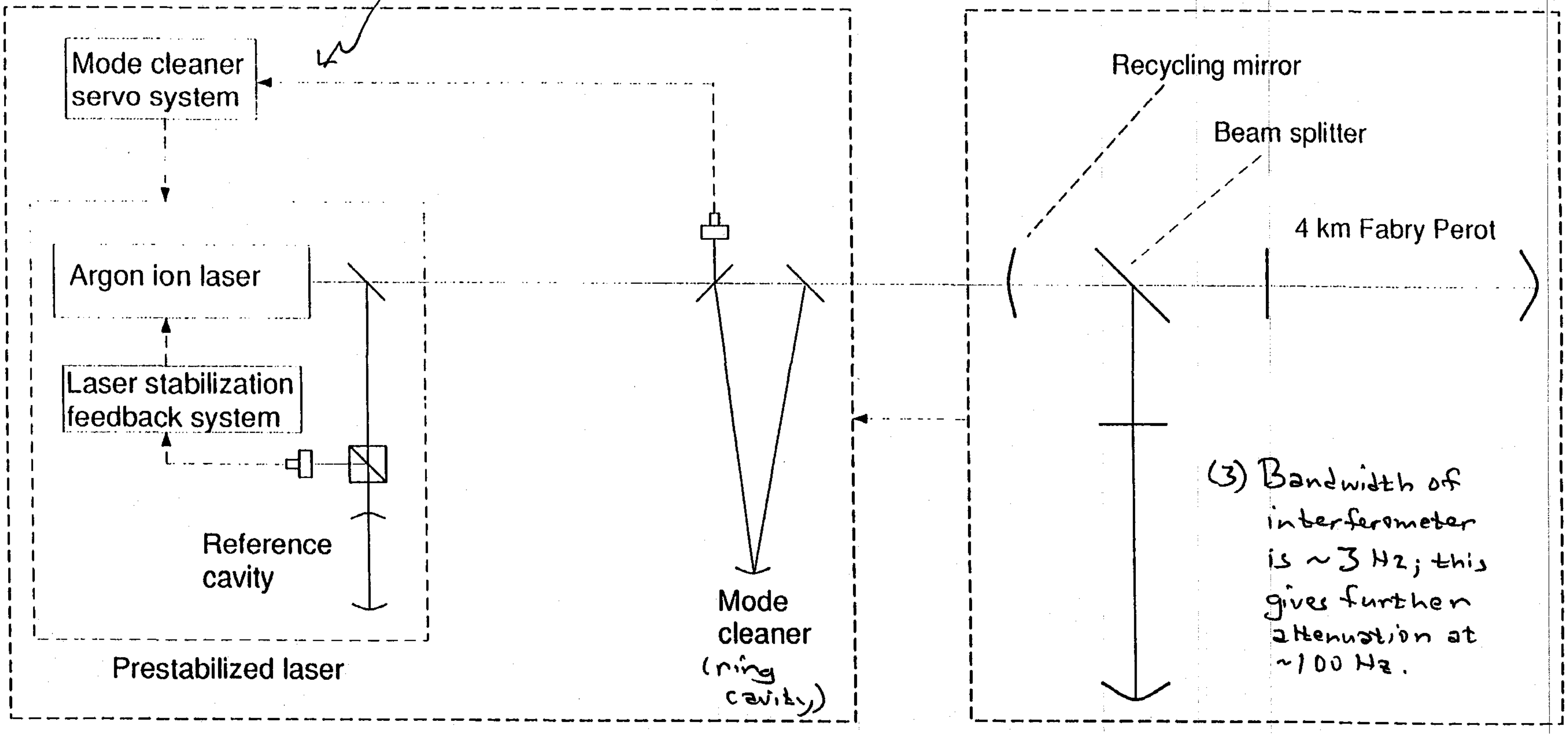


+
-
TEM$_{00}$ + STATIC TEM$_{01}$
TEM$_{00}$ + FLUCTUATING TEM$_{01}$

amplitude of TEM01 mode varying with time ⟺ beam jitter

RECOMBINED, POWER RECYCLED FABRY-PEROT INTERFEROMETER

(1) Prestabilized laser is locked to Mode Cleaning Cavity, producing an additional factor ~1000 in frequency stability



Mode cleaner servo system

Argon ion laser

Laser stabilization feedback system

Reference cavity

Prestabilized laser

Mode cleaner (ring cavity)

Recycling mirror

Beam splitter

4 km Fabry Perot

(3) Bandwidth of interferometer is ~3 Hz; this gives further attenuation at ~100 Hz.

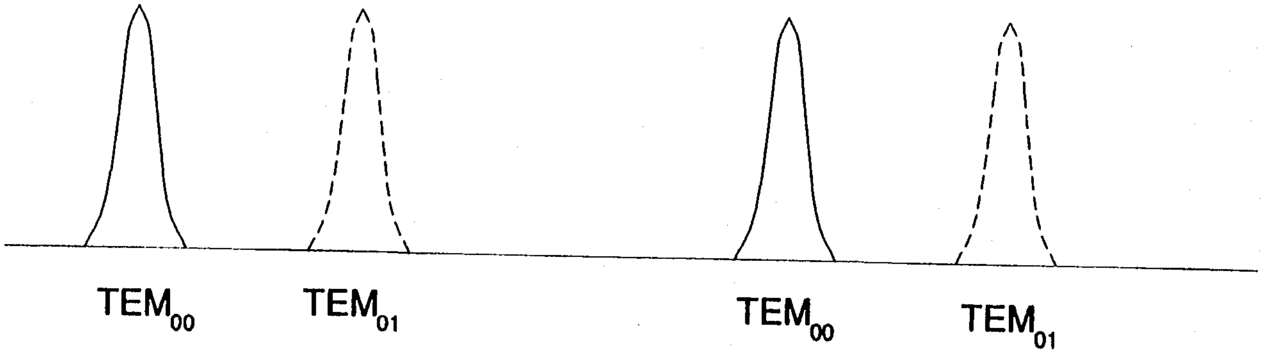
Light source: laser and input optics

Interferometer

(2) "Light source" is locked to one arm of interferometer, producing an additional ~100 of frequency stability

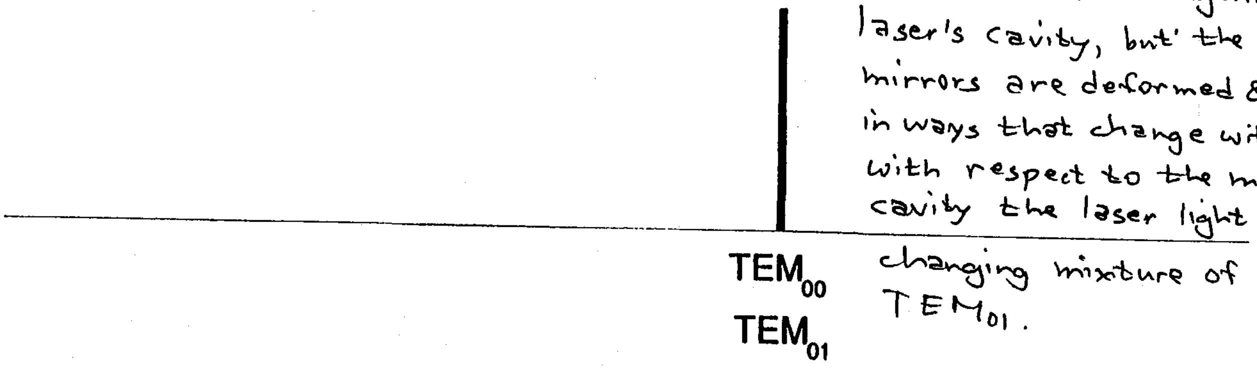
HOW A MODE CLEANER WORKS

SPECTRUM OF CAVITY RESONANCES



TEM$_{00}$ TEM$_{01}$ TEM$_{00}$ TEM$_{01}$

SPECTRUM OF LASER BEAM ...



TEM$_{00}$

TEM$_{01}$

Laser beam is an eigenmode of laser's cavity, but the laser's mirrors are deformed & misaligned in ways that change with time; so with respect to the mode cleaning cavity the laser light is a dynamically changing mixture of TEM$_{00}$ and TEM$_{01}$.

TEM$_{00}$ is kept locked to the cavity resonance by a powerful feedback system.

TEM$_{01}$ is thus kept off **its** cavity resonance, and is not resonating. This mode will thus be considerably weakened in the transmitted beam.

BATCH
START

9. Optical Elements

STAPLE
OR
DIVIDER

LECTURE 9.
OPTICAL ELEMENTS

Lecture by Rick Savage

Assigned Reading:

Y. W. Winkler, K. Danzmann, A. Rüdiger and R. Schilling, "Optical Problems in Interfereometric Gravitational Wave Antennas," in *The Sixth Marcel Grossmann Meeting*, eds. H. Sato and T. Nakamura (World Scientific, Singapore, 1991), pp. 176–191.

H. A. E. Siegman, *Lasers* (University Science Books, Mill Valley CA, 1986), chapter 17 "Physical Properties of Gaussian Beams": Section 17.1 "Gaussian Beam Propagation," (pages 663-674); section 17.4 "Axial Phase Shifts: The Guoy Effect," (pages 682-685), and section 17.5 "Higher-Order Gaussian Modes," (pages 685-691). [This material was suggested reading in Lecture 4.] If you did not read it then, you should read it now.]

Suggested Supplementary Reading:

Z. D. Malacara, *Optical Shop Testing* (John Wiley and Sons, New York, 1978), section 1.2, "Fizeau Interferometer," pp. 19–37.

AA. H. A. Macleod, *Thin-Film Optical Filters,* 2nd edition (Adam Hilger Ltd., Bristol, 1986), "Introduction," pp. 1–10.

SS. J. M. Elson, H. E. Bennett, and J. M. Bennett, "Scattering from Optical Surfaces," in *Applied Optical Engineering*, Vol. VII (Academic Press 1979), Chapter 7, page 191. [This is reproduced later, in connection with Lecture 15.]

A Few Suggested Problems

1. *Gaussian Beam Propagation.* The present conceptual design for the 4 km long LIGO arm cavities specifies that the input mirror be flat and the end mirror curved, with a radius of curvature of 6 km. The beam waist is therefore located on the flat input mirror and the spot size is significantly larger on the curved mirror than on the flat.
   a. Consider employing a symmetrical curved-curved mirror configuration instead of the flat-curved geometry. What is the required radius of curvature of the mirrors ($R_1 = R_2$) to maintain the cavity g factor product at $g_1 g_2 = 1/3$?
   b. Calculate the spot size at the beam waist and on the mirrors. What is the Rayleigh range for this configuration?
   c. What factors might influence the decision to adopt either the flat-curved or the symmetrical, curved-curved geometry?
2. *Scattering from mirror surface irregularities.* Consider the following simple model for scattering from mirror surface irregularities. Represent the mirror surface height $z = \mu(x, y)$ as a superposition of a number of spatially monochromatic terms. Assume, for the moment, that only one term is non-zero, and let that term have peak height a and wavelength Λ; i.e. set

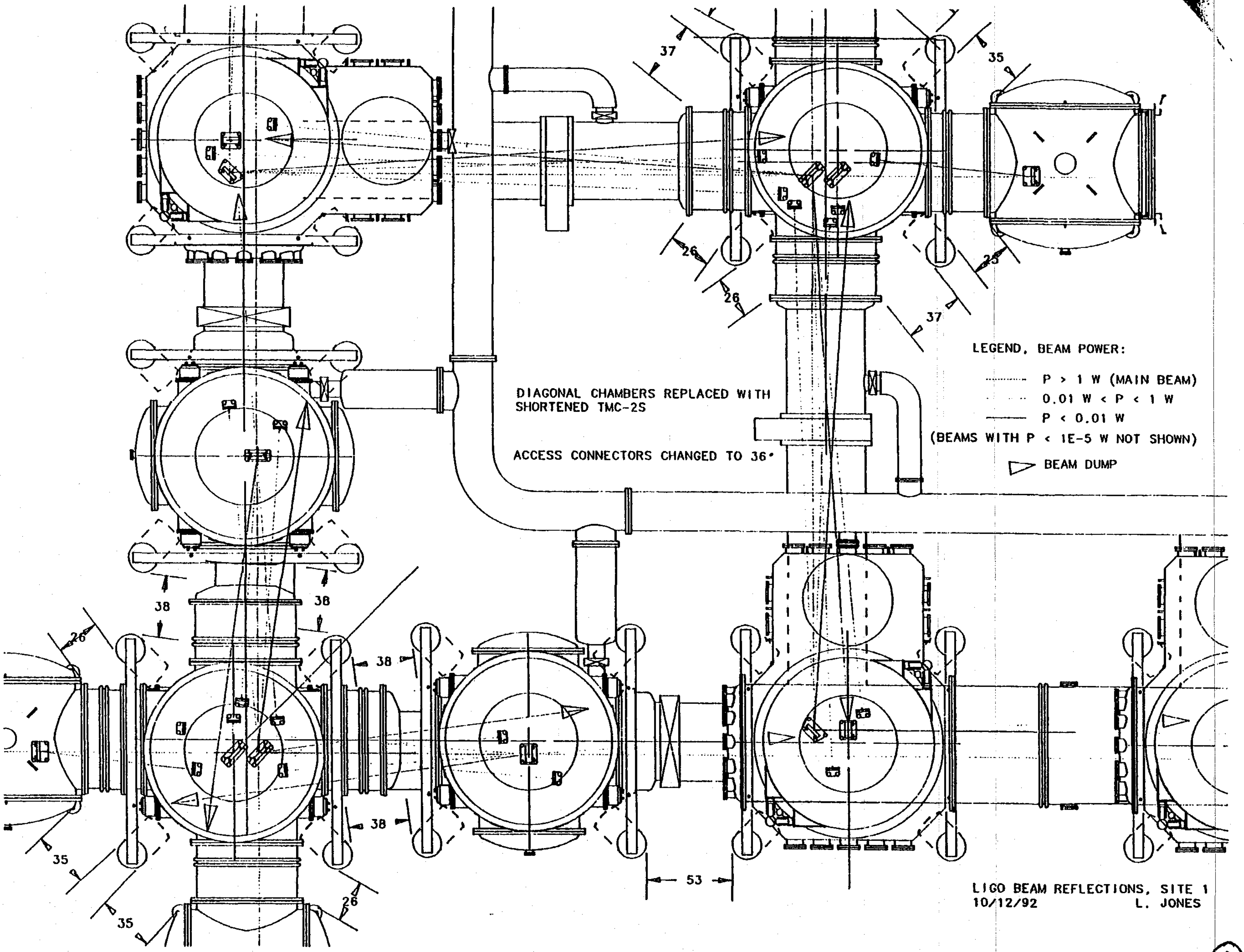
$$\mu(x, y) = a\cos(2\pi x/\Lambda).$$

   Idealize the mirror to be of infinite extent and irradiated by a plane wave at normal incidence.
   a. Show that the wave reflected from the surface has spatial sidebands that propagate at some angle θ relative to the specularly reflected beam. What is θ? [Hint: This can be regarded as an exercise in Fraunhofer diffraction (Why?); see, e.g., Section 7.3 of chapter 7, "Diffraction", of Blandford and Thorne, *Applications of Classical Physics* (which was passed out in Lecture 4).]
   b. Find the power scattered into these sidebands as a function of the amplitude of the surface variation, a.
   c. Generalize to find the total light scattered into all angles from a surface with a total rms irregularity σ.

Lecture 9
Optical Elements

by Rick Savage, 27 April 1995

Savage lectured from the following transparencies. Kip has added a few annotations, based on Savage's lecture.



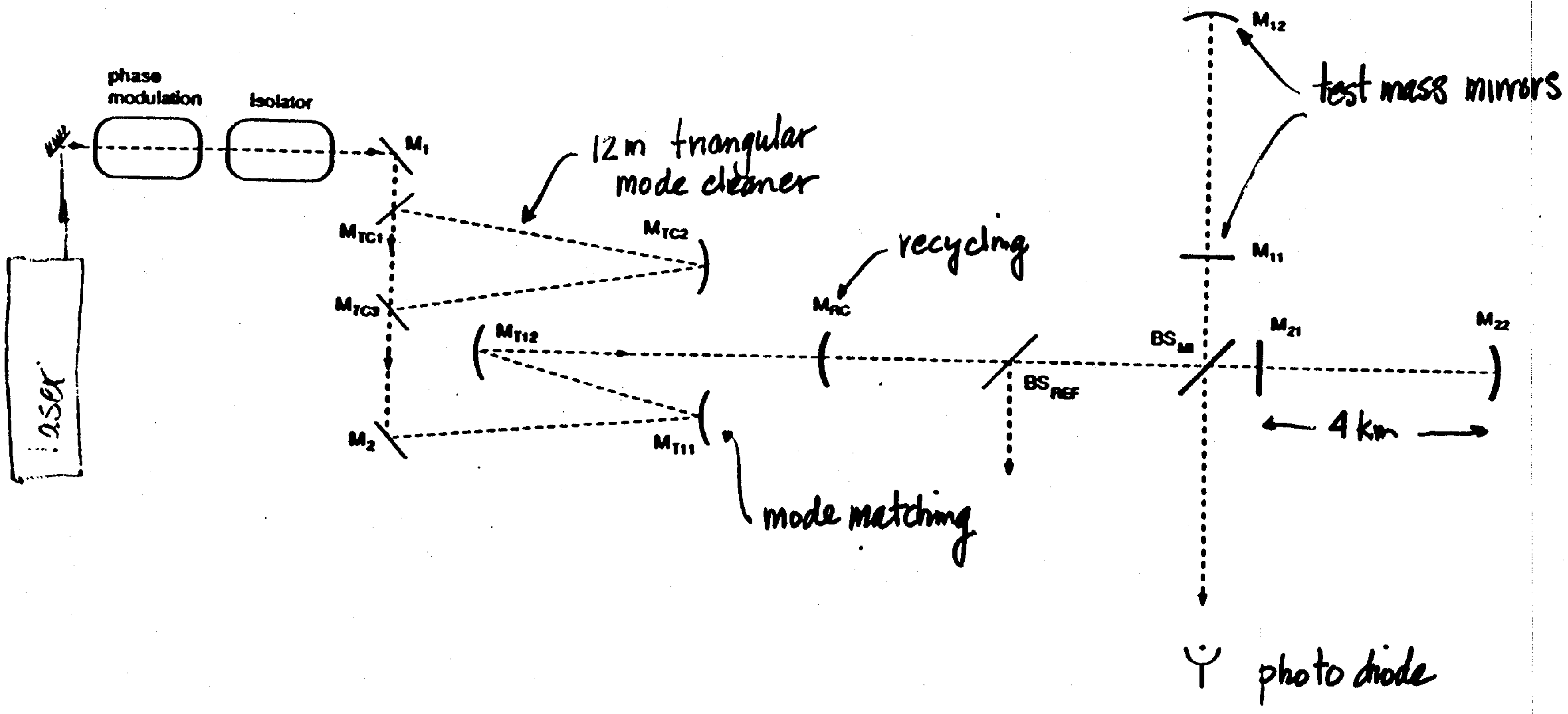
DIAGONAL CHAMBERS REPLACED WITH SHORTENED TMC-2S

ACCESS CONNECTORS CHANGED TO 36"

LEGEND, BEAM POWER:

········ P > 1 W (MAIN BEAM)
· · · · 0.01 W < P < 1 W
——— P < 0.01 W
(BEAMS WITH P < 1E-5 W NOT SHOWN)
▷ BEAM DUMP

LIGO BEAM REFLECTIONS, SITE 1
10/12/92 L. JONES



Initial interferometer optical schematic

Kip

TABLE 2

OPTICAL POWER AND INTENSITY AT VARIOUS COMPONENTS

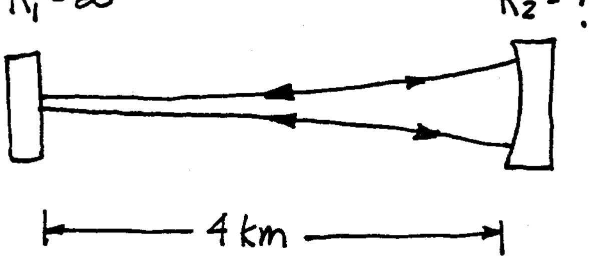
component	ω (cm)	Power (W)	Intensity (W/cm^2)
ϕ modulator	7×10^{-2}	4	4×10^2
isolator	5×10^{-2}	4	8×10^2
mode filter (flat)	1×10^{-1}	4×10^3	2×10^5
mode filter (curved)	2×10^{-1}	4×10^3	6×10^4
telescope out. mir.	2.1	3	4×10^{-1}
recycling mir.	2.1	8×10^1	1×10^1
beam splitter	2.1	8×10^1	1×10^1
arm cavity input mir.	2.1	4.5×10^3	6.7×10^2
arm cavity far mir.	3.8	4.5×10^3	2×10^2
main frame laser	5×10^{-2}	1×10^2	3×10^4

Arm cavity optical Parameters.

1) Arm cavity length = L_{arm} = $4 \cdot 10^5$ cm.

2) TEM_{00} mode

3) Arm input mirror flat : $R_{front\ arm} = R_1 = \infty$



To choose R_2, look at higher-order gaussian beam mode frequencies:

number of half wavelengths between mirrors

$$\nu_{qmn} = \frac{c}{2L}\left(q + \frac{1}{\pi}(m+n+1)\cos^{-1}\sqrt{g_1 g_2}\right)$$

$$g_{1,2} = 1 - L/R_{1,2} \qquad g_1 = 1 \ (R_1 = \infty)$$

* avoid having higher-order transverse modes degenerate with higher-order longitudinal modes.

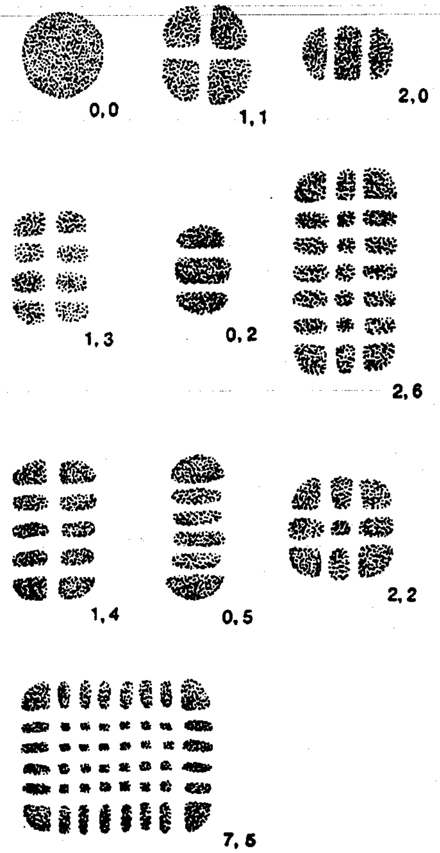


FIGURE 17.20
Transverse mode patterns for Hermite-gaussian modes of various orders.

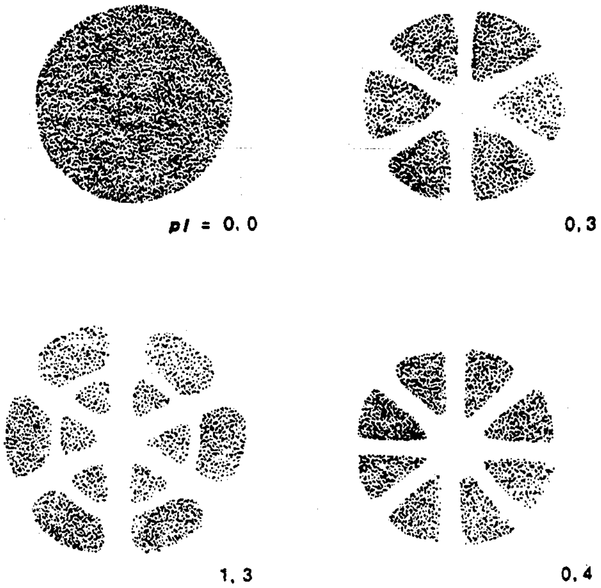


FIGURE 17.21
Transverse mode patterns for Laguerre-gaussian modes of various orders.

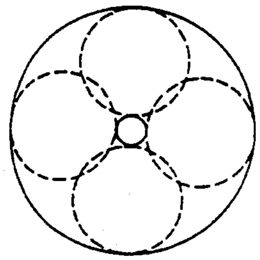


FIGURE 17.22
The "donut" mode is a linear superposition of 10 and 01 Hermite-gaussian modes.

for $R_2 = 6 \cdot 10^5$ cm, $g_2 = 1/3 = g_1 g_2$

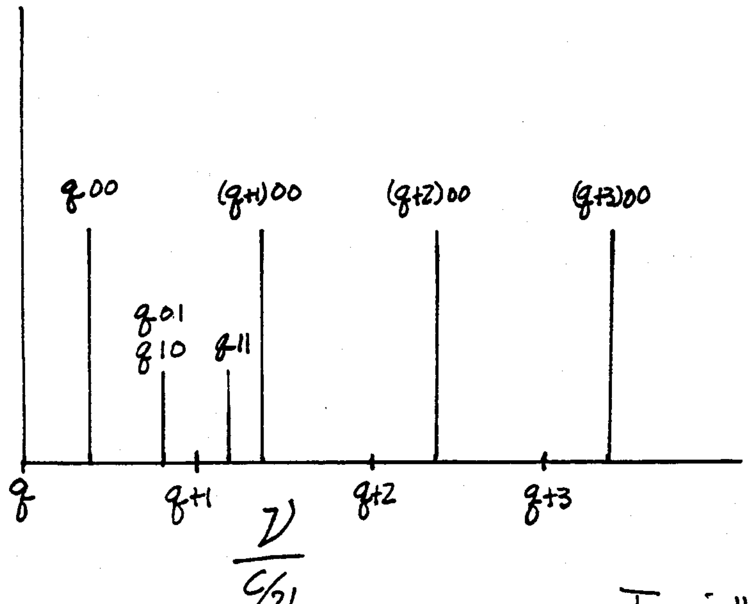
$$\frac{\cos^{-1}\sqrt{1/3}}{\pi} = .392.$$

$$\nu_{q00} = \frac{c}{2L}(q + .392)$$

$$\nu_{q01} = \nu_{q10} = \frac{c}{2L}(q + .784)$$

$$\nu_{q11} = \frac{c}{2L}(q + 1.176)$$

$$\vdots$$



$R_1 = \infty$, $R_2 = 6 \cdot 10^5$ cm

Spot sizes on mirrors ?

Typically $m+n \sim 15$ is the lowest order that will resonate along with $m=n=0$ (TEM$_{00}$)

Gaussian Beam Propagation

$$I(r) \propto E(r)^2 = I_0 e^{-2r^2/w^2}$$

$w \equiv$ spot size $\qquad E(r) = E_0 e^{-r^2/w^2}$

for $r = w \quad E(w) = 1/e\, E_0 \qquad I(w) = 1/e^2 I_0.$

$$w(z) = w_0 \left[1 + \frac{z^2}{z_0^2}\right]^{1/2}$$

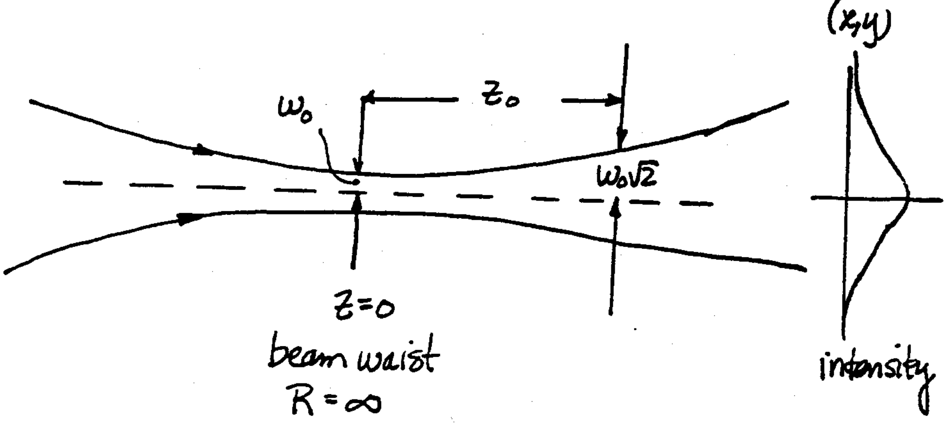
w_0 = spot size at beam waist (minimum spot size).

$$z_0 = \frac{\pi w_0^2}{\lambda}$$

z_0 = Rayleigh range

$w(z_0) = w_0 \sqrt{2}$

$$R(z) = z + \frac{z_0^2}{z}$$



$w_1 = w_0 = 2.15$ cm

$w_2 = 3.73$ cm

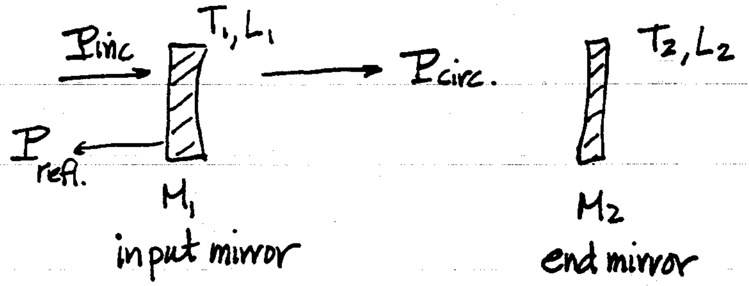
$z_0 = 2.83 \cdot 10^5$ cm

$\lambda = 5.145 \cdot 10^{-5}$ cm.

Mirror Transmission and Reflection

T = transmission
R = reflection
L = losses

$T + R + L = 1$ Cons. of energy.



Throw away as little light as possible.

$\Rightarrow$ minimize $T_2 \longrightarrow T_2 \cong 10\,ppm \quad (1 \cdot 10^{-5})$

minimize $L_1, L_2 \longrightarrow$ supermirrors $\quad L_1, L_2 \cong 100\,ppm$

what about T_1?

maximize number of bounces (cavity gain)?

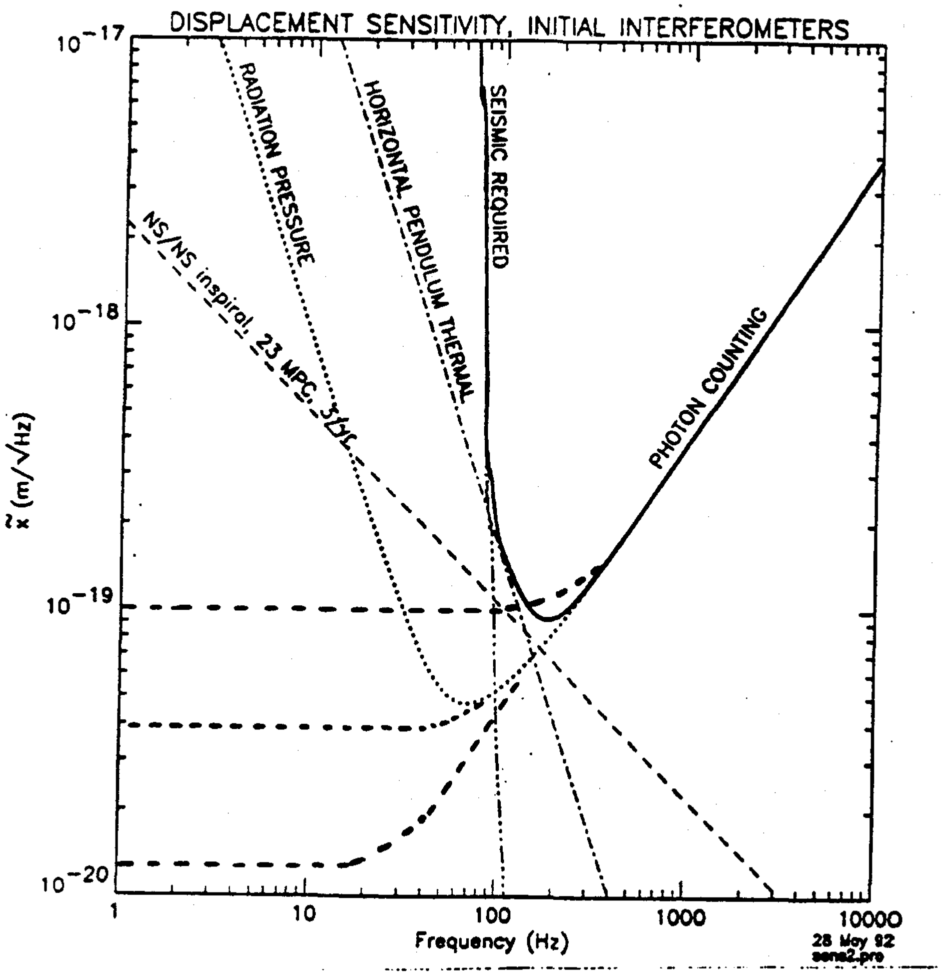
$$T_{1\,max} \cong T_2 + L_1 + L_2 = 210\,ppm.$$

$$\text{cavity gain} = \frac{P_{circ}}{P_{inc}} = 4700$$

photon life time τ_p : $P(t) = P_0 e^{-t/\tau_p}$

$$\tau_p = 63\,msec.$$

cavity pole frequency $f_c = \frac{1}{4\pi \tau_p} = \underline{\underline{1.3\,Hz}}$


DISPLACEMENT SENSITIVITY, INITIAL INTERFEROMETERS
10^{-17}
10^{-18}
10^{-19}
10^{-20}
$\tilde{x}$ (m/√Hz)
RADIATION PRESSURE
HORIZONTAL PENDULUM THERMAL
SEISMIC REQUIRED
NS/NS inspiral, 23 MPC, 3/yr
PHOTON COUNTING
1
10
100
1000
10000
Frequency (Hz)
28 May 92
sens2.pro

Optimum T_1

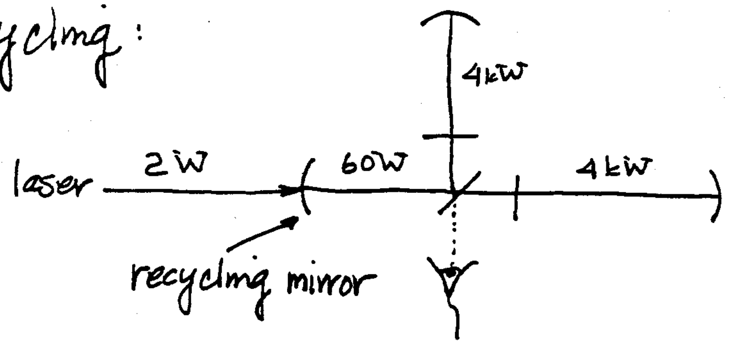
* Store light in arm cavities no longer than necessary to keep shot noise limited sensitivity below other noise sources.

$$\Rightarrow \quad f_c = 90\text{ Hz} \quad \longrightarrow \quad T_1 \cong .03 \quad (30{,}000\text{ppm})$$

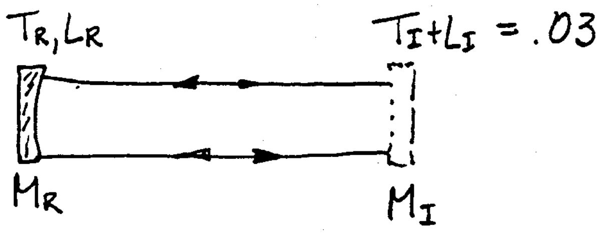
arm cavity gain $\cong$ number of bounces (round trip)
$= 130. \qquad (\tau_p \simeq .9\text{ msec})$

arm cavity reflectivity $\cong 1 - [130 \times (T_2 + L_1 + L_2)]$
$= .973 \quad$ (3% loss in arms)

Power recycling:



recycling cavity:

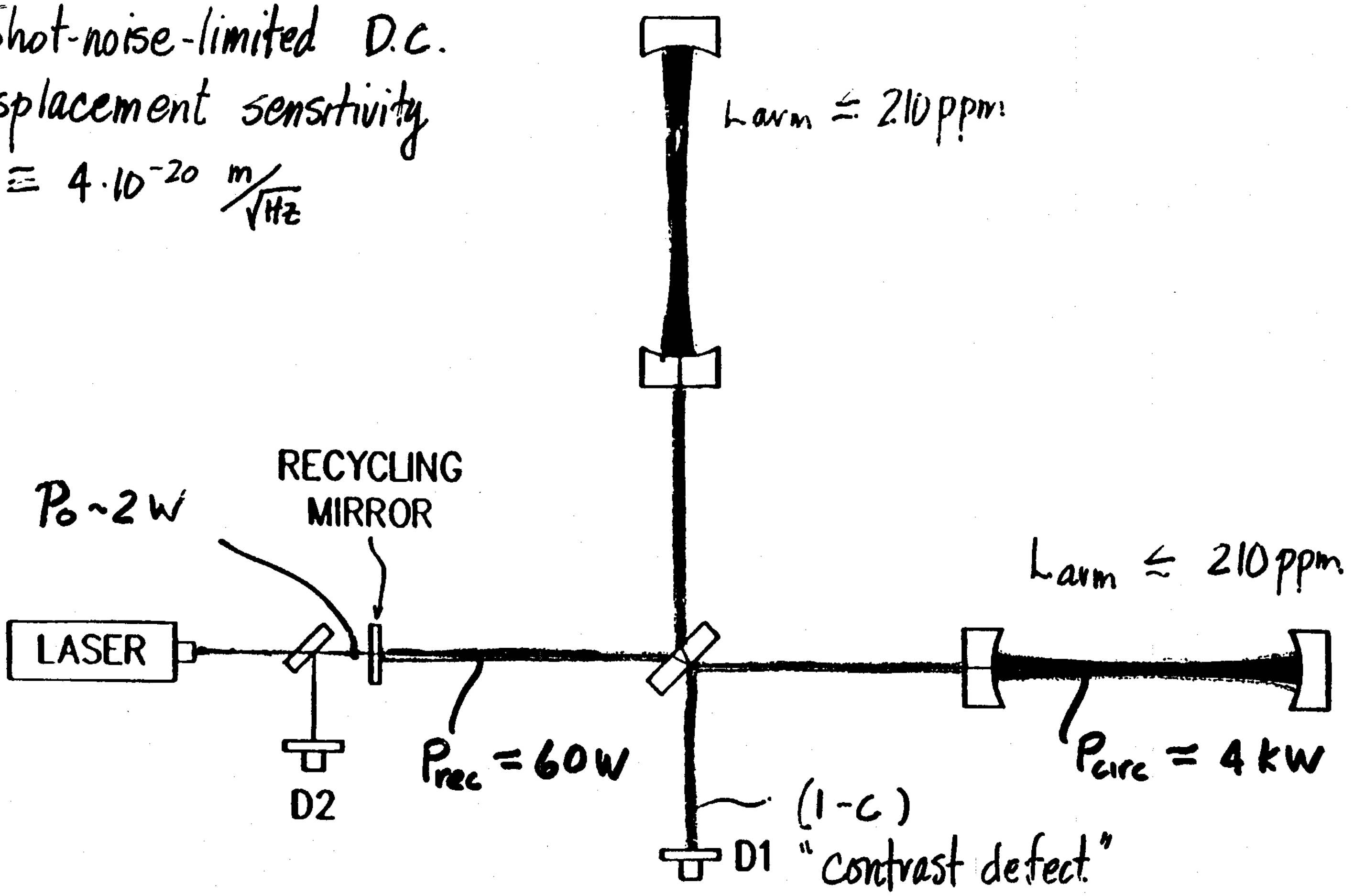


$L_R \cong 100$ppm

$T_{1\,max} = T_I + L_I + L_R$
$\cong 3\%$

$\Rightarrow$ recycling cavity gain $\cong \underline{\underline{30}}$

Shot-noise-limited D.C. displacement sensitivity $\cong 4 \cdot 10^{-20}\ \mathrm{m}/\sqrt{\mathrm{Hz}}$



Mirror imperfections —

Cause → 1) loss due to scattering and absorption

2) contrast defect (imperfect interference at antisymmetric port,

Contrast defect $\equiv 1-C$ $C = \text{Contrast} = \dfrac{\text{bright} - \text{dark}}{\text{bright} + \text{dark}}.$

$$1-C = \frac{\text{bright}+\text{dark}-\text{bright}+\text{dark}}{\text{bright}+\text{dark}} \cong \frac{2\,\text{dark}}{\text{bright}}$$

goal $(1-C) \leq 3\cdot10^{-3}$

for first LIGO interferometers

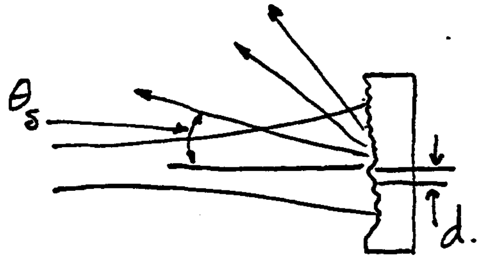
$$\Rightarrow \text{Power}_{\text{dark}} \cong \frac{3\cdot10^{-3}\cdot60\text{W}}{2.} \cong \underline{\underline{90\text{mW}}}$$

this is the resulting power onto photodetectors

$L_{\underset{\text{rec.}}{1-C}} \cong 1500\,\text{ppm}$ $L_{\underset{1-C}{\text{arm}}} \lesssim \dfrac{1500}{130} = 12\,\text{ppm}.$

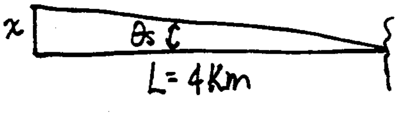
Scattering from surface imperfections —

reflection grating



$n\lambda = d(\sin\theta_i + \sin\theta_s)$

$\sin\theta_s = \dfrac{\lambda}{d}$ d = period of surface imperfection



$\sin\theta_s \simeq \theta_s = \dfrac{\lambda}{d} \cong \dfrac{x}{L}$

$x \simeq \dfrac{\lambda L}{d} \simeq \dfrac{20\,\text{cm}^2}{d}$

For $d \lesssim 1$ cm, the scattered Light misses the mirror at the other end of the interferometer arm

Spatial scales –

1) .05 – 3 cm^{-1} $(\frac{1}{d})$

Scattered light stays on mirrors - is coupled into higher-order transverse modes. - small contribution to total arm cavity loss

- primary effect - contrast defect

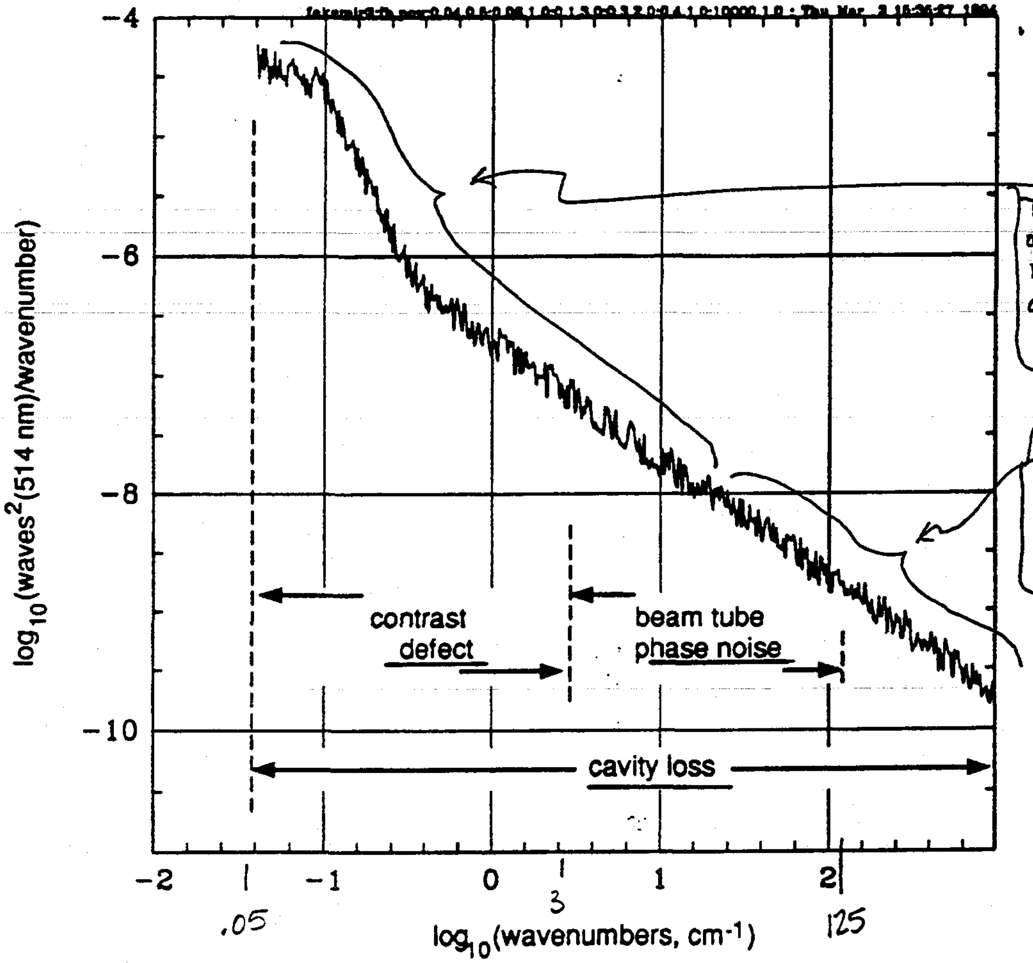
2) 3 - 125 cm^{-1} scattered light in the ligo beam tubes

- contributes to total arm cavity loss
- primary effect - phase noise due to scattering from moving beam tubes and baffles.

3) > 125 cm^{-1} ["microroughness"]

- large component of total arm cavity loss
- supermirrors

Composite hypothetical 1-D power spectrum
low spatial freq. – high quality small telescope mirror
high spatial freq – from literature : $\sigma_{>125 cm^{-1}}$ ~2.3 Å



• Cal flat mirror
AXAf mirror

This is based on actual measurements of the "Calflat" mirror

This is what the LIGO team thinks can be done

total losses from > 3 cm⁻¹ irregularities ≅ 130 ppm.

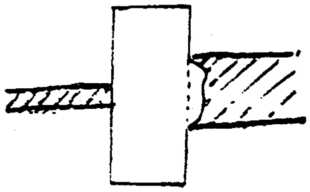
contrast defect in LIGO ~ $1 \cdot 10^3$

surface rms over spot diameter ≅ $\frac{\lambda}{500}$ (10 Å)

Mirror Absorption

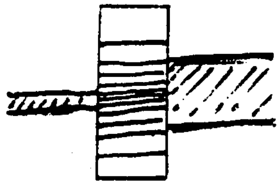
- Supermirrors have typical absorptions of < 10 ppm.
- In a high-finesse Fabry-Perot cavity, absorption in the coating is more important than absorption in the substrate.
- Absorption causes a temperature gradient in a mirror.
- Effects of temperature gradient:
    - Distortion of the mirror surface due to thermal expansion
    - Thermal lensing due to the temperature dependence of the index of refraction
- These effects limit our ability to couple power into a cavity.
- Thermal lensing is the dominant limitation.

(because fused silica substrates have very low coefficient of thermal expansion)



Less important

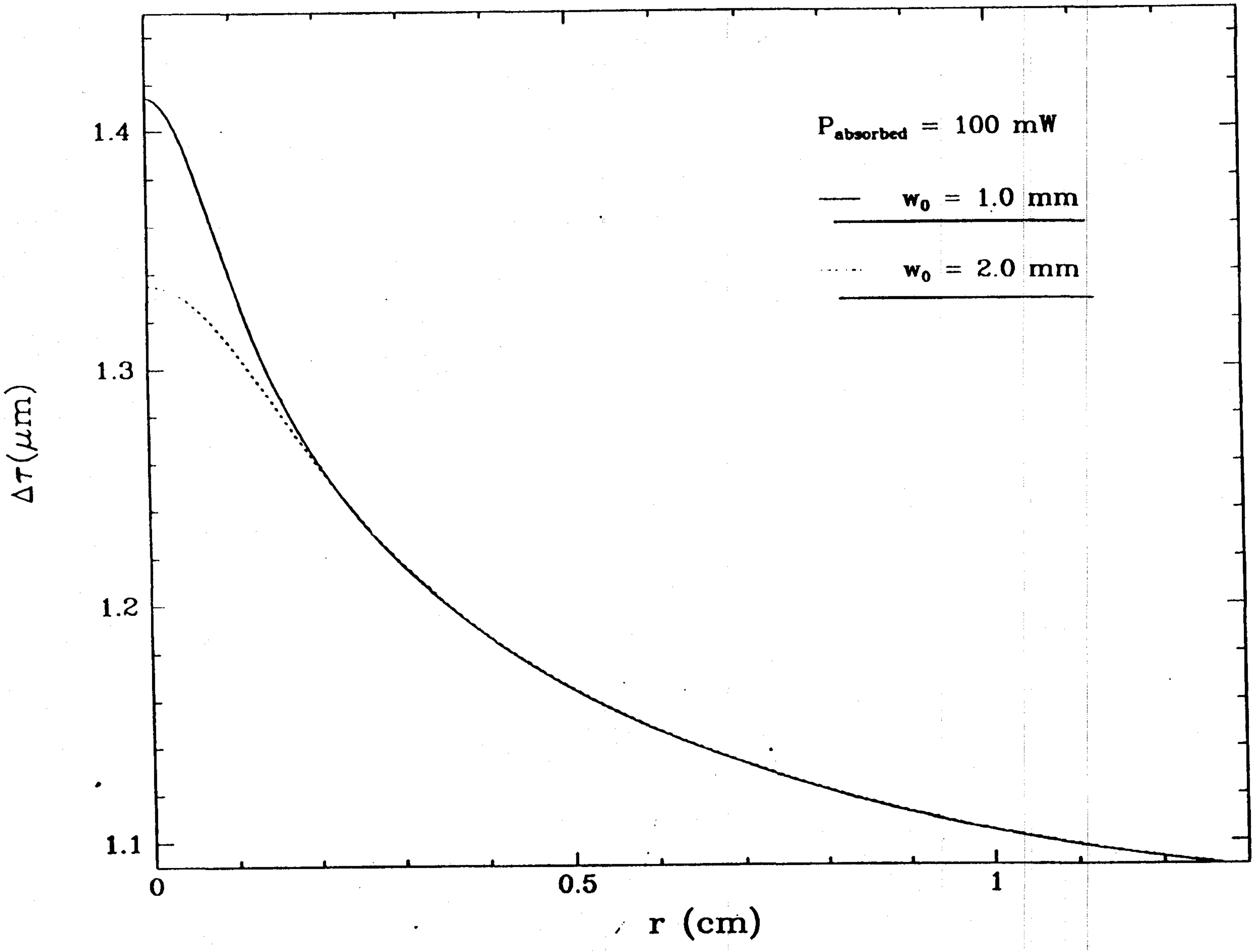
$$\frac{1}{\ell}\frac{d\ell}{dT} \approx 0.5 \times 10^{-6} K^{-1}$$



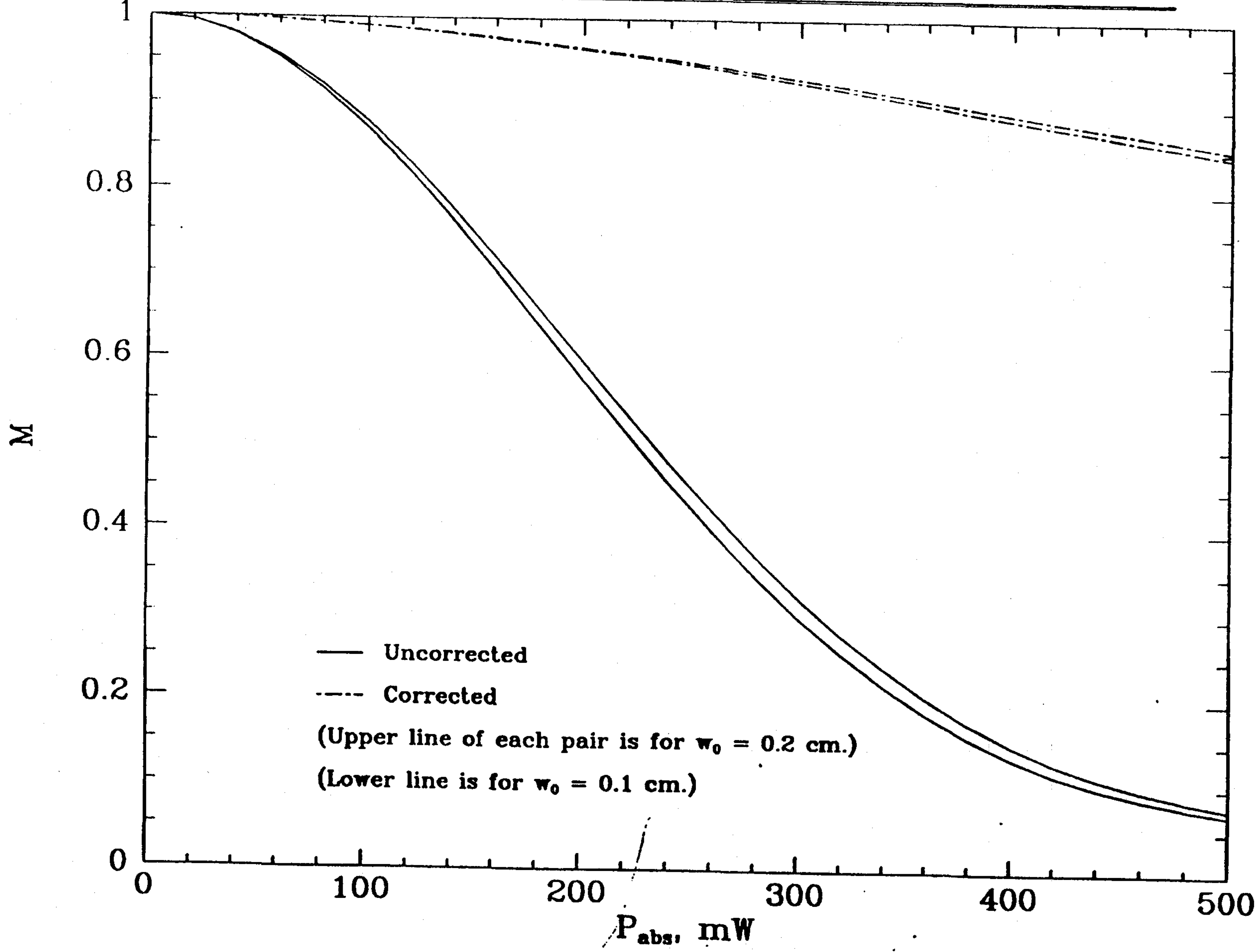
Dominant

$$\frac{dn}{dT} \approx 10^{-5} K^{-1}$$

Thermal Lensing in 1" Mirror



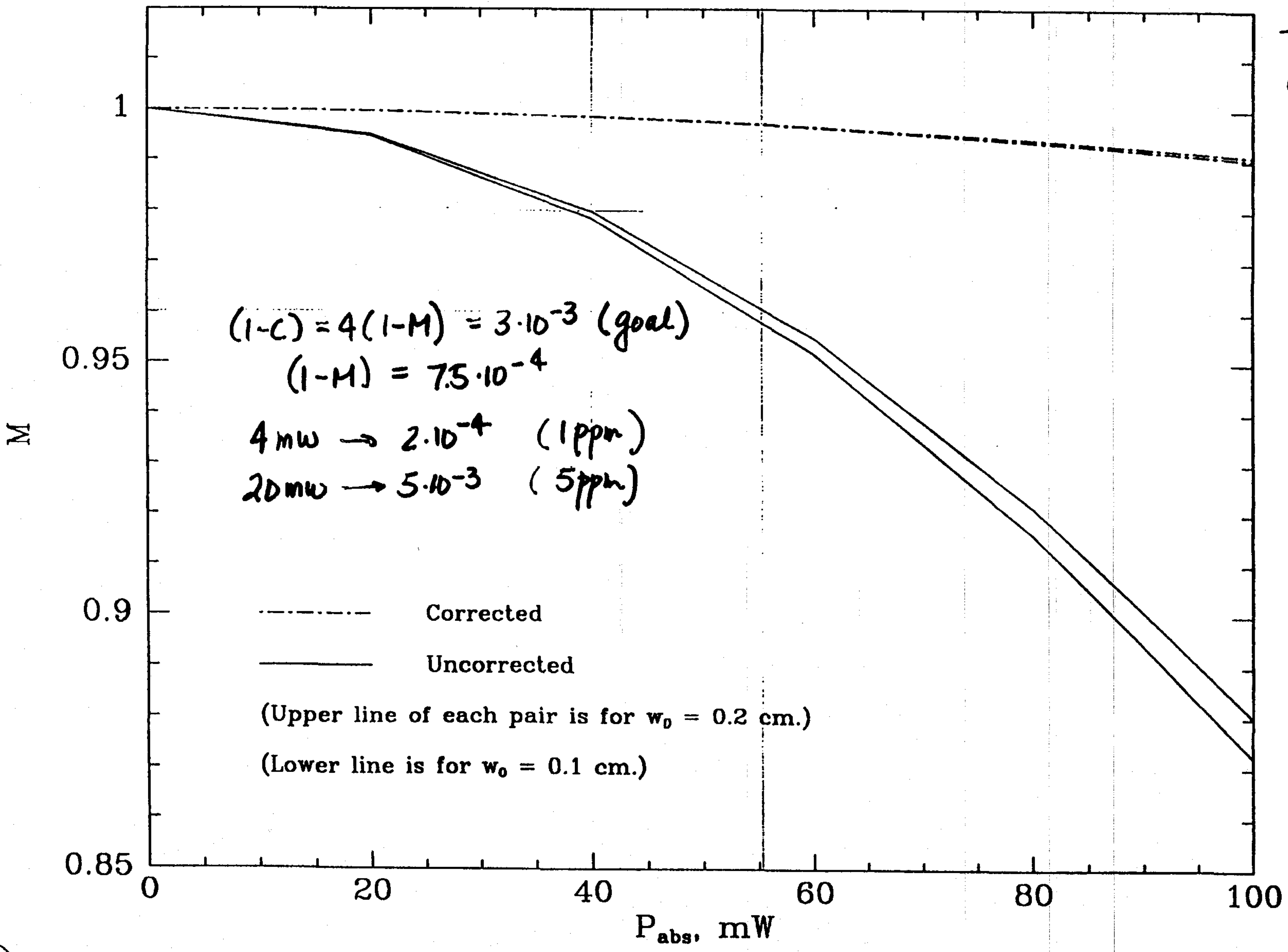
Mode Matching Efficiency vs. Absorbed Power



Mode Matching Efficiency vs. Absorbed Power

Fig.

Torrey
Lyons

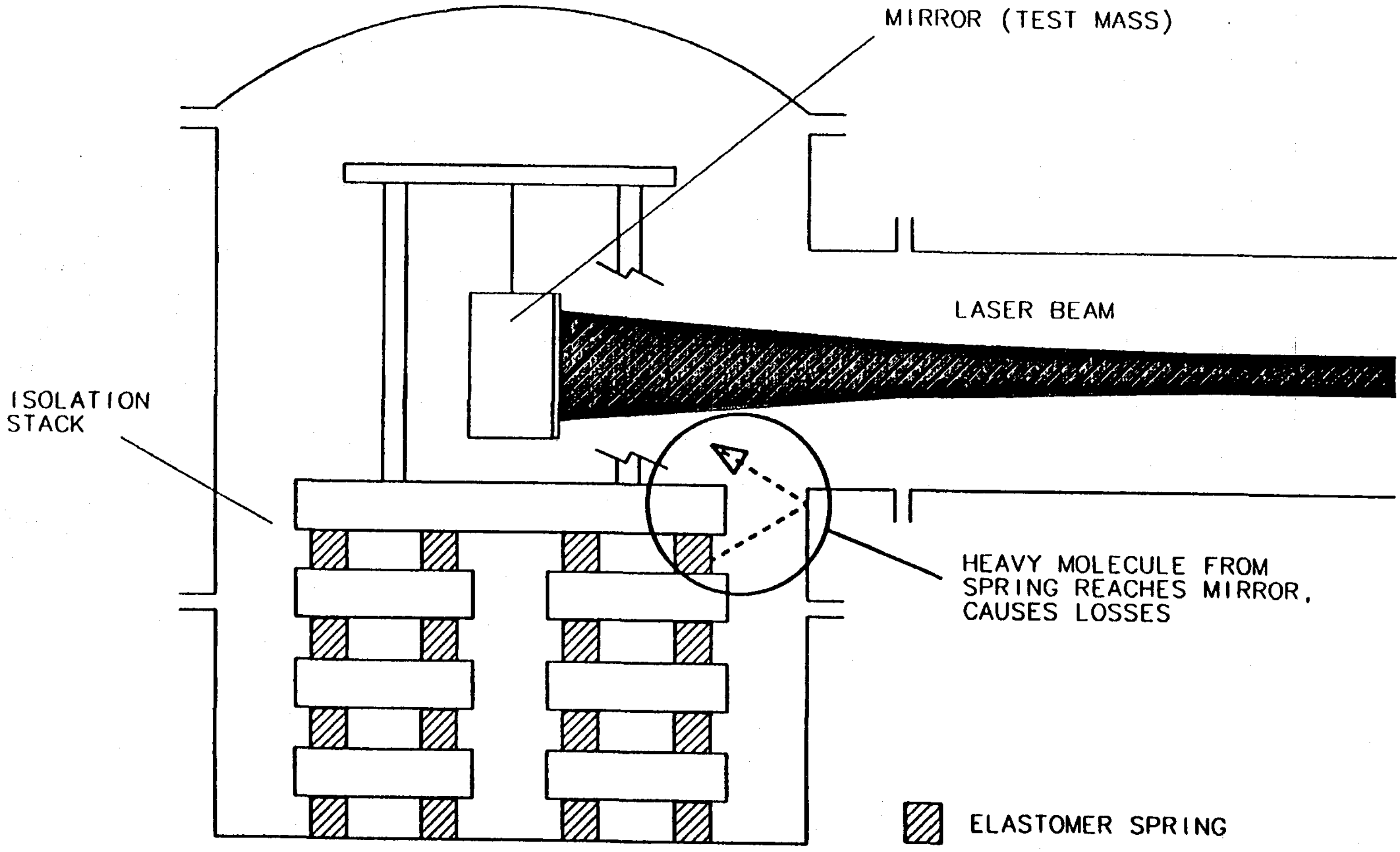


$(1-C) = 4(1-M) = 3 \cdot 10^{-3}$ (goal)

$(1-M) = 7.5 \cdot 10^{-4}$

4 mw → $2 \cdot 10^{-4}$ (1 ppm)

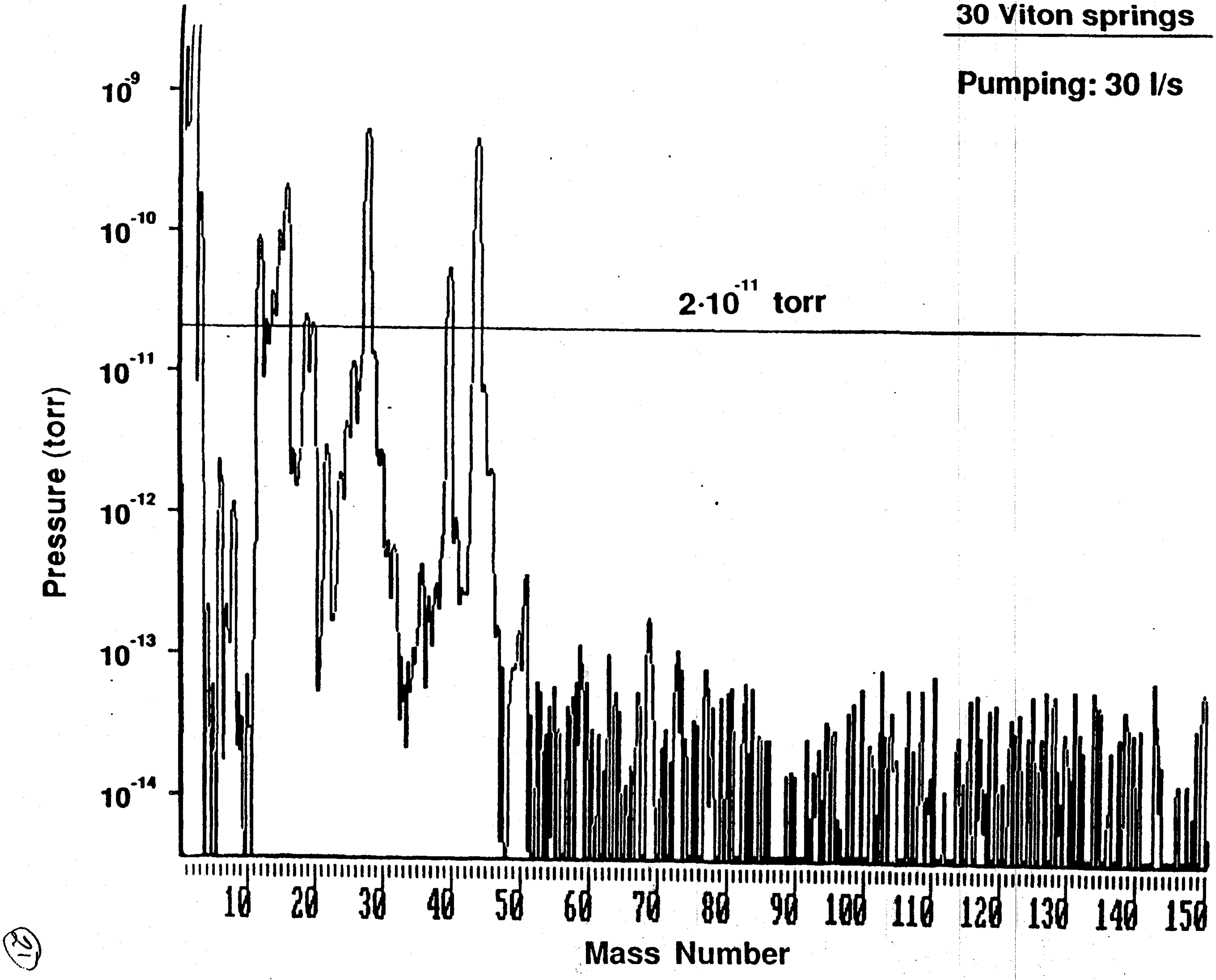
20 mw → $5 \cdot 10^{-3}$ (5 ppm)



Requirement: Sum of pressures from M=41,43,53,55,57 $<2*10^{-11}$ torr

(this gives $\lesssim$ 1 monolayer of deposited molecules on mirrors)

Requirement based on lab data and calculations. It is still undergoing refinement, and is believed to include a safety margin of one order of magnitude.



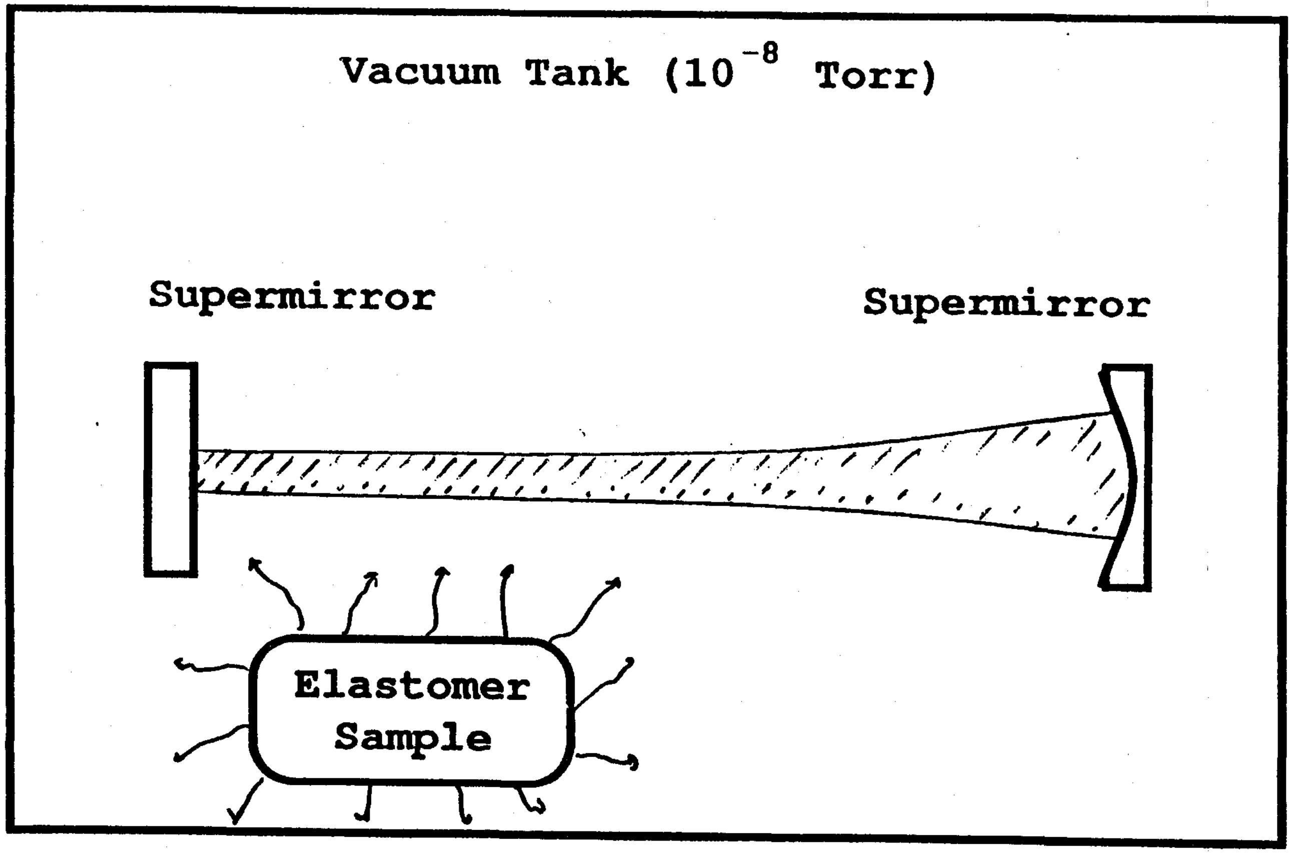
30 Viton springs

Pumping: 30 l/s

12

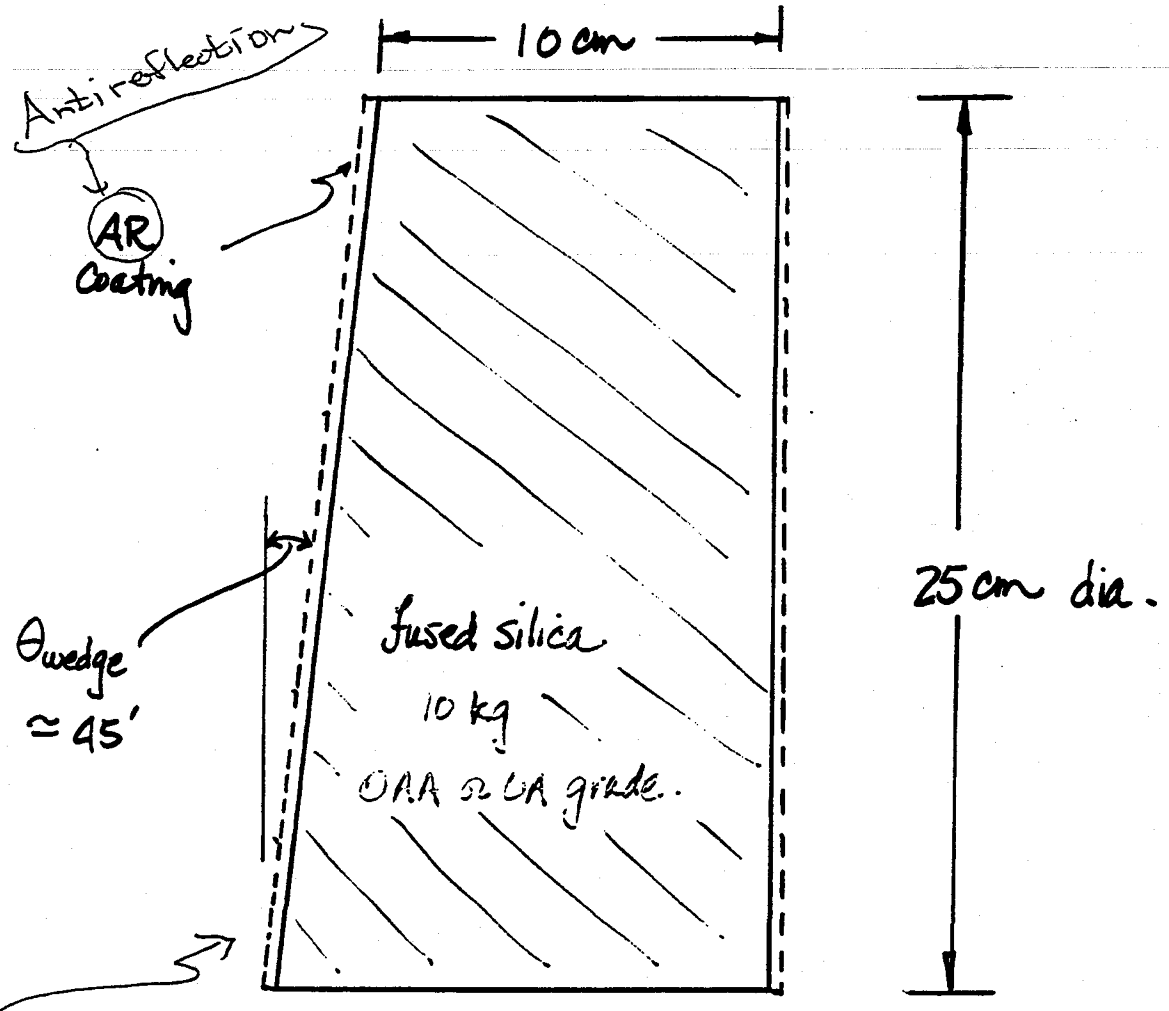
Material qualification Tests.

METHOD: Measure optical losses from storage time of light between two mirrors.



LIGO Test Mass Optics

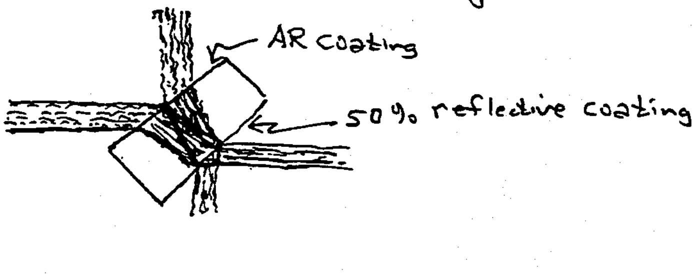
Mass = 10 kg



very high purity. → low scattering + absorption.

very low coefficient of thermal expansion

Used both for testmass/mirrors and for beam splitter. Beam-splitter configuration:



a few ppm per cm losses in substrate

Pathfinder Goals

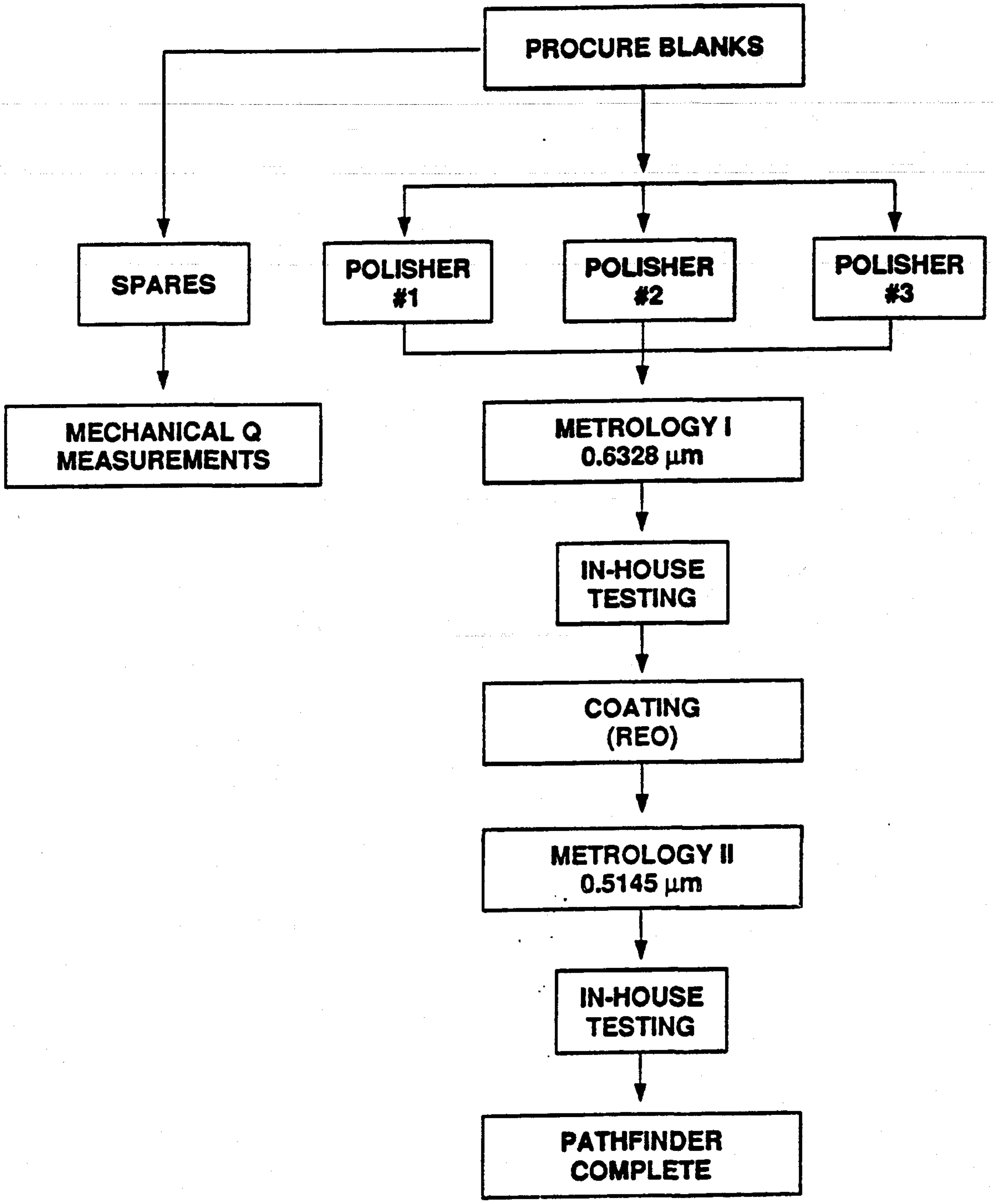
Primary

- Specify performance requirements of large aperature optics.
- Evaluate industrial polishing, coating and measuring capabilities.
- Scope costs for fabrication of the LIGO optics.

Secondary

- Develop and determine the actual processes and techniques for fabrication and mensuration of LIGO optics.

PATHFINDER PROCESS



PROCURE BLANKS

SPARES

POLISHER #1

POLISHER #2

POLISHER #3

MECHANICAL Q MEASUREMENTS

METROLOGY I 0.6328 μm

IN-HOUSE TESTING

COATING (REO)

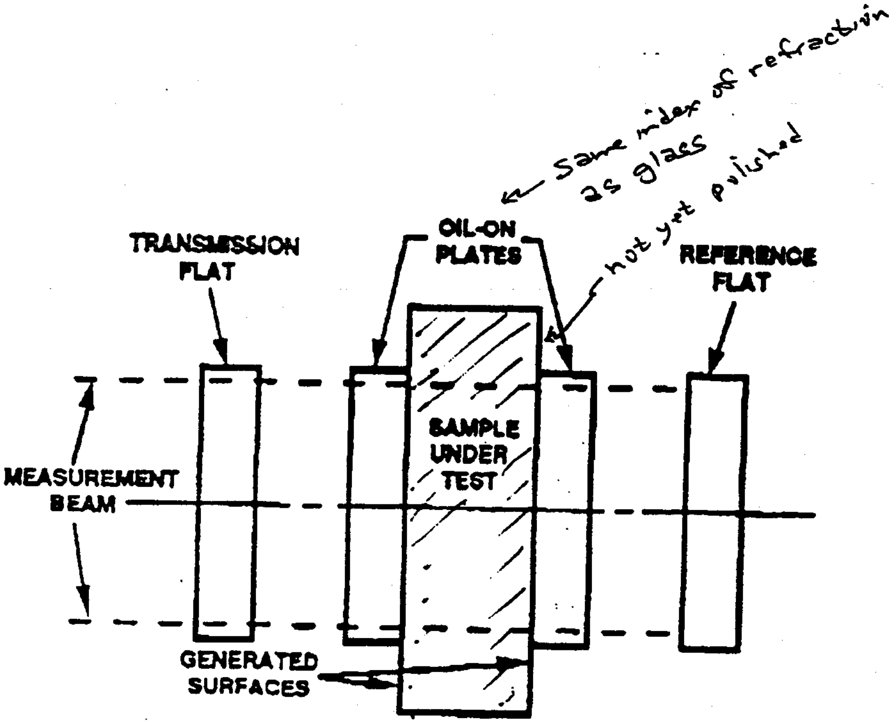
METROLOGY II 0.5145 μm

IN-HOUSE TESTING

PATHFINDER COMPLETE

LIGO

Substrate Homogeniety



$$OPL = 2 \cdot n \cdot t$$

$$OPD = \Delta OPL = 2 \cdot n \cdot \Delta t + 2 \cdot t \cdot \Delta n$$

remove $2 \cdot n \cdot \Delta t$

$$OPD = 2 \cdot t \cdot \Delta n \qquad \Delta n = \frac{OPD}{2 \cdot t} = \text{line-averaged index variation}$$

price per substrate

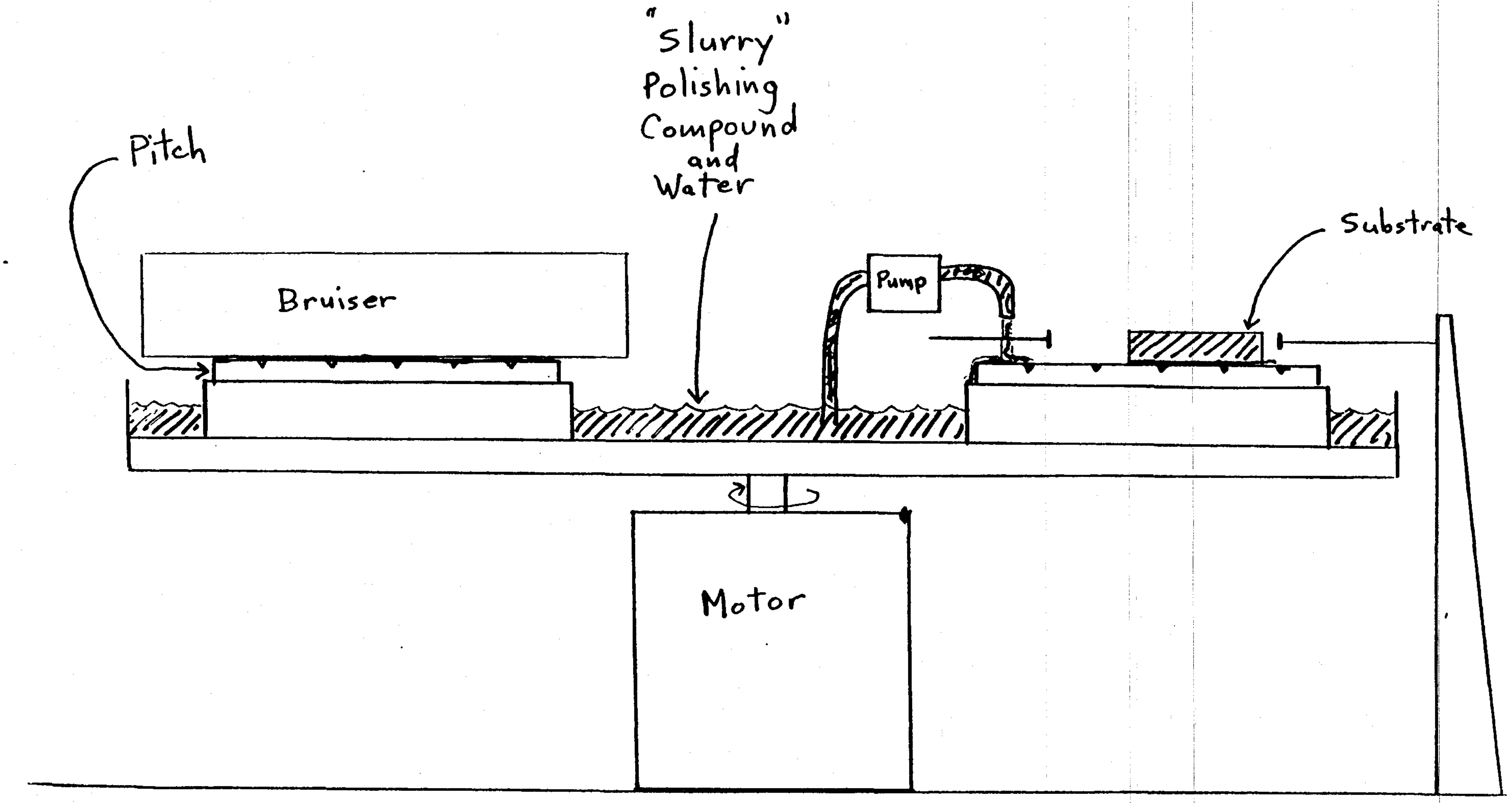
\$12k → OA grade $\Rightarrow \Delta n \leq 1 \cdot 10^{-6}$

\$18k → OAA grade $\Rightarrow \Delta n \leq 5 \cdot 10^{-7}$

for $2 \cdot t = 20$ cm, $\lambda = 514$ nm

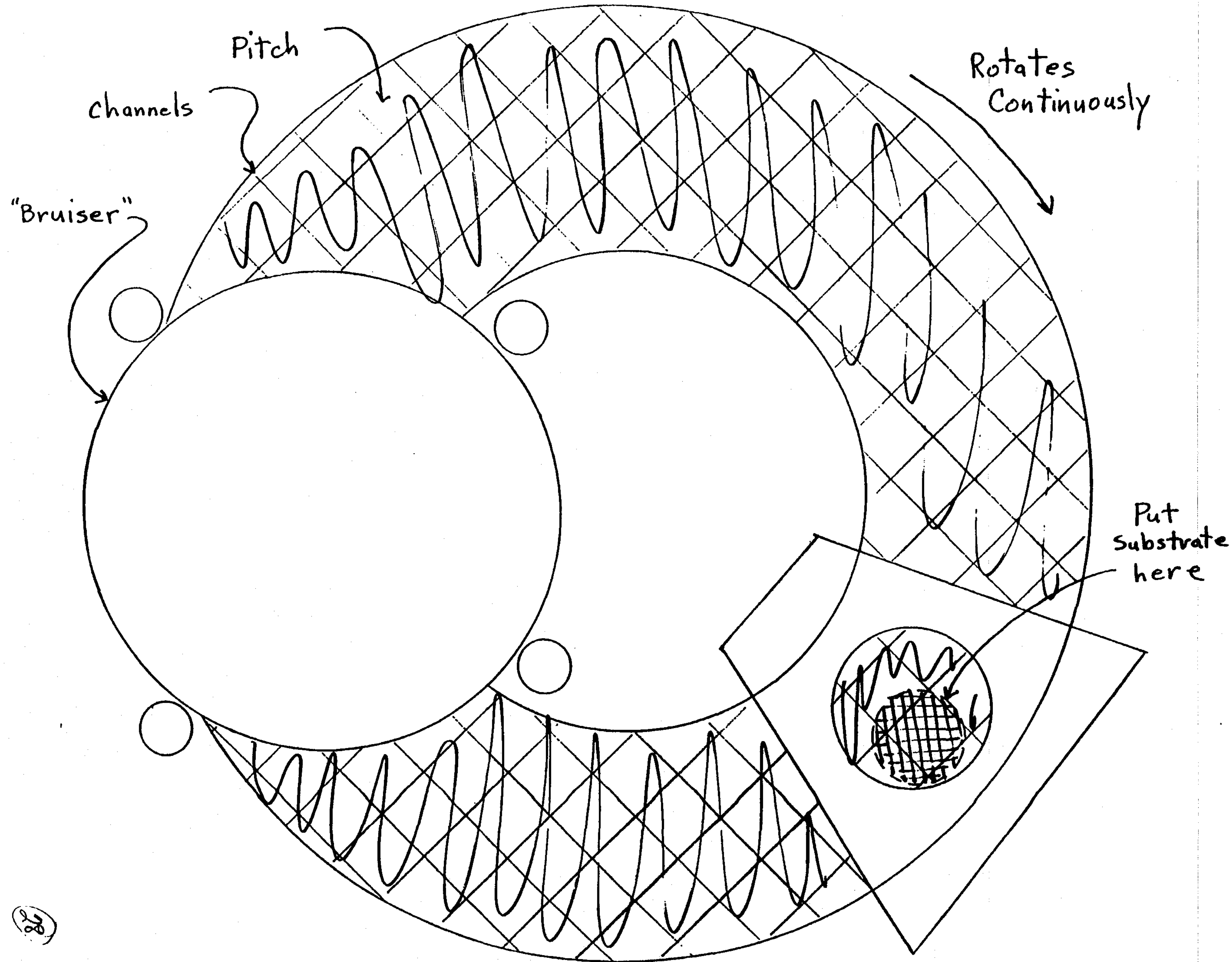
$2 \cdot t \cong 4 \cdot 10^{5} \lambda$ OAA grade $\Rightarrow OPD \leq \frac{\lambda}{5}$

Continuous Polisher



superpolish ⇒ $\sigma_{>125 cm^{-1}} \leq$ few Å

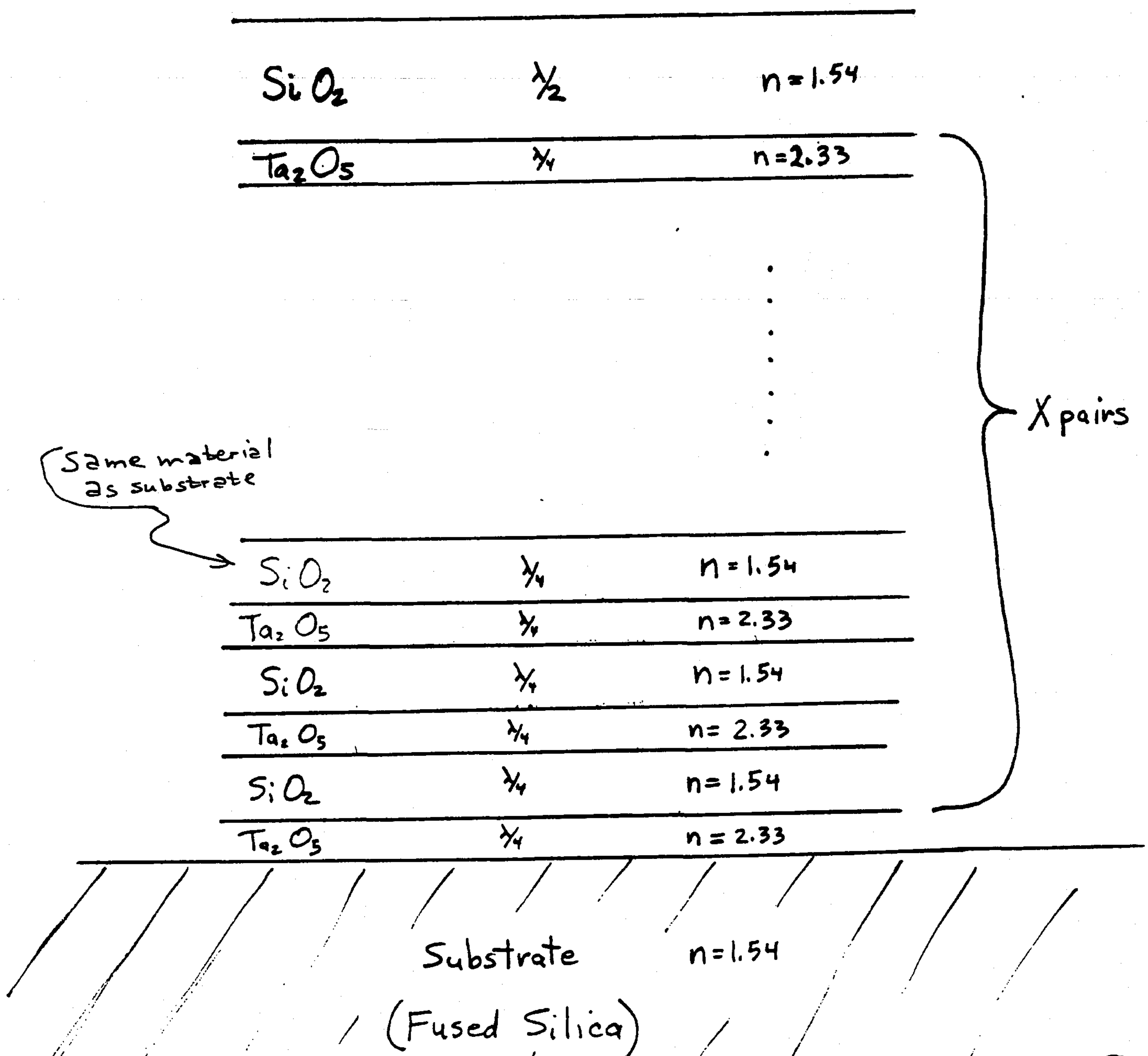
$<$ .5 Å possible losses $\leq$ 10 ppm


Pitch
Channels
"Bruiser"
Rotates Continuously
Put substrate here

High Reflector Coating Design

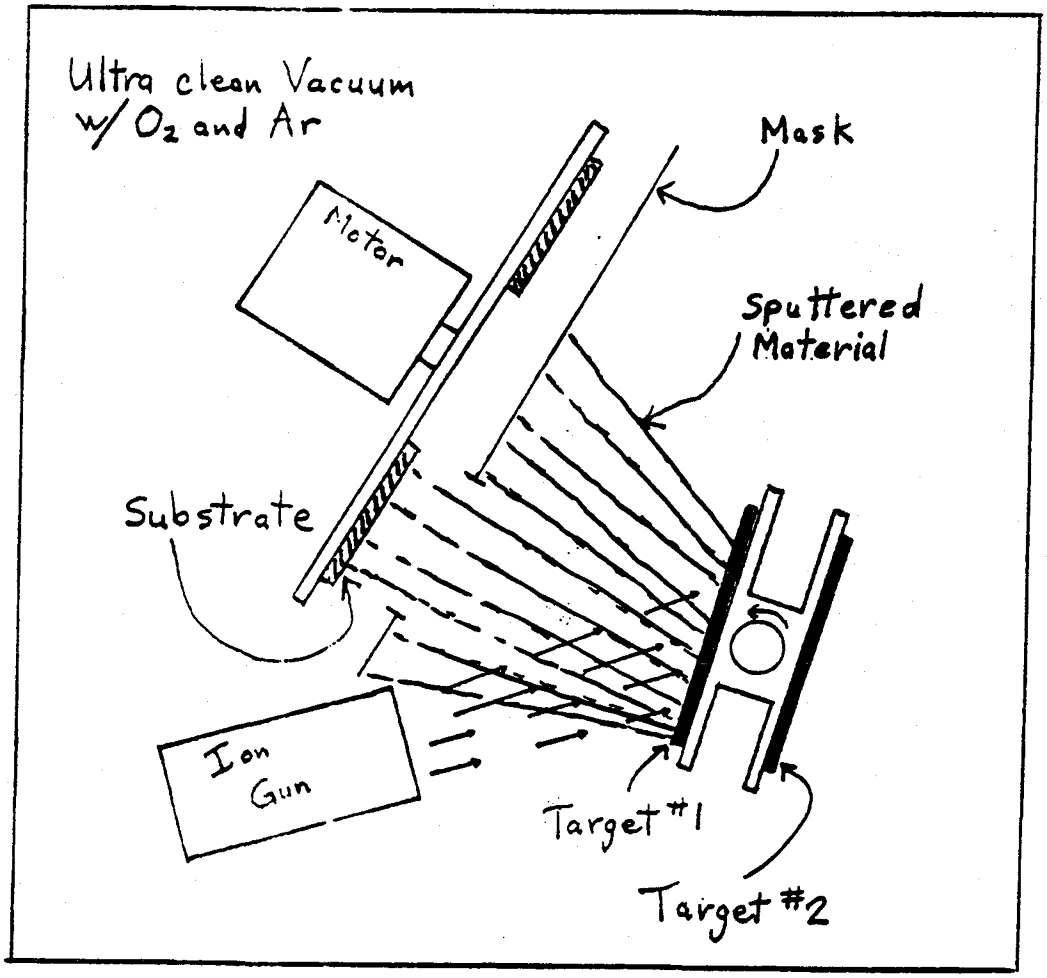
$S{:}H(LH)^{x}LL$

(Coated by ion beam sputtering)



SiO_2	$\lambda/2$	n=1.54
Ta_2O_5	$\lambda/4$	n=2.33
⋮		
SiO_2	$\lambda/4$	n=1.54
Ta_2O_5	$\lambda/4$	n=2.33
SiO_2	$\lambda/4$	n=1.54
Ta_2O_5	$\lambda/4$	n=2.33
SiO_2	$\lambda/4$	n=1.54
Ta_2O_5	$\lambda/4$	n = 2.33
Substrate (Fused Silica)		n=1.54

Ion Beam Sputtering
Coating Chamber



Target #1 SiO_2

Target #2 Ta_2O_5

uniformity > $\lambda/100$

In-House Measurements.

1) Bulk absorption
2) Bulk scatter
3) Transmission maps
4) Ringdown maps
5) Scatter Map
6) Surface absorption
7) Mechanical Q

LIGO optics lab - 058 West Bridge

BATCH START

⑩ Control Systems

STAPLE
OR
DIVIDER

Lecture 10
Control Systems for Test-Mass Position and Orientation

by Seiji Kawamura, 29 April 1994

Kawamura lectured from the following transparencies. Kip has annotated them, based on Kawamura's lecture.

Ph 103c: The Physics of LIGO 29 April 1994

LECTURE 10.
Control Systems for Test-Mass Position and Orientation
Lecture by Seiji Kawamura

Assigned Reading:

BB. S. Kawamura and M. E. Zucker, *Applied Optics*, in press. [This paper explains the influence of angular mirror orientation errors on the length of a Fabry-Perot resonator.]

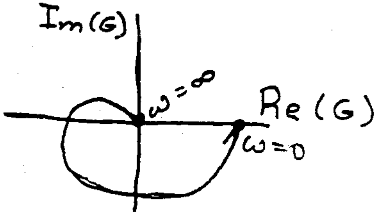
Read either item CC. below or item DD. [Item DD. is highly recommended, since feedback loops are so important; but for some students it may entail a fair amount of work, and CC. might be preferred.]

CC. M. Stephens, P. Saulson, and J. Kovalik, "A double pendulum vibration isolation system for a laser interferometric gravitational wave antenna," *Rev. Sci. Instrum.*, **62**, 924–932 (1991). [Here you are asked to focus on the control of the pendulum, rather than on the penedulum's role in vibration isolation.]

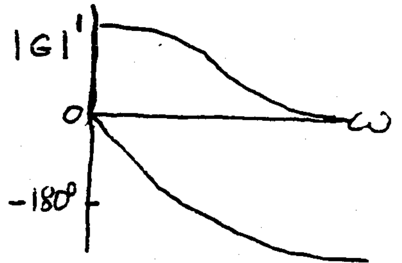
DD. Read, in your favorite control theory book [e.g., R. C. Dorf, *Modern Control Systems* 5th editon (Addison-Wesley, 1989), cited as *Dorf* below] or elsewhere, about the following issues:

a. The relationship of Laplace transforms to Fourier transforms [e.g., *Dorf* pp. 264–266]. Control theory is often formulated in terms of Lapace transforms rather than Fourier transforms because Lapace transforms are more naturally suited to describing the transient response of a system to some input; the reason is that they entail only the behavior of the system between some initial time $t = 0$ and $t = \infty$, by contrast with Fourier transforms which involve the behavior over all time. In this course we will probably *not* deal with any issues where the Laplace transform has an advantage; and we will most always discuss things in terms of Fourier transforms and thus in terms of the response of a system at some frequency ω. However, in order to read control theory books on these issues, it is necessary to understand Laplace transforms and their relation to Fourier transforms. [Note that, although theoretical physicists normally use the form $e^{-i\omega t}$ for the time dependence of a Fourier component of frequency ω, engineers, and control theorists normally use $e^{+j\omega t}$ (where $i = j = \sqrt{-1}$). In this course we shall use the engineers' conventions.]

b. The use of complex frequency-response plots to describe the ratio of the output amplitude $V_{\rm out}$ of a linear system such as a control loop, to its input amplitude $V_{\rm in}$, when the input and output have frequency ω [e.g., read *Dorf*, pp. 266–283]. In these plots, $V_{\rm out}/V_{\rm in} \equiv G(\omega)$, which is a complex quantity, is plotted as a curve in the complex plane parametrized by ω, for real ω. Such a plot contains the same information as a Bode diagram, in which one gives two plots, one of $|G(\omega)|$ plotted upward and ω horizontally; the other of the phase $\phi(\omega)$ of G plotted upward and ω horizontally; for example:

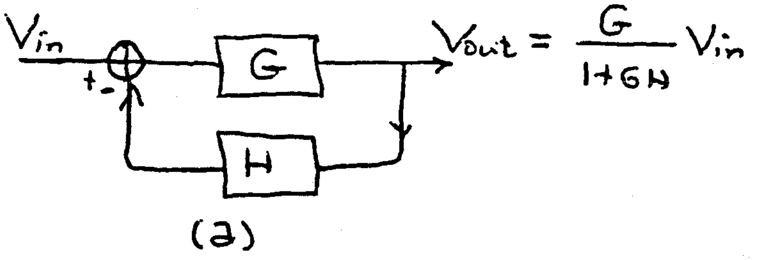


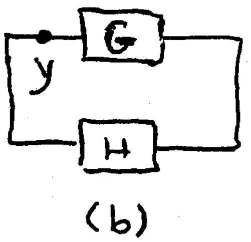
Frequency-Response Plot



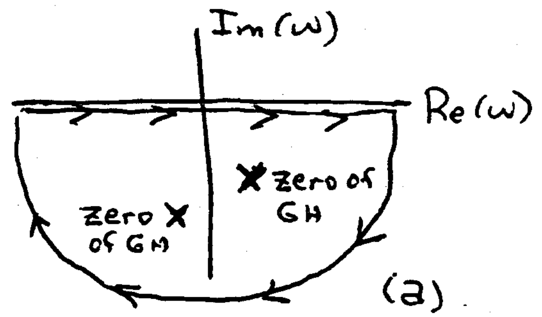
Bode Diagram

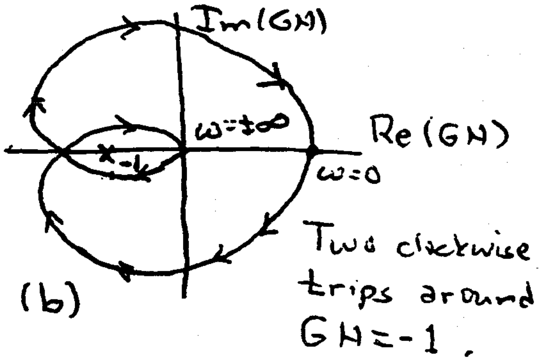
c. The Nyquist criterion for the stability of a control loop [e.g., read *Dorf*, pp. 309–333]. [The Nyquist criterion, in a nutshell, is this: Consider a simple feedback loop of form shown in (a) below. If the input and output ports are shut, the resulting closed loop shown in (b) can oscillate at certain complex eigenfrequencies without any stimulus. Those frequencies are easily deduced from the requirement that the amplitude y at the indicated point must satisfy $y = G(\omega)H(\omega)y$, and therefore $y(1 + GH) = 0$, and therefore *the loop's frequencies of self oscillation are the zeroes of* $1 + G(\omega)H(\omega)$.





Since the time dependence of these oscillations is $e^{+j\omega t}$, if there are any zeroes of $1 + GH$ in the *lower-half* complex frequency plane (any eigenfrequencies ω with negative imaginary parts), then the amplitude of the closed loop's oscillations will grow in time; i.e., the closed loop will be unstable. The number of zeroes in the lower-half frequency plane can be inferred from the Cauchy theorem of complex variable theory: Construct the curve $G(\omega)H(\omega)$ in the complex plane, with ω running along the real axis from $-\infty$ to $+\infty$, and then swinging down around the lower half frequency plane and back to $-\infty$; see drawing (a) below. The number of times that this curve, $G(\omega)H(\omega)$ encircles clockwise the point $GH = -1$ (on the real axis) is the number of zeroes of $1 + GH$ minus the number of poles of $1 + GH$; see drawing (b) below. For feedback loops there usually are no poles of $1 + GH$ [such a pole would give precisely zero output/input in the feedback loop of (a) above], so usually the number of clockwise trips around $GH = -1$ is the number of zeroes in the complex frequency plane. Thus, if there are no clockwise trips, the closed loop is stable; if there are some, it is unstable. This is the Nyquist criterion for stability.]



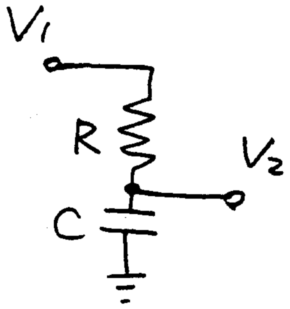


Suggested Supplementary Reading:

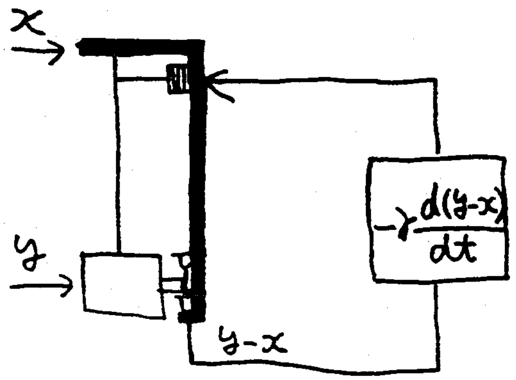
5. Read whichever of items 3. and 4. you did not do as "assigned reading".

A Few Suggested Problems

1. Use the Nyquist criterion for the stability of a feedback loop to show that, when the Bode diagram has the qualitative form shown on transparency 23 of Kawamura's lecture (where $f = \omega/2\pi$), then the loop is stable if the phase of GH at the unity gain point is $\phi > -180^o$, and unstable if $\phi < -180^o$. [Hint: show that, because in the time domain the equations describing most any servo loop are real, when ω is real then $G(-\omega)H(-\omega)$ is the complex conjugate of $G(+\omega)H(+\omega)$. This permits you to construct the Nyquist curve in the frequency-response plot for both positive and negative ω from Kawamura's positive-frequency Bode diagram.] For what shapes of Bode diagrams will this $\phi > -180^o$ stability criterion remain true?(Consider, for example, the issue of how many unity gain points there are).
2. Construct a complex frequency-response curve and also a Bode diagram for the following pass $R-C$ circuit. From the Bode diagram infer that this circuit is a low-pass filter.



3. In his lecture [transparencies numbered 15–17], Kawamura described the damping of the swing of a pendulum via a feedblack loop that produces a displacement $\delta x = -\gamma dy/dt$ of the pendulum's support point, where γ is the damping constant and y is the horizontal position of the pendulum's mass. Of course, in order to implement this, one needs some fixed object with respect to which y is measured. In transparency 15 that object is the shadow sensor, but nothing is said about what that sensor is attached to. A practical approach is to attach the sensor to the pendulum's support point, as shown below. Then the feedback displacement is $\delta x = -\gamma d(y-x)/dt$, where x is the instantaneous horizontal position of the support point. Repeat Kawamura's analysis [transparencies 15–17]] for this feedback system.



4. Suppose that one were to try to damp the (low-frequency, 1 Hz) swing of the pendulum in problem 3 not with a feedback displacement $\delta x = -\gamma d(y - x)/dt$, but instead with a feedback displacement that is $-ay$ (for some constant $a > 1$) at low frequencies (near 1 Hz) but that shuts off at higher frequencies (above 10 Hz), where the gravity waves are to be measured. Suppose one implements this feedback displacement by simply passing a voltage, proportional to y, through a low-pass $R-C$ filter of the sort discussed in problem 2. Show that the resulting damping system will be unstable.

5. Derive the relation $\delta l = d_1 \delta\theta_1 + d_2 \delta\theta_2$ on transparency 28 of Kawamura's lecture.

LECTURE 10

CONTROL SYSTEMS FOR TEST-MASS POSITION AND ORIENTATION

Seiji Kawamura

APR. 29, 1994

You will learn ...

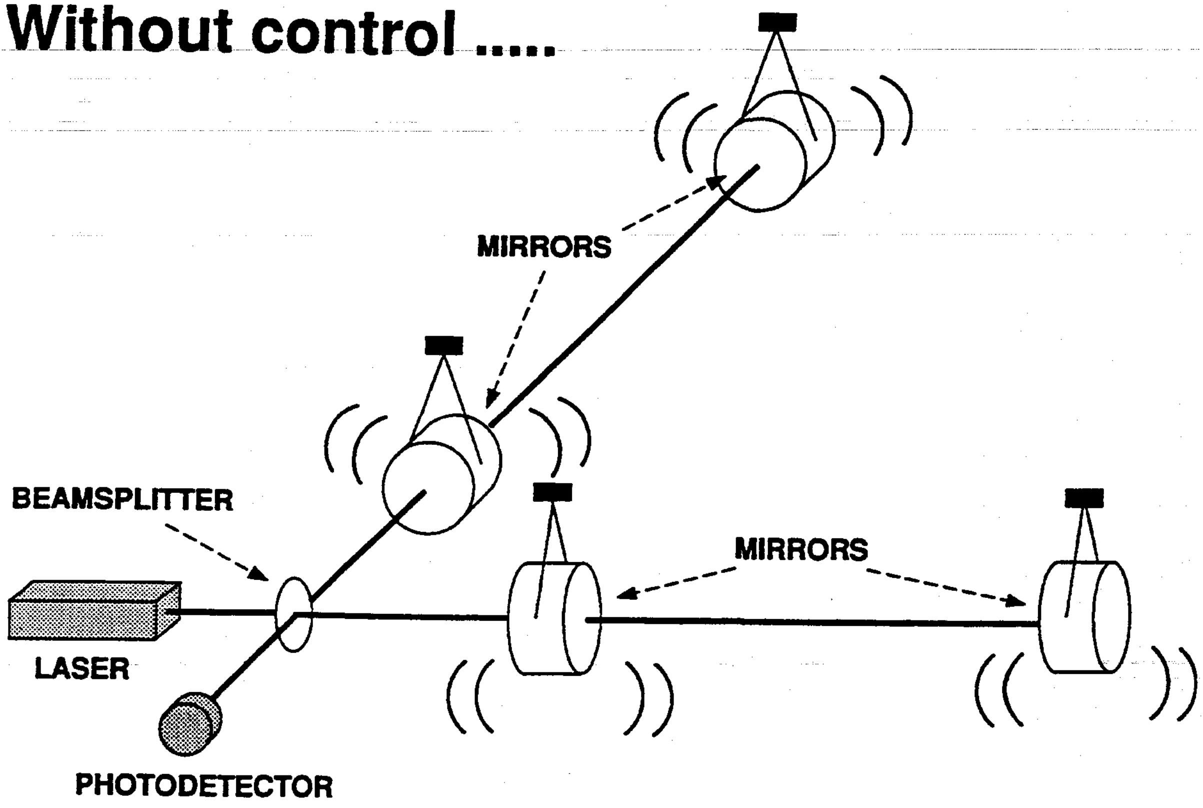
1. What is test mass position / orientation control ?

2. How to damp a test mass without adding extra noise ?

3. How to predict test mass orientation noise in a Fabry-Perot cavity ?

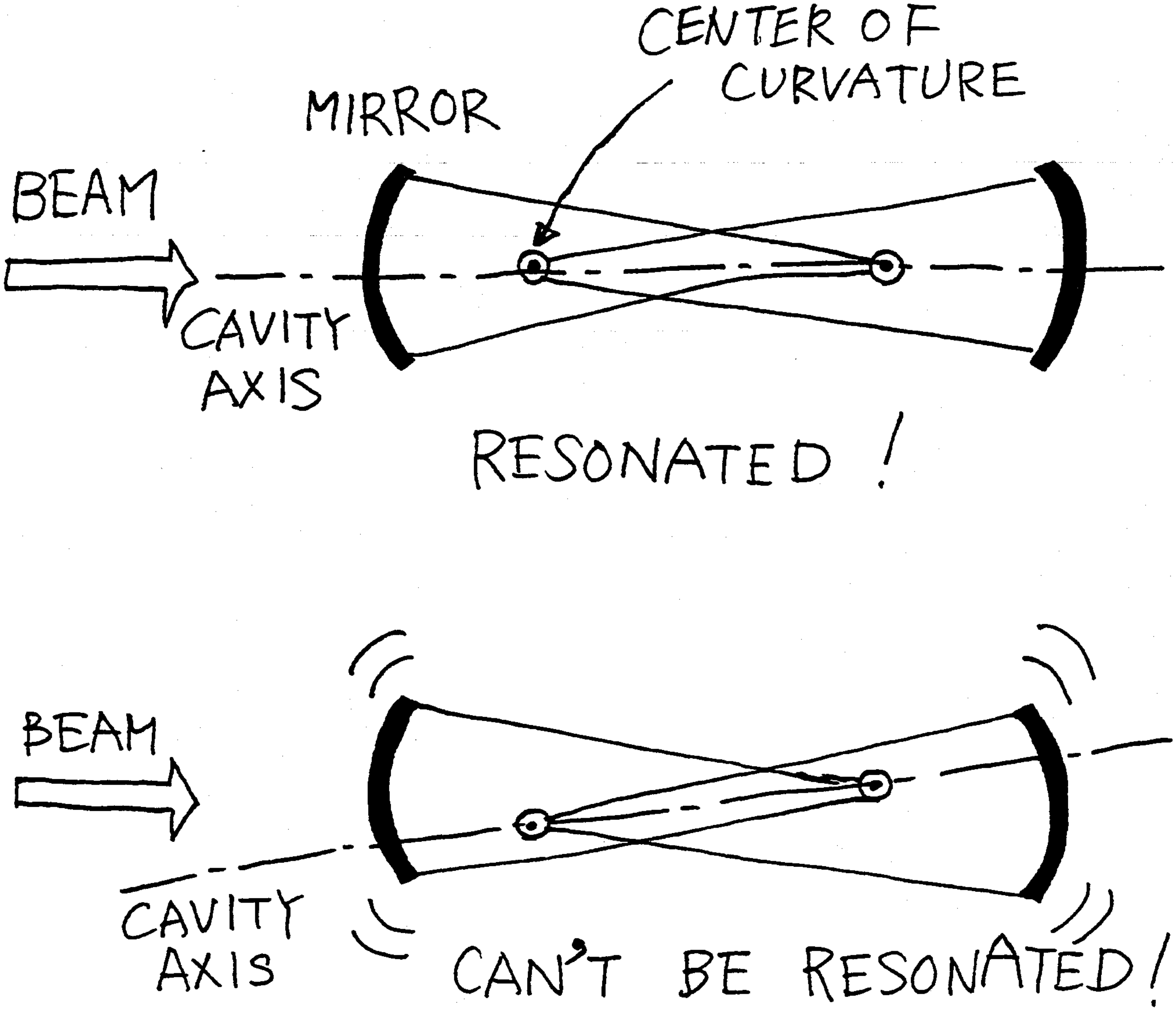
1. What is test mass position/orientation control?

Without control .....



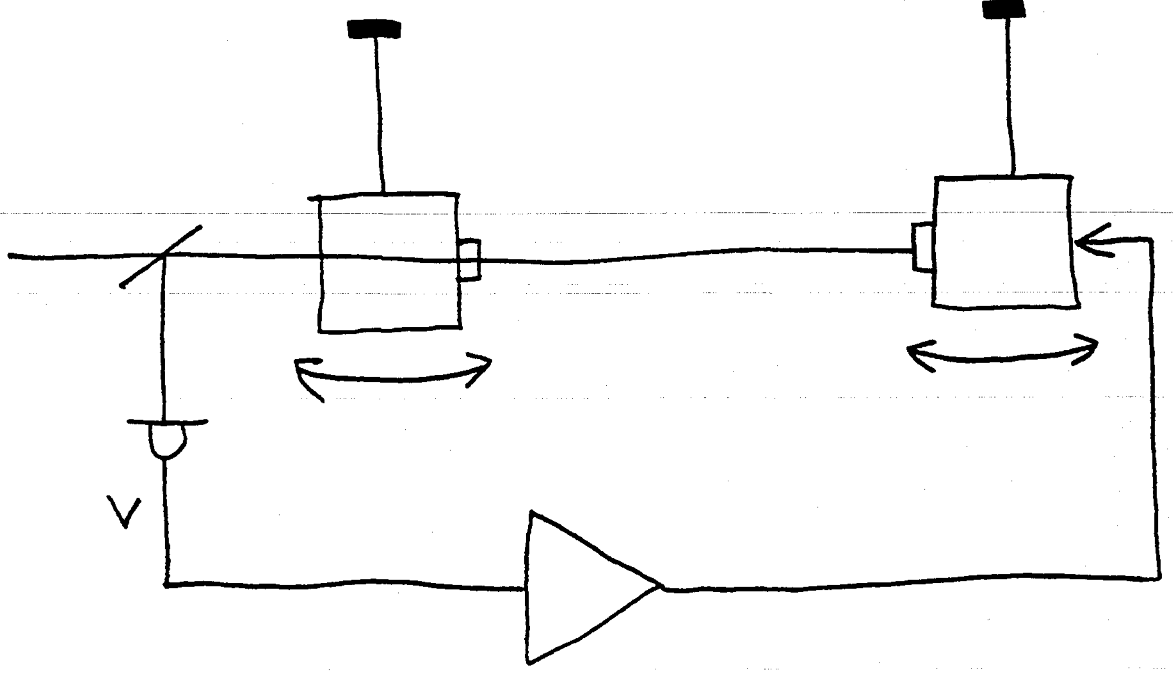
It doesn't work !

TEST MASS POSITION / ORIENTATION CONTROL SYSTEM

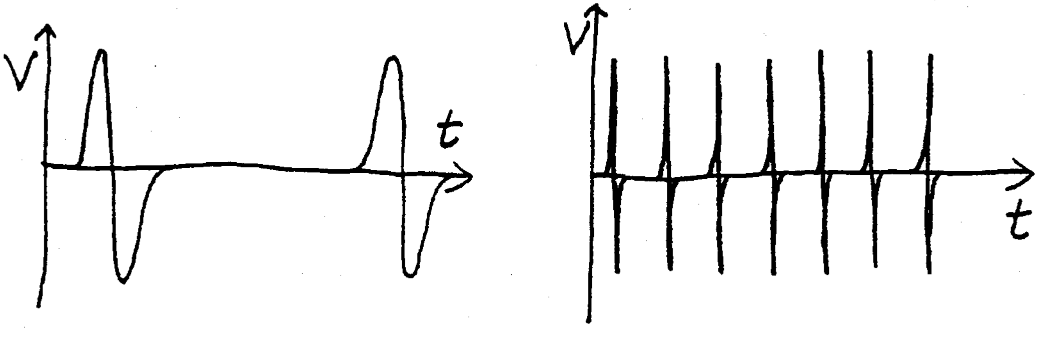


ORIENTATION CONTROL

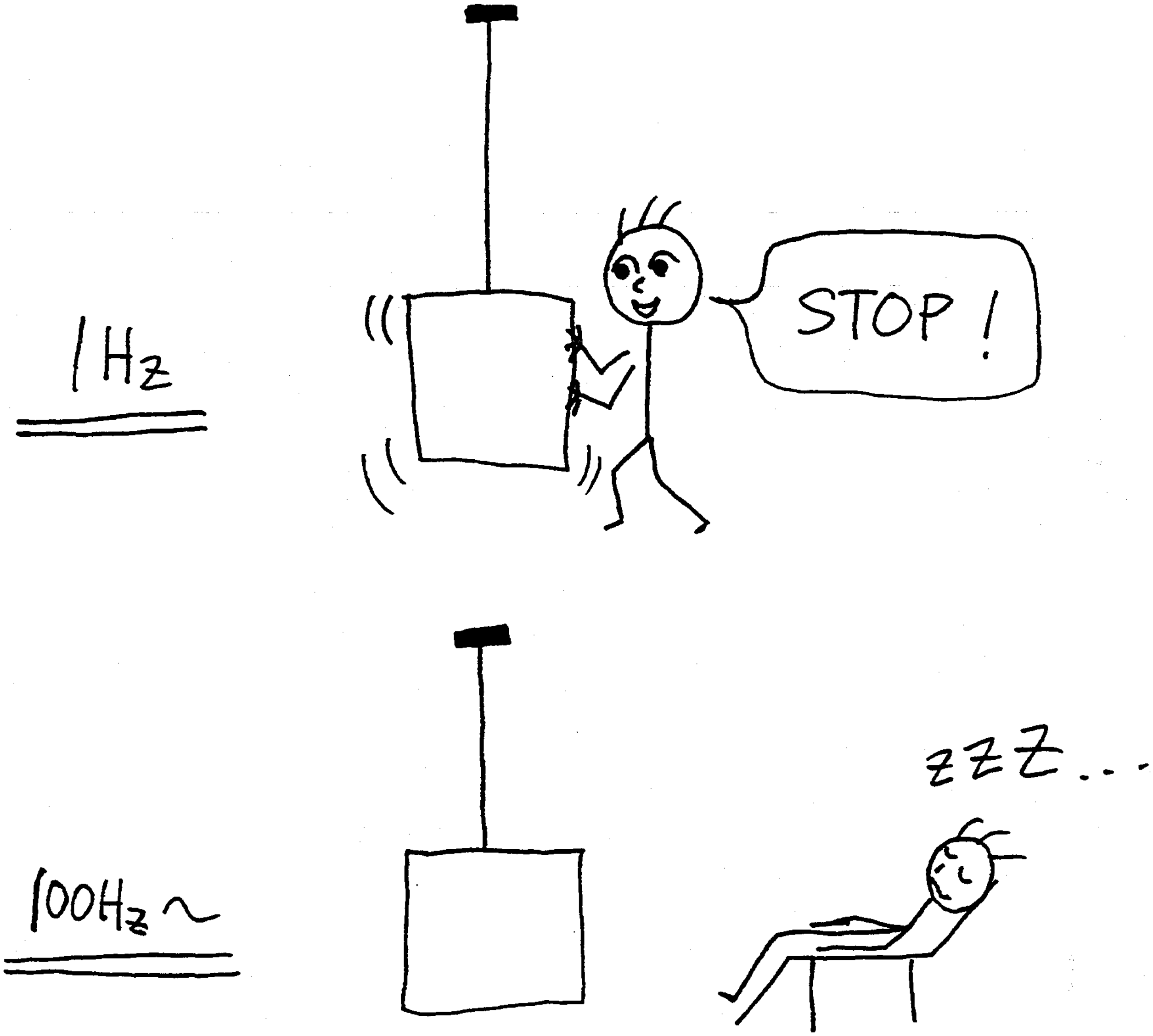
POSITION CONTROL



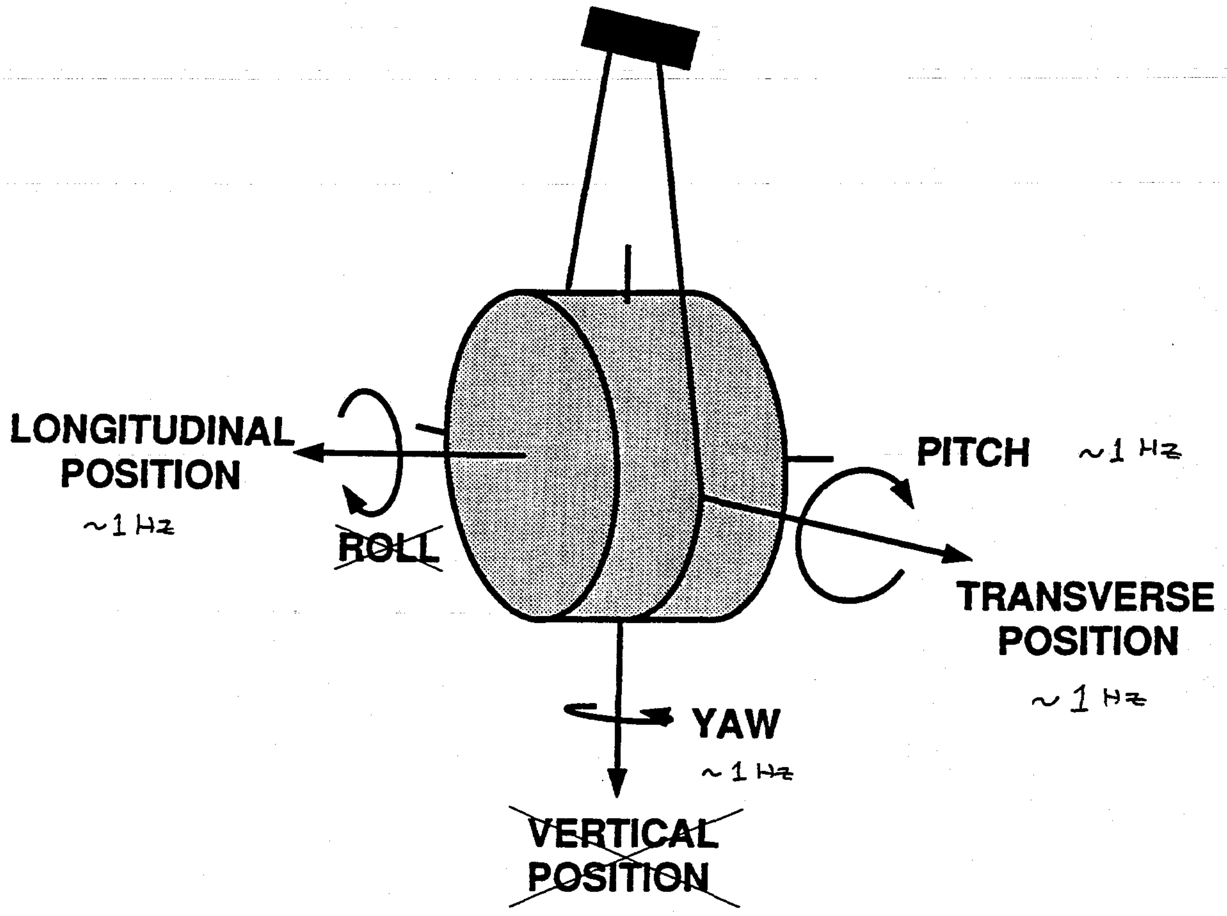
MOTION SMALL LARGE



LOCK EASY DIFFICULT

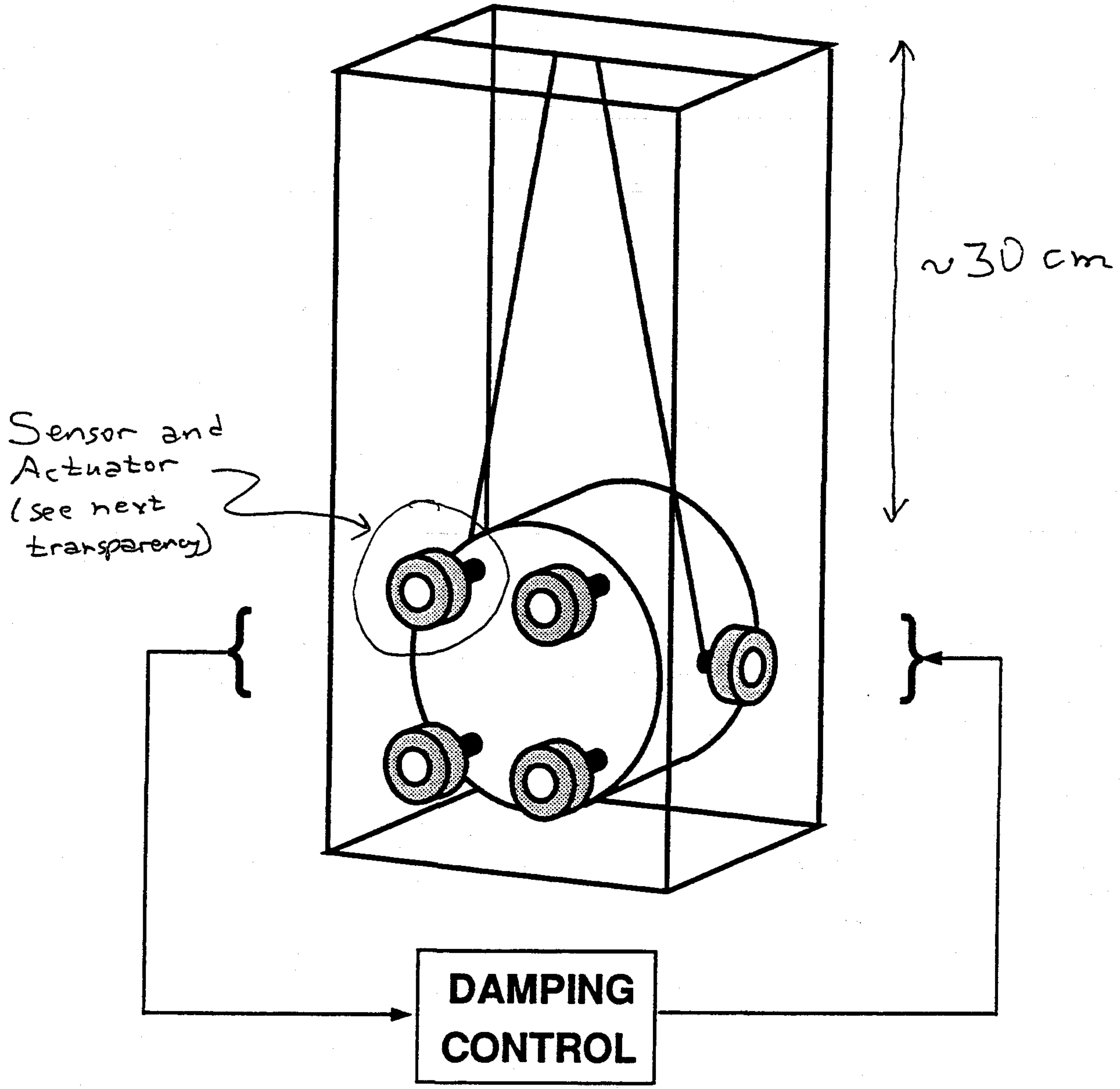


FREQUENCY DEPENDENCE OF CONTROL

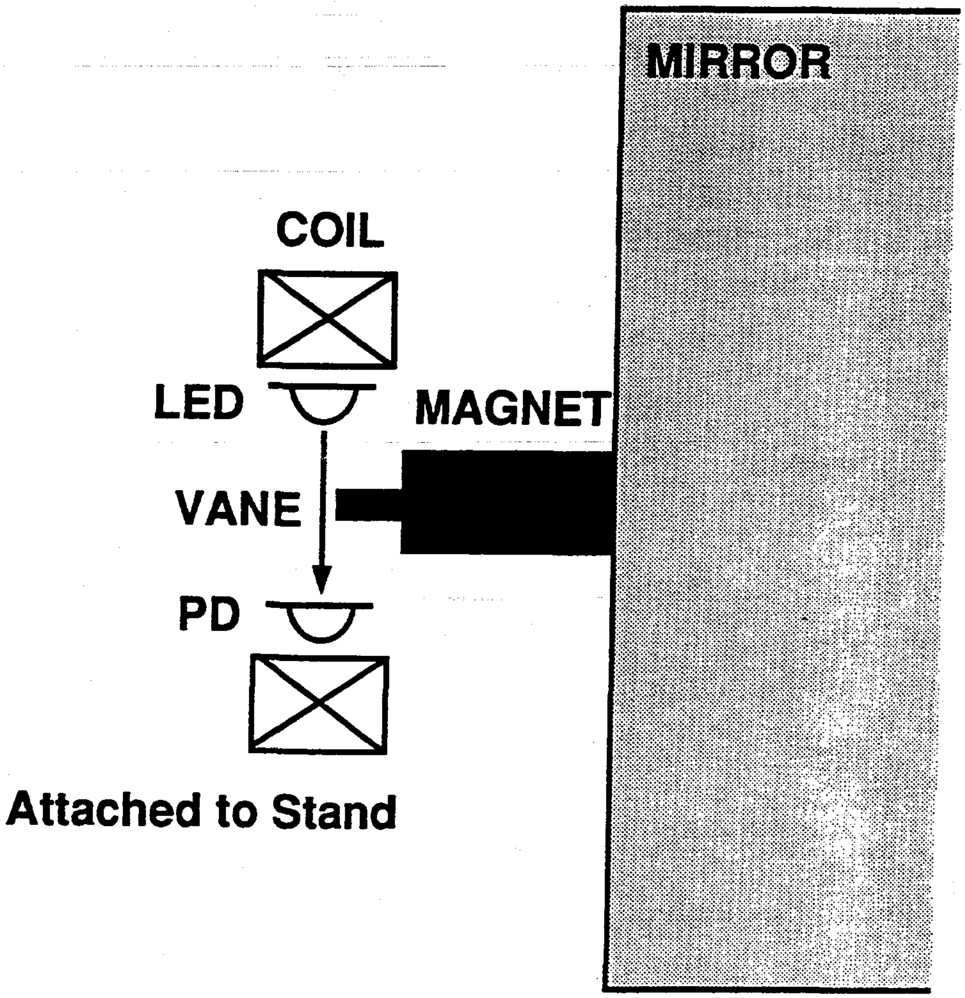


POSITION AND ORIENTATION

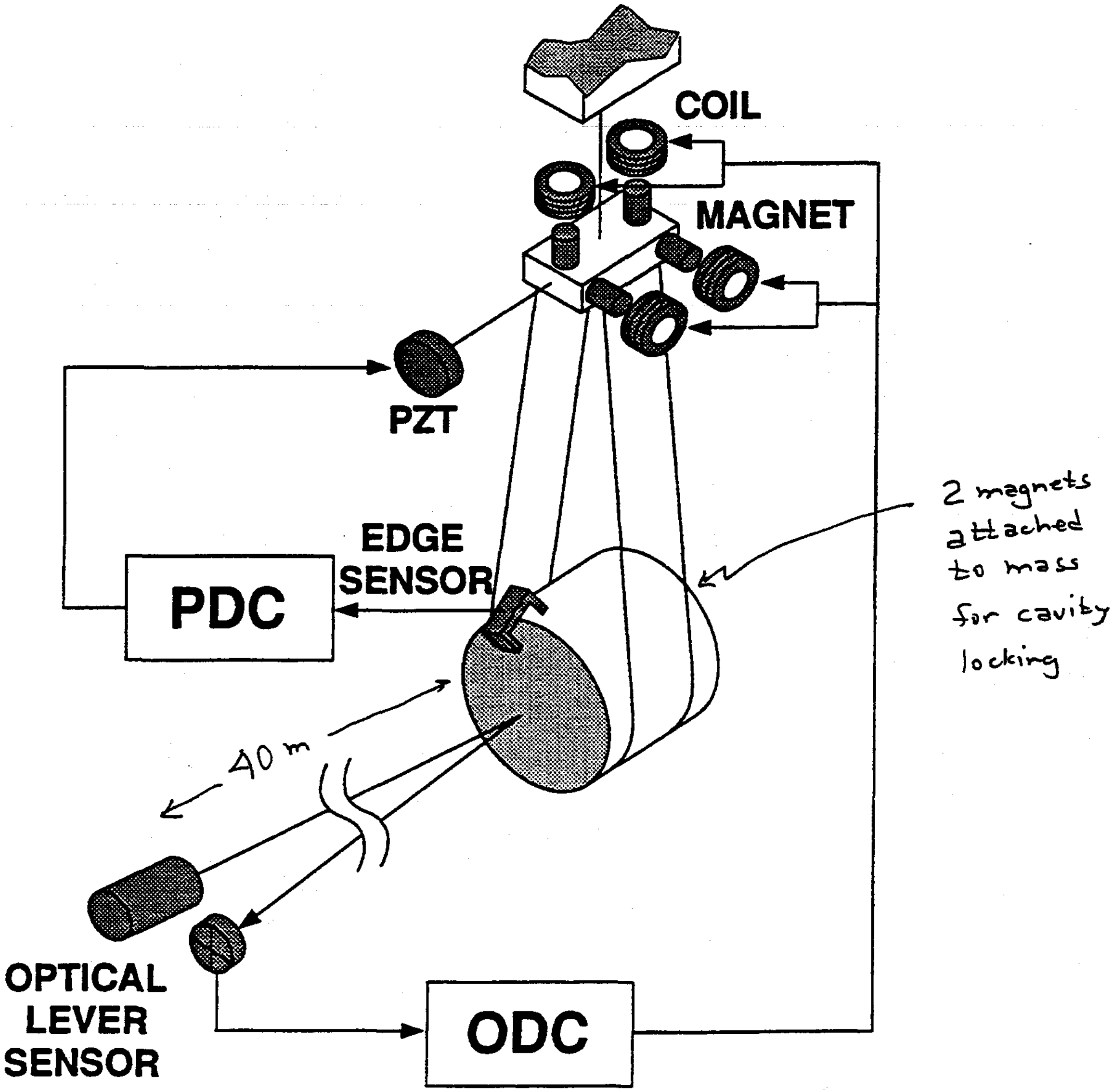
Motions that Must Be Controlled



TEST MASS DAMPING CONTROL SYSTEM 1

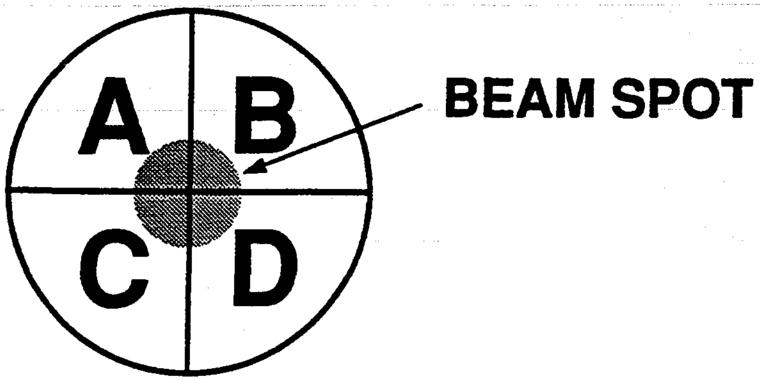


SENSOR AND ACTUATOR



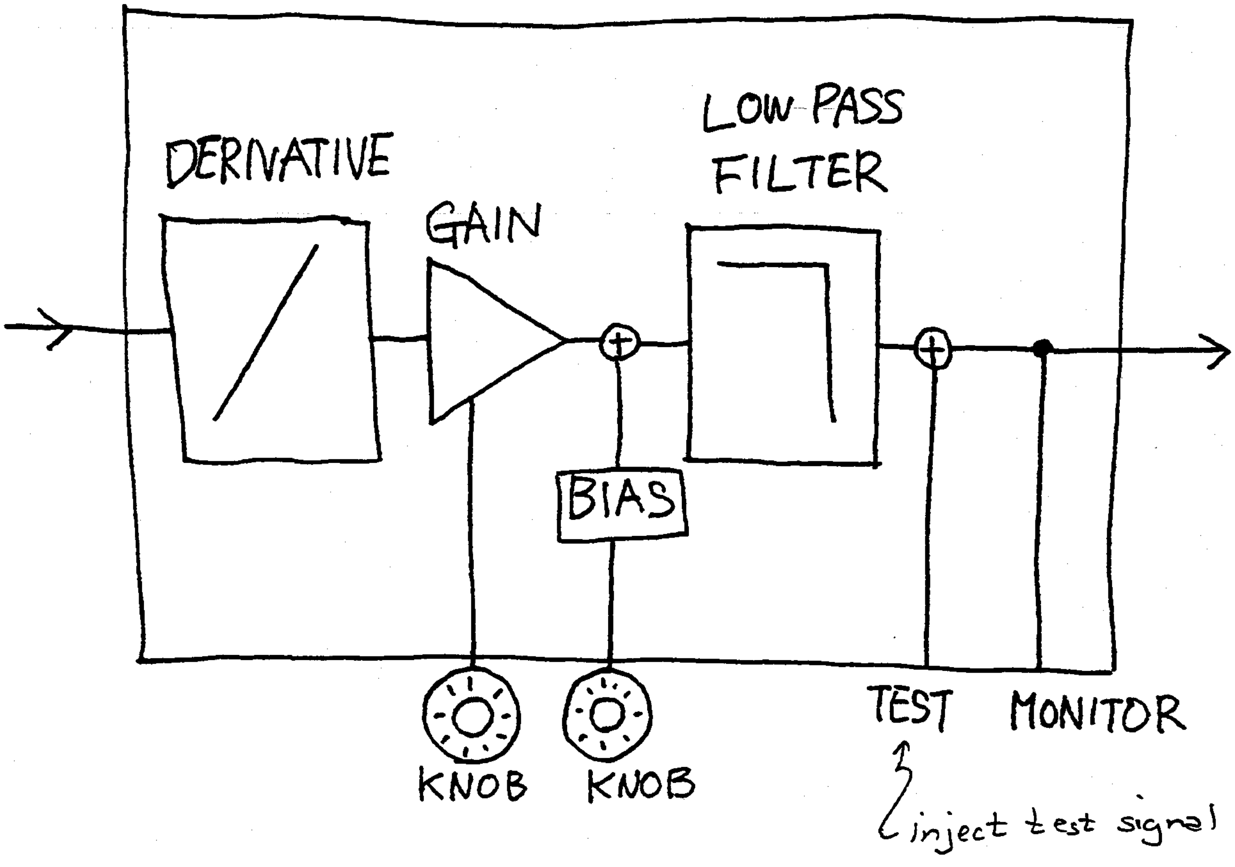
TEST MASS DAMPING CONTROL SYSTEM 2

for Mark II 40m Prototype



Pitch	**(A+B) - (C+D)**
Yaw	**(A+C) - (B+D)**

QUADRANT PHOTODIODE

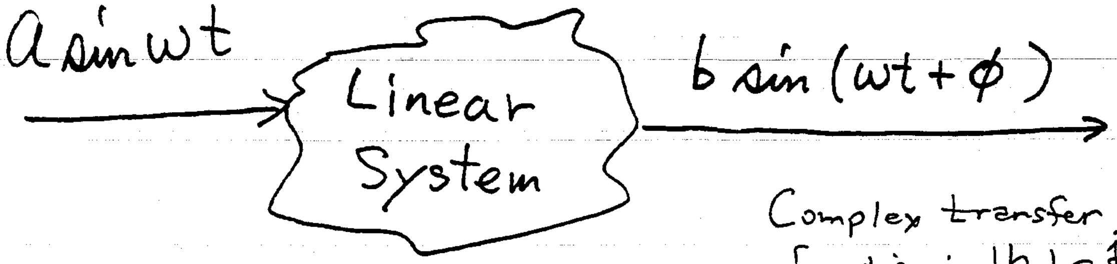


ORIENTATION DAMPING

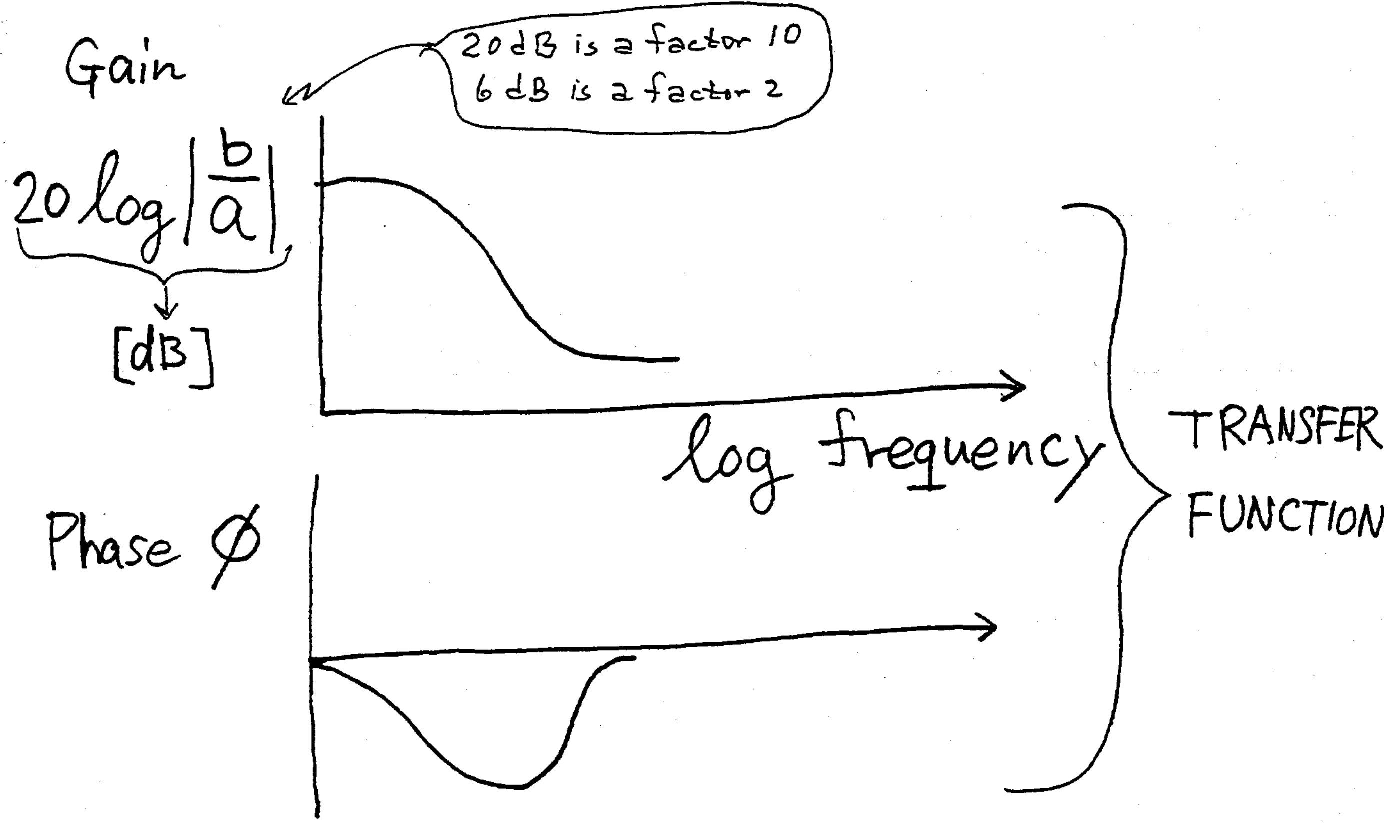
CONTROLLER

2. How to Damp a Test Mass Without Adding Extra Noise

FREQUENCY RESPONSE

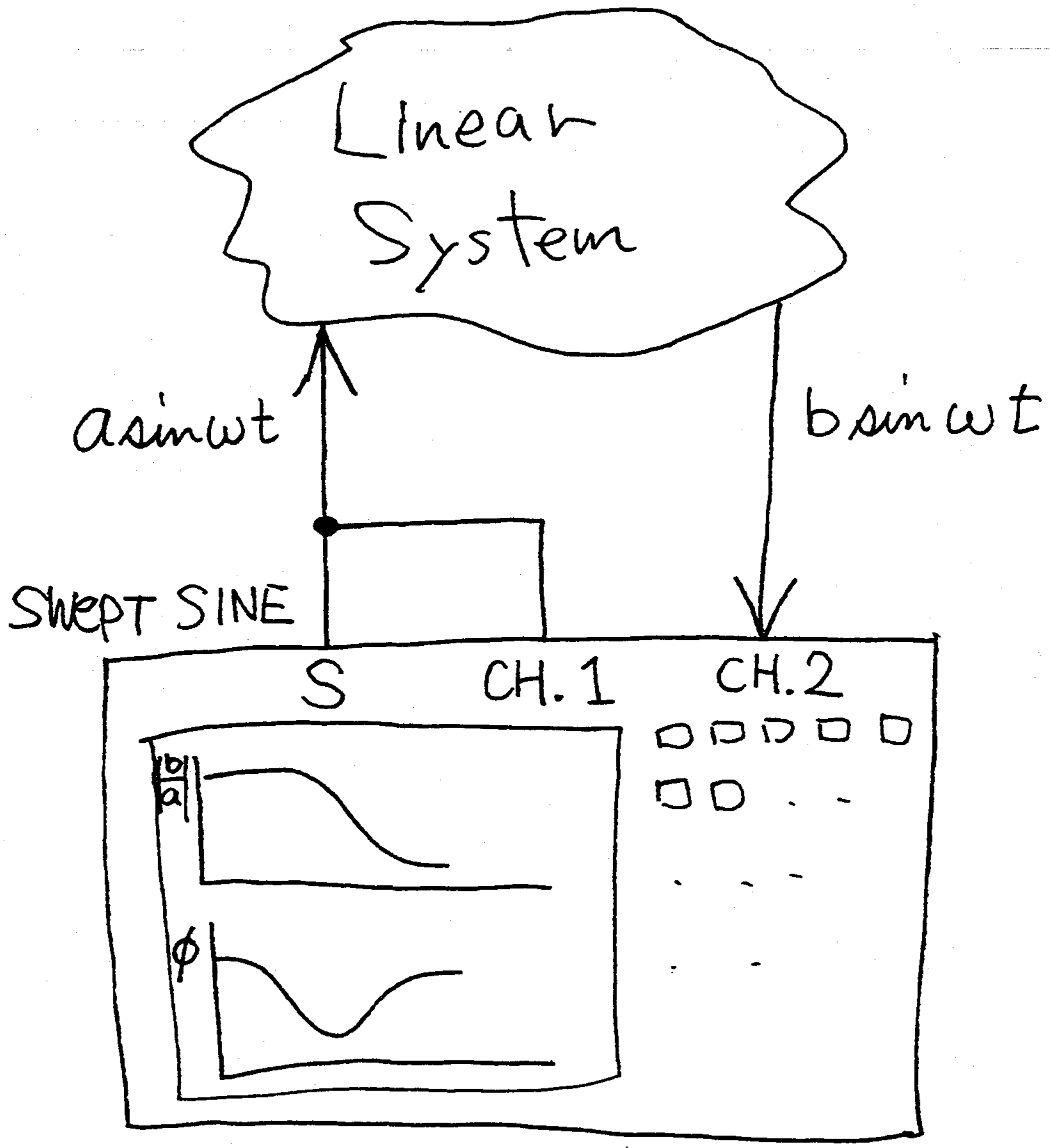


Complex transfer function: $\left|\frac{b}{a}\right| e^{j\phi}$

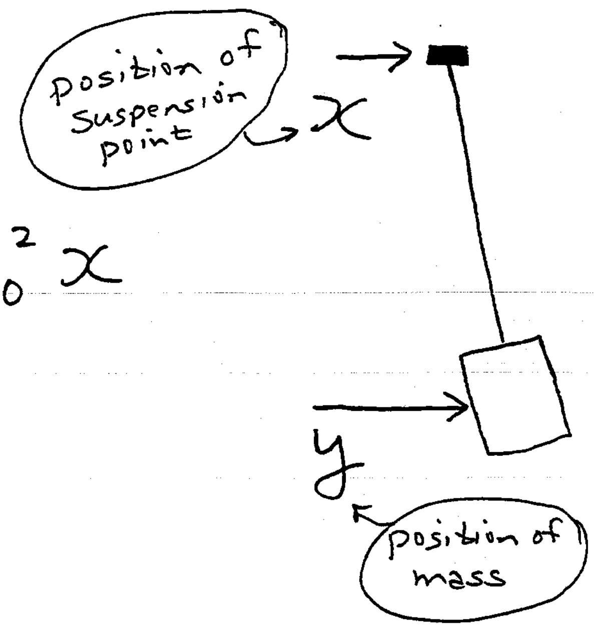


20 dB is a factor 10
6 dB is a factor 2

Bode Plot



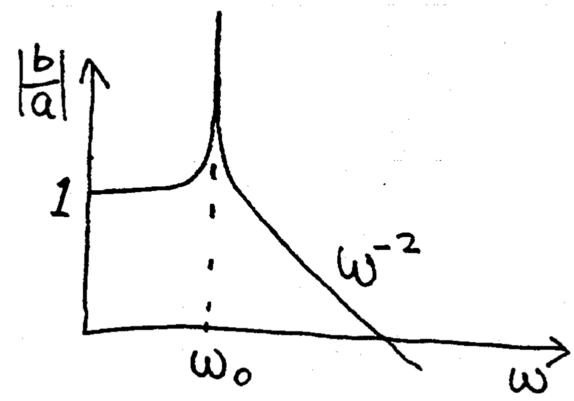
NETWORK ANALYZER



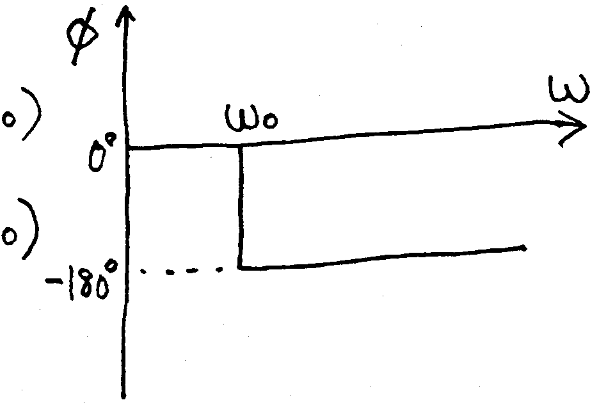
$$\ddot{y} + \omega_0^2 y = \omega_0^2 x$$

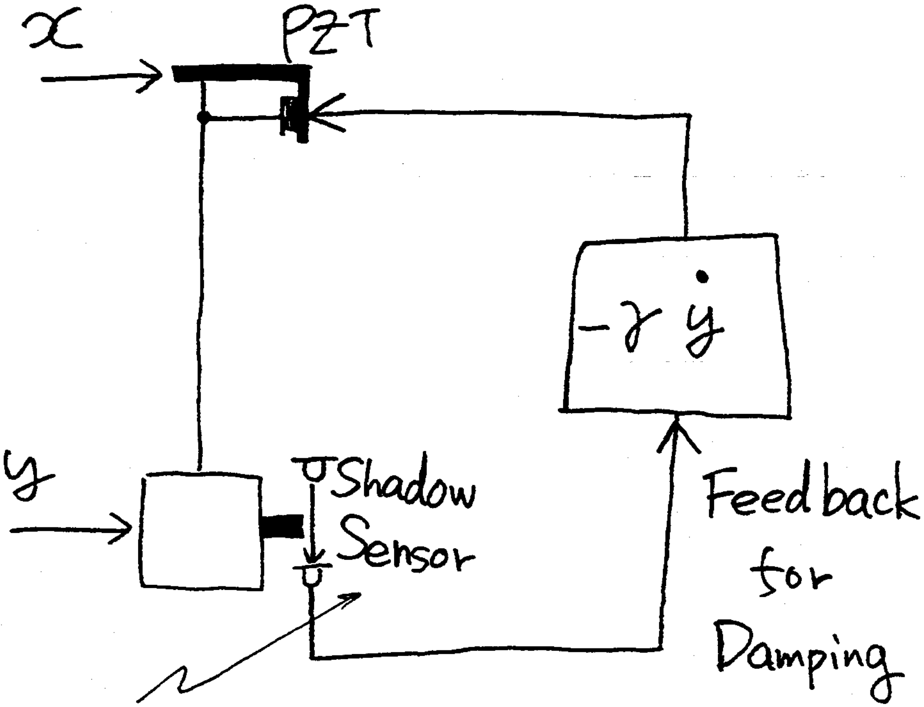
$$\left(\omega_0^2 = \frac{g}{l} \right)$$

$$\left|\frac{b}{a}\right| = \left|\frac{\omega_0^2}{-\omega^2+\omega_0^2}\right|$$



$$\phi = \begin{cases} 0^\circ & (\omega < \omega_0) \\ -180^\circ & (\omega > \omega_0) \end{cases}$$





Shadow sensor is attached to a fixed, quiet object (in problem set: attached to suspension point)

$$\ddot{y} + \frac{\omega_0}{Q}\dot{y} + \omega_0^2 y = \omega_0^2 x$$

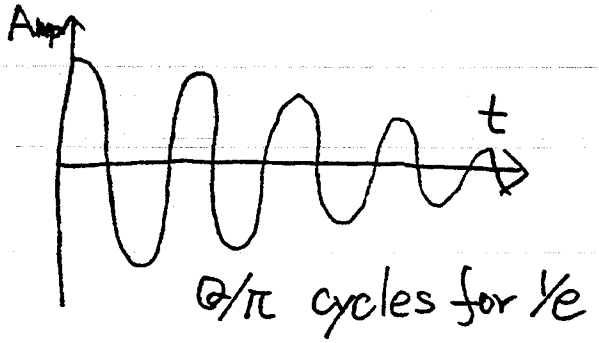
$$\left(Q = \frac{1}{\gamma\omega_0}\right)$$

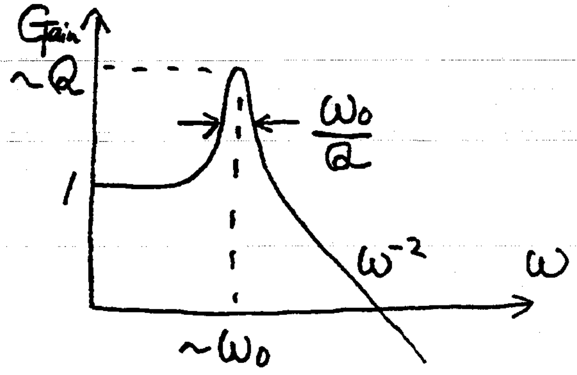
Quality Factor

Damped Pendulum

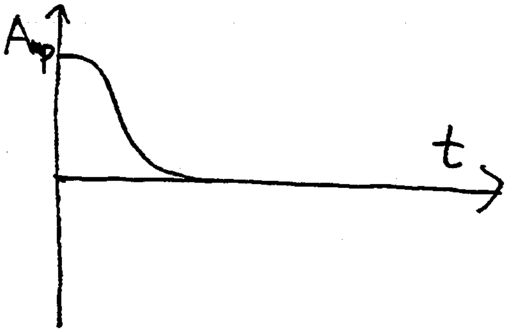
TIME RESPONSE to a short kick δx & FREQUENCY RESPONSE

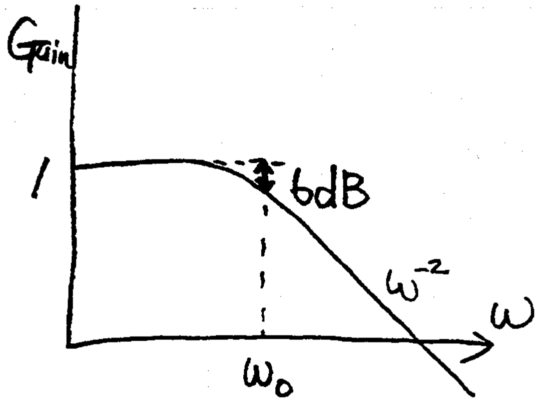
i) $Q > 0.5$ (Under-damping)



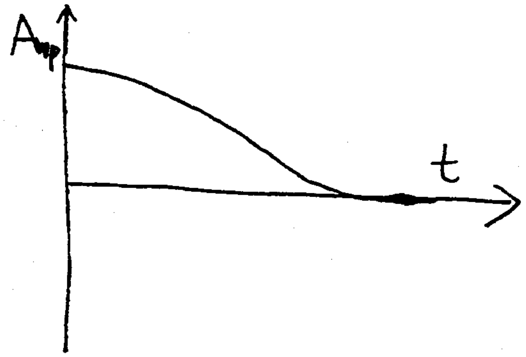


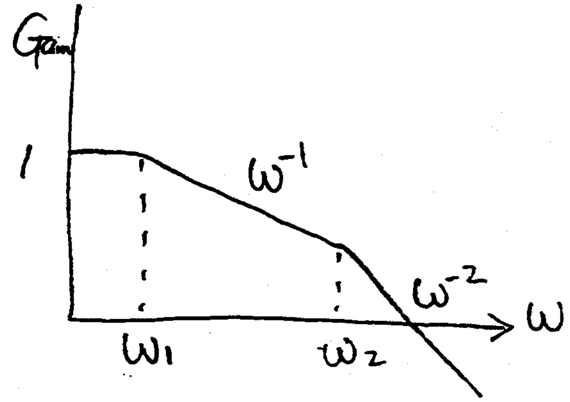
ii) $Q = 0.5$ (Critical-damping)



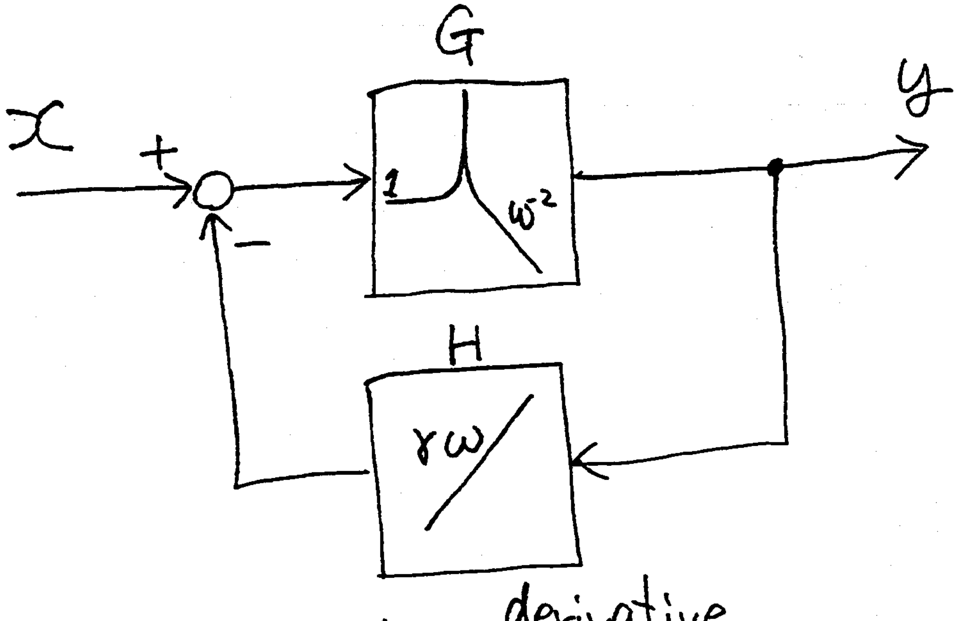


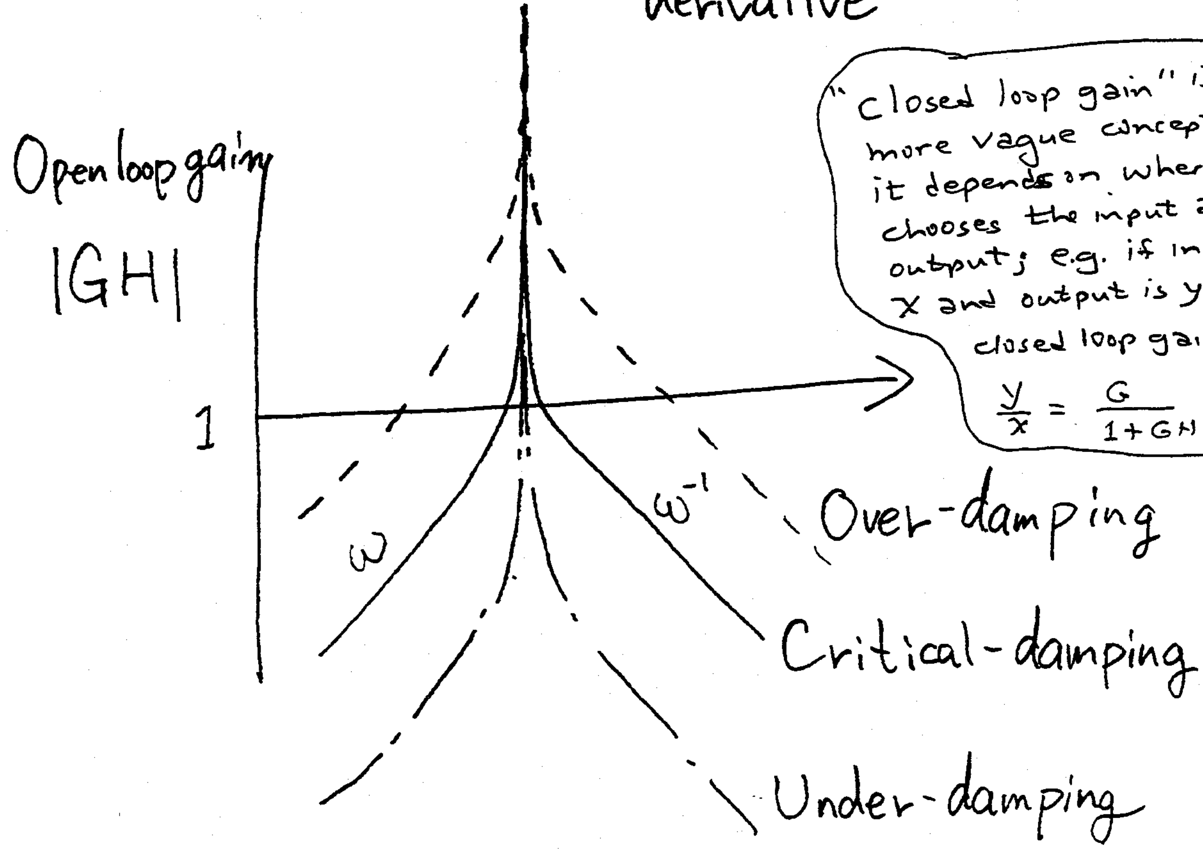
iii) $Q < 0.5$ (Over-damping)





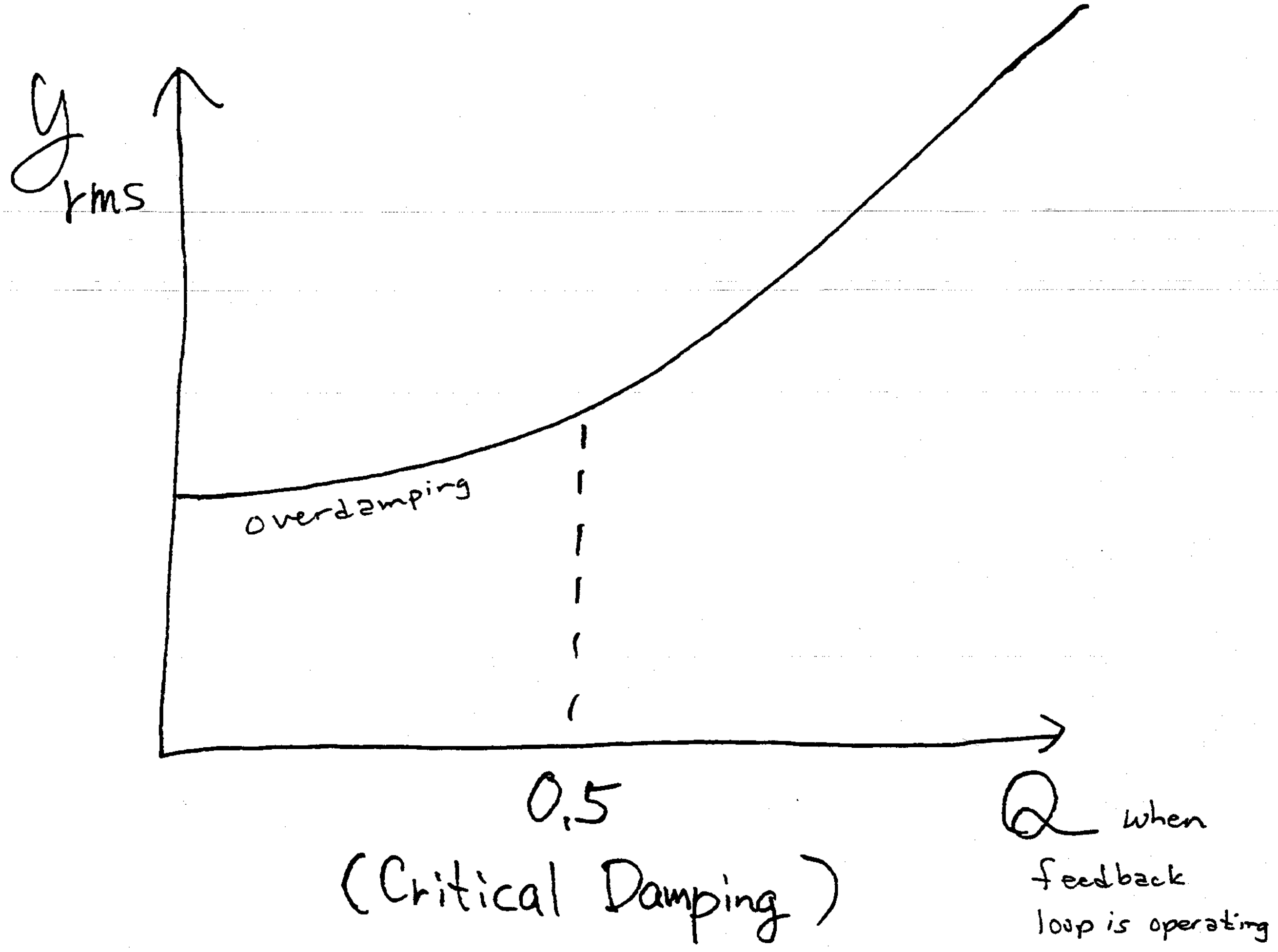
Open Loop Gain





"Closed loop gain" is a more vague concept; it depends on where one chooses the input and output; e.g. if input is x and output is y, then closed loop gain is

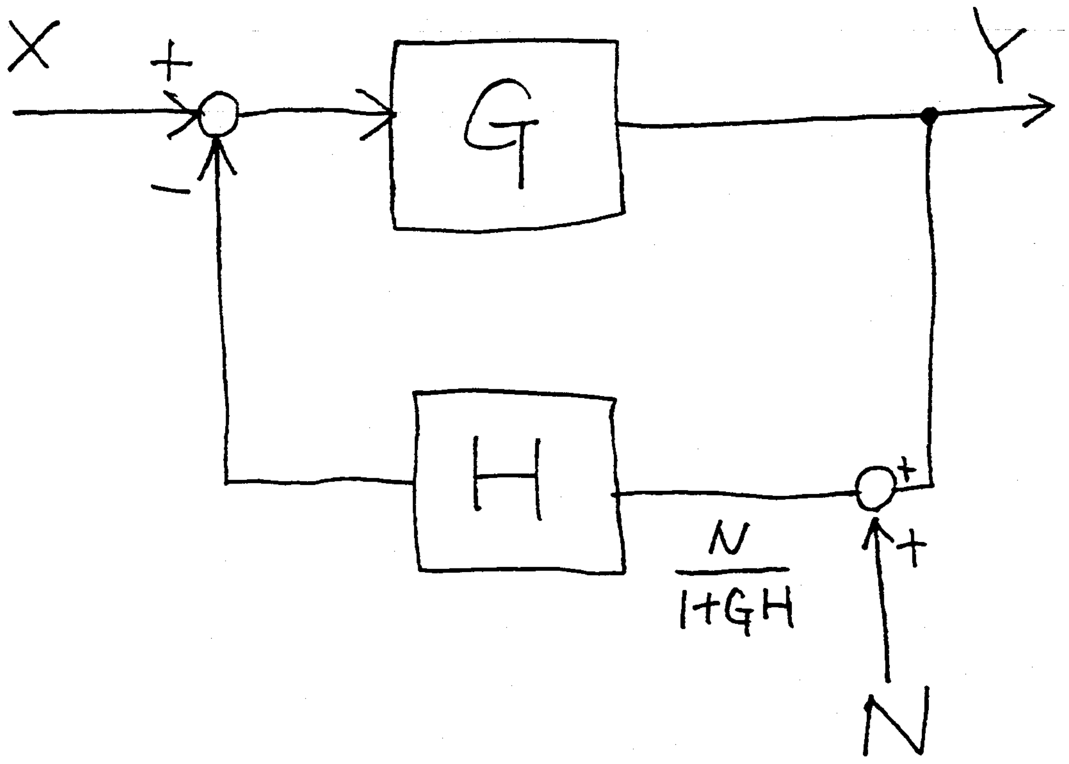
$$\frac{y}{x} = \frac{G}{1+GH}$$



As damping is increased (Q is decreased), Y_{rms} decreases. Not much is achieved by going ~~to Q~~ beyond $Q = 1/2$, to overdamped regime.

Critical damping is enough for the 40m interferometer!

Henceforth assume critical damping...



(Sensor Noise)
e.g. shot noise in sensor's photodiode, or pointing noise of sensor's beam

$$\frac{Y}{N} = \frac{-GH}{1+GH} \sim -GH \quad (\overbrace{|GH|}^{\text{open loop gain}} \ll 1)$$

To suppress sensor noise, must make $|GH| \ll 1$. If $|GH| \gtrsim 1$, then $Y/N \sim -1 \Rightarrow$ no suppression.

Sensor Noise

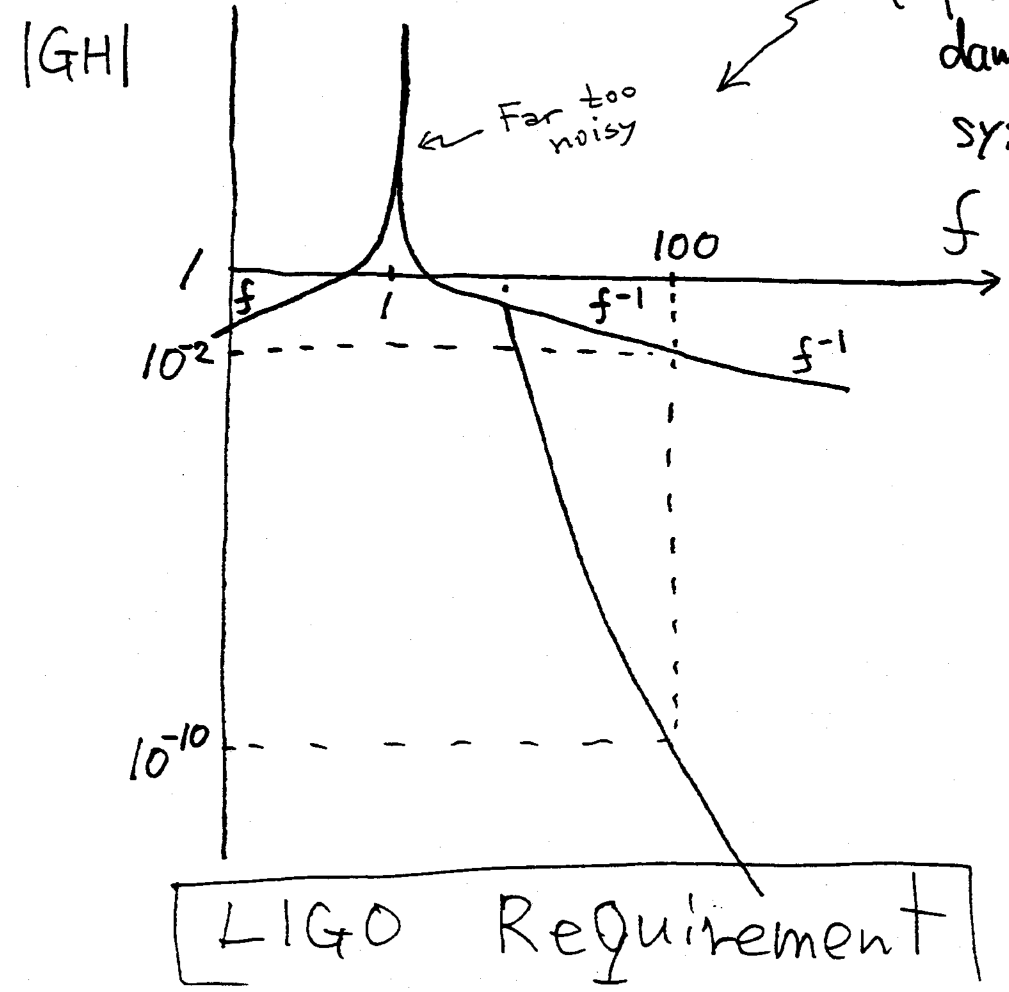
$\tilde{N} \sim 10^{-10}\ m/\sqrt{Hz}$ ← Typical shot noise achieved with a red laser sensing beam of < 1 mW power

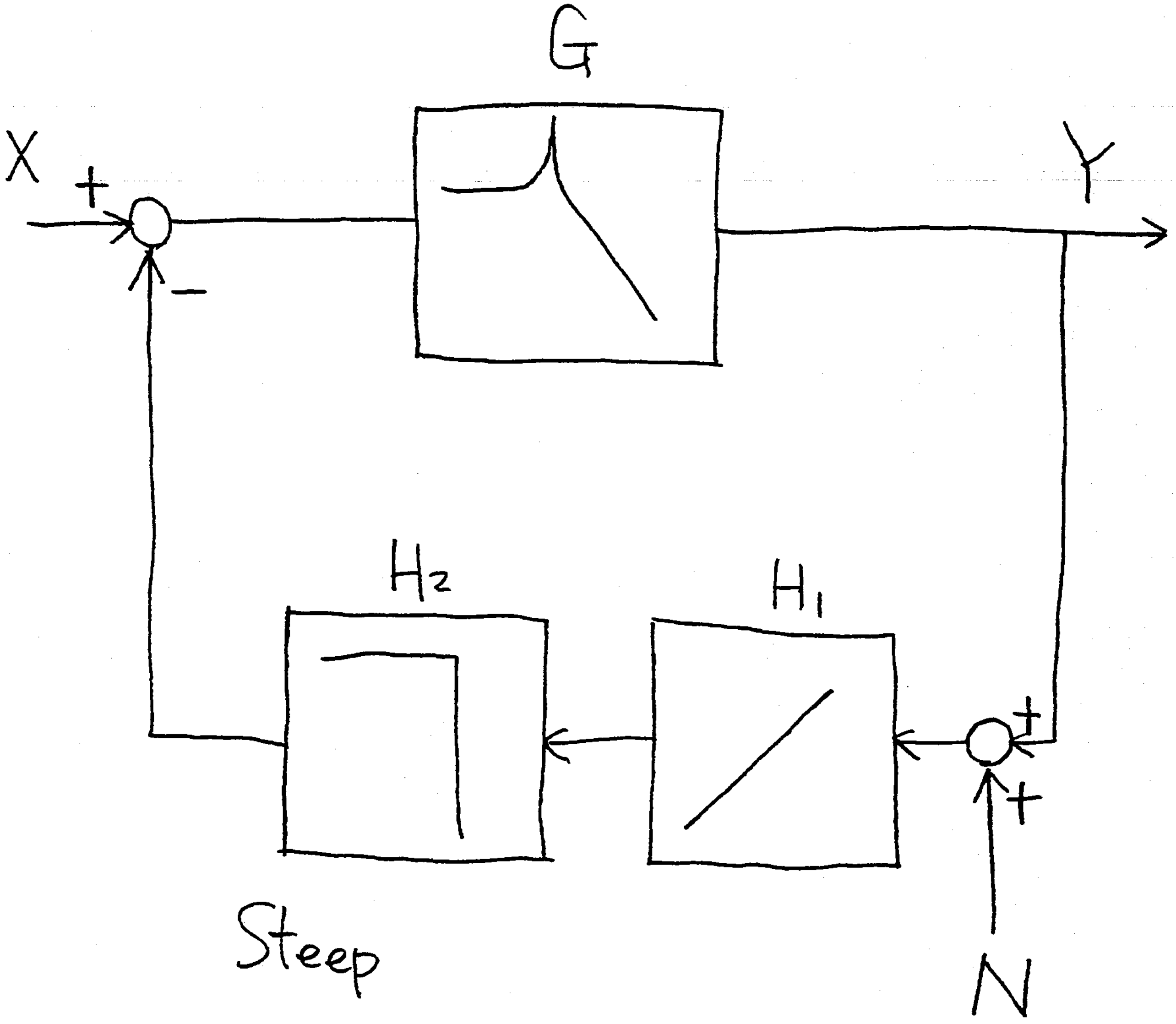
Initial LIGO → $\tilde{Y} \sim 10^{-20}\ m/\sqrt{Hz}$ at 100 Hz

$$\Downarrow$$

Open Loop Gain $|GH| < 10^{-10}$

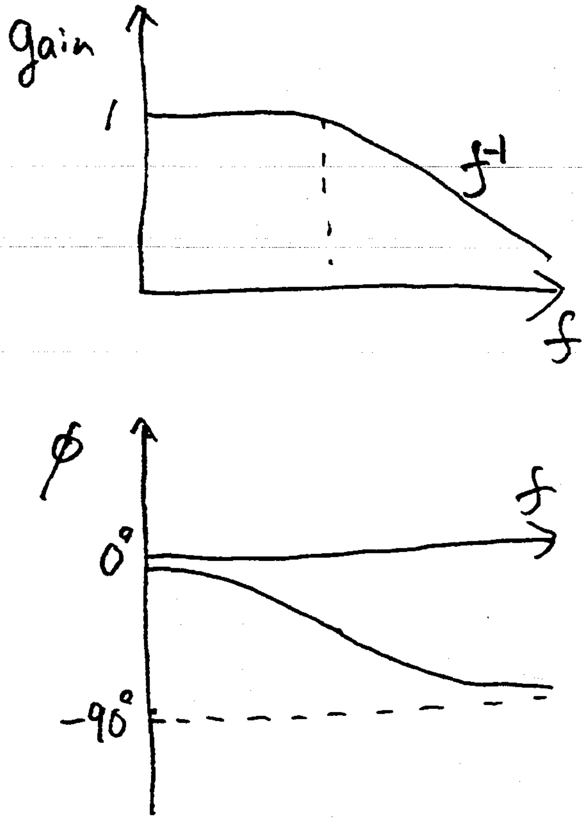
in pendulum damping system



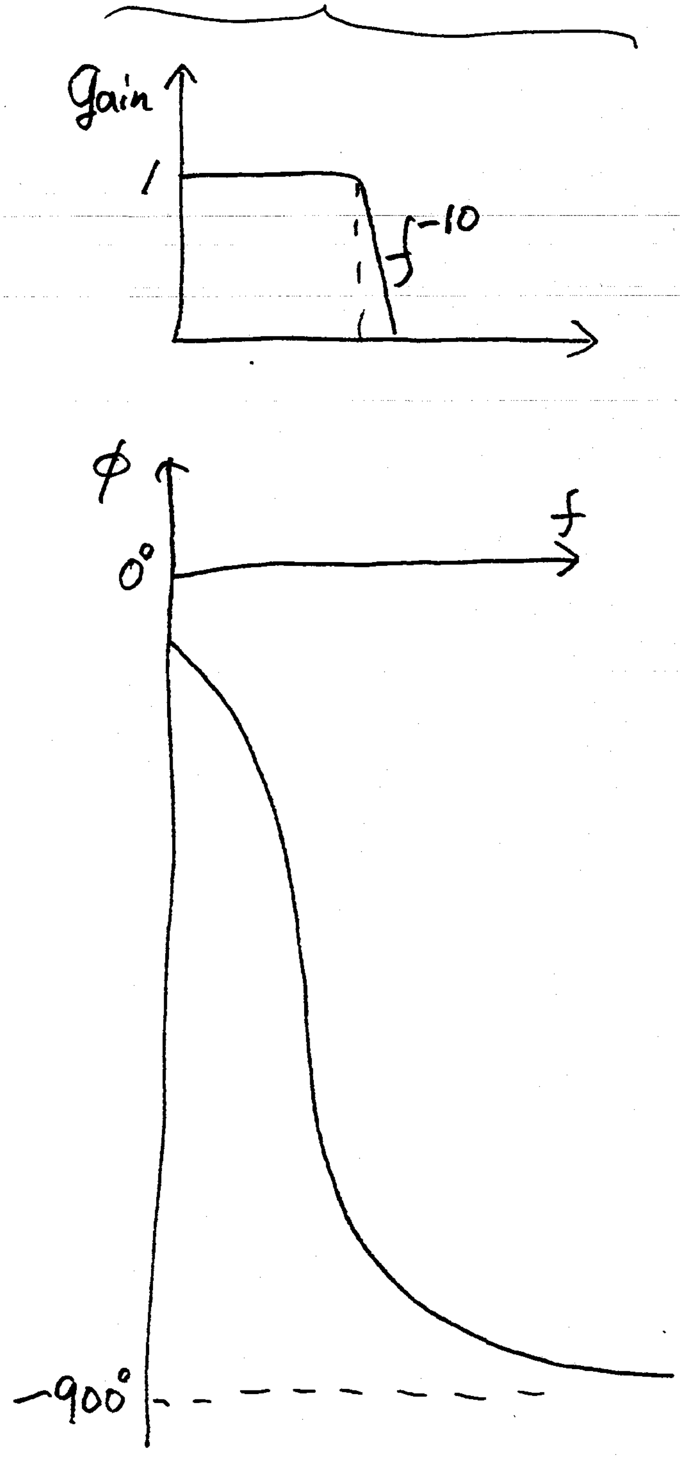


Steep
Low-Pass-Filter
to suppress sensor noise at ~100Hz

How to attack sensor noise.

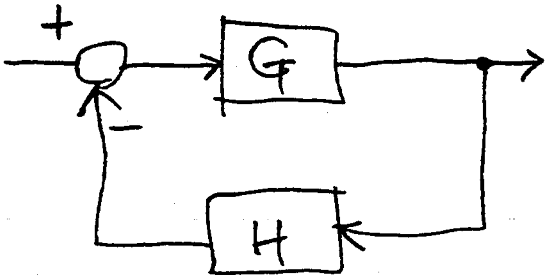


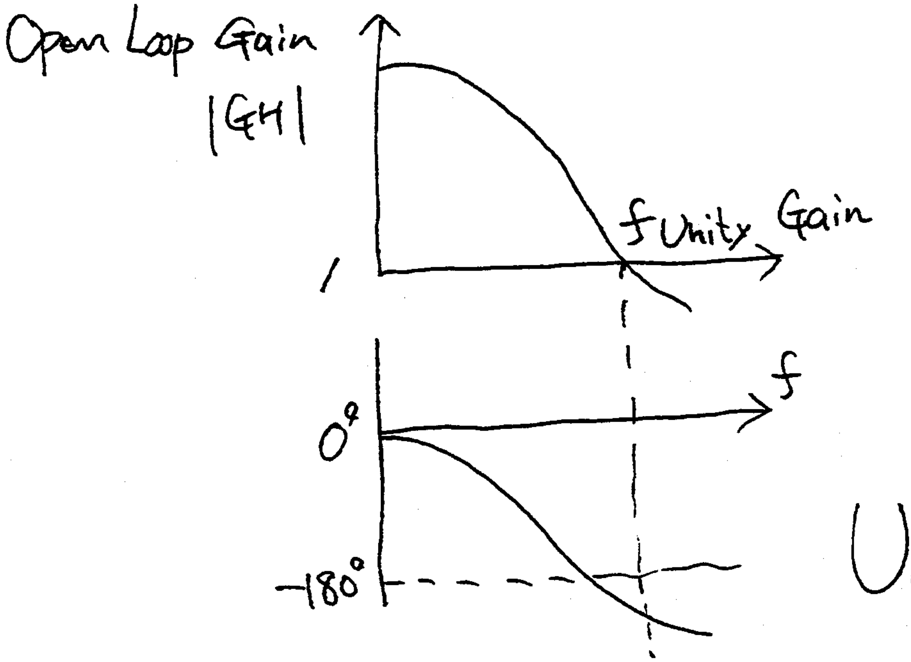
Not acceptable



Huge, unacceptable phase delay (unstable)

Phase Delay





Unstable !

If $\phi < -180^\circ$ at $f_{U.G.}$

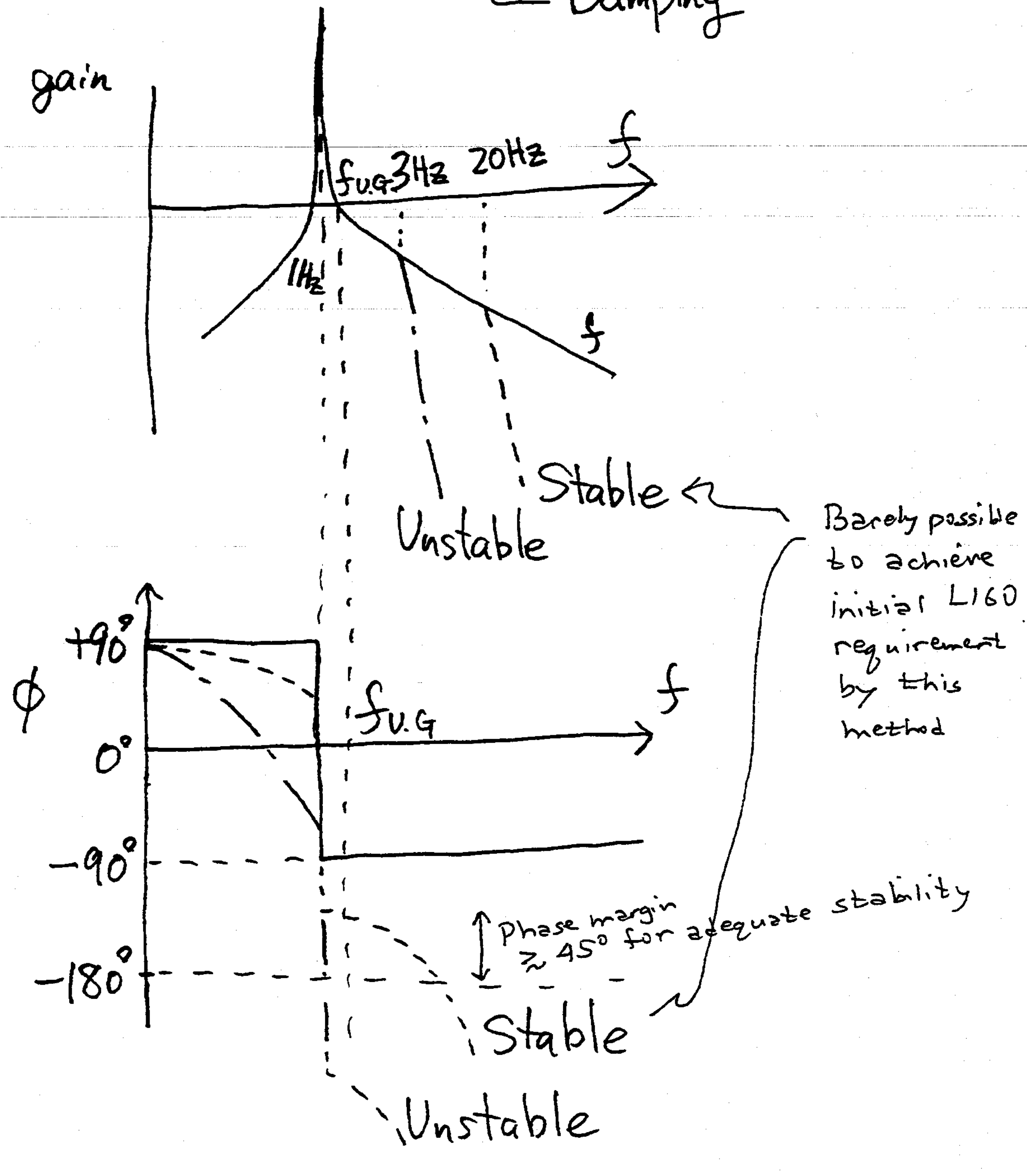
$\Longrightarrow$ Unstable

See problem set

$\boxed{\text{Stability of Feedback}}$.

STABILITY OF SYSEM

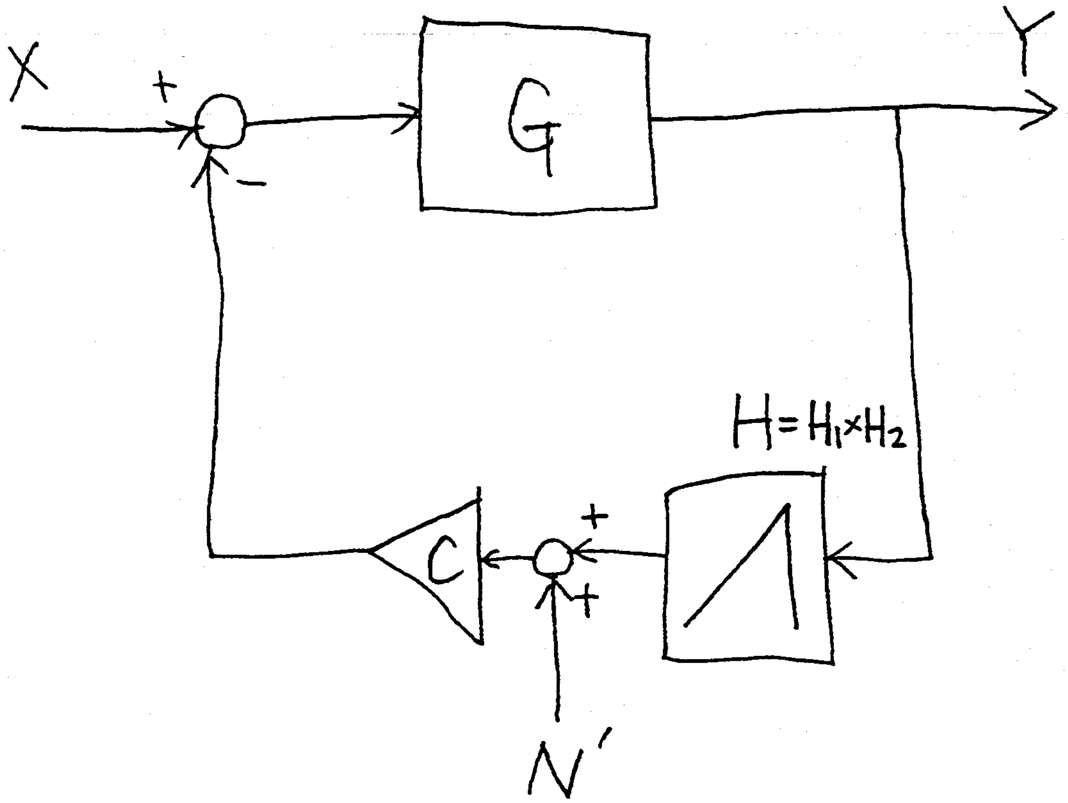
Damping



Barely possible to achieve initial LIGO requirement by this method

$\phi > -180°$ at $f_{U.G.}$ for stable system

[In practice, want $\Phi + 180° \gtrsim 45°$]



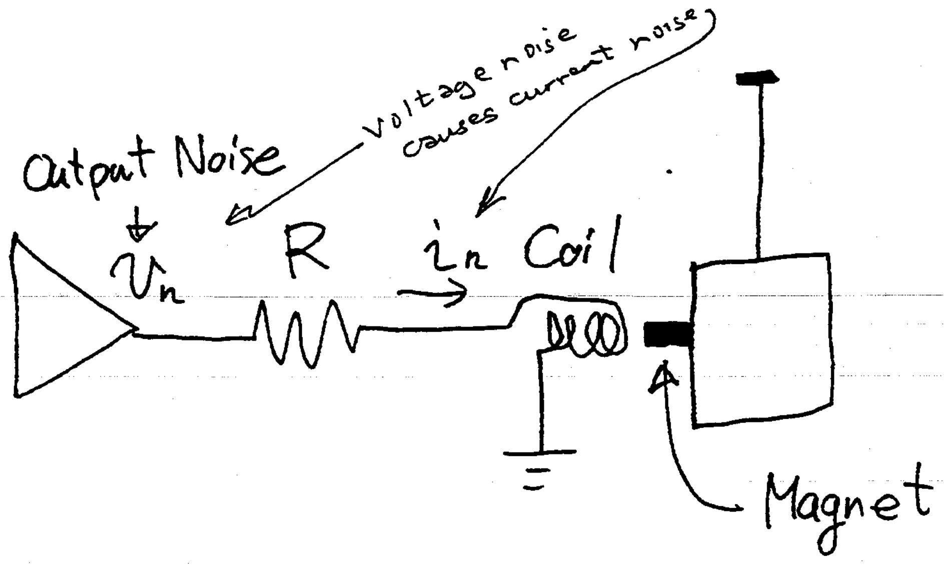
OUTPUT NOISE ← Noise produced by the steep low-pass filter

(Johnson Noise
Electronic Noise)

$$\frac{Y}{N'} \sim -CG \quad \text{when} \quad \left(|GHC| \ll 1\right)$$

$\boxed{\text{Output Noise}}$

We must reduce C to suppress low-pass filter's output noise



$R \longrightarrow$ Large

By doing some small things, like putting in this resistor R, we can keep the output noise below ~~the first~~ LIGO first interferometer requirements

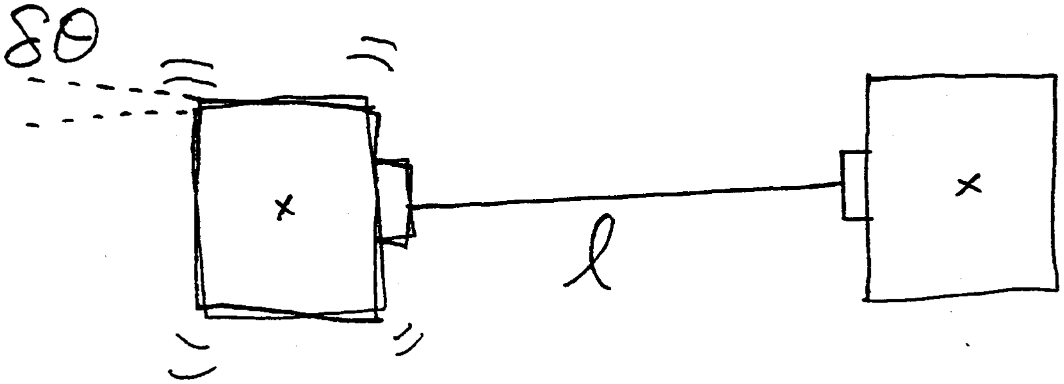
$i_n = \dfrac{V_n}{R} \longrightarrow$ Small (Good!)

Dynamic Range for i ← which produces feedback force $\longrightarrow$ Small (Bad!)

Example of Output Noise

3 How to Predict Test Mass Orientation Noise in a Fabry Perot Cavity

How test mass orientation noise affects cavity length?

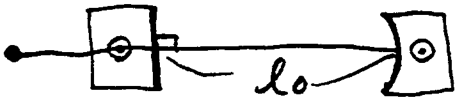


$$\delta\theta \xrightarrow{\quad} \delta\ell$$

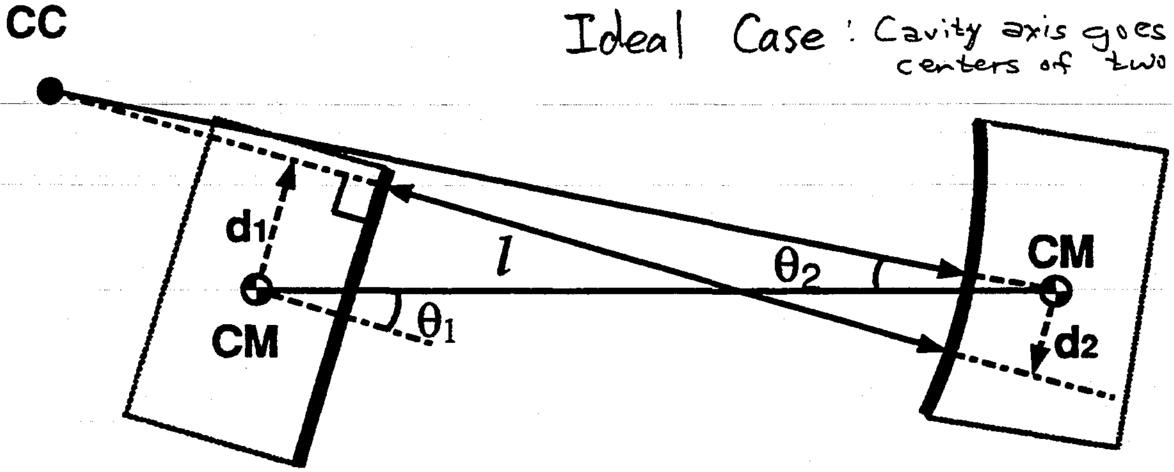
?

How does fluctuation $\delta\theta$ of test mass orientation affect cavity length ℓ?

Here we assume one mirror flat, the other curved; we could also use curved/curved.



Ideal Case: Cavity axis goes through centers of two masses



General Case

$$l = l_0 + \alpha\theta_1^2 + \beta\theta_2^2 + \gamma\theta_1\theta_2$$

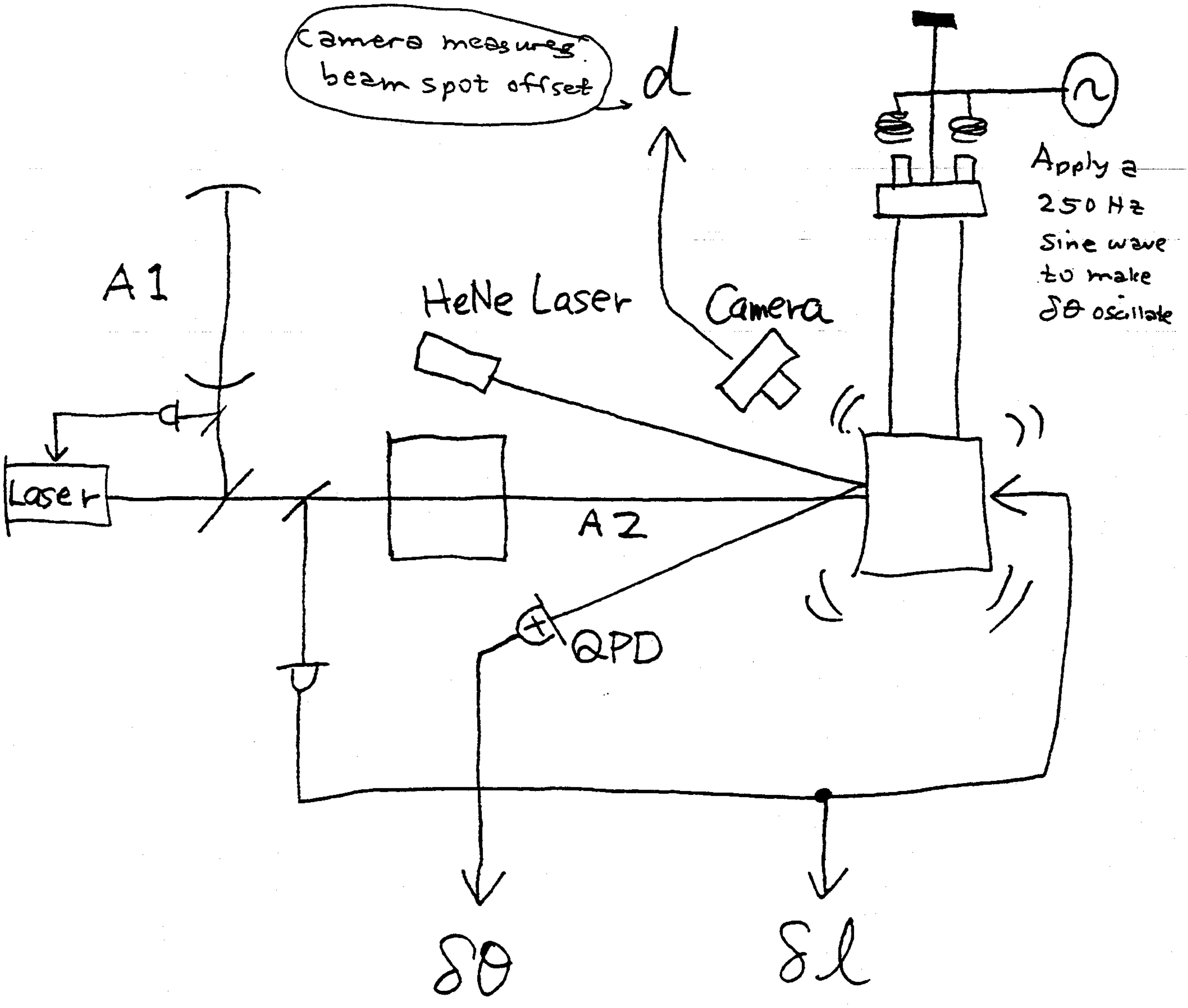
When $\theta = \underbrace{\Delta\theta}_{DC} + \underbrace{\delta\theta}_{100Hz\sim}$

produces d_1 and d_2

$$\delta l = \underbrace{d_1}_{\text{Beam spot position}}\delta\theta_1 + d_2\,\delta\theta_2$$

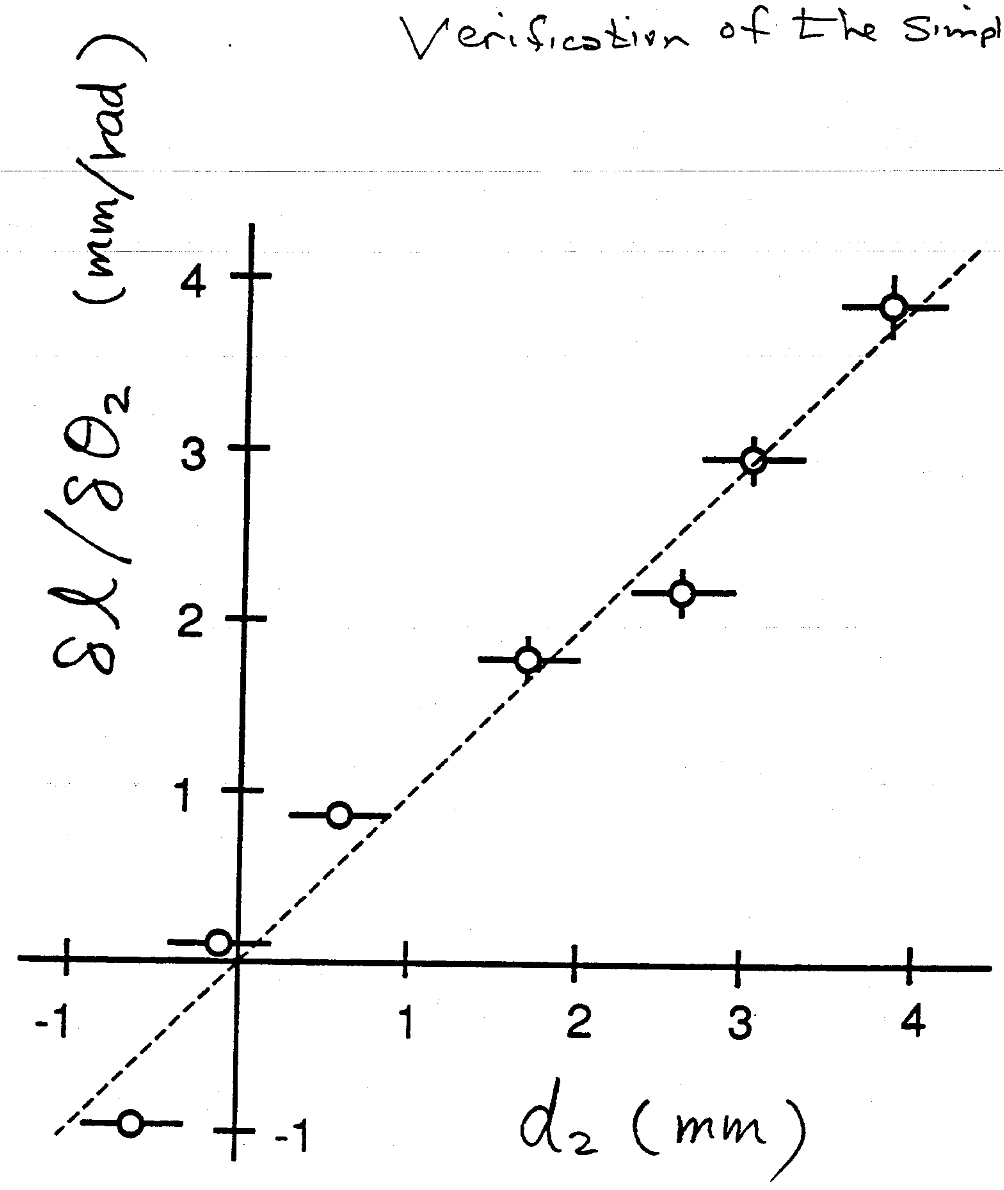
$$\boxed{\text{Linear Effect of Orientation Noise}}$$

Test of Above Theory


Camera measures beam spot offset
d
A1
HeNe Laser
Camera
Apply a 250 Hz Sine wave to make δθ oscillate
Laser
A2
QPD
δθ
δl

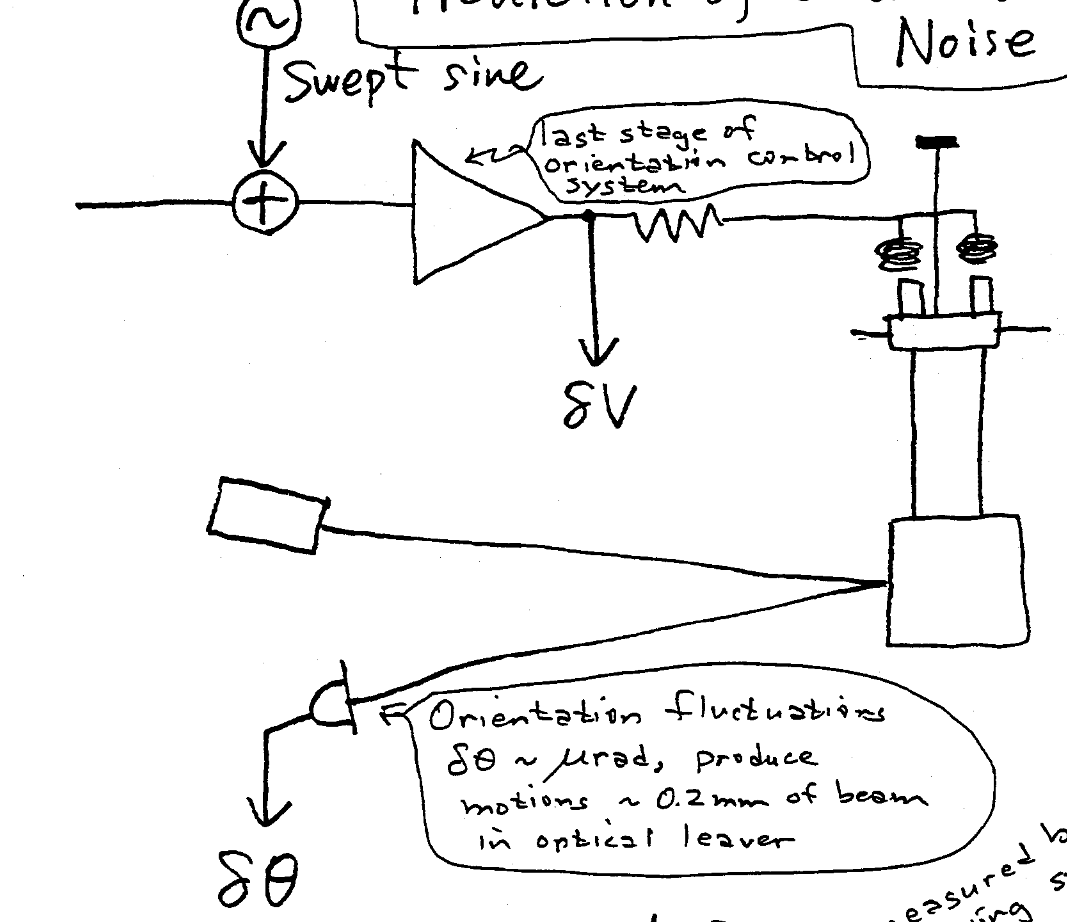
Measure δl / δθ for different beam spot position

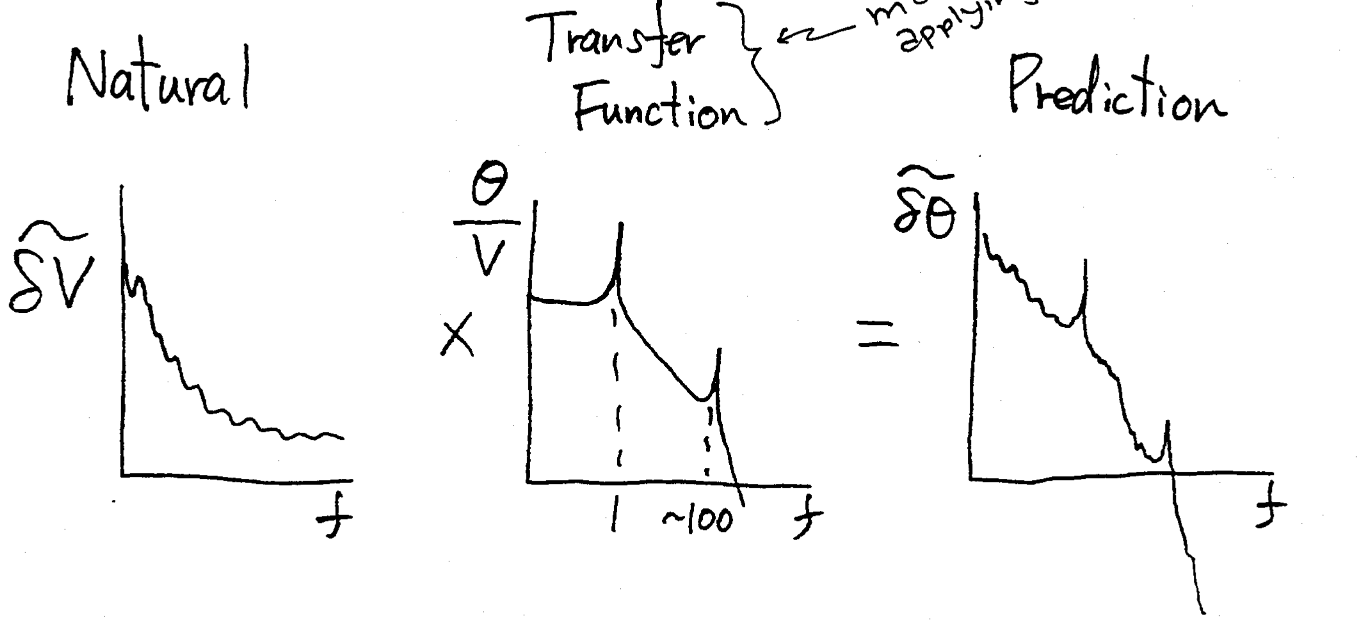
Verification of the simple theory



Linear Effect

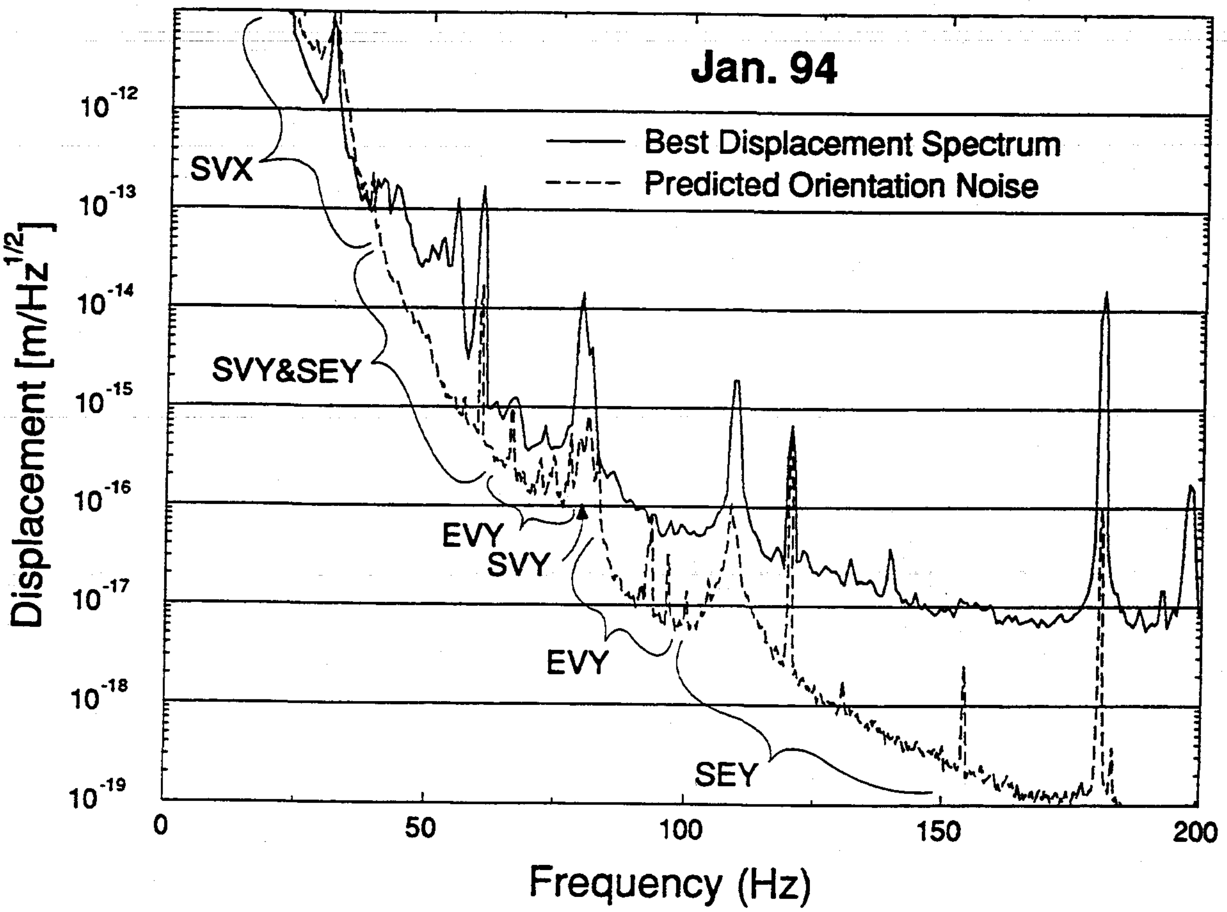
Prediction of Orientation Noise


Swept sine
last stage of orientation control system
δV
Orientation fluctuations δθ ~ μrad, produce motions ~ 0.2 mm of beam in optical leaver
δθ

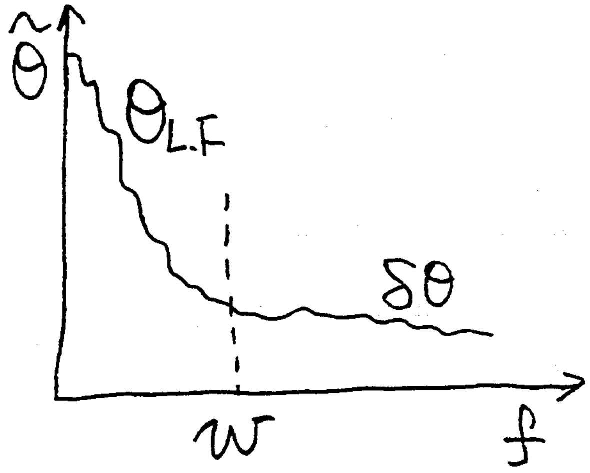


$$\widetilde{\delta l} = \underset{\sim 1\,\text{mm}}{d} \times \widetilde{\delta\theta}$$

40 m Prototype



$\theta = \theta_{L.F} + \delta\theta$

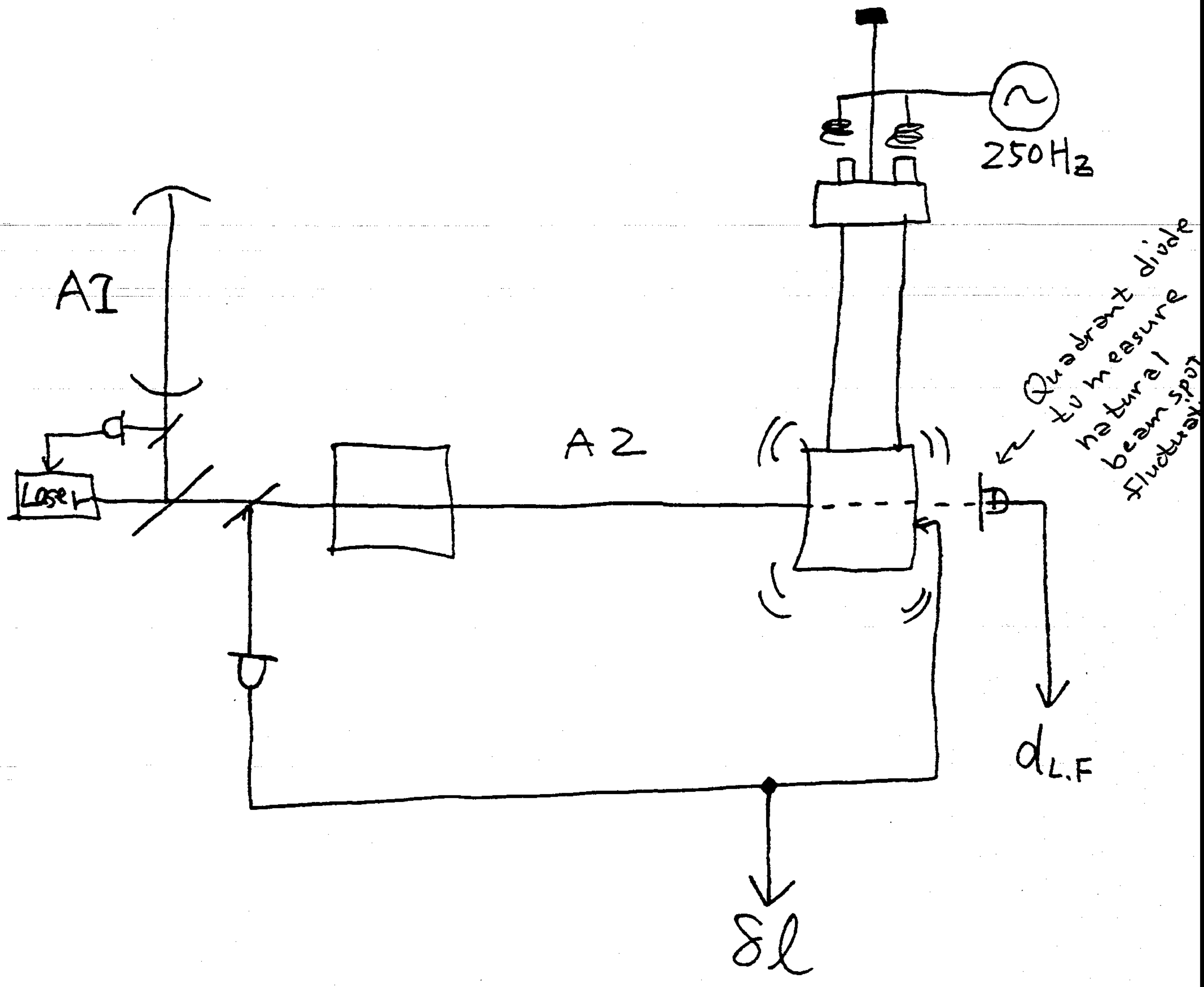


$$\delta l \sim \int_{-w}^{w} \underbrace{d_{L.F}(f')}_{\uparrow \text{low-frequency motion of beam spot at mirror}} \underbrace{\delta\theta(f-f')}_{\uparrow \text{high-frequency wiggle of mirror}} df'$$

$\boxed{\text{"CONVOLUTION EFFECT"}}$

$$\left.\begin{array}{l} d\,(10\text{Hz}) \\ \delta\theta\,(100\text{Hz}) \end{array}\right\} \rightarrow \delta l\,(90\text{Hz}, 110\text{Hz})$$

Experiments to Test Convolution Effect

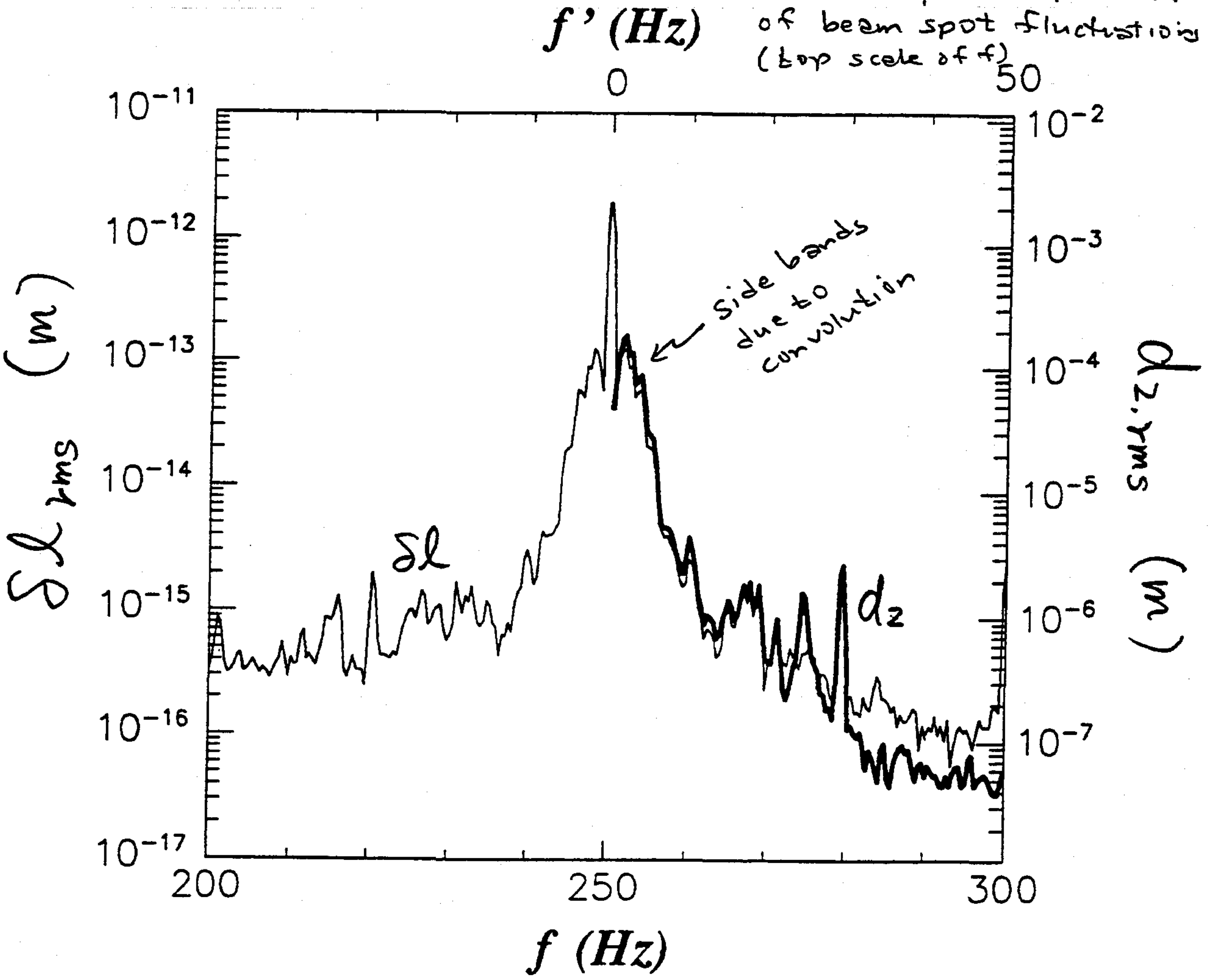


Measure δl around 250Hz and $d_{L.F}$ simultaneously

Convolution Effect I

Thin curve is interferometer noise spectrum. (bottom f scale)

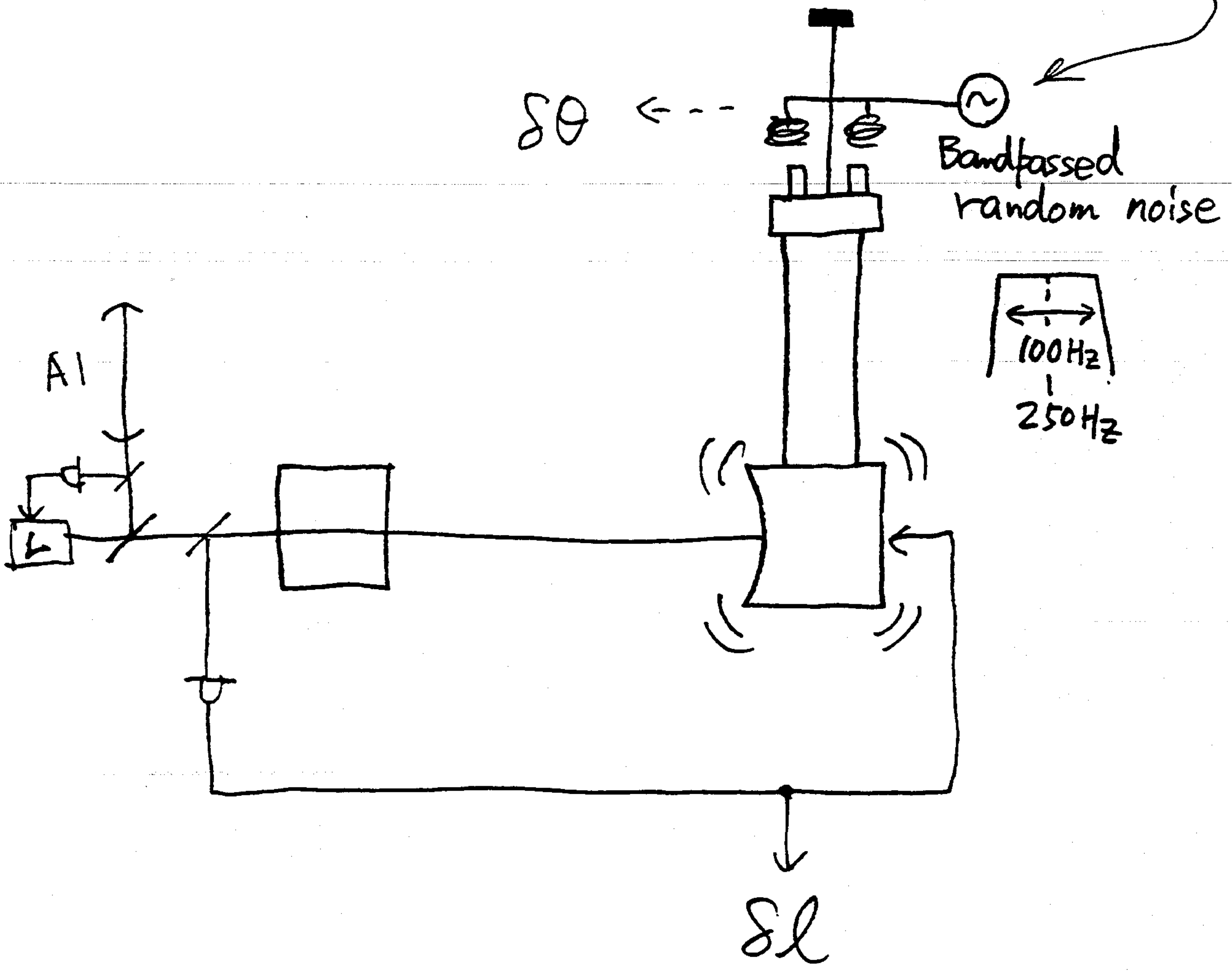
Thick curve is measured power spectrum of beam spot fluctuations (top scale of f)



side bands due to convolution

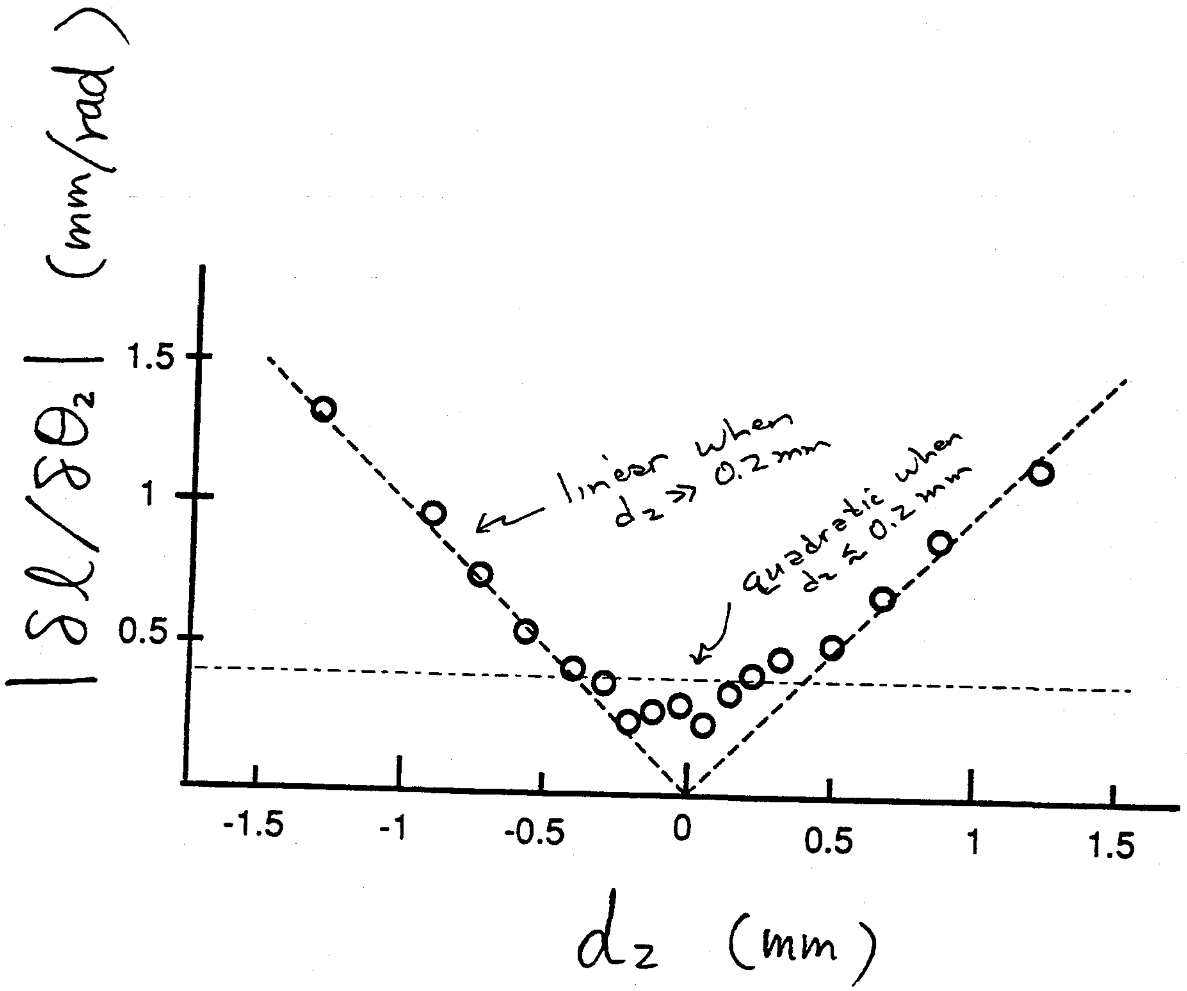
Convolution Effect I

Same as previous experiment, but now with broad-band noise driving the orientation



Measure $\delta l / \delta \theta$ for different beam spot positions

Convolution Effect II



beam spot position
(DC offset from ideal location)

$d_{z\,rms} \sim 0.2$ mm

Important !

To reduce this noise, we must reduce $d_{z\,rms}$ as well as reducing the offset

(But $d_{z\,rms} \sim 0.2$ mm is adequate for first LIGO interferometers)

Convolution Effect II

You have learned ...

1. Test mass position / orientation control is necessary !

2. To damp a test mass without adding extra noise is possible !

3. To predict test mass orientation noise in a Fabry-Perot cavity is fun !

BATCH START

11. Optical Topology

STAPLE
OR
DIVIDER

LECTURE 11.

Optical Topology for Locking and Control of an Interferometer, and Signal Extraction

Lecture by Martin Regehr

Assigned Reading:

EE. P. W. Milonni and J. H. Eberly, *Lasers* (Wiley, New York, 1988): sections 12.9 "AM Locking" and 12.10 "FM Locking," pp. 385–390. [Here you are asked to focus on the description of AM modulation and FM modulation as putting side bands onto a carrier frequency. Of particular interest is the fact that a sinusoidal FM modulation produces a whole series of side bands, whose strengths are described by Bessel functions. When the modulation amplitude is small compared to a radian, only the first side bands dominate.]

FF. C. N. Man, D. Shoemaker, M. Pham Tu and D. Dewey, "External modulation technique for sensitive interferometric detection of displacements," *Physics Letters A*, **148**, 8–16. [This paper describes in detail a technique used in LIGO to circumvent laser noise that is seriously in excess of standard photon shot noise in the gravitational wave's kHz band. The trick is to upconvert the gravitational-wave signal to ~ 10 MHz frequency, where the laser's noise is near the shot-noise level. This is achieved by modulating the laser light at ~ 10 MHz (i.e. put 10 MHz side bands on the light's $\sim 10^{15}$ Hz carrier frequency), and then arranging that the gravitational-wave signal becomes a ~ 1 kHz side band of the 10 MHz side band.]

Suggested Supplementary Reading:

Read in one or more electronics or laser textbooks about places elsewhere in experimental physics and engineering where noise is circumvented by upconverting a signal to higher frequency via modulation, and then recovering the signal by synchronous demodulation. For example, read about "lock-in amplifiers," which do this. Two references dealing with this were passed out in class:

GG. John H. Moore, Christopher C. Davis, and Michael A. Coplan, *Building Scientific Apparatus: A Practical Guide to Design and Construction* (Addison-Wesley, 1983), Sec. 6.8.3 "The lock-in amplifier and gated integrator or boxcar," (pp. 435–437).

HH. Paul Horowitz and Winfield Hill, *The Art of Electronics* (Cambridge University Press, Cambridge, 1980), Sec. 14.15 "Lock-in detection" (pp. 628–631) and an earlier section to which it refers, Sec. 9.29 "PLL components, Phase detector" (pp. 429–430).

A Few Suggested Problems: See the next page.

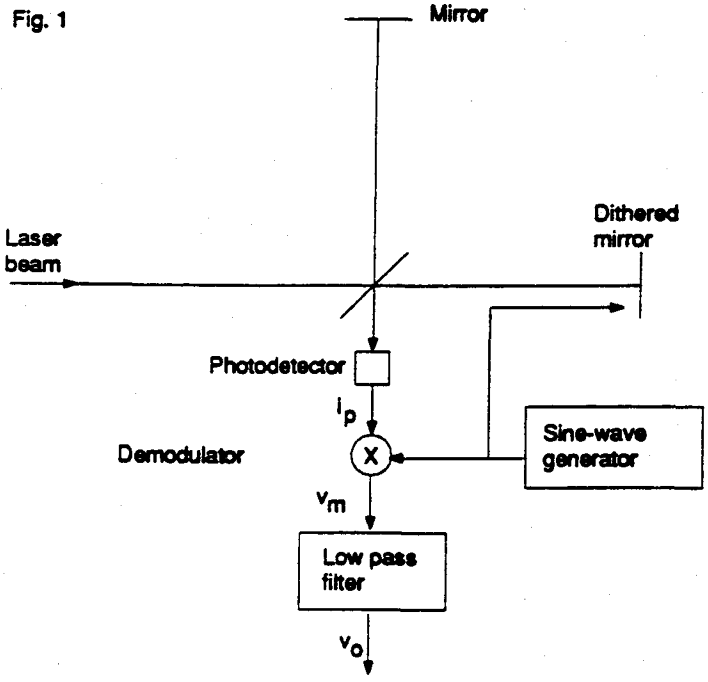
Problems for Ph 103 5/4/94

1. In this problem we will calculate the shot noise limited sensitivity of an ideal Michelson interferometer which is modulated around the dark fringe by dithering one of the mirrors with $\delta \sin \omega t$. We model the demodulator as a device which multiplies its input by $\sin \omega t$ and the low-pass filter as a device which averages over an interval T:

$$v_o(t) = \frac{1}{T} \int_{t-T}^{t} v_m(t')\,dt'$$

and for convenience we choose T to be an integral number of modulation periods $T = \frac{2\pi n}{\omega}$. Assume that the interferometer is small enough that we can neglect the light travel time from the dithered mirror to the photodetector, and that δ is very small $\delta \ll \lambda$.

Fig. 1



a. Find the derivative of the low-pass filter output with respect to displacement of the mirror which is not being dithered. It should be a function of the amplitude δ of the dithering.

b. Find the shot noise in $v_m(t)$ at the demodulator output (assuming that the ouput is at a dark fringe except for the dither):

   i. Use the time averaged photocurrent to calculate the shot noise $S_{i_p}(f)$.

   ii. Assume that the shot noise at the mixer output is related to $S_{i_p}(f)$ by the time average of the square of the mixer gain, i.e.:

$$S_{v_m}(f) = \overline{S_{i_p}(f)(\sin^2 \omega t)}$$
$$= \frac{1}{2}S_{i_p}(f)$$

It should also be a function of the dithering amplitude. Finally $S_{v_o}(f) = S_{v_m}(f)$ since the low-pass filter passes noise in the signal band virtually unattenuated.

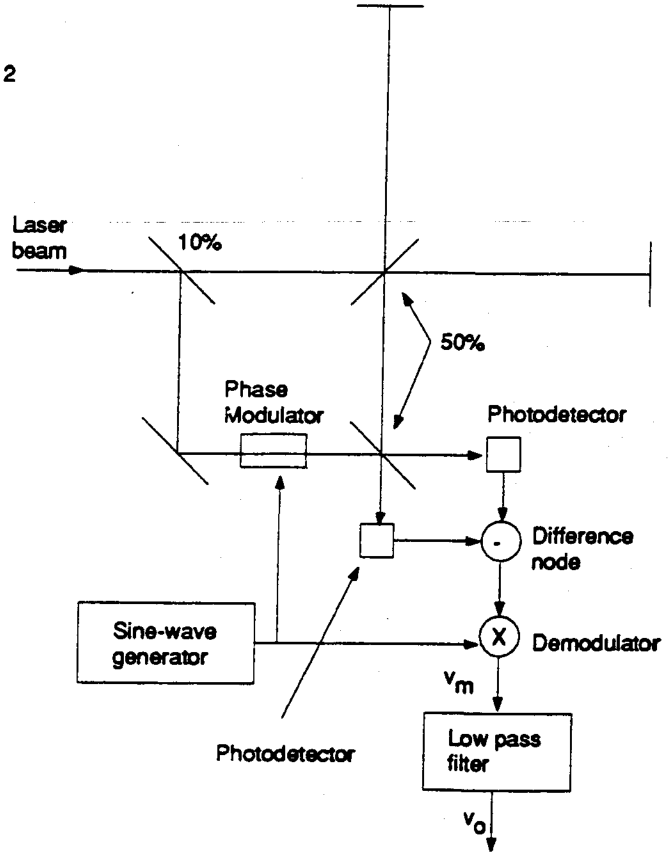
c. Take the ratio of the above two quantities to find the shot noise limited displacement sensitivity

$$S_x^{\frac{1}{2}}(f) \equiv \frac{\sqrt{S_{v_m}(f)}}{\frac{dv_o}{dx}}$$

It should be a function of the optical power and wavelength, and independent of the dithering amplitude.

2. Consider the externally modulated Michelson interferometer shown in Figure 2. Find the derivative of the low-pass filter output with respect to displacement of one of the Michelson end mirrors, assuming a pick-off which diverts 10% of the power from the main beam, a 50/50 beam splitter, a 50/50 beam combiner, a demodulator modeled as in question 1, the input to which is the difference in the photocurrents, and a low-pass filter modeled as above. Write your answer in terms of the optical power and Bessel functions of the modulation index.

Fig. 2


Laser beam
10%
50%
Phase Modulator
Photodetector
-
Difference node
Sine-wave generator
X
Demodulator
v_m
Low pass filter
v_o
Photodetector

Lecture 11
Optical Topology for Locking and Control of an Interferometer and for Signal Extraction

by Martin Regehr, 4&25 May 1994

Regehr lectured at the blackboard. His own lecture notes are illegible; the best available notes from this lecture are the ones that Kip scrawled down during it; they follow.

Regehr's lecture came in two parts: Part I, on 4 May, was a general introduction to the use of frequency modulation and demodulation to liberate signals from low-frequency noise; cf. the assigned reading. In this Part, Regehr confined attention, for pedagogical simplicity, to a simple one-bounce Michaelson interferometer.

Part II, given in the last half hour of the 25 May class, described how one can use modulation and demodulation to acquire information about the four key lengths that must be controlled in a power recycled gravity-wave interferometer (i.e., in LIGO's first interferometers). Two optical topologies for doing this are described.

Lecture 11 - Regehr: Optical Topology 5/4/94 1

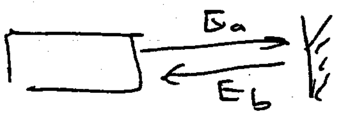
PART I

1. Laser puts out beam

$$\vec{E} = \vec{p}\left(\sqrt{\frac{2h\nu}{c\epsilon_0}}\right) \text{Re}\left\{E_a e^{i(2\pi\nu t - kz)}\right\}$$

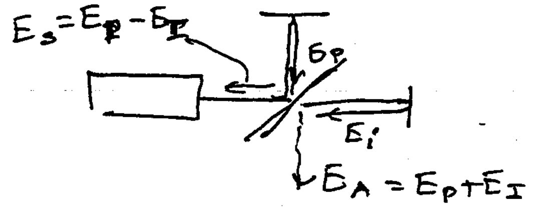
$\Rightarrow$ E_a ~~has~~ ~~is~~ has dimensions $\sqrt{\frac{dN_\gamma}{dt}} = \sqrt{\frac{\#(\text{number of photons})}{\text{Second}}}$

2. If have a mirror



$$\vec{E} = \vec{p}\sqrt{\frac{2h\nu}{c\epsilon_0}}\ \text{Re}\left\{E_a e^{i(2\pi\nu t - kz)} + E_b e^{i(2\pi\nu t + kz)}\right\}$$

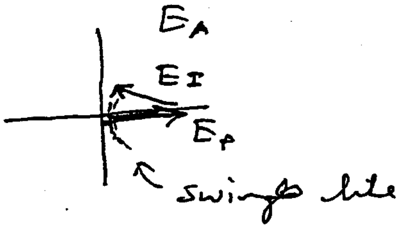
3. Michelson Interferometer



$$E_A = E_P + E_I$$

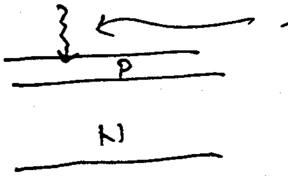
Treat beam splitter as ~~having~~ combining waves with ± on opposite sides

Describe $E_A = E_P + E_I$, $E_S = E_P - E_I$ } by phasors



4. Photodetector

a. Is a semiconductor

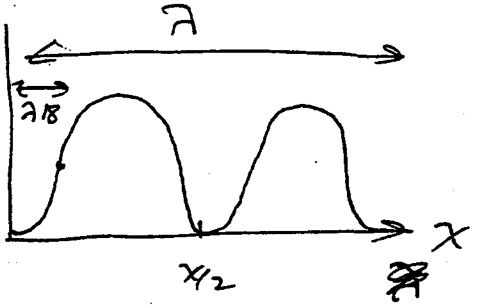


one photon kicks ~~one~~ electron up into conduction band, and it flows

- a good photodetector gives one electron per photon!

5. If move one mirror by x, get current @ photodiode

a. $i_p = |E_A|^2$
$= |E_{la}|^2 \sin^2\left(\frac{2\pi x}{\lambda}\right)$



b. Best place to operate is place of maximum di_p/dx ⇒ @ ~~λ~~ $x = \frac{\lambda}{8}$

c. But want maximum Signal to noise.

d. Suppose operate @ $x = \delta \ll \lambda$

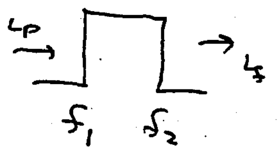
Then $\left.\frac{di_p}{dx}\right|_{x=\delta} = |E_e|^2 \frac{2\pi}{\lambda} \sin\frac{4\pi\delta}{\lambda} \simeq |E_e|^2 \cdot 2\delta \left(\frac{2\pi}{\lambda}\right)^2$

e. Shot noise in photocurrent

$S_{i_p}(f) = 2\,|E_\delta|^2$ [units $\frac{(\text{photons}/s)^2}{Hz}$

Why 2, intuitively?

i. ~~For understanding~~: Send current i_p into a band pass filter



Signal out of filter has rms noise

$$\sigma_{i_p} = \left[\int_{f_1}^{f_2} S_{i_p}(f)\,df\right]^{1/2}$$

When filter: $S_{i_f}(f) = |\tilde{H}(f)|^2 S_{I_p}(f)$

↑ transfer function of filter

f. Equivalent noise of mirror motion

$S_x^{1/2}(f) \equiv \frac{S_{i_p}^{1/2}(f)}{d i_p/dx} = \frac{1}{|E_e|}\frac{\lambda}{\sqrt{8\pi}}$ independent of δ

$\propto 1/\sqrt{\text{laser power}}$

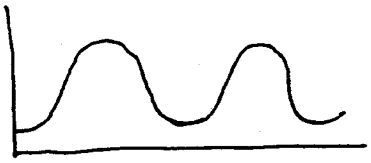
$\propto \lambda$

6. How choose δ?

a. We've learned it doesn't matter, for small δ
For large δ we get reduced sensitivity

b. If make δ too small, then signal is very small, noise is very small .... & one can be sensitive to imperfections in the interferometer.

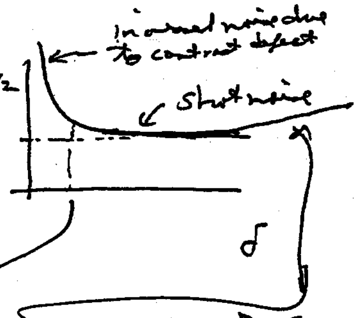
7. Example: If mirrors are irregular, get phase front mismatching @ beam splitter, & dark fringe is not fully dark.



This changes ~~sta~~ the analysis

$$S_{L_P}(f) = 2\left(|E_\Delta|^2 + \underbrace{|E_{cD}|^2}_{\text{extra light due to contrast defect}}\right)$$

$$\rightarrow S_x^{1/2}(f) = \frac{\sqrt{2|E_\ell|^2\left(\frac{2\pi\delta}{\lambda}\right)^2 + |E_{cD}|^2}}{2|E_\ell|^2\,\frac{2\pi\delta}{\lambda}\,\frac{2\pi}{\lambda}}$$



rises again @ $\delta \sim \frac{\lambda}{10}$

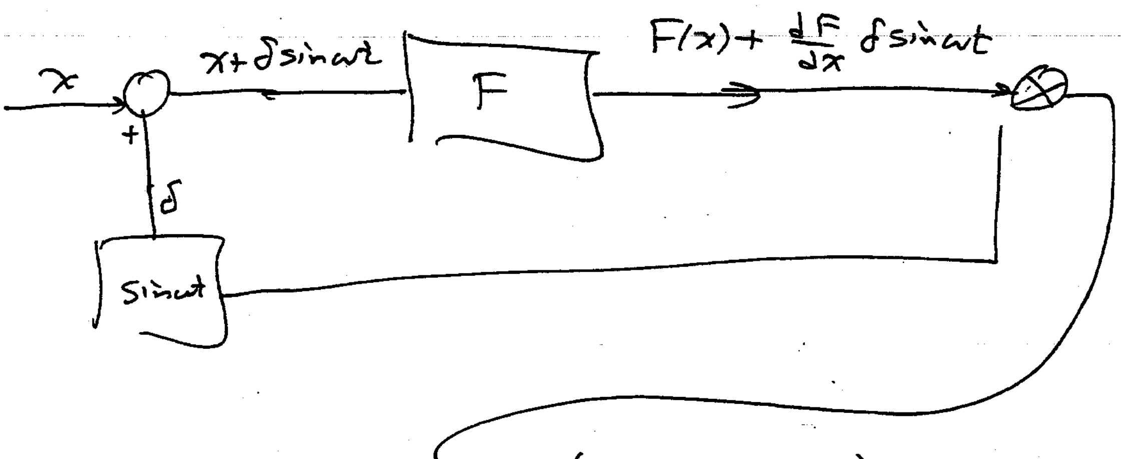
$\sqrt{2} \times$ min @ $\delta = \frac{\lambda}{\sqrt{8}\pi}\,\frac{|E_{cD}|}{|E_\ell|}$

8. Suppose the laser power fluctuates. This → false signal, since $L_P \propto |E_\ell|$. How deal with this?

9. Detour before answering.

a. Send x into a linear filter $\xrightarrow{x}\boxed{F(x)}\xrightarrow{F(x)}$

b. Put a dither on x



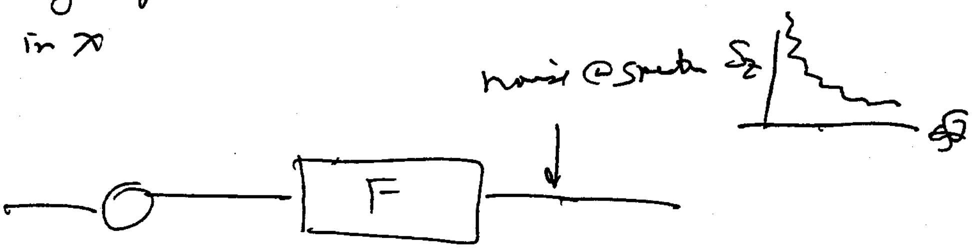
$$\left(F+\frac{dF}{dx}\delta\sin\omega t\right)\sin\omega t$$

$$= F\sin\omega t + \delta^2\frac{dF}{dx}\left(\frac{1}{2}+\frac{1}{2}\cos 2\omega t\right)$$

Send them a low-pass filter

$$\rightarrow \frac{1}{2}\delta^2\frac{dF}{dx}$$

c. Why useful? ... immune to ~~some~~ low frequency noise in x



i. This process gets rid of low frequency noise

ii

10. ~~Now~~ let $L_p = F(x)$, x = ~~motor~~ position of mirror, L_p: interferometer photocurrent

$\frac{dF}{dx} = \frac{dL_p}{dx}$ has finite slope @ dark fringe

Dither must be small compared to wavelength

11). For Interferometer the x is really $L_I - L_p$ (difference of arm lengths), and it doesn't matter which mirror we dither.

12. Key here is that we are operating @ dark fringe, where $dL_p/\Delta x = 0$ is independent of L_p. This gets rid of laser power fluctuations. – This is different from simply upconverting, ~~then down~~ via modulation, then demodulating ... difference is then $f(x)$ is nonlinear, Bob says =.)

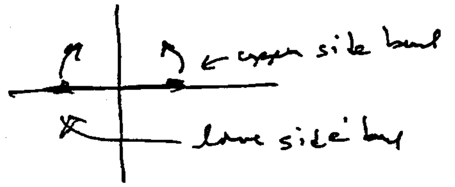
13. Note that dithering produces a series of side bands

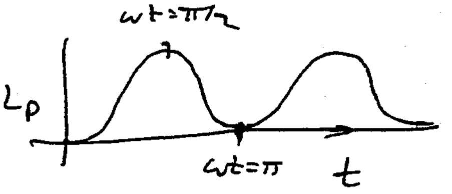
$$E_I = -\tfrac{1}{2}E_e\left[e^{ik\delta\sin\omega t}\right]$$

$$= -\tfrac{1}{2}E_e\Big[\underbrace{J_0(2k\delta)}_{\text{carrier}} + \underbrace{J_1(2k\delta)e^{i\omega t}}_{\text{upper side band}} + \underbrace{J_2(2k\delta)e^{-i\omega t}}_{\text{lower side band}}\Big]$$

14. Suppose there's no contrast defect; no GW; $E_A = 0$

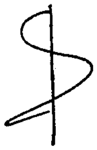
In terms of ~~[illegible]~~ phasors

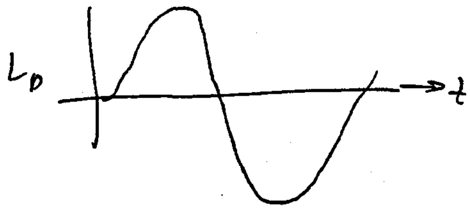




Now demodulate by $X\sin\omega t$

get

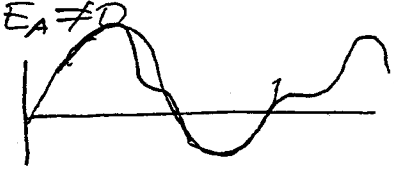




→ zero after the low pass filter

15. Suppose ~~[illegible]~~ we are off the dark fringe so $E_A \neq 0$

nonzero after low pass filter & positive



11. In homework, show that
get shot noise @ essentially same level

12. $E_{PD} = E_A + E_+ e^{i\omega t} + E_- e^{-i\omega t}$

$\rightarrow I_p = |E_{PD}|^2 = |E_A|^2 + |E_+|^2 + |E_-|^2$ DC

$+ 2\,\mathrm{Re}\left(E_A^* E_+ + E_-^* E_A\right) e^{i\omega t}$

$+ 2\,\mathrm{Re}\left(E_-^* E_+ e^{2i\omega t}\right)$

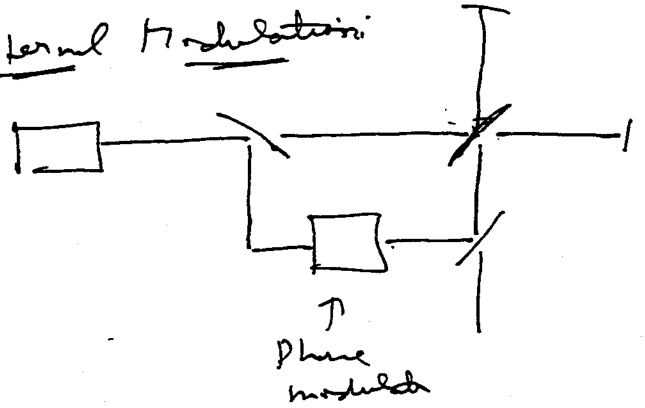
When (demodulate & low pass filter), see only → the ω term

@ $E_- = -E_+$ real, get

$$\frac{dV_{out}}{dx_p} = E_+ E_\ell \cdot (2k)$$ ... proportional to position of mirrors.

13. Methods of Modulation

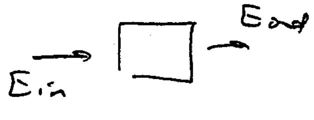
a. External Modulation:



b. Phase ~~modulator~~ modulator is a piezoelectric crystal

-- index of refraction $n(V)$ ← applied voltage

$V = V_0 \sin\omega t$



$E_{out} = E_{in}\, e^{-ikn_0 d}\, e^{-ik \frac{dn}{dV} V_0 \sin\omega t}$

c. get side bands on output light

- first side band do the job we want
- other side bands are ~~[illegible]~~ wasteful of power, but don't do damage.

d. Asymmetry Modulation

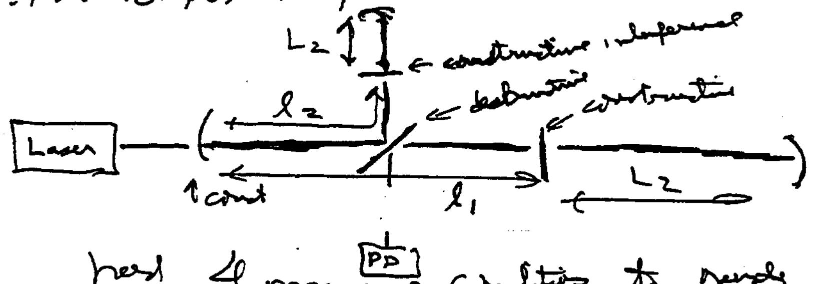
14. What happens if replace mirrors of Michelson Interferometer by cavities

- essentially the same as before
- get enhanced d phase / d (arm length)
- but also get ~~other~~ necessity to extract information about where mirrors are, and apply feedback.
    - more things to control

Martin Regehr : Optical Topologies 5/25/94

Part II

1. For simplest interferometer



$L_1 \simeq L_2 \simeq 4\,\mathrm{km}$

$l_1 \sim L_2 \sim$ meters

need 4 resonance conditions to operate

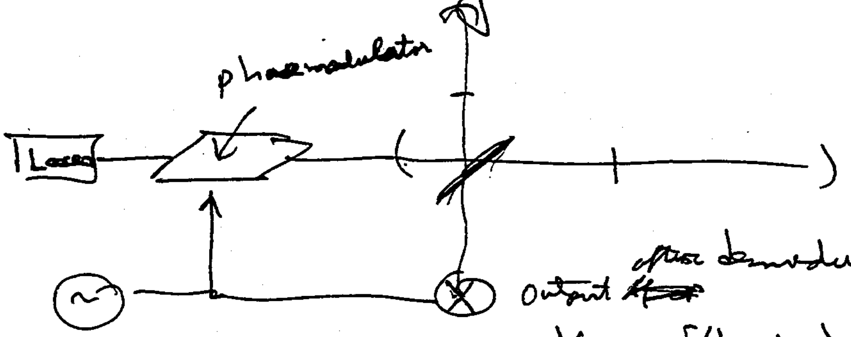
— corresponding to controlling $L_1, L_2,$ ~~l_1, l_2~~ $\frac{l_1+l_2}{2}$, l_1-l_2

2. Plan to Sense & control

$L_1-L_2,\ L_1+L_2,\ l_1-l_2,\ l_1+l_2$

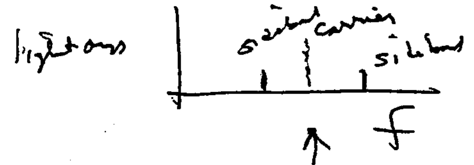
[can replace feedback to one of these by feedback to laser frequency]

3. L_1-L_2 is the GW signal



Output ~~[illegible]~~ after demodulating

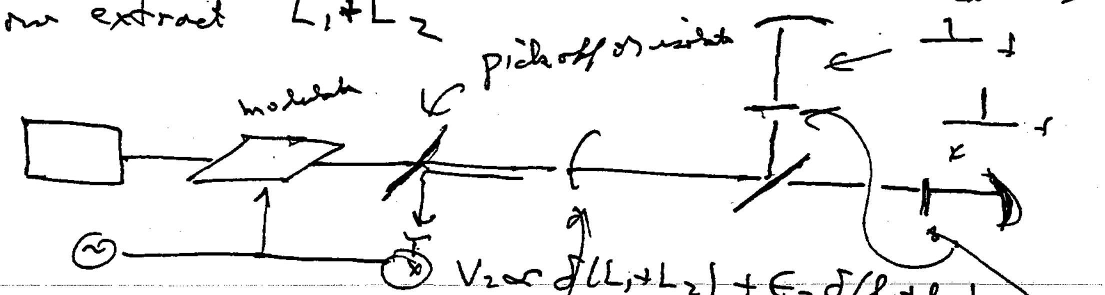
$V_1 \propto \delta(L_1-L_2) + \epsilon\,\delta(l_1-l_2)$



absent when no GW signal; GW produces some carrier

V_1

4. How extract L_1+L_2



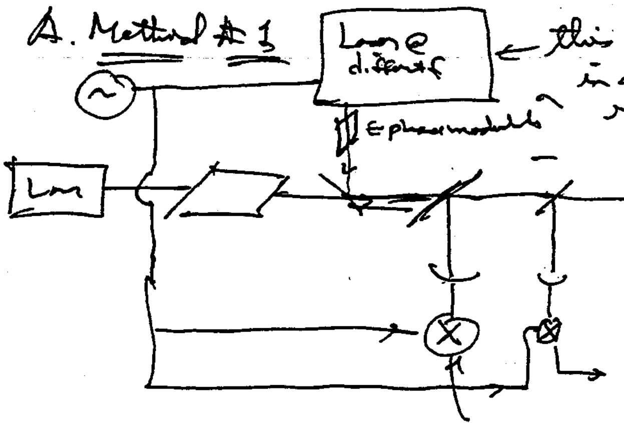
$V_2 \propto \delta(L_1+L_2) + \epsilon_2 \delta\underbrace{(l_1+l_2)}_{\text{recycling mirror location}}$

Same modulation, pick off @ different location

for a given distance change, far less effect than when move these mirrors

5. l_1+l_2, l_1-l_2

A. Method #1



this f chosen to not resonate in arm cavities, but does resonate in the main cavity (l_1 & l_2)

$\to V_3 \propto \delta(l_1-l_2)$

$\delta(l_1+l_2)$

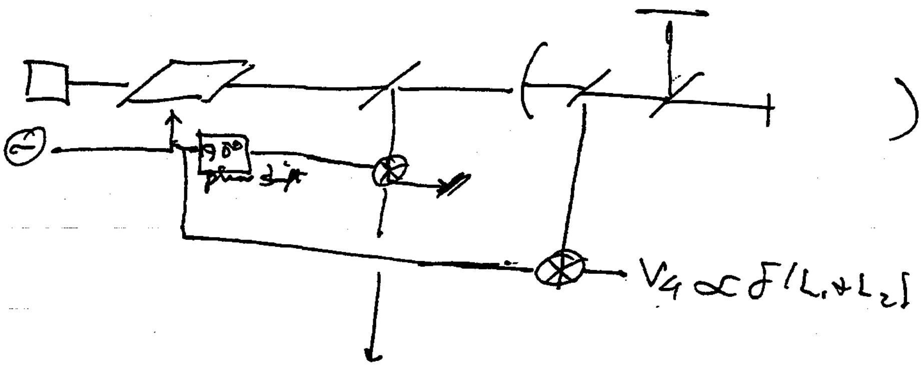
[choose mode cleaner length @ twice the average recycling cavity length]

~~B. Method~~ #2

→ Instead of using a second laser can divert some light from first laser, frequency shift it by a few 100 MHz

B. Method #2

Add a pickoff as before, ~~[illegible]~~ ∩



$V_4 \propto \delta(L_1+L_2) + \epsilon_4\,\delta(l_1+l_2)$

$V_3 \propto \delta(l_1-l_2) + \epsilon_3\,\delta(L_1-L_2)$

Both methods will be carried along for now, &
LIGO vacuum envelope will be ~~[illegible]~~
designed to handle both

BATCH START

12. Seismic Isolation

STAPLE
OR
DIVIDER

6 May 1994

LECTURE 12.

Seismic Isolation

Lecture by Lisa Sievers

Assigned Reading:

II. Leonard Meirovitch, *Elements of Vibration Analysis* (McGraw-Hill, 1986), pp. 48–58. [This reference develops the basic concepts of using mass-spring-damper systems for vibration isolation; and it discusses the measurement of vibrations, and two types of damping that can occur in mechanical systems: viscous damping and structural damping. These two types of damping will play an important role in the lectures on thermal noise next week.]

JJ. R. del Fabbro, A. di Virgilio, A. Giazotto, H. Kautzky, V. Montelatici, and D. Passuello, "Three-dimensional seismic super-attenuator for low frequency gravitational wave detection," *Physics Letters A*, **124**, 253–257 (1987). [This reference describes and analyzes an early version of the ambitious mass-spring-damper vibration-isolation stack that is being developed by the Pisa, Italy group as their prime contribution to the VIRGO Project. The analysis of the LIGO isolation stacks is similar, though their initial design is less ambitious.]

Suggested Supplementary Reading:

II. Leonard Meirovitch, *Elements of Vibration Analysis* (McGraw-Hill, 1986), pp. 39–48. [This is largely foundational material underlying the assigned reading (item 1. above); you may find it helpful.

KK. C. A. Cantley, J. Hough, and N. A. Robertson, "Vibration isolation stacks for gravitational wave detectors—Finite element analysis," *Rev. Sci. Instrum.*, **63**, 2210–2219 (1992). [This paper, by the Glasgow gravity-wave group, illustrates an isolation-stack analysis that is more sophisticated than the simple models used in class and in reference 2, and that reveals pitfalls in the design of a stack.]

LL. M. Stephens, P. Saulson, and J. Kovalik, "A double pendulum vibration isolation system for a laser interferometric gravitational wave antenna," *Rev. Sci. Instrum.*, **62**, 924–932 (1991). [This paper, passed out for other reasons in Lecture 10, analyses the use of compound pendula for vibration isolation.]

MM. L. Ju, D. G. Blair, H. Peng, and F. van Kann, "High dynamic range measurements of an all metal isolator using a sapphire transducer," *Mass. Sci. Technol.*, **3**, 463–470 (1992). [This paper, by the Perth resonant-bar gravitational-wave-detector group, describes a type of all-metal isolator which might be a precursor to an isolation stack for advanced LIGO detectors; see transparencies 20, 21, and 22 of Sievers' lecture.]

A Few Suggested Problems: See the next page.

1. You have the 2 stage spring/mass stack shown in Figure 1 and want to decide the best mass ratio m_1/m_2 in the 2 stages so that you achieve maximum isolation at frequencies well above $w_o^2=k_1/m_1$. A good designer would assume that the springs are compressed to their maximum limit in order to get the most bang for their buck, therefore the strain energy in each spring should be assumed equal ($k_1/m_1 = k_2/(m_1+m_2)$). Show that the transmissibility $X_1(f)/X_g(f)$ is maximized as the mass ratio m_1/m_2 goes to zero but that a point of diminishing returns is reached when the ratio is about 1.

2. Work out the equations of motion for the 1 and 2 stage pendula shown in Figure 2. Compare the amount of isolation achieved at two different frequencies: $\omega = 2\sqrt{\frac{g}{l}}$ and $\omega = 10\sqrt{\frac{g}{l}}$

3. A method for mechanically damping a high Q mechanical resonance is to use a "proof mass damper" as shown in Figure 3. The proof mass damper is a damped oscillator whose mass is much smaller than the mass to be damped and whose resonant frequency and damping coefficient is tuned specifically to damp the system in the most effective way. Assume $m_1=20m_2$, f_1=4Hz, and $f_2=(\frac{1}{1+m_2/m_1})f_1$. Plot the transmissibility function $X_1(f)/X_g(f)$ for 3 different damping coefficients, c. [Definition of damping coefficient, c: If a particle of mass m moves under the combined influence of a linear restoring force -kx and a resisting force $-c\dot{x}$, the differential equation which describes the motion is $m\ddot{x} + c\dot{x} + kx = 0$ .... c is inversely proportional to Q]

   1. c=0
   2. c=infinity
   3. $c=2m_2(2\pi f_1)\sqrt{\frac{3m_2/m_1}{8(1+m_2/m_1)^3}}$

   The third case is the case where you get the maximum attenuation possible (i.e. $X_1(f_1)/X_g(f_1)=\sqrt{1+2m_1/m_2}$)

   [A proof mass damper has been experimentally implemented in Mark I. One of the stacks (i.e. optics plate mounted on rubber), had a high Q horizontal resonance at f_1=4Hz. In a compact vacuum sealed vessel, we built a pendulum whose bob was 1/20 the mass of the offending optics plate. The pendulum was partially submerged in motor oil whose damping coefficient was given in (3). The length of the pendulum bob was tuned to the resonant frequency of f_2. The stack resonance was damped without compromising the isolation at higher frequencies.]

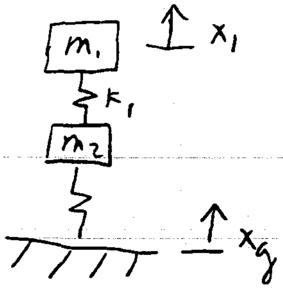


Figure 1

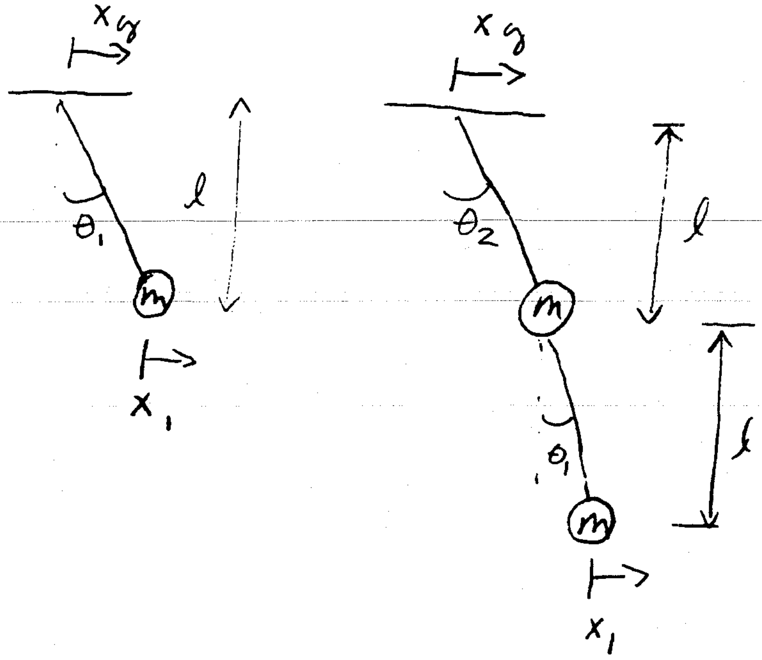
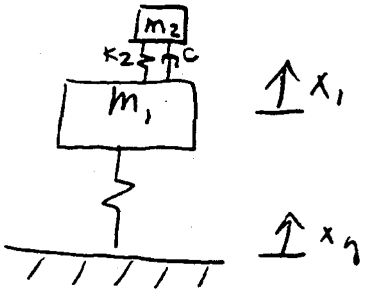


Figure 2



$$2\pi f_1 = \sqrt{\frac{k_1}{m_1}}$$

$$2\pi f_2 = \sqrt{\frac{k_2}{m_2}}$$

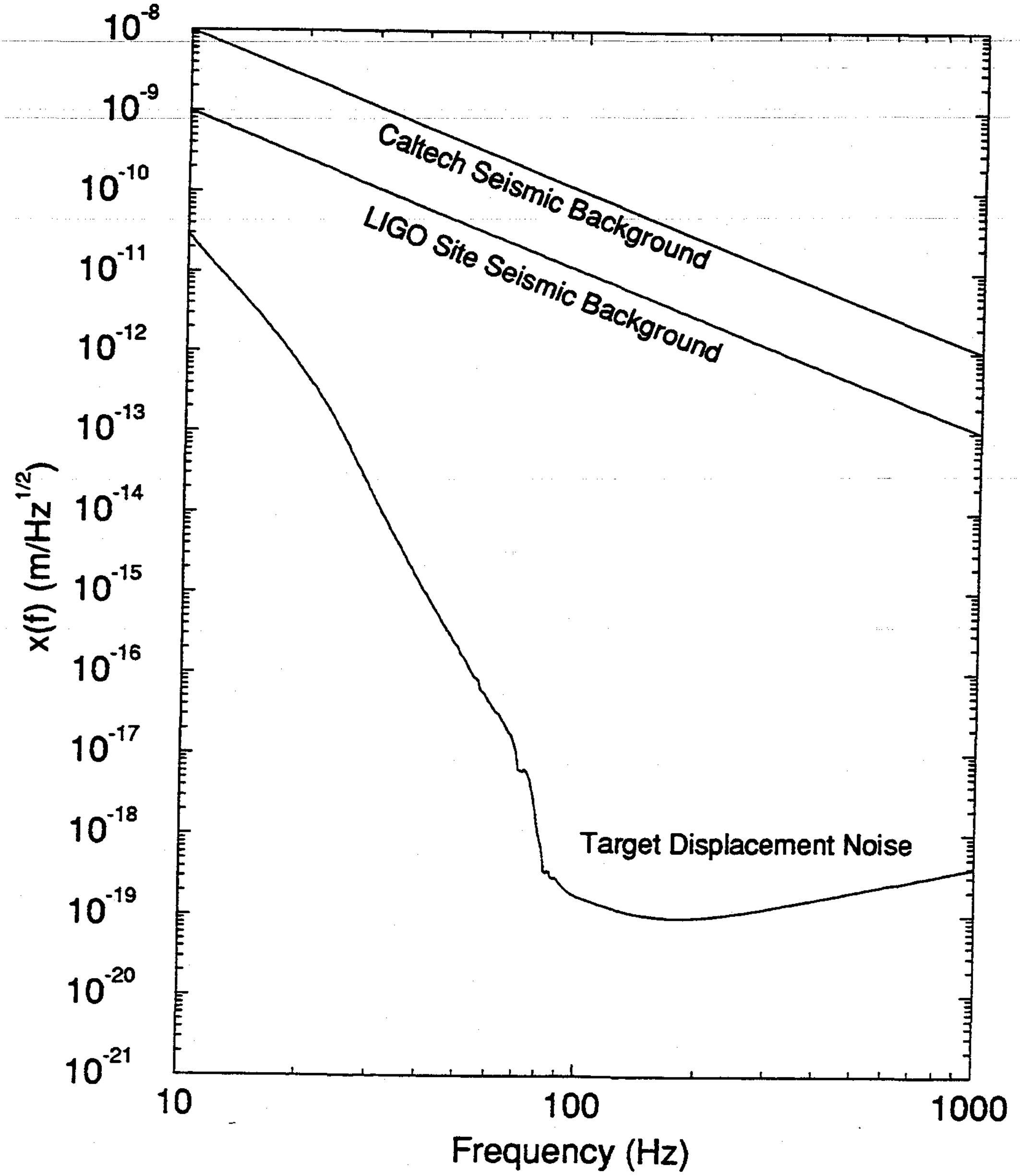
Figure 3

Lecture 12
Seismic Isolation

by Lisa Sievers, 6 May 1994

Sievers lectured from the following transparencies.

LIGO DISPLACEMENT NOISE

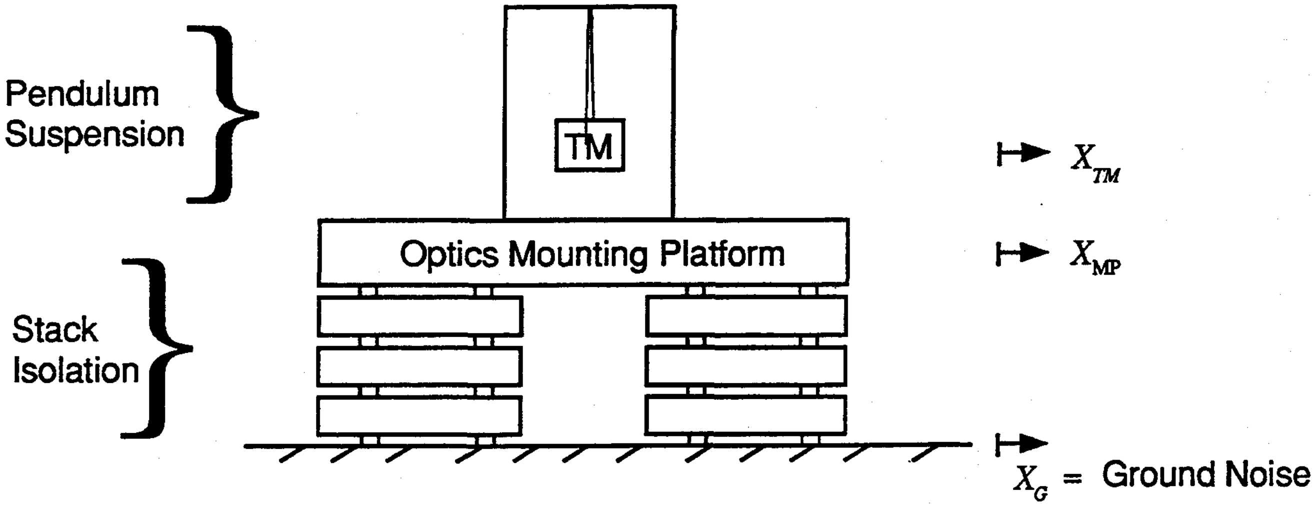


SEISMIC ISOLATION OF TEST MASS

Seismic Isolation of test mass is composed of 2 components:

- **Stack Isolation**
- **Pendulum Suspension**

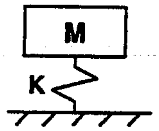
The ground noise at the sites drives the requirement on the amount of isolation necessary



LIGO

CONCEPTS FOR DESIGNING SPRING/MASS PASSIVE ISOLATION SYSTEMS

MODEL OF
1 LAYER STACK }

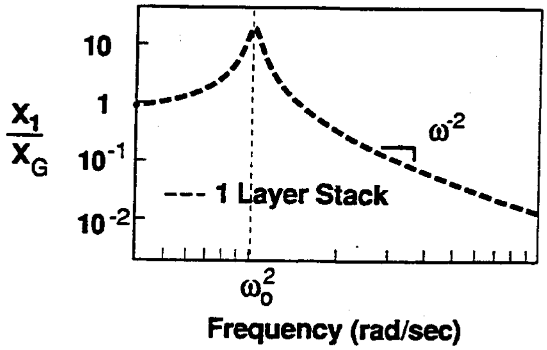


X_1 = Displacement of mass

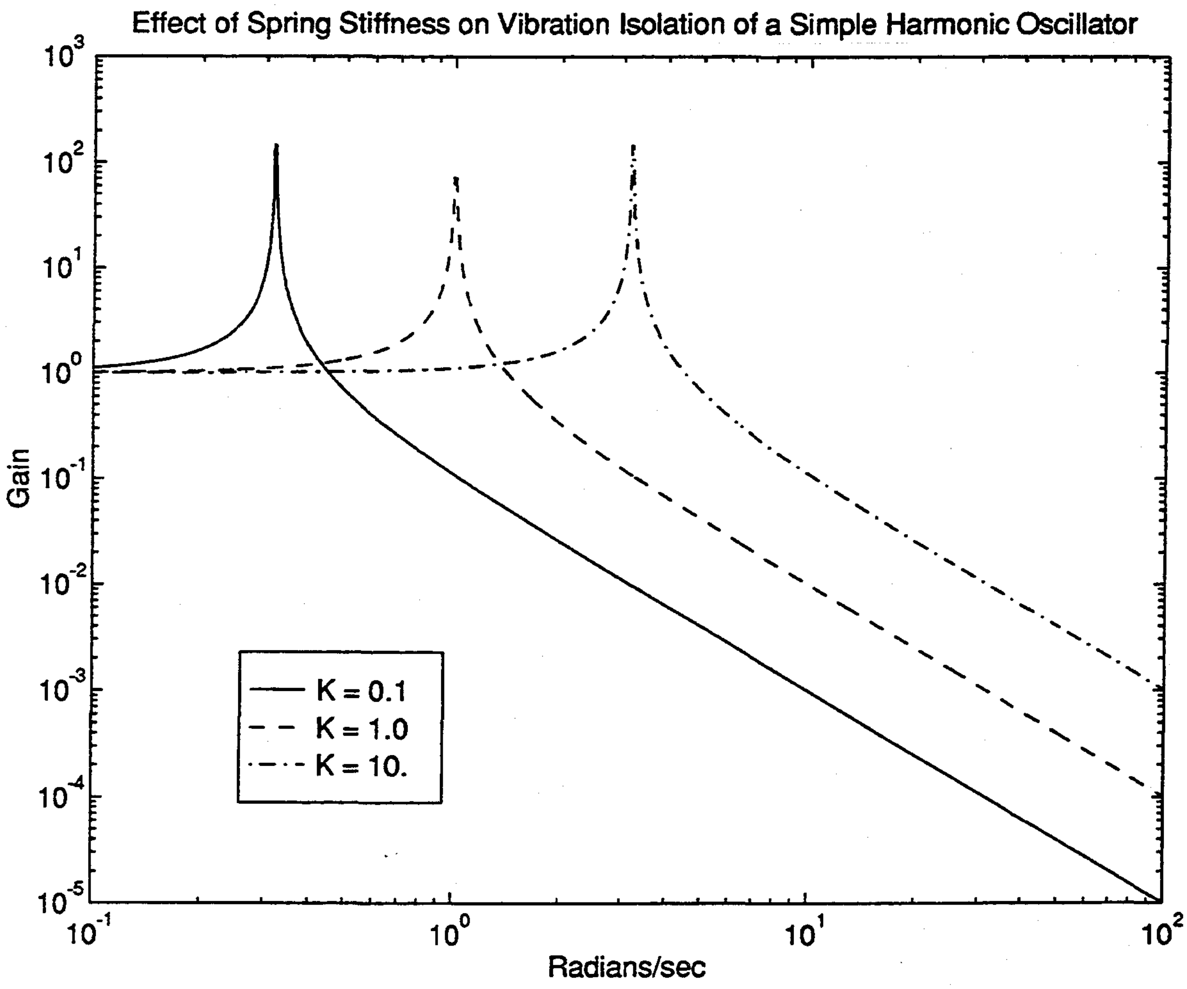
X_G = Ground noise displacement

- MEASURE OF ISOLATION IS THE TRANSMISSIBILITY FUNCTION:

$$\frac{X_1}{X_G} = \frac{K/M}{(j\omega)^2 + K/M}$$

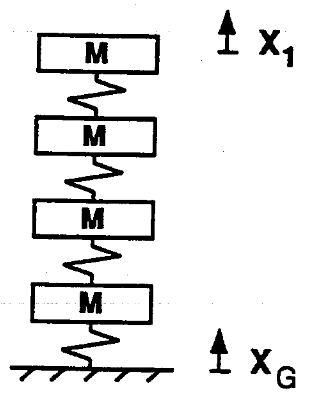


- $\omega_o^2 = K/M$
- As $\omega \to \infty$ slope is ω^{-2}

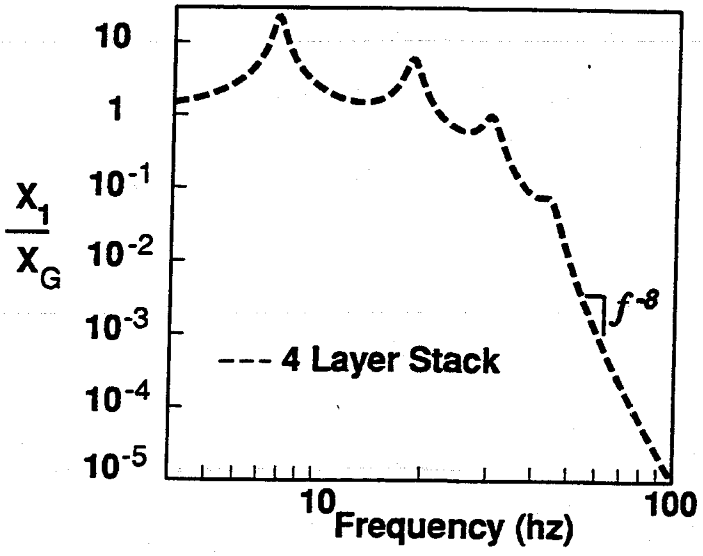

Effect of Spring Stiffness on Vibration Isolation of a Simple Harmonic Oscillator
Gain
10^3
10^2
10^1
10^0
10^-1
10^-2
10^-3
10^-4
10^-5
10^-1
10^0
10^1
10^2
Radians/sec
K = 0.1
K = 1.0
K = 10.

- **EFFECT OF MULTIPLE LAYERS:**

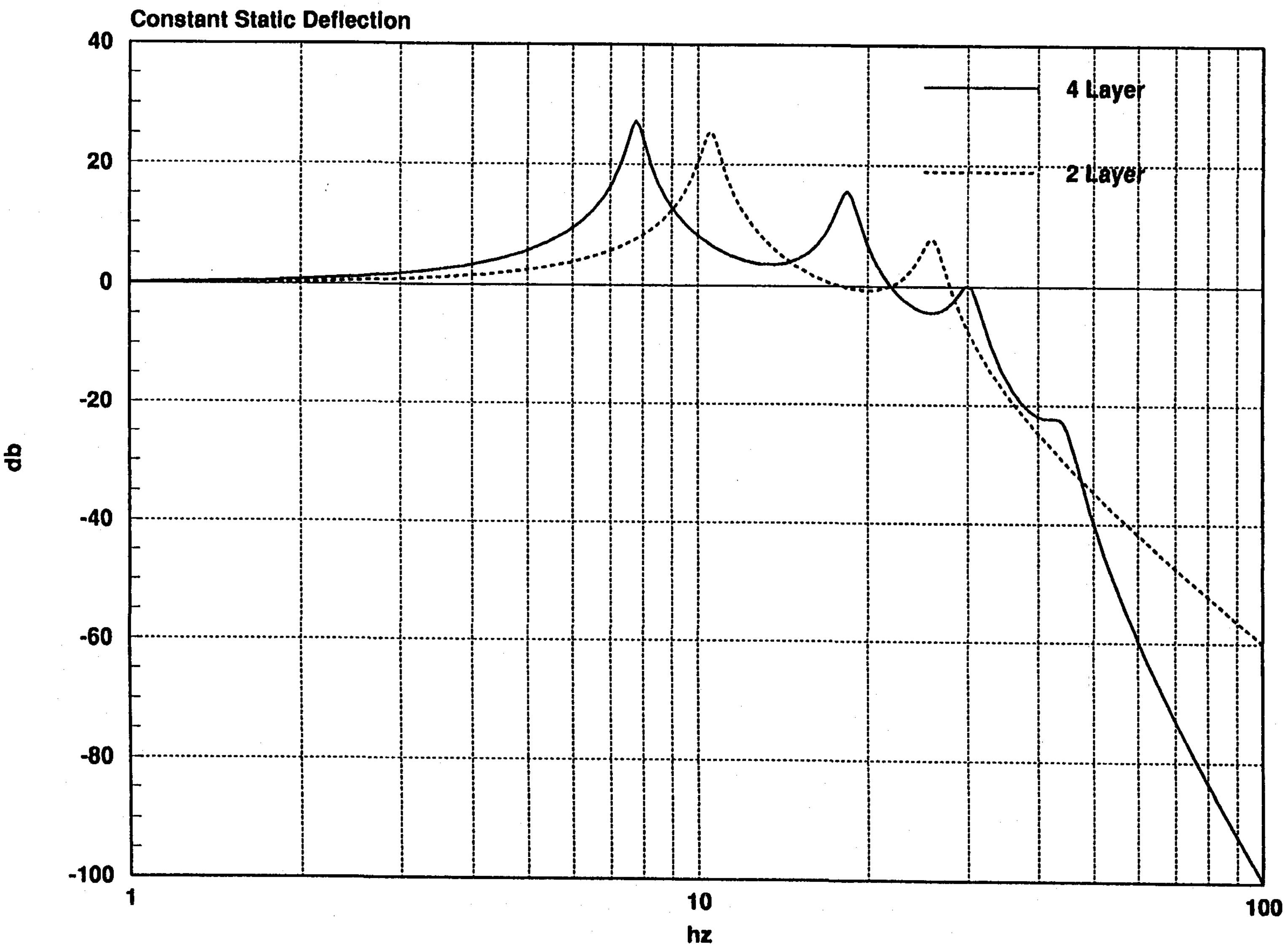
  $\left| \frac{X_1}{X_G} \right|$ **ROLLS-OFF AS** $\omega^{-2\times N}$ **WHERE N IS THE NUMBER OF LAYERS.**



Simple Model of Mark 2 Stack Isolation (vertical)



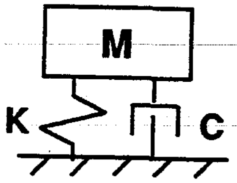
- **WHY NOT USE A STACK WITH MANY MANY LAYERS TO MAXIMIZE ROLL-OFF?**



4 and 2 Layer Stacks with silicone springs

5a

EFFECTS OF DAMPING ON ISOLATION



↑ X_1 = Displacement of mass

↑ X_G = Ground noise displacement

- **TRANSMISSIBILITY FUNCTION:**

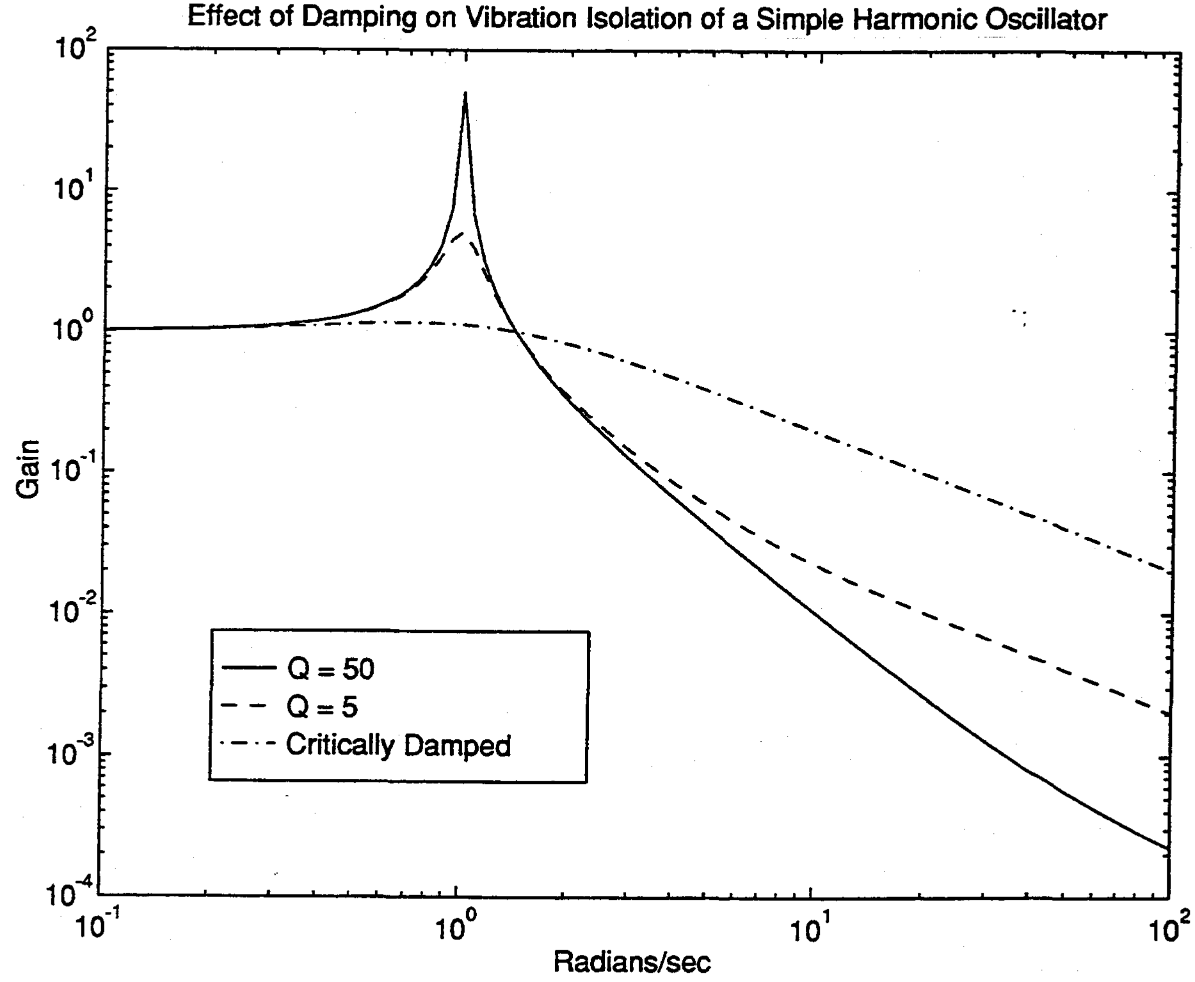
$$\frac{X_1}{X_G} = \frac{K/M + C/M}{(j\omega)^2 + C/M(j\omega) + K/M} = \frac{\omega_o^2 + \frac{\omega}{Q}}{(j\omega)^2 + \frac{\omega_o}{Q}(j\omega) + \omega_o^2}$$

$$\omega_o^2 = K/M$$

$$Q = \frac{\omega_o}{C/M}$$

As $\omega \rightarrow \infty$ slope is ω

- **WHY DO WE NEED DAMPING IN STACK?**
  - In prototypes, damping is essential; seismic noise at Caltech and MIT is high enough that cavities would be much more difficult to lock and stay in lock
  - May also be nonlinear coupling into gravity wave signal due to seismic peak motion
  - Have no verification for minimum damping required in LIGO but believe that Q < 10 is more than adequate


Effect of Damping on Vibration Isolation of a Simple Harmonic Oscillator
10^2
10^1
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
Gain
10^{-1}
10^0
10^1
10^2
Radians/sec
Q = 50
Q = 5
Critically Damped

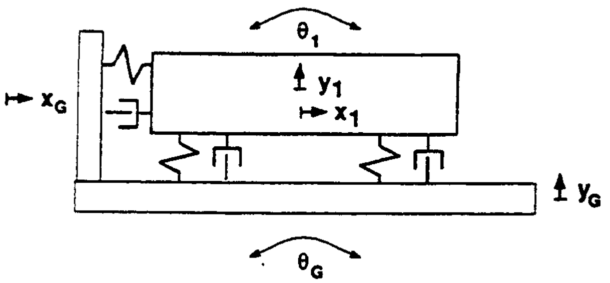
CONSIDERATIONS FOR CHOOSING SPRING/DAMPER MATERIAL

1. VACUUM COMPATIBILITY?
2. DAMPING?
3. STIFFNESS?
4. LOAD (TOTAL LOAD IN MARK II IS OVER A TON)

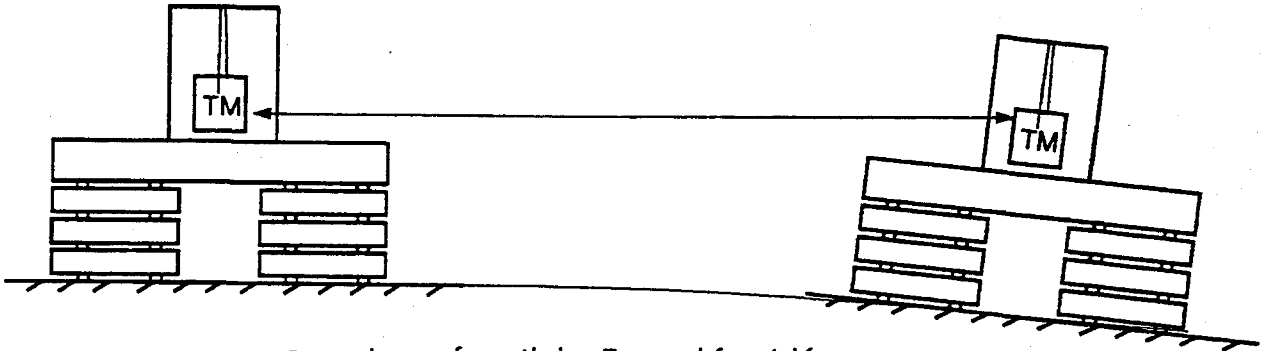
- METAL SPRINGS ARE VACUUM COMPATIBLE BUT DON'T PROVIDE MUCH DAMPING
- ELASTOMER SPRINGS PROVIDE DAMPING BUT NEED TO BE SPECIALLY PROCESSED BEFORE THEY ARE VACUUM COMPATIBLE
- ELASTOMERS HAVE NICE PROPERTY THAT DAMPING IS FREQUENCY DEPENDENT; LOTS OF DAMPING AT LOW FREQUENCIES AND LITTLE DAMPING AT HIGHER FREQUENCIES
  - RTV: FLEXIBLE (ABOUT 40% DEFLECTION) BUT LITTLE DAMPING
  - VITON: STIFFER (ABOUT 20 % DEFLECTION) BUT MORE DAMPING
- PICKED A SIZE FOR THE SPRINGS SO COULD LOAD WITH 55 KG PER SPRING

REAL SPRING/MASS ISOLATORS MAY BE MORE DIFFICULT TO MODEL

- REAL SYSTEMS ARE 6 DIMENSIONAL, NOT 1; TILTS AND TRANSLATIONS COUPLE INTO TEST MASS MOTION

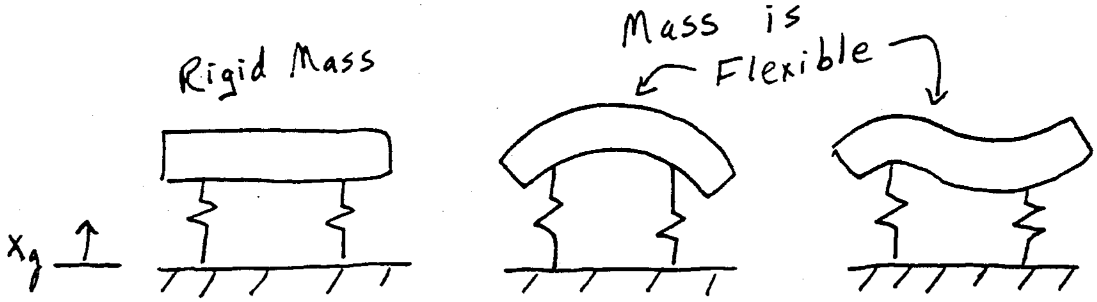


- Why be concerned with vertical isolation?



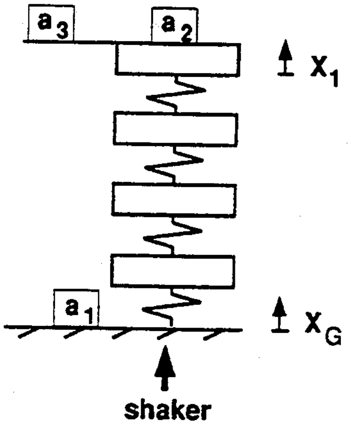
curvature of earth is .5 mrad for 4 Km

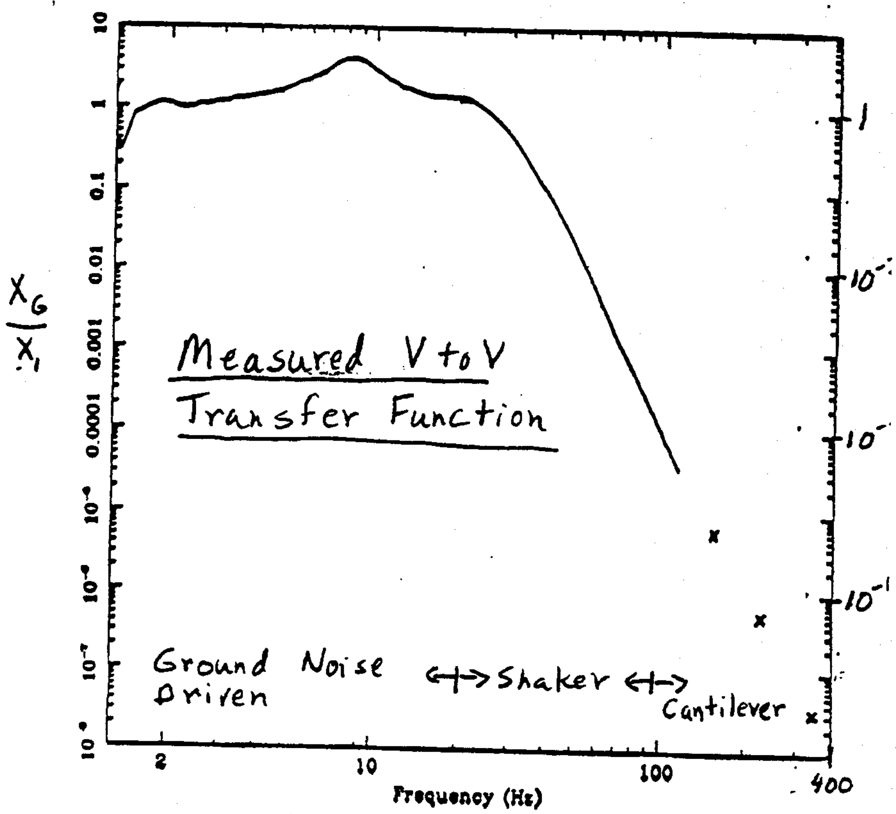
- SPRINGS AND DASHPOTS ARE NOT NECESSARILY "BEST MODEL" FOR REAL SPRING AND DAMPING ELEMENTS
- MASSES ARE NOT RIGID AT ALL FREQUENCIES



Measuring Stack Transfer Functions

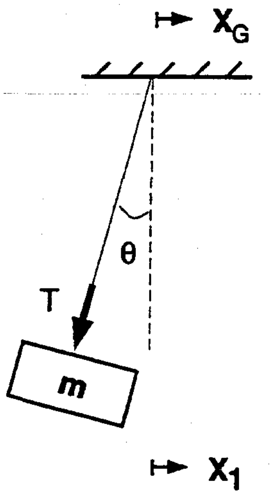
- Drive base of stack with shaker and measure ratio between a1 and a2
- Above about 40 Hz, signal in a2 is mainly acoustic pickup so must do measurements in vacuum
- Above about 100 Hz (10^{-5} attenuation) signal is mainly sensor electronics noise
- To get higher frequency points use mechanical amplifier

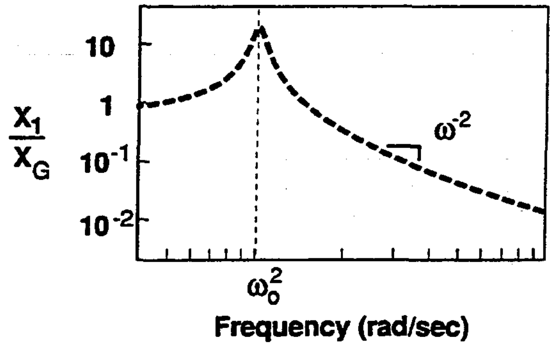




LIGO

PENDULA AS VIBRATION ISOLATORS





$$\omega_o^2 = g / l$$

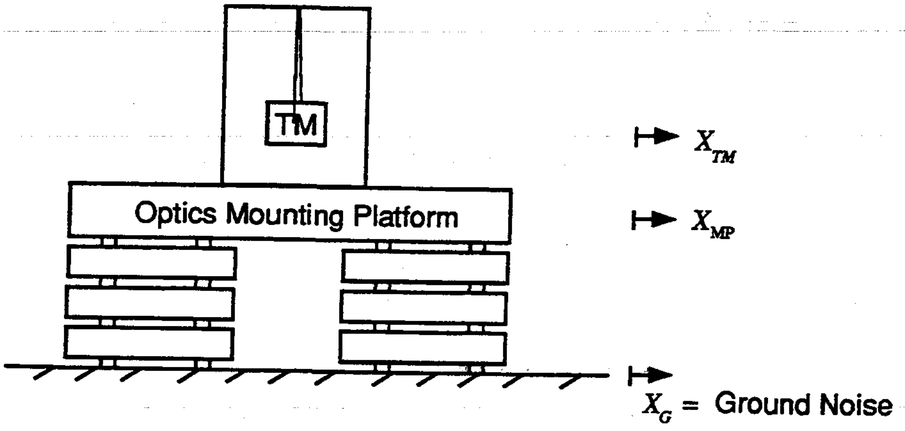
TRANSMISSIBILITY FUNCTION:

$$\frac{X_1}{X_G} = \frac{g/l}{(j\omega)^2 + g/l}$$

Effects of :

- **shortening pendulum ⇒ same as stiffening spring**
- **adding damping (e.g. air) ⇒ same as in spring/mass case; roll-off varies between ω^{-1} and ω^{-2} depending on level of damping**
- **pendulum in series ⇒same as adding more layers (get $\omega^{-2\times N}$ roll-off where N is the number of pendulum stages)**

CALCULATING DIRECT TRANSMISSION OF GROUND MOTION TO TEST MASS MOTION



- MEASURE POWER SPECTRAL DENSITY OF GROUND MOTION:

$$\mathrm{X_G(f)}$$

- MEASURE STACK TRANSFER FUNCTION:

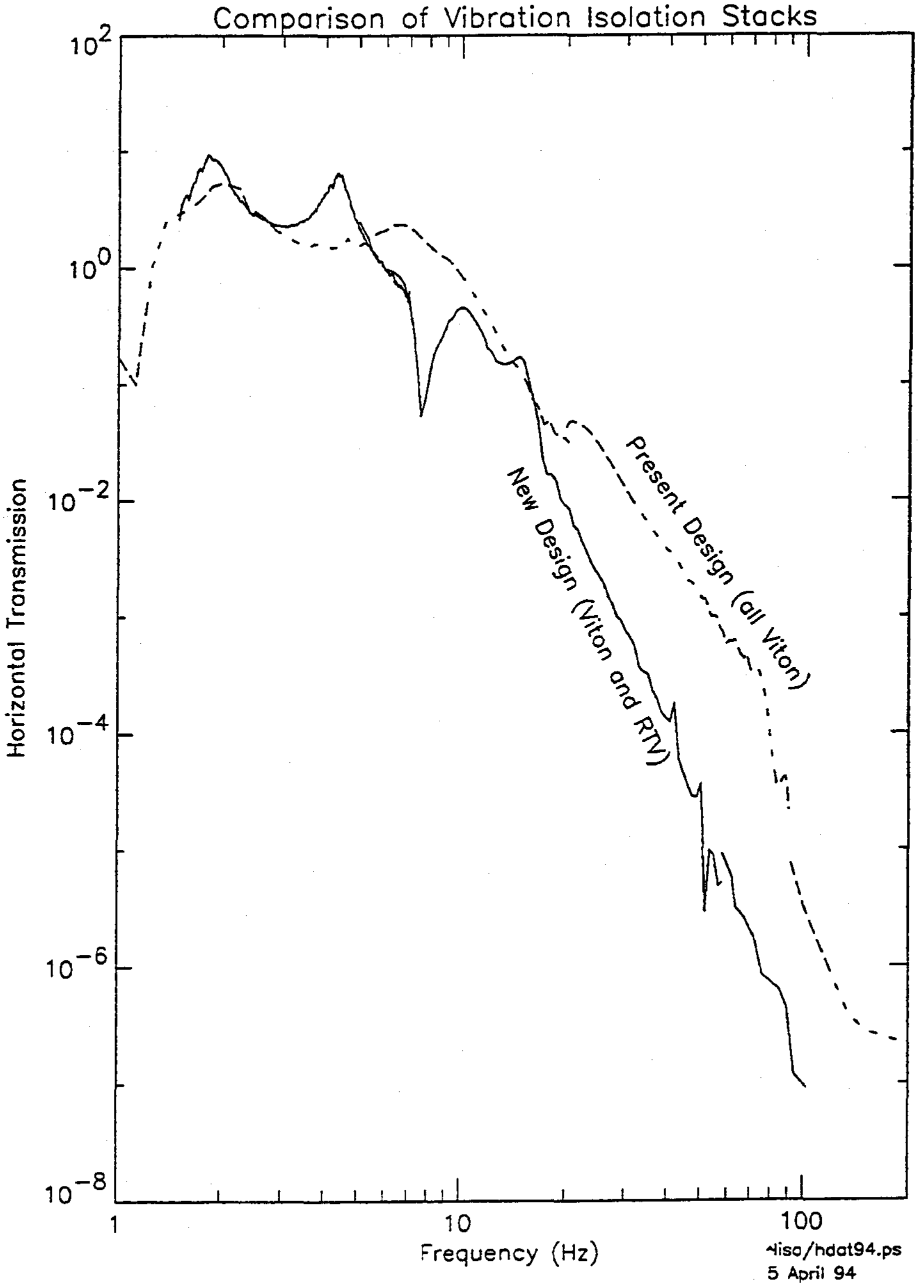
$$\frac{\mathrm{X_{MP}(f)}}{\mathrm{X_G(f)}}$$

- MEASURE PENDULUM TRANSFER FUNCTION:

$$\frac{\mathrm{X_{TM}(f)}}{\mathrm{X_{MP}(f)}}$$

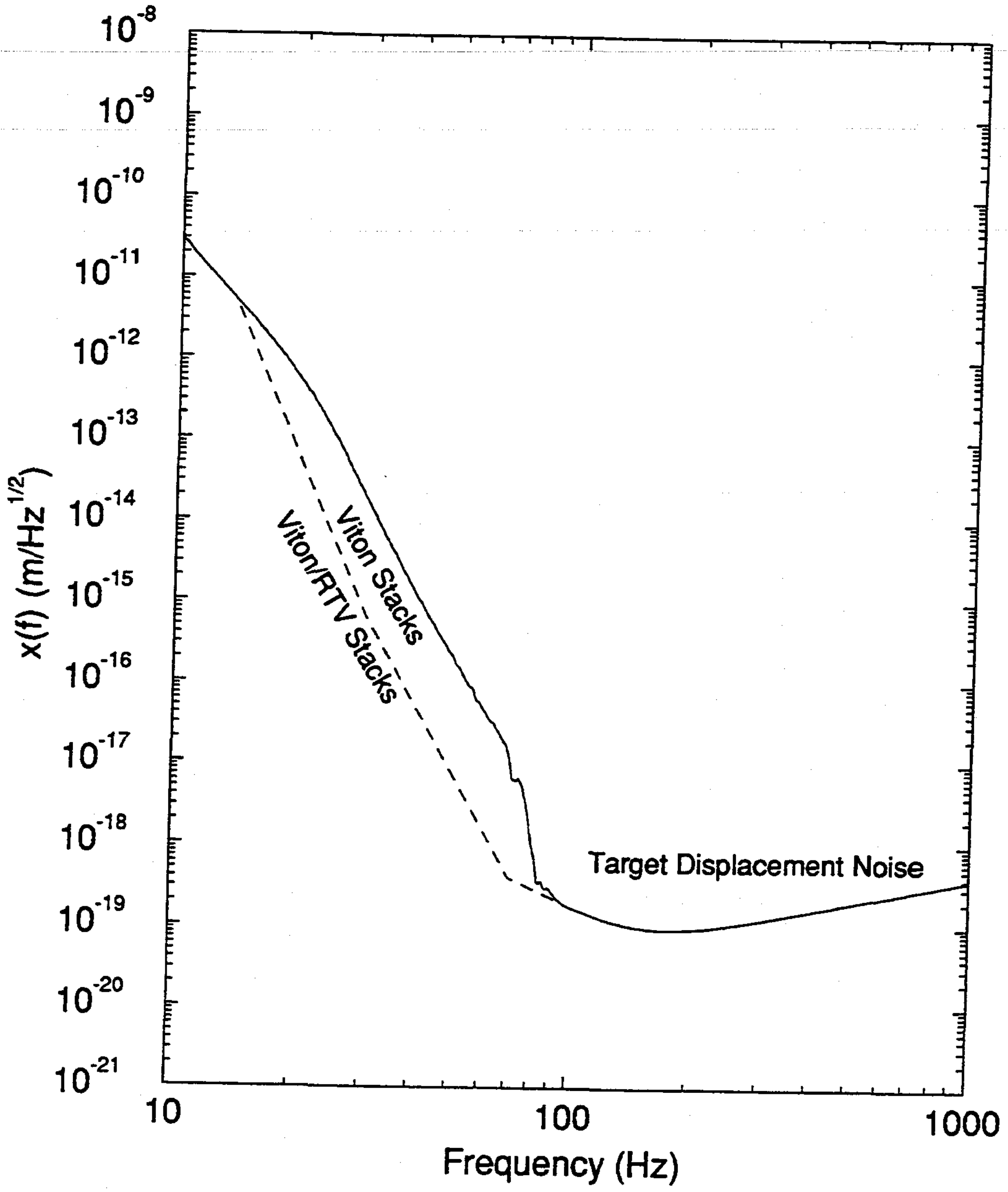
$$\mathrm{X_{TM}(f)} = \frac{\mathrm{X_{TM}(f)}}{\mathrm{X_{MP}(f)}}\frac{\mathrm{X_{MP}(f)}}{\mathrm{X_G(f)}}\mathrm{X_G(f)}$$

LIGO



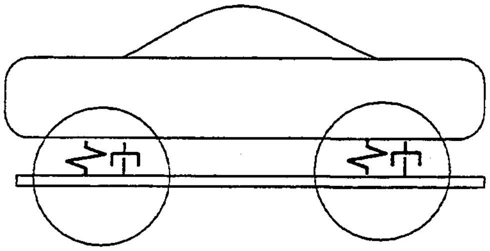
~iso/hdat94.ps
5 April 94

LIGO DISPLACEMENT NOISE



OTHER APPLICATIONS FOR VIBRATION ISOLATION

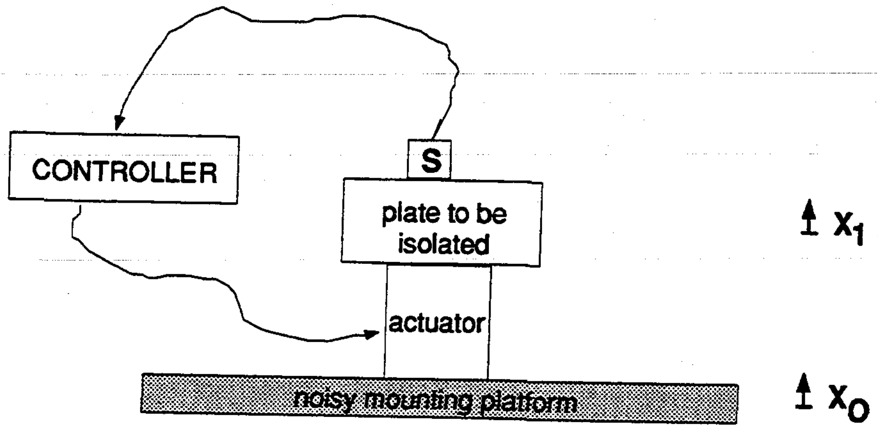
1. CAR SUSPENSION



2. MACHINERY RAFTS
3. DENTAL DRILLS
4. Many many others

LIGO

METHODS FOR FUTURE VIBRATION ISOLATION SYSTEMS



ACTIVE VIBRATION ISOLATION

WHERE SHOULD ACTIVE CONTROL BE APPLIED?

1. ON TEST MASS DIRECTLY

   NO NO NO!!!! Have to reduce seismic noise before get to test mass or can't detect gravity wave signal

2. AT SUSPENSION POINT

   Need a very sensitive sensor. Have to worry about tampering with Q of suspension. No work in progress

3. ACTIVE STACKS

   Need a very good 6–D model of stack to design controller (think you are driving translations but really driving tilts). Work in progress at JILA

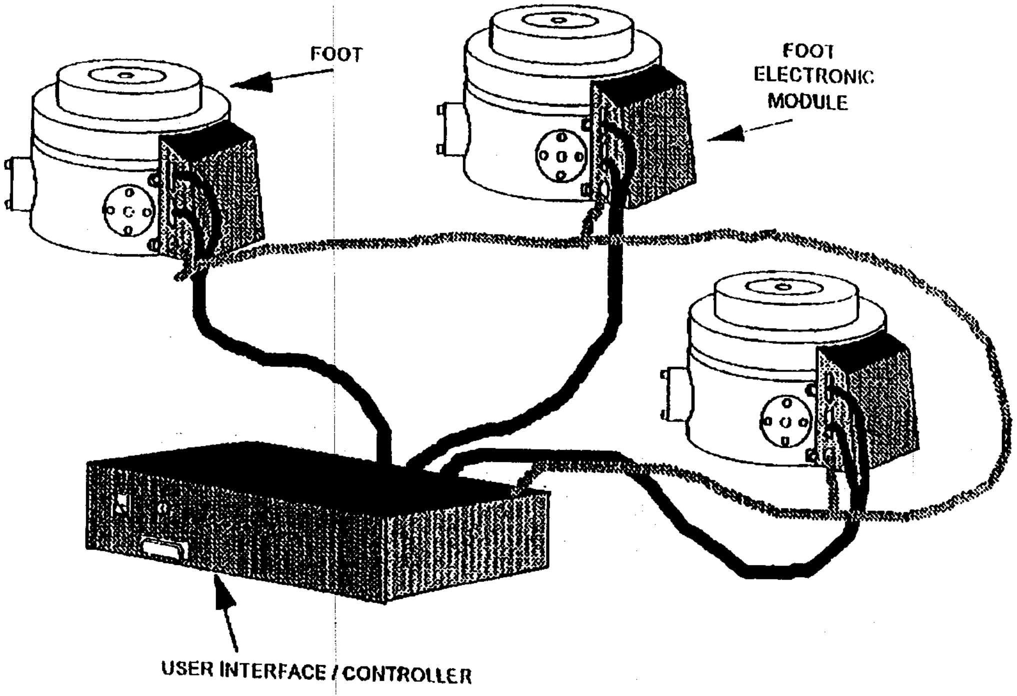
4. ISOLATION OUTSIDE VACUUM AT SUPPORT POINTS

   Don't have to worry about vacuum compatibility issues. Actuators need to support loads in the tons with 10 micron stroke. Work in progress at MIT

LIGO

Bedrock

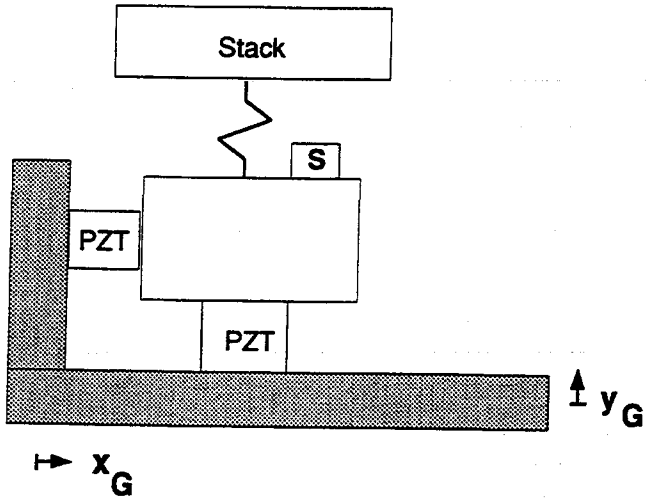
System Configuration



17

BARRY CONTROLS.
A UNIT OF APPLIED POWER INC.

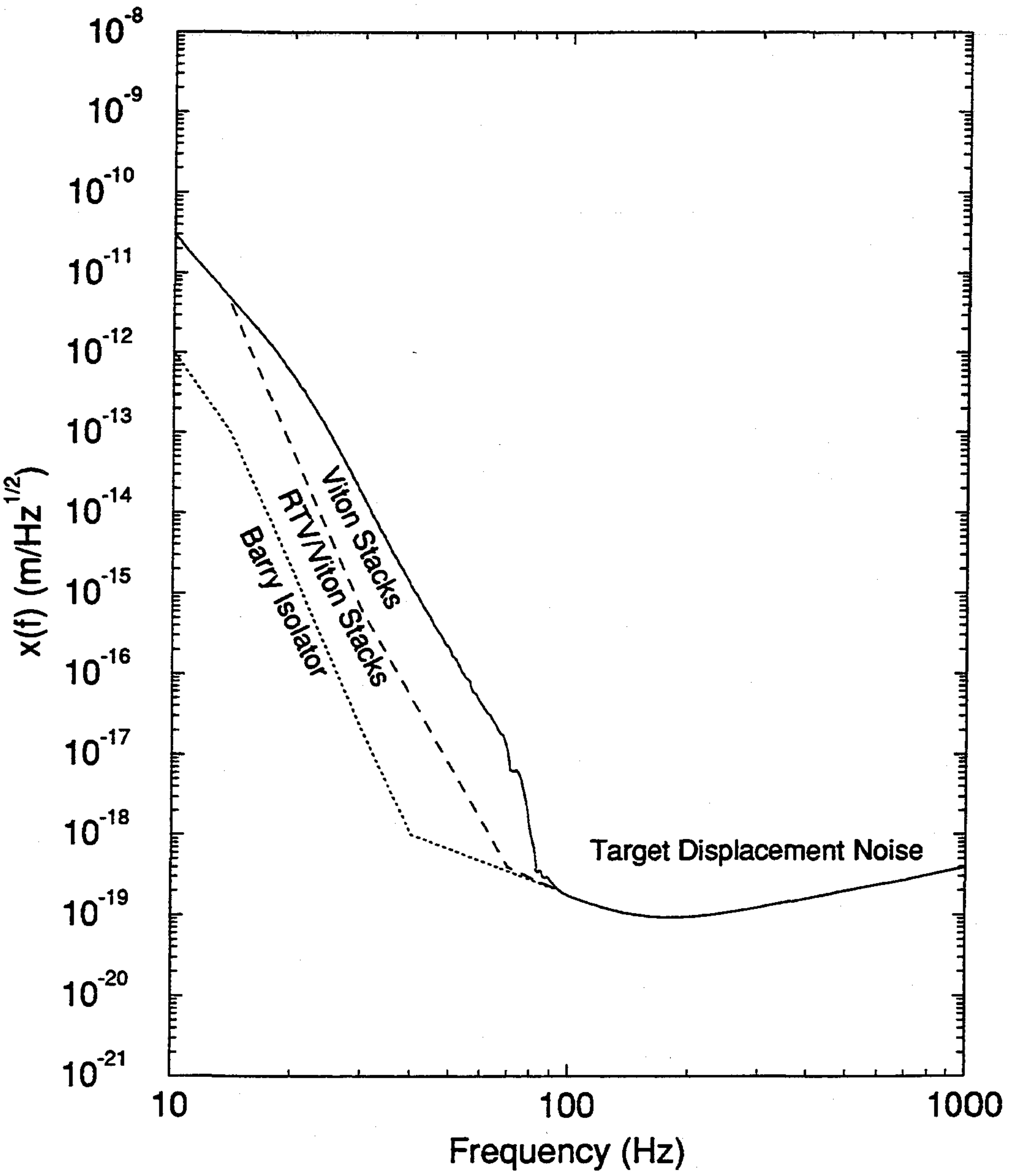
ACTIVE ISOLATION OUTSIDE VACUUM AT SUPPORT POINTS (BARRY MOUNTS)



- Sensor and PZT actuator pair for each of 3 translational degrees of freedom
- Provides about a factor of 30–100 isolation between 3 Hz and 100 Hz

LIGO

LIGO DISPLACEMENT NOISE



PASSIVE METHODS (POSSIBLY COMBINED ACTIVE) FOR FUTURE VIBRATION ISOLATION SYSTEMS

- 5–STAGE PENDULUM HORIZONTAL ISOLATION (VIRGO PROJECT)
- 5–STAGE BLADE SPRING ISOLATION FOR VERTICAL (AUSTRALIAN PROJECT)

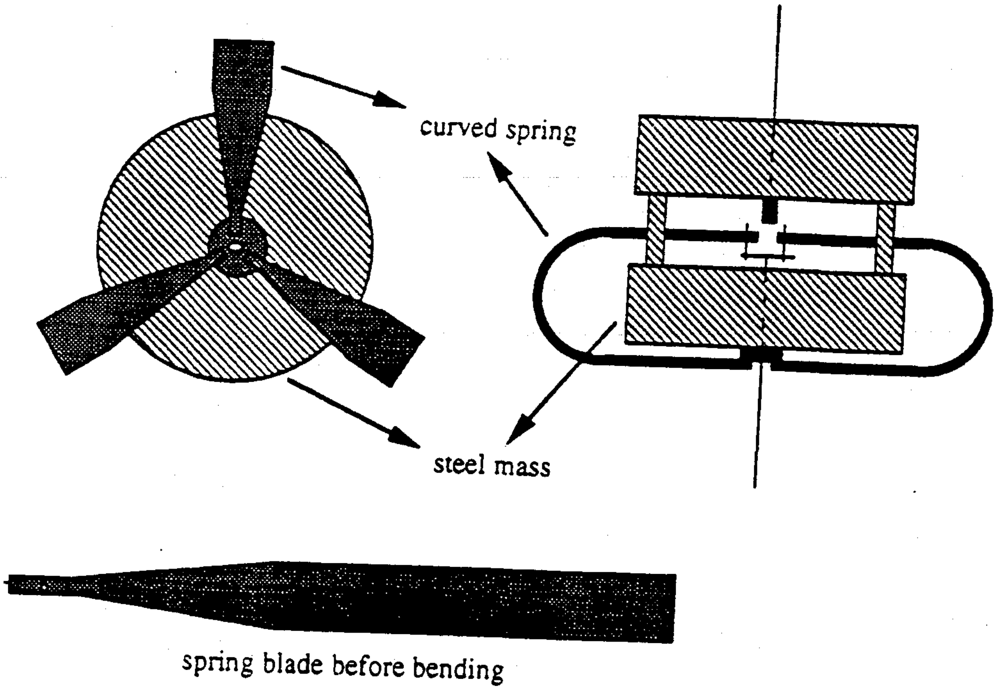


Figure 2. Configuration of 1 isolator element.

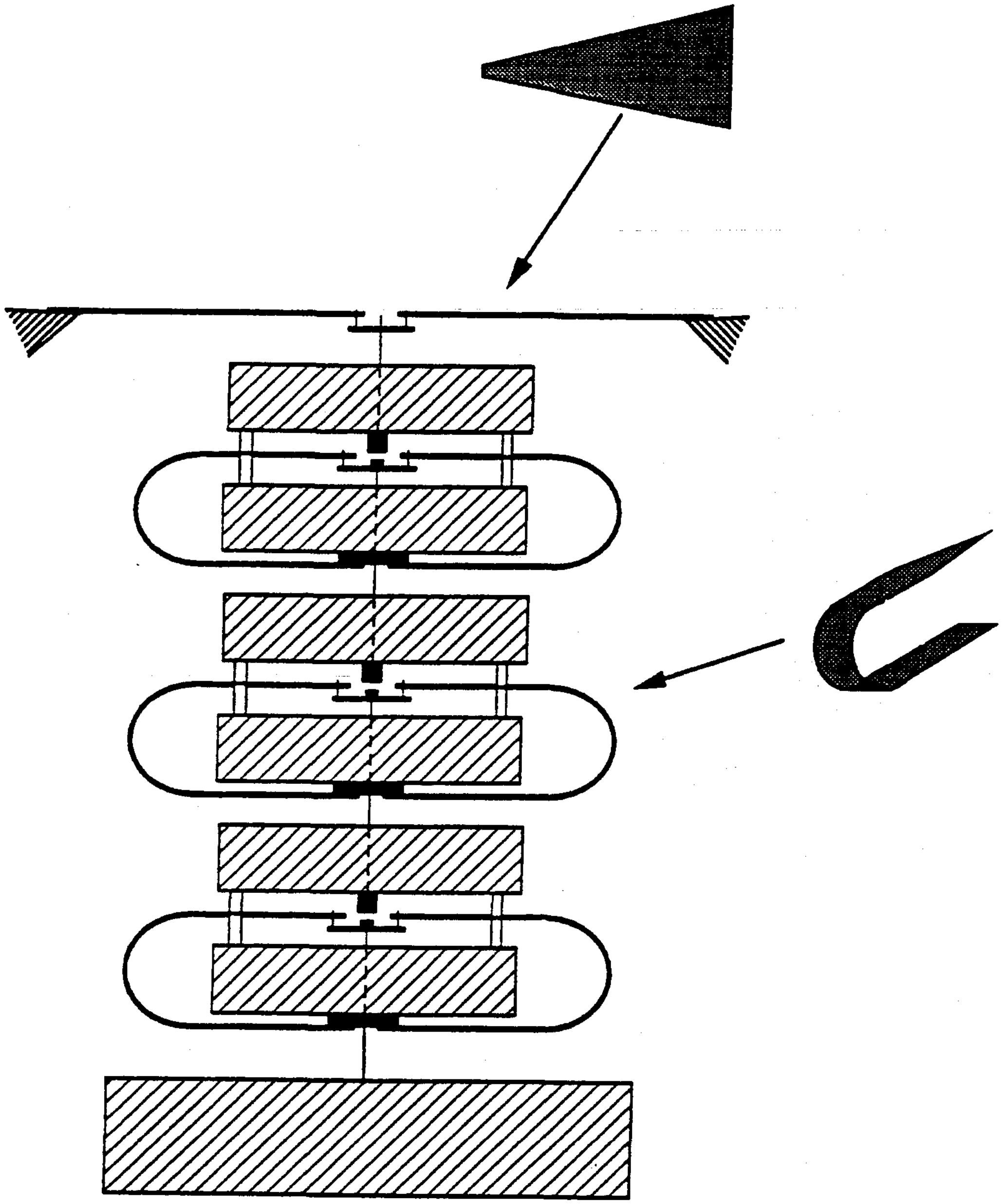
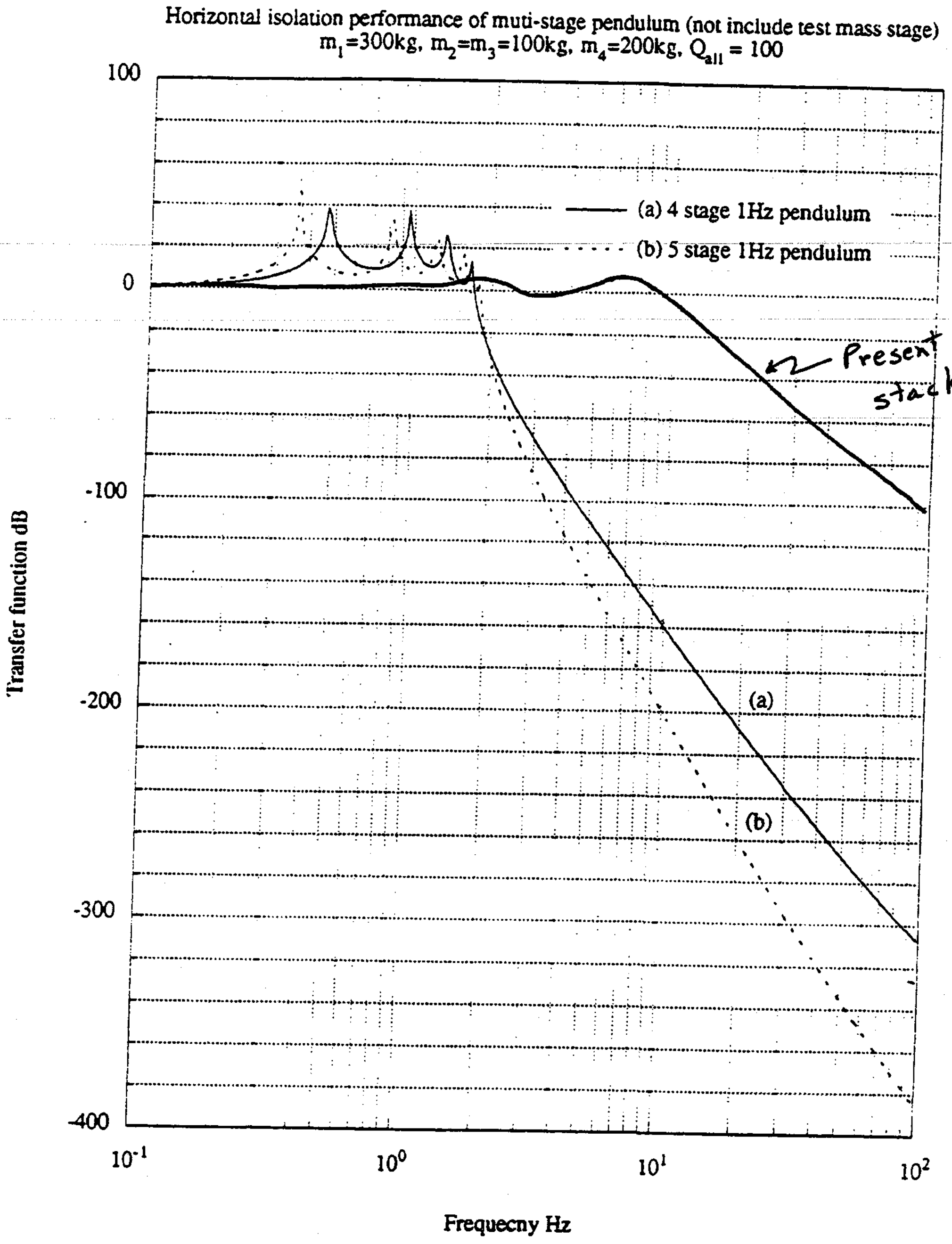


Figure 3. Four element stack.



Present LIGO all-vi
stack design

Figure 5. Comparison of horizontal isolation performance for a 4 and 5 element stack design.

LIGO

BATCH START

13. & 14. Thermal Noise

STAPLE
OR
DIVIDER

11 & 13 May 1994

Ph 103c: The Physics of LIGO

LECTURES 13 & 14

Thermal Noise

Lectures by Aaron Gillespie

Assigned Reading:

NN. Herbert B. Callen and Theodore A. Welton, "Irreversibility and generalized noise," *Phys. Rev.* **83**, 34–40 (1951). [This paper derives a generalized version of the fluctuation-dissipation theorem (Nyquist's theorem), cf. Lecture 2. The terminology and notation are quite different from what is now standard. Equations (4.8) in the quantum regime and (4.11) in the classical limit describe the mean square value $\langle V^2\rangle$ of the fluctuating "generalized force" V in terms an integral over the real part $R(\omega)$ of a complex generalized impedance $Z(\omega)$. In modern language, one switches from ω to $f = \omega/2\pi$ and thereby rewrites (4.11) as $\langle V^2\rangle = 4kT\int R(f)df$, and one then identifies the contribution at frequency f as the "Spectral density" of V:

$$G_V(f) \equiv S_V(f) \equiv \tilde{V}^2(f) = 4kTR(f); \qquad (1)$$

and similarly for the quantum-regime formula (4.8).]

OO. Peter R. Saulson, "Thermal noise in mechanical experiments," *Phys. Rev. D* **42**, 2437–2445 (1990). [This paper applies the generalized fluctuation-dissipation theorem of paper 1. to thermal noise in a mechanical oscillator. The fluctuation-dissipation theorem is Eq. (5) of this paper; and it implies that the spectral density of the oscillator's displacement $x(t)$ has the form (16),

$$G_x(f) = \tilde{x}^2(f) = \frac{4k_BTk\phi(\omega)}{\omega[(k - m\omega^2)^2 + k^2\phi^2]}. \qquad (2)$$

Here $k = m\omega_o^2$ is the oscillator's spring constant with ω_o its angular eigenfrequency; $\omega \equiv 2\pi f$ is angular frequency; and $k\phi(\omega)$ is $R(\omega)$, the real part of the generalized impedance. The key issue raised in this paper is "What is the frequency dependence of $\phi(\omega)$?" For viscous damping, $\phi \propto \omega$; for structural damping, ϕ is independent of ω.]

PP. Aaron Gillespie and Frederick Raab, "Thermal noise in the test mass suspensions of a laser interferometer gravitational-wave detector prototype," *Phys. Lett. A*, **178**, 357–363 (1993). [In this paper strong evidence is given that for the flexural motion of the wire from which a test mass hangs, the damping is structural, i.e. ϕ is independent of ω. Note that in this paper the phrase "lineshape" is sometimes used for the spectral density $\tilde{x}^2(f)$.]

Suggested Supplementary Reading:

e. More on the fluctuation-dissipation theorem:

Herbert B. Callen and Richard F. Greene, "On a theorem of irreversible thermodynamics," *Phys. Rev.*, **86**, 702 (1952).

f. Elasticity theory:

L. D. Landau and E. M. Lifshitz, *Theory of Elasticity* (Pergamon Press, New York, 1959).

g. Losses in materials:

g1. A.S. Nowick and B.S. Berry, *Anelastic Relaxation in Crystalline Solids*, (Academic Press, New York, 1972). [A good, general book.]

g2. A.L. Kimball and D.E. Lovell, "Internal friction in solids," *Phys. Rev.* **30** 948 (1927). [An early reference for losses independent of frequency, i.e. structural damping, in solid materials.]

g3. Clarence Zener, "Internal friction in solids: I. Theory of internal friction in reeds," *Phys. Rev.* **52** 230 (1937); "Internal friction in solids: II. General theory of thermoelastic internal friction," *Phys. Rev.* **53**, 90 (1938). [The original references for thermoelastic damping.]

h. Vibrations of Cylinders:

QQ. Aaron Gillespie and Frederick Raab, "Thermally excited vibrations of the mirrors of a laser interferometer gravitational-wave detector," unpublished (1994).

h2. James R. Hutchinson, "Vibrations of solid cylinders," *J. Appl. Mech.*, **47**, 901 (1980).

h3. James R. Hutchinson, "Axisymmetric vibrations of free finite-length rod," *J. Acoust. Soc. Am.* **51**, 233 (1972).

h4. G. W. McMahon, "Experimental study of vibrations of solid, isotropic elastic cylinders," J. Acoust. Soc. Am. **36**, 85 (1964).

i. "Exact" solution for a pendulum with a finite size, lossy wire:

Gabriela I. Gonzalez and Peter R. Saulson, "Brownian motion of a mass suspended by an anelastic wire," *J. Acoust. Soc. Am.*, in press (1994). [See Aaron Gillespie (x2128) for a copy.]

j. Ultra-High Q pendula:

j1. V.B. Braginsky, V.P. Mitrofanov, and O.A. Okhrimenko, "Pendulum fused silica oscillators with small dissipation," *Phys. Lett. A*, **175**, 82 (1993).

j2. D.G. Blair, L. Ju, and M. Notcutt, "Ultra high Q pendulum suspensions for gravitational wave detectors," *Rev. Sci. Instrum.*, **64**, 1899 (1993).

k. Some current experimental work:

k1. J.E. Logan, N.A. Robertson, J. Hough, and P.J. Veitch, "An investigation of coupled resonances in materials suitable for test masses in graviational wave detectors," *Phys. Lett. A* **161**, 101 (1991).

k2. J.E. Logan, N.A. Robertson, and J. Hough, "An investigation of limitations to quality factor measurements of suspended masses due to resonances in the suspension wires," *Phys. Lett. A*, **170**, (1992).

RR. A. Gillespie and F. Raab, "Suspension losses in the pendula of laser interferometer gravitational wave detectors," *Phys Lett A*, in press (1994).

A Few Suggested Problems:

1. *Relationship between the equipartition theorem and the fluctuation-dissipation theorem.* Consider a pendulum with mass $m = 1kg$ and swing frequency $f_o = 2\pi\omega_o = 1$ Hz, and with damping such that, when it is driven at angular frequency ω, its equation of motion is

$$m\ddot{x} = -k(1 + i\phi(\omega))x + Fe^{i\omega t}; \tag{3}$$

where $k = m\omega_o^2$ and x is the transverse position of the pendulum mass. Recall from Gillespie's lecture or the above assigned reading that the pendulum's position will exhibit fluctuations with spectral density given by Eq. (2) above.

   a. If the pendulum is set swinging freely, what is its damping time, i.e. the time τ_* for its energy of swing to be damped by $1/e$?

   b. Take $\phi(\omega) = 10^{-7}\omega/\omega_o$, corresponding to weak frictional damping. Integrate the thermal noise spectrum to find the pendulum's rms velocity $v_{\rm rms} = \langle\dot{x}^2\rangle^{1/2}$. Compare your result with the rms velocity predicted by the equipartition theorem.

   c. What is the full width at half maximum (FWHM) of the thermal noise spectral density $\tilde{x}^2(f)$ in terms of f_o and $\phi(\omega)$?

   d. What fraction of the pendulum's total rms energy lies in the frequency band defined by the FWHM around the resonant frequency? What fraction lies within 10 FWHM?

   e. Take $\phi(\omega) = 10^{-7}$ independent of ω, corresponding to structural damping. Notice that at $\omega \ll \omega_o$, the pendulum's motions are characterized by flicker noise, $\tilde{x}^2(f) \propto 1/f$, and that as a result the integral of the spectral density diverges. Can you explain physically how this comes about? Take as a lower cutoff frequency the inverse of the lifetime of the universe (10^{10} years). What then is the pendulum's rms velocity?

2. *Johnson noise and thermal noise due to eddy current damping.* In Gillespie's lecture the eddy current damping of a pendulum due to a simple current loop and a resistor was found; see his transparency number 20.

   a. What is the spectral density of the pendulum's displacement, $\tilde{x}^2(f)$?

   b. An electrical resistor has Johnson noise, $\tilde{V}^2(f) = 4k_BTR$ where R is its resistance. Compute the pendulum's spectral density $\tilde{x}^2(f)$ due to the Johnson noise associated with the flow of current in the resistor.

   c. Is there a difference, physically, between the damping processes in parts a. and b.?

3. *Pendulum losses due to flexing of the wire material.* This problem examines how the losses in the pendulum due to flexing of the wire material and the associated thermal noise scale with the parameters of the pendulum.

   a. Show that, for a pendulum of fixed mass, the thinner one makes the support wire, the weaker will be the damping of the pendulum's swing.

   For parts b–d, assume that the wire is stressed to its maximum safe value, i.e. that its tension per unit area is held at the maximum (a fixed constant independent of the pendulum's other parameters).

b. How do the pendulum's losses, i.e. $\phi(\omega)$, scale with its mass? How does the off-resonance thermal noise scale with the mass?

c. One way of getting high Q pendulums is to use ribbons with rectangular cross sections rather than circular wires. How do the losses in the pendulum scale with the ribbon thickness (its short dimension)? How does the off-resonance thermal noise scale?

d. How do the losses and thermal noise in the vertical mode scale with mass? with ribbon thickness?

4. *Effective mass coefficients in a simple model of a test mass.* The concept of "effective mass coefficients" was introduced in Gillespie's lecture (see his transparency numbers 40-41 and also see reference 7.a of the supplementary reading). Explicit calculation of these effective mass coefficients is a recent development in the gravitational-wave field. Previously, experimenters used a model which assumed that the laser was an ideal one-dimensional beam and the mass was one dimensional. In this model, the modes can be found by solving the one-dimensional acoustic wave equation:

$$\frac{\partial^2 u}{\partial z^2} = \frac{1}{c^2}\frac{\partial^2 u}{\partial t^2}, \tag{4}$$

where c is the sound velocity and $u(z,t)$ is the longitudinal displacement of the mass's material.

a. What are the eigenfunctions of the modes? (The boundary condition is no stress at the end faces: $\partial u/\partial z = 0$ at $z = \pm h/2$ where h is the length of the mass.)

b. What are the effective mass coefficients as a function of ω_n the resonant frequency of the $n'th$ mode? How do these compare to the actual effective mass coefficients of the 40m prototype's test masses for the lowest-frequency mode, ω_0? for higher-frequency modes?

c. What is the mode density $\rho(\omega)$?

d. What is the low-frequency ($\omega \ll \omega_0$) thermal noise as a function of the highest frequency mode included in the analysis?

e. Experimenters knew that this model was flawed in that the axisymmetric modes comprised a two-dimensional system ($\rho(\omega) \propto \omega$), so they used the effective mass coefficients of the one-dimensional model but changed the mode density to $\rho(\omega) \propto \omega$. What is the low-frequency thermal noise in this case as a function of the highest frequency mode included?

Lectures 13 & 14
Thermal Noise

by Aron Gillespie, 11 & 13 May 1994

Gillespie lectured from the following transparencies. Kip has added a few annotations to them.

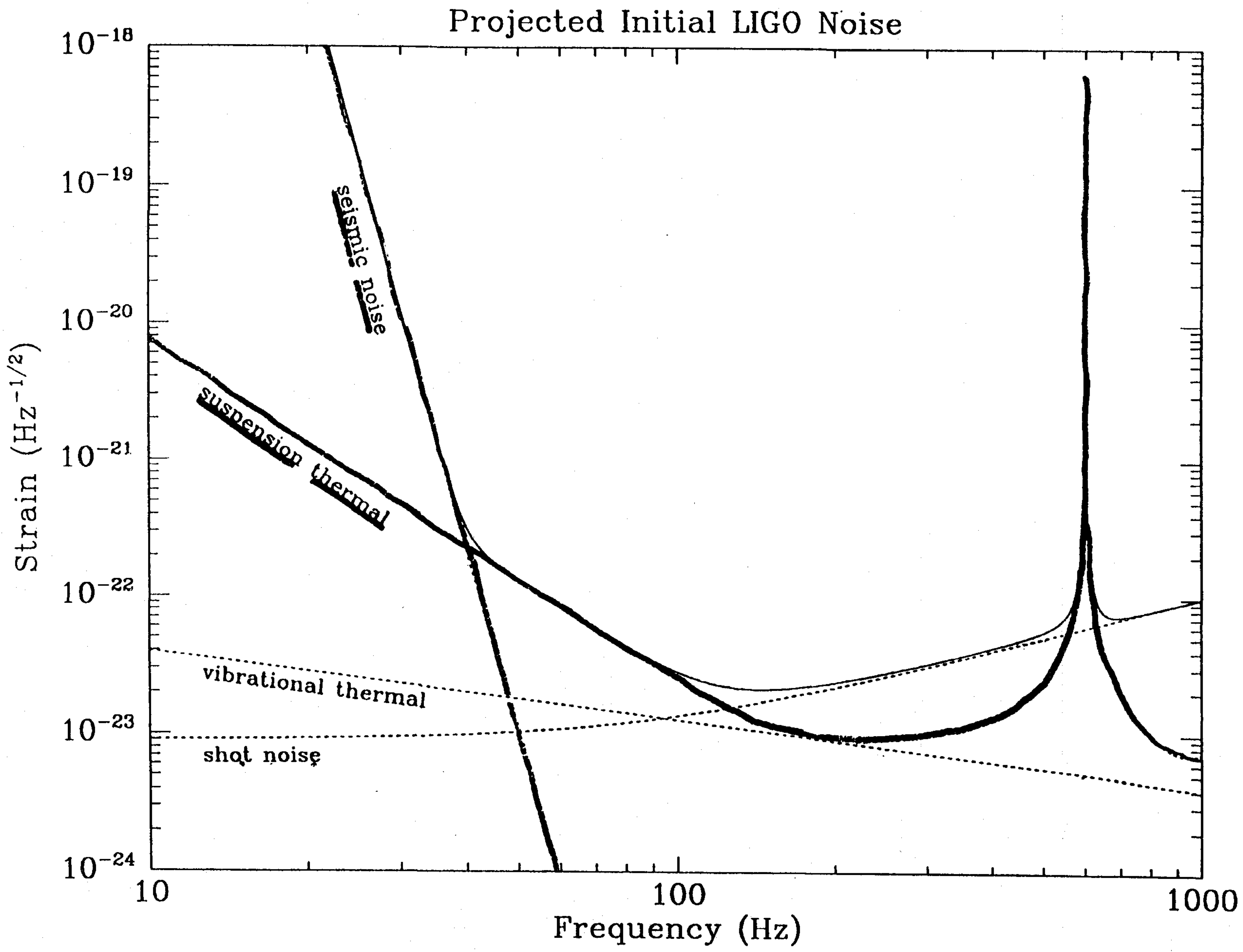
LECTURES 13&14

Test Mass and Suspensions and Thermal Noise

May 11 & 13

Aaron Gillespie

- Equipartition Theorem
- Key Issues in elasticity
- Fluctuation Dissipation Theorem & frequency dependence of the noise
- Suspension modes
  - damping mechanisms
  - noise levels
- Test Mass Vibrational Modes
  - coupling to interferometer
  - damping mechanisms
- Issues for Advanced Detectors
- "Excess" Noise


Projected Initial LIGO Noise

Equipartition Theorem

One definition of a body being at a given temperature is that each mode of the system contains on average $\frac{1}{2}k_BT$ thermal energy.

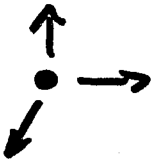
$k_B = 1.38 \times 10^{-23}\ J/K$ is Boltzman's constant.

There are then two problems to be solved:

- identify the relevant modes of the gravitational-wave detector
- evaluate the consequences of those modes having $\frac{1}{2}k_BT$ thermal energy.

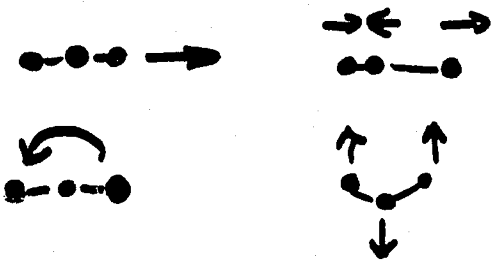
Identifying the Modes

I. Monatomic Gas (ie He)



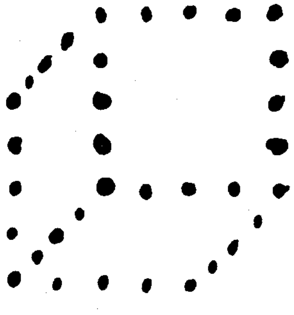
3 translational modes

II. Molecular Gas (ie CO_2)



3 translation
2 rotation
2 vibration
2 bending

III. Microscopic Crystal
(ie. 5x5x5 cubic crystal)



need to solve ~350 coupled harmonic oscillators

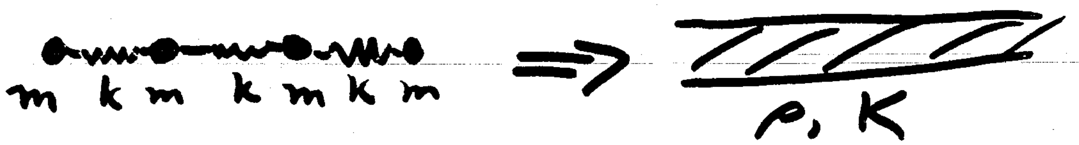
hard problem

IV. Macroscopic Body
~10^{26} atoms

need a new theory!

Theory of Elasticity

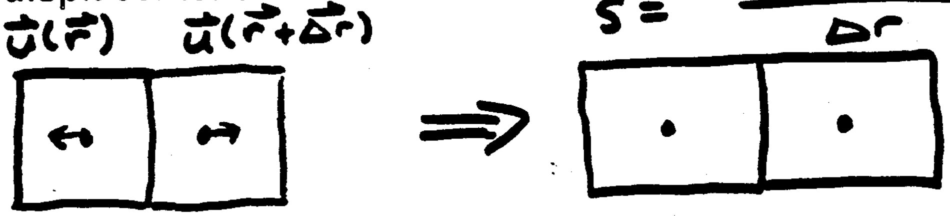
Elaticity theory models a solid not as individual atoms, but rather as a continuum with a mass density, ρ, and a spring constant per unit length, called a modulus.



Motion is described in terms of a displacement vector, $\vec{u}(\vec{r})$, which gives the distance of a particular volume element from its equilibrium value.

Energy and forces are described in terms of strain, which is the relative motion of adjacent volume elements, which can be described mathematically as the spatial derivative of the displacement.

$$S = \frac{u(r) - u(r+\Delta r)}{\Delta r}$$

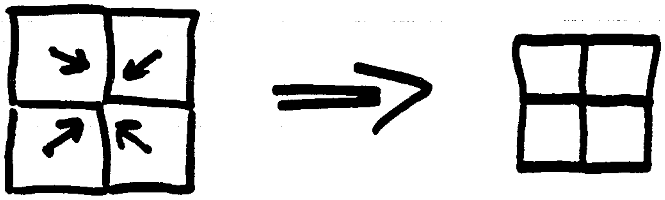


In a three dimensional body, the strain forms a 3×3 tensor, called the strain tensor.

$$S = \begin{bmatrix} \frac{\partial u_1}{\partial x_1} & \frac{\partial u_1}{\partial x_2} & \frac{\partial u_1}{\partial x_3} \\ \frac{\partial u_2}{\partial x_1} & \frac{\partial u_2}{\partial x_2} & \frac{\partial u_2}{\partial x_3} \\ \frac{\partial u_3}{\partial x_1} & \frac{\partial u_3}{\partial x_2} & \frac{\partial u_3}{\partial x_3} \end{bmatrix}$$

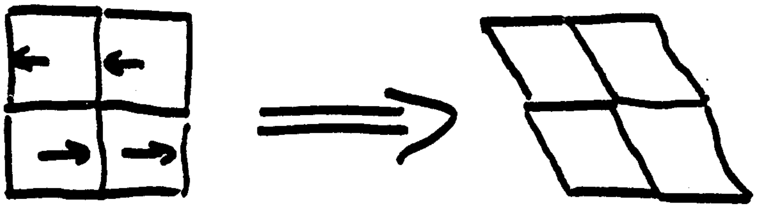
2 Different Types of Motion:

1. Bulk Motion:



Volume Change

2. Shear Motion:



No Volume Change

Mathematically,

Bulk Motion is trace of strain tensor

Shear Motion is traceless symmetric part of strain tensor

(Antisymmetric part describes uniform rotation which has no internal energy)

Bulk and Shear Motions have different moduli associated with them:

K : Bulk Modulus

μ : Shear Modulus

Stress Tensor describes force between adjacent volume elements:

$$T = -K \oplus I - 2\mu \Sigma$$

trace of S

Identity Matrix

traceless symmetric part of S:

$$\Sigma_{ij} = \frac{1}{2}\left(S_{ij} + S_{ji} - \frac{1}{3} \oplus \delta_{ij}\right)$$

Elastic Energy per unit volume

$$U = \frac{1}{2} K \oplus^2 + \mu \Sigma_{ij} \Sigma_{ij}$$

Dimensions:

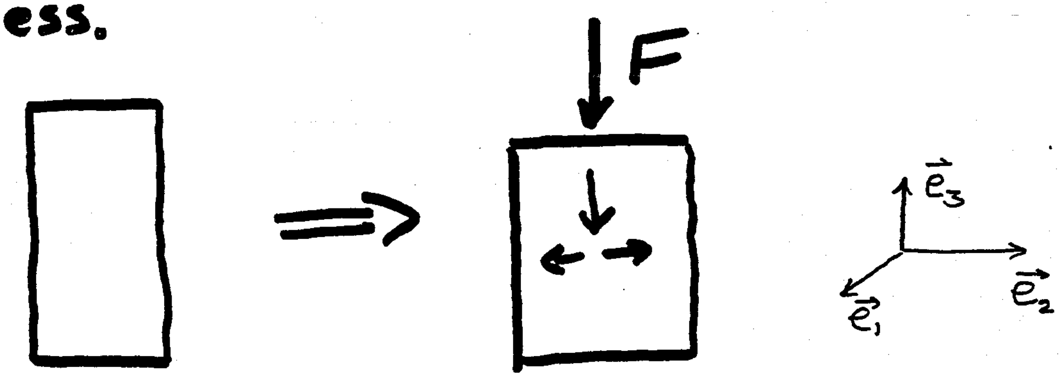
K, μ : $\frac{\text{spring constant}}{\text{length}}$ = Pressure: F/A

T : Force across area between 2 volume elements: Pressure

U : Energy/volume

Often, decomposing motion into shear and bulk is awkward.

Usually we are concerned with a force in one direction or homogenous stress.



Young's Modulus, E relates stress and Strain:

$$E = \frac{T_{33}}{S_{33}} = \frac{9\mu K}{3K+\mu}$$

Complementary Quantity Poisson's Ratio

$$\nu = \frac{1}{2}\frac{3K-2\mu}{3K+\mu}$$

Example:

What is extension and vertical frequency of 40m test mass suspension?

$l = 25\,cm$

$d = 75\,\mu m$

$m = 1.6\,kg$

$E(\text{steel}) = 210\,GPa$

$$T_{33} = \frac{F}{A} = \frac{mg}{4 \cdot \pi \left(\frac{d}{2}\right)^2} = 8.9 \times 10^8 \frac{N}{m^2}$$

All numbers are in mks units

$$S_{33} = \frac{\Delta l}{l} = 4.0\,\Delta l$$

$$E = T_{33}/S_{33} \implies 4.0\,\Delta l = \frac{8.9 \times 10^8}{210 \times 10^9}$$

$\Delta l = 1.1\,mm$ $\Delta l / l \simeq 1/250$ [half the breaking strain]

$$\omega = \sqrt{\frac{k}{m}} = \sqrt{\frac{F/\Delta x}{m}}$$

$\omega = 96\ ^{rad}/sec$

$f = 15\,Hz$

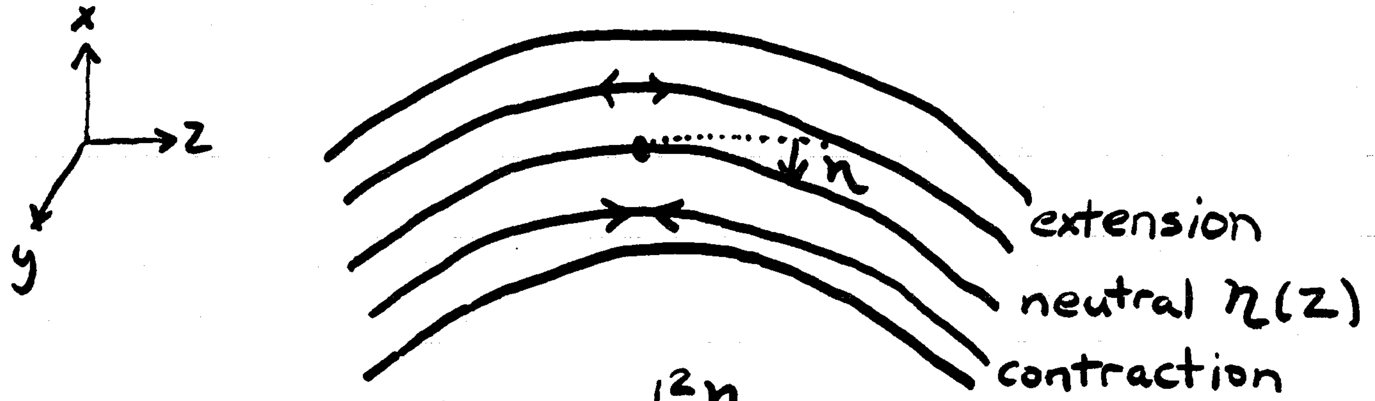
$$T_{33} = E S_{33}$$

$$F/A = E \frac{\Delta x}{l}$$

$$F/\Delta x = EA/l$$

$$= 1.5 \times 10^4\ N/m$$

Bending Moment



$$S_{zz} = \frac{x}{R} = x\frac{d^2\eta}{dz^2}$$

↳ radius of curvature

$$F_z = E S_{zz} dA = E x \frac{d^2\eta}{dz^2} dA$$

"Torque" about neutral surface:

$$N_y = F_z x = E x^2 \frac{d^2\eta}{dz^2} dA$$

Bending Moment

$$M = E \int x^2 dA \frac{d^2\eta}{d^2 z} = E J \frac{d^2\eta}{d^2 z}$$

Geometrical moment of inertia

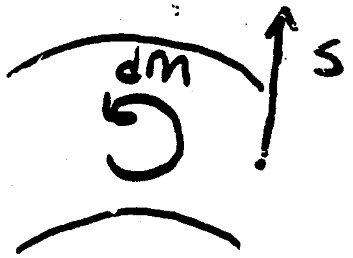
$$J = \int x^2 dx dy$$

$J = \frac{\pi}{4} r^4$ rod

$= a^3 b/12$ rectangular cross section

Bent Rods

If the bending moment, M, varies along the length of the rod, then their will be a net torque which can be described by a shearing force, S.



$$S_x = \frac{dM}{dz} = EJ\frac{d^3\eta}{dz}$$

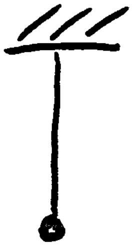
Shearing force varies along the length of the rod to balance extenal loads.

$$\frac{dS}{dz} = EJ\frac{d^4\eta}{dz} = W(z)$$

↳ load in force/length

Fourth order differential equations are the norm in elastostatics.

Example: Pendulum



Point mass supported by massless wire with finite stiffness.

Eq. of Motion:

tension in wire

$$-EJ\overset{\partial^4\eta/\partial z^4}{\eta^{(iv)}}(z)+\overset{\partial^2\eta/\partial z^2}{T\eta''}(z)=0$$

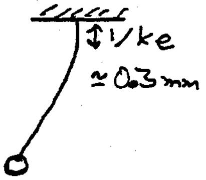
Boundary conditions:

$$\eta(0)=\eta'(0)=0$$
$$EJ\eta'''(L)-T\eta'(L)=M\ddot{\eta}(L)$$
$$-EJ\eta''(L)=0$$

Solution:

$$\eta(z)=\eta_0\left(\frac{\sinh k_eL}{\cosh k_eL}(\cosh k_ez-1)+(k_ez-\sinh k_ez)\right)\times e^{\pm i\omega_0 t}$$

$$\omega_0=\sqrt{\frac{g/l}{1-\frac{\sinh k_el}{k_el\cosh k_el}}}\approx\sqrt{\frac{g}{l}\left(1+\frac{1}{k_el}\right)}$$



$k_e=\sqrt{\frac{I}{EJ}}=3.5\times10^{3}\ m^{-1}$ for 40 m parameter:

$4\times\frac{1}{k_eL}=4.6\times10^{-3}$: not much change in ω_0, but this term will become very important later on.

Internal Modes:

- thus far discussed modes due to external forces using quasi-static thoory
- to describe internal modes, must solve time dependent equations of elasticity

these equations are <u>complicated</u> and <u>difficult</u> to solve because:

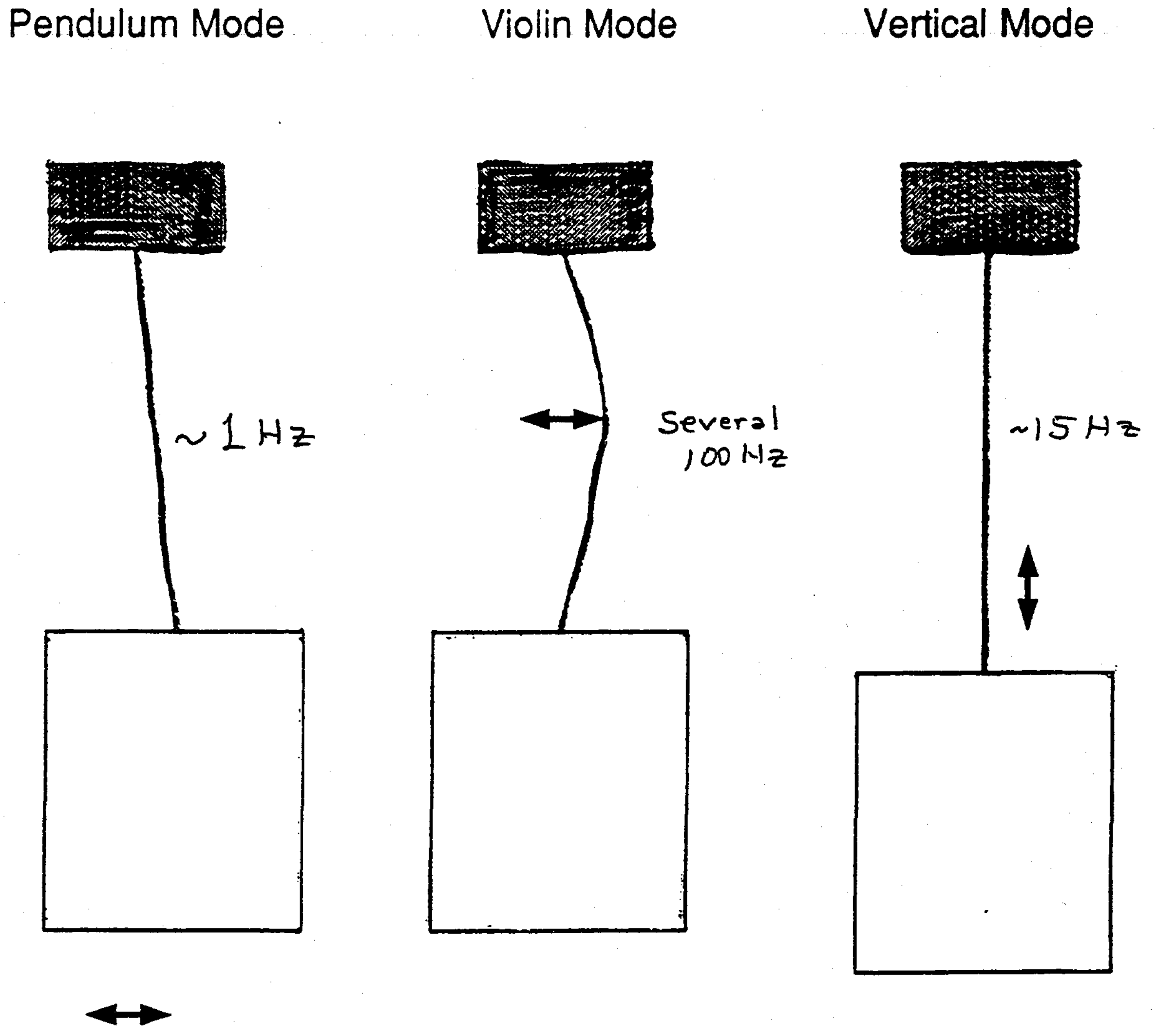
1) must decompose arbitrary displacements into shear and bulk motion
2) shear and bulk motion have different moduli and hence different sound velocities and time dependences

Luckily, right solid cylinder has been solved (but not until 1980 after over 100 years of work). Our mirrors can be approximated by right solid cylinders.

In real world, internal modes are found using finite element analysis software tools.

Here ends elasticity!

Suspension Modes



Thermal Motion of Suspension Modes.

1) Pendulum: $\frac{1}{2} k_B T \approx \frac{1}{2} m \omega_0^2 x^2$ $\quad$ $\omega_0 = 6.28$, $m = 1.6\,\mathrm{kg}$, $T = 300\,\mathrm{K}$

$$x_{rms} \cong 8.1 \times 10^{-12}\,m$$

2) Violin: $\frac{1}{2} k_B T \sim \frac{1}{2} m \omega_0^2 x^2$ $\quad$ $\omega_0 = 2\pi \cdot 600$, $m = 9 \times 10^{-6}\,\mathrm{kg}$

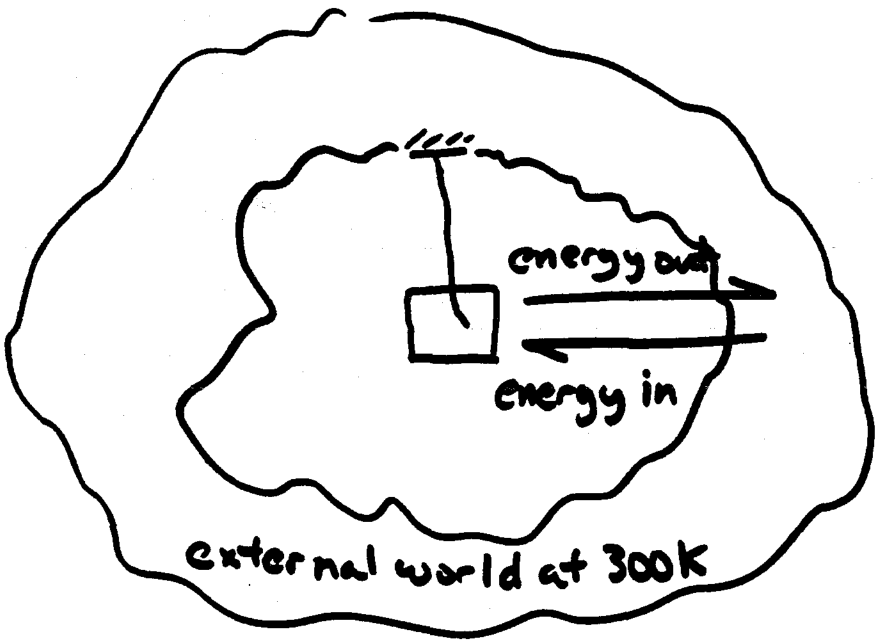
$$x_{rms} \sim 6 \times 10^{-12}\,m$$

(wire amplitude)

$$x_{mass} \approx \frac{m_{wire}}{m_{mass}} x_{wire} \sim 3 \times 10^{-17}\,m$$

These motions, if in the wrong bandwidth, could affect LIGO. Need to know the spectral density of the motion, which can be derived from Fluctuation-Dissipation theorem.

Fluctuation - Dissipation



at equilibrium, average rate of flow of energy out of pendulum (dissipation) = heat flow into pendulum (fluctuation)

Level of equilibrium is

$$\tilde{F}^2 = 4 k_B T \, \mathrm{Re}(Z)$$

Generalized Force $\swarrow$ $\searrow$ Impedance

Mechanical System; $Z = \frac{F}{V}$ $\frac{\text{force}}{\text{velocity}}$

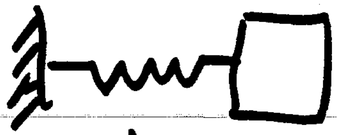
$$\tilde{F}^2 = 4 k_B T \, \mathrm{Re}\left(\frac{F}{V}\right)$$

Electrical System: $Z = \frac{V}{I}$ $\frac{\text{Voltage}}{\text{Current}}$

Johnson Noise : $\tilde{V}^2 = 4 k_B T R$

Impedance of a Mechanical System

model each mode as harmonic oscillator.



$$x = x_0 e^{i\sqrt{\frac{k}{m}}\, t}$$

add loss by giving spring constant a small imaginary part.

$$k \rightarrow k(1+i\phi)$$

$$x = x_0 e^{i\sqrt{\frac{k(1+i\phi)}{m}}\, t} \approx e^{i\sqrt{\frac{k}{m}}\, t} \underbrace{e^{-\sqrt{\frac{k}{m}}\frac{\phi}{2} t}}_{\text{lossy term}}$$

equation of motion

$$-m\omega^2 x + k(1+i\phi(\omega))\,x = F$$

impedance

$$Z = \frac{F}{v} = \frac{F}{i\omega x} = i m\omega - \frac{ik}{\omega} + \frac{k\phi(\omega)}{\omega}$$

$$\mathrm{Re}(Z) = \frac{k\phi(\omega)}{\omega}$$

ϕ is fractional loss of Energy per radian of oscillation at a given frequency ω.

$$Q = \frac{1}{\phi(\omega_0)}$$

General "Thermal Noise Lineshape"

↰ $\tilde{x}^2(f)$ [also called "Spectral density of $x(t)$]

$$\ast\ \tilde{x}^2(f) = \frac{4k_BT}{m\omega}\,\frac{\omega_0^2\,\phi(\omega)}{(\omega_0^2-\omega^2)^2+\omega_0^4\,\phi^2(\omega)}\ \ast$$

$\omega \gg \omega_0$ (ie pendulum mode)

$$\ast\ \tilde{x}^2(f) = \frac{4k_BT}{m}\,\frac{\omega_0^2\,\phi(\omega)}{\omega^5}\ \ast$$

$\omega \ll \omega_0$ (ie test mass vibrational modes)

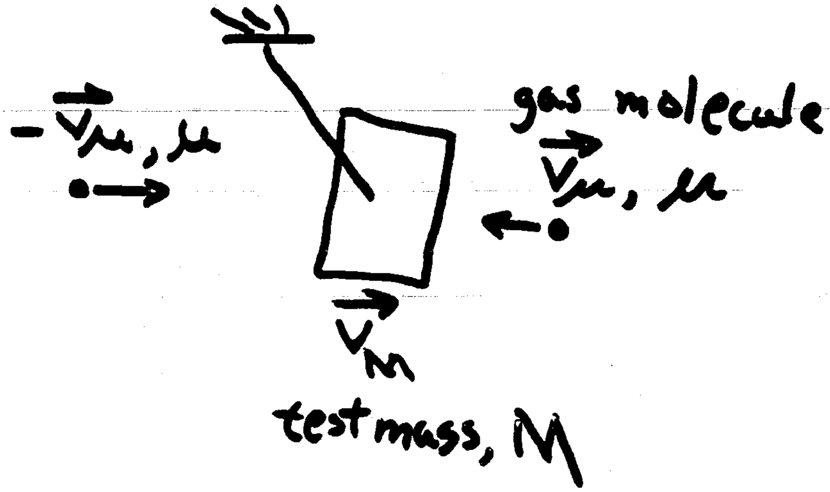
$$\ast\ \tilde{x}^2(f) = \frac{4k_BT}{m}\,\frac{\phi(\omega)}{\omega_0^2\,\omega}\ \ast$$

All you need to know is

$\phi(\omega)$!

Damping Mechanisms

I. Residual Gas Damping



$$\phi(\omega) \approx \frac{2A_\perp P}{M}\sqrt{\frac{\mu}{k_B T}}\,\frac{\omega}{\omega_0^2}$$

← for situation where mean free path of molecules $\gg x$ = displacement of mass

Viscous Loss Mechanism, $\phi(\omega) \propto \omega$

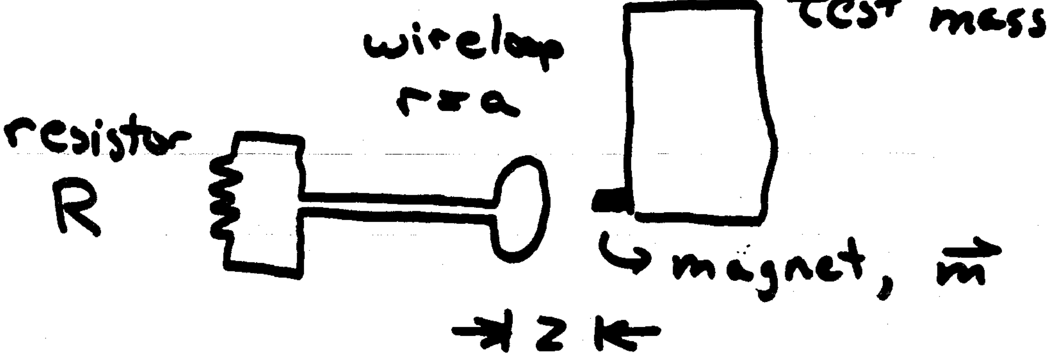
for LIGO like parameters,

$$\phi(\omega) \approx 10^{-10}\left(\frac{P}{10^{-6}\,\mathrm{Torr}}\right)\omega$$

$$\tilde{x}(100\,\mathrm{Hz}) \approx 10^{-20}\sqrt{\frac{P}{10^{-6}}}\ \mathrm{m}/\sqrt{\mathrm{Hz}}$$

$\sim 10^{-6}$ torr would be the necessary vacuum for first LIGO detectors

II. Eddy Current Damping



$$F = \vec{m}\cdot\vec{\nabla}\vec{B} = \vec{m}\left[\frac{3}{2}\frac{\mu_0 I a^2 z}{(a^2+z^2)^{5/2}}\right]$$

when magnet moves, it induces a current + voltage

$$V = \frac{F}{I}\dot{x} = \frac{3}{2}\frac{m\mu_0 a^2 z}{(a^2+z^2)^{5/2}}\dot{x} \equiv \alpha\dot{x}$$

$$I = \frac{\alpha\dot{x}}{R}$$

Power Dissipation + damping force

$$P = I^2 R = \frac{\alpha^2\dot{x}^2}{R}$$

$$F = \frac{P}{v} = \frac{\alpha^2\dot{x}}{R}$$

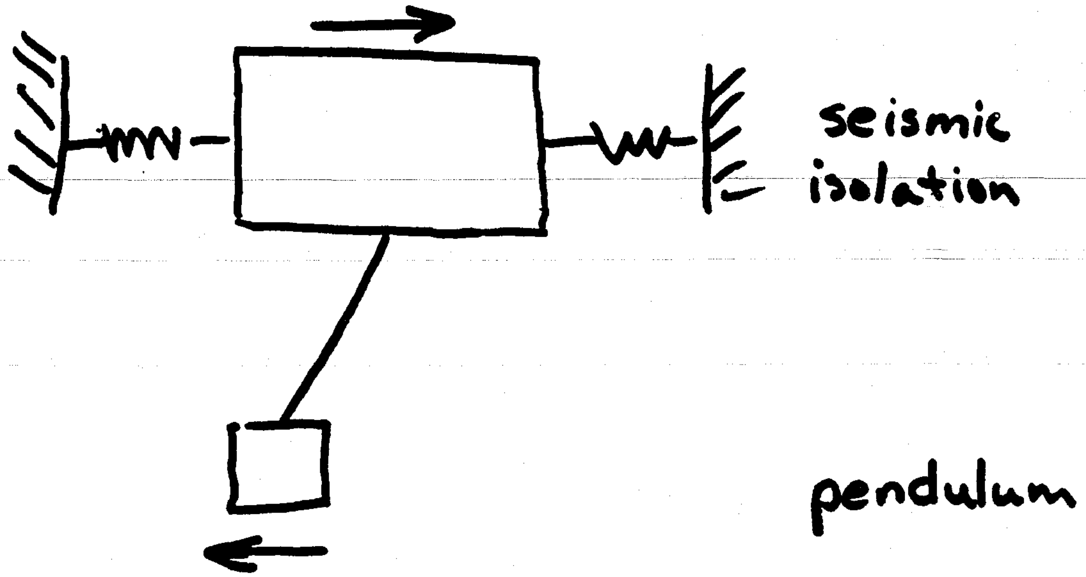
So

$$\phi(\omega) = \frac{\alpha^2\omega}{mR\omega_0^2} \quad \text{(also viscous)}$$

Actual interferometer has more complicated geometries.

Eddy current damping requirements are difficult to meet for initial interferometer, and perhaps impossible for advanced detectors.

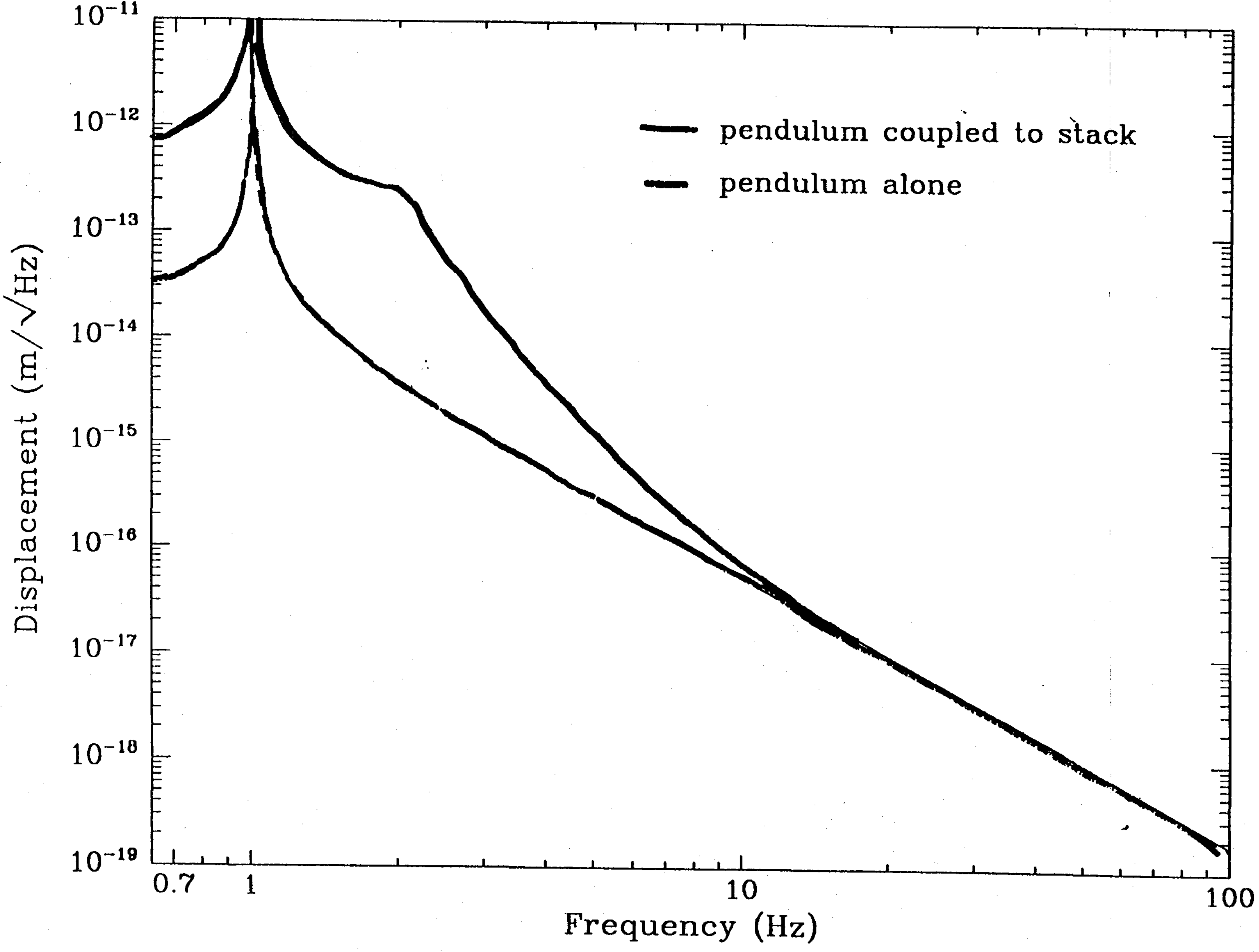
III. Recoil Damping



$$\phi_p(\omega) \approx \frac{\left(\frac{m_p}{m_s}\right)\omega_p^2\,\omega_s^2\,\phi_s(\omega)}{(\omega_s^2-\omega^2)^2+\omega_s^4\,\phi_s^2(\omega)}$$

↑ includes only the topmost oscillator in the seismic isolation stack

Thermal Noise of Pendulum Coupled to Mark II Stack



IV Material Losses

Recall the resonant frequency of a pendulum: [cf. transparency 12]

$$\omega_0 \approx \sqrt{\frac{g}{\ell}\left(1+\frac{1}{k_e\ell}\right)}$$

Define spring constant:

$$k = m\omega_0^2 = \underbrace{\frac{mg}{\ell}}_{\text{due to gravity, lossless}}\left(1+\underbrace{\frac{1}{k_e\ell}}_{\text{due to material, has loss}}\right)$$

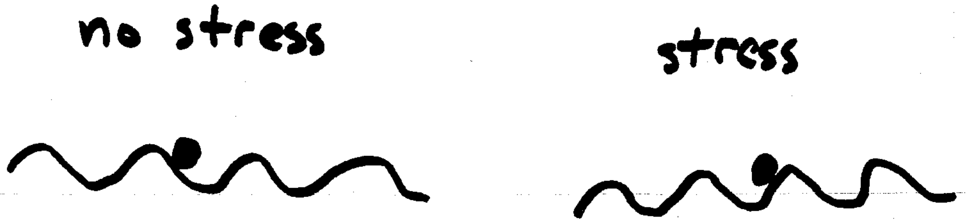
$$k = \frac{mg}{\ell}\left(1+\frac{1}{k_e\ell}\left[1+i\phi(\omega)\right]\right) \approx \frac{mg}{\ell}\left(1+i\frac{\phi(\omega)}{k_e\ell}\right)$$

$$\phi_{pend} \approx \frac{\phi_{wire}}{k_e\ell}$$

$k_e\ell \simeq 10^{-3}$ when wire is strained to within 1/2 its breaking strain

Pendulum losses are ~1000x less than intrinsic losses of wire.

Material Losses



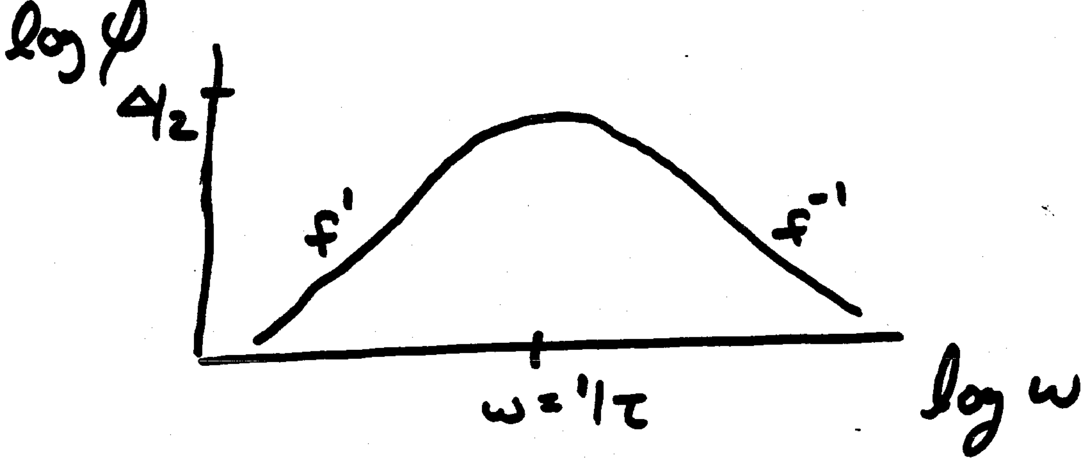
Concept: Some defect, vacancy, impurity or other structural characteristic that has a different equilibrium when material is under stress.

Very Broad Class can be described by 2 parameters:

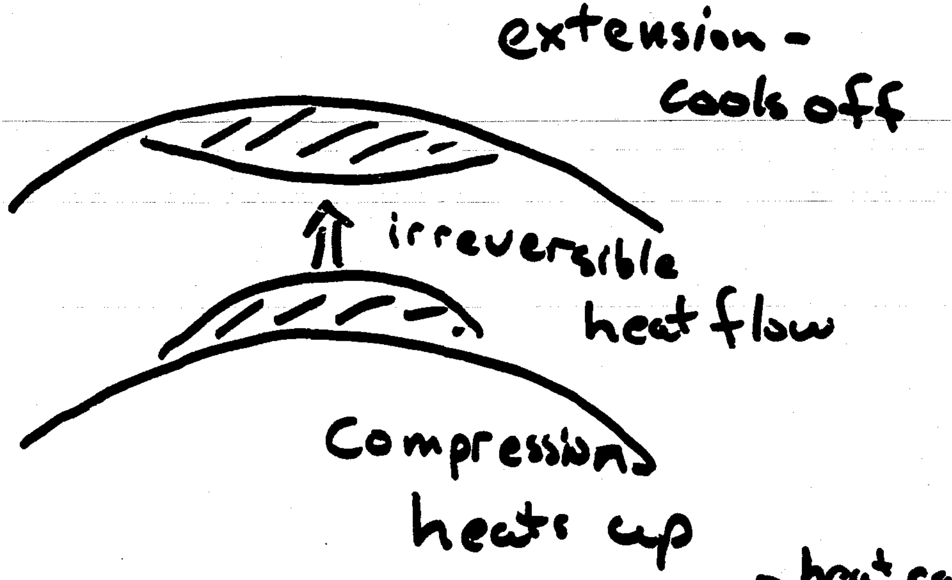
τ: timescale to move from one state to another

Δ: overall stength of loss mechanism

$$\phi(\omega) = \Delta \frac{\omega\tau}{1+\omega^2\tau^2}$$ "DeBye Peak"



Thermoelastic Damping



timescale: $$\tau = \underbrace{.074}_{\text{geometrical factor}} \frac{\overset{\text{heat capacity}}{C}\,\overset{\text{diameter}}{d^2}}{\underset{\text{thermal conductivit}}{K}}$$

strength: $$\Delta = \frac{E\,\overset{\text{thermal expansion}}{\alpha^2}\,T}{c}$$

$$\phi = \frac{\Delta\,\omega\tau}{1+\omega^2\tau^2}$$ A DeBye Peak!

Material Losses (cont.)

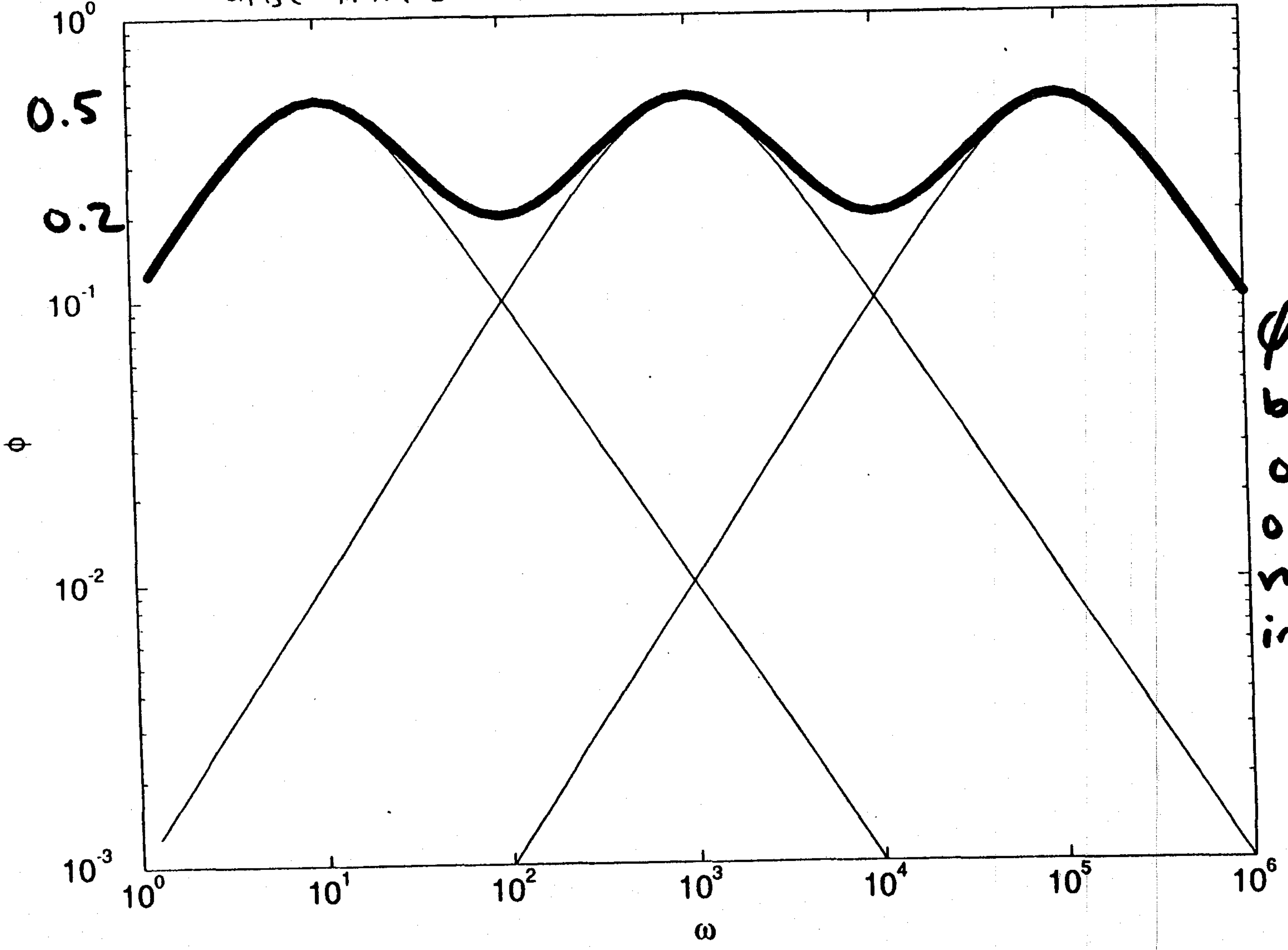
Generally observed behavior in many solids over large frequency bands is losses which vary very slowly as a function of frequency:

$$* \quad \phi(\omega) \approx \text{constant} \quad *$$

No known mechanism gives this behavior. General belief is that solids have many types of defects with wide variety of timescales which when added together give a nearly frequency independent loss function.

Multiple DeBye Peaks

All the DeBye peaks have about the same strength because they all arise from similar atomic processes



ϕ varies by 2.5 over 6 orders of magnitud in frequen

Given the difficulty of experimental measurements of $\Phi(\omega)$, this would look like $\Phi(\omega)$ independent of ω

Measuring ϕ

- on resonance: measure Q by either decay of free oscillation or by FWHM of frequency spectrum

$$\phi(\omega_0) = \frac{1}{Q} \quad \text{easy!}$$

- off resonance: measure phase shift between driving force and forced oscillation. Need to know phase shift to better than ϕ.

difficult to impossible if $\phi \sim 10^{-5} - 10^{-7}$!

- multiple resonance system: derive a relationship between losses of different resonances and intepolate to find general frequency dependence

works in some cases!

Pendulum and Violin Modes

If the principal damping mechanism is due to forces acting through the wire, then the losses in the pendulum mode can be calculated from the measured losses in the violin modes.

- Energy in the pendulum mode as a function of the bending angle:

$$E_p(\theta) \approx \frac{1}{2} m g l \theta^2$$

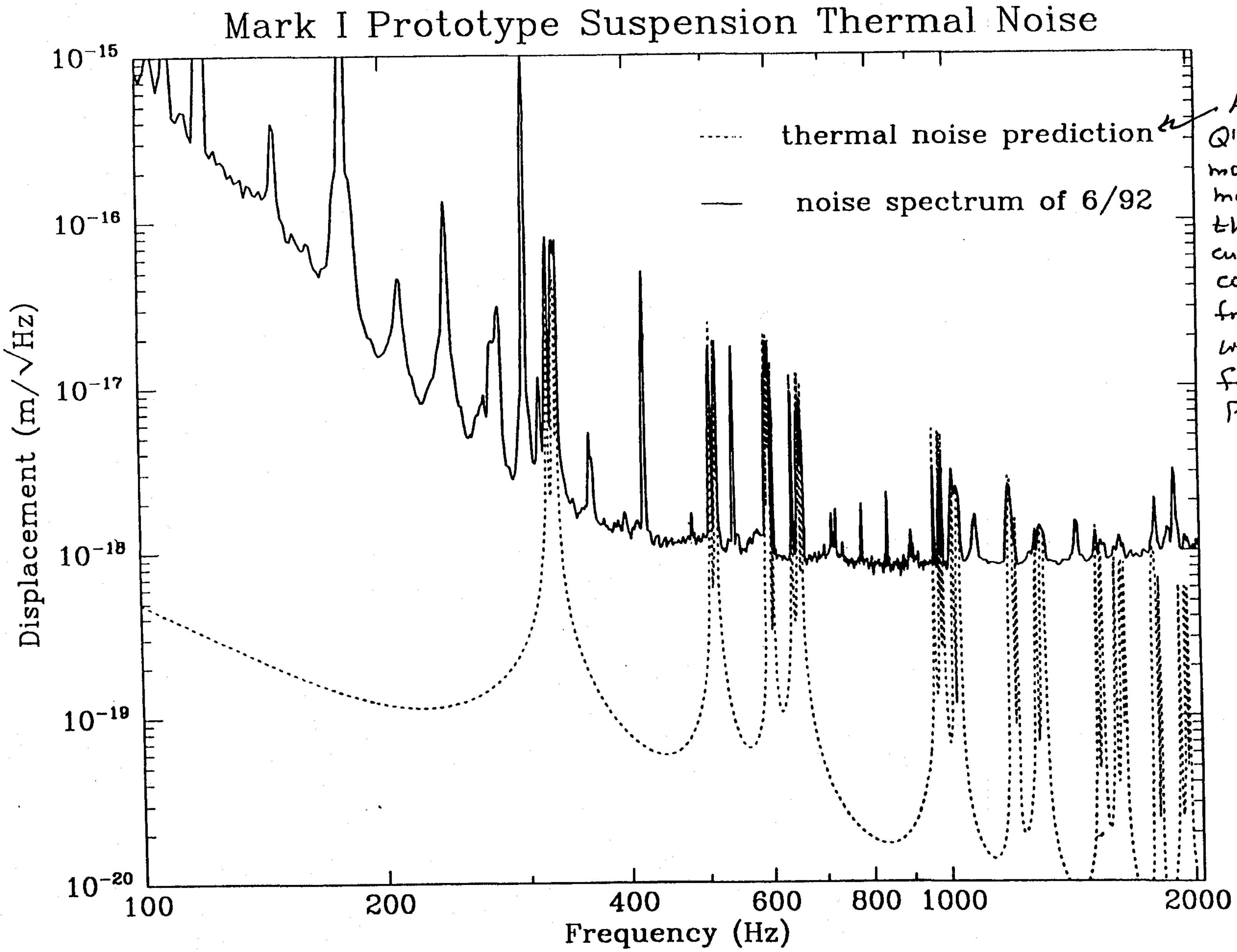
- Energy in a violin mode:

$$E_v(\theta) \approx \frac{1}{4} T l \theta^2$$

- For a 4–wire system with the losses concentrated at the endpoints,

$$\varphi_{pend}(\omega) \approx \frac{1}{8} \times \sum_{4\, wires} \varphi_{violin}(\omega)$$

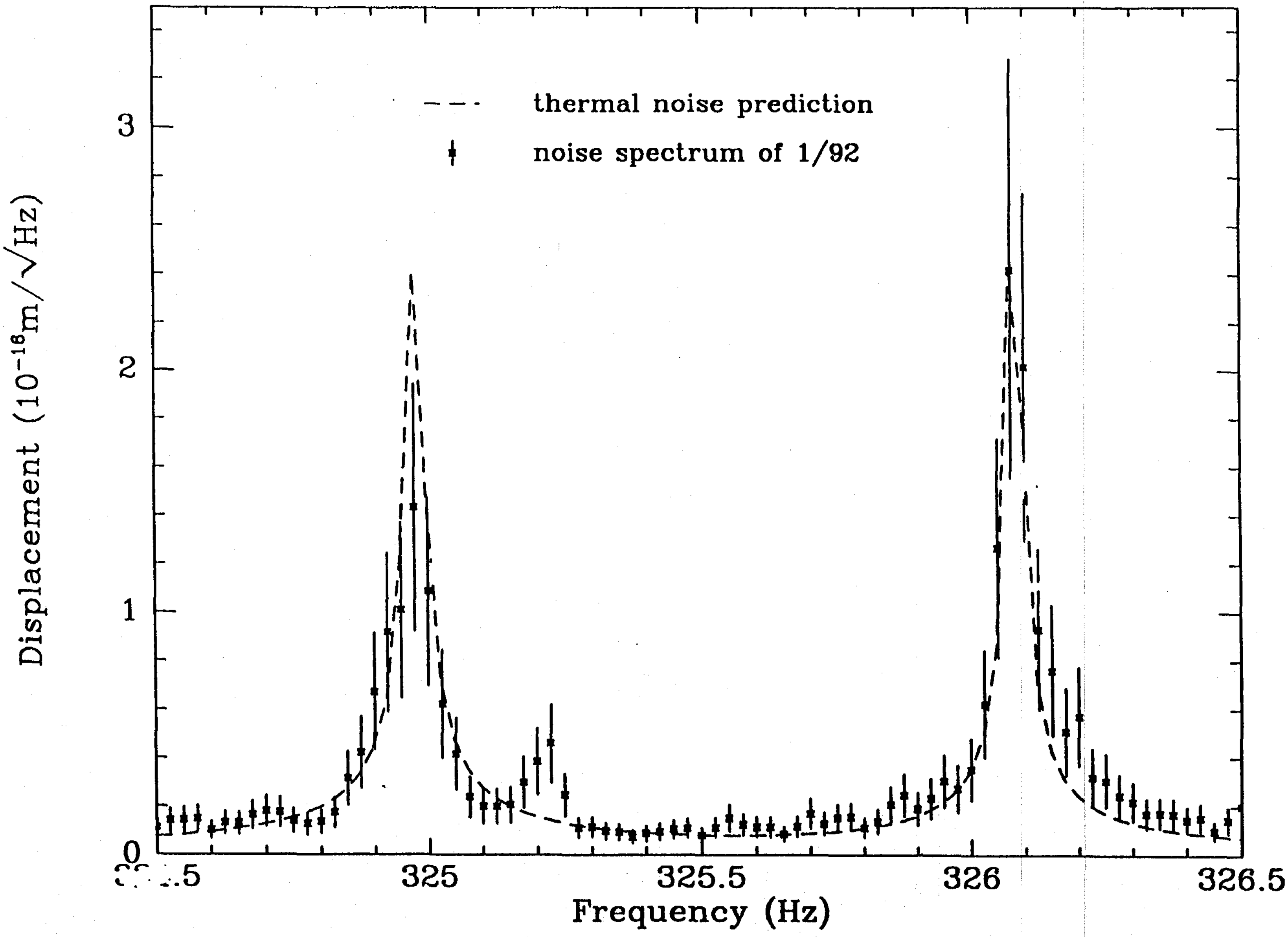
Valid, for 40m prototype suspensions, for violin modes of order $n \lesssim 16$. For $n \gtrsim 16$, the losses in the bulk of the wire become $\gtrsim$ those at its ends, so this fails



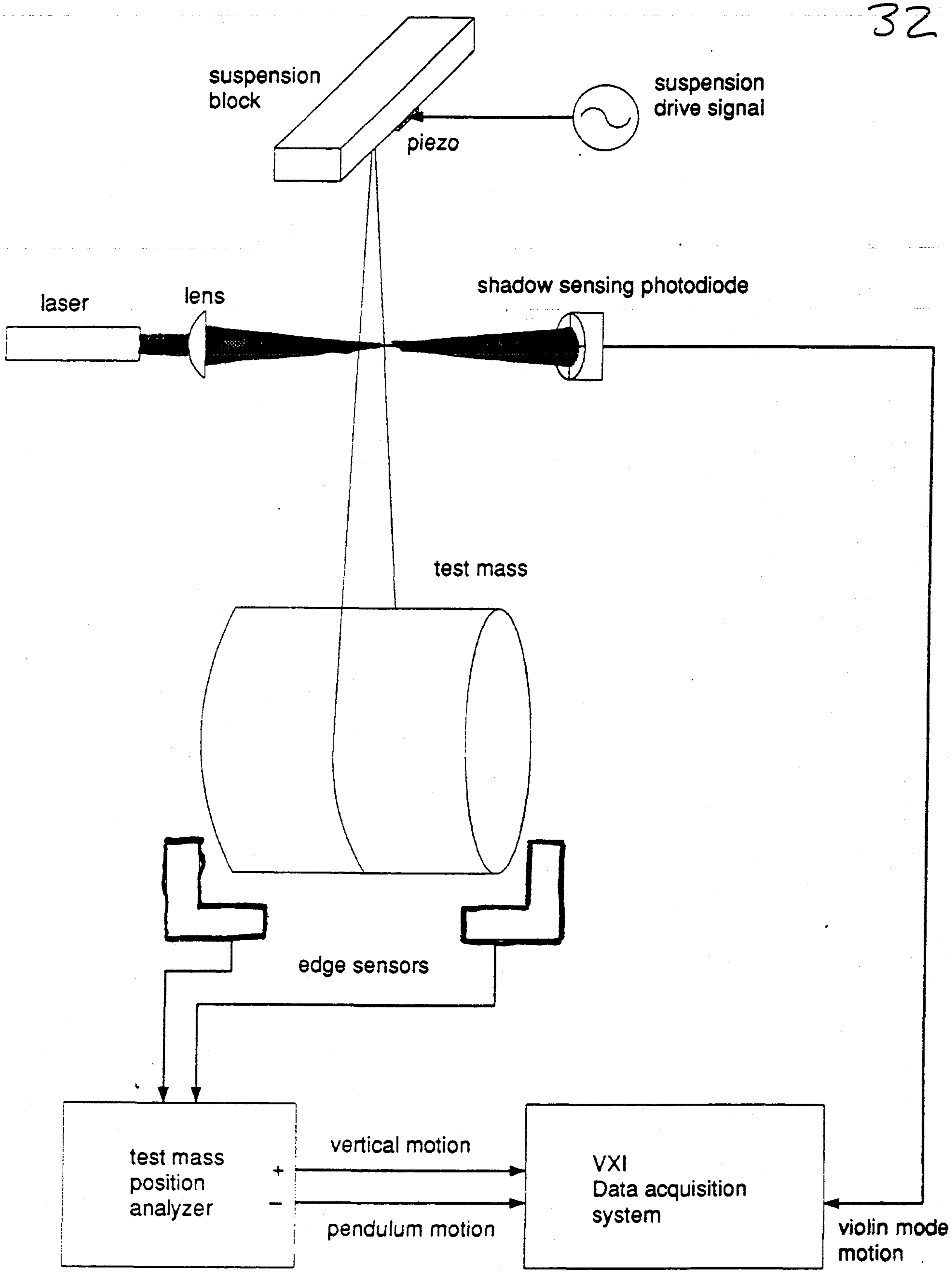
All ω's and Q's of violin modes are measured, then this curve is computed from them with no free parameters.

W
O

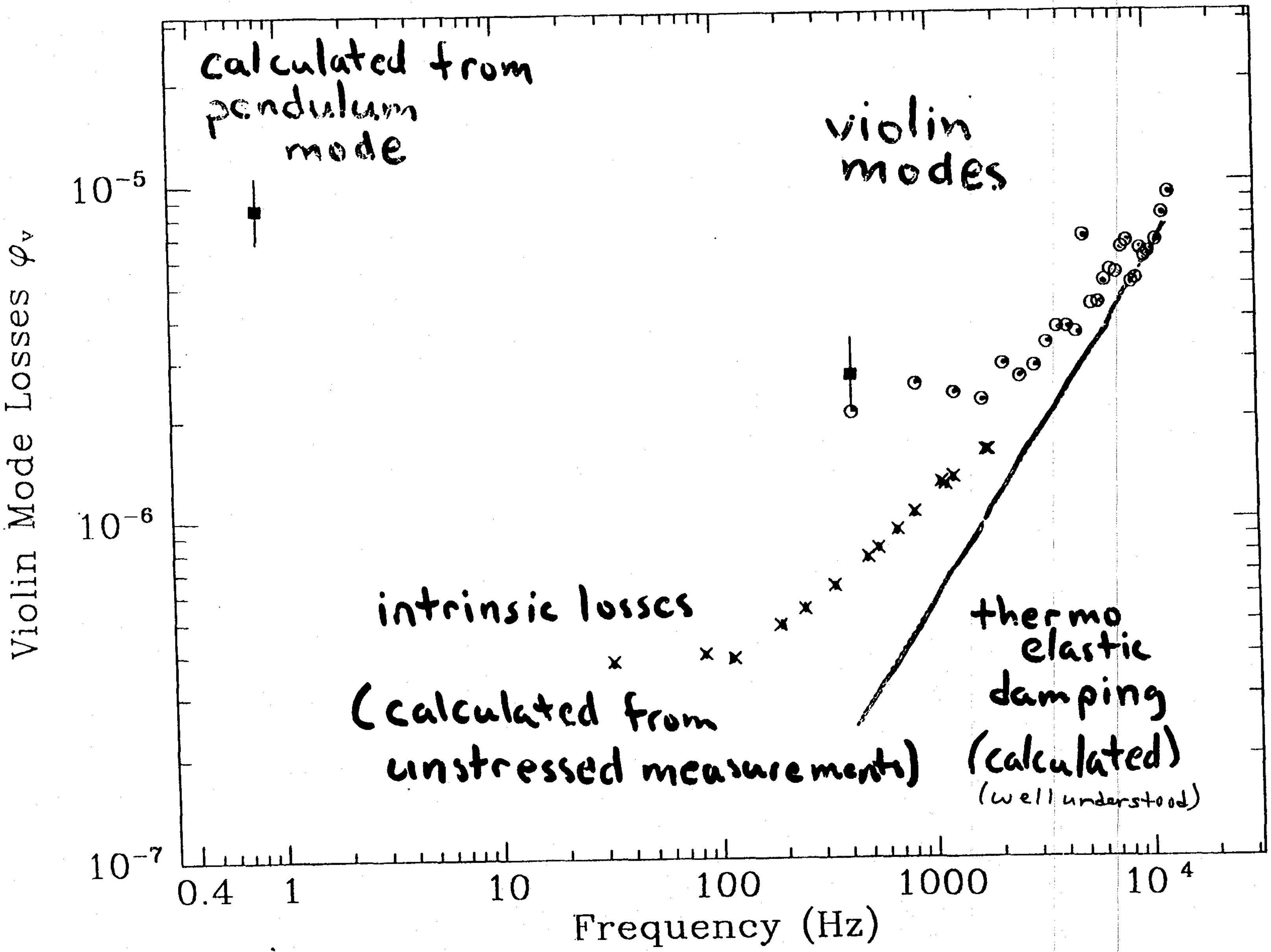
East End Mass Violin Resonances



Suspension Development Apparatus

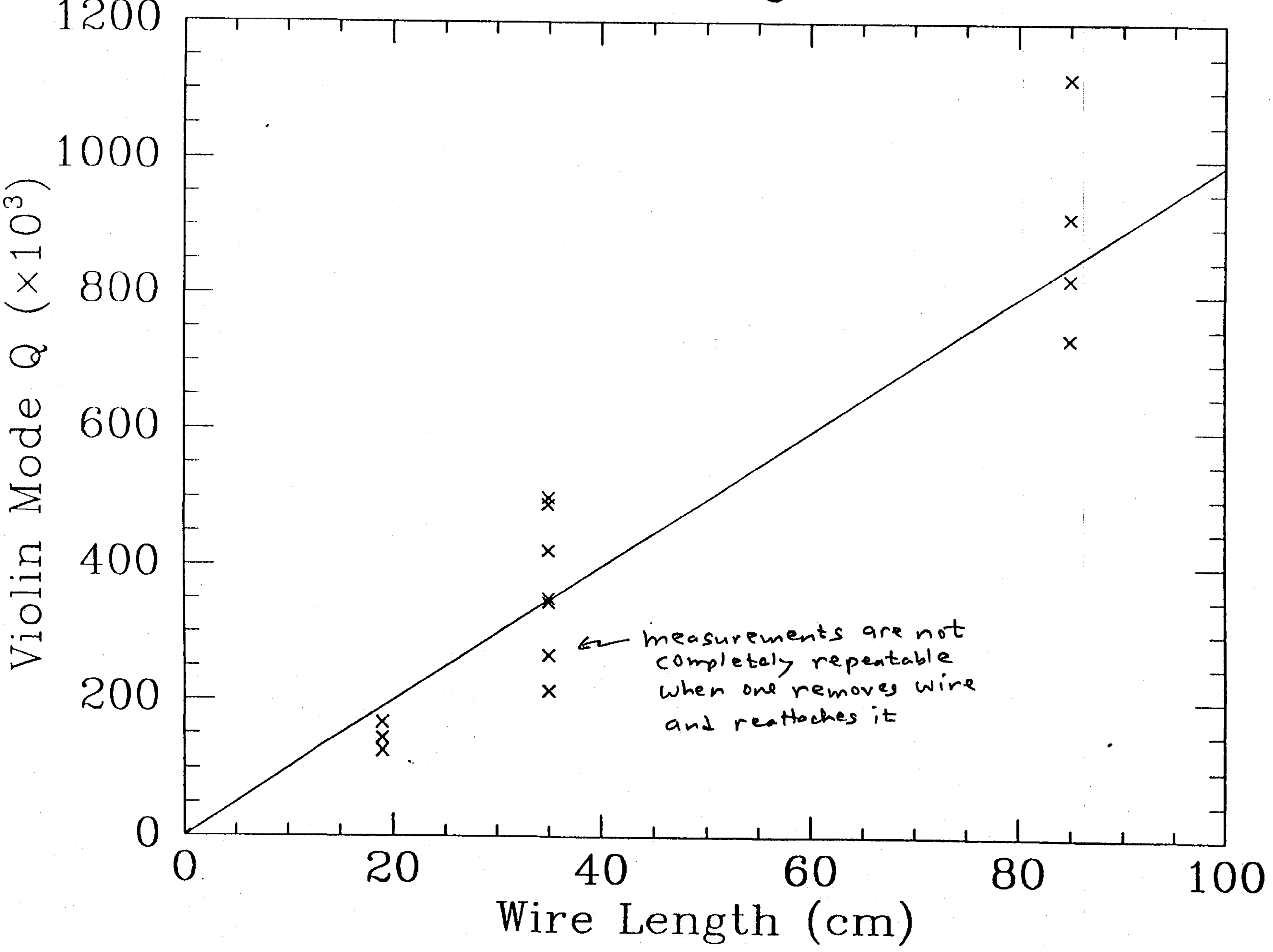


Suspension Losses

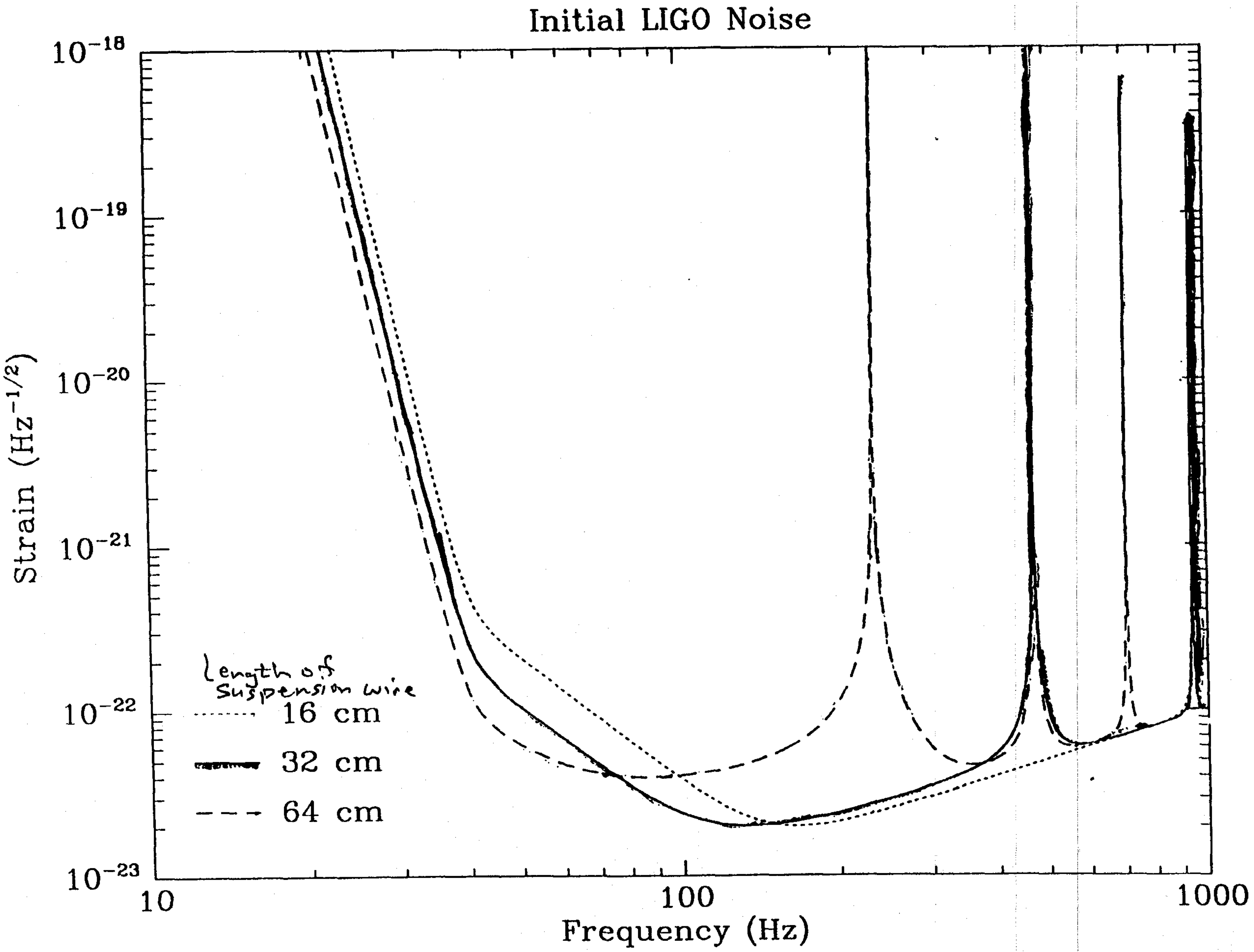


The disagreement between ⊙ & ■ on one hand, and X on other, suggests there are additional stress loss [illegible]

Q vs. Length

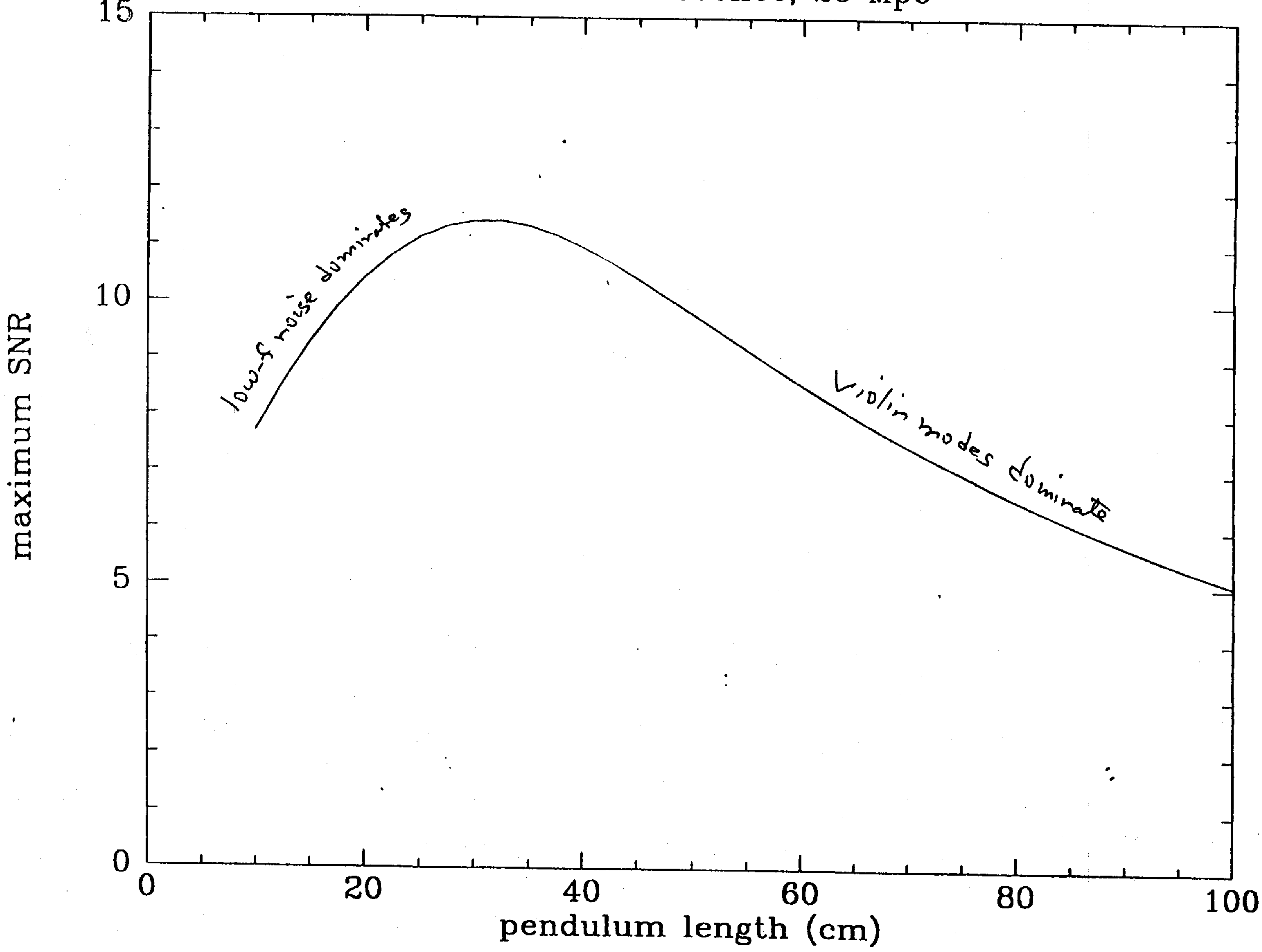


measurements are not completely repeatable when one removes wire and reattaches it



Initial LIGO Noise
10^{-18}
10^{-19}
10^{-20}
10^{-21}
10^{-22}
10^{-23}
Strain ($Hz^{-1/2}$)
10
100
1000
Frequency (Hz)
Length of suspension wire
16 cm
32 cm
64 cm

NS–NS Coalescence, 23 Mpc



36

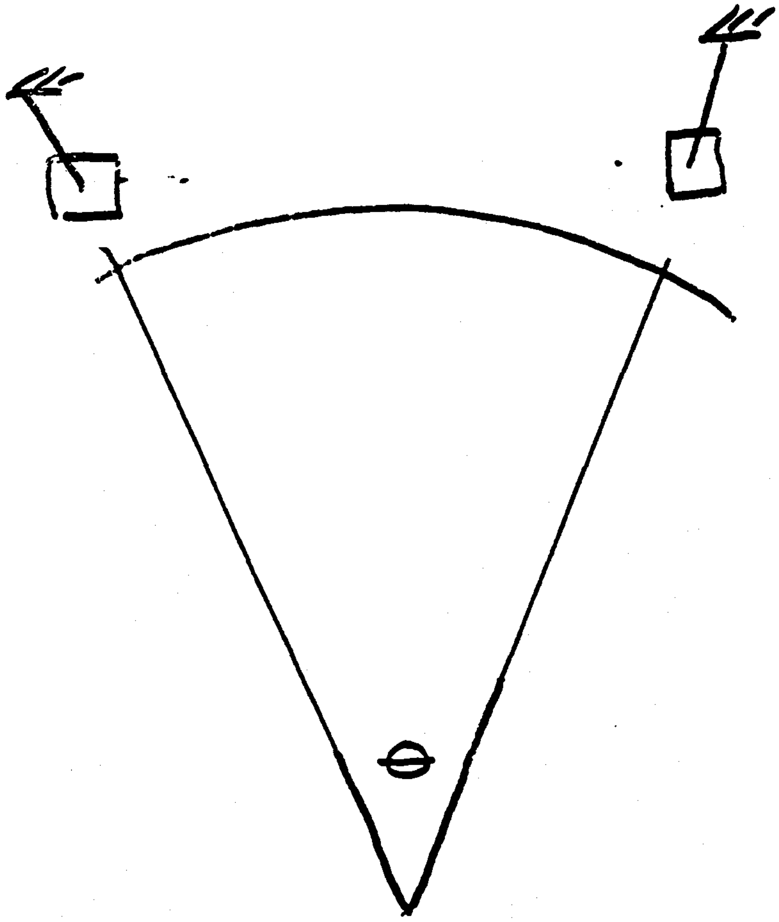
Vertical Resonance

Due to the curvature of the earth, vertical motion is coupled to the interferometer.

$$\frac{\tilde{x}_{vert}(f)}{\tilde{x}_{pend}(f)} = \Theta\sqrt{\frac{\omega_v^2\varphi_v(\omega)}{\omega_p^2\varphi_p(\omega)}} \approx \mathbf{0.1}$$

$Q \ll Q_{pendulum}$ because here <u>all</u> restoring force is in material an none in gravi

Q	3,000
resonant frequency	13 Hz
Θ	0.6 mrad



LIGO

Vibrational Thermal Noise of Monolithic Test Masses

Aaron Gillespie and Fred Raab

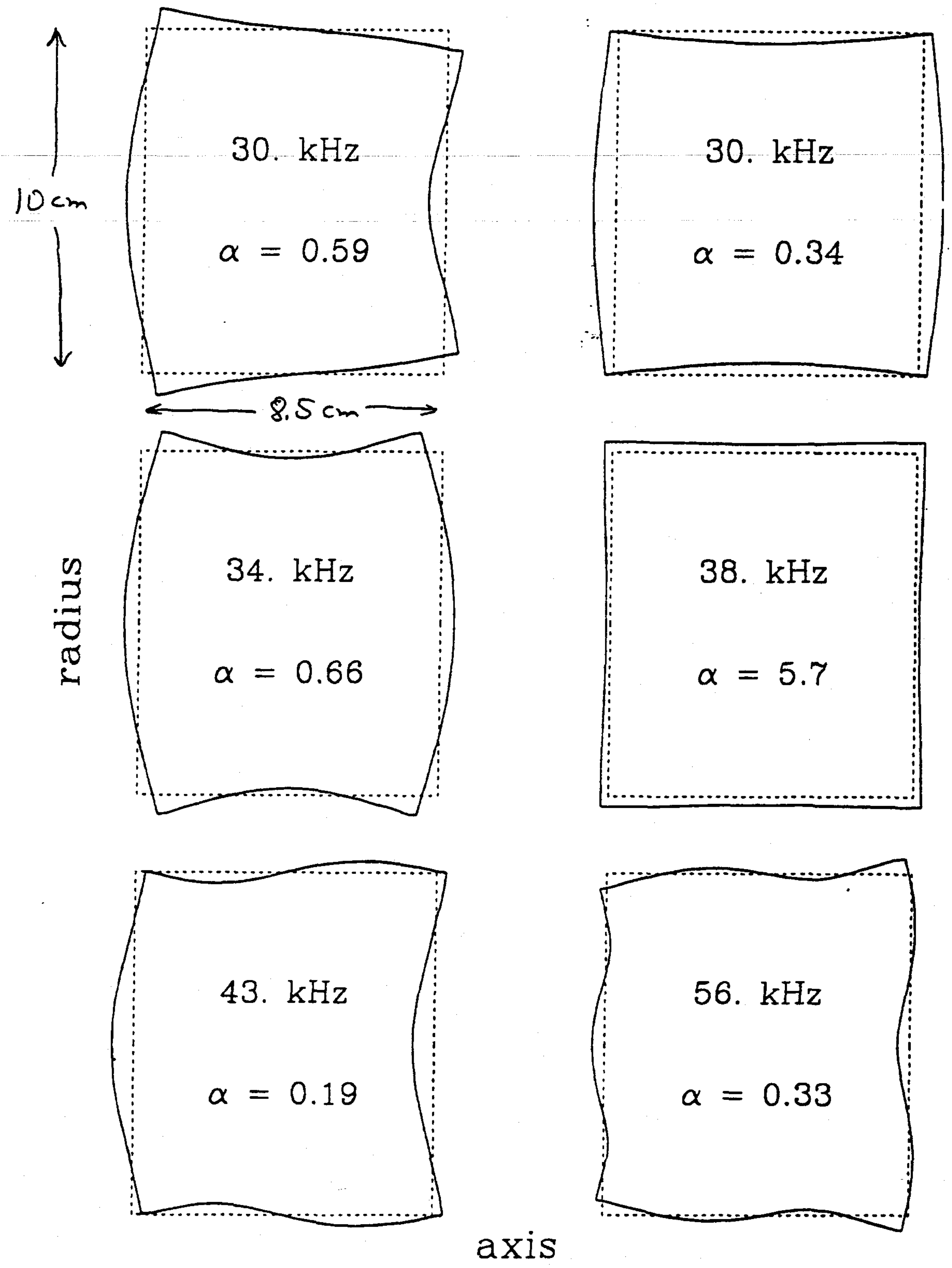
Goal: predict the thermal noise in interferometers due to vibrational modes of test masses:

$$\tilde{x}^2(f) \approx \sum_{n=1}^{\underline{N}} \frac{4k_B T \varphi_n(\omega)}{\alpha_n m \underline{\omega_n^2} \omega}$$

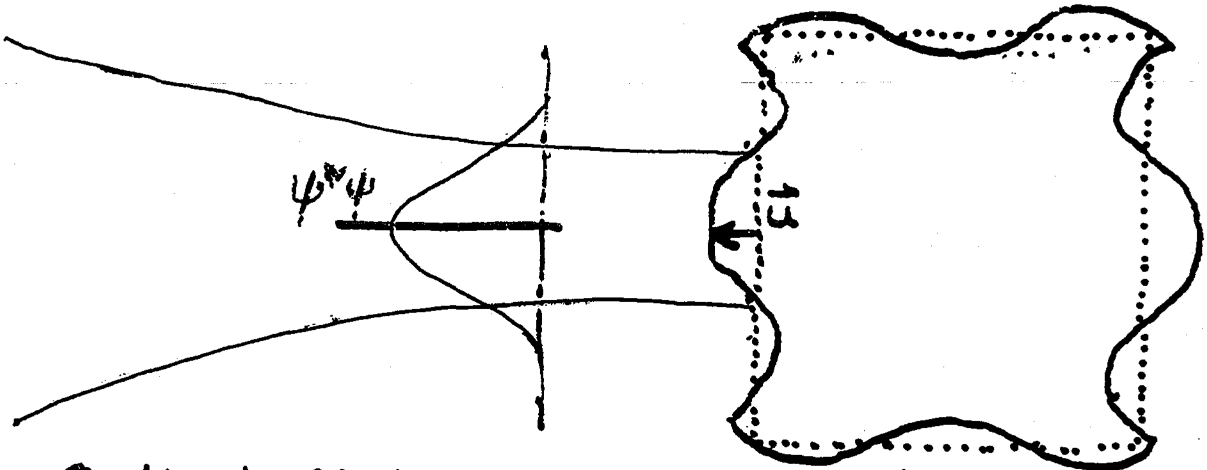
parameters to be determined:

- **the resonant frequencies of all the modes,** $\underline{\omega_n}$
- **the coupling between each mode and the interferometer signal, a parameter which we call the effective mass coefficient,** $\underline{\alpha_n}$
- **the damping of each mode,** $\varphi_n(\omega)$
- **the number of modes which must be included,** $\underline{N}$

Axisymmetric Modes of 40m Monolithic Test Mass



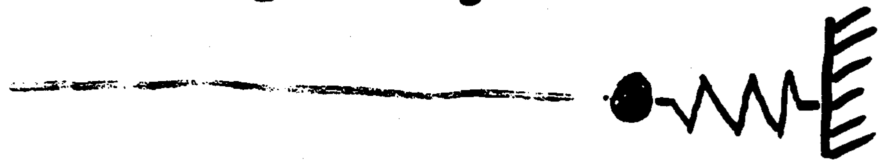
Calculation of Effective Mass Coefficient, α



Optical Mode		Vibration Mode	
Hermite Gaussian Function	$\psi_{lm}(\vec{r})$	Displacement Vector	$\vec{u}(\vec{r})$
Wave Vector	$\vec{k}$	Resonant Frequency	ω_0

$\Downarrow$

New System



ideal one-dimensional
laser beam

Simple Harmonic Oscillator
mass αm
frequency ω_0

Calculation of α

phase shift on reflection:

$$\psi_{00}(\vec{r}) \rightarrow \psi_{00}(\vec{r})\, e^{i2\vec{k}\cdot\vec{u}(\vec{r})}$$

write perturbed mode in terms of unperturbed modes:

$$\psi = \sum_i \sum_j c_{ij}\psi_{ij} \qquad c_{ij} = \int_S \psi_{ij}^* \psi_{00}\, e^{i2\vec{k}\cdot\vec{u}}\, dA$$

only TEM$_{00}$ part still resonates, so

$$\psi_{new} = \psi_{00}\left[\int_S \psi_{00}^* \psi_{00}\, e^{i2\vec{k}\cdot\vec{u}}\, dA\right]$$

$$\approx \psi_{00}\Big[1 + \underbrace{i2\int_S \psi_{00}^* \psi_{00}\, \vec{k}\cdot\vec{u}\, dA}_{\text{phase shift}} - \underbrace{2\int_S \psi_{00}^* \psi_{00} (\vec{k}\cdot\vec{u})^2 dA}_{\text{scattered light}}$$

Apperent Length Change:

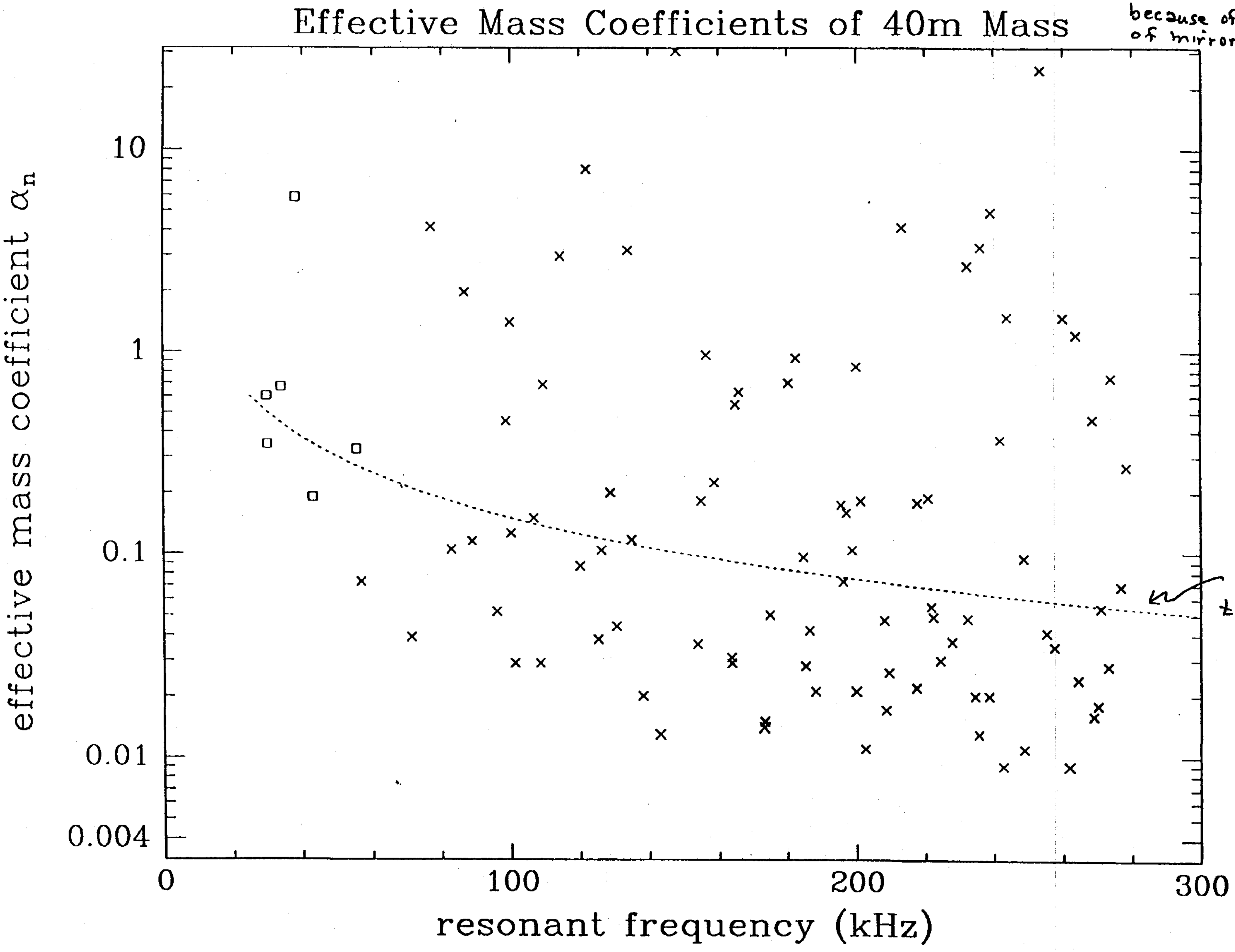
$$\Delta l_n = \frac{\int_S \psi_{00}^* \psi\, \vec{k}\cdot\vec{u}_n(\vec{r})\, dA}{|\vec{k}|}$$

Energy Normalization:

$$\alpha_n = \frac{\frac{1}{2}\int_M \rho\, \omega_n^2\, \vec{u}\cdot\vec{u}\, dV}{\frac{1}{2} m\, \omega_n^2\, \Delta l_n^2}$$

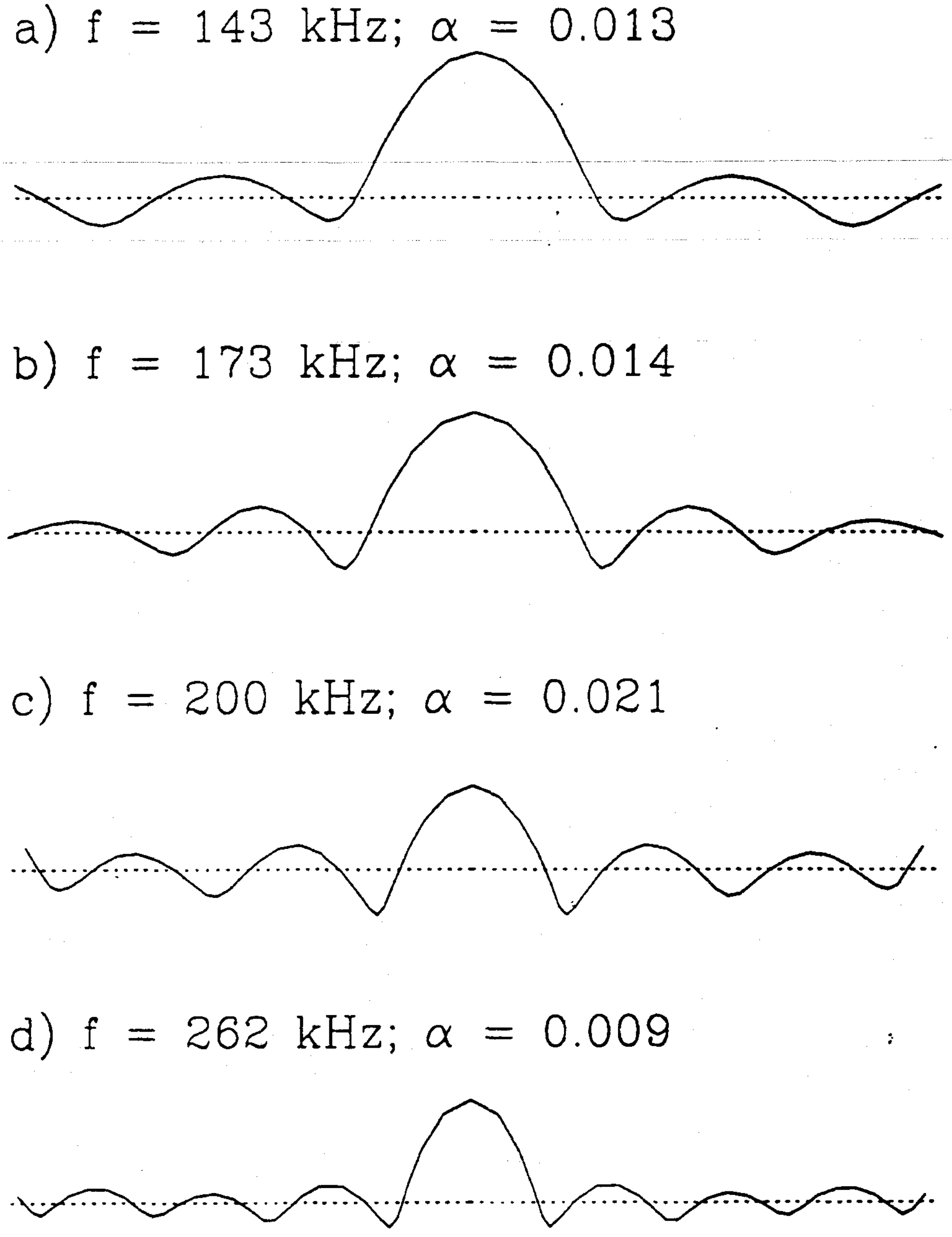
Squares □ are modes pictured in previous transparencies

Density of modes increases @ ω because of 2-dimensions of mirror face.



general trend ∝ 1/ω

Mirror Surfaces of Some Small α Modes



a) $f = 143$ kHz; $\alpha = 0.013$

b) $f = 173$ kHz; $\alpha = 0.014$

c) $f = 200$ kHz; $\alpha = 0.021$

d) $f = 262$ kHz; $\alpha = 0.009$

ϕ in Fused Silica

Current available data

ϕ(1kHz - 100 kHz) ~ (0.3 - 1.0)×10^{-7}

compression resonances in bulk samples

ϕ(~1Hz) ~ 1-2 ×10^{-7}

torsion resonances in fibers

routinely acheived

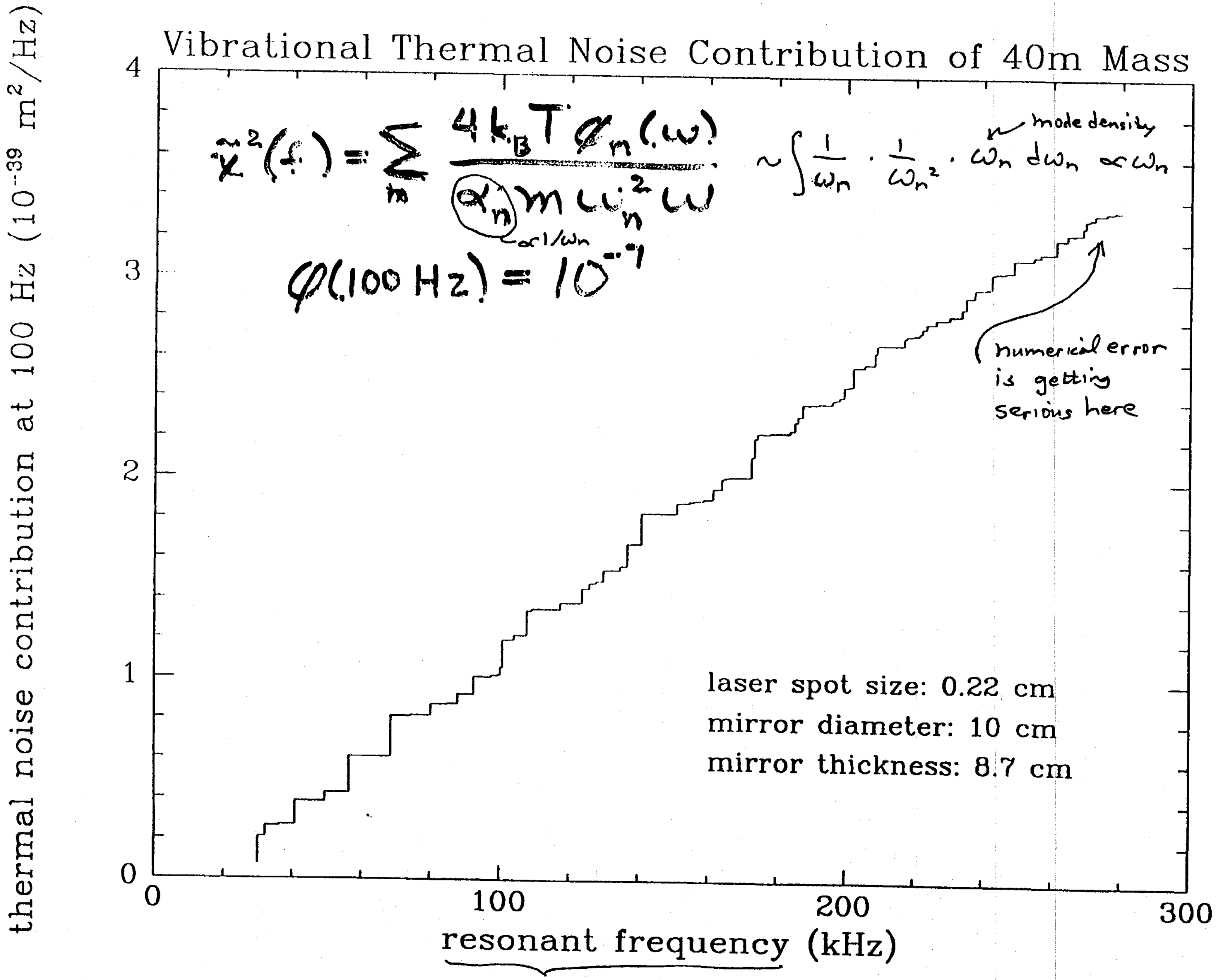
ϕ(100Hz) ~ no data !

ϕ believed to depend on surface effects, impurities

fundamental level of ϕ unknown for Fused Silica

Best Guess

ϕ(100Hz) ≈ 10^{-7}, independent of frequency



Vibrational Thermal Noise Contribution of 40m Mass

$$\tilde{x}^2(f) = \sum_n \frac{4 k_B T \phi_n(\omega)}{\underbrace{\boxed{\alpha_n}}_{\propto 1/\omega_n} m \omega_n^2 \omega} \sim \int \frac{1}{\omega_n} \cdot \frac{1}{\omega_n^2} \cdot \omega_n \, d\omega_n \propto \omega_n$$

mode density

$$\phi(100\,\text{Hz}) = 10^{-7}$$

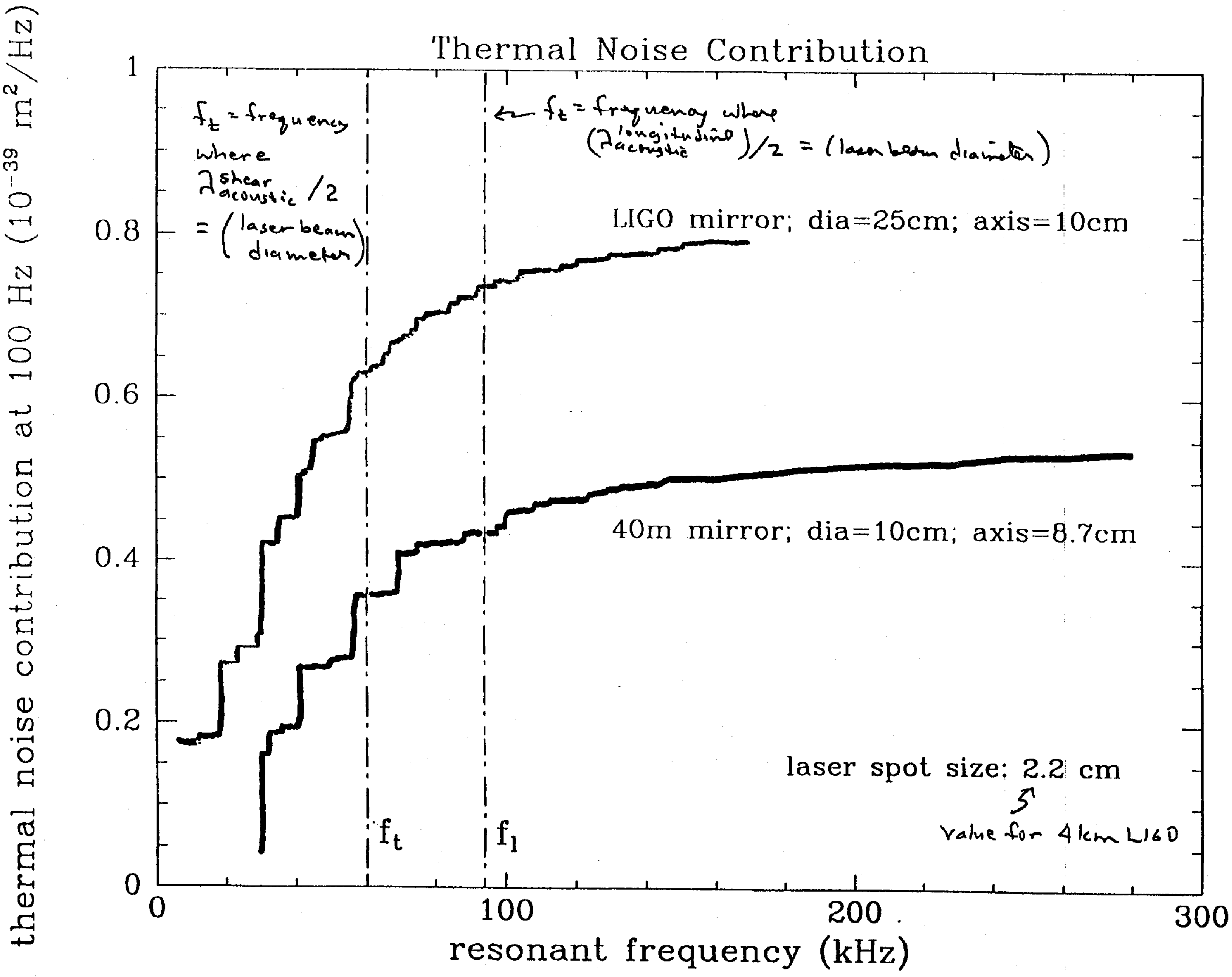
numerical error is getting serious here

laser spot size: 0.22 cm

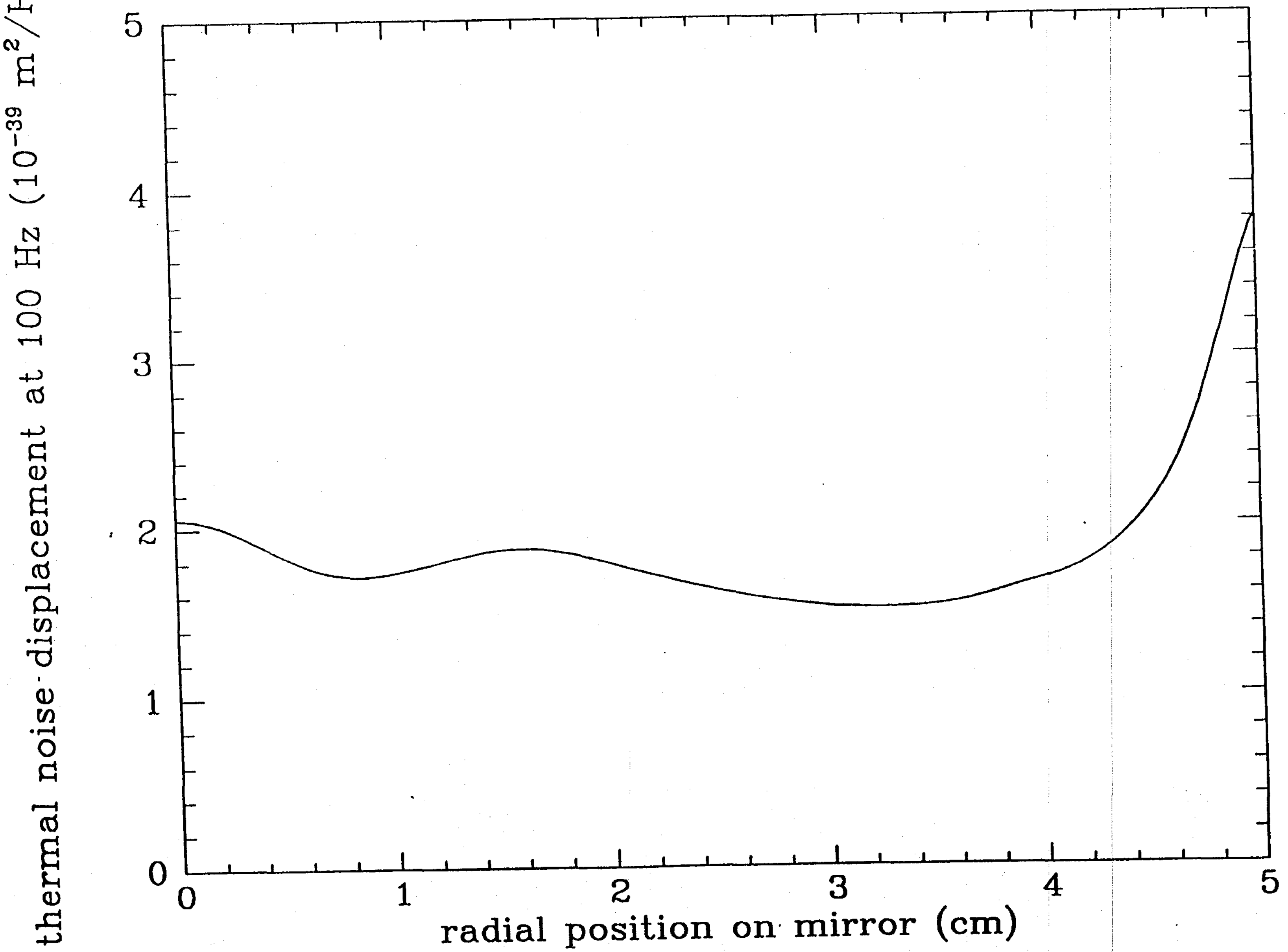
mirror diameter: 10 cm

mirror thickness: 8.7 cm

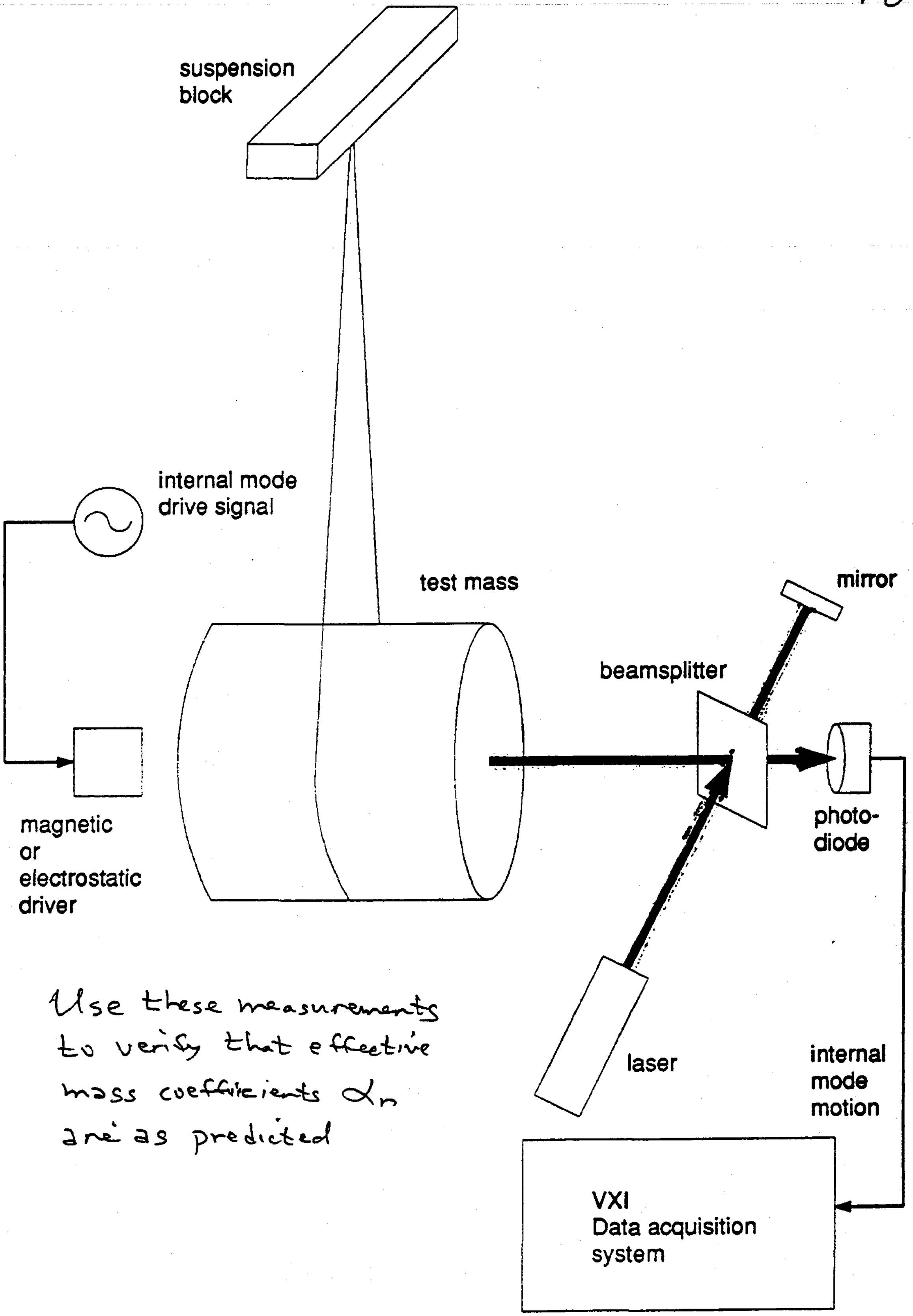
$\underbrace{\text{resonant frequency}}_{\uparrow\,\omega_n}$


Thermal Noise Contribution
thermal noise contribution at 100 Hz (10^{-39} m^2/Hz)
resonant frequency (kHz)
0
0.2
0.4
0.6
0.8
1
0
100
200
300
f_t = frequency where $\lambda^{shear}_{acoustic}/2$ = (laser beam diameter)
← f_t = frequency where $(\lambda^{longitudinal}_{acoustic})/2$ = (laser beam diameter)
LIGO mirror; dia=25cm; axis=10cm
40m mirror; dia=10cm; axis=8.7cm
laser spot size: 2.2 cm
value for 4 km LIGO
f_t
f_l

Thermal Noise Dependence on Position of Laser Spot on 40m Mass

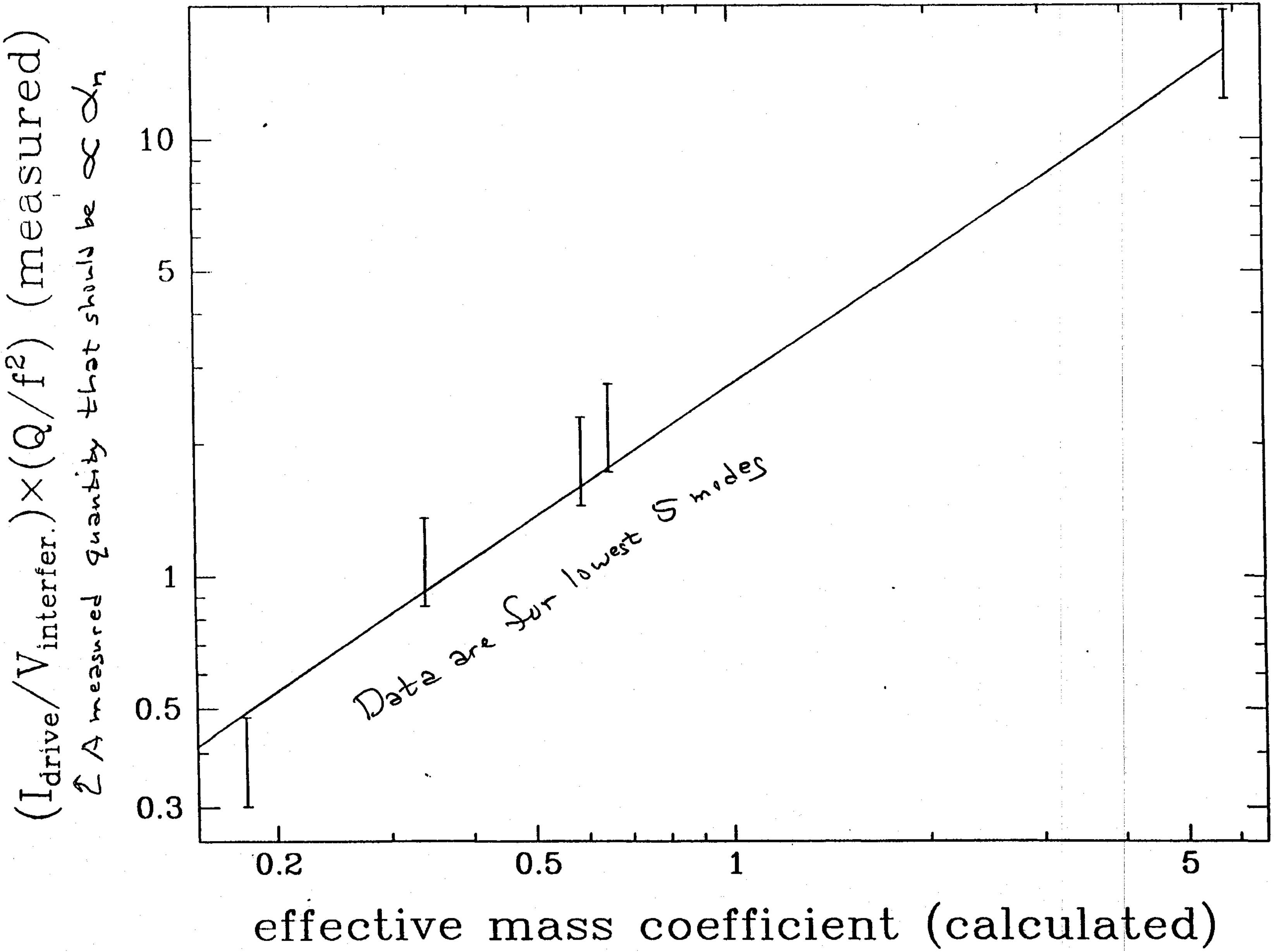


Suspension Development Apparatus


suspension block
internal mode drive signal
test mass
mirror
beamsplitter
photo-diode
magnetic or electrostatic driver
laser
internal mode motion
VXI Data acquisition system

Use these measurements to verify that effective mass coefficients α_n are as predicted

Measurement of Effective Mass Coefficients



Ø in Real Mirrors

Real test masses have

1) wires
2) coated surfaces
3) magnets

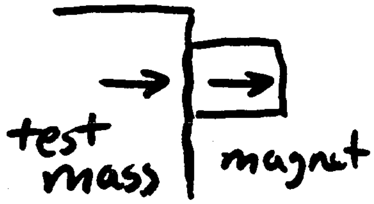
All of these effects can add loss which may affect each mode differently and may be frequency dependent.

Careful consideration of placement and geometry of these effects can minimize losses.

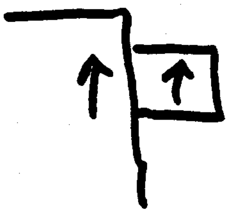
Example: Attaching magnets is knownto increase losses.

3 types of motion at magnets:

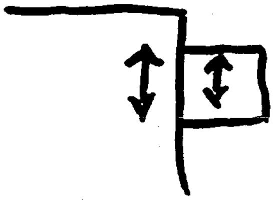
1) axial



2) radial

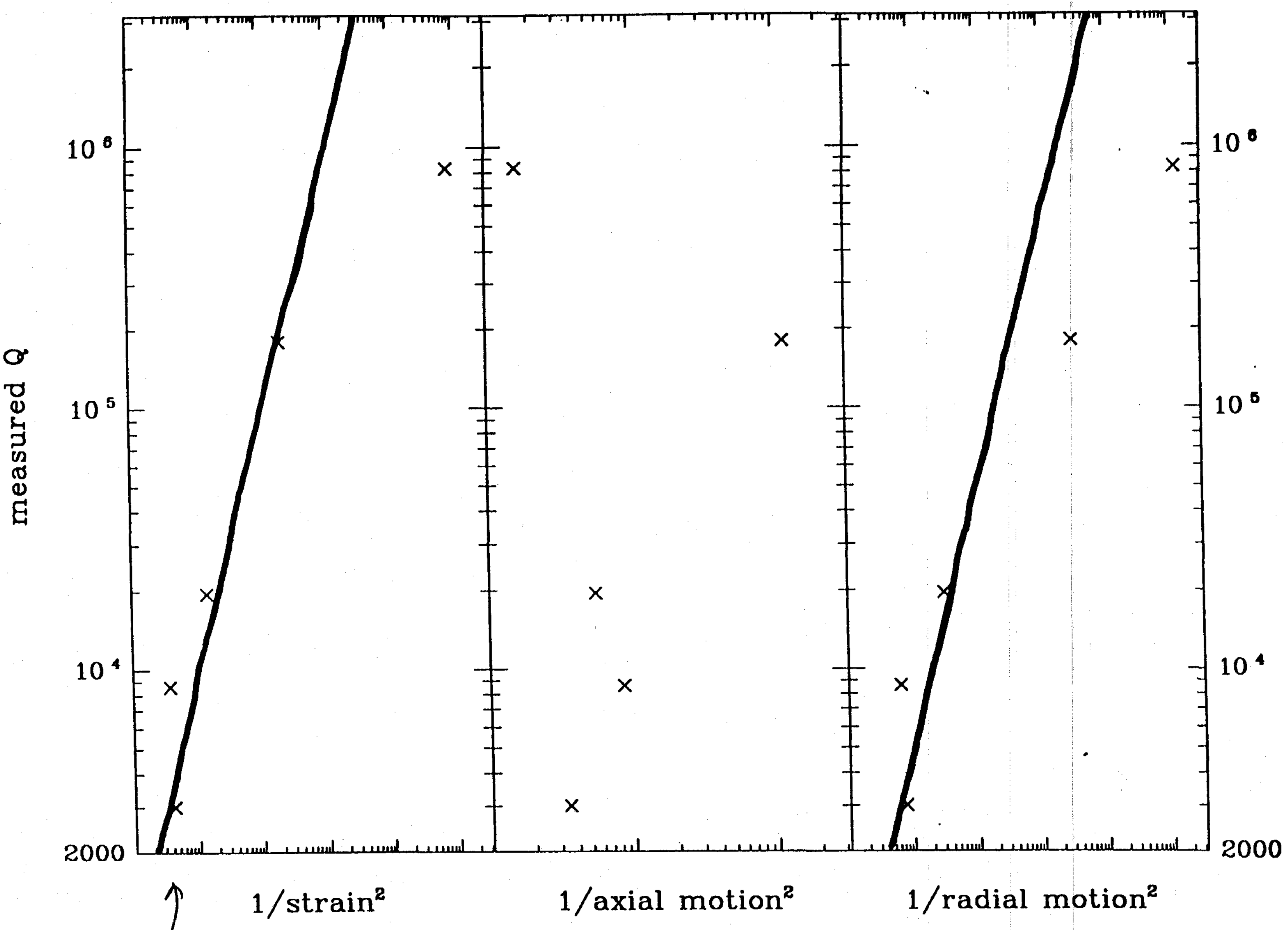


3) strain



By correlating motion to losses, we can find and fix loss mechanism!

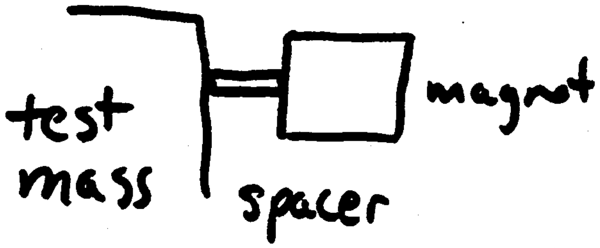
Damping Due to Magnets



correlation suggests that strain is culprit

Real World Loss Mechanisms

- magnet solution: put low loss spacer between magnet and test mass to reduce strain and improve damping by 100



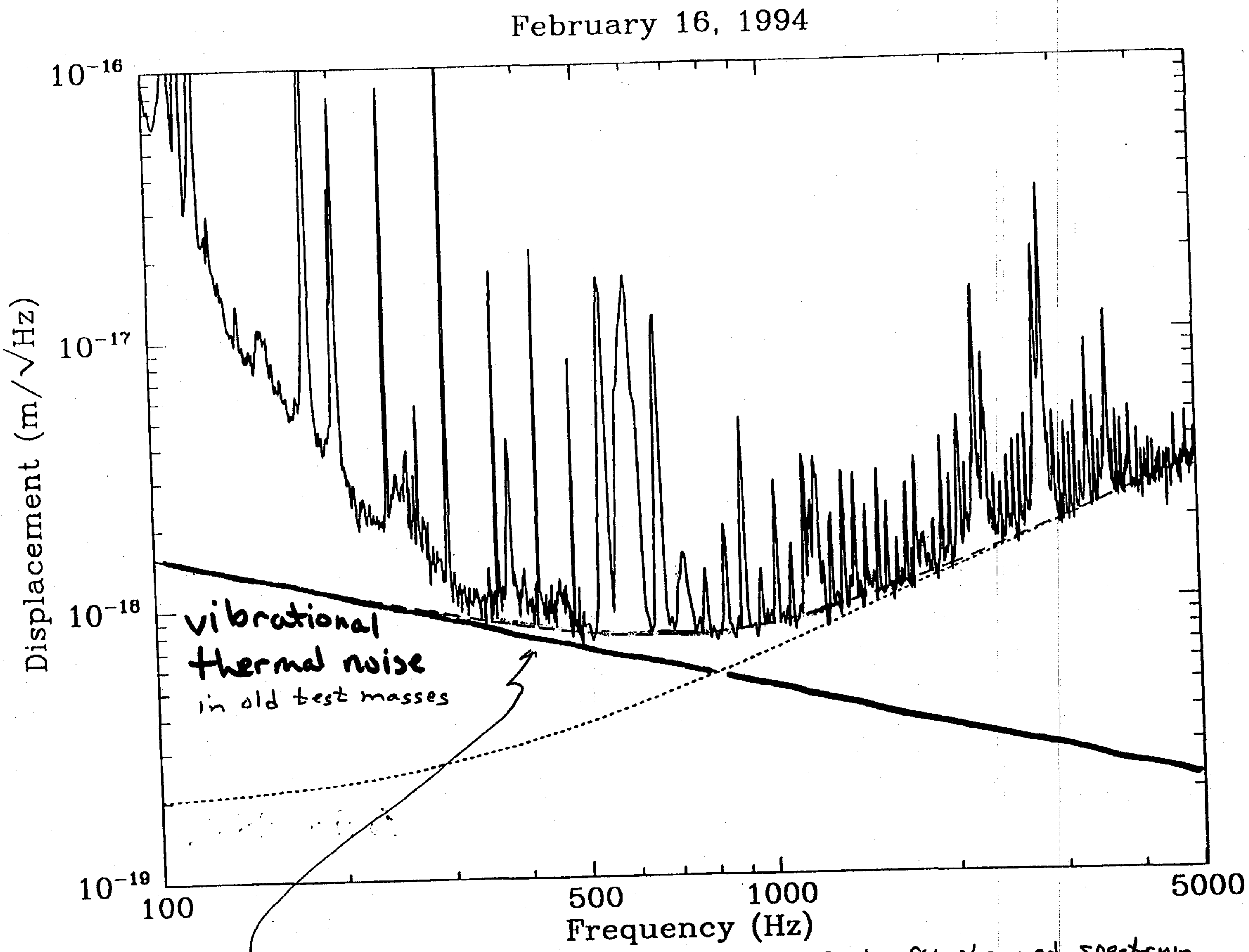
- other loss mechanism which has been found is [by Glasgow group] resonant coupling of energy to violin modes

- no loss has been found to be correlated with coated surfaces at level of

$\phi \sim 10^{-7}$

Good News!

Worry that mirror coating might strain the test mass, and thermal vibrations might then produce creep in the coating



level (but not slope) is adjusted by ~2 to fit observed spectrum.
– the computations of level are good only to factor ~2.

Other Ways of Measuring $\phi(\omega)$

1) Phase Measurements

$$\frac{\tilde{F}}{\tilde{x}} = -m\omega^2 + k(1+i\phi(\omega))$$

$$\text{Phase} = \text{Arctan}\left(\frac{\text{Im}(F/x)}{\text{Re}(F/x)}\right)$$

$$= \text{Arctan}\left(\frac{\phi(\omega)}{1-\omega^2/\omega_0^2}\right)$$

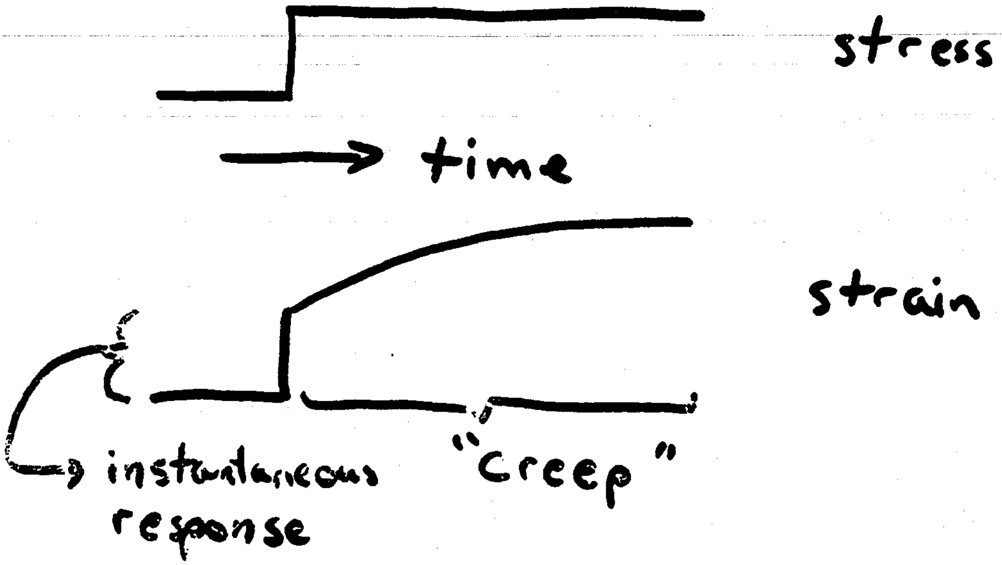
Proposed to measure pendulum losses near 10 Hz. ~ VIRGO project is attempting such measurements; very difficult

If $\phi \sim 10^{-7}$, $f = 10\,\text{Hz}$, $f_0 = 1\,\text{Hz}$,

Phase lag: $180 - 6\times10^{-8}$ Degrees

Requires 100 ps timing - routine in High Energy Physics, however more difficult to get timing on mechanical action.

2. time domain measurements



$$x = [\varepsilon_\infty + \psi(t)]\,F$$

There exists a transformation

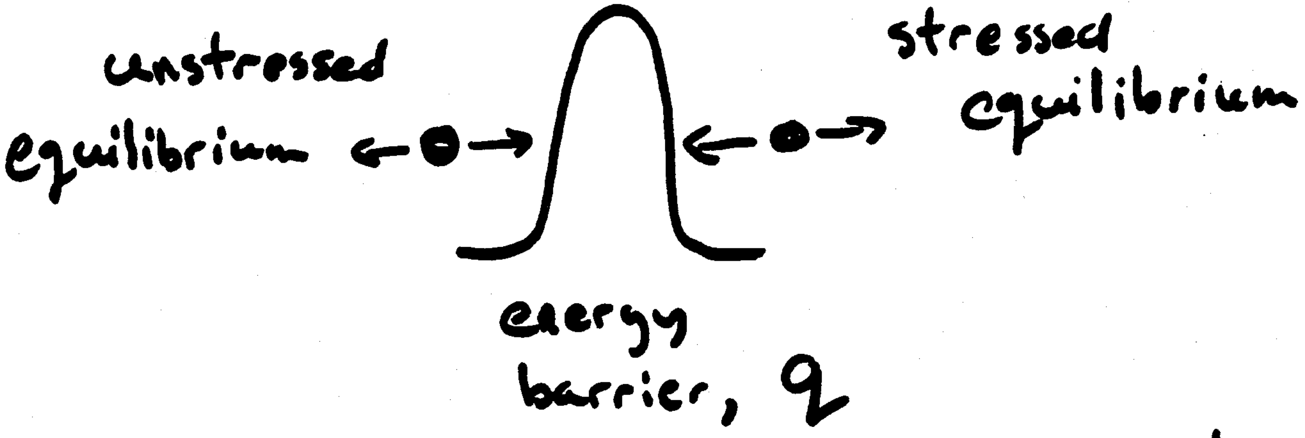
$$(\varepsilon_\infty, \psi(t)) \leftrightarrow (k, \phi(\omega))$$

2 The dominant contribution to ϕ at ω comes from $\psi(t)$ @ $t \sim 1/\omega$

Proposed application: measure $\psi(t)$ in Fused Silica over timescales of 1-100 ms to determine ϕ around 100 Hz.

3. Temperature Measurements

If you know that dominant damping mechanism is of the form of microscopic defect damping making DeBye peaks, then changing temperature is equivalent to changing frequency.



timescale to cross barrier

$$\tau \propto e^{q/k_BT}$$

remember $\left(\phi(\omega) \approx \Delta \frac{\omega\tau}{1+\omega^2\tau^2}\right)$

or

$$f \propto e^{-q/k_BT}$$

Example: Fused Silica has

$$\phi(50\,\text{kHz}) \approx 10^{-3} \text{ @ } 30\,\text{K}$$

due to absorption peak at 10^{12} Hz at 300K.

No active research in gravity in this area.

Issues for Advanced Detectors

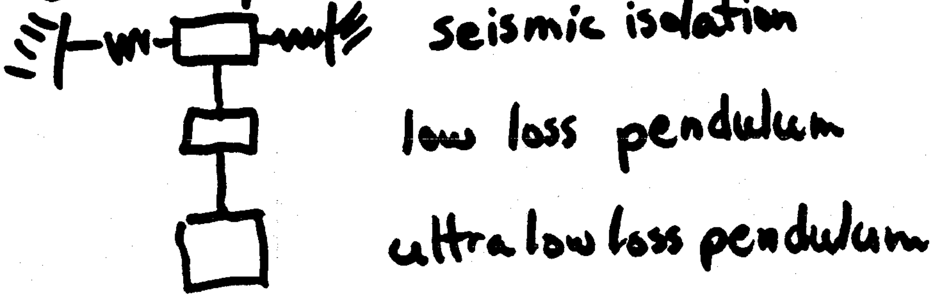
I) Suspension

A. External Damping Mechanisms

1. Residual Gas Damping: Plenty of margin to reduce pressure via better vacuum materials or larger pump speeds. EASY!

2. Eddy Current Damping: Specifications for initial detectors hard to meet. Advanced detectors will probably have to be built without magnets on mass. This requires new control systems. Moderately Hard!

3. Recoil Damping: Advanced pendulum cannot be attached directly to the damped seismic isolation stack. Requires double pendulum.



B. Suspension - Material Losses

1. Wide variety of metals available, some of which may be an order of magnitude better than steel.
   e.g. Tungsten, Niobium, Copper-Berrylium
   Easy short term solution to swap one wire for another, could be used in very early detectors (no technology change).
   But Need to spend resources to test each candidate material (tedious work) and probably only interim solution - will not meet advanced goals.

2. Use ribbons (wires with rectangular cross section) to reduce wire stiffness in beam direction.
   Problem: Does not help vertical mode thermal noise which quickly becomes dominant. Not much gain here.
   [See homework]

3. Use better material for suspension systems ie Fused Silica

Pendulum Q's in excess of 10^8 have been built with fused silica fibers. GOOD!

Pendulum fiber can be fused to mirror to make one monolithic piece and hence reduce potential for losses at the joint. GOOD!

Fused silica is brittle and hard to work with. BAD!

Fusing fiber to mirror may distort mirror optically. [Has not been explored experimentally] BAD!

Current Fused Silica suspensions use a single wire and thus do not fit with current control systems. — Will have to switch to several wires. BAD!

BUT all bad points are merely technical (ie can probably be fixed with time and money). Although fused silica probably won't be fully developed in time for initial detectors, it looks very promising for advanced detectors. VERY VERY VERY GOOD!

4. Lower Temperature

Instead of reducing ϕ, one can reduce noise by reducing T.

Some metals have the property that ϕ decreases as T decreases. BONUS!

One such metal is Niobium, which can have low temperature Q's of 10^8!

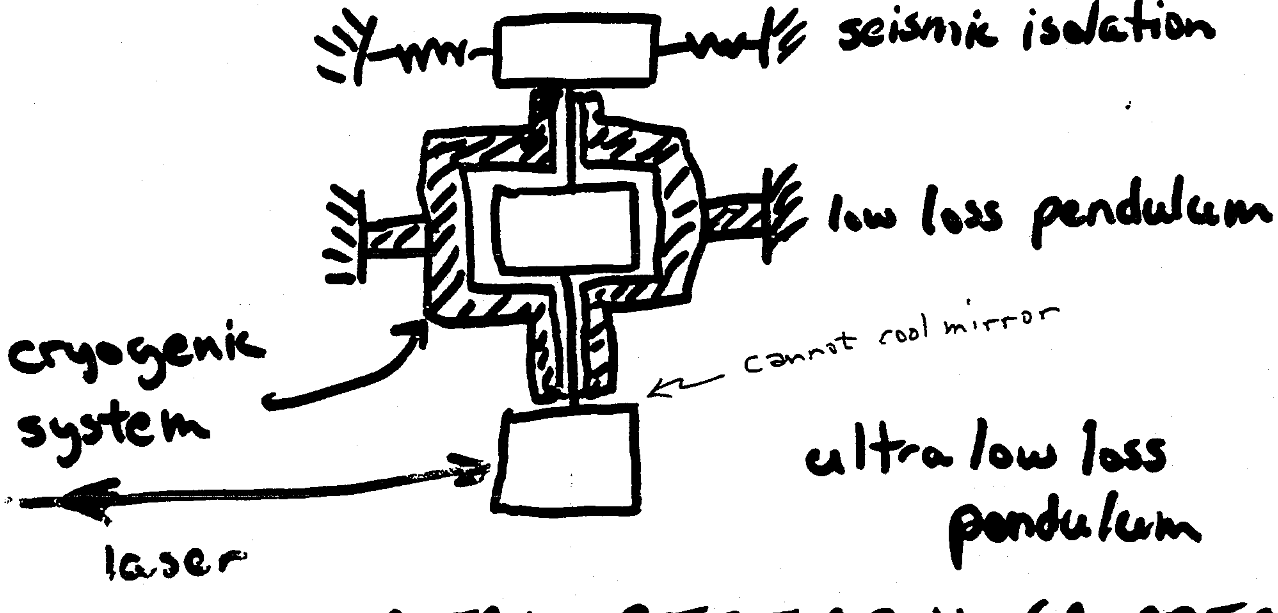
This is better than fused silica at room temp.

So, Niobium could give a pendulum

$$\phi \sim 10^{-11} \text{ at } 4K$$

or $\tilde{x}(10\,Hz) \approx 3\times10^{-21}\, m/\sqrt{Hz}$

Far exceeds advanced LIGO goals.



NO REAL RESEARCH (A DREAM)

II. Advanced Detectors: Vibrational Thermal Noise

Taking Initial LIGO geometry and $\phi(100\,Hz) \sim 10^{-7}$

$$\tilde{x}(100\,Hz) \sim 5\times10^{-20}\ m/\sqrt{Hz}$$

Advanced LIGO goal

$$\tilde{x}(100\,Hz) \sim 3\times10^{-21}\ m/\sqrt{Hz}$$

Need factor of 16 improvement.

1) Optimize geometry: Perhaps a factor of 2 or so.

2) Understand and Reduce ϕ: Fundamental limits of ϕ unknown. Possibility that increasing purity will decrease ϕ. Need factor of $8^2 = 64$! Very Hard!

- Lower Temperatures are no help; ϕ increases at T decreases.
- Possibility of other higher Q materials eg Sapphire, but these do not have necessary optical quality.

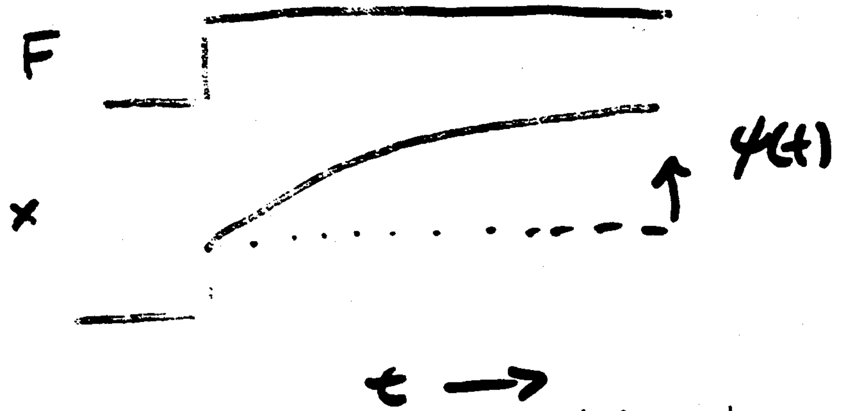
No clear way to achieve advanced goals; Vibrational thermal noise may limit advanced detectors.

Excess Noise

Excess noise consists of relatively rare but large relaxations which cause non-Gaussian noise bursts in gravity wave signal.

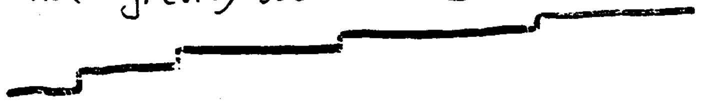
Fits into thermal noise because

1) property of materials
2) generally thermally activated
3) related to creep, $\psi(t)$ and hence $\phi(\omega)$.



Creep is generally logarithmic in time and can continue for years, eventually it is believed to become discreet.

If there are ~10 msec steps, than they would look like gravity wave bursts



What is happening?

1) Initially, individual vacancies and defects migrate under stress. There are so many of them that $\psi(t)$ appears continuous.

2) Vacancies and defects pile up at crystal grain boundaries, where there may be larger energy barriers.

3) Crystal grains rotate or slip past each other to realign themselves in stress. Some of those events will have large activation energies and hence large timescales.

4) The adjustment of a large grain may significantly change stress around it and cause an avalanche of smaller events. These may be big enough to be noticeable in interferometer signal if the response time of the avalanche is of order 10 ms.

Example: suppose a large grain ($1\mu m \times 1\mu m \times 1\mu m$) slips past another grain

mass $\sim 10^{-14}$ kg
distance $\sim 10^{-6}$ m

test mass recoils: $x = 10^{-6}\left(\frac{10^{-14}}{10\,kg}\right) \approx 10^{-21}$ m

Hypothotical - Never Been Seen

Summary

1) Thermal noise is a fundamental noise source which can easily be estimated once the damping mechanism is known.

2) Damping mechanisms can be difficult to understand and are the subject of current research.

3) Suspension thermal noise goals can be met for initial LIGO interferometers and very likely can be met or exceeded in advanced detectors.

4) Vibrational thermal noise goals for the advanced interferometers may be difficult to meet.

BATCH
START

(15) Light Scattering

STAPLE
OR
DIVIDER

Ph 103c: The Physics of LIGO

18 May 1994

LECTURE 15

Light Scattering and its Control

Lecture by Kip Thorne

Assigned Reading:

G. Chapter 7, "Diffraction" from the manuscript *Applications of Classical Physics* by Roger Blandford and Kip Thorne. [This material is needed as the foundation for the scattering analyses of Kip's lecture and for the Suggested Problem 2. at the end of this assignment. Sections 7.2 and 7.5 were assigned previously, in Lecture 4, and the manuscript was passed out then. If you have mastered the theory of diffraction, in some other course, in comparable detail to that given in this chapter, then you do not need to do this reading.]

Suggested Supplementary Reading:

SS. J. M. Elson, H. E. Bennett, and J. M. Bennett, "Scattering from Optical Surfaces," in *Applied Optical Engineering*, Vol. VII (Academic Press 1979), Chapter 7, page 191. [This was suggested previously, in Lecture 9. It is a review with few equations and with many references to the literature. The focus is on scattering from surfaces that are quite smooth (rms fluctuations in height much less than the wavelength of light, e.g., the LIGO mirrors).

l. Petr Beckmann and André Spizzichino, *The Scattering of Electromagnetic Waves from Rough Surfaces* (Macmillan/Pergamon, New York, 1963). [This is the classic treatise on the subject, with extensive equations. It deals with scattering from rough surfaces (rms fluctuations in height larger than a wavelength) as well as smooth ones. Unfortunately, it is written in such a way that one cannot readily understand later chapters without reading earlier ones.]

m. Kip S. Thorne, *Light Scattering and Proposed Baffle Configuration for the LIGO*, preprint GRP-200, available upon request from Kip. [This was the original, analytic calculation of the "gravity-wave" noise $\tilde{h}(f)$ caused by light scattering in LIGO both with and without baffles. It has two defects that make it not directly useful: (i) subsequent analytic calculations by Jean-Yves Vinet of the VIRGO Project ferreted out a serious error (a missing factor B inside the square brackets of Eqs. (4.6) and (4.7), which then propagated throughout GRP-200); and (ii) the final LIGO baffle configuration is rather different from the one in GRP-200. Kip's lecture is based on GRP-200, with the error corrected and the baffle configuration changed to the new one. The resulting noise spectrum $\tilde{h}(f)$, as discussed in Kip's lecture, is in good agreement with numerical simulations by the Breault Research Organization (BRO), under contract from LIGO. Eanna Flanagan and Kip are in the process of a final, thorough analytic reanalysis, which they plan to publish.]

A Few Suggested Problems:

1. *Backscatter off Baffles.* The dominant scattered-light noise source, according to the calculations by Kip, by Eanna Flanagan, by Jean-Yves Vinet, and by BRO, is backscatter off vibrating baffles; see the last of Kip's lecture transparencies.
   a. Give a list of factors that make the backscattered light coming from different directions superpose incoherently.
   b. Compute the "gravity-wave" noise $\tilde{h}(f)$ due to baffle backscatter, assuming incoherent superposition. Kip gives the answer (accurate to within a factor ~ 2) on his last transparency, when the mirrors are as close to the vacuum pipe wall as we expect them ever to be, $Y \simeq 20$ cm. You may prefer, for simplicity, to treat the case of mirrors centered in the beam tube, which has a radius $R = 60$ cm. [*Note:* In Kip's answer on his last transparency, $\alpha \lesssim 10^{-6}$ is the mirror's light-scattering coefficient (the probability for a photon to scatter from the main beam into a unit solid angle is $dP/d\Omega = \alpha/\theta^2$); $L =$ 4km is the length of the beam tube, l_1 =100m is the distance from the mirror to the nearest baffle, $\lambda = 0.4\mu$m is the wavelength of the laser light, $d\sigma/dAd\Omega \simeq 10^{-2}$ is the baffle's differential scattering cross section per unit area of baffle per unit solid angle into which the light goes (equivalently it is the probability that a photon, hitting the baffle at an angle of a few tens of degrees from its normal, gets backscattered into the direction from which it came); and $\tilde{\xi}(f)$ is the square root of the spectral density of the baffle's seismically induced displacement.
2. *Diffraction Off Baffles.* Consider the "gravity-wave" noise produced by diffraction of scattered light off vibrating baffles (the first process on Kip's next-to-the-last transparency.
   a. Compute $\tilde{h}(f)$ for the extreme worst-case scenario in which coherence increases the noise: Place the mirrors precisely at the center of the beam tube, assume the baffles are perfectly round and not serrated, and assume for each baffle that all points on the baffle's edge vibrate radially in phase with each other. Then light from all points on any chosen baffle will superpose coherently in $\tilde{h}(f)$. Give arguments why the various baffles should contribute incoherently with respect to each other. [Hint: one factor is the speed of sound along the vacuum pipe, which is $\sim$ 0.4km/sec; another deals with the baffle spacings.] Your final answer for $\tilde{h}(f)$ should be somewhat worse than the baffle backscatter noise of problem 1.
   b. The following factors mitigate the noise due to diffraction. For each factor make an estimate of the resulting reduction in $\tilde{h}(f)$. [Note that these mitigating factors do not act multiplicatively; the reduction in $\tilde{h}(f)$ is not equal to the product of the reductions due to the various factors. However, the net reduction makes $\tilde{h}(f)$ much less than baffle backscatter.] (i) The baffles will be serrated (jagged) with peak-to-valley serration heights of 3.5mm, which is somewhat larger than the width of a Fresnel zone (so as a baffle vibrates, some locations are alternately covering and uncovering an even numbered Fresnel zone, thereby producing phase shifts of one sign, while other locations are alternately covering and uncovering an odd numbered Fresnel zone, producing phase sifts of the opposite sign, and the two effects tend to cancel). There will be a $\sim$ 5 per cent irregularity in the

serrations on scales $\sqrt{\lambda L} \sim$ 4cm. (ii) The mirrors will generally not be centered in the vacuum pipe, but rather will be off center by $\gtrsim 10$; and as a result, different regions of a baffle will intercept different Fresnel zones. (iii) The various points on a baffle do not vibrate in phase with each other. (iv) Each baffle will be out of round by a few millimeters in some random way.

Lecture 15
Light Scattering and its Control

by Kip Thorne, 18 May 1994

Thorne lectured from the following transparencies. His lecture covered only the first half of the class on 18 May.

Lecture 15

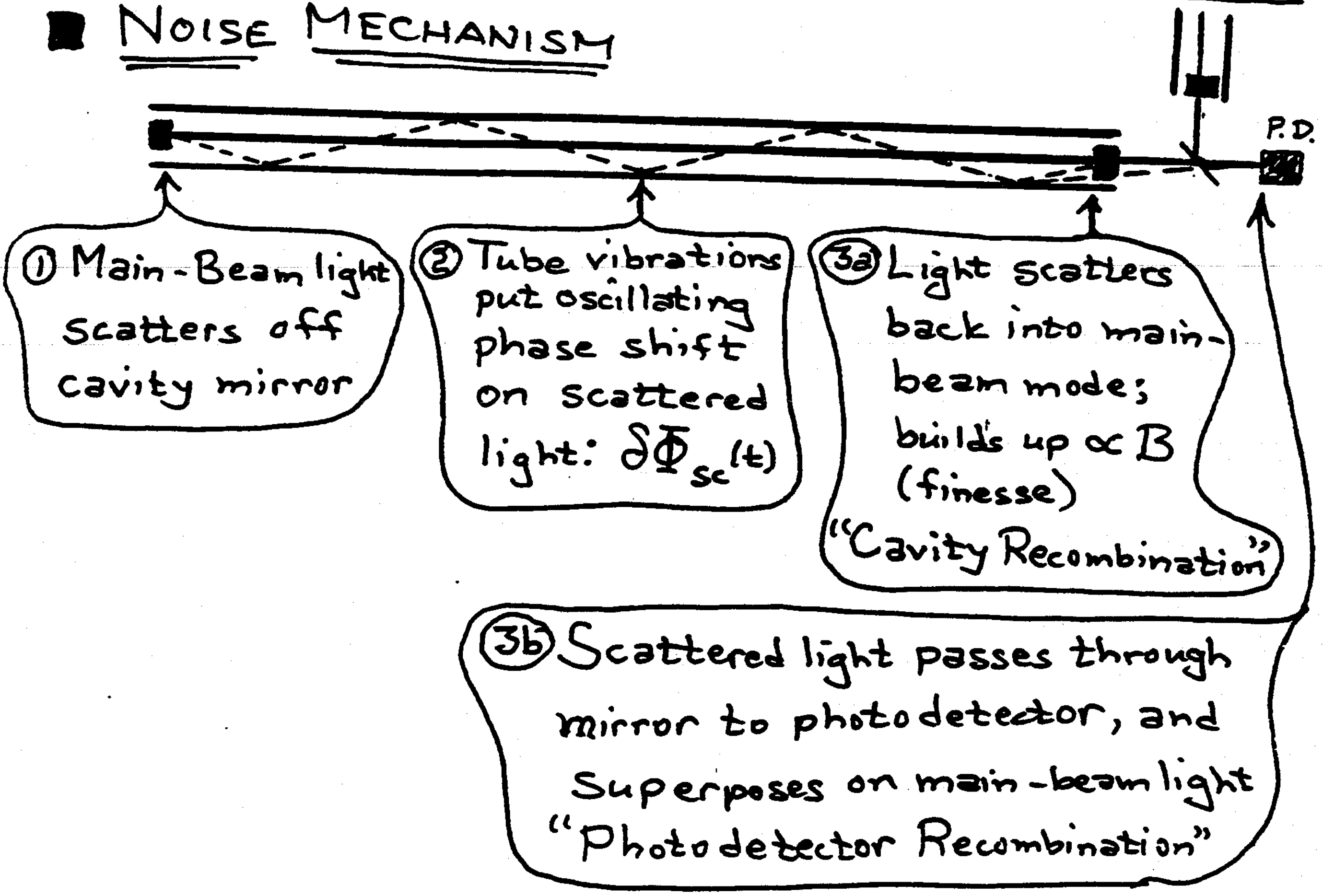
"Light Scattering and Its Control"

by

Kip Thorne

STRAY LIGHT & SCATTERING [OVERVIEW]

■ NOISE MECHANISM



① Main-Beam light scatters off cavity mirror

② Tube vibrations put oscillating phase shift on scattered light: $\delta\Phi_{sc}(t)$

③a Light scatters back into main-beam mode; builds up $\propto \mathcal{B}$ (finesse) "Cavity Recombination"

③b Scattered light passes through mirror to photodetector, and superposes on main-beam light "Photodetector Recombination"

$$\text{Photodetector Current} \propto \int |\Psi_{mb} + \Psi_{sc} e^{i\delta\Phi_{sc}(t)}|^2 \eta \, dA$$

$$= \text{constant} + \underbrace{2\,\mathrm{Im}\int \Psi_{mb}\Psi^*_{sc}\,\delta\Phi_{sc}(t)\,\eta\, dA}_{\text{interpreted as } \propto h(t)}$$

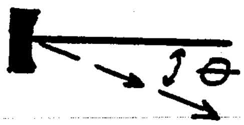
$$\Rightarrow \boxed{\tilde{h}(f) \propto \delta\tilde{\Phi}_{sc}(f)}$$

- Differs from usual light-scattering noise by the essential role of the fluctuating phase shift, $\delta\Phi_{sc}(t)$
- DC scattered light is NOT a problem.

Mitigation

- Minimize Scattering:

$\boxed{\text{Supermirrors}}$



$$\frac{dP_{\text{scatter}}}{d\Omega} \simeq \frac{\alpha}{\theta^2}; \quad \alpha \lesssim 10^{-6}; \quad \theta \sim 10^{-2} = \frac{1m}{100m} \text{ to } \sim 3\times10^{-4} = \frac{1m}{3km}$$

$$TIS \simeq 2\pi\alpha \ln\left(\frac{\theta_2}{\theta_1}\right) \lesssim 2\times10^{-5}$$

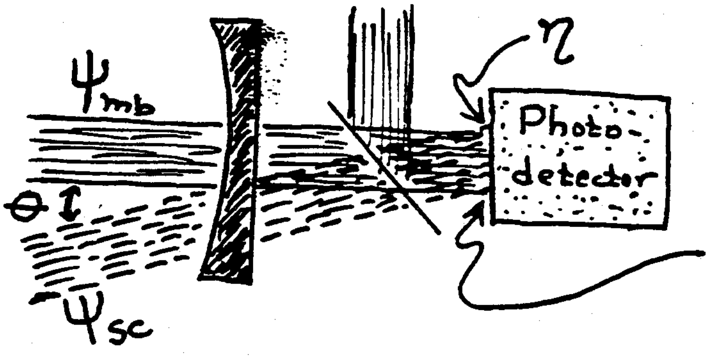
Small θ very important

- Minimize Cavity Recombination

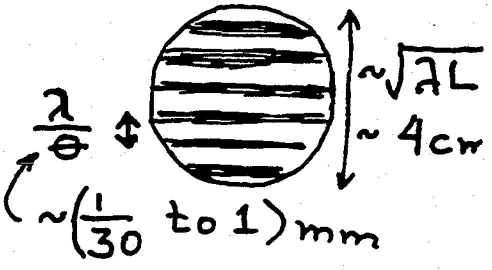
$\boxed{\text{Supermirrors}}$ $P_{rec} \simeq \frac{\alpha}{\theta^2} \cdot \frac{\lambda L}{L^2}$

- Minimize Photodetector Recombination

$$h(t) \propto \int \Psi_{mb} \Psi_{sc}^{*} \eta \, dA$$



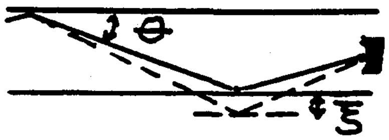
Fringes:



$\boxed{\begin{array}{l}\bullet \text{ Spatially smooth } \eta \\ \bullet \text{ Output mode cleaner}\end{array}}$

SPECIAL DANGERS AND THEIR MITIGATION

Reflections on Tube Wall



$$\delta\Phi = 4\pi\theta\frac{\xi}{\lambda} \leftarrow \text{(10×ground)}$$

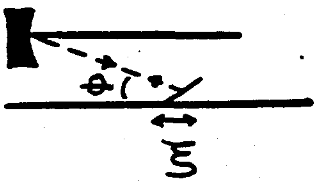
- Mitigation:

  Baffles to remove reflections with $\theta < \theta_0 \simeq 0.02$ radians

  Rough Walls to attenuate light with $\theta > \theta_0$

- Price:

  Backscatter from baffles



$$\delta\Phi = \frac{4\pi\xi}{\lambda}$$

- Mitigation:

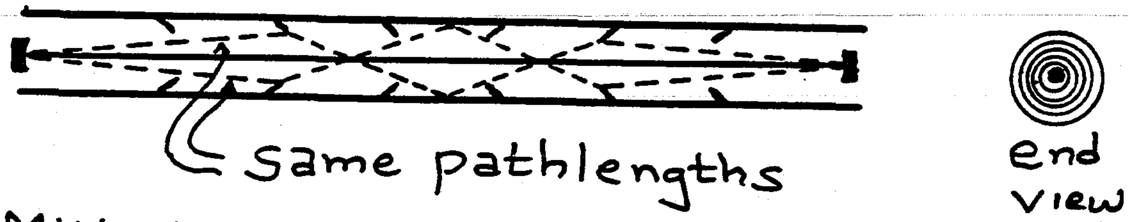
  "Black" baffles

■ SPECIAL DANGERS AND THEIR MITIGATION

[continued]

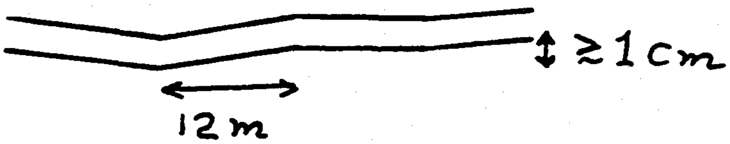
Coherent Scattering

especially if
- mirror is at center of tube
- tube is perfectly straight
- tube & baffles are perfectly round

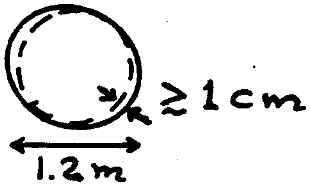


- Mitigation:

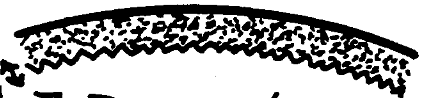
$\boxed{\text{Crooked tube}}$



$\boxed{\text{tube \& baffles out of round}}$



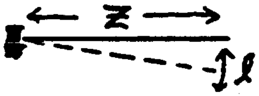
$\boxed{\text{baffles serrated slightly irregularly}}$



$\gtrsim 3.5$ mm > (Fresnel Zone Width)

≳5% irregularity on scales of few cm

- Natural Mitigation:
  - Incoherence of wall & baffle motions $\rightarrow$ incoherence of $\delta\Phi$
  - Fresnel-zone action if mirror is off center
  - Transverse incoherence of scattered light

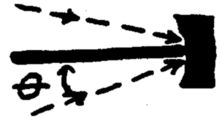


coherence length: $\ell_c \gtrsim \frac{z}{L}\sqrt{\lambda L}$

QUANTITATIVE DETAILS

- **Analytic Analysis** [by Thorne; Flanagan]
  - Amplitude analysis, allowing for coherence
  - Intensity analysis, assuming incoherence
  - Semi-Independent, analytic intensity analysis [by Vinet, of VIRGO Project]
    - found one significant error
- **Monte Carlo Intensity Analysis**
  - LIGO: Weiss & Whitcomb ↔ BRO
  - VIRGO
- **Foundational Formula for Intensity Analysis**

$$\tilde{h}(f) = \frac{\lambda}{2\pi BL}\left[\frac{dI_{recombined}/df}{I_{main\ beam}}\right]^{1/2}$$



↑ effective # of bounces = 4/(1-R)

$$\boxed{\tilde{h}(f) = \frac{\lambda}{2\pi BL}\left[\int P_{\substack{rec.\\ B.U.}}(\theta)\,\frac{dE_{return}/dt\,dA\,d\Omega\,df}{I_{mb}/\lambda L}\,d\Omega\right]^{1/2}}$$

↑

(Probability to scatter photon photon back into main-beam mode) x (subsequent energy buildup as if main beam were absent)

Cavity Recombination & Buildup

$$\boxed{P_{\substack{\text{rec}\\ \text{B.U.}}} = \frac{\alpha}{\theta^2}\cdot\left(\frac{\lambda L}{L^2}\right)\cdot B^2}$$

$\sim 10^{-4}$ typical ("Advanced" detectors)

$\frac{dP_{scatt}}{d\Omega} \sim 1$

$(\Delta\Omega)_{MB} \sim 10^{-10}$

$\sim 10^{6}$

Photodetector Recombination

[referred back to equivalent recombination in cavity]

$$\boxed{P_{rec} = \frac{[\sqrt{\nu}\,\tilde{\eta}(\nu)]^2}{\theta}\sqrt{\frac{\lambda}{L}}}$$

$\sim 10^{-6}$ to 10^{-8} typical [I had expected 10^{-4}]

$\sim 10^{-3}$

$\sim 10^{-5}$

$$\left(\begin{array}{l}\eta_{rms} \text{ @ wave number } \nu = \theta/\lambda \\ \qquad\qquad \sim(10 \text{ to } 300/\text{cm}) \\ \text{in bandwidth } \Delta\nu \simeq \nu\end{array}\right)^2$$

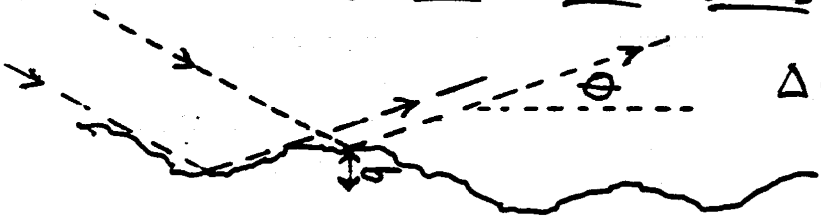
$\sim \underbrace{10^{-4} \text{ to } 10^{-6}}_{\text{Weiss}}$ [I had expected 10^{-2}]

- with such smooth photodetectors:

$$\boxed{\text{Cavity recombination is likely to dominate even without an output mode cleaner}}$$

QUANTITATIVE DETAILS [CONTINUED]

■ Reflections off Tube Walls



$$\Delta\Phi = \frac{2\pi}{\lambda}\cdot 2\sigma\theta$$

Rayleigh Criterion: $\frac{4\pi}{\lambda}\sigma\theta << 1$: Highly

$\frac{4\pi}{\lambda}\sigma\theta << 1$: Highly Reflecting

$>> 1$: Poorly Reflecting

Dividing line: $\theta_{crit} = \frac{\overset{\leftarrow 0.4\mu m}{\lambda}}{\underset{\nwarrow 5\mu m}{4\pi\sigma}} = 0.006 \text{ rad}$

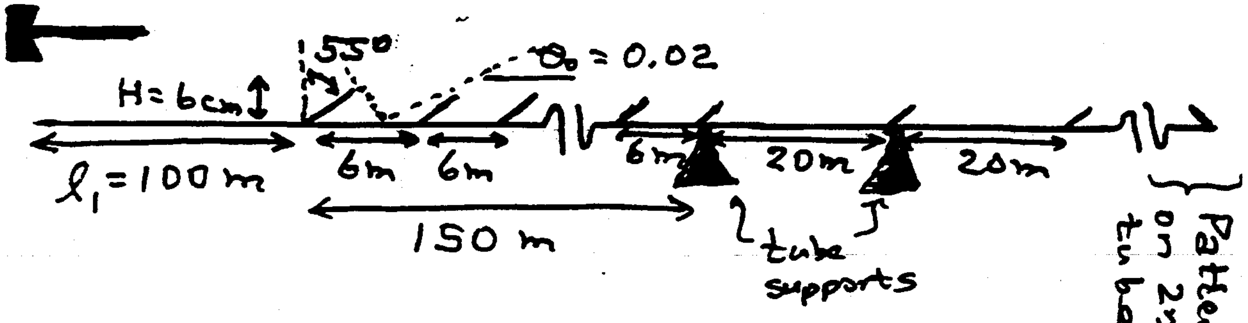
Construct Baffles to intercept all light with $\theta < \theta_0 \simeq 0.02 \text{ rad} = 3\,\theta_{crit}$

For $\theta \simeq \theta_0$:

$$\left(\begin{matrix}\text{Number of}\\ \text{reflections}\end{matrix}\right) = \mathcal{N} \simeq \frac{\overset{4km}{L}\,\overset{0.02}{\theta_0}}{2\underset{60cm}{R}} \simeq 70$$

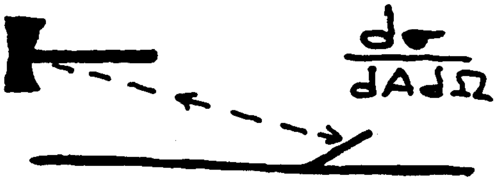
$$\left(\frac{\text{Surviving flux}}{\text{incident flux}}\right) \simeq \underset{< 0.8}{R}^{\mathcal{N}} \ll 10^{-7}$$

Baffle Configuration



} Pattern reversed on 2nd half of tube

Scattering off baffles



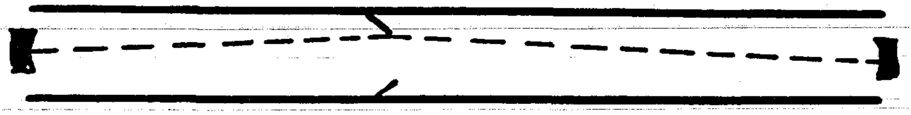
$$\frac{d\sigma}{dA\,d\Omega} = \frac{dP}{d\Omega} \simeq$$

0.01 /sterr
wall material
0.001/sterr
"Martin Black"

QUANTITATIVE DETAILS [CONTINUED]

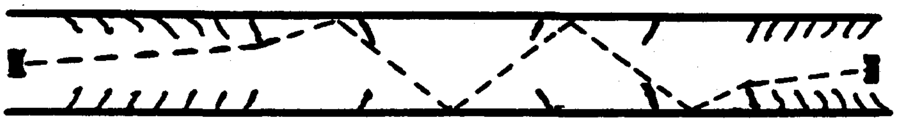
SCATTERING PATHS

- Diffraction:



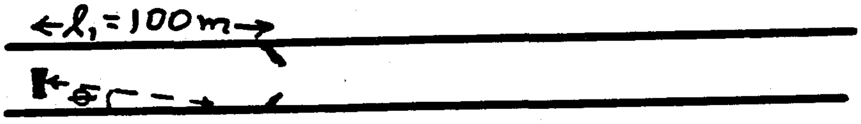
$\ll$ baffle backscatter
[Suggested Problem 2]

- Diffraction-Aided Reflection:



For chosen baffles
& wall roughness } less important than baffle backscatter

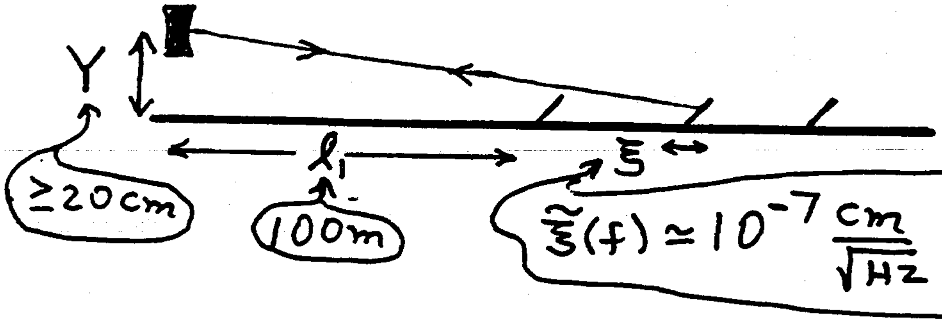
- Scatter off Nearby Tube:



Also less important than baffle backscatter

• Back Scatter off Baffles

[The Dominant Noise Source]



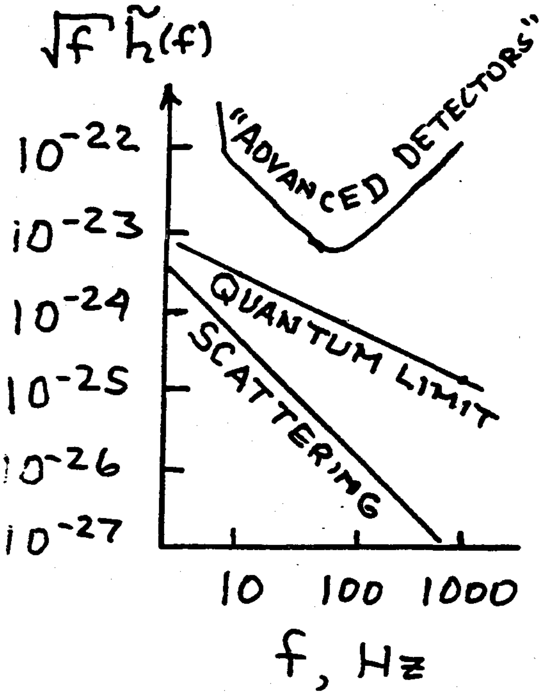
$$\tilde{\xi}(f) \simeq 10^{-7}\frac{\text{cm}}{\sqrt{\text{Hz}}}\left(\frac{10\,\text{Hz}}{f}\right)^2$$

Cavity Recombination

$$\boxed{\tilde{h}(f) \simeq 4\underbrace{\alpha}_{\lesssim 10^{-6}}\underbrace{\sqrt{\ln\left(\frac{L}{l_1}\right)}}_{2}\,\underbrace{\frac{\partial}{Y}}_{\lesssim 2\times10^{-6}}\underbrace{\sqrt{\frac{d\sigma}{dA\,d\Omega}}}_{10^{-2}}\;\frac{\tilde{\xi}(f)}{L}}$$

$$\lesssim \frac{3\times10^{-25}/\sqrt{\text{Hz}}}{(f/10\,\text{Hz})^2}$$

$$= \frac{\text{Standard Quantum Limit, }10^3\text{kg}}{10\times(f/10\,\text{Hz})}$$



BATCH
START

16. Squeezed Light

STAPLE
OR
DIVIDER

Lecture 16
Squeezed Light and its Potential Use in LIGO

by Jeff Kimble, 18 & 20 May 1994

Kimble lectured from the following transparencies, which Kip has annotated a bit. Kimble's lecture came in two parts, one covering the second half of the class on 18 May; the other covering the full class on 20 May.

LECTURE 16

Squeezed Light and its Potential Use in LIGO

Lecture by H. Jeff Kimble

Assigned Reading:

TT. C. M. Caves, "Quantum mechanical noise in an interferometer," *Phys. Rev. D*, **23**, 1693–1708 (1981).

UU. D. F. Walls, "Squeezed states of light," *Nature*, **306**, 141–146 (1983).

VV. M. Xiao, L. A. Wu, and H. J. Kimble, "Precision measurement beyond the shot-noise limit," *Phys. Rev. Lett.*, **59**, 278–281 (1987).

Suggested Supplementary Reading:

l. H. J. Kimble, "Quantum fluctuations in quantum optics—Squeezing and related phenomena," in *Fundamental Systems in Quantum Optics*, eds. J. Dalibard, J. M. Raimond, and J. Zinn-Justin, (Elsevier, Amsterdam, 1992), pp. 545–674.

m. "Squeezed States of the Electromagnetic Field," Feature Issue, *J. Opt. Soc. Amer.*, B4, 1450–1741 (1987).

n. "Squeezed Light," Special Issue, *J. Modern Optics*, **34**, 709–1020 (1987).

o. "Quantum Noise Reduction," Special Issue, *Appl. Phys. B*, **55**, 189ff. (1992).

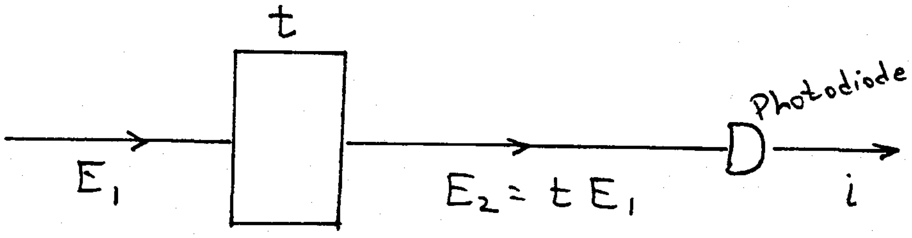
p. S. Reynaud, A. Heidman, E. Giacobino, and C. Fabre, "Quantum fluctuations in optical systems," in *Progress in Optics*, XXX, ed. E. Wolf (Elsevier, 1992), pp. 1–85.

A Few Suggested Problems:

1. *Detection of Modulation in a Squeezed State.* An electromagnetic field propagates through a medium whose transmission coefficient is given by $t = t_0 e^{-\gamma(t)}$, where $\gamma(t) \equiv \gamma_0 \cos(\Omega_0 t)$ (i.e., sinusoidally modulated absorption with amplitude γ_0 and frequency Ω_0).

   a. Assuming that $\gamma_0 \ll 1$ and that the input field is in a coherent state (with frequency $\gg \Omega_0$), derive an expression for the minimum detectable value of γ_0, for a fixed input energy flux $\langle |E_1|^2 \rangle$ and a fixed bandwidth $B \equiv \Delta f$ (corresponding to a photodiode integration time $\hat{\tau} = 1/B$).

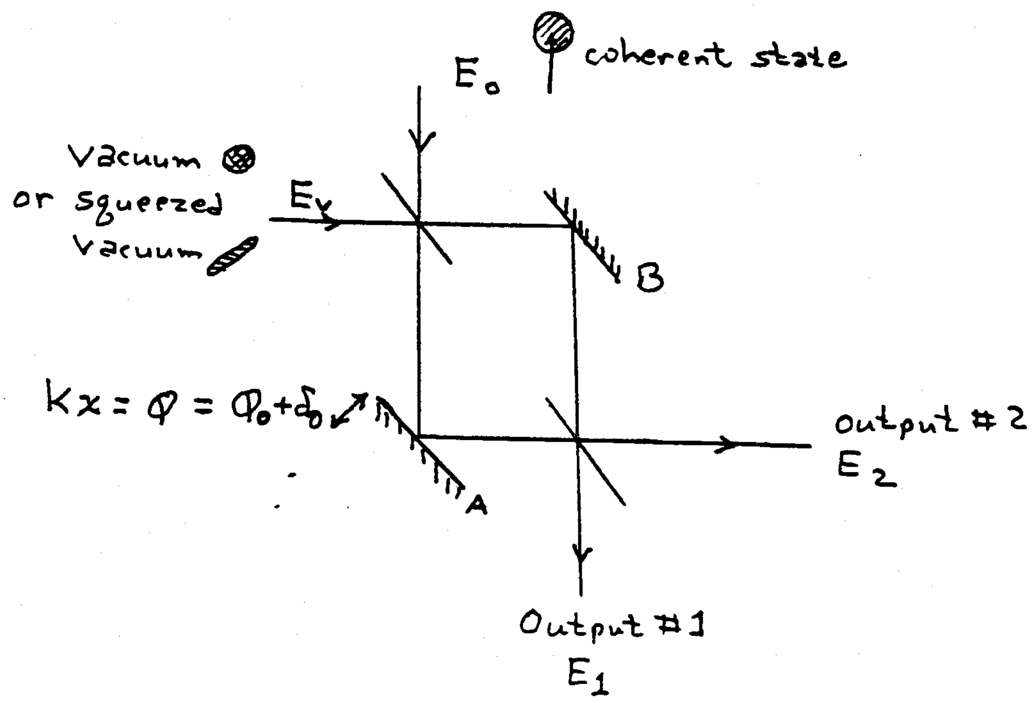
   b. If the input field instead is in a squeezed state, derive an expression for the minimum detectable amplitude γ_0. Illustrate in a "ball-and-stick" sketch the dependence of your answer on the orientation of the squeezing ellipse.



2. *Squeezed Vacuum in an Interferometer.* In Part IV of Kimble's lecture transparencies, he sketches a calculation of the minimum detectable phase deviation δ_0 when a coherent state is put into one port of the Mach-Zehnder interferometer shown below, and either the vacuum state or the squeezed vacuum state is put into the other port. His answer was $\delta_0 = 1/\sqrt{N}$ for the vacuum state, and $\delta_0 = (1 + \xi S)^{1/2}/\sqrt{N}$ for the squeezed vacuum, where N is the total number of available photons, S is the squeeze factor ($-1 < S \leq 0$), and $\xi < 1$ is the efficiency of the squeezing. Show, in a phasor diagram, the relative phase relationships for the fields that emerge from the outputs, and from your diagrams infer that to achieve the above optimal sensitivities with readout at output #1, the unperturbed position of mirror A should be adjusted so that the phase difference between the two paths along the two arms is $\phi_0 = \pi/2$. More specifically:

   a. Show the orientation of the squeezing ellipses relative to the coherent amplitudes for each of the two fields E_a, E_b that contribute to the total field E_1 at the output #1.

   b. Show how these two fields with their fluctuations sum to give a resultant E_1 that (for $\phi_0 = \pi/2$) produces noise in the photodetector below the standard shot-noise level $1/\sqrt{N}$ and a signal proportional to the phase deviation δ_0.



c. Note that for an efficiency $\xi \to 1$ and for perfect squeezing $S \to -1$, the above analysis and diagrams predict that the minimum detectable phase deviation becomes arbitrarily small, $\delta_0 \to 0$. Show that, in fact, if the interferometer system is perfectly lossless, and δ_0 is modulated so $\delta_0 = \Delta_0 \cos(\Omega_0 t)$, the minimum detectable modulation amplitude Δ_0 is actually $\Delta_0 \sim 1/N$. Calculate the corresponding length sensitivity Δx for the displacement of mirror A. Estimate the laser power required to achieve the sensitivity of the advanced LIGO, *if* this limit could be achieved.

d. In the above discussion it was tacitly assumed that the interferometer mirrors are so massive that light pressure fluctuations do not disturb them significantly. Suppose now that mirror A has a finite, small mass and is free to move in response to light pressure, and that we apply a feedback force to the back of the mirror, to counteract the time-averaged light-pressure force on its front. Show, using the phasor diagrams of parts a. and b., that when we improve our measurement of δ_0 (and hence of the mirror position x) by increasing the amount of squeezing, we increase the random light-pressure perturbations of the mirror, thereby enforcing the uncertainty principle. Relate this result to the standard quantum limit for sensing the position of the small mass, and thence to the curve labeled "Quantum Limit" in the plots of LIGO noise sources that were shown in earlier lectures. [For a quantitative analysis, in the context of a Michelson interferometer, see C. M. Caves, *Phys. Rev. D*, **23**, 1693 (1981). In this problem you are supposed to be ignoring the possibility of going beyond the standard quantum limit as discussed by Jackel and Reynaud, *Europhys. Lett.*, **13**, 301 (1990).]

LECTURE 16

"Squeezed Light and its Potential Use in LIGO"

- PART I

by

JEFF KIMBLE

5-18, 5-20-94

Quantum $\left\{\begin{matrix}\text{Measurement}\\ \text{Fluctuations}\\ \text{Noise}\end{matrix}\right\}$ in Quantum Optics

I. A vocabulary for quantum noise

- Wigner distributions

II. Generation, propagation, & detection of squeezed light

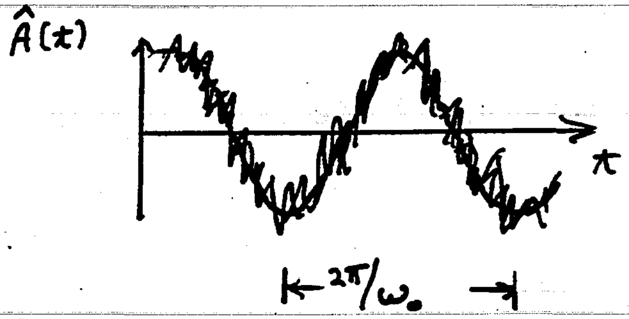
III. Quantum Measurement with Squeezed Light

- Interferometry

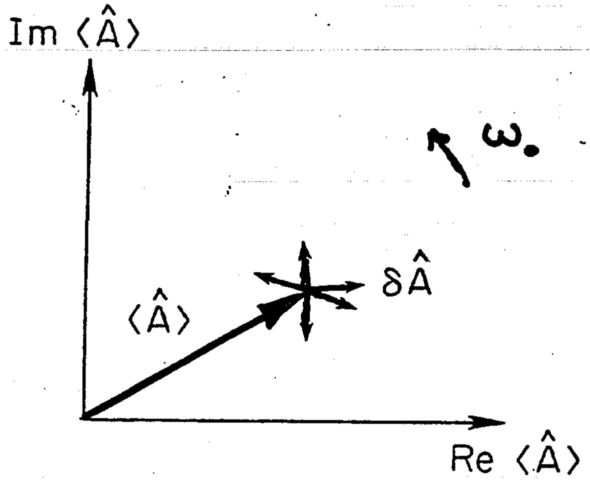
IV. Beyond the SQL for a free mass

H.J. Kimble
Caltech

I. Field Fluctuations



Phasor diagram-



$\langle\hat{A}\rangle$ - mean amplitude of field

$\delta\hat{A}$ - fluctuations of field

"Distribution" of fluctuations?

- Begin with Wigner phase space function

$$W(x_+, X_-)$$

where

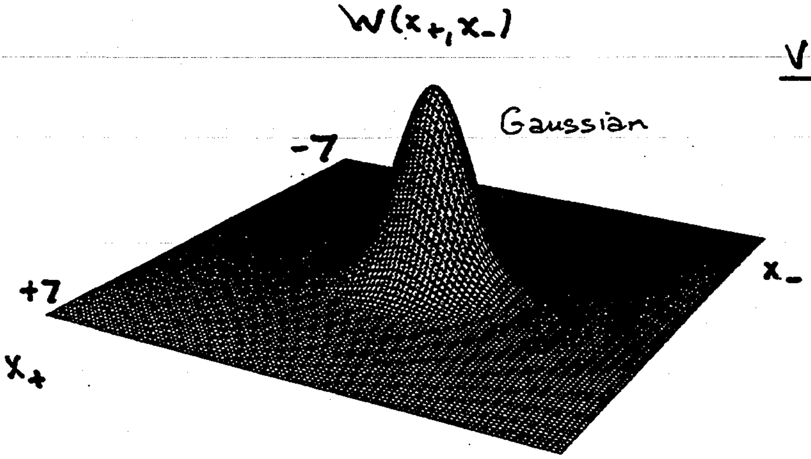
$$\left.\begin{aligned}\hat{x}_+ &\equiv \delta\hat{A} + \delta\hat{A}^+ \\ \hat{X}_- &\equiv \frac{1}{i}(\delta\hat{A} - \delta\hat{A}^+)\end{aligned}\right\}$$

Note - $\hat{x}_+, \hat{x}_-$ canonical variables with $[\hat{x}_+, \hat{x}_-] = 2i$

Wigner Distributions - $W(x_+, x_-)$

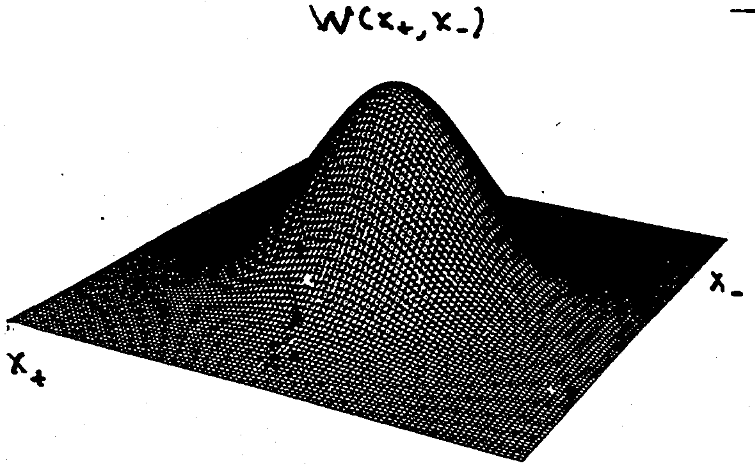
Recall that $[\hat{x}_+, \hat{x}_-] = 2i$

$$\Rightarrow \quad \Delta x_+^2 \, \Delta x_-^2 \geq 1$$



Vacuum State $|0\rangle$

- $\langle \hat{A} \rangle = 0$
- zero-point fluctuations $\sigma_\pm^2 = 1$
- $\langle n \rangle = 0$ photons

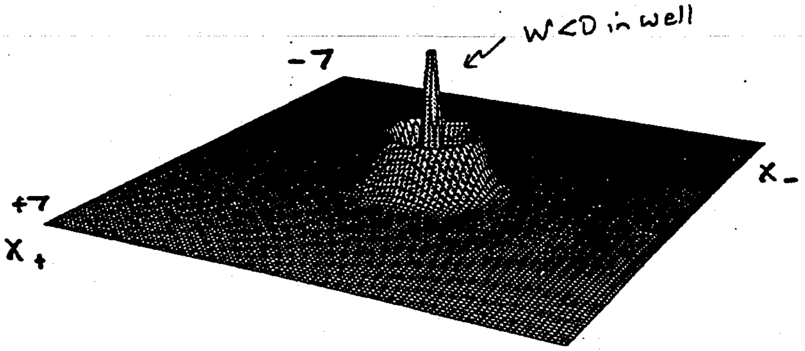


Thermal Field $\hat{\rho}_{th}$

- $\langle \hat{A} \rangle = 0$
- increased, symmetric fluctuations $\sigma_\pm^2 = 5$
- $\langle n \rangle = 2$ photons

$W(x_+, x_-)$, cont.

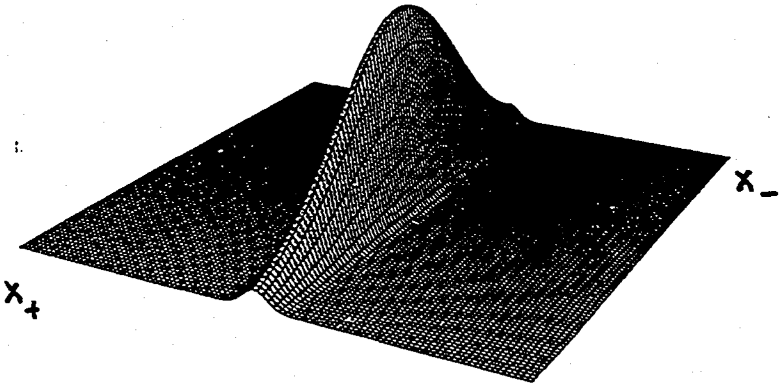
$W(x_+, x_-)$



Fock (number) State

$|n\rangle = |2\rangle$

- $\langle \hat{A} \rangle = 0$
- $n = 2$ exactly



Squeezed State

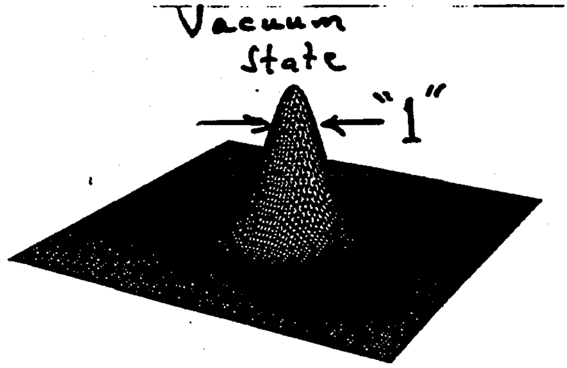
$|r\rangle$

- $\langle \hat{A} \rangle = 0$
  "Squeezed vacuum"
- phase dependent redistribution of quantum fluctuation
  $\sigma_+^2 = 1/10$
  $\sigma_-^2 = 10$
  $\{\sigma_+ \sigma_- = 1\}$
- $\langle n \rangle = 2$ photons

Manifestly Quantum or Nonclassical States of the Electromagnetic Field

OPTICAL EQUIVALENCE THEOREM

Manifestly Quantum	Classical
Fock State	Coherent State
Squeezed State	Thermal State
⋮	⋮
Los Angeles	Boston



Scale for variation << 1 ←

→ Scale for variation >> 1

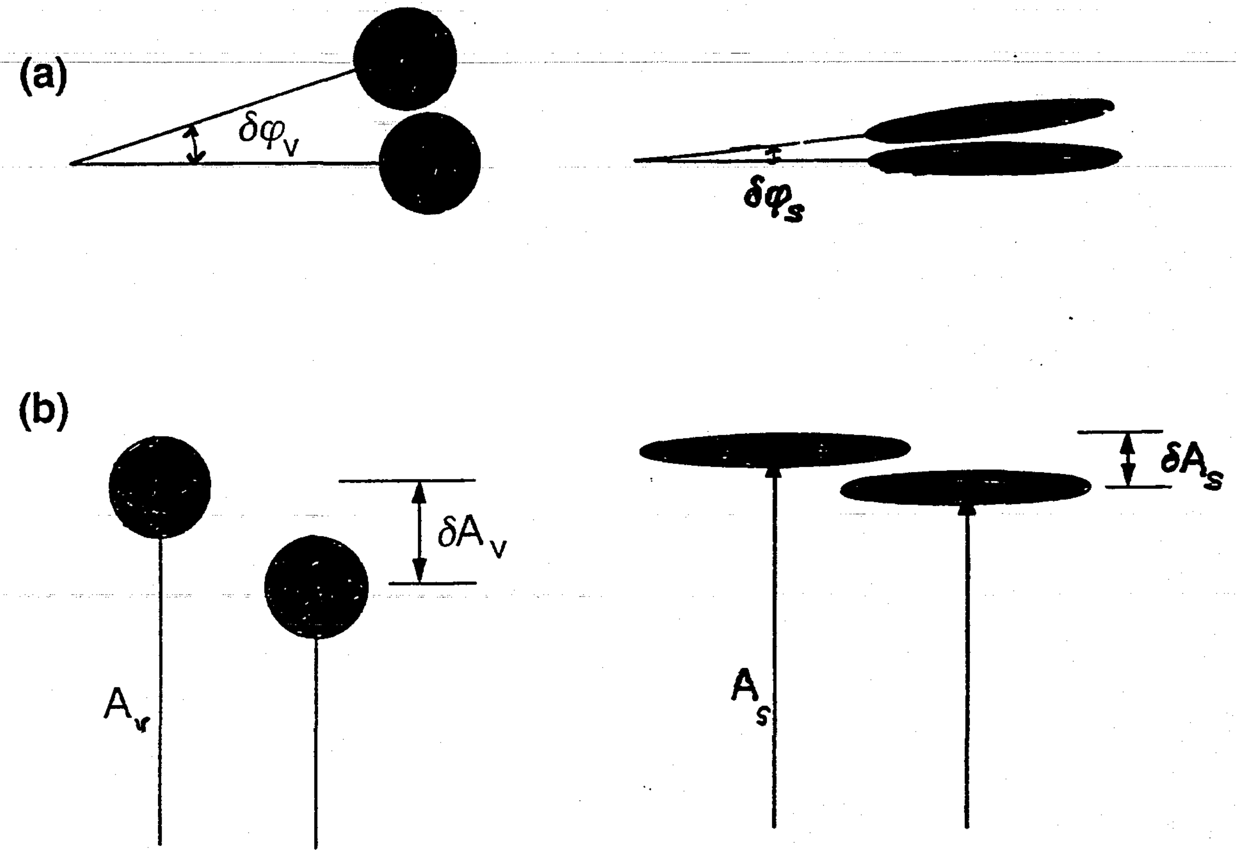
→ Fields from classical (stochastic) current sources

Fields which require quantum probability amplitudes ←

Squeezed (Nonclassical) Light for Sensitivity Beyond Standard Quantum Limit

Phase Measurements

$$\delta \varphi_v \simeq 1/\sqrt{N} \qquad\qquad \delta \varphi_s \simeq \Delta X_- / \sqrt{N}$$

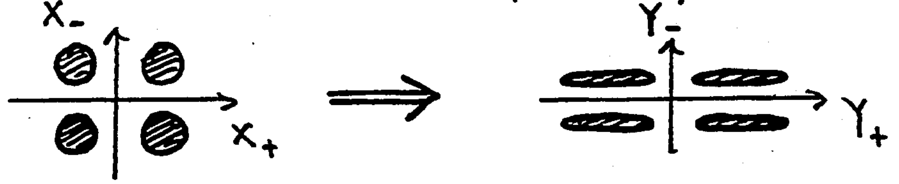


Amplitude Measurements

$$\delta A_v / A_v \simeq 1/\sqrt{N} \qquad\qquad \delta A_s / A_s \simeq \Delta X_+ / \sqrt{N}$$

Generation of Squeezed States

→ "Elastic" deformation of phase space



with
$$Y_- = e^{-r}\, X_- = S^{\dagger}(r)\, X_-\, S(r)$$
$$Y_+ = e^{+r}\, X_+ = S^{\dagger}(r)\, X_+\, S(r)$$

Squeezing generated by transformation

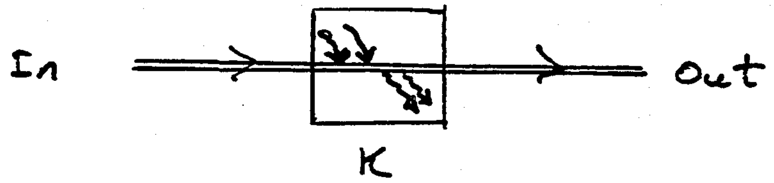
$$S(r) = \exp\left[\tfrac{1}{2}\left(r\,\hat{a}^2 - r\,\hat{a}^{\dagger 2}\right)\right]$$

Annihilate ↗   ↖ create

Correlated pairs of photons

Optical Interactions

$$\hat{H}_I \sim i\hbar K\,\hat{a}^2 - i\hbar K\,\hat{a}^{\dagger 2}$$

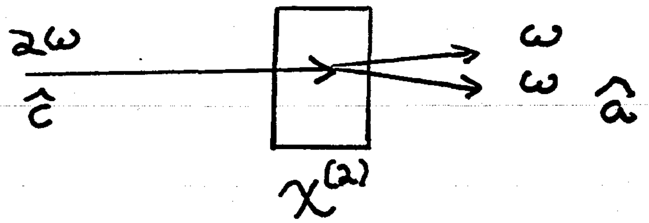


Time evolution operator

$$U(t) \sim \exp\left\{ (Kt)\,\hat{a}^2 - (Kt)\,\hat{a}^{\dagger 2} \right\} \longleftrightarrow S(Kt)$$

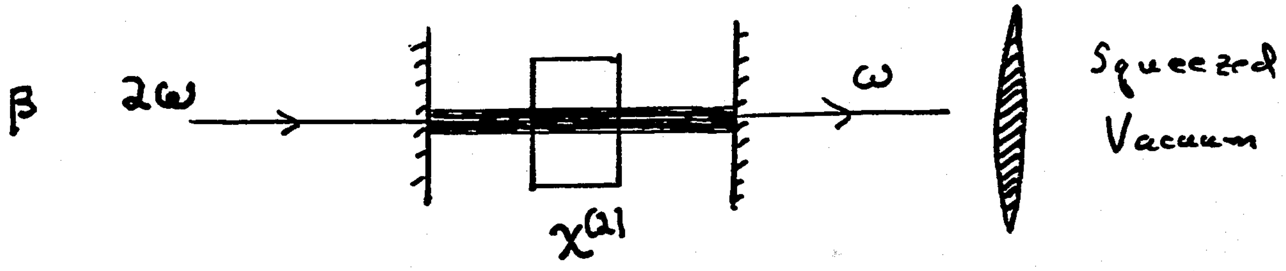
$U(t)$ generates squeezed state !

Squeezed State Generation by Parametric Down Conversion

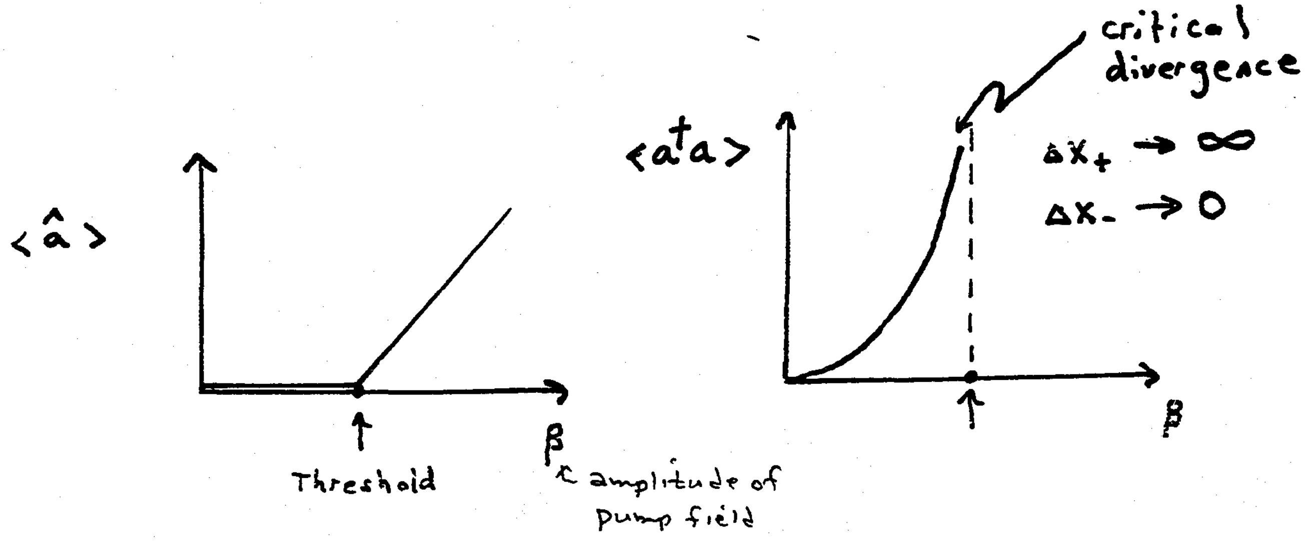


$$\hat{H}_0 \sim \chi^{(2)} \; \hat{c} \; \hat{a}^{\dagger 2} + \text{H.c.}$$

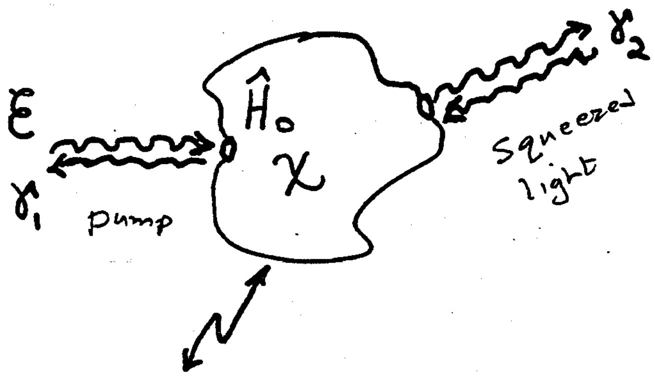
Experiment



(Subthreshold) Optical Parametric Oscillator - OPO



Dynamics of Open Quantum Systems



$\hat{\rho}$ - Reduced density operator

A well-worn path -

- Schroedinger Eqn.

$$\Big\downarrow \ \omega_0 >> (\chi, \gamma)$$

- Master Eqn.

$$\dot{\hat{\rho}} = \frac{1}{i\hbar}[\underset{\substack{\uparrow \\ \text{reversible evolution}}}{\hat{H}_0}, \hat{\rho}] + \underset{\substack{\uparrow \\ \text{irreversible decay}}}{\hat{\mathcal{L}}(\gamma_1, \gamma_2)}\hat{\rho} + \hat{\mathcal{G}}(\underset{\substack{\uparrow \\ \text{excitation}}}{\varepsilon})\hat{\rho}$$

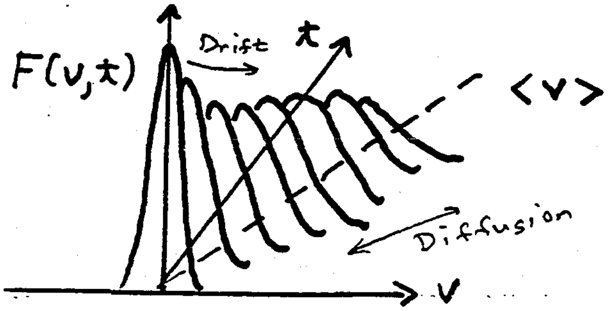
$$\Big\downarrow \ \chi << \gamma$$

- Dynamics for distribution $F(v, v^*)$ ← Wigner Distribution

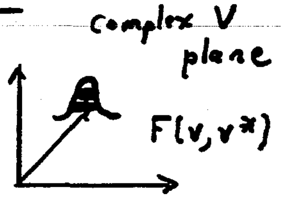
$$\frac{\partial F(v,v^*,t)}{\partial t} = D(\chi, \gamma, \varepsilon)\, F(v, v^*, t)$$

← Quantum fluctuations are here

Fokker-Planck type of equation



Quasiprobability Distributions



- Consider single mode of EM field

  $\hat{a}, \hat{a}^{+}$ ↔ creation, annihilation operators

  $\hat{\rho}$ ↔ density operator

  $= |\psi\rangle\langle\psi|$ for pure state

- "Distribution" $F(v,v^*)$ from $\hat{\rho}$?

$$F(v,v^*) = \frac{1}{\pi^2}\int d^2z\, e^{-i(z^*v^*+zv)}\,\mathrm{Tr}\left[\hat{\rho}\, e^{i(z^*\hat{a}^{+}+z\hat{a})}\right]$$

C-#'s ↑ ↑ operators

- **Note**

  1. Association of c-numbers with operators

     e.g. $\left\{\begin{matrix} v \leftrightarrow \hat{a} \\ v^* \leftrightarrow \hat{a}^{+}\end{matrix}\right\}$, $\left\{\begin{matrix} x_+ \leftrightarrow \hat{X}_+ \\ x_- \leftrightarrow \hat{X}_-\end{matrix}\right\}$

  2. Wavefunction $|\psi\rangle$?

     e.g. $|\psi(x_{\pm})|^2 = \int W(x_+, x_-)\, dx_{\mp}$

  3. Field commutation relations for "distributions"

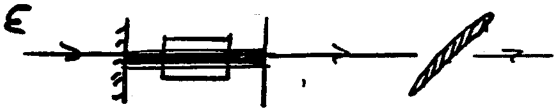
     $$e^{i(z^*\hat{a}^{+}+z\hat{a})} \neq e^{iz^*\hat{a}^{+}}e^{iz\hat{a}} \neq e^{iz\hat{a}}\,e^{iz^*\hat{a}^{+}}$$

     ∴ Can build many distributions from $\hat{\rho}$

  * Nonuniqueness in discussion of physical processes

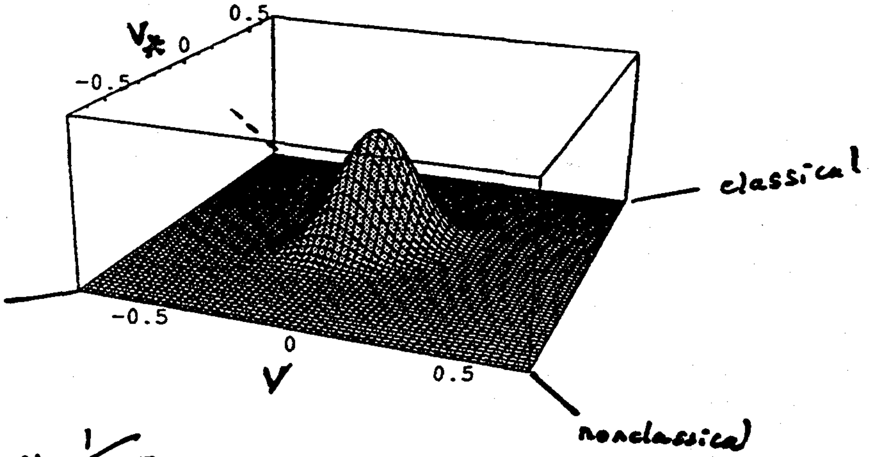
Positive P Representation for Optical Parametric Oscillator (OPO)*

$\left. \begin{array}{l} V \\ V_* \end{array} \right\} \rightarrow P(V, V_*)$, with line $V_* = V$ as "classical" dimension and $V_* = -V$ as "nonclassical" dimension



Below threshold –

$\mathcal{E}/\mathcal{E}_{th} = 0.5$

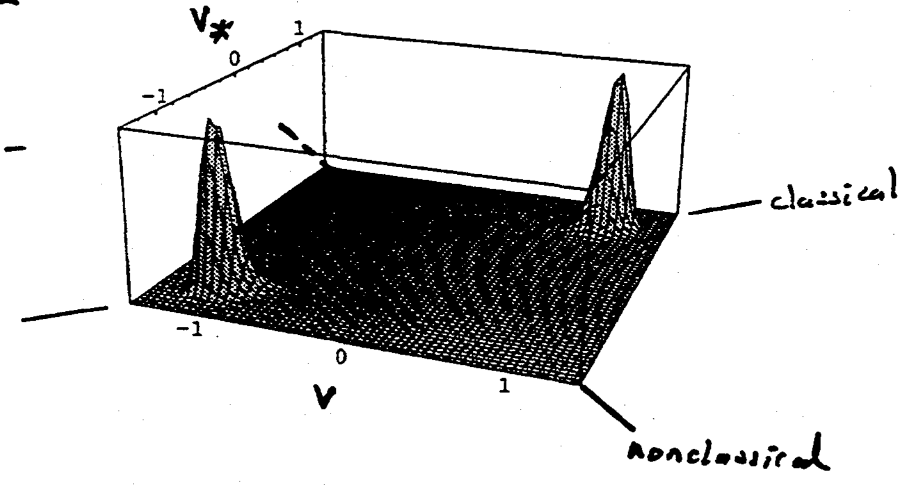


$$g \equiv \frac{\chi}{\sqrt{\gamma_2 \gamma_1}} \sim \frac{1}{\sqrt{n_0}}$$

$g = 0.2$

Above threshold –

$\mathcal{E}/\mathcal{E}_{th} = 2$



[*] Wolinsky and Carmichael, PRL 60, 1836 (1988).

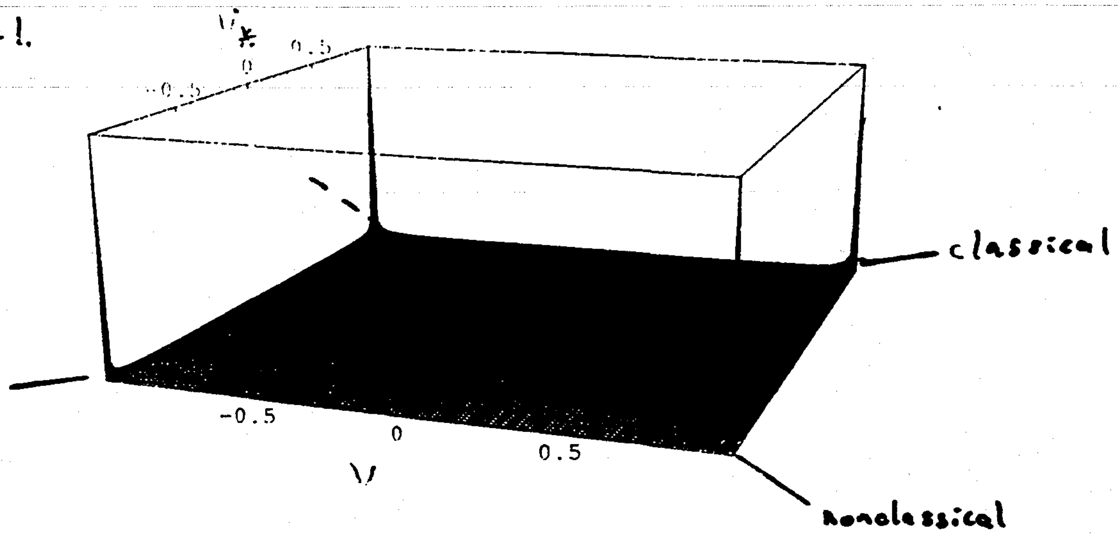
From Squeezed States to Coherent Superpositions

Strong Coupling $g \sim \frac{\chi}{\gamma} \sim \frac{1}{\sqrt{n_0}}$

$g = 5$

Pathological Distribution!
Positive $P(v, v_*)$

$E/E_{th} = 1.$



$P(v, v_*) \rightarrow$ delta functions at 4 corners

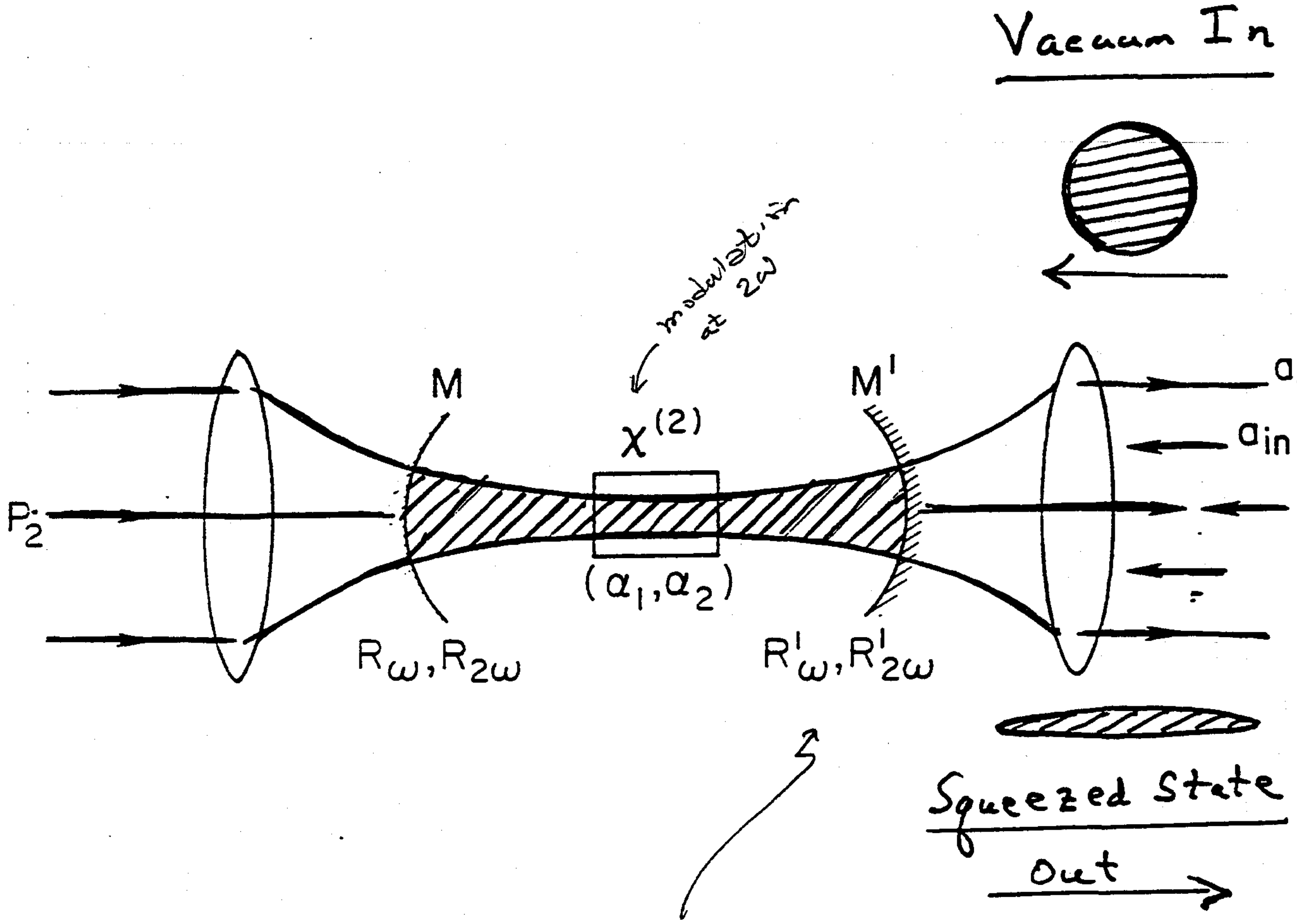
- Significance?
  Coherent superposition $|\psi\rangle = \frac{1}{\sqrt{2}}\left[|\alpha_0\rangle + |-\alpha_0\rangle\right]$

$\alpha_0 = 1/g$

From a dissipative dynamical system!
{See also Carmichael, Haroche, Meystre, ...
within context of cavity Q.E.D. for large g}

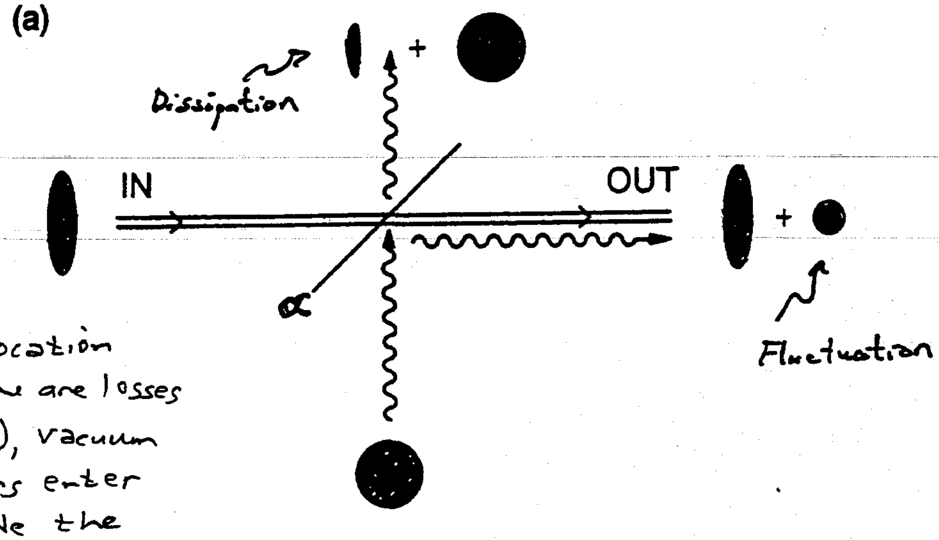
Wolinsky ; Carmichael

Generation of Squeezed Light

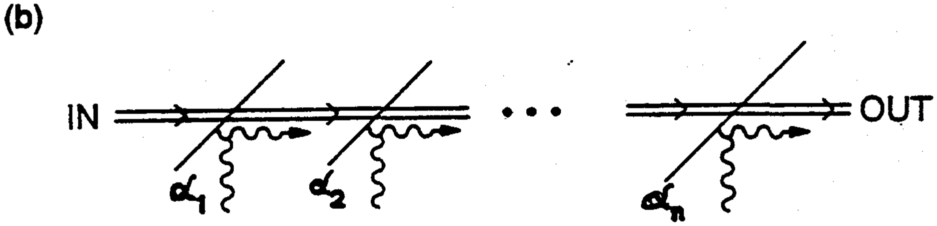


Would like to just wiggle this mirror, but cannot do so fast enough

Fluctuation - Dissipation Theorem - The Grim Reaper for Nonclassical Fields



At any location where there are losses (dissipation), vacuum fluctuations enter and degrade the squeezing



$$\Delta X^2_{out} = (1-\bar{\alpha}) + \bar{\alpha}\,\Delta X^2_{in}$$

$\bar{\alpha}$ - overall efficiency $\quad 0 \leq \bar{\alpha} \leq 1$

$$\bar{\alpha} = \prod_{i=1}^{n} \alpha_i$$

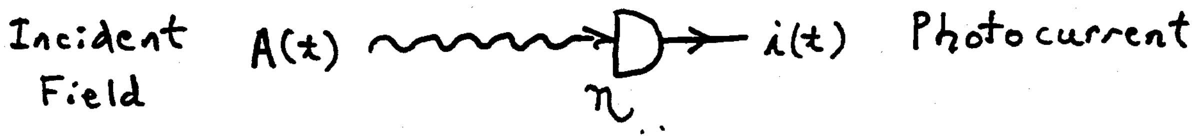
Many tiny losses kill you!

Every kind of loss is bad news!

$\underline{\underline{\bar{\alpha} \to 0}}$

$$\Rightarrow \quad \Delta X^2_{out} \longrightarrow \underset{\uparrow \text{ vacuum-state limit}}{1}$$

Photoelectric Detection of Squeezed Light

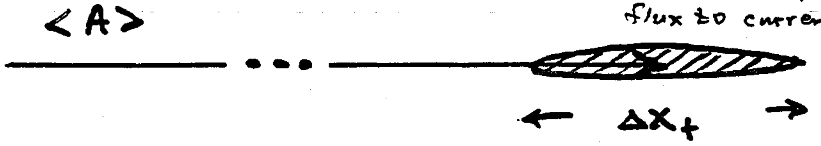


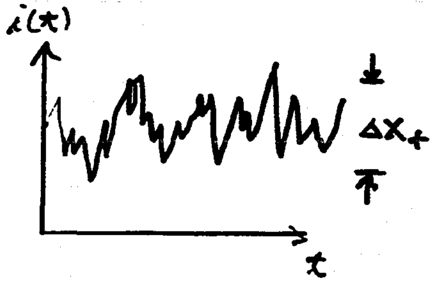
$$i(t) \sim e\eta\,|A(t)|^2$$

efficiency

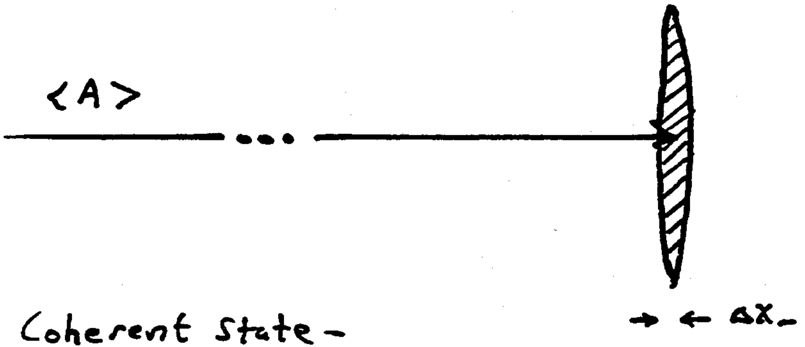
electron charge to convert photon flux to current

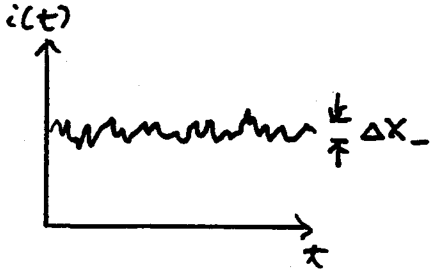
Squeezed State -



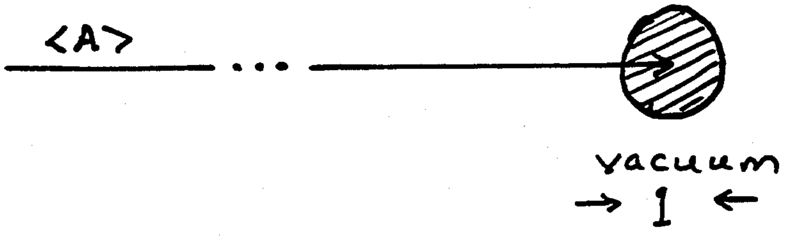


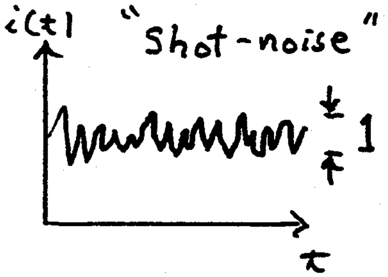
$\{\langle A\rangle >> \Delta A\}$



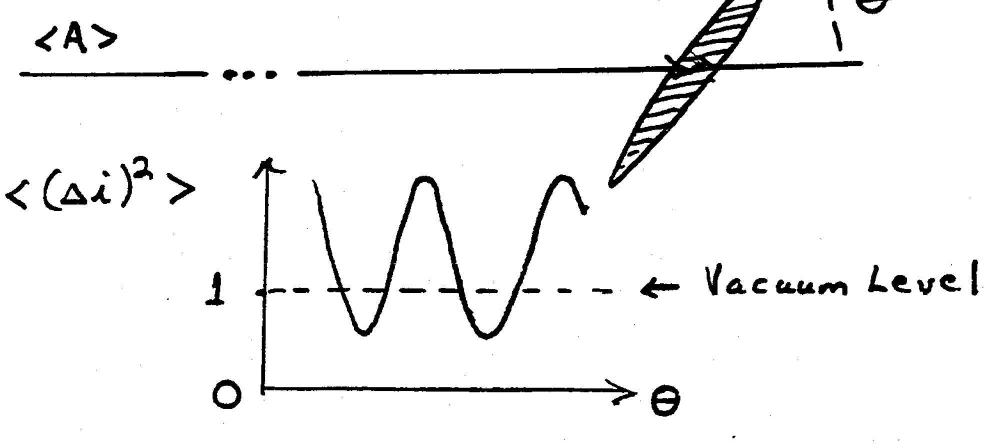


Coherent State -





More generally,



"Quantum Shish Kebab"

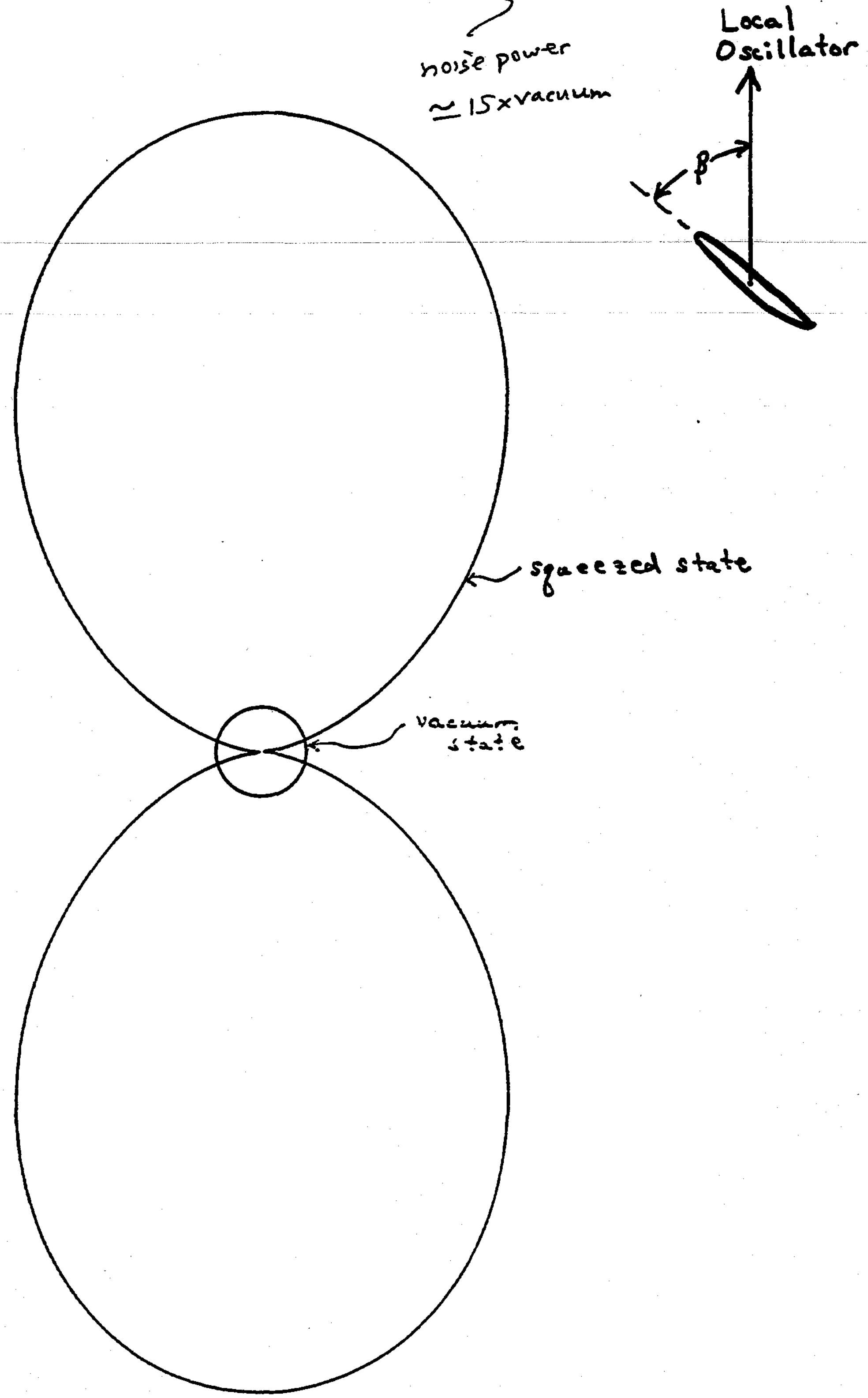
gradually change θ to see the squeezing

Variance vs. Phase β for Squeezed State

$\sigma_+^2 = 0.07$

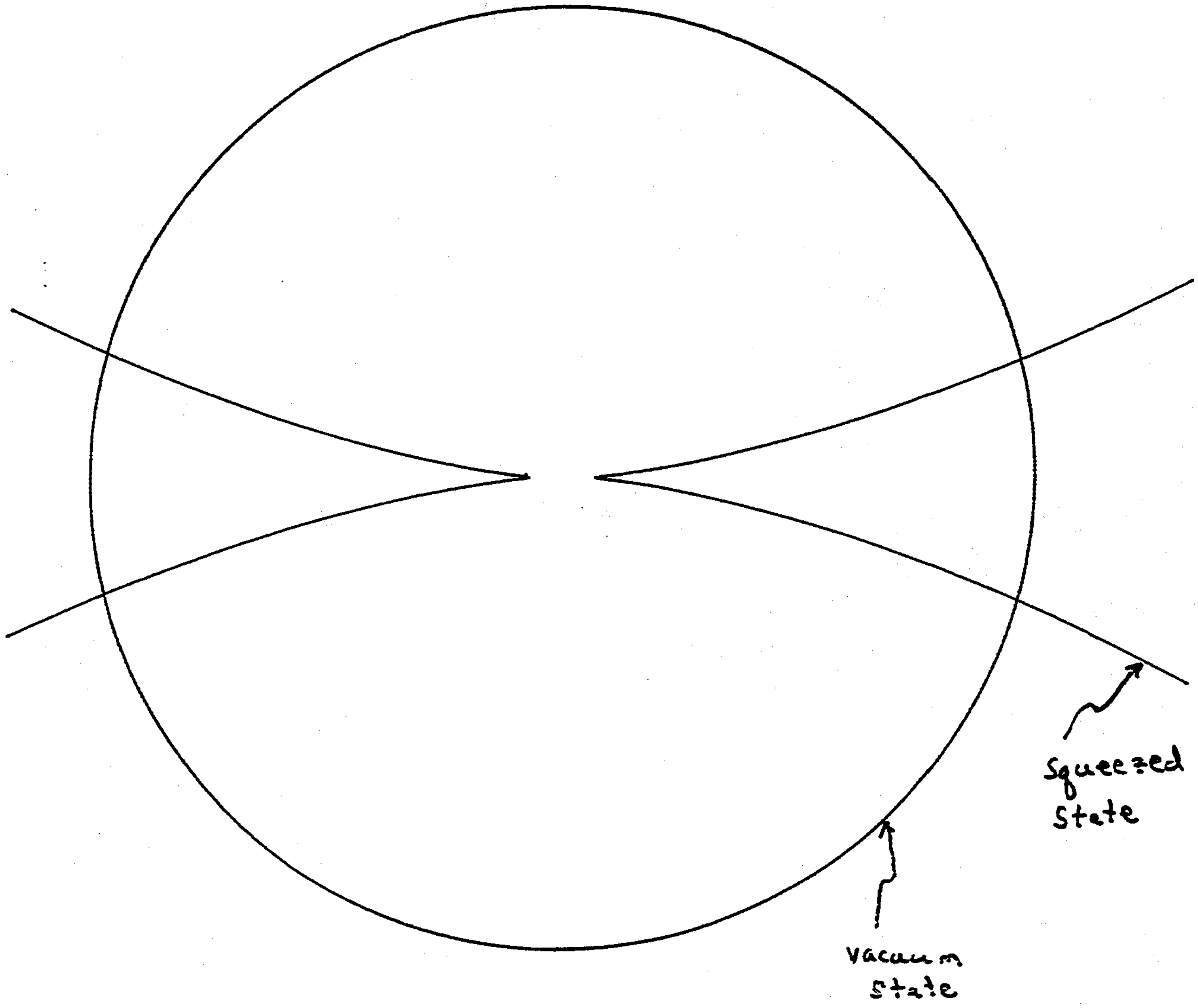
$\sigma_-^2 = 14.3$

noise power $\simeq 15\times$vacuum

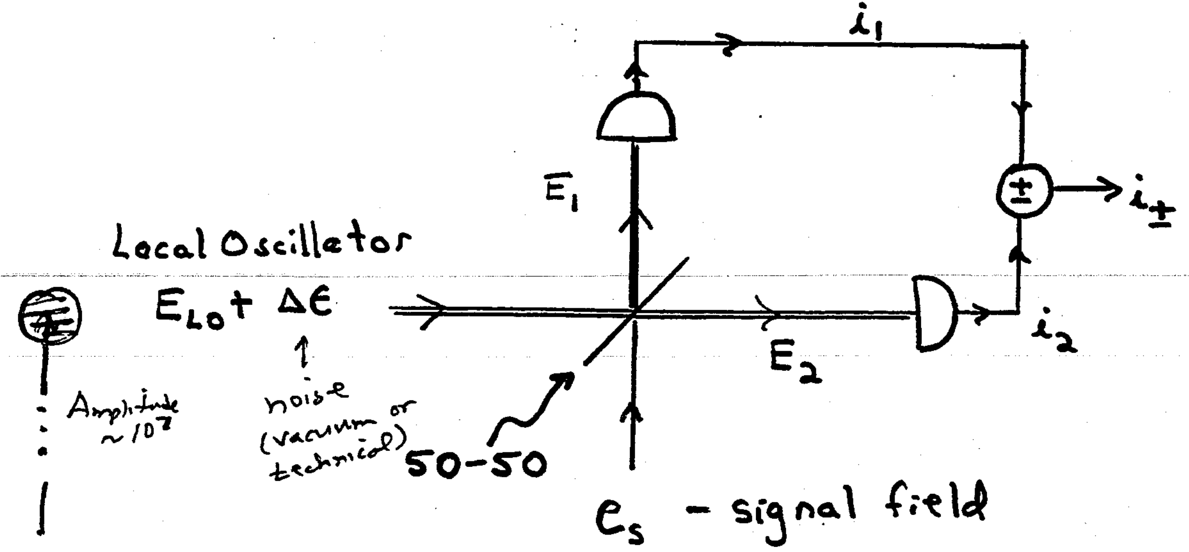


Variance vs. Phase β – Enlarged View

$\sigma_+^2 = 0.07$
$\sigma_-^2 = 14.3$



Balanced Homodyne Detector - The Basic Idea



Assume $E_{LO} \gg (\Delta\epsilon, e_s)$

Photocurrents i_1, i_2

$$i_1 \propto |E_1|^2 \quad , \quad E_1 = \frac{1}{\sqrt{2}}\left(E_{LO} + \Delta\epsilon + e_s\right)$$

$$i_2 \propto |E_2|^2 \quad , \quad E_2 = \frac{1}{\sqrt{2}}\left(E_{LO} + \Delta\epsilon - e_s\right)$$

Hence $\Delta i_+ \sim \mathrm{Re}[E_{LO}\Delta\epsilon^*]$ Homodyne of LO with its fluctuations

$\Delta i_- \sim \mathrm{Re}[E_{LO} e_s^*]$ Homodyne of LO with signal field

Noise power in photocurrent $\sim i^2$

Hence $\Delta i_-^2 \sim \mathcal{E}_0^2 \left[X_+ \cos\theta + X_- \sin\theta \right]^2$

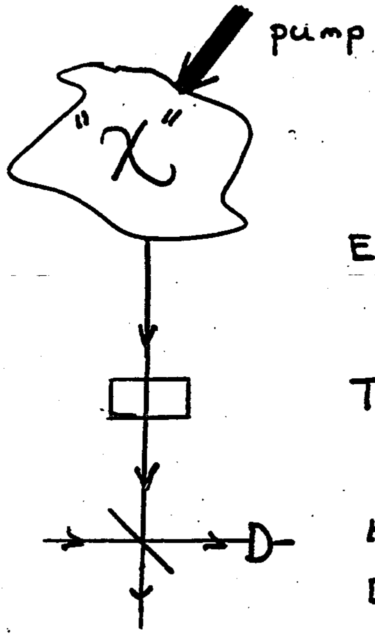
where $E_{LO} = \mathcal{E}_0 e^{i\theta}$

$X_+ = e_s + e_s^*$, $X_- = \frac{1}{i}(e_s - e_s^*)$ ⟷ quadrature phase amplitudes

- Interrogate signal quadratures by varying LO phase β

- Quality of balancing

Generic Experiment



Generation — Escape Efficiency ρ

Propagation — Transmission T_0

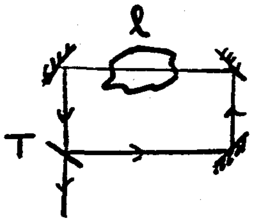
Detection — Heterodyne Efficiency η
Detector q.e. α

Recall $$R = \xi \Delta X^2 + (1-\overset{\text{efficiency}}{\xi})$$

with $$\xi = \rho T_0 \eta^2 \alpha$$

$\rho \leftrightarrow$ losses in medium

* $\rho \simeq 0.98$ [-17 dB]



$l \sim 2\times10^{-3}$
$T \sim 0.1$

$\rho \sim \frac{T}{T+l}$

$T_0 \lesssim 0.99$ Transmission

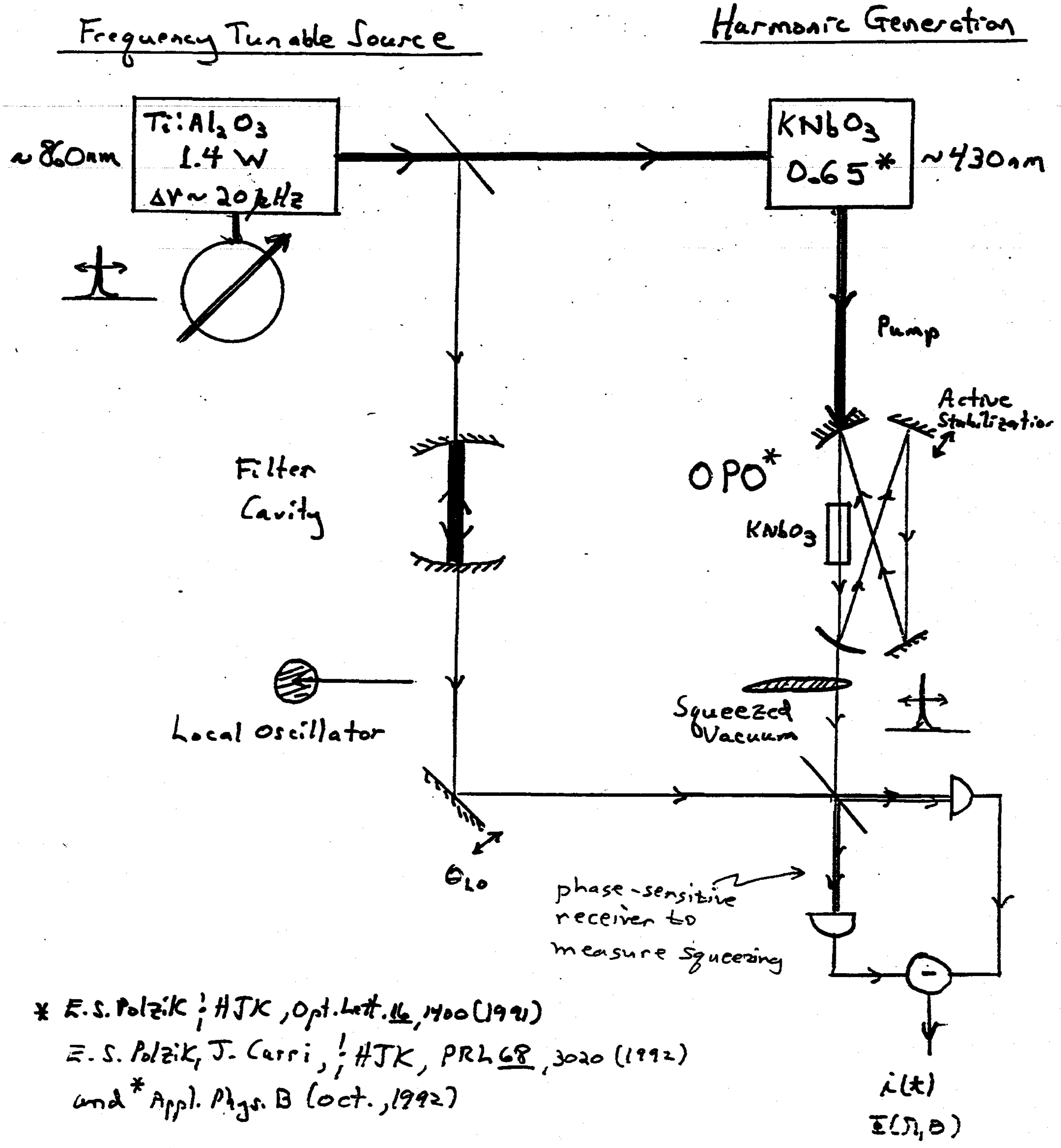
$\eta \gtrsim 0.99$ Heterodyne efficiency

$\alpha \gtrsim 0.99$ Photodetector efficiency

Total $\xi \lesssim 0.94$ {17×, -12 dB}
observed

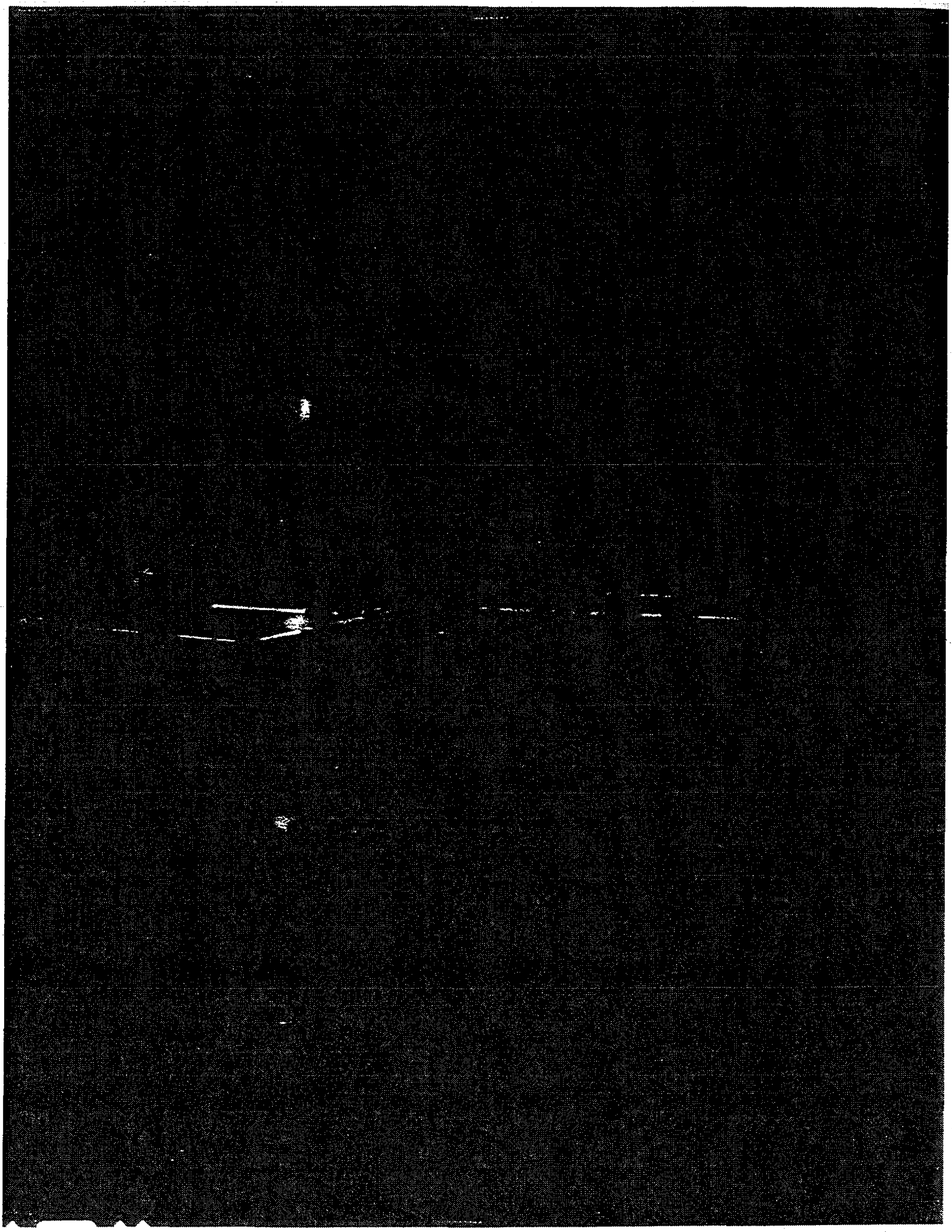
Experimental Arrangement -
Atomic Spectroscopy with Squeezed Light

[in Kimble's lab]

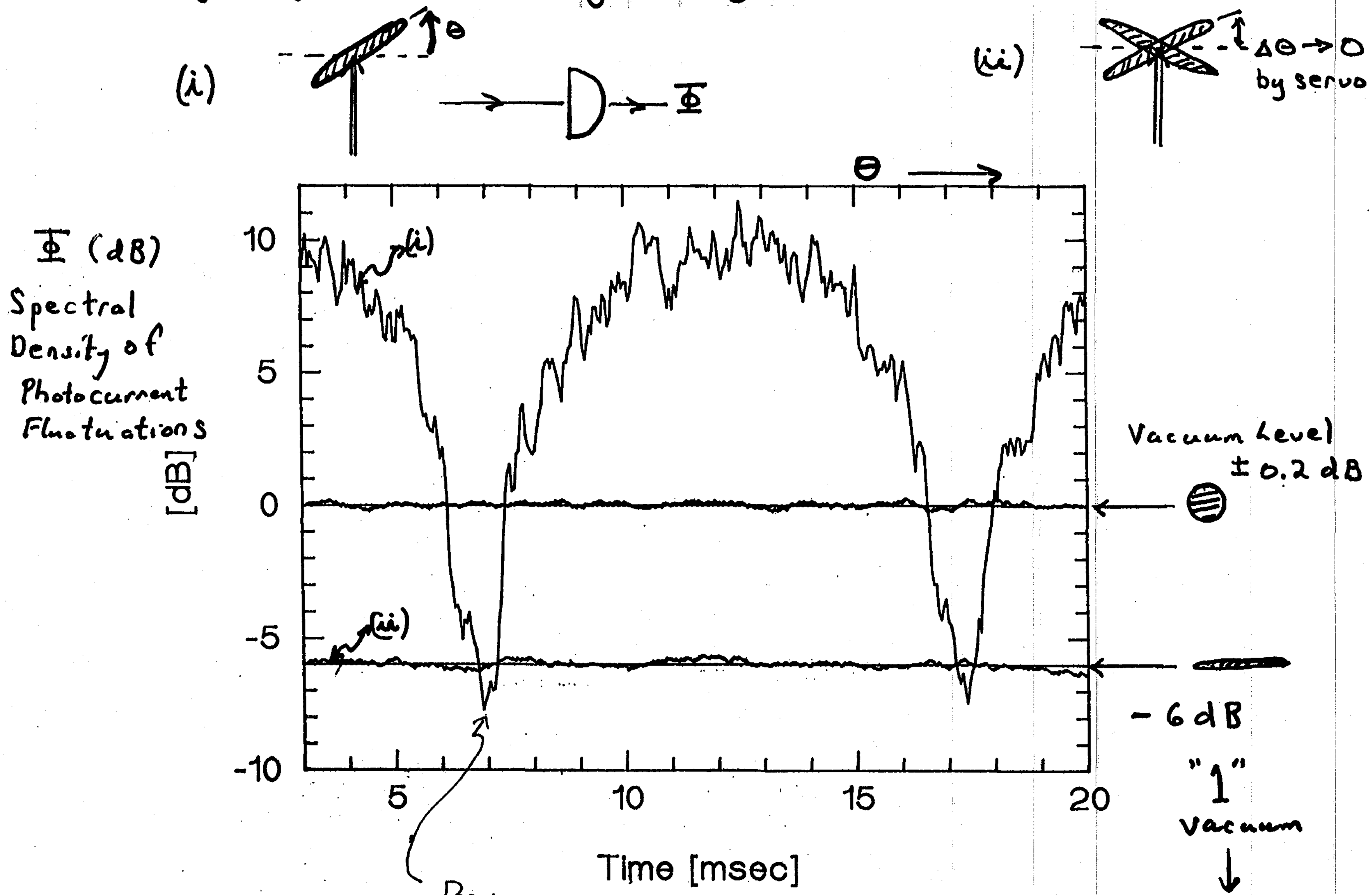


* E.S. Polzik & HJK, Opt. Lett. 16, 1400 (1991)
E. S. Polzik, J. Carri, & HJK, PRL 68, 3020 (1992)
and * Appl. Phys. B (Oct., 1992)

°1=17



Frequency Tunable Squeezing



"1"
vacuum
↓
0.25
"quantum quietness"

Best squeezing that has ever been achieved: Factor 4 in noise power

λ = 856nm

Polzik, Carri, AJK

Appl. Phys. B 55, 279 (1992)

Status of Observed Noise Levels R

[circa 1992

			Vacuum-State Limit 1.0
85-86			
	Slusher et al.	$\chi^{(3)}$, Na beam	0.75
	Shelby et al.	$\chi^{(3)}$, fiber	0.87
•	Wu et al.	$\chi^{(2)}$, subthreshold OPO	0.37
	Maeda et al.	$\chi^{(3)}$, Na Vapor	0.96
•	Raizen et al.	$\chi^{(3)}$, Na beam	0.70
87			
	Schumaker et al.	$\chi^{(3)}$, 4-mode	0.80
	Grangier et al.	$\chi^{(2)}$, subthreshold OPO	0.63
	Slusher et al.	$\chi^{(2)}$, pulsed squeezing	0.87
88→			
•	Pereira et al.	$\chi^{(2)}$, frequency doubling	0.87
•	Pereira et al.	$\chi^{(2)}$, subthreshold OPO	0.45
	Ho et al.	$\chi^{(3)}$, Na vapor	0.75
	Kumar et al.	$\chi^{(2)}$, incoherent pulse	0.83
	Sizmann et al.	$\chi^{(2)}$, frequency doubling	0.60
	Movshovich et al.	$\chi^{(2)}$, Josephson paramp	0.99896
	Hirano, Matsuoka	$\chi^{(2)}$, pulsed	0.78
	Rosenbluh, Shelby	$\chi^{(3)}$, solitons	0.68
	Bergman, Haus	$\chi^{(3)}$, pulsed	0.32
	Hope et al.	$\chi^{(3)}$, Ba beam	0.87
•	Polzik et al.	$\chi^{(2)}$, subthreshold OPO	0.25

⋮

↑ power below vacuum

• { UT Austin
Caltech

Omnivorous Creatures

Not Included in Theoretical Model Community

[Practical factors that impede better squeezing]

- Thermally excited noise {TEXAS, GAWBS, ...}

- Transverse effects in propagation

  $\chi^{(2)}$ - Slusher et al

  $\chi^{(3)}$ - Shapiro et al.

- Raman Scattering

  $\chi^{(3)}$ - Solitons (Shelby, Drummond, ...)

- Light induced absorption

  $\chi^{(2)}$ - Polzik

  ⋮

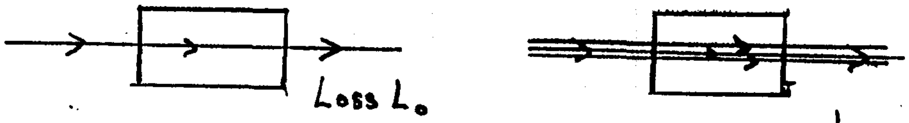
→ Nature of microscopic processes underlying χ

Thus far only for $\chi^{(3)}$ in atomic vapors

→ Self consistent treatment of propagation and quantization

Solitons; quantization in dielectric

Trouble in River City!



Loss

$L_0 + L_2(P_2)$

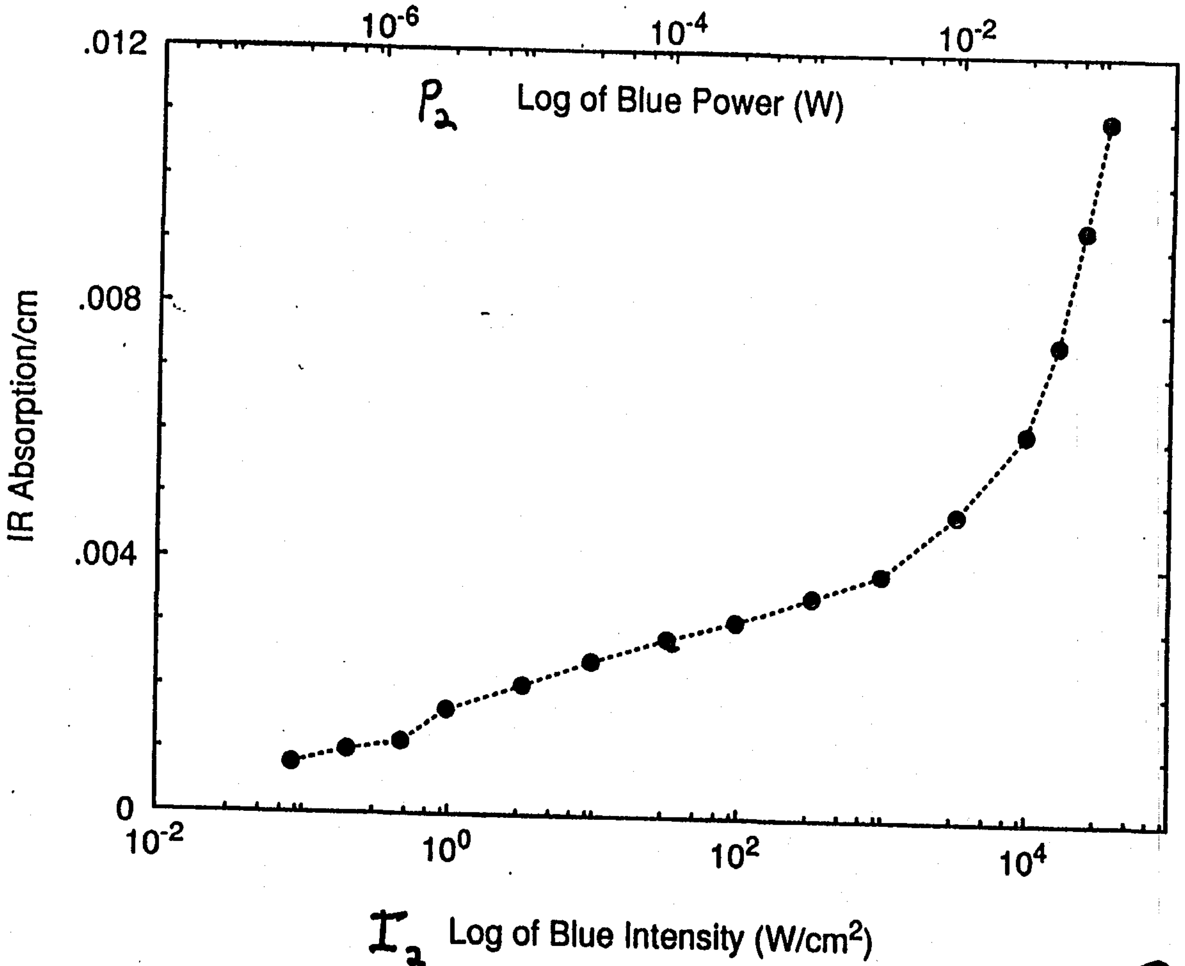
↑

Light-induced absorption

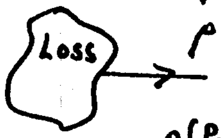
Blue Light Induced Infrared Absorption in $KNbO_3$

$L_2(P_2)$

(1/cm)



30x squeezing until turn up the power



$\rho = 0.972$

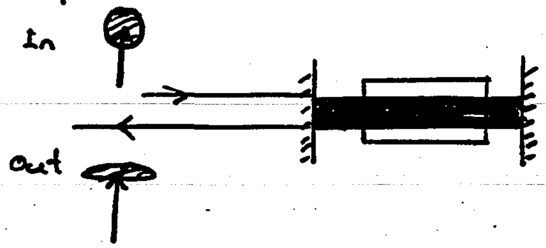
↓

$\rho(P_2) \simeq 0.875$

result of turning up power

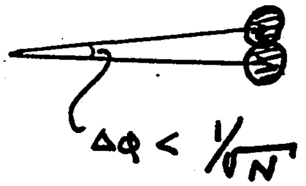
Moral ?

→ Squeezing with "large" coherent amplitude invites other problems

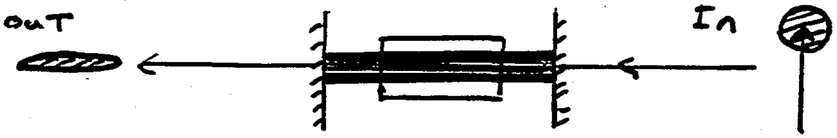


Technical noise makes squeezing much harder when the amplitude is large than when squeezing the vacuum.

Allowed $\Delta\phi$?



→ By contrast, squeezed vacuum by way of parametric down conversion

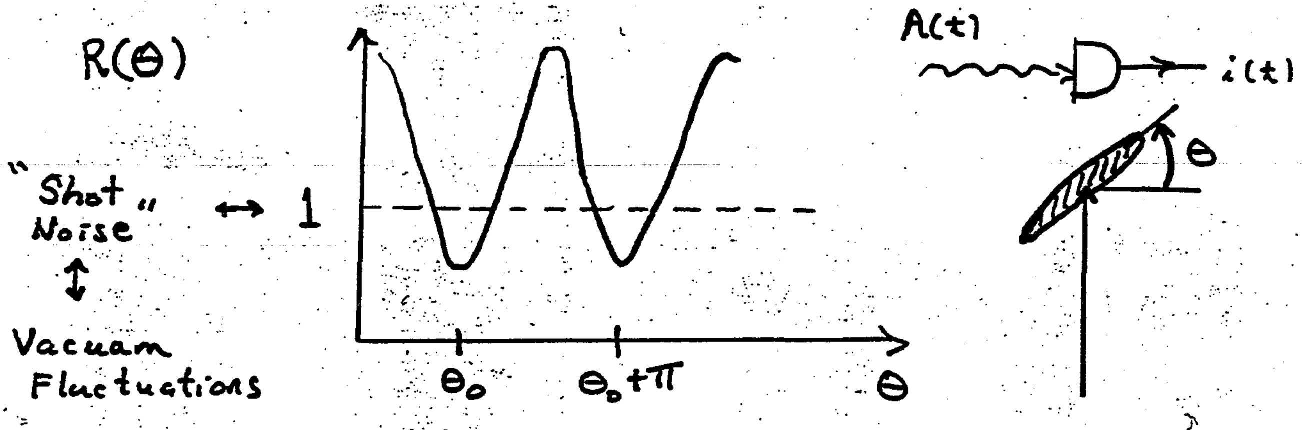


Allowed $\Delta\phi$?

$$\Delta\phi \sim \left[R_- / R_+ \right]^{1/2}$$

{Recall pump quantization, Caves et al.}
laser linewidth effects, Walls et al.

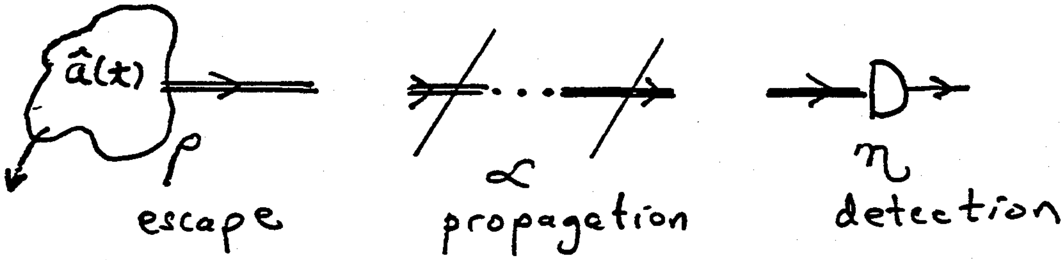
Inference from Photocurrent Statistics to Field Statistics



$$R(\Theta) = \left[(1-\xi) + \xi \langle (\Delta X(\Theta))^2 \rangle\right]$$

$\uparrow$

ξ - overall system efficiency



$$\xi = \rho \alpha \eta$$

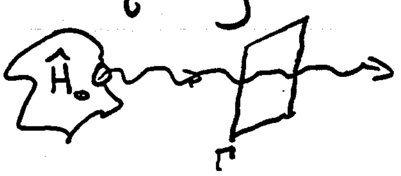
- From $R(\Theta)$ and ξ extract $\langle(\Delta X(\Theta))^2\rangle$

  $\Theta = 0$, ΔX_+

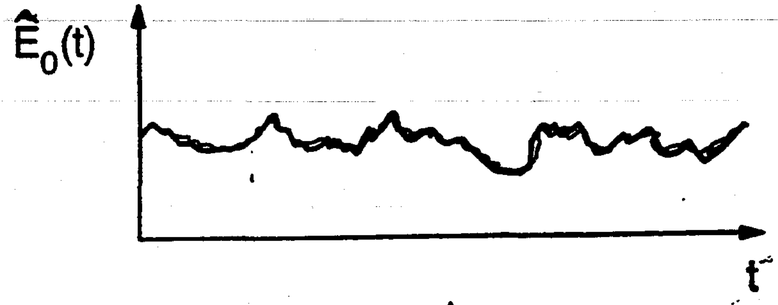
  $\Theta = \pi/2$, ΔX_-

Minor Complication -

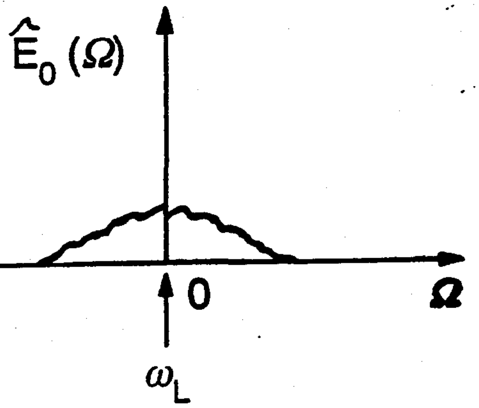
Distribution in frequency Ω of field fluctuations



At plane Γ, take $\hat{E}_0 e^{-i\omega_L t}$



Fourier transform



In practice, one does not use a degenerate pump; instead $\omega \pm \Omega$. Squeezing correlates the fields at $\omega+\Omega$ and $\omega-\Omega$

- **Spectrum of Squeezing $S(\Omega)$***

Quadrature amplitudes $\hat{X}_\theta(\Omega) = \overset{\text{annihilation at } \Omega}{\overset{\downarrow}{\hat{E}_0(\Omega)}} e^{-i\theta} + \overset{\text{creation at } -\Omega}{\overset{\downarrow}{\hat{E}_0^+(-\Omega)}} e^{i\theta}$

$$\Rightarrow \quad \langle \hat{X}_\theta(\Omega), \hat{X}_\theta(\Omega') \rangle = \big[1 + \underset{\uparrow}{S_\theta(\Omega)} \big] \delta(\Omega+\Omega')$$

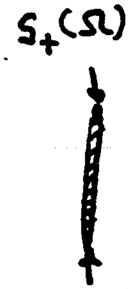
Spectral density for quantum noise at $\pm\Omega$ relative to ω_L

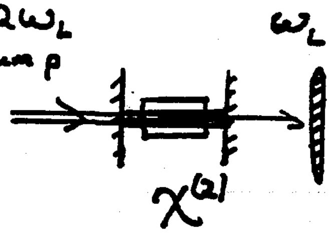
Units ~ (photons/sec)/bandwidth

– dimensionless

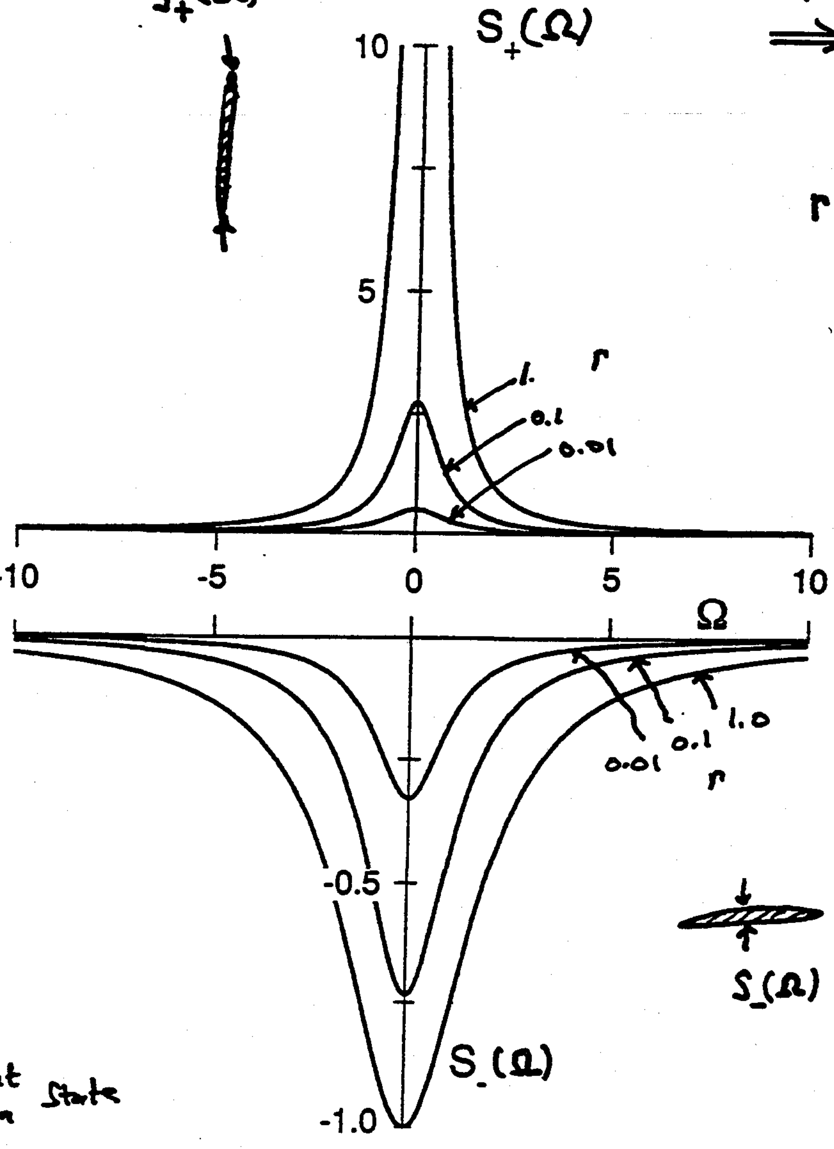
* Gardiner et al.

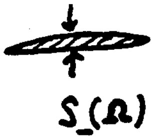
Spectrum of Squeezing $S_\pm(\Omega)$ for Subthreshold OPO





$r = P_2 / P_{2\,\text{threshold}}$





Note -

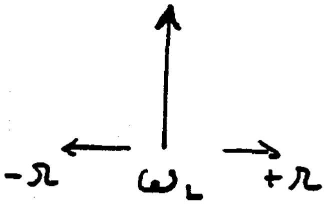
$S(\Omega) = 0$

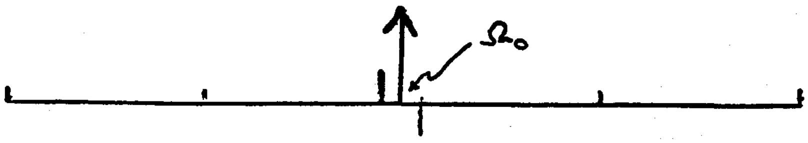
coherent state
vacuum

$S(\Omega, \theta) \rightarrow -1$

$S(\Omega, \theta + \frac{\pi}{2}) \rightarrow +\infty$

limit of perfect squeezing





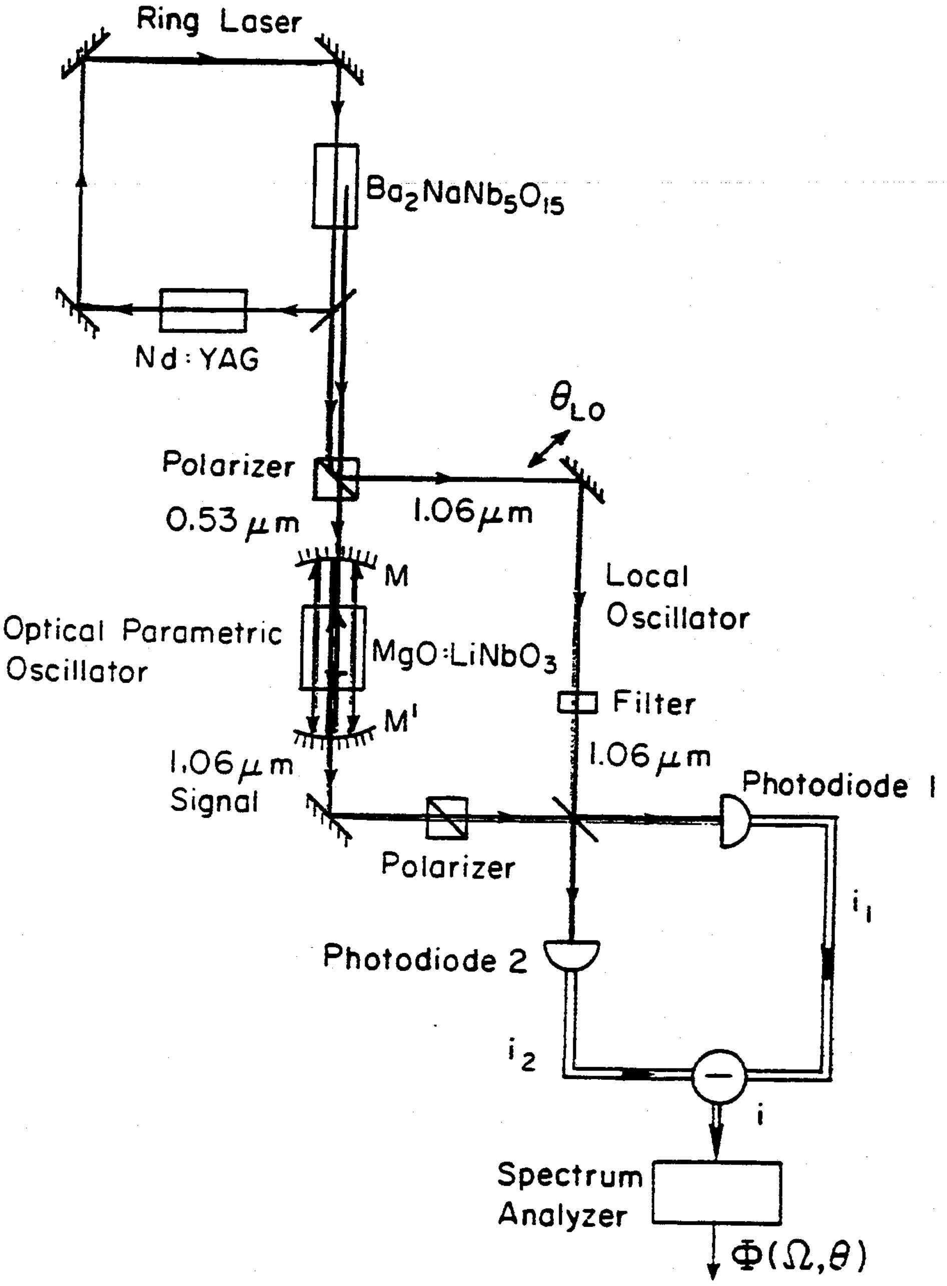
Ω_0 must sit under the squeeze curve — a technical headache

FM Probe Sidebands

at ± Ω_0

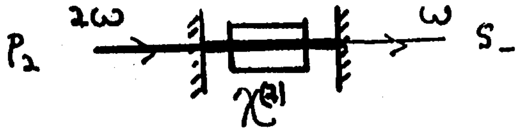
Relative to $S_\pm(\Omega)$

Experiment - Squeezing with an OPO

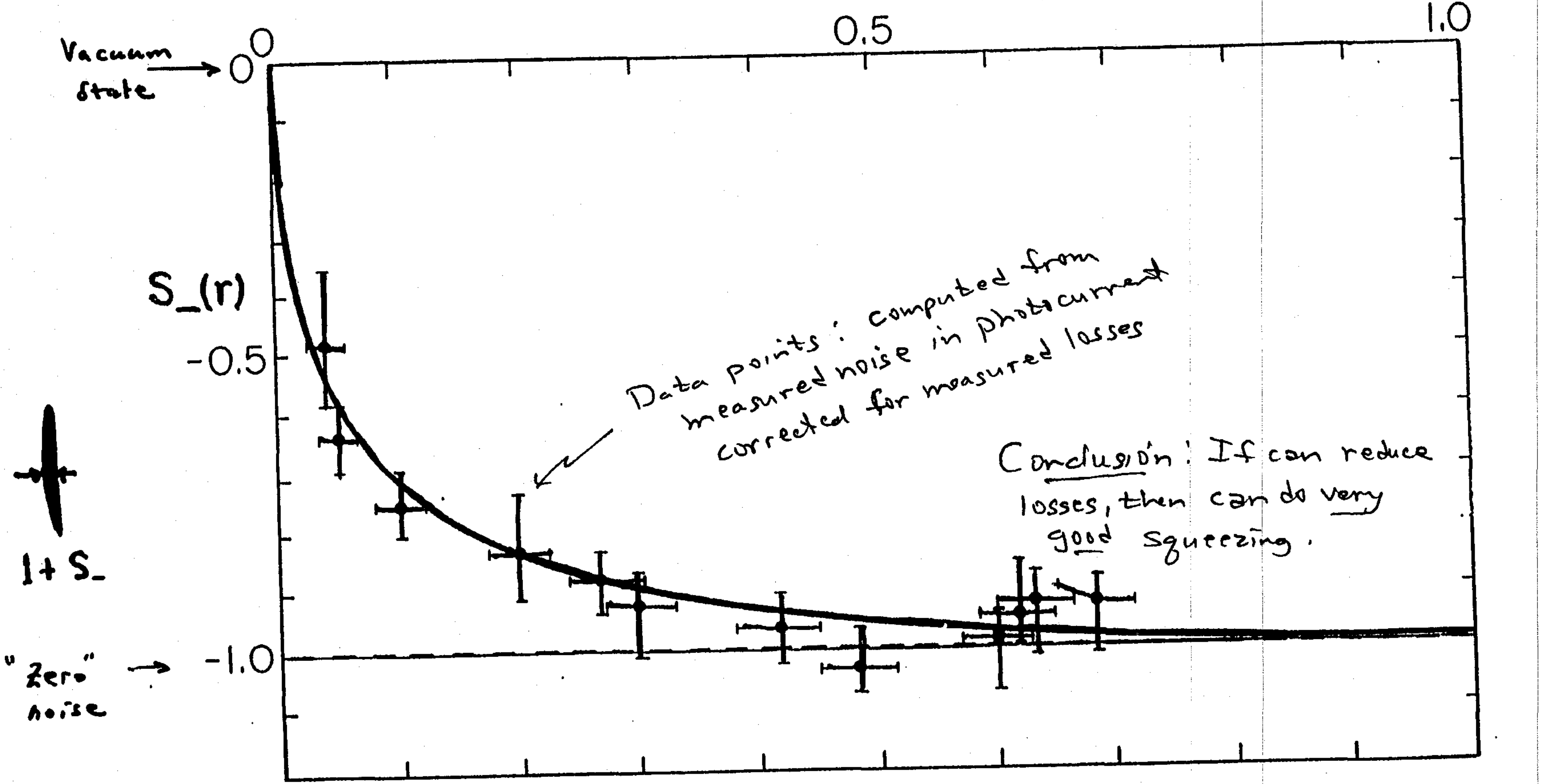


OPO

Inferred Degree of Squeezing vs. Pump Power - Absolute Measurement



How good would the amount of squeezing be if we could get rid of all the losses?



Data points: computed from measured noise in photocurrent corrected for measured losses

Conclusion: If can reduce losses, then can do very good squeezing.

$$R_- = 1 + \xi S_-$$
$$S_- = (R_- - 1)/\xi$$

Wu et al., PRL 57, 2520 (1986).

LECTURE 16

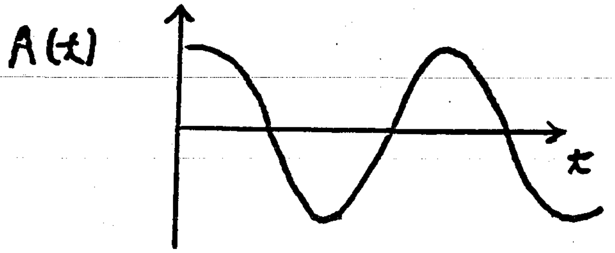
"Squeezed Light and its Potential Use in LIGO - PART II

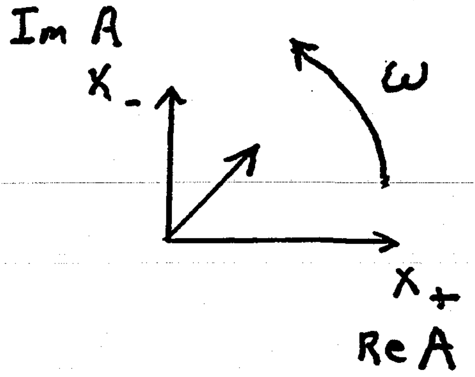
by

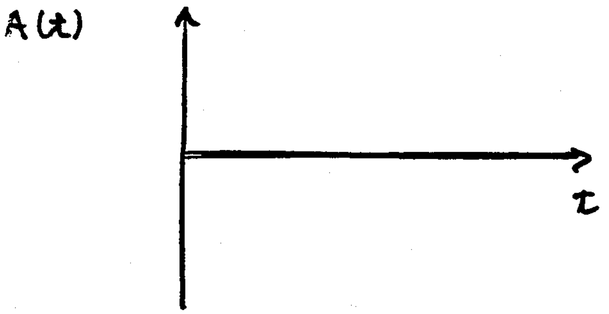
JEFF KIMBLE

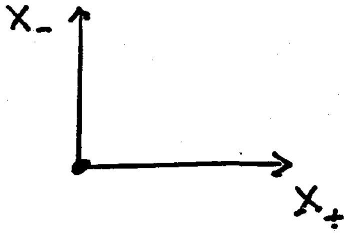
Pictographs for Squeezing

Classical Field

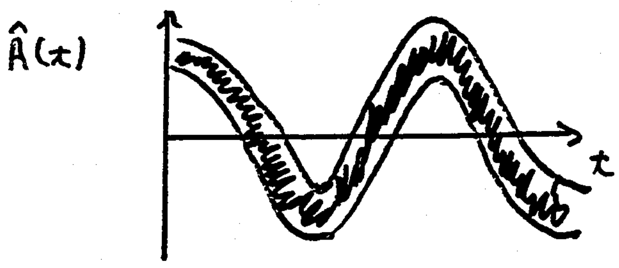


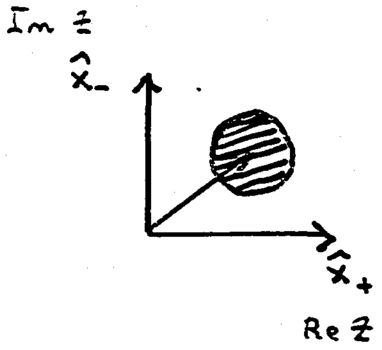




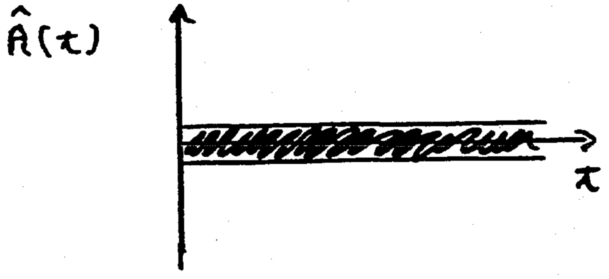


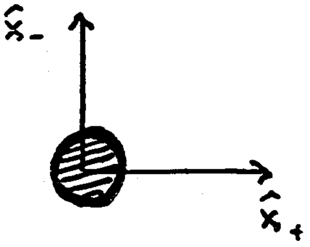
Quantum Field





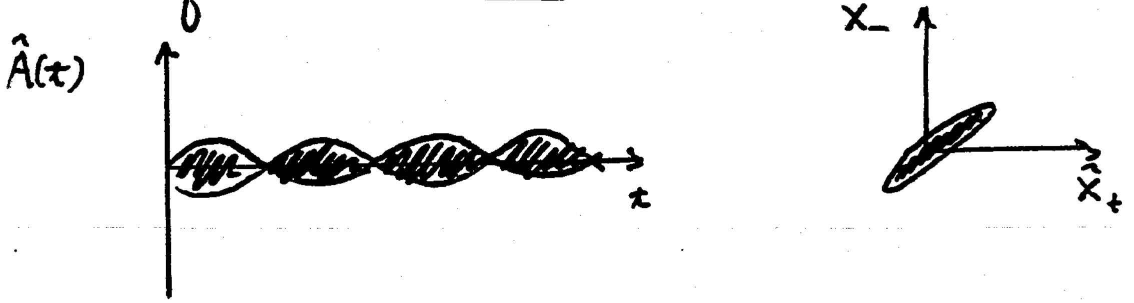
Quantum Vacuum



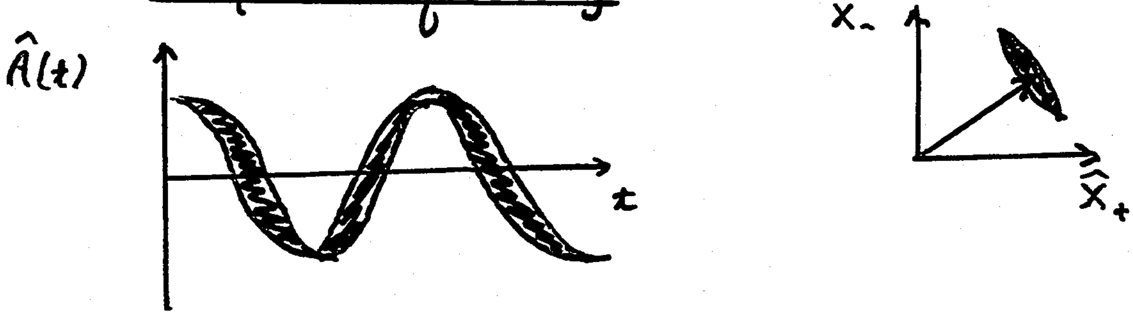


Other Possibilities

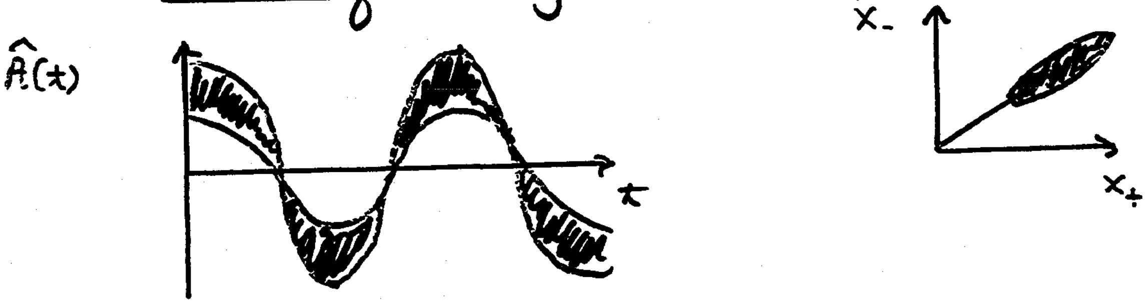
Squeezed Vacuum



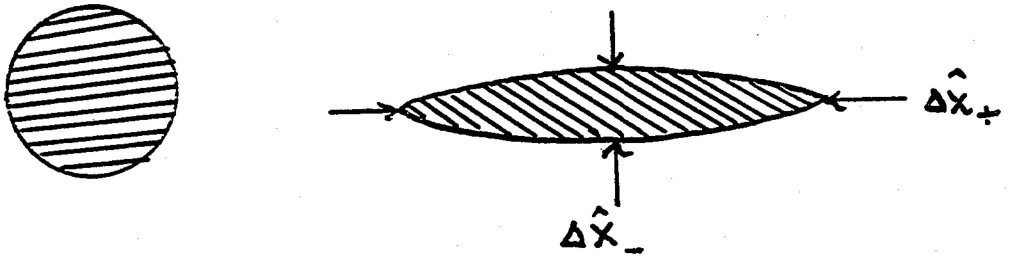
Amplitude Squeezing



Phase Squeezing



Squeezing - Rules and Regulations for Quantum "Fuzz Balls"



$$\langle(\Delta\hat{X}_+)^2\rangle\langle(\Delta\hat{X}_-)^2\rangle \geq 1$$

Frequency Correlations for Squeezed Light

Introduce spectral decomposition of field

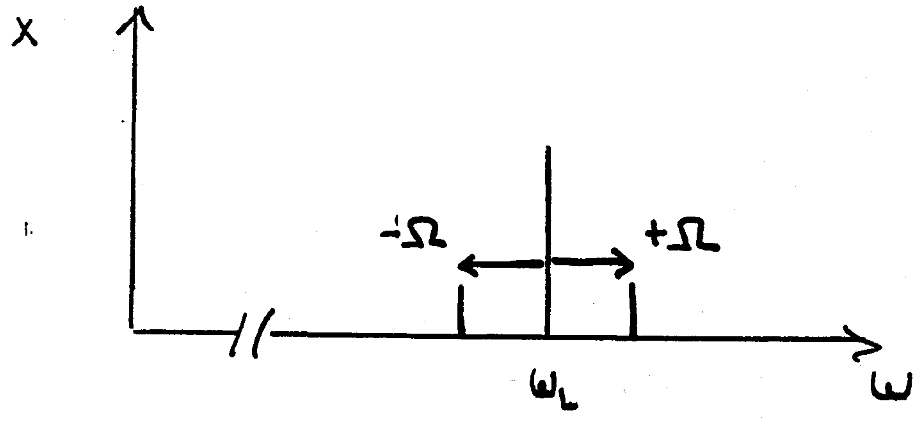
$$a(t) = \frac{1}{2\pi}\int d\Omega\, \underset{\substack{\uparrow\\ \text{annihilation}}}{a(\Omega)}\, e^{-i\Omega t}\ ,\qquad a^\dagger(t) = \frac{1}{2\pi}\int d\Omega\, \underset{\substack{\uparrow\\ \text{creation}}}{a^\dagger(\Omega)}\, e^{i\Omega t}$$

and

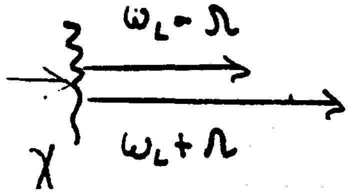
$$X_\theta(t) = \frac{1}{2\pi}\int d\Omega\, \underset{\substack{\uparrow\\ \text{significance?}}}{X_\theta(\Omega)}\, e^{-i\Omega t}$$

In terms of $\{a(\Omega), a^\dagger(\Omega)\}$,

$$\underset{\substack{\uparrow\\ [\text{illegible}]}}{X_\theta(\Omega)} = e^{-i\theta}\, \underset{\substack{\uparrow\\ \text{annihilation}\\ \text{at } \Omega}}{a(\Omega)} + e^{i\theta}\, \underset{\substack{\uparrow\\ \text{creation}\\ \text{at } -\Omega}}{a^\dagger(-\Omega)}$$



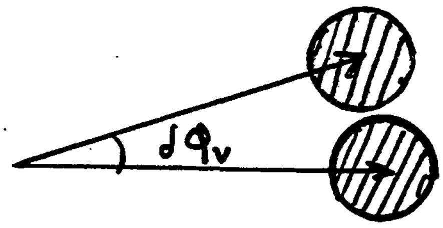
* Parametric process introduces correlations between fields at $\omega_L+\Omega$ and $\omega_L-\Omega$

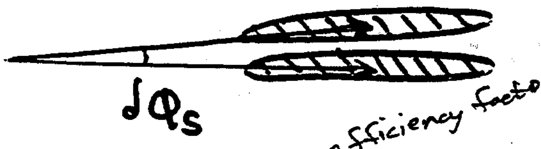


Squeezed Light for

Sensitivity Beyond the Vacuum-State Limit

Phase Changes



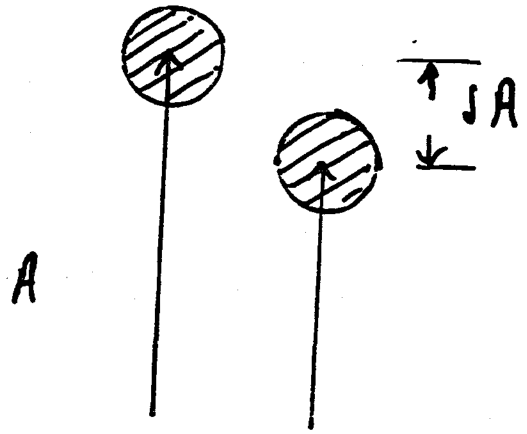


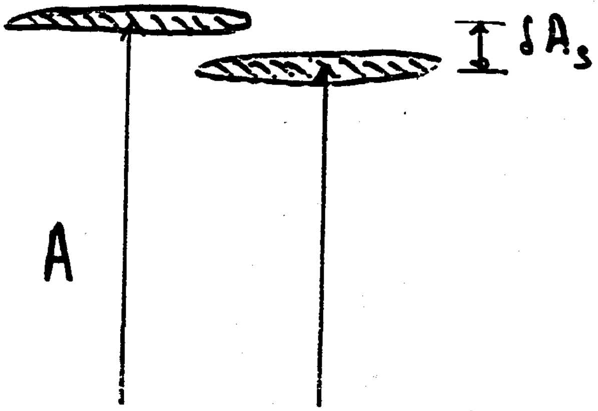
$$\delta\varphi_v \sim 1/\sqrt{N}$$

$$\delta\varphi_s \sim \left[1+\overset{\text{efficiency factor}}{\alpha} S_-\right]^{1/2}/\sqrt{N}$$

$$\left.\begin{array}{l}\alpha \simeq 1\\ S_- \rightarrow -1\end{array}\right\}\ \delta\varphi_s \rightarrow 0$$

Amplitude Changes





$$\delta A_v/A \sim 1/\sqrt{N}$$

$$\delta A_s/A \sim \left[1+\alpha S_+\right]^{1/2}/\sqrt{N}$$

PHYSICAL REVIEW D VOLUME 23, NUMBER 8 15 APRIL 1981

Quantum-mechanical noise in an interferometer

Carlton M. Caves

W. K. Kellogg Radiation Laboratory, California Institute of Technology, Pasadena, California 91125

(Received 15 August 1980)

The interferometers now being developed to detect gravitational waves work by measuring the relative positions of widely separated masses. Two fundamental sources of quantum-mechanical noise determine the sensitivity of such an interferometer: (i) fluctuations in number of output photons (photon-counting error) and (ii) fluctuations in radiation pressure on the masses (radiation-pressure error). Because of the low power of available continuous-wave lasers, the sensitivity of currently planned interferometers will be limited by photon-counting error. This paper presents an analysis of the two types of quantum-mechanical noise, and it proposes a new technique—the "squeezed-state" technique—that allows one to decrease the photon-counting error while increasing the radiation-pressure error, or vice versa. The key requirement of the squeezed-state technique is that the state of the light entering the interferometer's normally unused input port must be not the vacuum, as in a standard interferometer, but rather a "squeezed state"—a state whose uncertainties in the two quadrature phases are unequal. Squeezed states can be generated by a variety of nonlinear optical processes, including degenerate parametric amplification.

I. INTRODUCTION

The task of detecting gravitational radiation is driving dramatic improvements in a variety of technologies for detecting very weak forces.[1] These improvements are forcing a careful examination of quantum-mechanical limits on the accuracy with which one can monitor the state of a macroscopic body on which a weak force acts.[2] One promising technology uses an interferometer to monitor the relative positions of widely separated masses. This paper analyzes the quantum-mechanical limits on the performance of interferometers, and it introduces a new technique that might lead to improvements in their sensitivity.

The prototypal interferometer for gravitational-wave detection is a two-arm, multireflection Michelson system, powered by a laser (see Fig. 3 below). The intensity in either of the interferometer's output ports provides information about the difference $z \equiv z_2 - z_1$ between the end mirrors' positions relative to the beam splitter, and changes in z reveal the passing of a gravitational wave. The first interferometer for gravitational-wave detection was built and operated at the Hughes Research Laboratories in Malibu, California, in the early 1970's (Ref. 3); this first effort was small-scale and had modest sensitivity. Now several groups around the world are developing interferometers of greatly improved sensitivity.[4-6] A long-range goal is to construct large-scale interferometers, with baselines $l \sim 1$ km, in order to achieve a strain sensitivity $\Delta z/l \sim 10^{-21}$ for frequencies from about 30 Hz to 10 kHz. This sensitivity goal is based on estimates for the strength of gravitational waves that pass the Earth reasonably often.[1]

It has been known for some time that quantum mechanics limits the accuracy with which an interferometer can measure z—or, indeed, the accuracy with which any position-sensing device can determine the position of a free mass.[2,5,7] In a measurement of duration τ, the probable error in the interferometer's determination of z can be no smaller than the "standard quantum limit":

$$(\Delta z)_{\mathrm{SQL}} = (2\hbar\tau/m)^{1/2}, \qquad (1.1)$$

where m is the mass of each end mirror $[(\Delta z)_{\mathrm{SQL}} \sim 6 \times 10^{-18}$ cm for $m \sim 10^5$ g, $\tau \sim 2 \times 10^{-3}$ sec]. The validity of the standard quantum limit is unquestionable, resting as it does solely on the Heisenberg uncertainty principle applied to the quantum-mechanical evolution of a free mass.

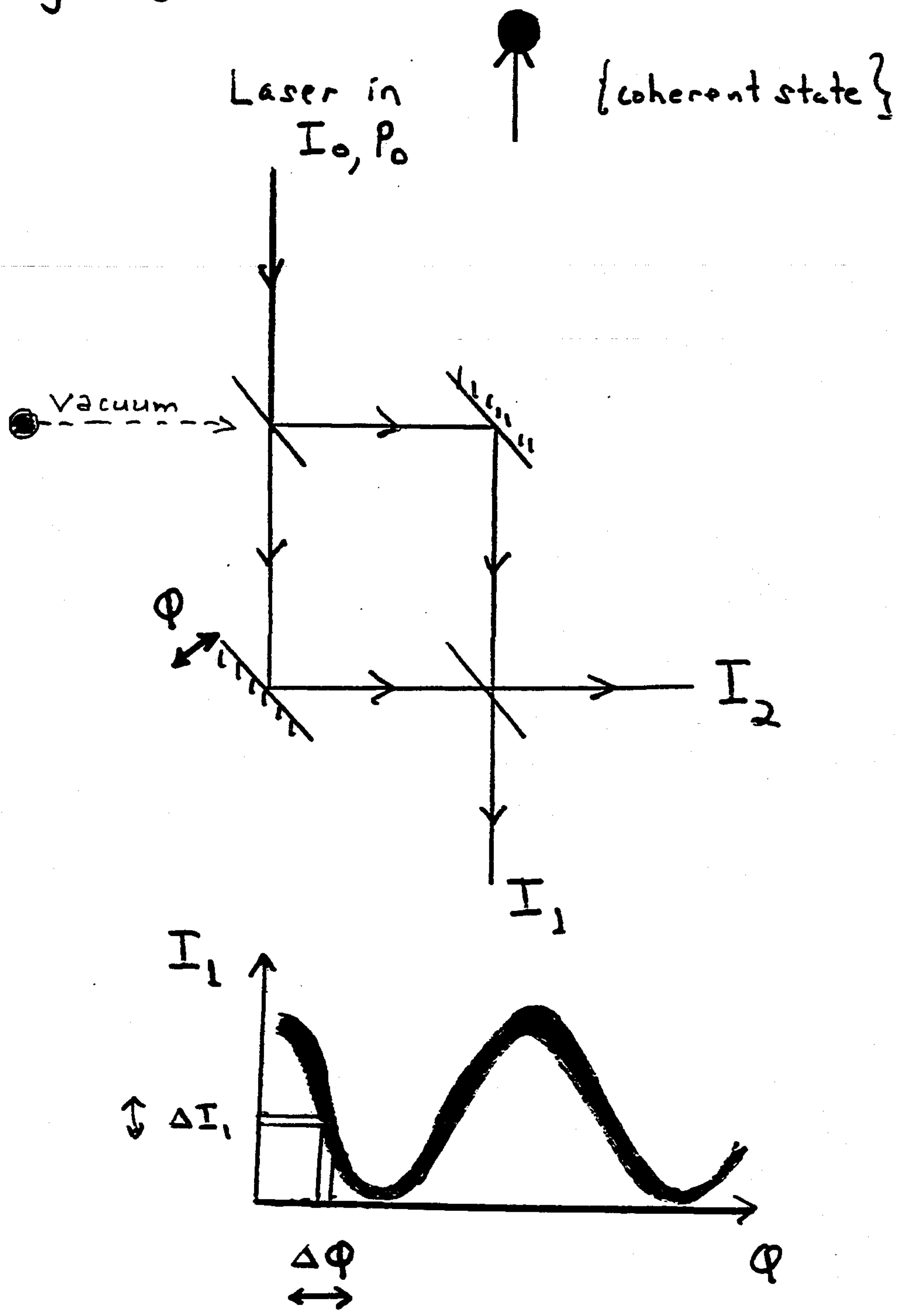
The standard quantum limit for an interferometer can also be obtained from a more detailed argument[5,8-10] that balances two sources of error: (i) the error in determining z due to fluctuations in the number of output photons (photon-counting error) and (ii) the perturbation of z during a measurement produced by fluctuating radiation-pressure forces on the end mirrors (radiation-pressure error). As the input laser power P increases, the photon-counting error decreases, while the radiation-pressure error increases. Minimizing the total error with respect to P yields a minimum error of order the standard quantum limit and an optimum input power[9,11]

$$P_0 \simeq \tfrac{1}{2}(mc^2/\tau)(1/\omega\tau)(1/b^2) \qquad (1.2)$$

at which the minimum error can be achieved. Here ω is the angular frequency of the light, and b is the number of bounces at each end mirror.

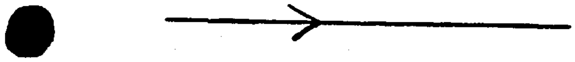
© 1981 The American Physical Society

IV. Interferometry Beyond Vacuum-State Limit ↔ Caves

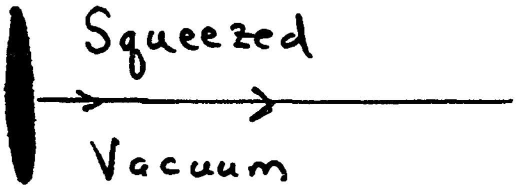


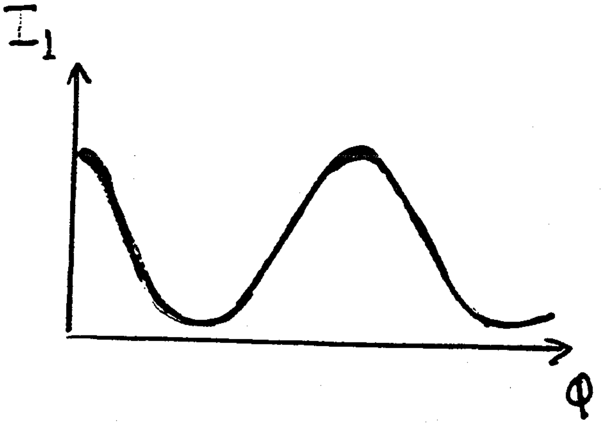
[Overlay for previous transparency]

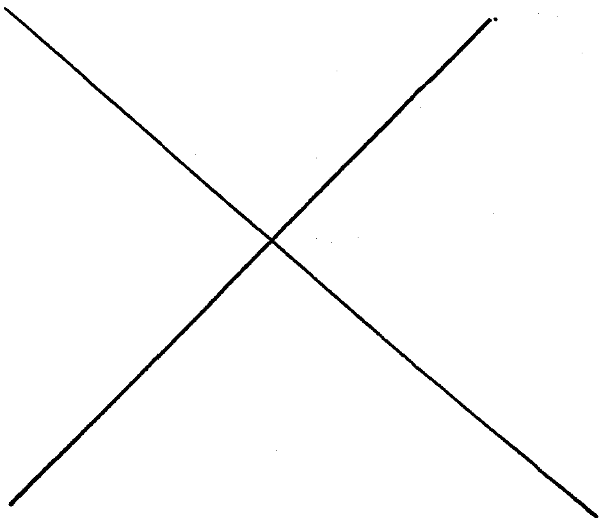
Vacuum



[overlay again]







Xiao, Wu, Kimble
PRL 59, 278 (1987).

$$\left.\begin{array}{l}\text{Vacuum - State}\\ \text{Shot - Noise}\end{array}\right\}\text{Limit}$$

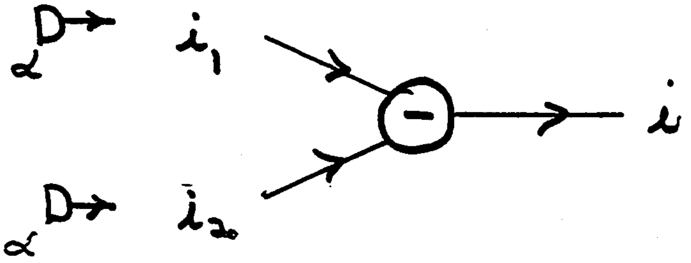
(How one does the calculations corresponding to previous diagrams.)

<u>Power</u> $P_{1,2} = T\frac{P_0}{2}\left[1 \pm \cos\varphi\right]$

$$\varphi = \frac{\pi}{2} + 2\delta_0 \cos\Omega t$$

↑ (sit on side of fringe)

<u>Current</u> $i_{1,2} = \alpha e P_{1,2}$



<u>Signal</u> Coherent modulation at Ω − $i_s = \sqrt{2}\, e T \alpha P_0 \delta_0$

<u>Noise</u> "Shot-Noise" $i_n^2 = 2eiB$

<u>**Signal to Noise**</u> $\psi = i_s^2 / i_n^2$

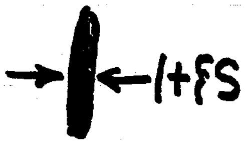
$$= \delta_0^2 N$$

where $N = T\alpha P_0 B^{-1}$ = (number of photons that get turned into photoelectrons)

$\Rightarrow \psi = 1$ for $\delta_0 = \frac{1}{\sqrt{N}}$

(Overlay)

Improvement with Squeezed Light



Spectrum of squeezing S

$-1 \le S \le 0$

Efficiency ξ

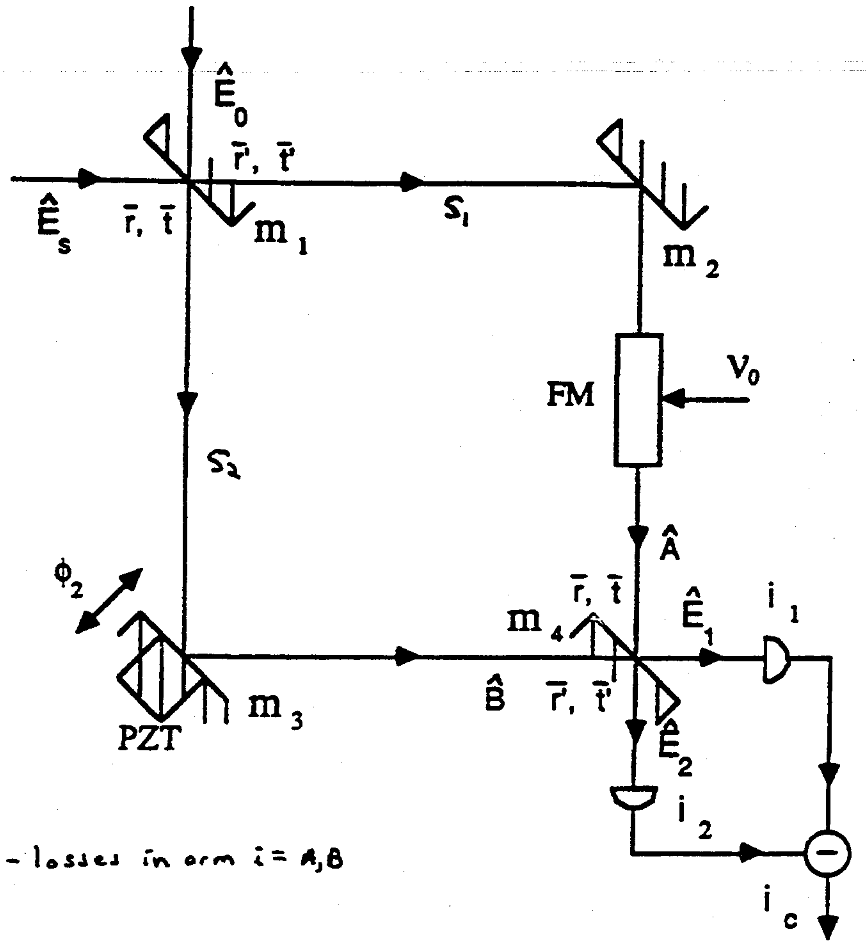
$0 \le \xi \le 1$

$[1+\xi S]$

$/[1+\xi S]$

$[1+\xi S]^{1/2}$

The Real Story



S_i - losses in arm $i = A, B$

Photocurrent Fluctuations $<\Delta i(t)\Delta i(t+\tau)>$

$$\begin{aligned}
\underline{<\Delta i_c(t)\Delta i_c(t+\tau)>} = {} & R_1(t)Q_1^2\delta(\tau) + R_2(t)Q_2^2\delta(\tau) \\
& + \underline{H_1}[\underline{<T:\hat{E}_s^\dagger(t)\hat{E}_s^\dagger(t+\tau)\hat{E}_s(t+\tau)\hat{E}_s(t)> - <\hat{E}_s^\dagger(t)\hat{E}_s(t)>^2}] \\
& + \underline{H_2}[\underline{<T:\hat{E}_0^\dagger(t)\hat{E}_0^\dagger(t+\tau)\hat{E}_0(t+\tau)\hat{E}_0(t)> - <\hat{E}_0^\dagger(t)\hat{E}_0(t)>^2}] \\
& + \underline{H_3}\ \underline{<\hat{E}_0(t+\tau)\hat{E}_0(t)><\hat{E}_s^\dagger(t)\hat{E}_s^\dagger(t+\tau)>} \\
& + \underline{H_4}\ \underline{<\hat{E}_0^\dagger(t)\hat{E}_0^\dagger(t+\tau)><\hat{E}_s(t+\tau)\hat{E}_s(t)>} \\
& + \underline{H_5}\ \underline{<\hat{E}_0^\dagger(t+\tau)\hat{E}_0(t)><\hat{E}_s^\dagger(t)\hat{E}_s(t+\tau)>} \\
& + \underline{H_6}\ \underline{<\hat{E}_0^\dagger(t)\hat{E}_0(t+\tau)><\hat{E}_s^\dagger(t+\tau)\hat{E}_s(t)>}, \qquad (4-44)
\end{aligned}$$

where

→ Quantum Statistical Characteristics of Incident Fields

$$\begin{aligned}
\underline{H_1} \equiv {} & \bar{\alpha}_1^2Q_1^2\underline{k_1(t)k_1(t+\tau)} + \bar{\alpha}_2^2Q_2^2\underline{\ell_1(t)\ell_1(t+\tau)} \\
& - \bar{\alpha}_1\bar{\alpha}_2Q_1Q_2\underline{k_1(t)\ell_1(t+\tau)} - \bar{\alpha}_1\bar{\alpha}_2Q_1Q_2\underline{k_1(t+\tau)\ell_1(t)}, \\
\underline{H_2} \equiv {} & \bar{\alpha}_1^2Q_1^2\underline{k_4(t)k_4(t+\tau)} + \bar{\alpha}_2^2Q_2^2\underline{\ell_4(t)\ell_4(t+\tau)} \\
& - \bar{\alpha}_1\bar{\alpha}_2Q_1Q_2\underline{k_4(t)\ell_4(t+\tau)} - \bar{\alpha}_1\bar{\alpha}_2Q_1Q_2\underline{k_4(t+\tau)\ell_4(t)}, \\
\underline{H_3} \equiv {} & \bar{\alpha}_1^2Q_1^2\underline{k_2(t)k_2(t+\tau)} + \bar{\alpha}_2^2Q_2^2\underline{\ell_2(t)\ell_2(t+\tau)} \\
& - \bar{\alpha}_1\bar{\alpha}_2Q_1Q_2\underline{k_2(t)\ell_2(t+\tau)} - \bar{\alpha}_1\bar{\alpha}_2Q_1Q_2\underline{k_2(t+\tau)\ell_2(t)}, \\
\underline{H_4} \equiv {} & \bar{\alpha}_1^2Q_1^2\underline{k_3(t)k_3(t+\tau)} + \bar{\alpha}_2^2Q_2^2\underline{\ell_3(t)\ell_3(t+\tau)} \\
& - \bar{\alpha}_1\bar{\alpha}_2Q_1Q_2\underline{k_3(t)\ell_3(t+\tau)} - \bar{\alpha}_1\bar{\alpha}_2Q_1Q_2\underline{k_3(t+\tau)\ell_3(t)}, \\
\underline{H_5} \equiv {} & \bar{\alpha}_1^2Q_1^2\underline{k_2(t)k_3(t+\tau)} + \bar{\alpha}_2^2Q_2^2\underline{\ell_2(t)\ell_3(t+\tau)} \qquad (4-45) \\
& - \bar{\alpha}_1\bar{\alpha}_2Q_1Q_2\underline{k_2(t)\ell_3(t+\tau)} - \bar{\alpha}_1\bar{\alpha}_2Q_1Q_2\underline{k_3(t+\tau)\ell_2(t)}, \\
\underline{H_5} \equiv {} & \bar{\alpha}_1^2Q_1^2\underline{k_2(t+\tau)k_3(t)} + \bar{\alpha}_2^2Q_2^2\underline{\ell_2(t+\tau)\ell_3(t)} \\
& - \bar{\alpha}_1\bar{\alpha}_2Q_1Q_2\underline{k_3(t)\ell_2(t+\tau)} - \bar{\alpha}_1\bar{\alpha}_2Q_1Q_2\underline{k_2(t+\tau)\ell_3(t)},
\end{aligned}$$

Mach-Zehnder with Squeezed Light

Photocurrent Fluctuations, cont.

$$\hat{E}_s(r,t) = \hat{\mathcal{E}}_s(t)\, V(\vec{r})$$
$$\hat{E}_o(r,t) = \hat{E}_o(t)\, u(\vec{r})$$

$$\begin{aligned}
\bar{\eta} &\equiv \left|\iint_S V_i^*(\vec{r})U_i(\vec{r})dS\right| = \left|\iint_S U_i^*(\vec{r})V_i(\vec{r})dS\right|,\\
\bar{\nu} &\equiv \left|\iint_S V_i^*(\vec{r})U_j(\vec{r})dS\right| = \left|\iint_S U_i^*(\vec{r})V_j(\vec{r})dS\right|,\\
\bar{\mu} &\equiv \left|\iint_S V_i^*(\vec{r})V_j(\vec{r})dS\right|,\\
\bar{\varepsilon} &\equiv \left|\iint_S U_i^*(\vec{r})U_j(\vec{r})dS\right|,\\
i,j &= 1,2 \quad \text{and} \quad i \neq j.
\end{aligned} \tag{4-34}$$

$$\begin{aligned}
&\underline{k_1(t)} - \bar{R}\bar{T}\left[S_1^2 + S_2^2 + 2S_1S_2\bar{\mu}\,\underline{\cos\Delta\phi(t)}\right]\\
&\underline{k_2(t)} = \sqrt{\bar{R}\bar{T}}e^{i(-\varphi_1+\varphi_2)}\left[\bar{R}S_1^2\bar{\eta} - \bar{T}S_2^2\bar{\eta} - S_1S_2\bar{\nu}\left(\bar{T}e^{-i\underline{\Delta\phi(t)}} - \bar{R}e^{i\underline{\Delta\phi(t)}}\right)\right]\\
&\underline{k_3(t)} = \sqrt{\bar{R}\bar{T}}e^{-i(-\varphi_1+\varphi_2)}\left[\bar{R}S_1^2\bar{\eta} - \bar{T}S_2^2\bar{\eta} + S_1S_2\bar{\nu}\left(\bar{R}e^{-i\underline{\Delta\phi(t)}} - \bar{T}e^{i\underline{\Delta\phi(t)}}\right)\right]\\
&\underline{k_4(t)} = \bar{R}^2S_1^2 + \bar{T}^2S_2^2 - 2\bar{R}\bar{T}S_1S_2\bar{\varepsilon}\,\underline{\cos\Delta\phi(t)},
\end{aligned} \tag{4-38}$$

and

$$\begin{aligned}
&\underline{\ell_1(t)} = \bar{T}^2S_1^2 + \bar{R}^2S_2^2 - 2\bar{R}\bar{T}S_1S_2\bar{\mu}\,\underline{\cos\Delta\phi(t)}\\
&\underline{\ell_2(t)} = \sqrt{\bar{R}\bar{T}}e^{i(-\varphi_1+\varphi_2)}\left[\bar{T}S_1^2\bar{\eta} - \bar{R}S_2^2\bar{\eta} + S_1S_2\bar{\nu}\left(\bar{T}e^{-i\underline{\Delta\phi(t)}} - \bar{R}e^{i\underline{\Delta\phi(t)}}\right)\right]\\
&\underline{\ell_3(t)} = \sqrt{\bar{R}\bar{T}}e^{-i(-\varphi_1+\varphi_2)}\left[\bar{T}S_1^2\bar{\eta} - \bar{R}S_2^2\bar{\eta} - S_1S_2\bar{\nu}\left(\bar{R}e^{-i\underline{\Delta\phi(t)}} - \bar{T}e^{i\underline{\Delta\phi(t)}}\right)\right]\\
&\underline{\ell_4(t)} = \bar{R}\bar{T}\left[S_1^2 + S_2^2 + 2S_1S_2\bar{\varepsilon}\,\underline{\cos\Delta\phi(t)}\right].
\end{aligned} \tag{4-39}$$

$\underline{\qquad\qquad}$ signal

Min Xiao, PhD thesis

Interferometry Beyond The Vacuum State Limit → Caves

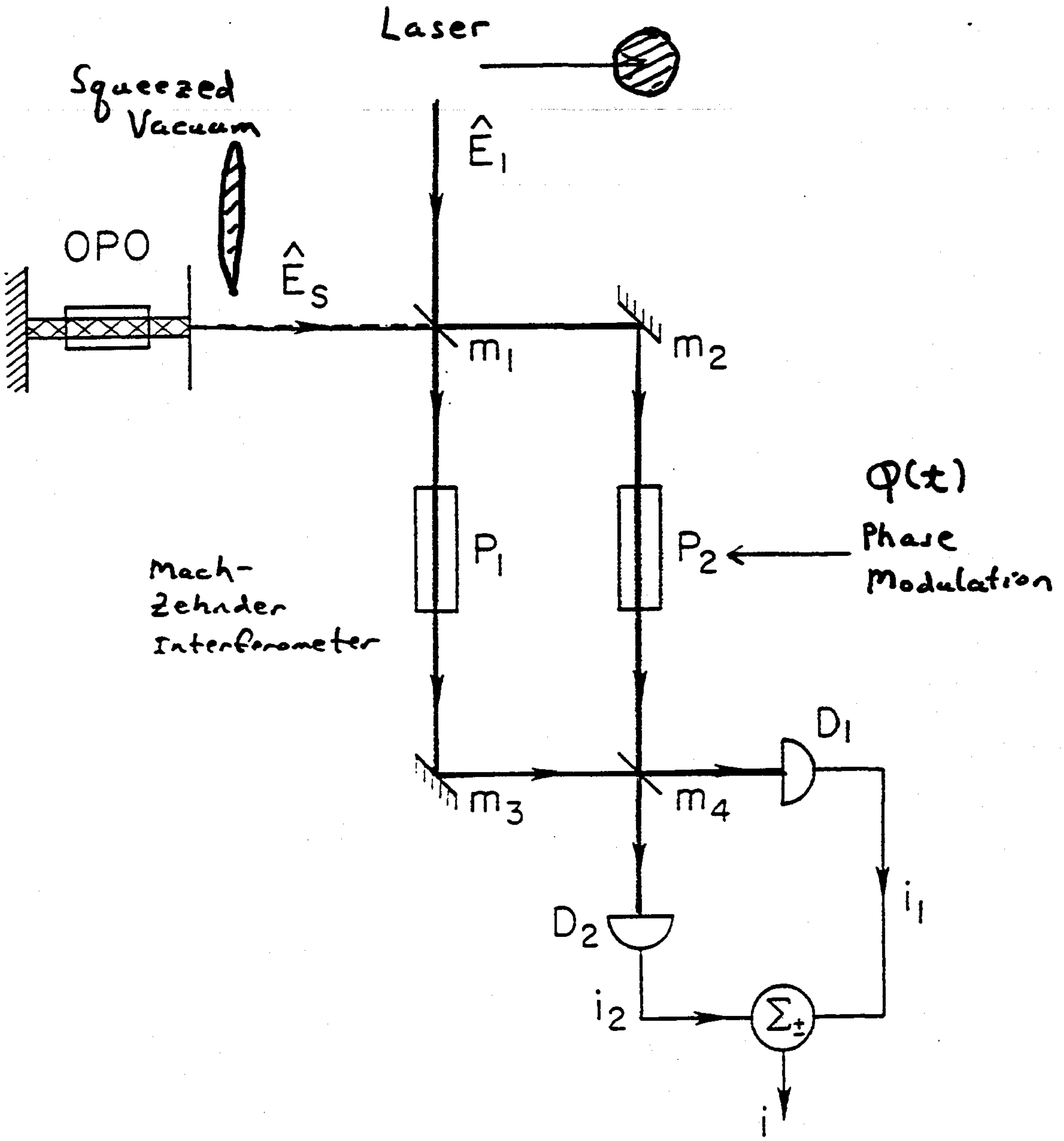


FIGURE 1

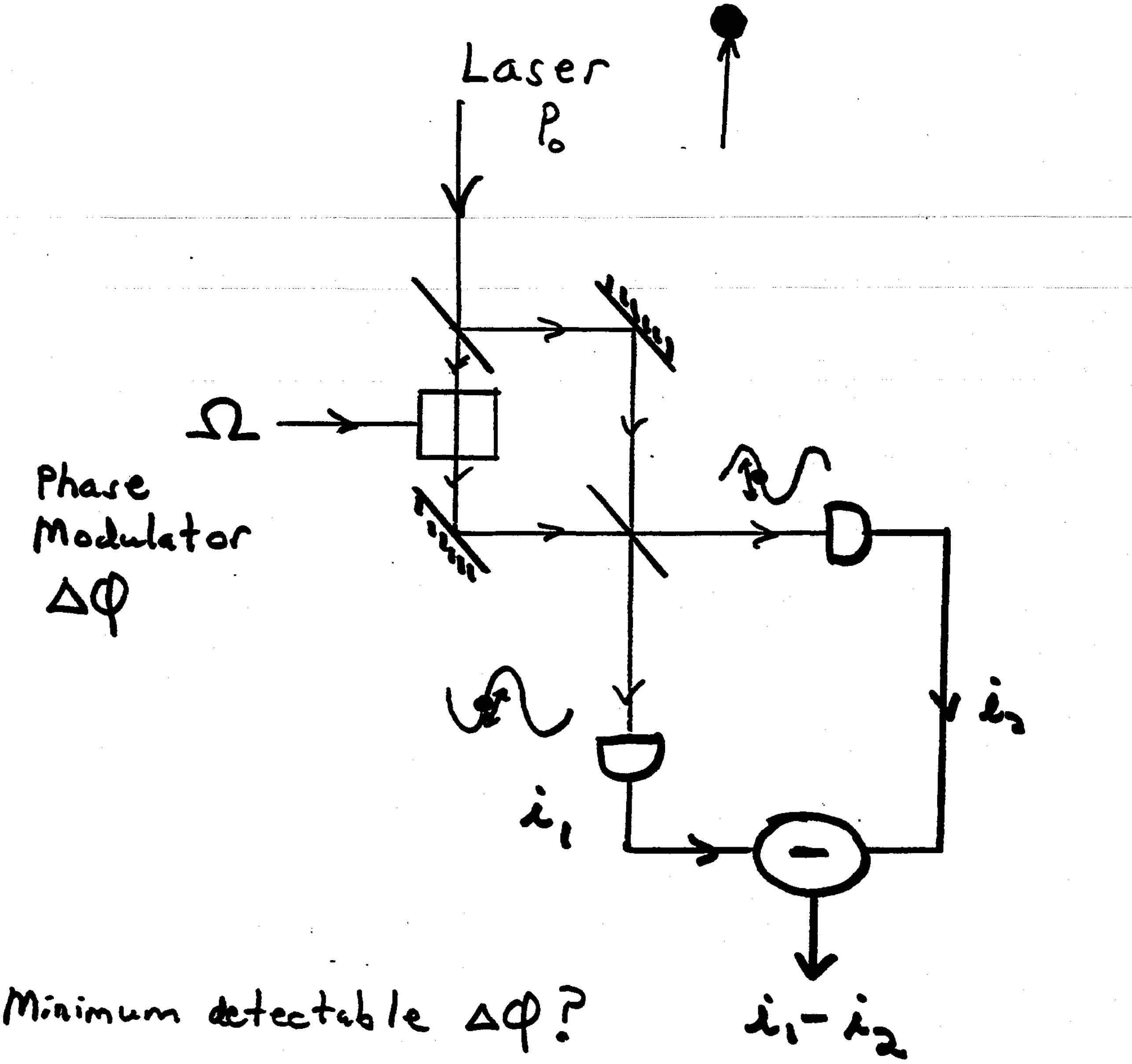
Photocurrent

Signal $\varphi(t)$

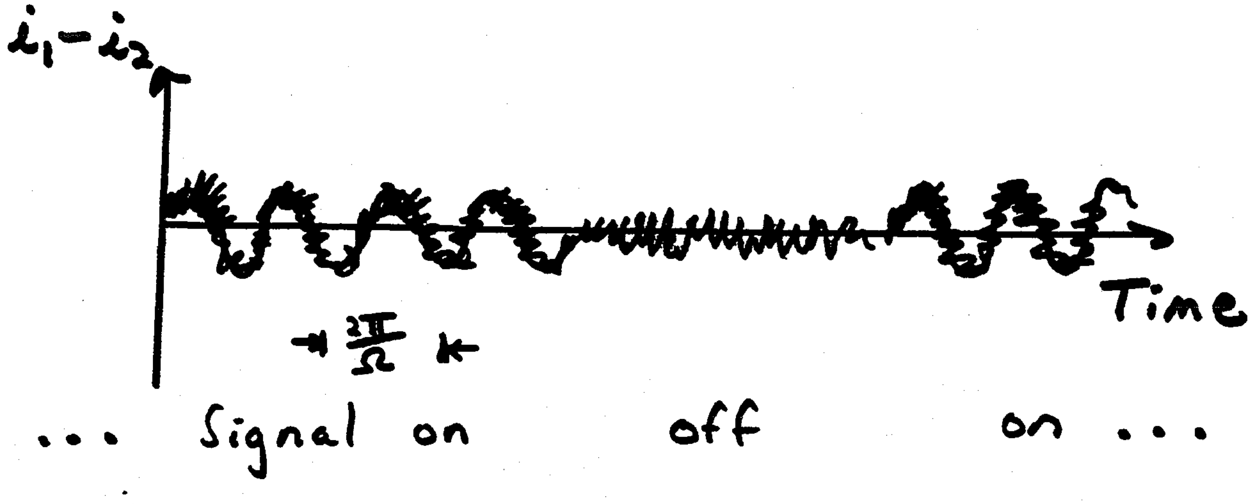
Noise ⬭ ◍

Xiao, Wu, HJK, Phys. Rev. Lett. 59, 278 (1987)

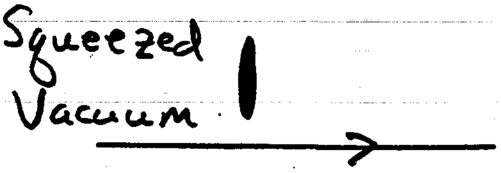
Overview of Experiment {Caves}

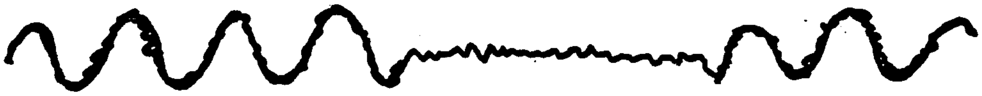


Minimum detectable $\Delta\phi$?



(Overlay)





Vacuum State Input ●

⇒ Shot-Noise Limit ~ 5 μrad rms

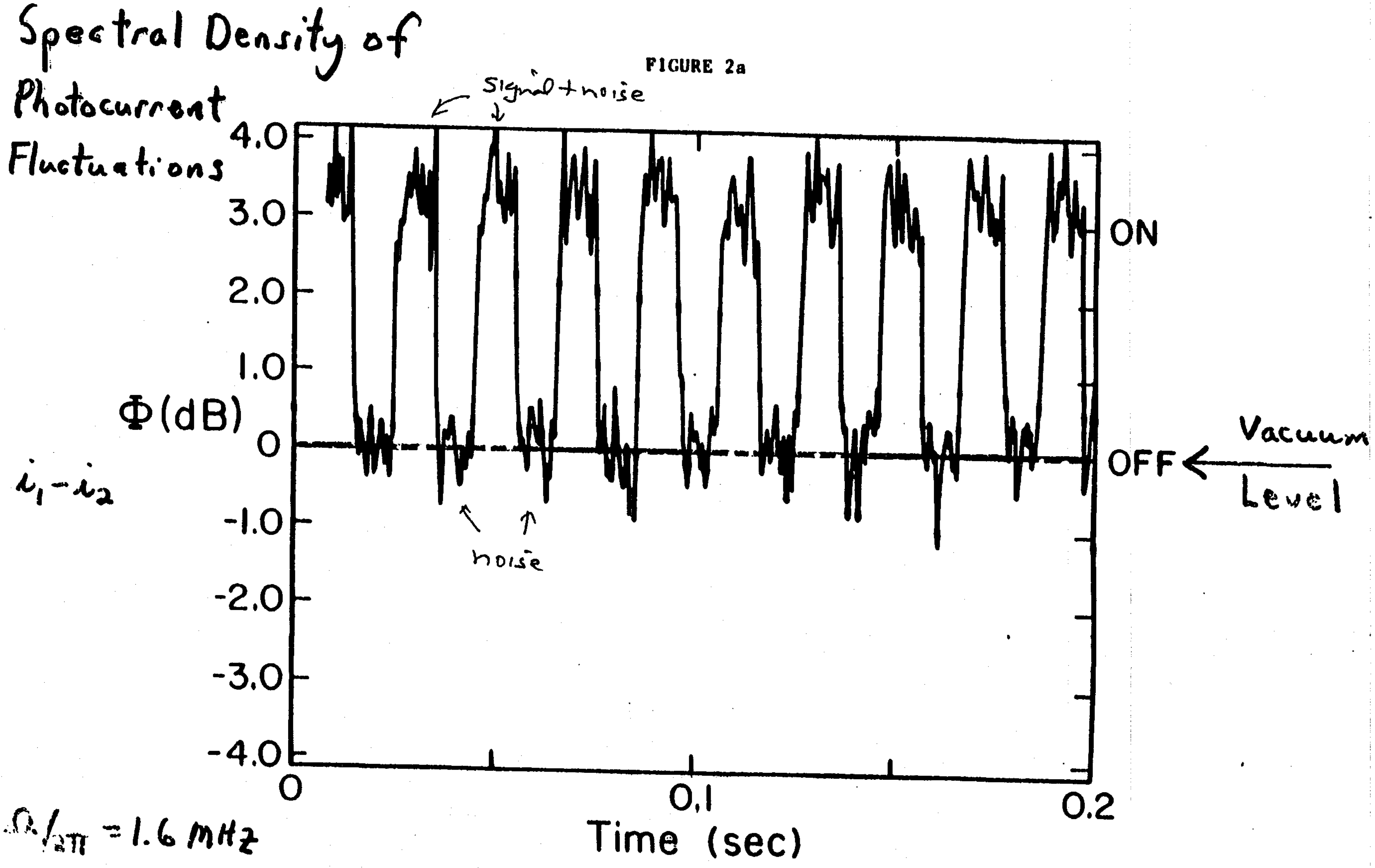


FIGURE 2a

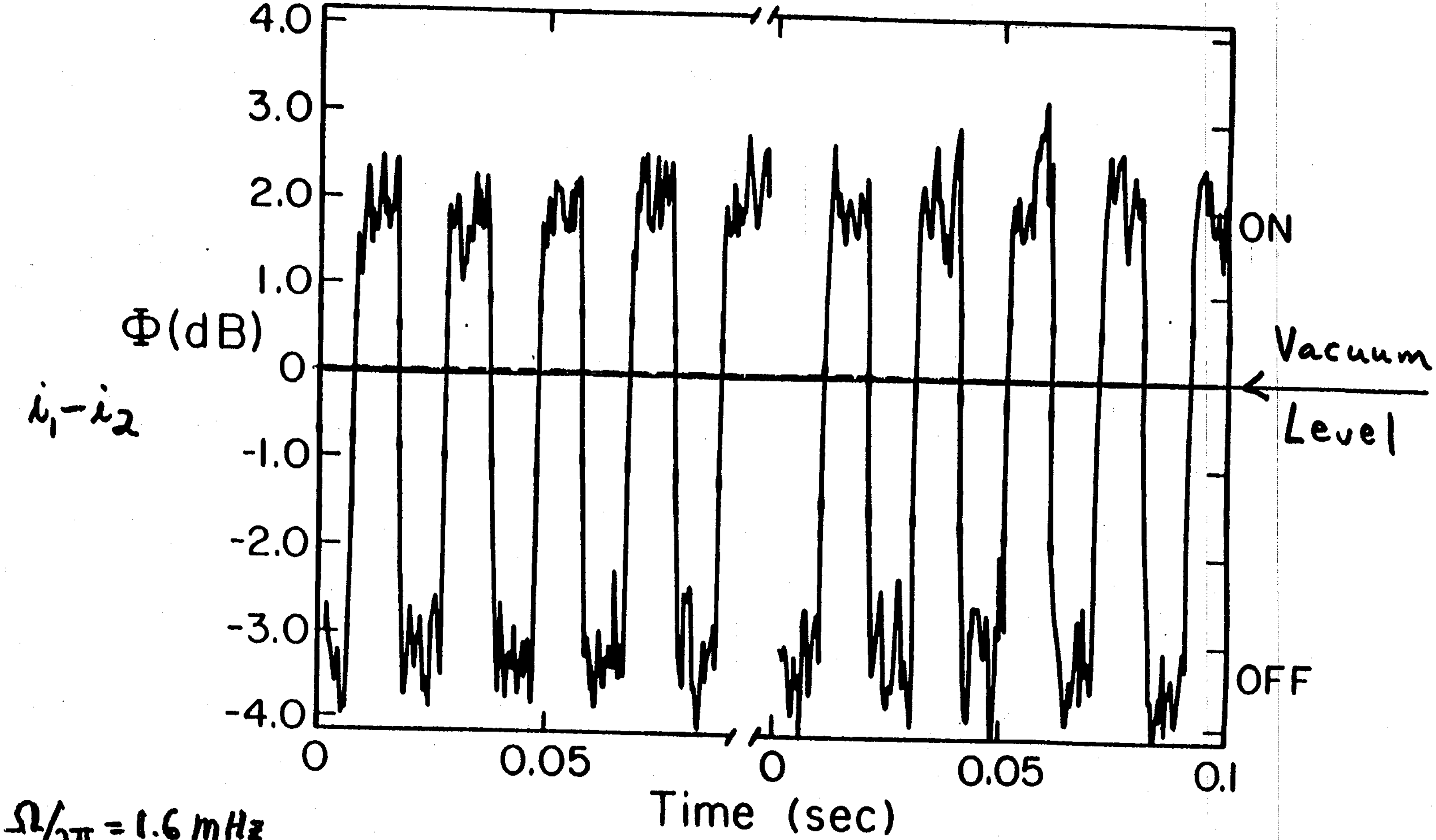
$\Omega/2\pi = 1.6$ MHz

$\Delta f = 100$ kHz $P = 800$ μW

Squeezed Input [redacted] $\{\sim 10^{-12}$ WATT$\}$

⇒ Improvement in S/N of 3.0 dB beyond Shot-Noise Limit

FIGURE 2b

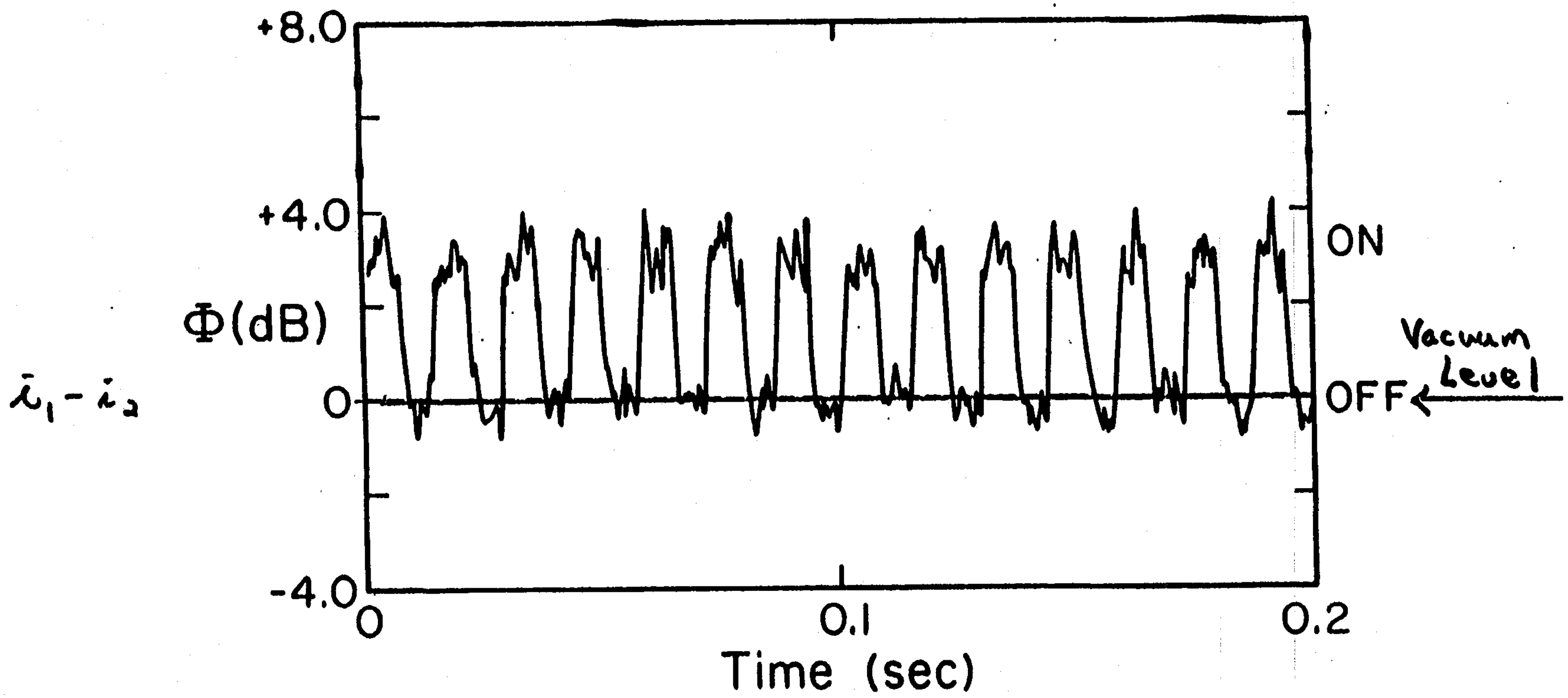


$\Omega/2\pi = 1.6$ mHz

$\Delta f = 100$ kHz

Vacuum State Input

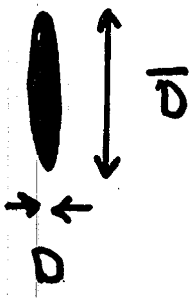
Spectral Density of photocurrent Fluctuations

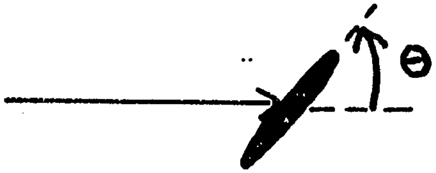


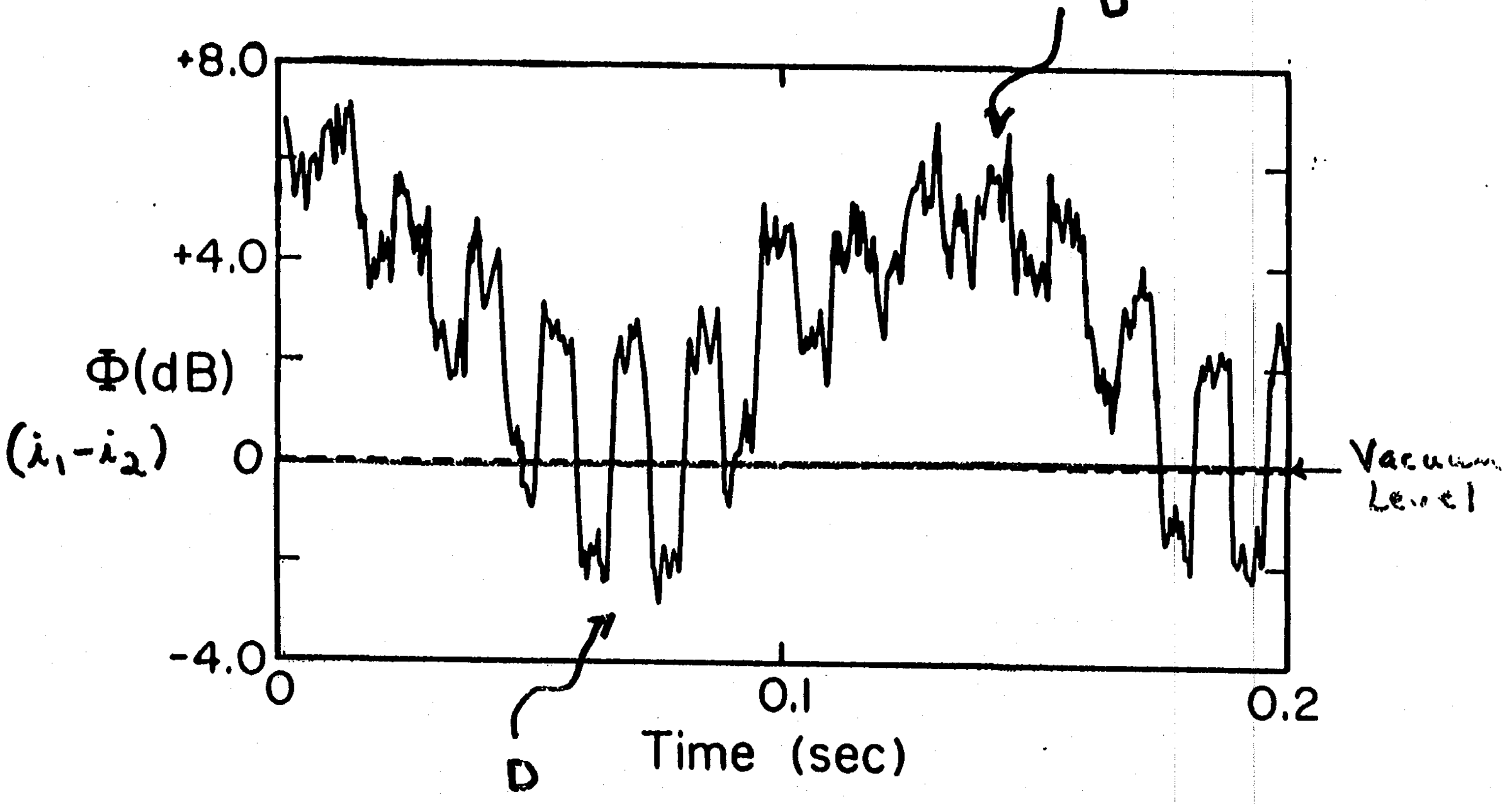
$\Omega/2\pi \simeq 1.6$ MHz
$\Delta F = 100$ kHz

Two Sides to the Coin:

"Darkness" and "Antidarkness"





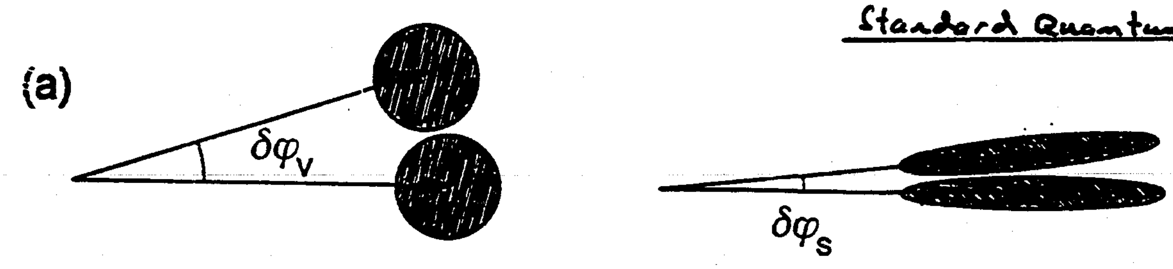


$\Omega/2\pi = 1.6$ MHz
$\Delta f = 100$ kHz

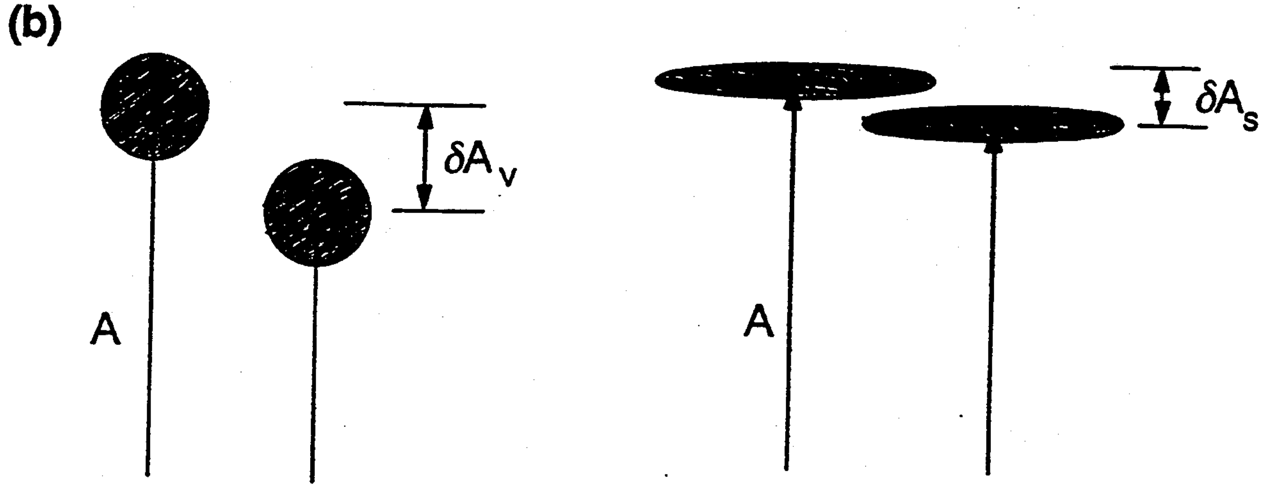
→ Phase Θ

Squeezed Light for Sensitivity Beyond Vacuum-State Limit

aka: Shot-Noise Limit,
Coherent-state Limit,
Standard Quantum Limit



$$\delta\varphi_v \simeq 1/\sqrt{N} \qquad\qquad \delta\varphi_s \simeq \Delta X_- / \sqrt{N}$$

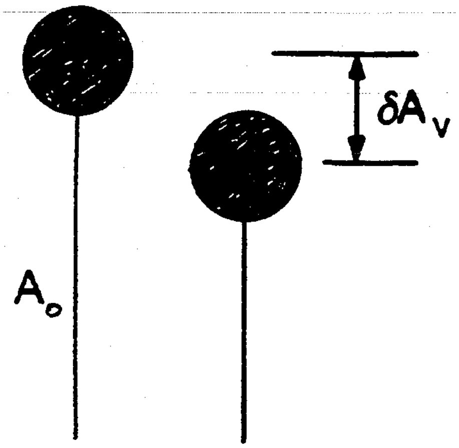


$$\delta A_v / A \simeq 1/\sqrt{N} \qquad\qquad \delta A_s / A \simeq \Delta X_+ / \sqrt{N}$$

(a)	Xiao, Wu, Kimble (87)	-3.0 dB	•		• Caltech U.T. Austin
	Grangier et al. (87)	-2.0 dB			
	Bergman et al. (92)	-3. dB			
(b)	Xiao, Wu, Kimble (87)	-2.5 dB	•		
	Polzik, Carri, Kimble (92)	-3.1 dB	•	[-3.8dB]	

Also, Nabors & Shelby (90) - Twin beams
Rarity & Tapster (90) - Photon pairs
Hong, Friberg, Mandel (85) - Photon pairs

"Ultimate" Limit for Sensitivity Enhancement with Squeezing



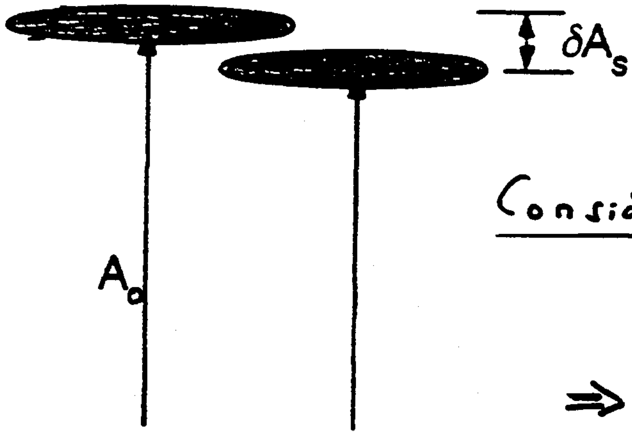
Sample loss γ

$$\gamma_v = \frac{\delta A_v}{A_o} \sim \frac{1}{\sqrt{N}}$$

# photons in measurement interval

Degree of Squeezing

Total loss = $1-\xi$



$$\gamma_s = \frac{\delta A_s}{A_o} \sim \frac{\left[1+\xi S_-\right]^{1/2}}{\sqrt{N}}$$

Consider limit $1-\xi \rightarrow \gamma_s$

$S_- \rightarrow -1$

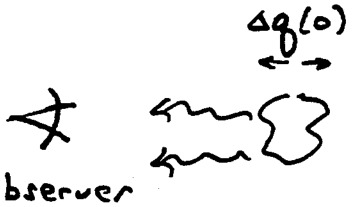
$$\Rightarrow \quad \gamma_s \sim \frac{\left[1-(1-\gamma_s)\right]^{1/2}}{\sqrt{N}}$$

$$\gamma_s \sim \frac{1}{N}$$

Enhancement

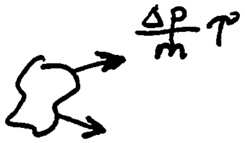
$\gamma_v / \gamma_s \sim \sqrt{N}$!

The Standard Quantum Limit (SQL) for the Position of a Free Mass

1. 

Measurement with accuracy $\Delta q(0)$

$$\Rightarrow \quad \Delta p(0) \simeq \frac{\hbar}{2\Delta q(0)}$$

2. 

Free evolution for time τ

3.

2nd measurement with accuracy

$$\Delta q^2(\tau) = \Delta q^2(0) + \frac{\Delta p^2(0)}{m^2}\tau^2$$

$$\geqslant 2\,\underset{\substack{\uparrow \\ \text{sensing error}}}{\Delta q(0)}\;\underset{\substack{\uparrow \\ \text{back reaction error}}}{\Delta p(0)}\,\frac{\tau}{m}$$

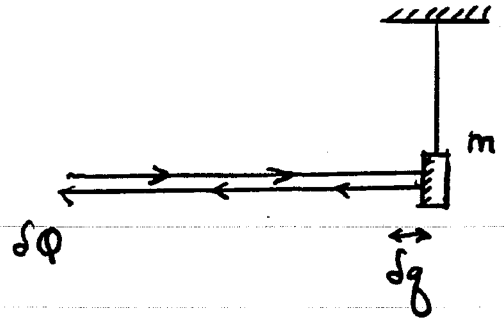
$$\geqslant \hbar\tau/m$$

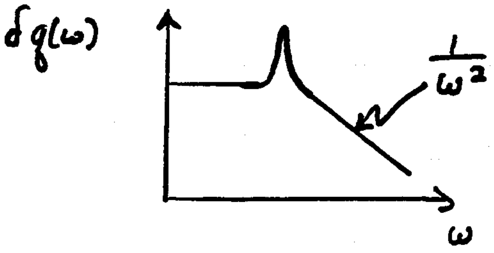
SQL -

$$\boxed{\Delta q_{SQL} \simeq \sqrt{\hbar\tau/m}}$$

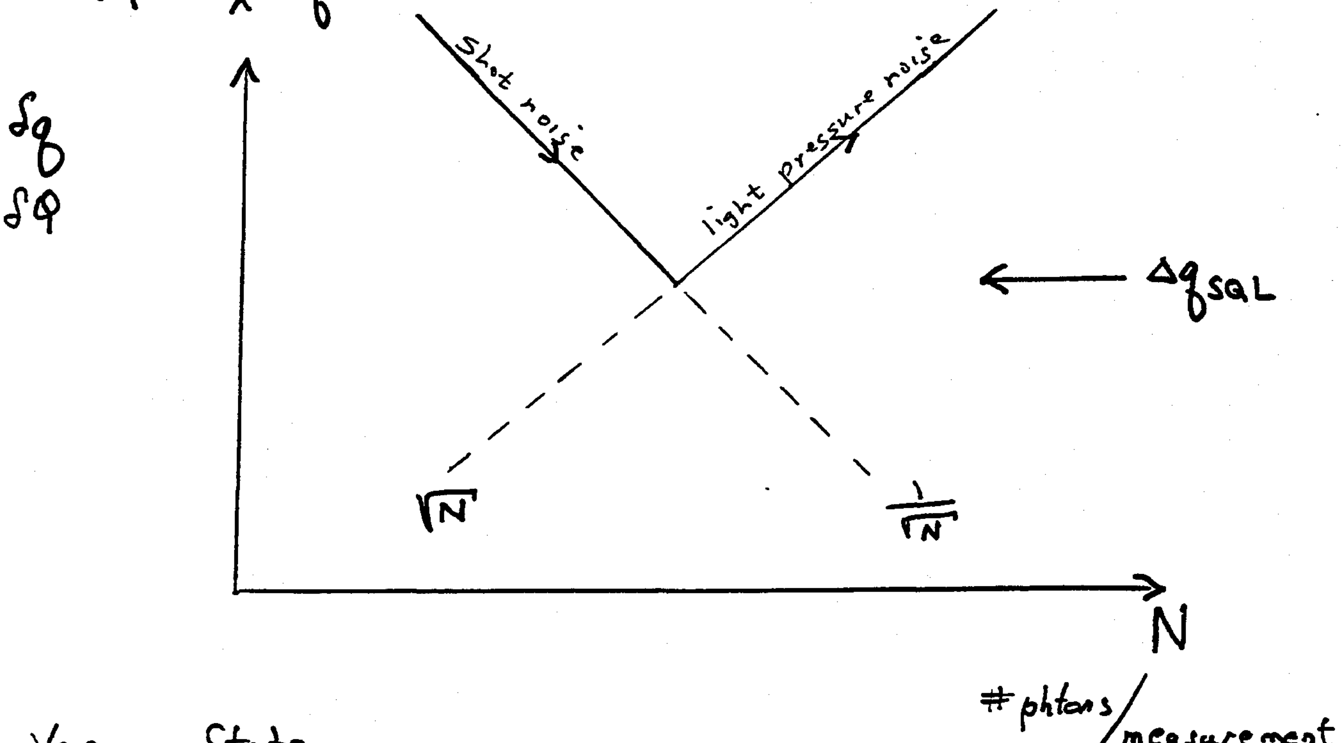
Expressed Explicitly for Interferometry

"Free" Mass

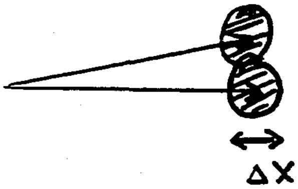




$$\delta\phi = \frac{4\pi}{\lambda}\,\delta q$$



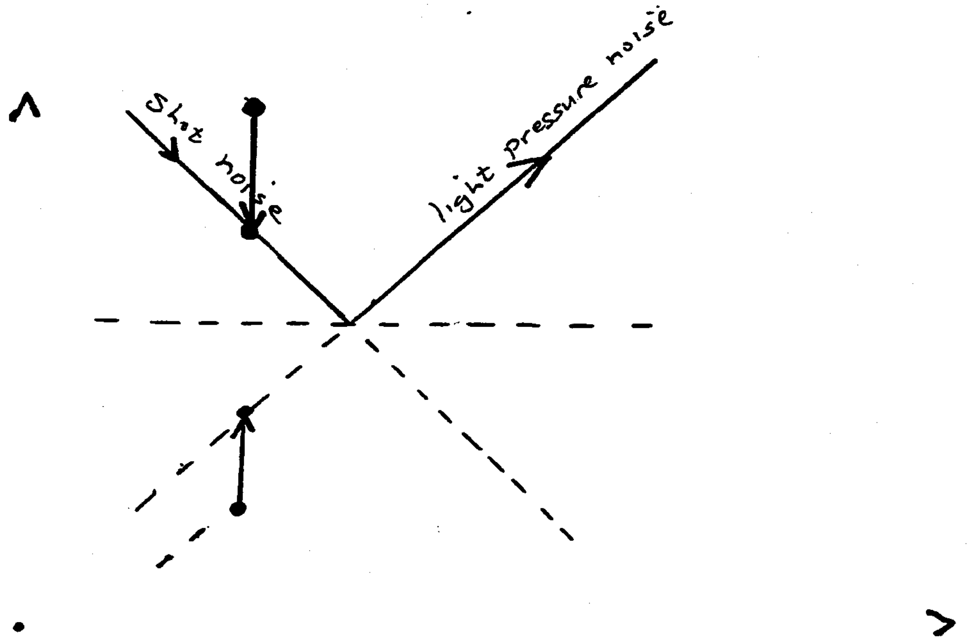
Vacuum State -



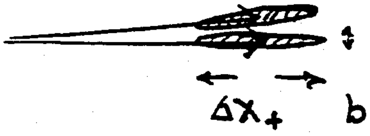
↕ ΔX_- sensing $\propto \frac{1}{\sqrt{N}} \leftrightarrow \Delta q(0)$

↔ ΔX_+ back reaction $\propto \sqrt{N} \leftrightarrow \Delta p(\tau)$

(Overlay)



Squeezed State -



good for Δq sensing

bad for Δp back reaction

PHYSICAL REVIEW LETTERS

VOLUME 51 29 AUGUST 1983 NUMBER 9

Contractive States and the Standard Quantum Limit for Monitoring Free-Mass Positions

Horace P. Yuen

Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, Illinois 60201

(Received 27 June 1983)

The familiar minimum-uncertainty wave packets for masses are generalized in analogy with the two-photon coherent states of the radiation field. The free evolution of a subclass of these states, the contractive states, leads to a narrowing of the position uncertainty in contrast with the usual minimum-uncertainty wave packets. As a consequence the standard quantum limit for monitoring the positions of a free mass can be breached. Further implications on quantum nondemolition measurements are discussed.

PACS numbers: 03.65.Bz, 04.80.+z

There has been considerable recent interest in ascertaining and achieving the fundamental quantum limits on signal processing and precision measurements, in particular for applications to optical communications[1-3] and gravitational-wave detection.[4-8] A major result of this work is that one can beat the so-called standard quantum limit for amplitude measurements on harmonic oscillators. However, for the gravitational-wave interferometer[9] it is usually supposed[7,8] that the resolution is limited by the "standard quantum limit" (SQL) for measuring the positions of a free mass.[4-5] In this paper it is shown that the latter SQL is also *not* generally valid; it can be breached by a specific quantum measurement without special preparation of the free-mass quantum state. Toward this end I will describe a class of generalized minimum-uncertainty wave packets for masses, to be called twisted coherent states, which are also of interest in their own right. The breakdown of the SQL for free-mass position measurements demonstrates the fact that back actions from a conjugate observable do *not* necessarily, at least in accordance with the principle of quantum mechanics, limit the accuracy of subsequent measurements on an observable.

The evolution of a free mass is given by $X(t) = X(0) + P(0)t/m$, so that the position fluctuation at time t is

$$\langle \Delta X^2(t)\rangle = \langle \Delta X^2(0)\rangle + \langle \Delta P^2(0)\rangle t^2/m^2 + \langle \Delta X(0)\Delta P(0) + \Delta P(0)\Delta X(0)\rangle t/m. \quad (1)$$

In the previous derivation[4-5] of the general SQL for monitoring free-mass positions, it is implicitly assumed that the $t=0$ state of the mass (or the state after measurement) is such that the last term in (1) either vanishes or is positive. Under this assumption the uncertainty principle can be applied to minimize (1) at any time t with the resulting SQL

$$\langle \Delta X^2(t)\rangle_{\rm SQL} = \hbar t/m. \quad (2)$$

On the other hand, it is clear that $\langle \Delta X^2(t_0)\rangle = 0$ if the initial state is an eigenstate of the self-adjoint operator $X(0)+P(0)t_0/m$. Thus, the last term in (1) can surely be negative and the SQL is not generally valid. However, $\langle \Delta X^2(t)\rangle = 0$ implies $\langle \Delta P^2(t)\rangle = \infty$ so that $\langle P^2(0)/2m\rangle = \langle P^2(t)/2m\rangle = \infty$, i.e., an infinite average energy is needed to produce such a state.[3] A more realistic description can be developed as follows.

For an oscillator of mass m and frequency ω, the twisted or two-photon coherent states (TCS)[1,3]

© 1983 The American Physical Society

$|\mu\nu\alpha\rangle$ are the eigenstates of $\mu a+\nu a^\dagger$:

$$(\mu a+\nu a^\dagger)|\mu\nu\alpha\rangle=(\mu\alpha+\nu\alpha^*)|\mu\nu\alpha\rangle,$$
$$|\mu|^2-|\nu|^2=1, \tag{3}$$

where a is the annihilation operator of the oscillator mode. Here we adopt them to yield a class of states for a mass m with position X and momentum P. Define the following operator a on the Hilbert space of states for the mass:

$$a\equiv X(m\omega/2\hbar)^{1/2}+iP/(2\hbar m\omega)^{1/2},\quad [a,a^\dagger]=I, \tag{4}$$

where ω is now an *arbitrary* parameter with unit sec^{-1}. The *twisted coherent states* (TCS) $|\mu\nu\alpha\omega\rangle$ of a mass are defined to be the eigenstates of $\mu a+\nu a^\dagger$, $|\mu|^2-|\nu|^2=1$, in analogy with (3) but with a given by (4). The free-mass Hamiltonian can be expressed

$$H=P^2/2m=\tfrac{1}{2}\hbar\omega(a^\dagger a-\tfrac{1}{2}a^2-\tfrac{1}{2}a^{\dagger 2}+\tfrac{1}{2}). \tag{5}$$

The wave function $\langle x|\mu\nu\alpha\omega\rangle$, $X|x\rangle=x|x\rangle$, can be found through Eq. (3.24) of Ref. 1. Within the choice of a constant phase it is given by

$$\langle x|\mu\nu\alpha\omega\rangle=\left[\frac{m\omega}{\pi\hbar|\mu-\nu|^2}\right]^{1/4}\exp\left\{-\frac{m\omega}{2\hbar}\frac{1+i\xi}{|\mu-\nu|^2}\left[x-\left(\frac{2\hbar}{m\omega}\right)^{1/2}\alpha_1\right]^2+i\left(\frac{2m\omega}{\hbar}\right)^{1/2}\alpha_2\left[x-\left(\frac{2\hbar}{m\omega}\right)^{1/2}\alpha_1\right]\right\}, \tag{6}$$

where

$$\xi\equiv\mathrm{Im}(\mu^*\nu);\quad \alpha\equiv\alpha_1+i\alpha_2,\quad \alpha_1,\alpha_2\ \text{real}. \tag{7}$$

The wave functions (6) constitute a generalization of the usual minimum-uncertainty wave packets treated in every quantum mechanics textbook, which are given by (6) with $\xi=0$. In the context of oscillators, "squeezing" is obtained when $\nu\neq 0$ in $|\mu\nu\alpha\rangle$, and ξ is related to the direction of minimum squeezing. As will be seen in the following, when $\xi>0$ the x-dependent phase in (6) leads to a narrowing of $\langle\Delta X^2(t)\rangle$ from $\langle\Delta X^2(0)\rangle$ during free evolution, in direct contrast with the well-known spreading of $\langle\Delta X^2(t)\rangle$ for minimum-uncertainty wave packets.[10] Because of this behavior, mass states (6) with $\xi>0$ will be called *contractive states*.

The first two moments of (6) are

$$\langle X\rangle\equiv\langle\mu\nu\alpha\omega|X|\mu\nu\alpha\omega\rangle=(2\hbar/m\omega)^{1/2}\alpha_1,\quad \langle P\rangle=(2\hbar m\omega)^{1/2}\alpha_2, \tag{8}$$
$$\langle\Delta X^2\rangle\equiv\langle(X-\langle X\rangle)^2\rangle=2\hbar\zeta/m\omega,\quad \langle\Delta P^2\rangle=2\hbar m\omega\eta, \tag{9}$$
$$\zeta\equiv|\mu-\nu|^2/4,\quad \eta\equiv|\mu+\nu|^2/4;\quad \zeta\eta=(1+4\xi^2)/16, \tag{10}$$
$$\langle\Delta X\Delta P\rangle=i\hbar/2-\xi\hbar,\quad \langle\Delta P\Delta X\rangle=-i\hbar/2-\xi\hbar, \tag{11}$$
$$\langle P^2/2m\rangle=\hbar\omega(\alpha_2{}^2+\eta). \tag{12}$$

The average mass energy (12) is finite when ω, α_2, and $|\nu|$ are finite. From (9)–(10) it follows that the minimum-uncertainty product $\langle\Delta X^2\rangle\langle\Delta P^2\rangle=\hbar^2/4$ is achieved if and only if $\xi=0$.

The position fluctuation for a free mass starting in an arbitrary TCS (6) is immediately obtained from (1) and (9)–(11),

$$m\langle\Delta X^2(t)\rangle/2\hbar=\zeta/\omega-\xi t+\eta\omega t^2. \tag{13}$$

If $\xi\leq 0$, $\langle\Delta X^2(t)\rangle$ increases monotonically. In contrast to this usual situation, Eq. (13) is plotted in Fig. 1 for contractive states at $t=0$ (i.e., for $\xi>0$). The minimum fluctuation $1/16\omega\eta$ can be made arbitrarily small even for fixed ω by letting η (and thus also $\langle H\rangle$) become arbitrarily large. The time t_m at this fluctuation level is $t_m=\xi/2\eta\omega$ so that $m\langle\Delta X^2(t_m)\rangle/2\hbar t_m=1/8\xi$. If $\langle\Delta X^2(t)\rangle$ is minimized with respect to ω at any given t simi-

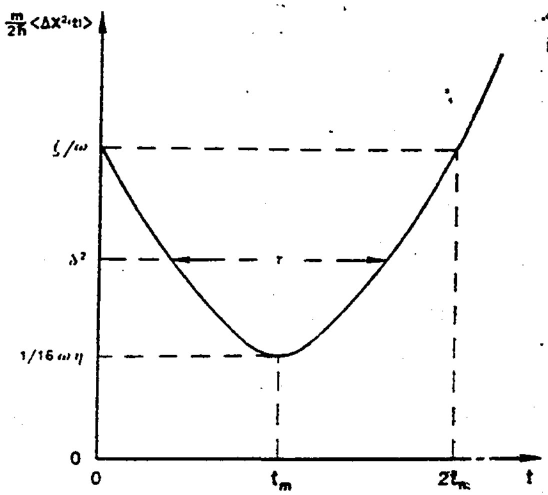


FIG. 1. The position fluctuation of a contractive state from (13); $t_m\equiv\zeta/2\eta\omega$, $t_m\to 0$ when $\zeta\to 0$.

Comment on "Contractive States and the Standard Quantum Limit for Monitoring Free-Mass Positions"

In a recent Letter,[1] Yuen has considered the so-called twisted or two-photon coherent states to show that the free evolution of certain of such states (contractive states) leads to a narrowing of the position uncertainty wave packets. It is the purpose of this Comment to stress that this narrowing property has nothing to do with Yuen's coherent states and has an almost twenty-year-old history. The general criterion for narrowing of the free-motion position uncertainty has been obtained in some of the standard textbooks on quantum mechanics where the following expression has been derived[2]:

$$\langle \Delta\hat{x}^2(t)\rangle = \langle \Delta\hat{x}^2(0)\rangle + (\langle \Delta\hat{p}^2(0)\rangle/m^2)t^2 + 2t\int dx\,[x - \langle\hat{x}(0)\rangle]\,j(x). \qquad (1)$$

In this equation, j is the standard quantum mechanical probability current of the initial wave function. If we assume (without loss of generality) that initially $\langle\hat{x}(0)\rangle = 0$, a narrowing of $\langle \Delta x^2(t)\rangle$ is obtained if and only if

$$\int dx\, x j(x) < 0. \qquad (2)$$

There is, of course, an infinite number of states that satisfy this condition. As an example, the initial wave function

$$\psi(x,0) = f(x)\exp\left(-\frac{i|\lambda_n|}{\hbar}\frac{x^{2n}}{2n}\right) \quad n = 1,2,\dots \qquad (3)$$

with λ_n arbitrary complex numbers and with $f(x)$ a real L^2 normalizable function leads to the narrowing effect. The wave function with $n = 2$ is especially instructive. For any $f(x)$ this wave function, which has a contractive phase similar to Yuen's wave packet, is not a twisted coherent state but nevertheless leads to a narrowing effect.

A general discussion on how to realize experimentally the initial wave function (3) and how to obtain the narrowing effect (including the one given by Yuen) was presented by Lamb in 1969.[3]

This research has been partially supported by the U. S. Department of Energy and the U. S. Office of Naval Research.

K. Wódkiewicz

Department of Physics and Astronomy
University of Rochester, Rochester, New York 14627,
and Institute of Theoretical Physics
Warsaw University, Warsaw 00-681, Poland[a]

Received 07 November 1983
PACS numbers: 03.65.Bz, 04.80.+z

[a] Permanent address.

[1] H. P. Yuen, Phys. Rev. Lett. 51, 719 (1983).

[2] See, for example, K. Gottfried, *Quantum Mechanics* (Benjamin, New York, 1966), p. 27, Eq. (41).

[3] W. E. Lamb, Jr., Phys. Today 22, No. 4, 23 (1969). See pp. 25 and 26 for details.

© 1984 The American Physical Society

Yuen Responds: Contractive twisted coherent states (TCS) comprise the first explicit class of states that was shown to lead to a narrowing of the free-mass position fluctuation $\langle \Delta X^2(t) \rangle$, as far as I know. It makes little sense to say that they have nothing to do with such narrowing. Nowhere in my paper is it stated or implied that these states are the only ones leading to such narrowing, or that they are somehow essential for that purpose. In fact, when I first mentioned the possibility of such narrowing I used the eigenstate of the self-adjoint operator $X+Pt/m$ as an example, which is strictly speaking not a TCS. Among all the possible states that exhibit such narrowing, contractive TCS form a natural generalization of the usual minimum-uncertainty wave packets. In addition, the time duration of their contraction and the associated $\langle \Delta X^2(t) \rangle$ can be conveniently parametrized. Calling such states "contractive states" when other states may also contract is like calling TCS "squeezed states" when they are not the only states that exhibit squeezing.

A main objective of my paper is to give a measurement scheme that can directly monitor the positions of a free mass in a quantum-nondemolitional way. For this purpose, the states which contract are to be the ones in which a free mass would be left after a certain measurement, without additional intervention. For a discussion of this point see Caves in Ref. 8 of my paper. For contractive TCS such a measurement is the one described by $|\mu\nu\alpha\omega\rangle\langle\mu\nu\alpha\omega|$ as discussed in my paper; the possible realization of this measurement I merely stated without proof because of space limitation. Thus, contractive TCS turn out to be essential in my quantum-nondemolitional position measurement scheme. Dr. Wodkiewicz did not give a measurement which would leave the free mass in his more general contractive states. On the other hand, I would be very surprised if TCS are essential in all possible quantum-nondemolitional position measurements.

Horace P. Yuen
Department of Electrical Engineering
and Computer Science
Northwestern University
Evanston, Illinois 60201

Received 21 November 1983
PACS numbers: 03.65.Bz, 04.80.+z

Comment on "Contractive States and the Standard Quantum Limit for Monitoring Free-Mass Positions"

Recently, Yuen has published a very interesting paper[1] in which he gives a non-QND method for beating the standard quantum limit when measuring the position of a free mass (QND stands for quantum nondemolition). The technique utilizes the so-called two-photon coherent state (TCS).

There is a difficulty with his repeated measurement scheme, however. The reason for this Comment is to call attention to this difficulty, and also to show that TCS can be used to make finite-energy QND-type measurements.

Consider the following recapitulation of Yuen's paper. At $t=0$, an arbitrary free mass state $|\psi\rangle$ is prepared into a TCS, $|\mu\nu\alpha\omega\rangle = |\alpha\rangle$, by interaction of the system with a generalization of the two-meter detector of Arthurs and Kelly,[2] followed by subsequent meter reduction. This measurement can be described in the Gordon and Louisell[3] terminology (which Yuen prefers) by $|\alpha\rangle\langle\alpha|$. It is important to note that the acutal state $|\alpha\rangle$ which obtains after meter reduction is only probabilistically determined, depending on the overlap between $|\alpha\rangle$ and $|\psi\rangle$. Finally, one may also look upon this state preparation as the measurement of the non-self-adjoint operator $A(0)$, where

$$\begin{aligned} A(t) &= A(x(t),p(t)) \\ &= (\mu+\nu)(m\omega/2\hbar)^{1/2}x(t) \\ &\quad + i(\mu-\nu)p(t)/(2\hbar m\omega)^{1/2}. \end{aligned} \qquad (1)$$

Thus at $t=0$ the system is found in some eigenstate $|\alpha\rangle$ of $A(0)$ with eigenvalue $\alpha=\alpha_1+i\alpha_2$.

As shown by Yuen, from the measured eigenvalue one can read off the mass's position $\langle x(0)\rangle = (2\hbar/m\omega)^{1/2}\alpha_1$ and momentum $\langle p(0)\rangle = (2\hbar m\omega)^{1/2}\alpha_2$, with uncertainties $\langle \Delta x^2(0)\rangle = 2\hbar\xi/m\omega$ and $\langle\Delta p^2(0)\rangle = 2\hbar m\omega\eta$, where ξ and η are functions of μ, ν. He also shows that as t goes from 0 to a time $2t_m$, the position uncertainty $\langle\Delta x^2(t)\rangle$ first decreases, and then increases to its $t=0$ value. So far, no difficulties.

However, at $t=2t_m$ Yuen calls for another measurement on the system, presumably of $A(2t_m)$. But we should not describe this measurement by $|\alpha\rangle\langle\alpha|$ as Yuen does, but as $|\alpha'\rangle\langle\alpha'|$, assuming in general that $\alpha\neq\alpha'$. This assumption is correct since one finds $[A(0),A(2t_m)] = i(\mu+\nu)^2\omega t_m \neq 0$. One can easily show by calculating $\langle\alpha|A(2t_m)|\alpha\rangle$ and the nonzero $\langle\alpha|\Delta A(2t_m)|\alpha\rangle$ that the system will "jump" to a range of states centered about $\alpha_1' = \alpha_1 + (m\omega/2\hbar)^{1/2}\langle p(0)\rangle 2t_m/m$, $\alpha_2' = \alpha_2$. The width of the range of states and the magnitude of the effect that it has on Yuen's proposal are difficult to calculate because of the non-self-adjoint nature of $A(2t_m)$. Nevertheless this "back-action" mechanism will contaminate the measurements, and has not been included in Yuen's scheme. A complete evaluation of the extent of this difficulty will presumably involve a lengthy calculation of the "meter-interaction-and-reduction" type pioneered by Caves.

It is interesting to note that one could achieve a continuous QND-like measurement with the operator $A(0)$. This is because the operators $x(0)=x(t)-p(t)t/m$ and $p(0)=p(t)$ are separately QND. Measurement of $A(0)$ has the additional desirable property that the resulting state has finite energy, quite properly one of Yuen's motivations for examining TCS in his original proposal. However, with $x(0)$ mixing position and momentum operators, measurement of $A(0)$ could no longer be described as a position measurement, a view also taken of Yuen's scheme by Caves recently.[4]

Robert Lynch
University of Petroleum and Minerals
Dhahran, Saudi Arabia, and
Blackett Laboratory[(a)]
Imperial College
London SW7 2BZ, United Kingdom

Received 30 January 1984
PACS numbers: 03.65.Bz, 04.80.+z

[(a)] Address during 1983–1984 sabbatical leave.

[1] H. P. Yuen, Phys. Rev. Lett. 51, 719 (1983).

[2] E. Arthurs and J. L. Kelly, Jr., Bell Syst. Tech. J. 44, 725 (1965).

[3] J. P. Gordon and W. H. Louisell, in *Physics of Quantum Electronics*, edited by P. L. Kelly *et al.* (McGraw-Hill, New York, 1966), pp. 833–840.

[4] C. M. Caves, "A Defense of the Standard Quantum Limit for Free-Mass Position" (to be published).

© 1984 The American Physical Society

Yuen Reponds: I am grateful to Dr. Lynch for providing me with the opportunity to clarify certain points in connection with the standard quantum limit (SQL), quantum nondemolition (QND), and my paper. Terminology and concepts such as back action, position measurement, QND, etc., are fraught with ambiguity and imprecision, both in the QND literature and in my own paper. Before they are definitively cleared up, it is important to attend to the actual content of a result rather than its verbal representation.

Using the notations of my paper,[1] I would like to first describe more accurately my principal results as they relate to SQL and QND: (1) The previous derivation of the SQL for monitoring free-mass positions is not generally valid. In particular, a specific realizable measurement described by $|\mu\nu\alpha\omega\rangle \times \langle\mu\nu\alpha\omega|$ could leave the free mass in a contractive two-photon coherent state (TCS). The SQL can then be broken to an arbitrary degree in a second sufficiently accurate position measurement. Note that the SQL would *not* be broken by the same $|\mu\nu\alpha\omega\rangle\langle\mu\nu\alpha\omega|$ measurement. As explained later, another one with $|\Psi^M\rangle$ having a smaller associated position fluctuation $\langle\Psi^M|\Delta X^2|\Psi^M\rangle$ is required, for example, $|\mu'\nu'\alpha\omega\rangle\langle\mu'\nu'\alpha\omega|$ with $|\mu'-\nu'|^2 < |\mu-\nu|^2$ performed at $t \sim t_m$. (2) The $|\mu\nu\alpha\omega\rangle \times \langle\mu\nu\alpha\omega|$ measurement can be used to monitor the free-mass positions in an arbitrary sequence of measurement, with a limitation $\hbar/m$ on the ratio of the position resolution to the time lapse between two measurements. No such limitation exists for the measurement $|\mu\nu\alpha\omega\rangle\langle\mu'\nu'\alpha\omega'|$, of which no realization is known, however. If such measurement is indeed realizable, there can be no quantum limit of any kind on position monitoring.

Dr. Lynch's main point[2] appears to result from a combination of both his confusion about approximate simultaneous measurements and my overly condensed presentation as a Letter. [There are also a number of misprints in my paper, some of which have been corrected in the erratum.[1] I have since found four more: "this work" should read "these works" in the second sentence of the paper; $1+i\xi$ should read $1+i2\xi$ in Eq. (6); $\zeta/2\eta\omega$ should read $\xi/2\eta\omega$ in the figure caption; and $\hbar\xi/m\omega$ should read $4\hbar\xi/\eta\omega$ in line 12 of the second column of p. 721.] The measurement described by $|\mu\nu\alpha\omega\rangle \times \langle\mu\nu\alpha\omega|$, the $A(0)$ measurement in Lynch's terminology, is an approximate simultaneous measurement of position and momentum: α being a variable whose real and imaginary parts provide the *measurement readings* corresponding to the position and momentum estimates. The state after a measurement with reading α is just $|\mu\nu\alpha\omega\rangle$; it is no probabilistically determined. In the case of point (2) above corresponding to that discussed by Lynch, the same $|\mu\nu\alpha\omega\rangle\langle\mu\nu\alpha\omega|$ measurement is made at $t=0$ and $t=2t_m$ while the mass state has evolved. It is important to note that $\langle\Delta X^2\rangle$ gives only the state contribution to the position fluctuation in a measurement; it is the fluctuation observed in a perfect or "exact position measurement." In an "approximate measurement" there would be additional fluctuation from the measurement itself. In my paper, both the state and measurement contributions to the position fluctuation have been included through the resolution factor $4\hbar\xi/m\omega$ instead of $2\hbar\xi/m\omega$. This resolution value is obtained from the probability $|\langle\mu\nu\alpha'\omega|\mu\nu\alpha\omega\rangle|^2$; it makes no sense to set $\alpha'=\alpha$. [It turns out that this doubling of the position uncertainty exactly cancels out the factor of 2 advantage of Eq. (15) compared to the SQL. This explains why a second measurement with $|\psi^M\rangle$ having lower $\langle\Delta X^2\rangle$ is required for bleaching the SQL.] The momentum reading needs never be made; it has no effect on the position fluctuation during the sequence of measurements. Thus, whatever "back action" there is has already been accounted for.

The $|\mu\nu\alpha\omega\rangle\langle\mu\nu\alpha\omega|$ measurement without the α_2 reading is emphatically a position measurement on all grounds: physical, formal, and the purpose of such measurements. The α_1 reading indicates the free-mass position before and after measurements within prescribed uncertainties. It is mathematically equivalent to an exact position measurement in the presence of meter-reading fluctuation, as far as the measurement probability is concerned. There is no reason to call an expression a "quantum limit" if it does not cover this kind of approximate position measurements which serve the purpose of monitoring the mass positions. The SQL is meant to apply to *all* conceivable measurements.

Horace P. Yuen
Department of Electrical Engineering
and Computer Science
Northwestern University
Evanston, Illinois 60201

Received 21 February 1984
PACS numbers: 03.65.Bz, 04.80.+z

[1] H. P. Yuen, Phys. Rev. Lett. 51, 719, 1603(E) (1983).

[2] R. Lynch, preceding Comment [Phys. Rev. Lett. 52, 1729 (1984)].

PHYSICAL REVIEW
LETTERS

VOLUME 54 15 APRIL 1985 NUMBER 15

Repeated Contractive-State Position Measurements and the Standard Quantum Limit

Robert Lynch
Physics Department, University of Petroleum & Minerals, Dhahran, Saudi Arabia

(Received 23 April 1984)

It is shown that if the "standard quantum limit" is taken in a predictive sense, then a repeated measurement scheme involving contractive states, recently proposed by Yuen, does not break this limit.

PACS numbers: 03.65.Bz

Recently Yuen[1] has proposed a scheme to beat the "standard quantum limit" (SQL) on free-mass position monitoring by means of contractive states. There have been several unpublished responses[2] to Yuen's proposal which seek to defend the SQL on general grounds.

A Comment[3] I wrote takes a different view. A qualitative point made was that even assuming the validity of the framework adopted by Yuen, he has not fully considered the impact of measurements in his scheme. It is the purpose of this paper to flesh out the arguments of that Comment, and to show that if the SQL is taken in a predictive sense, Yuen's proposal fails to beat this limit precisely because of such measurement corrections.

Before turning to the detailed discussion it is worthwhile reviewing the reasoning which leads to the SQL. Suppose at $t=0$ one places a free mass approximately at the origin, with the intent of measuring its subsequent position in time (to see if a weak force is acting on it, for example.) How often, and how closely, should one monitor the particle's position? If it is decided to make a measurement every t seconds, one must make t short enough to counter any possible spreading of the wave function. On the other hand, each measurement to precision Δx produces a variance of momentum Δp (by means of the uncertainty principle), which then feeds back into the uncertainty of the position, $\Delta x(t)$, at the time of the next measurement. An analysis[4] of this "back action" leads to the SQL, $\Delta x(t) \geqslant (\hbar t/m)^{1/2}$.

Yuen seeks to beat the SQL by means of "contractive states," and a measurement formalism based on the work of Gordon and Louisell.[5] The contractive states are the so-called "two-photon coherent states" (TCS), $|\mu\nu\alpha\omega\rangle$. For a full discussion of these states the reader is referred to Yuen's original paper[1] and the references therein. Here I simply recall that this state may be taken to represent a free particle of mass m, whose expectation values of position and momentum are $\langle x\rangle=(2\hbar/m\omega)^{1/2}\operatorname{Re}(\alpha)$, $\langle p\rangle=(2\hbar m\omega)^{1/2}\times\operatorname{Im}(\alpha)$, with variances $\langle\Delta x^2\rangle=2\hbar\zeta/m\omega$, $\langle\Delta p^2\rangle=2\hbar m\omega\eta$. Here ω is an arbitrary parameter, and $\zeta=|\mu-\nu|^2/4$, $\eta=|\mu+\nu|^2/4$, subject to $|\mu|^2-|\nu|^2=1$.[6]

As Yuen has shown, for values of the parameter $\xi\equiv\operatorname{Im}(\mu^*\nu)>0$, the $|\mu\nu\alpha\omega\rangle$ states are *contractive*, that is, the initial position variance $\langle\Delta x^2\rangle$ narrows under free evolution for a time $t_m\equiv\xi/2\eta\omega$. This result is cleverly exploited by Yuen to avoid spreading of the wave function. The idea then is to make a sharp position measurement at time $t=t_m$ when the position uncertainty is a minimum, while leaving the system in a $|\mu\nu\alpha\omega\rangle$ state after the measurement, ready to undergo another contraction.

In the Gordon-Louisell terminology such a measurement is described by the projection operator, $|\mu\nu\alpha\omega\rangle\langle\mu'\nu'\alpha\omega'|$.[7] This notation is somewhat abstract—in fact, it is not clear that such measurements are physically possible, in the sense of a Hamiltonian realization, for example. Nevertheless, if one assumes the existence such measurements, and if one considers a system initially in the state $|\phi\rangle$, then according to the Gordon-Louisell theory a $|\mu\nu\alpha\omega\rangle\langle\mu'\nu'\alpha\omega'|$ measurement yields the value α' and the corresponding state $|\mu\nu\alpha'\omega\rangle$ after measure-

© 1985 The American Physical Society

J. Kimble

PHYSICAL REVIEW
LETTERS

VOLUME 54 | 10 JUNE 1985 | NUMBER 23

Defense of the Standard Quantum Limit for Free-Mass Position

Carlton M. Caves
Theoretical Astrophysics, California Institute of Technology, Pasadena, California 91125

(Received 6 April 1984)

Measurements of the position x of a free mass m are thought to be governed by the standard quantum limit (SQL). In two successive measurements of x spaced a time τ apart, the result of the second measurement cannot be predicted with uncertainty smaller than $(\hbar\tau/m)^{1/2}$. Yuen has suggested that there might be ways to beat the SQL. Here I give an improved formulation of the SQL, and I argue for, but do not prove, its validity.

PACS numbers: 03.65.Bz, 06.20.Dk

Conventional wisdom[1,2] holds that in two successive measurements of the position x of a free mass m, the result of the second measurement cannot be predicted with uncertainty smaller than $(\hbar\tau/m)^{1/2}$, where τ is the time between measurements. This limit is called the *standard quantum limit* (SQL) *for monitoring the position of a free mass.*

The standard "textbook" argument for the SQL runs as follows. Suppose that the first measurement of x at $t=0$ leaves the free mass with position uncertainty $\Delta x(0)$. This first measurement disturbs the momentum p and leaves a momentum uncertainty $\Delta p(0) \geq \hbar/2\Delta x(0)$. By the time τ of the second measurement the variance of x (squared uncertainty) increases to

$$(\Delta x)^2(\tau) = (\Delta x)^2(0) + [(\Delta p)^2(0)/m^2]\tau^2 \geq 2\Delta x(0)\Delta p(0)\tau/m \geq \hbar\tau/m. \tag{1}$$

The standard argument views the SQL as a straightforward consequence of the position-momentum uncertainty principle $\Delta x(0)\Delta p(0) \geq \frac{1}{2}\hbar$.

Yuen[3] has pointed out a serious flaw in the standard argument. Between the two measurements the free mass undergoes unitary evolution. In the Heisenberg picture the position operator $\hat{x}$ evolves as

$$\hat{x}(t) = \hat{x}(0) + \hat{p}(0)t/m. \tag{2}$$

Thus the variance of x at time τ is given not by Eq. (1), but by

$$(\Delta x)^2(\tau) = (\Delta x)^2(0) + \frac{(\Delta p)^2(0)}{m^2}\tau^2 + \frac{\langle \hat{x}(0)\hat{p}(0) + \hat{p}(0)\hat{x}(0)\rangle - 2\langle \hat{x}(0)\rangle\langle \hat{p}(0)\rangle}{m}\tau. \tag{3}$$

The standard argument assumes implicitly that the last term in Eq. (3) is zero or positive. Yuen's point[3] is that some measurements of x leave the free mass in a state for which this term is negative. He calls such states *contractive states* because the variance of x decreases with time, at least for a while. As a result, the uncertainty $\Delta x(\tau)$ can be smaller than the SQL. Yuen[3,4] concludes that there are measurements of x that beat the SQL. My conclusion is different: The flaw lies in the standard argument, not in the SQL. In this Letter I give a new, heuristic argument for the SQL, formulate an improved statement of the SQL, and analyze a measurement model that supports the heuristic argument.

© 1985 The American Physical Society

J. Kuroe

PHYSICAL REVIEW
LETTERS

VOLUME 60 1 FEBRUARY 1988 NUMBER 5

Measurement Breaking the Standard Quantum Limit for Free-Mass Position

Masanao Ozawa
Department of Mathematics, College of General Education, Nagoya University, Nagoya 464, Japan

(Received 2 July 1987)

An explicit interaction-Hamiltonian realization of a measurement of the free-mass position with the following properties is given: (1) The probability distribution of the readouts is exactly the same as the free-mass position distribution just before the measurement. (2) The measurement leaves the free mass in a contractive state just after the measurement. It is shown that this measurement breaks the standard quantum limit for the free-mass position in the sense sharpened by the recent controversy.

PACS numbers: 03.65.Bz, 04.80.+z

For monitoring the position of a free mass such as the gravitational-wave interferometer,[1] it is usually supposed[2,3] that the predictability of the results is limited by the so-called standard quantum limit (SQL). In the recent controversy,[4-8] started with Yuen's proposal[4] of a measurement which beats the SQL, the meaning of the SQL has been much clarified and yet no one has given a general proof nor a counterexample for the SQL. Recently, Ni[9] succeeded in constructing a repeated-measurement scheme to monitor the free-mass position to an arbitrary accuracy. However, it is open whether this scheme beats the SQL in the sense sharpened by the recent controversy. In particular, the following problem remains open: Can we realize a high-precision measurement which leaves the free mass in a contractive state?

In the present paper, I shall give a model of measurement of a free-mass position which breaks the SQL in its most serious formulation. An explicit form of the system-meter interaction Hamiltonian will be given and it will be shown that if the meter is prepared in an appropriate contractive state[4] then the measurement leaves the free mass in a contractive state and the uncertainty of the prediction for the next identical measurement decreases in a given duration to a desired extent. Thus Yuen's original proposal[4] is fully realized. This result will open a new way to an arbitrarily accurate non-quantum-nondemolition monitoring for gravitational wave detection and other related fields such as optical communications.

The precise formulation of the SQL is given by Caves[8] as follows: Let a free mass m undergo unitary evolution during the time τ between two measurements of its position x, made with identical measuring apparatus; the result of the second measurement cannot be predicted with uncertainty smaller than $(\hbar\tau/m)^{1/2}$ in average over all the first readout values. Caves[8] showed that the SQL holds for a specific model of a position measurement due to von Neumann[10] and he also gave the following heuristic argument for the validity of the SQL. His point is the notion of the imperfect resolution σ of one's measuring apparatus. His argument runs as follows: *The first assumption* is that the variance of the measurement of x is the sum of σ^2 and the variance of x at the time of the measurement; this is the case when the measuring apparatus is coupled linearly to x. *The second assumption* is that just after the first measurement, the free mass has position uncertainty $\Delta x(0) \leq \sigma$. Under these conditions, he derived the SQL from the uncertainty relation $\Delta x(0)\Delta x(\tau) \geq \hbar\tau/2m$.

However, his definition of the resolution of a measurement is ambiguous. In fact, he used three different definitions in his paper: (1) the uncertainty in the result, (2) the position uncertainty after the measurement, and (3) the uncertainty of the meter before the measurement. These three notions are essentially different, although they are the same for von Neumann's model. I

© 1988 The American Physical Society

Does a Conservation Law Limit Position Measurements?

Masanao Ozawa[a]

Lyman Laboratory of Physics, Harvard University, Cambridge, Massachusetts 02138

(Received 5 March 1990)

The demonstrations of Wigner and others, that observables which do not commute with additive conserved quantities cannot be measured precisely, are reexamined. A proposed new formulation of the claim is shown to be valid for observables with a continuous spectrum whenever the conserved quantities are bounded. However, a countermodel is constructed, and it suggests that the position can be measured as precisely as the momentum even though the measuring interaction conserves the total linear momentum.

PACS numbers: 03.65.Bz

Recently, there has been considerable interest in the analysis of fundamental quantum limits on measurements of unquantized quantities, such as positions of masses and amplitudes of harmonic oscillators, for applications, in particular, to optical communications [1] and gravitational wave detection [2]. A major achievement in this area is that we have breached the two types of quantum limits posed previously, the so-called standard quantum limit for amplitude measurements on harmonic oscillators [3] and the so-called standard quantum limit for monitoring of free-mass positions [4,5]. However, it has long been claimed by several authors [6–10] that *observables which do not commute with additive conserved quantities cannot be measured precisely*. This limit, which will be called the *conservation-law-induced quantum limit* (CQL) for measurements, implies that the conservation law of the linear momentum limits the accuracy of position measurements. Although implications of the CQL in measurements of the spin components have been examined in detail [9], those in position measurements have not been discussed seriously. An obvious difficulty for discussions about the limits on position measurements lies in the fact that the position observable has a continuous spectrum and that any observable with a continuous spectrum cannot be measured with absolute precision, whether it commutes with the additive conserved quantities or not. Thus, if one would claim the CQL for position measurements in a physically meaningful way, the claim would imply that the accuracy of position measurements has an apparent limitation compared with the accuracy of momentum measurements. In this Letter the validity of the CQL is examined from this point of view and it is shown that the CQL is *not* generally valid for position measurements.

We shall first give a rigorous statement of the CQL. Suppose that an observable (self-adjoint operator) A of a quantum system, called an *object*, represented by a Hilbert space $\mathcal{H}_1$, is actually measurable by a measuring instrument. Then we can describe the interaction between the object and the instrument by quantum mechanics in principle. Let $\mathcal{H}_2$ be the Hilbert space of the instrument system. The interaction is supposed to be turned on during a finite time interval from time $t=0$ to $t=\tau$ and represented by a unitary operator U on $\mathcal{H}_1 \otimes \mathcal{H}_2$. Just after the interaction is turned off the object is separated from the instrument and the observer measures an observable B of the instrument to get the outcome of this measurement. In the Heisenberg picture, we can write $A(0) = A \otimes 1$, $B(0) = 1 \otimes B$, $A(\tau) = U^\dagger (A \otimes 1) U$, and $B(\tau) = U^\dagger (1 \otimes B) U$. By saying that this measurement is an *exact* measurement of the observable A it is meant that the measurement satisfies (i) the statistical formula for the probability distribution of the outcome of a measurement (Ref. [11], pp. 200 and 201) and (ii) the repeatability hypothesis: *If an observable is measured twice in succession in the same individual system, then we get the same value each time* (Ref. [11], p. 335).

When does the measuring interaction U give an exact measurement of A? A simple but general condition sufficient for it is as follows: *There is some self-adjoint operator N, called the noise operator, in $\mathcal{H}_2$ for which U, A, and B satisfy the relations*

$$U^\dagger (1 \otimes B) U = A \otimes 1 + 1 \otimes N , \tag{1}$$

$$U^\dagger (A \otimes 1) U = A \otimes 1 . \tag{2}$$

Let ψ be the initial state of the object and ξ the initial state of the instrument. Note that we can assume without any loss of generality that 0 is in the spectrum of the noise operator N; otherwise, replace B in Eq. (1) by $B - \lambda 1$ for any λ in the spectrum of N. In our formulation, a B measurement at time τ gives the outcome of an A measurement at time 0; i.e., a $B(\tau)$ measurement gives the outcome of an $A(0)$ measurement in the initial Heisenberg state $\psi \otimes \xi$. Then, if we prepare the instrument in the eigenstate of N for the eigenvalue 0, i.e., $N\xi = 0$, it follows easily from Eq. (1) that the outcome of the $B(\tau)$ measurement has the same probability distribution as the $A(0)$ measurement. As to the repeatability hypothesis, the first measurement is the $A(0)$ measurement and the second measurement is an A measurement at the time just after the object system is separated from the first measuring instrument, so that the latter is just an $A(\tau)$ measurement. On the other hand, Eq. (2) assures that the $A(0)$ measurement and the $A(\tau)$ measurement give the same value. Thus a measuring interaction satisfying Eqs. (1) and (2) gives an exact measurement of an observable A. Indeed, the conditions are fulfilled in the conventional approach to measurements of discrete observables by a suitable relabeling of eigenvalues, where the unitary U is given by the relation $U(\varphi_m \otimes \xi_n) = \varphi_m$

© 1991 The American Physical Society

EUROPHYSICS LETTERS 15 October 1990

Europhys. Lett., 13 (4), pp. 301-306 (1990)

Quantum Limits in Interferometric Measurements.

M.-T. JAEKEL (*) and S. REYNAUD (**)

(*) *Laboratoire de Physique Théorique de l'Ecole Normale Supérieure* (§)
24 rue Lhomond, F-75231 Paris Cedex 05

(**) *Laboratoire de Spectroscopie Hertzienne* (§§), *Université Pierre et Marie Curie*
4 place Jussieu, F-75252 Paris Cedex 05

(received 25 June 1990; accepted in final form 7 August 1990)

PACS. 42.50 – Quantum optics.
PACS. 06.30 – Measurement of basic variables.
PACS. 03.65 – Quantum theory; quantum mechanics.

Abstract. – Quantum noise limits the sensitivity of interferometric measurements. It is generally admitted that it leads to an ultimate sensitivity, the «standard quantum limit». Using a semi-classical analysis of quantum noise, we show that a judicious use of squeezed states allows one in principle to push the sensitivity beyond this limit. This general method could be applied to large-scale interferometers for gravitational wave detection.

Quantum noise ultimately limits the sensitivity in interferometric detection of gravitational waves [1-3]. A gravitational wave is detected as a phase difference between the optical lengths of the two arms. It seems accepted that there exists a «standard quantum limit» (SQL), equivalent to an ultimate detectable length variation:

$$(\Delta z)_{\mathrm{SQL}} = \sqrt{\hbar\tau/M}, \tag{1}$$

where M is the mass of the mirrors and τ the measurement time [4]. The SQL can be derived by considering that the positions $z(t)$ and $z(t+\tau)$, which are noncommuting observables, are measured [5]. This interpretation of SQL has given rise to a long controversy [6].

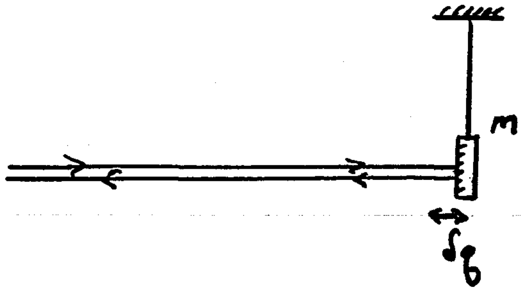
Alternatively, the SQL can be understood by considering the quantum noise as a sum of two contributions. Photon counting noise corresponds to fluctuations of the number of photons detected in the two output ports, while radiation pressure noise stems from the random motion of the mirrors which is sensitive to the fluctuations of the numbers of photons in each arm. The sum of these two contributions leads to an optimal sensitivity given by expression (1). This limit is reached for very large laser power which is not presently achievable.

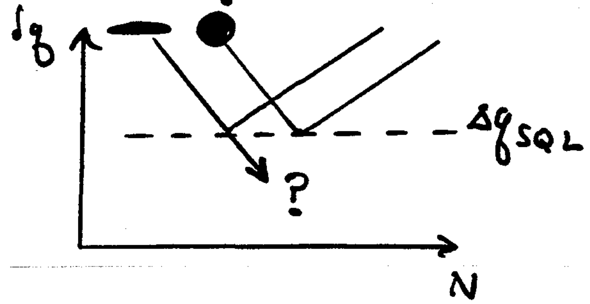
[§] Unité propre du Centre National de la Recherche Scientifique associée à l'Ecole Normale Supérieure et à l'Université de Paris-Sud.

[§§] Unité de l'Ecole Normale Supérieure et de l'Université Pierre et Marie Curie, associée au Centre National de la Recherche Scientifique.

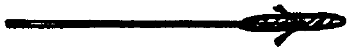
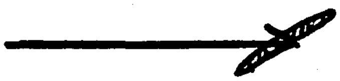
"Practical" Resolution

{ Unruh
Jaekel & Reynaud





At Δq_{SQL} -

- Instead of , try  !

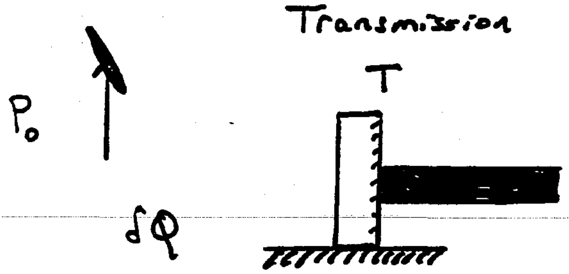
$$\left\langle \begin{pmatrix} \Delta \text{ phase} \\ \Delta X_- \end{pmatrix} \cdot \begin{pmatrix} \Delta \text{ amplitude} \\ \Delta X_+ \end{pmatrix} \right\rangle \rightarrow \text{Correlated}, < 0.$$

- Recall previous "derivation" -

$$\left\langle \left[\Delta q(0) + \frac{\Delta p(0)}{m} \tau \right]^2 \right\rangle \rightarrow \langle \Delta q(0) \Delta p(0) \rangle \begin{matrix} \neq 0 \\ < 0 \end{matrix}$$

- **Prospects ?**

(Overlay)



→ Use a resonant cavity ① $\delta\varphi \simeq \frac{4\pi}{\lambda}\delta q \cdot \frac{1}{T}$

$T = 1.6\times10^{-6}$

$R = 1-T = 0.9999984$

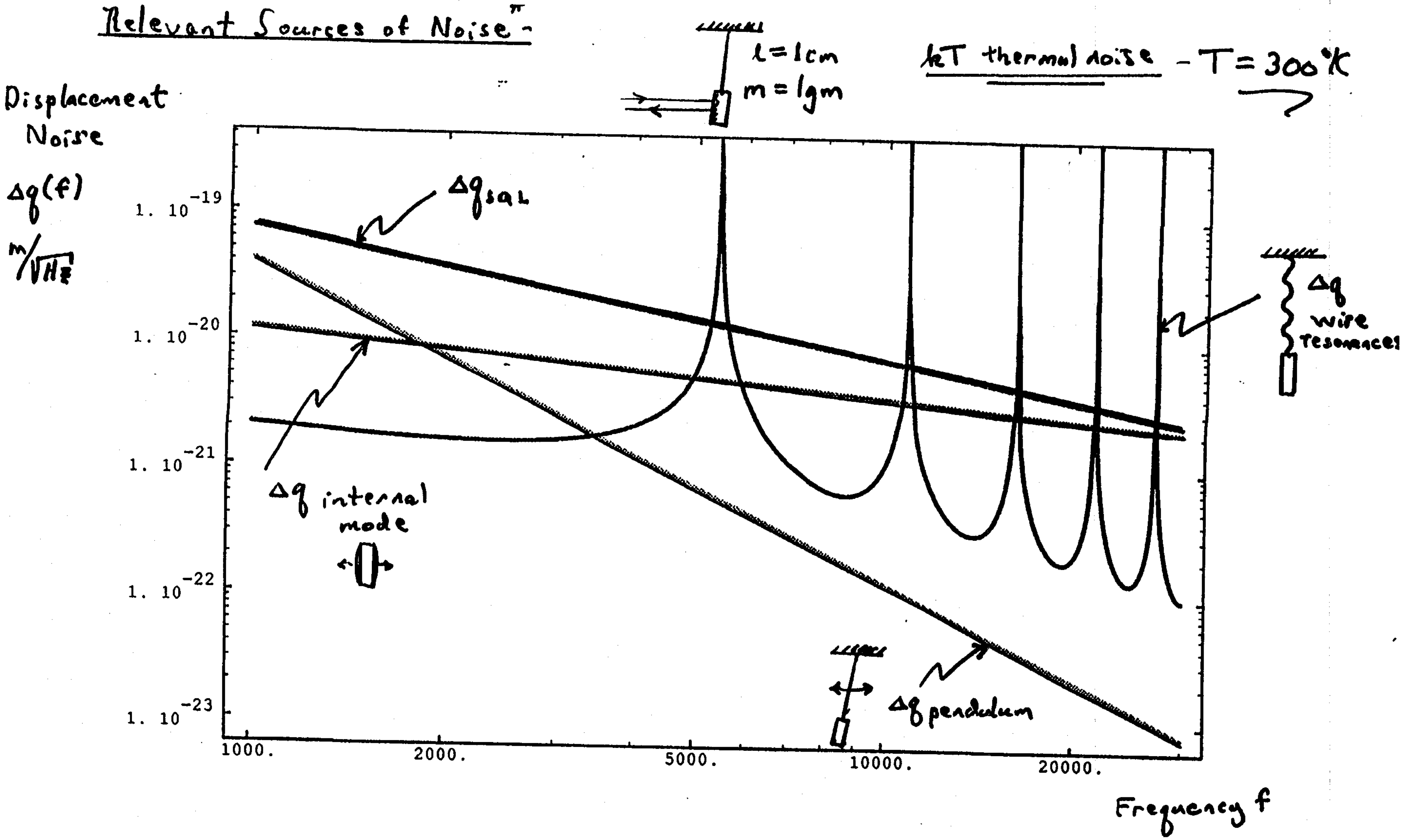
② $P_o \rightarrow P_o T^2$

—o—

Note: For $m = 1\,gm$, $\tau = 10^{-4}\,sec$,

$P_o \sim 10^6\,W$!!

Relevant Sources of Noise*



$Q_{pendulum} = 10^6$ (Braginsky - 10^8)

$Q_{wire} = 10^6$

$Q_{internal} = 10^7$ (" - 10^7 at 1 MHz)

* P. Saulson, Phys. Rev. D42, 2437 (1990).

BATCH START

17. Vacuum System

STAPLE
OR
DIVIDER

Ph 103c: The Physics of LIGO 18 & 20 May 1994

LECTURE 17

The LIGO Vacuum System

Lecture by Jordan Camp

Assigned Reading:

WW. J. H. Moore, C. C. Davis, M. A. Coplan, *Building Scientific Apparatus* (Addison-Wesley, 1983), Chapter 3. "Vacuum Technology." A good overview of the basic issues involved in vacuum system design, including gas kinetics, pressure measurement, and pumping.

Suggested Supplementary Reading:

q. F. Reif, *Fundamentals of Statistical and Thermal Physics* (McGraw-Hill), Chapter 7: "Kinetic Theory of Dilute Gases in Equilibrium." A discussion of basic kinetic theory, including molecular flux, effusion, and pressure and momentum transfer.

r. J. O'Hanlon, *A User's Guide to Vacuum Technology* (John Wiley and Sons). Considerably more detailed than the assigned reading. Includes material vapor pressures and outgassing, calculation of conductances and a chapter on residual gas analyzers.

s. K. Welch, "The pressure profile in a long outgassing vacuum tube", *Vacuum*, **23**, 271–276.

A Few Suggested Problems

1. *Residual gas damping of test mass*: In Lecture 13, the following expression for the losses due to residual gas damping was given:

$$\phi(\omega) \sim \frac{2AP}{M}\sqrt{\frac{\mu}{kT}}\,\frac{\omega}{\omega_o^2}$$

where A, M are the test mass area and mass, P is the residual gas pressure, μ is the mass of a gas molecule, and the gas molecules are thermalized at temperature T. (Recall that $\phi = \gamma\omega/\omega_o^2$, where γv is the acceleration due to gas damping and v is the test mass velocity.) Derive this expression. For simplicity, assume that the gas molecules are of uniform velocity and normally incident on the test mass.

2. *Pressure in LIGO beam tubes:* The final pressure achieved in the LIGO beam tubes will depend on the outgassing rates, conductances and pumping speeds, and available budget.

a. Conductance of an orifice: the ion pumps, which will provide quiet, high vacuum pumping for the beam tubes, will be connected to the tubes through 25 cm diameter orifices. The conductance of an orifice of area A for molecular nitrogen (atomic weight=28) is given by C (in Liter/sec) = 11.6 A (in cm^2). How does this value scale with the molecular mass, and what is the conductance for hydrogen (atomic weight=2)? (Hint: the conductance is linearly related to the flux of molecules across the aperture). Assuming that the ion pumps have pumping speeds of 10000 L/sec, what is the combined pumping speed of the orifice and pump?

b. Conductance of a beam tube: in paper 4 of the suggested reading K. Welch derives the following expression for the average pressure of a long outgassing tube of diameter D and tube length l:

$$P_{av} = P_p + \frac{\pi q l^2}{3kD^2}$$

Here q is the outgassing rate (torr l/(sec-cm^2)) and k is a function of temperature and molecular weight (k=45 for hydrogen at room temp). The first term, P_p, is the pressure at the ion pump, while the second term accounts for the finite conductance of the 1.2 m diameter beam tube.

1) for the special LIGO low-outgassing steel, q ~ 1.0 x 10^{-13} torr l/(sec-cm^2). Assume an initial pumping configuration of 2 end pumps per each 2 km long beam tube module. What is the total outgassing flux (torr l/sec) seen by each of the pumps? Using the earlier calculation of pumping speed, find P_p. What is P_{av}? This number should be close to the goal of 1.0 x 10^{-9} torr for the advanced interferometer.

2) assume that unprocessed steel with a higher outgassing rate is used, where q ~ 1.0 x 10^{-12} torr 1 / (sec cm^2). What is P_{av} for this beam module? How many additional equally spaced pumps would be necessary to recover the desired value of P_{av}? (The 2 pieces of P_{av} scale differently with the # of pumps.) With a cost of $35 K per additional pump station and a total of 8 beam tube modules for the two sites, how much additional cost would be incurred if this steel were used?

Lecture 17
LIGO Vacuum System

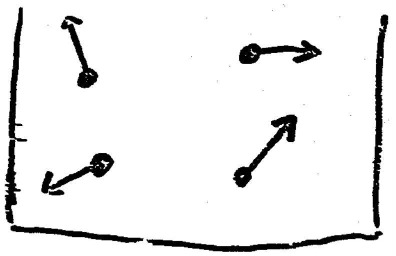
by Jordan Camp, 25 May 1994

Camp lectured from the following transparencies, which Kip has annotated a bit.

Thermal equilibrium:

$$N(v)\, d^3 v$$

$$= n \left(\frac{m}{2\pi kT}\right)^{3/2} e^{-mv^2/2kT}\, d^3 v$$



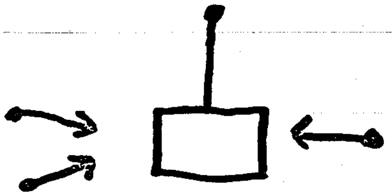
$$F_z = \int d^3 v \underbrace{m V_z}_{\text{momentum transfer}} \underbrace{N(v)\, V_z\, n\, dA}_{\text{Flux}}$$

$$\bar{P} = \frac{F_z}{dA} = n m \overline{V_z^2}$$

$$= \frac{1}{3} n m \overline{V^2}$$

$$= n k T$$

Inteferometer Noise From Gas Pressure : Displacement Noise



1) Acoustic Noise

$\sigma_0 = \pi d^2$

$l \equiv$ mean free path $\Rightarrow$ $n \sigma_0 l = 1$

$n = \dfrac{\bar{P}}{kT}$

$l \simeq 3 \times 10^{-5}$ cm   770 torr (STP)

~ 3 m 10^{-4} torr

2) Residual Gas Damping

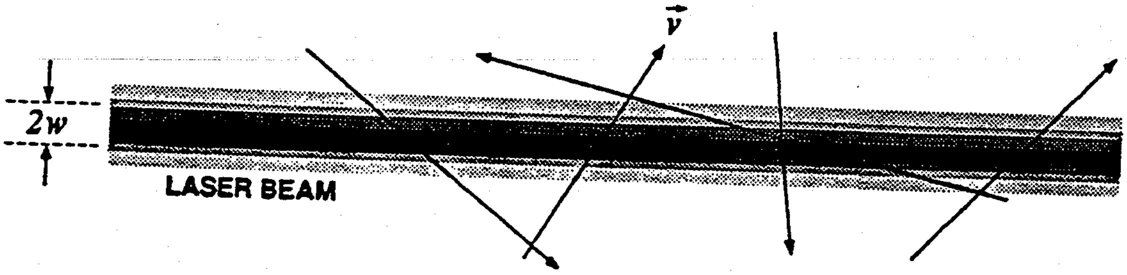
$$\phi(\omega) \sim \overbrace{\frac{2AP}{M}\sqrt{\frac{\mu}{kT}}}^{\gamma} \frac{\omega}{\omega_0^2}$$

$\Rightarrow \tilde{x}(100\ \mathrm{Hz}) \sim 10^{-20}\ \mathrm{m}/\sqrt{\mathrm{Hz}}$ @ 10^{-6} torr

Sensing Noise:

Phase Noise

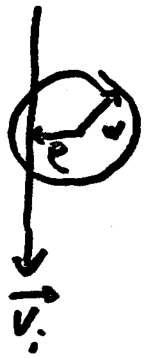
EFFECT OF GAS MOLECULES ON OPTICAL PHASE



$$\delta L_i = C_i e^{-\frac{2r^2}{w^2}}$$

$$= \frac{2\overset{\text{a few} \times 10^{-4} \text{ at STP}}{\overbrace{(n-1)}}}{n_0 \pi w^2} e^{-\frac{2\rho^2 + (t-t_i)^2 V_i^2}{w^2}}$$

$\uparrow$ (number density of molecules of species i)



$$G(f) = \int N(\rho, v) \left| \int \delta L(\rho, v) e^{i 2\pi f t} dt \right|^2 d\rho \, dv \, dz$$

$$\tilde{\Delta L} \sim \sqrt{G} \sim \frac{n_0^{1/2} \alpha}{V_0^{1/2} w} L^{1/2} e^{-2\pi f w / V_0}$$

note: $\frac{\overset{\leftarrow 10^4 \text{ cm/s}}{V_0}}{2\pi W} \sim 10 \text{ kHz}$

$$\sim \frac{p^{1/2}}{v^{1/2}} \alpha L^{1/2}$$

Beam tube partial pressure requirements:

GAS	INITIAL REQUIREMENT (TORR)	GOAL (TORR)
H_2	1×10^{-6}	1×10^{-9}
H_2O	1×10^{-7}	1×10^{-10}
N_2	6×10^{-8}	6×10^{-11}
CO	5×10^{-8}	5×10^{-11}
CO_2	2×10^{-8}	2×10^{-11}
CH_4	3×10^{-8}	3×10^{-11}
Ar	5×10^{-8}	5×10^{-11}
He	1×10^{-5}	1×10^{-8}

H_2 Outgases more easily thanother molecules, and thus is a special problem →

Also, hydrogen is infused into steel in large quantities in the production process.

4) Optical Contamination of mirrors

$$\sum_i P_{\underbrace{HC_i}_{\uparrow}} < 10^{-11} \text{ torr} \Rightarrow$$

Hydrocarbons

(working limit)

mirrors won't degrade faster than a few ppm/few months (rough rule of thumb from current studies)

Sources of Gas

this figure is for a typical unbaked vacuum wall; the numbers thus are not relevant to L160, but the qualitative behavior is relevant

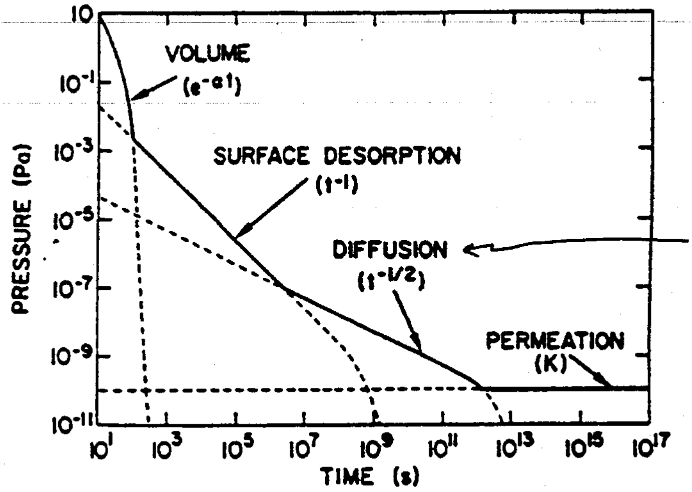


Fig. 4.6 Rate limiting steps during the pumping of a vacuum chamber.

out of bulk metal of vacuum walls

Volume: N_2, O_2, Ar, H_2O, CO_2

Planned L160 bake: ~1 month at 140° (cost of electricity is ~$500k)

Surface: H_2O, $(CH)_N$ ← In-Situ bake

To be baked before putting them into the vacuum

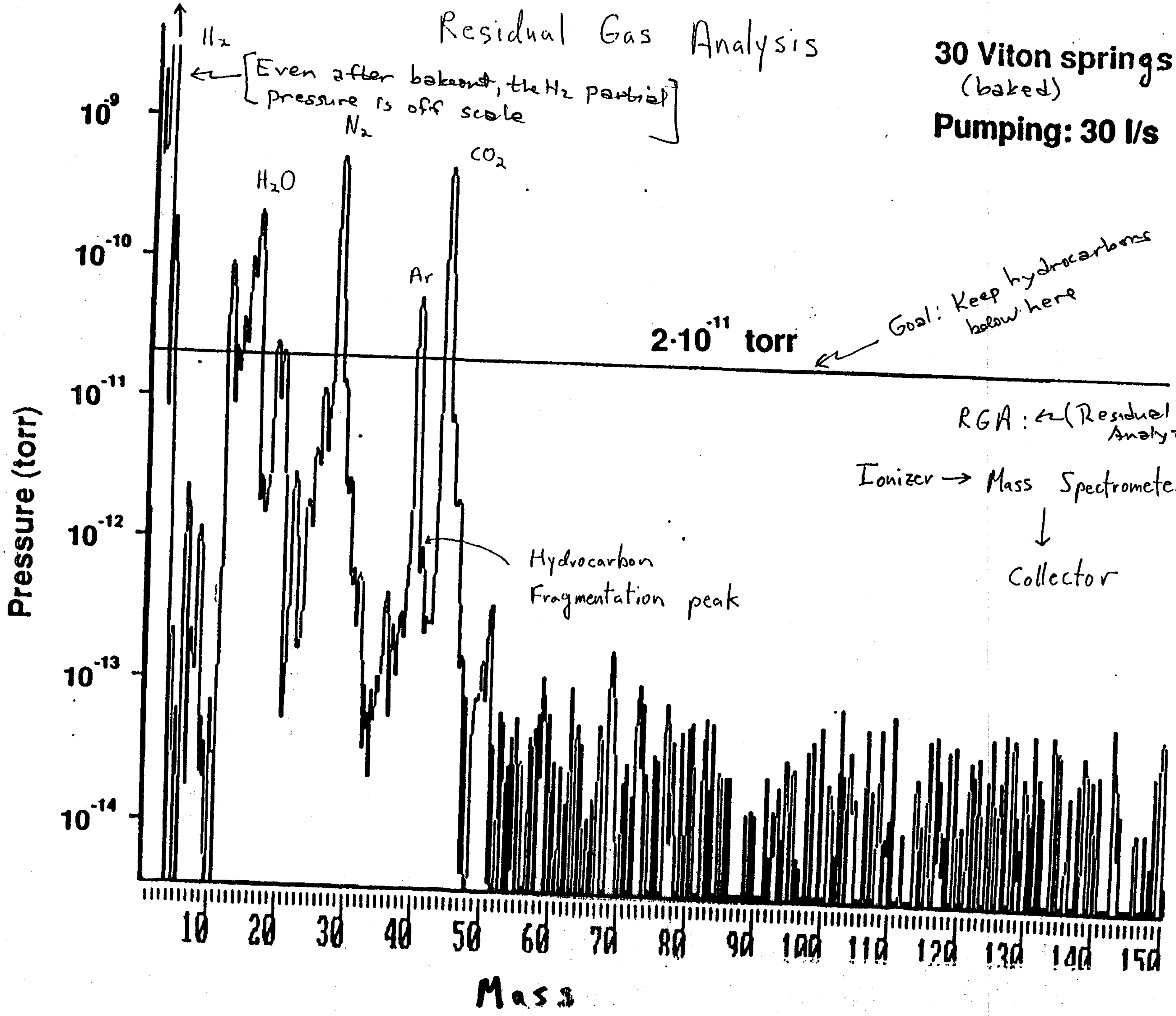
Diffusion: Elastomers : N_2, H_2O, $(CH)_N$

Metal : H_2 ← Pre-bake

Residual Gas Analysis

30 Viton springs
(baked)

Pumping: 30 l/s



← [Even after bakeout, the H_2 partial pressure is off scale]

$2 \cdot 10^{-11}$ torr ← Goal: Keep hydrocarbons below here

Hydrocarbon Fragmentation peak

RGA: ← (Residual Gas Analyzer)

Ionizer → Mass Spectrometer
↓
Collector

Roughing Pump

$10^{-2} < P < \text{atm}$

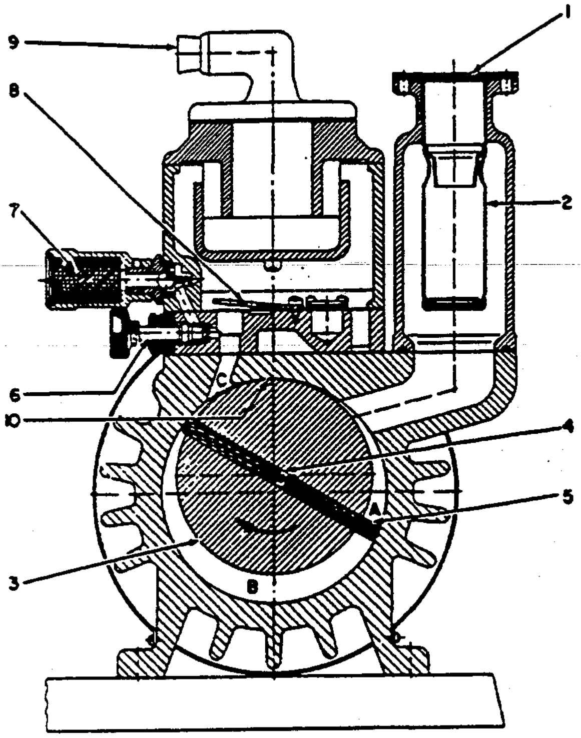


Fig. 10.1 Sectional view of the Pfeiffer DUO-35, 35-m^3/h double-stage, rotary vane pump: (1) intake, (2) filter, (3) rotor, (4) spring, (5) vane, (6) gas ballast valve, (7) filter, (8) discharge valve, (9) exhaust, (10) sealing surface. Reprinted with permission from A. Pfeiffer Vakuumtechnik G.m.b.H., Wetzlar, West Germany.

Ion Pump

$10^{-11} < P < 10^{-5}$ torr

Electrons from cathode ionize gas molecules, which then get swept out electrically

↑ B

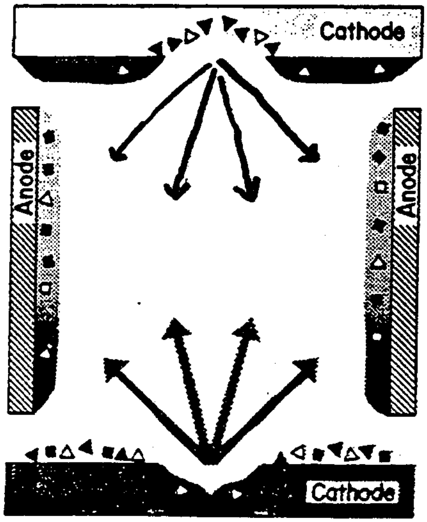
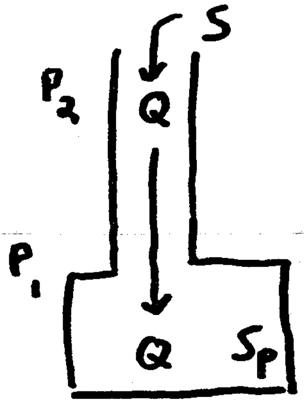


Fig. 14.6 Schematic diagram showing sputter deposition and pumping in a Penning cell: ■ chemically active gases buried as neutral particles; ▶ Chemically active gases ionized before burial; □ inert gases buried as neutral particles; △ inert gases ionized before burial. Reprinted with permission from *Proc. 4th Int. Vac. Congr. (1968)*, p. 325, D. Andrew. Copyright 1969, The Institute of Physics.

Pumping Speed and Conductance

Pump, in its design range, pumps a given <u>volume</u> of gas per unit time, independent of the pressure



$Q \equiv$ throughput (mass rate of flow) into pump

$S \equiv$ pumping speed (volume rate of flow)

$$S_p P_1 = Q$$

$C \equiv$ conductance (Ability of tube to transmit mass flow)

$$C(P_2 - P_1) = Q$$

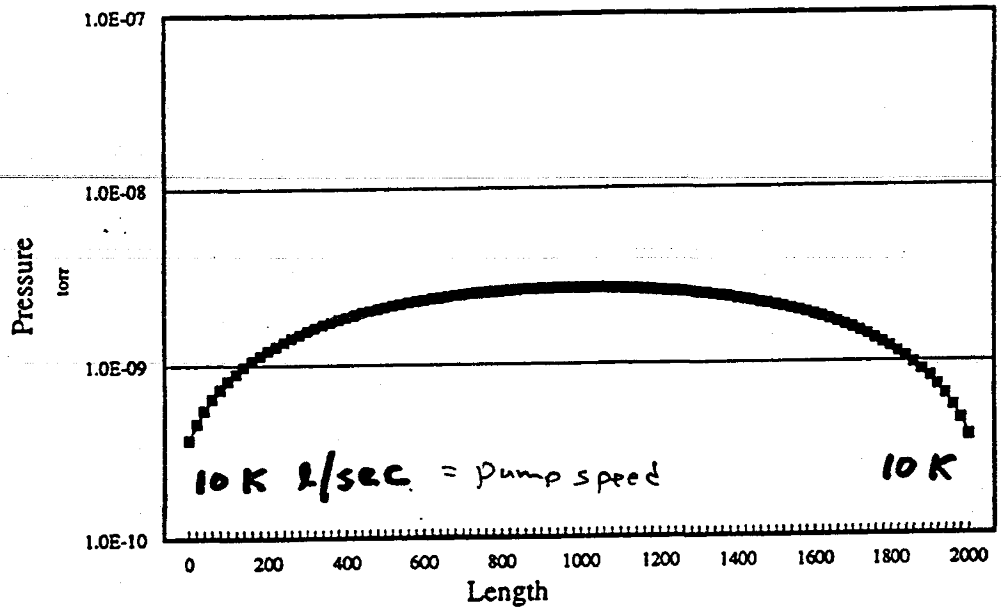
what is $S = \frac{Q}{P_2}$

$\frac{1}{S} = \frac{1}{S_p} + \frac{1}{C}$: Conductance limit

$P_1 = 0 \Rightarrow P_2 = Q/C$

Pressure Distribution

in LIGO vacuum tube

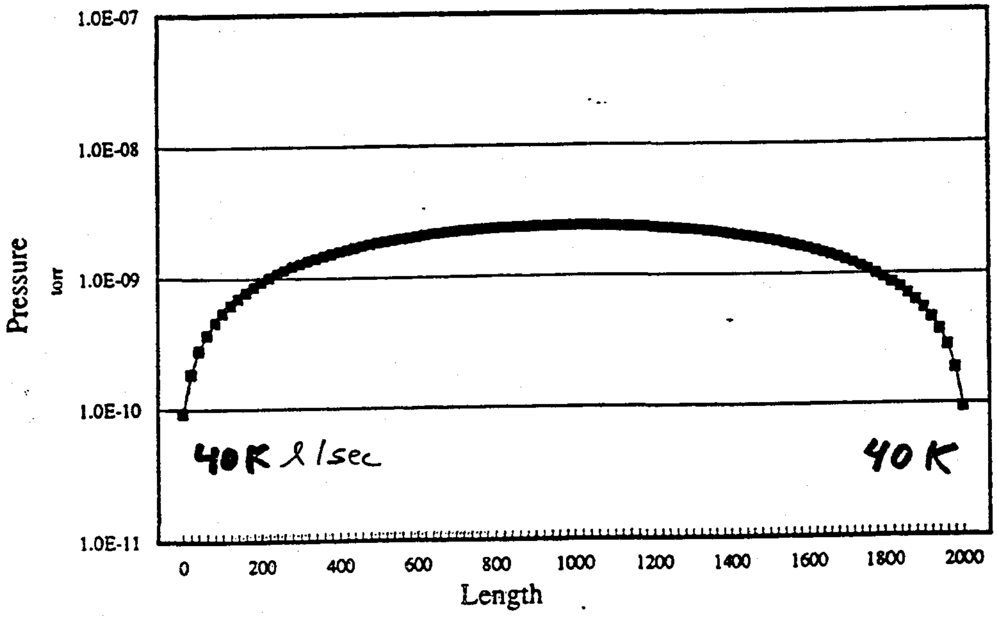


10 K l/sec = pump speed

10 K

mass 2, 300K, q=1E-13
10000 l/s pumps at 2000m

Pressure Distribution



40K l/sec

40 K

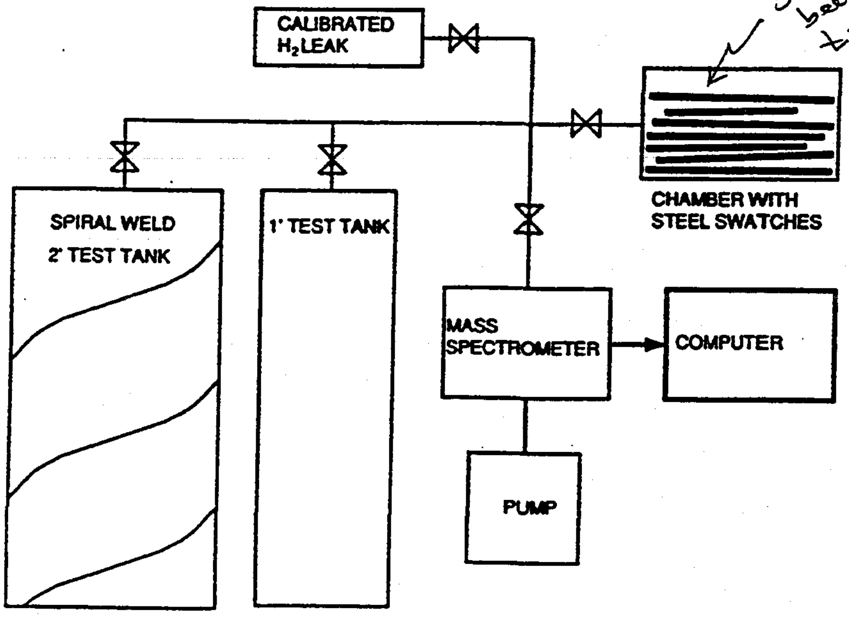
mass 2, 300K, q=1E-13
10000 l/s pumps at 2000m

↑
.. l/sec

Thus, going to higher pump speed improves pressure only a little.

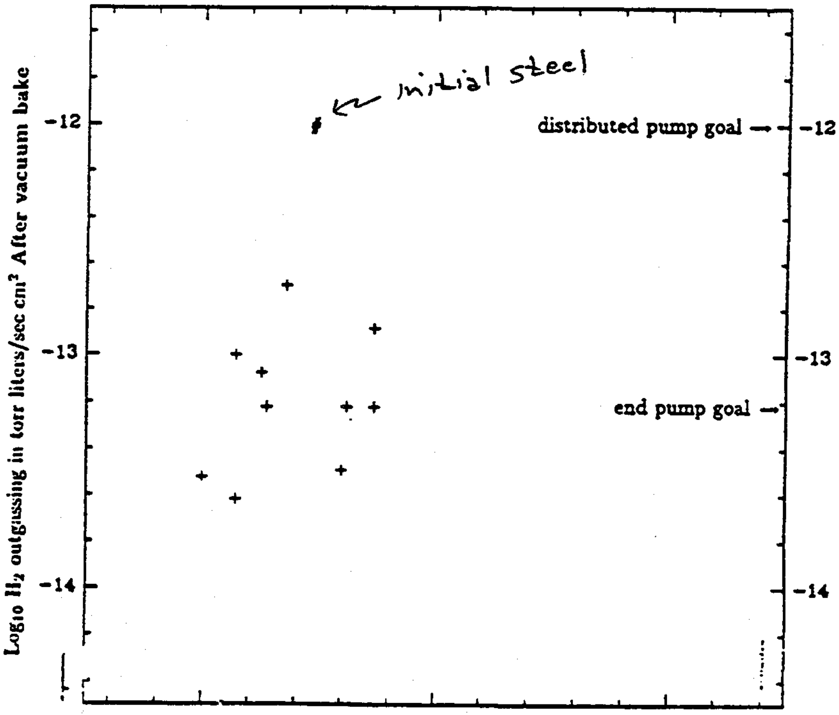
For ~~[illegible]~~ significant improvement, must add pumps along ~~[illegible]~~ tube.

Vacuum Test Facility Concept



steel that has been subjected to different kinds of bakes

HYDROGEN OUTGASSING TESTS

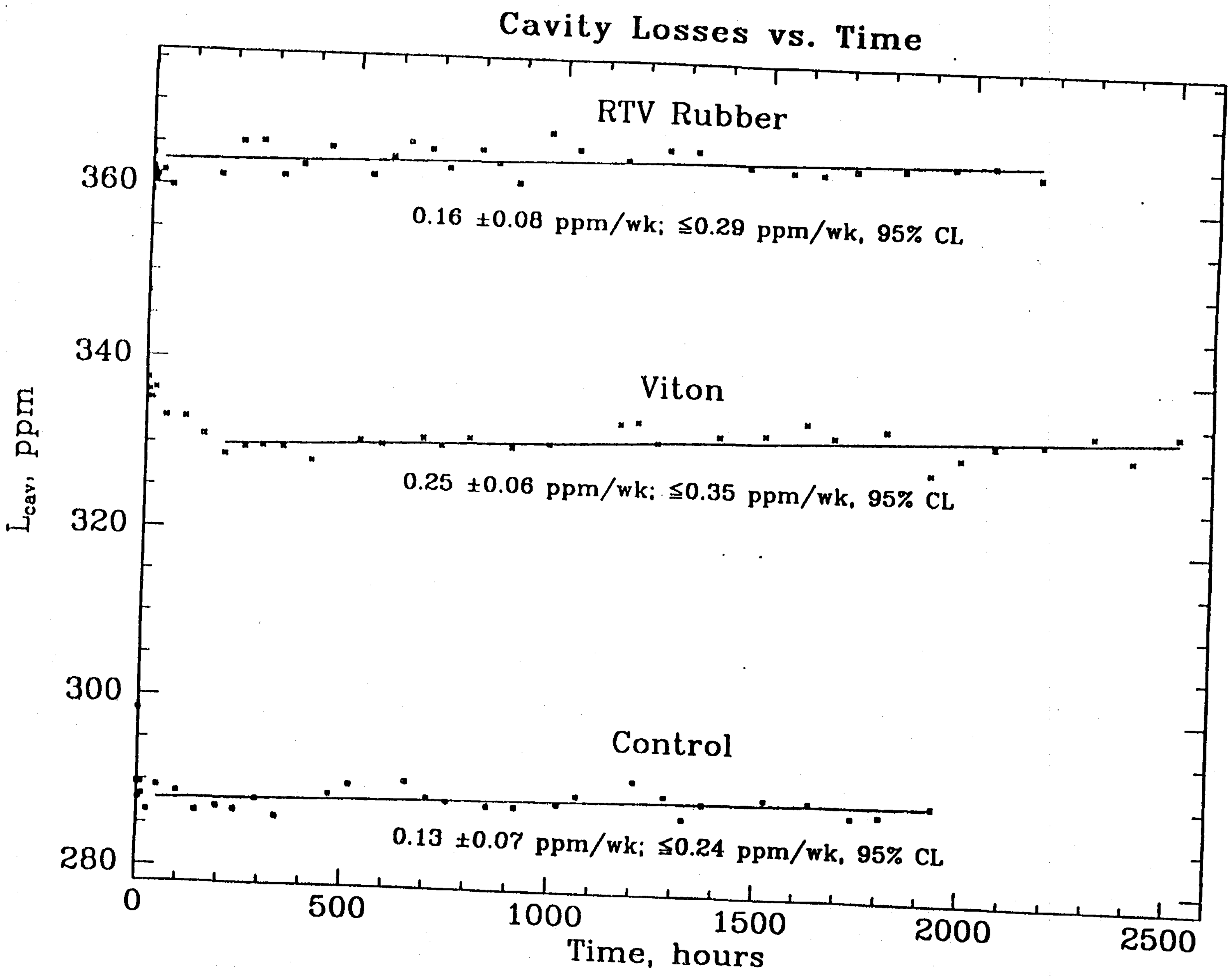


initial steel

air bake
36 hr, 450°C

Legend:

= Original Process


Cavity Losses vs. Time
RTV Rubber
0.16 ±0.08 ppm/wk; ≦0.29 ppm/wk, 95% CL
Viton
0.25 ±0.06 ppm/wk; ≦0.35 ppm/wk, 95% CL
Control
0.13 ±0.07 ppm/wk; ≦0.24 ppm/wk, 95% CL
L_{cav}, ppm
360
340
320
300
280
0
500
1000
1500
2000
2500
Time, hours

BATCH START

18. 40 meter Prototype

STAPLE
OR
DIVIDER

LECTURE 18

The 40 Meter prototype Interferometer as an Example of Many of the Issues Studied in this Course

Lecture by Robert Spero

Assigned Reading:

XX. R. Weiss, *Quart. Prog. Rep. Res. Lab. Electron. M.I.T.* **105**, 54 (1972). [The seminal paper presenting in detail how laser interferometers can be used for gravitational wave detection, including a comprehensive analysis of noise sources.]

YY. Robert L. Forward, "Wideband laser-interferometer gravitational-radiation experiment," *Phys. Rev. D* **17**(2), 379–390 (1977). [A description of the first interferometric gravitational-wave detector to be built (a 2 m Michelson interferometer) and the first search for gravitational waves using such a detector (a coincidence run conducted in 1972 between the interferometer and several bar detectors).]

A Few Suggested Problems: *Note: Your homework for Lectures 17 and 18 is to be turned in to Shirley Hampton in room 151 Bridge Annex before 1:00PM Friday June 3*

1. In this course you have encountered all the significant noise sources for interferometric detectors that the LIGO team is now aware of. Which of these were unanticipated by Weiss in Ref. 1 above; and are they "fundamental" or are they "technical"?
2. The interferometer described in Weiss's paper has a different optical configuration from the 40 m interferometer, but the shot noise calculated by Weiss is similar to that achieved in the 40 m. Why? Compare the shot noise limited sensitivity calculated by Weiss with the sensitivity achieved by Forward, and account for the difference.
3. The thermal noise calculated by Weiss is based on viscous damping [$\phi(\omega) \propto \omega$]. How does the thermal noise prediction change when if the damping is structural [$\phi(\omega)$ independent of ω]? cf. Lectures 13 and 14.
4. Weiss calculated noise from laser intensity fluctuations acting on the test masses via radiation pressure. Intensity fluctuations also result in noise (in a manner discussed in Spero's lecture) if the interferometer's operating point is offset from a "dark fringe" at the photodiode. Under what circumstances will this noise be larger than that due to radiation-pressure fluctuations.

Lecture 18
The 40 Meter Prototype Interferometer as an Example of Many of the Issues Studied in this Course

by Robert Spero, 27 May 1994

Spero lectured from the following transparencies. Kip has annotated them a bit.

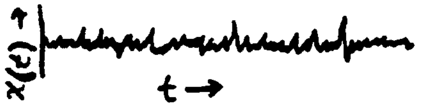
LECTURE 18

"THE 40 m INTERFEROMETER AS EXAMPLE OF ISSUES STUDIED IN COURSE"

R. SPERO
27 MAY 1994

Time Domain Data, Pulse Height Distribution, and Pulse Calibration [Lecture 3]

View Interferometer output in Time Domain,



and analyze as a train of pulses of varying height x(t)

Gaussian distribution: $N(x) \propto e^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2}$

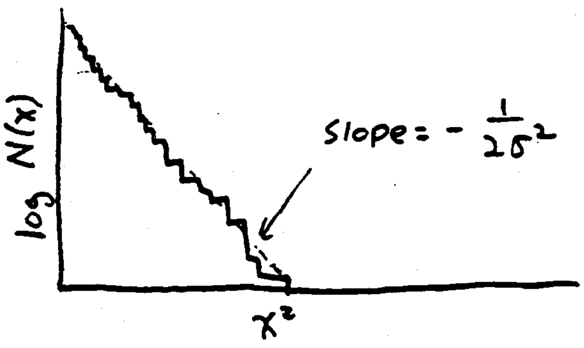
where $N(x)$ = "density" of pulses with amplitude x

σ = average (rms) pulse height

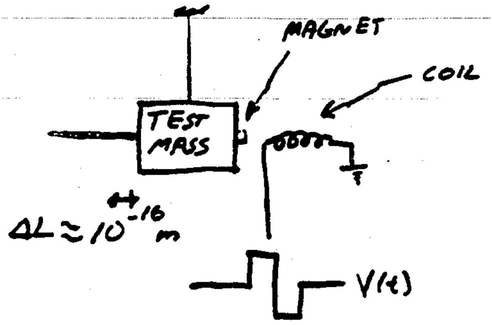
Then

signal-to-noise ratio (SNR) of a pulse = $\frac{x}{\sigma}$

Histogram the time series

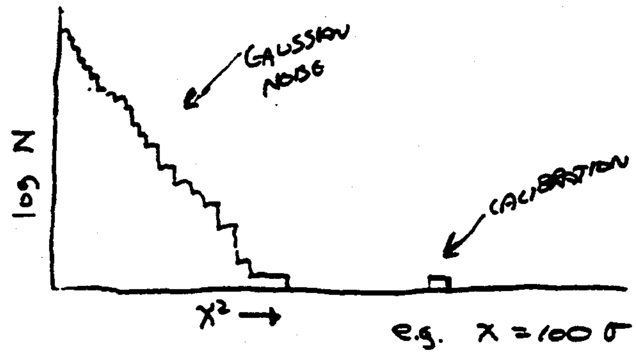


Calibration requires converting from voltage to ΔL or h: inject large $(x > 10\,\sigma)$ displacement pulse, of known amplitude



Q: HOW CAN ONE CALIBRATE THE CALIBRATOR; i.e. CONVERT FROM $V(t)$ TO $\Delta L(t)$

Calibration pulse is separate from noise:



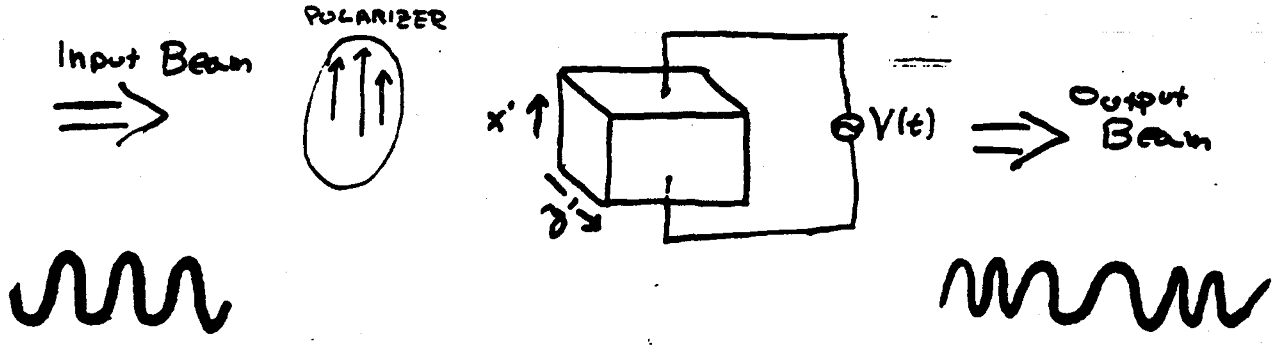
e.g. $x = 100\,\sigma \Rightarrow \sigma = 10^{-18}\,m$

σ = PULSE SENSITIVITY OF INTERFEROMETER?

IMPERFECTIONS IN PHASE MODULATION USING POCKELS CELLS

[LECTURES 7, 8]

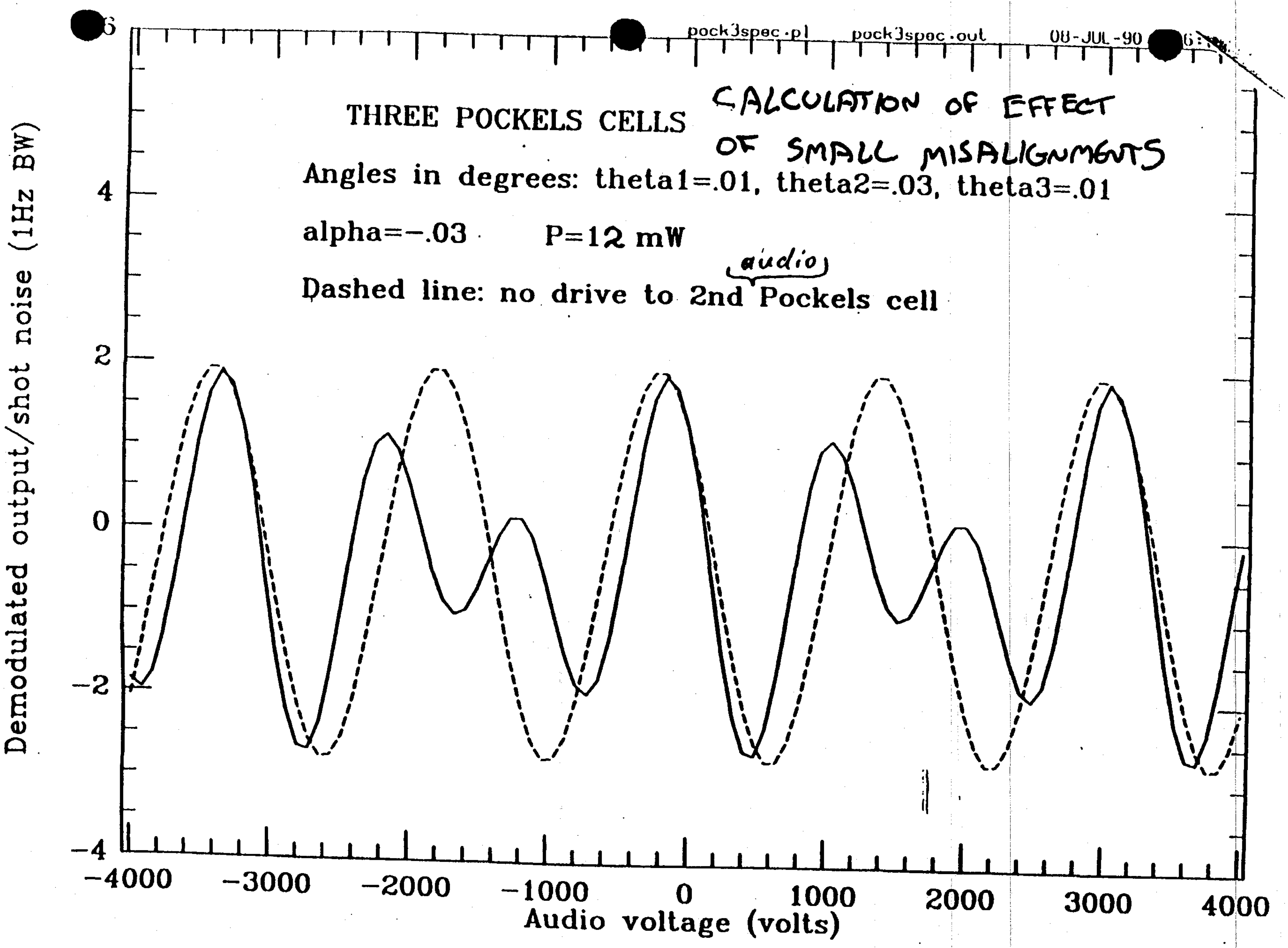
1. IDEAL POCKELS CELL PHASE MODULATOR



POLARIZER IS ALIGNED WITH CRYSTAL AXIS, SO POCKELS CELL DOES NOT ALTER POLARIZATION

ONLY EFFECT IS SMALL MODULATION OF INDEX OF REFRACTION n, RESULTING IN A CHANGE IN OPTICAL PATH LENGTH $\propto V(t)$

MISALIGNMENT OF AXES W.R.T. BEAM RESULTS IN INTENSITY MODULATION (MAXIMUM WHEN x', y' ROTATED BY 45°)



THREE POCKELS CELLS

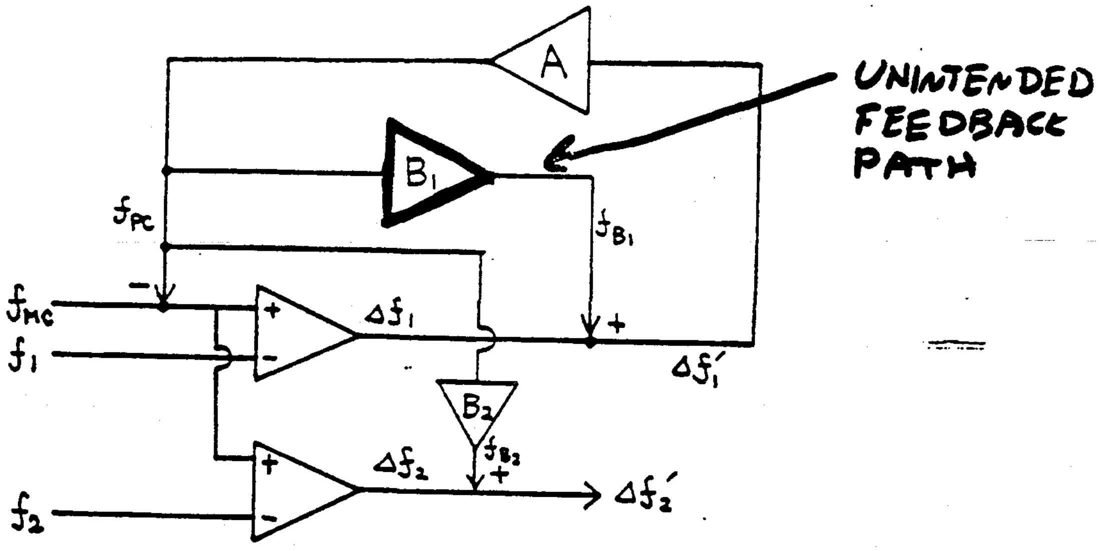
CALCULATION OF EFFECT OF SMALL MISALIGNMENTS

Angles in degrees: theta1=.01, theta2=.03, theta3=.01

alpha=−.03 P=12 mW

Dashed line: no drive to 2nd $\overset{\text{audio}}{\text{Pockels}}$ cell

SERVO MODEL OF POCKELS CELL MISALIGNMENT



f_{MC}: Frequency of the light out of the mode cleaner

f_1: Resonance frequency of the primary cavity

f_0: Stabilized frequency

Δf_1: True frequency deviation (between f_0 and f_1)

$\Delta f_1'$: Measurable frequency deviation (between f_0 and f_1)

A: Open loop transfer function of the primary servo

f_{PC}: Equivalent correction frequency to the PC

B_1: Transfer function from PC correction frequency to deviation frequency (in the primary) due to PC misalignment (through intensity noise around 12MHz)

f_{B1}: False frequency deviation due to PC misalignment

f_2: Resonance frequency of the secondary cavity

Δf_2: True frequency deviation (between f_0 and f_2)

$\Delta f_2'$: Measurable_frequency deviation (between f_0 and f_2)

B_2: Transfer function from PC correction frequency to deviation frequency (in the secondary) due to PC misalignment

f_{B2}: False frequency deviation due to PC misalignment

... AND CALCULATION OF EFFECT ON FREQUENCY STABILIZATION SERVO

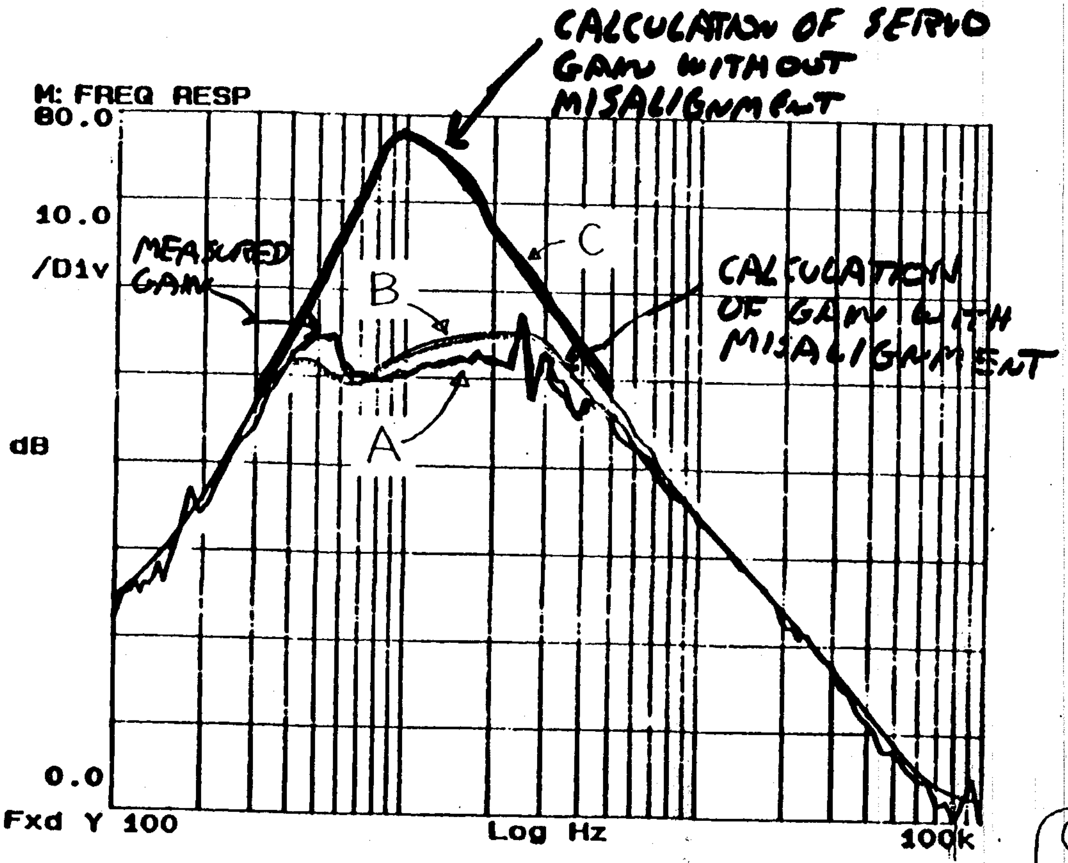


Fig. 1 Frequency suppression of the primary cavity servo.

A: Measured

B: Calculated (with spurious paths)

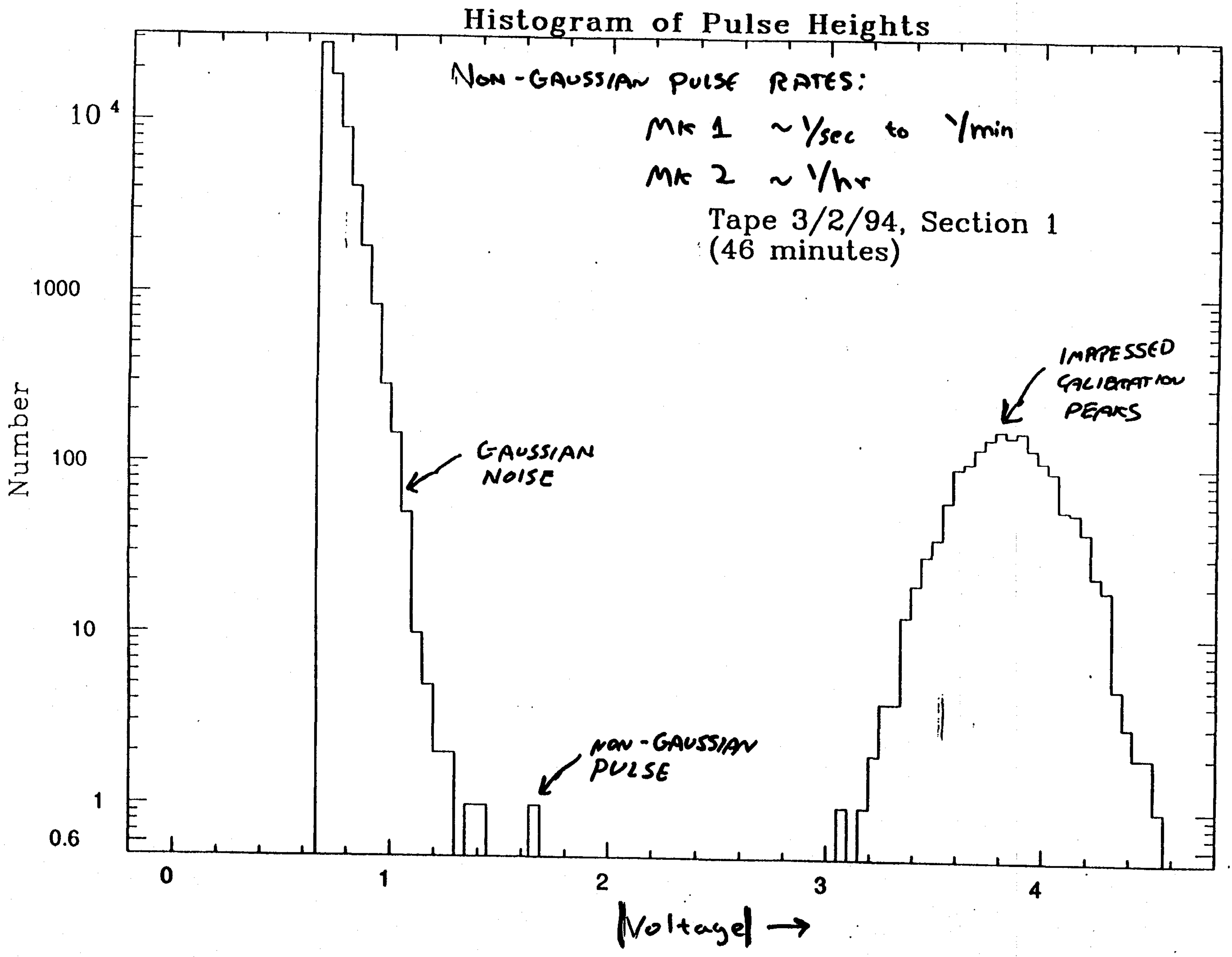
C: Calculated (without spurious paths)

Conclusion:

Pockels cells used in this configuration are unacceptable.

Solution:

Move Pockels cells out of the most sensitive locations



Histogram of Pulse Heights

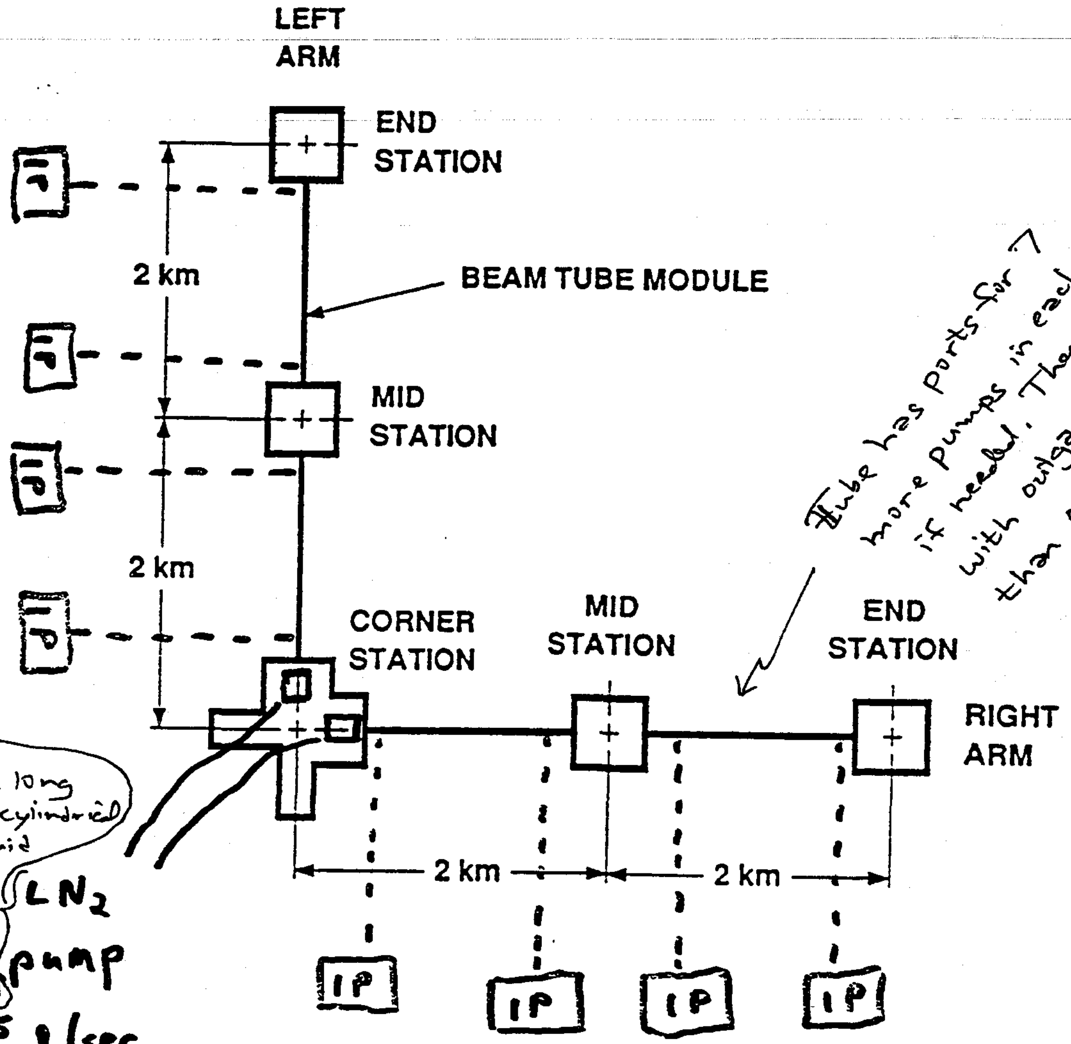
NON-GAUSSIAN PULSE RATES:

MK 1 ~1/sec to 1/min

MK 2 ~1/hr

Tape 3/2/94, Section 1
(46 minutes)

LIGO Initial Pumping



Tube has ports for 7 more pumps in each 2km, if needed. These could deal with outgasing 10x higher than our good steel giv

A few feet long section of cylindrical tube @ liquid Nitrogen temperature - catches molecules that hit wall

LN$_2$ pump

10^5 l/sec

For condensible gases (to keep hydrocarbons etc out of vacuum)

Ion pump

$S = 10^4$ l/sec

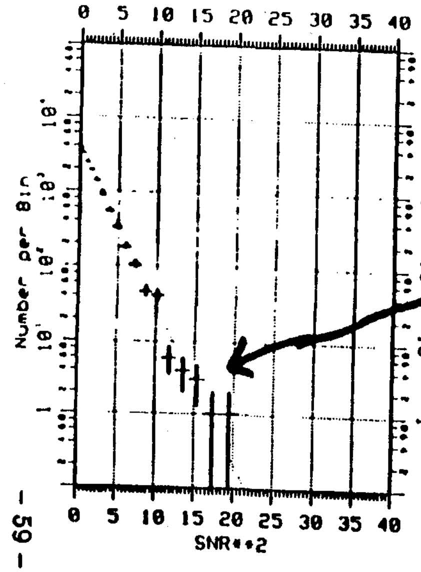
For H_2

MATCHED FILTER TO INCREASE SNR

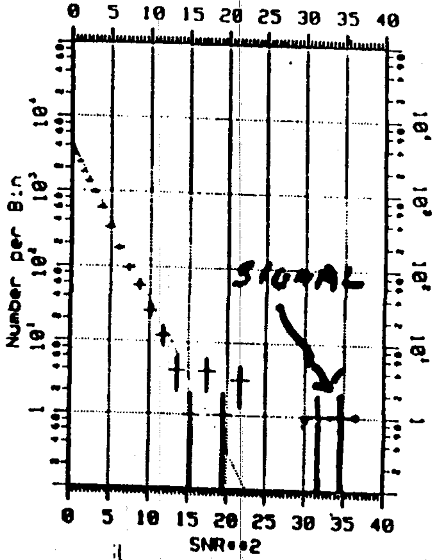
(A SIMULATION)

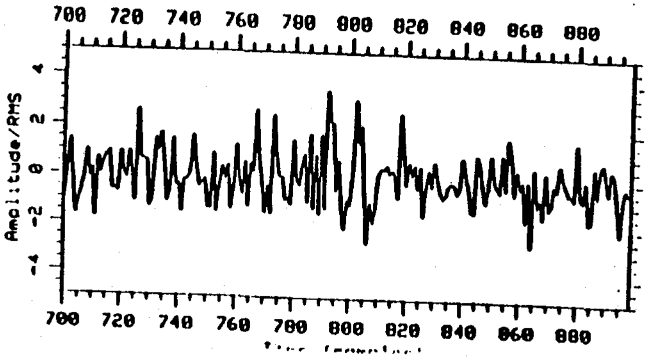
FILTER INPUT

FILTER OUTPUT

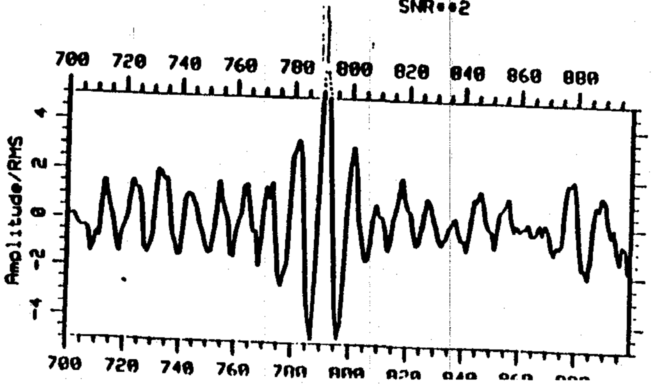


NO EVIDENCE OF SIGNAL ← PULSE HEIGHT DISTRIBUTIONS →



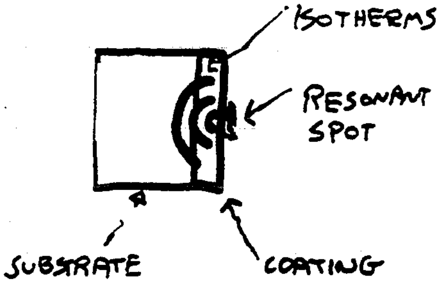


← TIME SERIES →



MIRROR HEATING AS LIMIT TO OPTICAL POWER

[LECTURE 9]



TWO HEATING EFFECTS

1) MECHANICAL DISTORTION VIA THERMAL EXPANSION (α)

2) OPTICAL DISTORTION VIA TEMPERATURE DEPENDENCE OF INDEX OF REFRACTION ($\Delta n/\Delta T$)

BOTH RESULT IN LENSING AND BEAM DISTORTIONS THAT ARE DIFFICULT TO CONTROL IF THE ABSORBED POWER IS $P_{ABS} \gtrsim 1$ W.

HEATING EFFECTS MINIMIZED BY

1) REDUCING ABSORPTION IN COATING

$\frac{P_{ABSORBED}}{P_{INCIDENT}} < 10^{-5}$ COATINGS ARE AVAILABLE

2) SELECTING SUBSTRATE MATERIAL

SMALL α, SMALL $\frac{\Delta n}{\Delta T}$, LARGE THERMAL CONDUCTIVITY λ

↑(thermal expansion)

MATERIAL	α $(10^{-6}/K)$	$\frac{\Delta n}{\Delta T}$ $(10^{-6}/K)$	λ $(W/m\text{-}K)$
FUSED SILICA	0.59	12	1.3
BK7	7.1	0.6	1.1
SAPPHIRE	8.4	13	35

It appears that coating heating will always dominate over substrate heating.

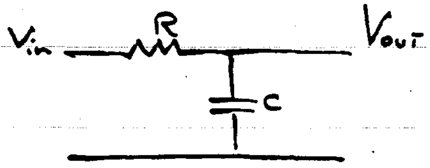
3) PIES IN THE SKY

a) ADVANCED OPTICAL ARRANGEMENTS, SUCH AS DUAL RECYCLING

b) ADAPTIVE OPTICS

PRACTICAL LIMITATIONS TO LOW PASS FILTER

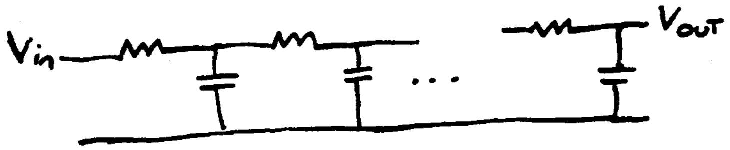
1) PHASE SHIFT AFFECTS SERVO PERFORMANCE
2) FILTER OUTPUT NOISE



"SINGLE POLE"

$$\frac{V_{OUT}}{V_{IN}} = \frac{Z_c}{R+Z_c} = \frac{1}{1+j\frac{\omega}{\omega_0}} \qquad \omega_0 = \frac{1}{RC}$$

"n poles" (heuristic only) ← because the successive stages are not sufficiently independent

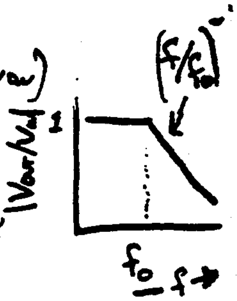


$$\frac{V_{OUT}}{V_{in}} = \frac{1}{\left(1+j\frac{\omega}{\omega_0}\right)^n}$$

$\omega \gg \omega_0 \Rightarrow \left|\frac{V_{OUT}}{V_{in}}\right| \simeq \left(\frac{\omega}{\omega_0}\right)^{-n}$ ("STOPBAND" ATTENUATION)

PHASE SHIFT IN "PASSBAND" – $\omega \lesssim \omega_0$:

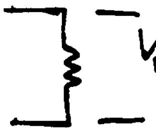
$\omega \ll \omega_0 \Rightarrow \frac{V_{OUT}}{V_{in}} \simeq 1 - nj\frac{\omega}{\omega_0}$; Phase $\simeq -n\frac{\omega}{\omega_0}$



EXAMPLE: $n=12$, $f_0 = 10$ Hz attenuates noise at 100 Hz by $\sim 10^{12}$,
BUT CAUSES ~70° phase shift at $f = 1$ Hz,
and may make servo with gain at 1 Hz
unstable

(In this case, lowering the bandwith is an adequate solution; role off the gain at $f \lesssim 1$ Hz and thereby get a lower phase shift at unity gain)

JOHNSON NOISE OF RESISTOR



$$V_{noise}^2 = 4kTBR$$

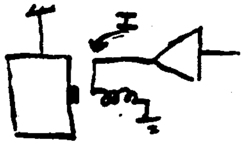
$$\tilde{V}(f) = \sqrt{4kTR}$$

$$\tilde{V}_{50\,\Omega} = 9\cdot 10^{-10}\,V/\sqrt{Hz}$$

PRACTICAL CONSEQUENCE: NOISE AT FILTER OUTPUT CAN BE SIGNIFICANT, ESPECIALLY AT FREQUENCIES WHERE LOOP GAIN IS SMALL

JOHNSON NOISE IN LONGITUDINAL CONTROL SERVO

$\ddot{x} \propto I$



[Examples of the kinds of things one must worry about.]

$R_{coil} \simeq 2\,\Omega$

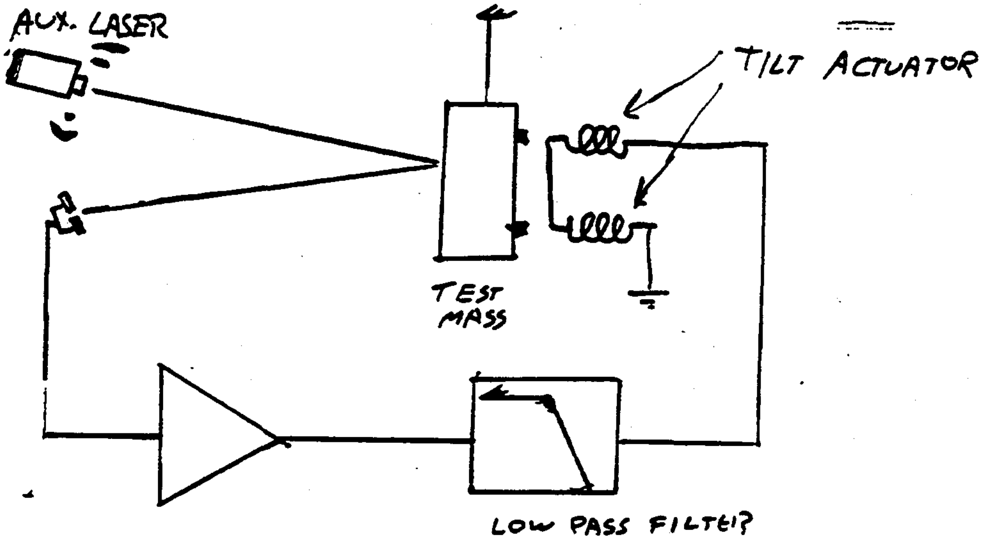
$$\tilde{V}(f) \simeq 2\cdot 10^{-10}\; V/\sqrt{Hz}$$

$$x = \frac{\beta V}{f^2} \qquad \beta = 5\cdot 10^{-5}\,\frac{m}{sec^2}\cdot\frac{1}{Volt}$$

$$\Rightarrow \tilde{x}(f) = 10^{-18}\,\frac{m}{\sqrt{Hz}} \text{ at } 100\text{ Hz}$$

ANTI-NOISE FILTERS AND SERVO BANDWIDTH

[LECTURE 10]



UNCONTROLLED TEST MASS IS QUIET IN TILTS IN SIGNAL BAND $f \gtrsim 100$ Hz, BUT HAS LARGE MOTION NEAR $f = 1$ Hz.

OPTICAL LEVER SENSING SYSTEM IS ADEQUATELY STABLE NEAR 1 Hz, BUT NOISY FOR $f \gtrsim 100$ Hz (VIBRATION OF LASER, CONVECTION CURRENTS DEFLECTING BEAM, ...

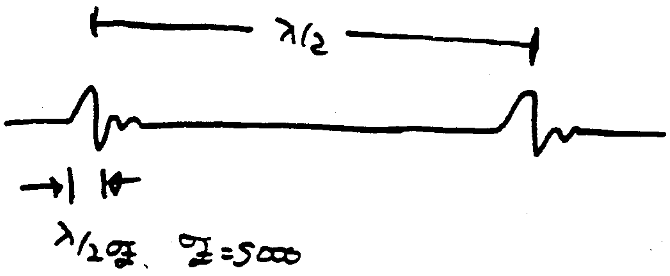
(Q: HOW CAN ONE MAKE A QUIET OPTICAL LEVER?)

LOW PASS FILTER IS INTENDED TO PASS THE LOW FREQUENCY CONTROL SIGNALS, AND BLOCK THE SENSING NOISE BEFORE IT CAN DISTURB THE TEST MASS.

STARTING UP THE CONTROL SYSTEM: LOCK ACQUISITION

[LECTURE 11]

PROBLEM: MOST (>99.9%) CAVITY IS OUT OF RESONANCE, AND THERE IS NO SIGNAL ~~FOR THE~~ CONTROL SYSTEM TO USE.



DESIGN STRATEGY:

1.) Slow down fringes by seismic isolation, active damping

2) Construct servo to have high gain and stable operation when in resonance

3) Attend to electronic saturation properties

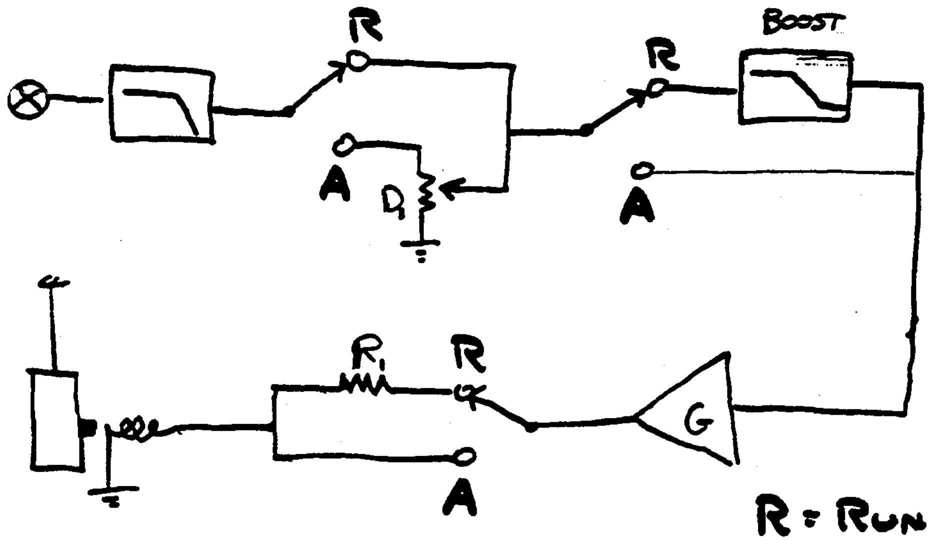
4) Maximize dynamic range

$\Delta x_{\text{BEFORE LOCK}} \approx 10^{-6}$ m

$\Delta x_{\text{IN LOCK}} \approx 10^{-12}$ m

Saturation is inevitable when trying to lock the interferometer; one must be careful of how the electronics behaves when saturated.

REQUIREMENTS FOR LOCK ACQUISITION COMPETE WITH REQUIREMENT OF LOW NOISE AFTER RESONANCE IS ESTABLISHED. TRICK: SWITCH FROM "ACQUIRE" MODE TO "RUN" MODE. SWITCH MUST BE GENTLE, OTHERWISE LOCK IS DISRUPTED



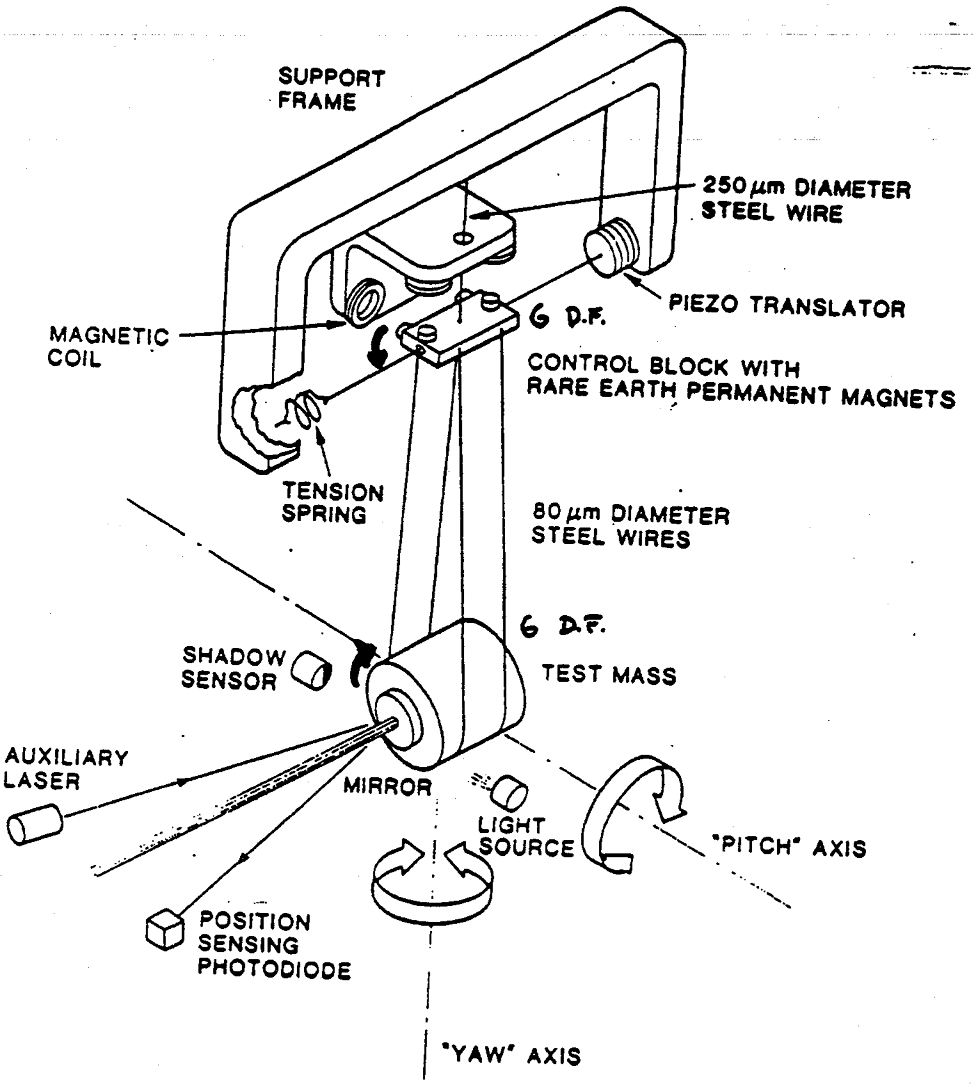
R = RUN MODE
A = ACQUIRE MOD

<u>ACQUIRE MODE</u>: INPUT TO AMPLIFIER G IS ATTENUATED BY D_1, TO REDUCE OVERLOADING

<u>RUN MODE</u>: ATTENUATE OUTPUT OF G FOR LESS NOISE (AND LESS DYNAMIC RANGE). ATTENUATION SAME AS D_1

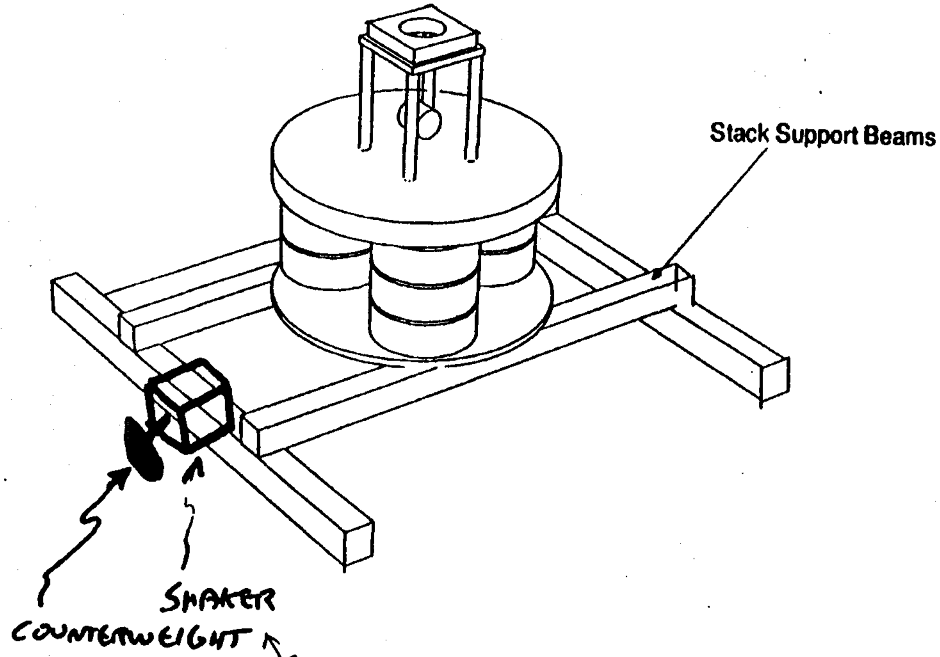
ALSO: LOW FREQUENCY BOOST INSERTED. INCREASES GAIN BELOW 100 Hz, DOES NOT AFFECT SERVO STABILITY

DEPARTURE FROM IDEAL $\left(\frac{1}{f^2}\right)$
SUSPENSION: RESONANT NORMAL MODES
OF COUPLED MECHANICAL OSCILLATORS
[LECTURE 12]



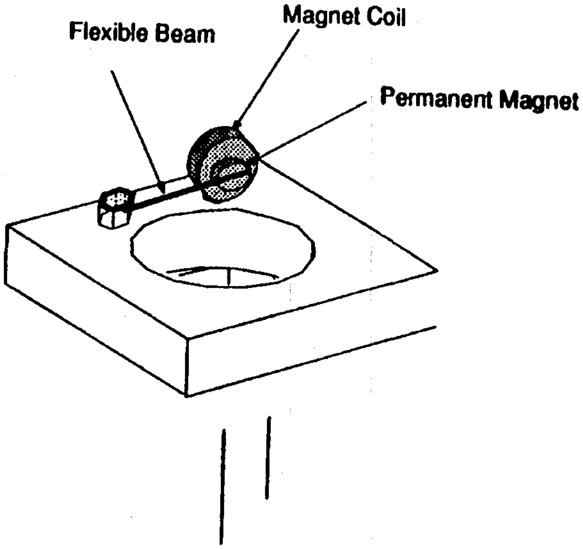
SUSPENSION ISOLATION MEASUREMENT

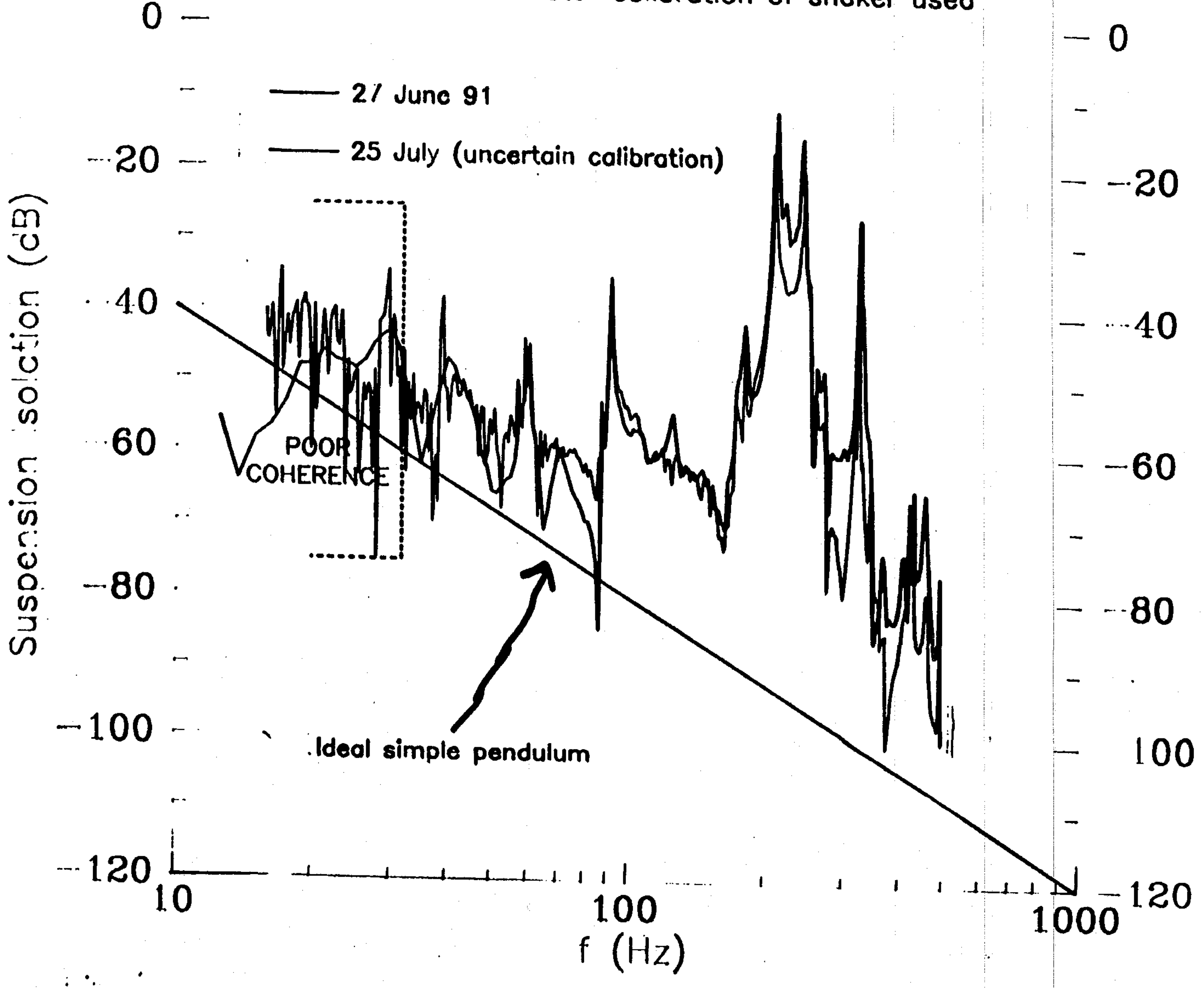
STACK EXCITATION



To test the suspension's performance (measure transfer function from ground motion to interferometer output)

SUSPENSION EXCITATION



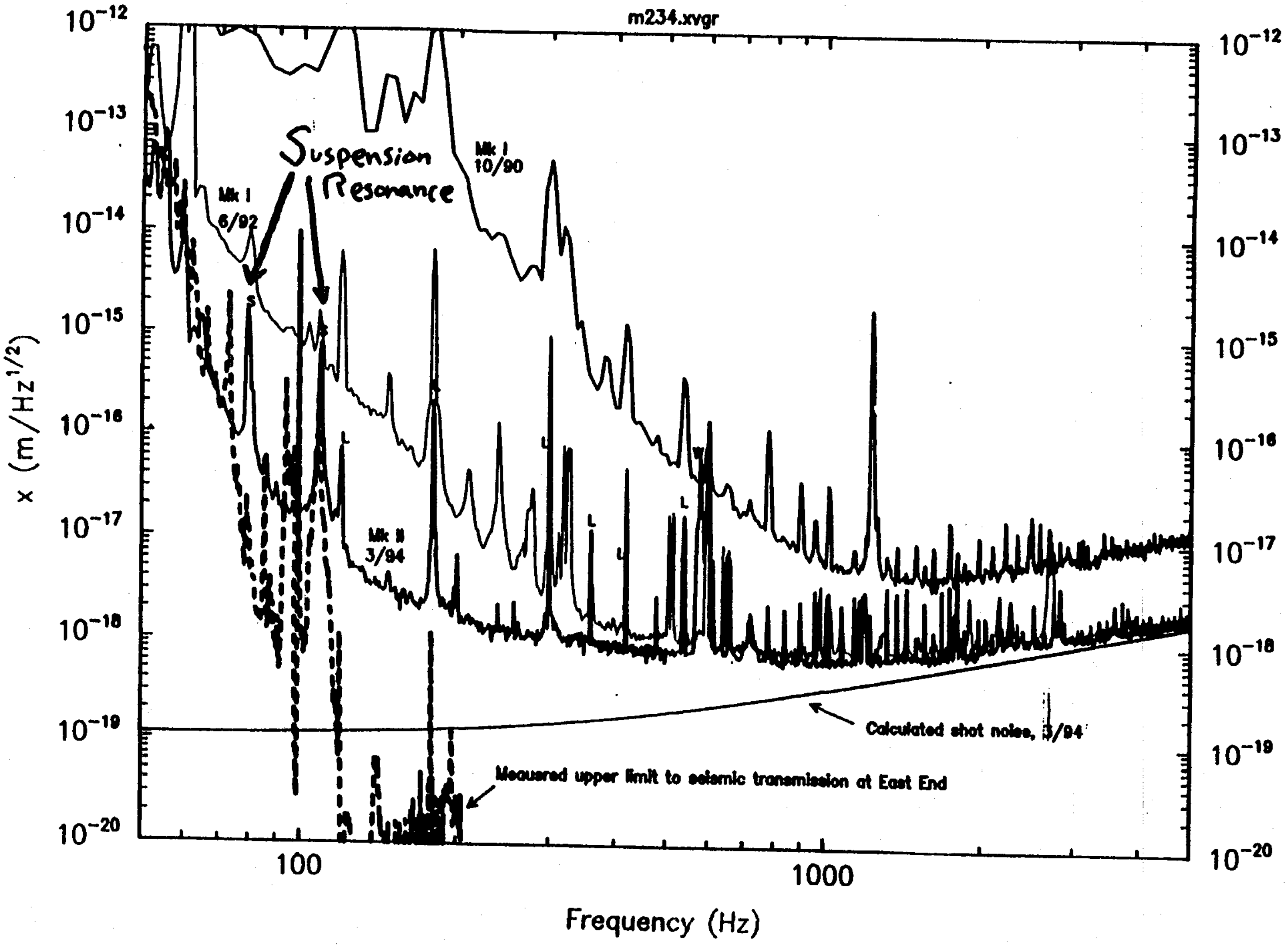


TOP PLATE SHAKER vs. INTERFEROMETER OUTPUT

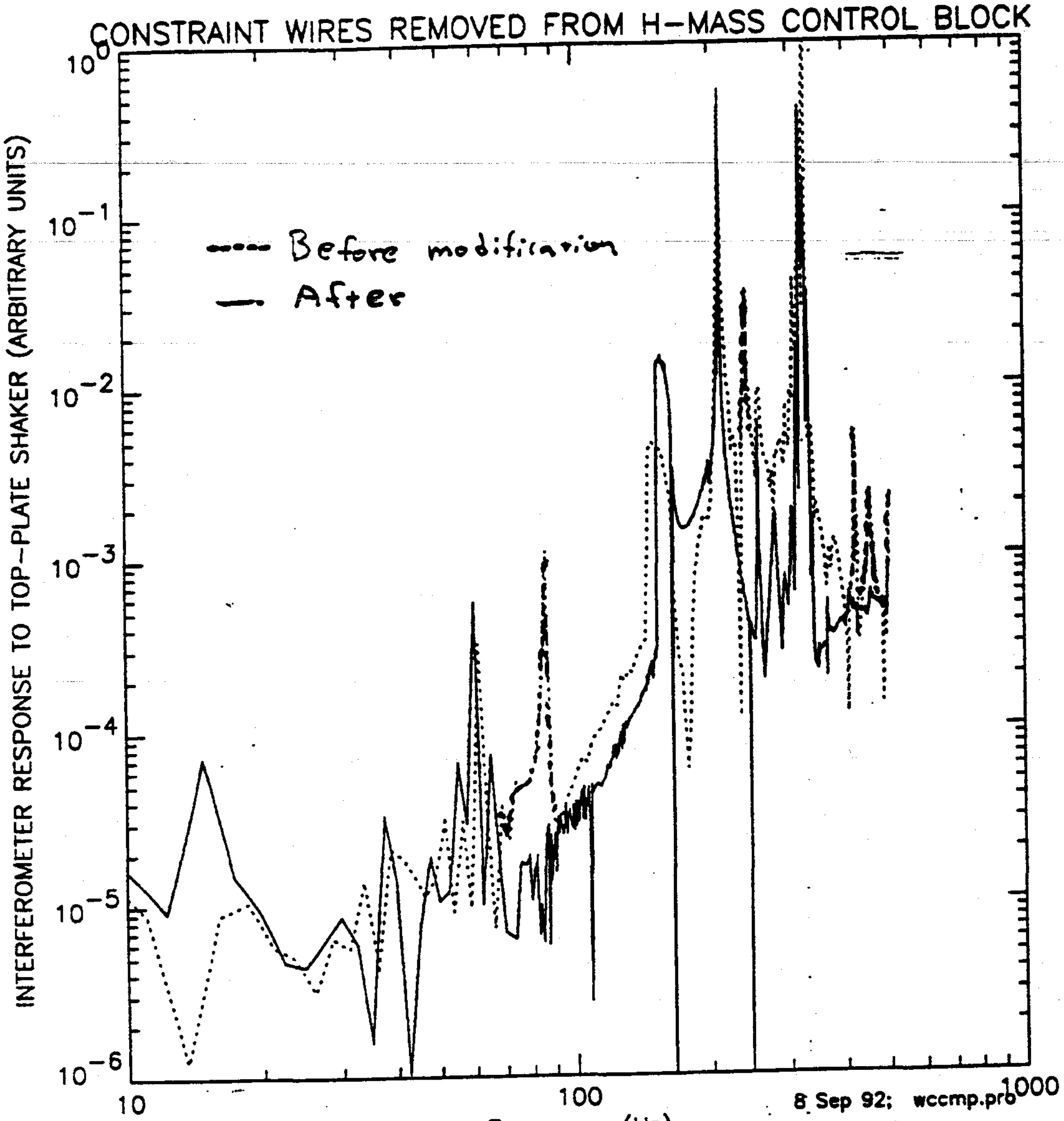
In–air accelerometer calibration of shaker used

SUSPENSION RESONANCES EVIDENT
IN INTERFEROMETER OUTPUT

Displacement Sensitivity of Caltech 40 m Interferometer



(SOME) REDUCTION OF SUSPENSION RESONANCES
BY (SLIGHT) SIMPLIFICATION OF SUSPENSION

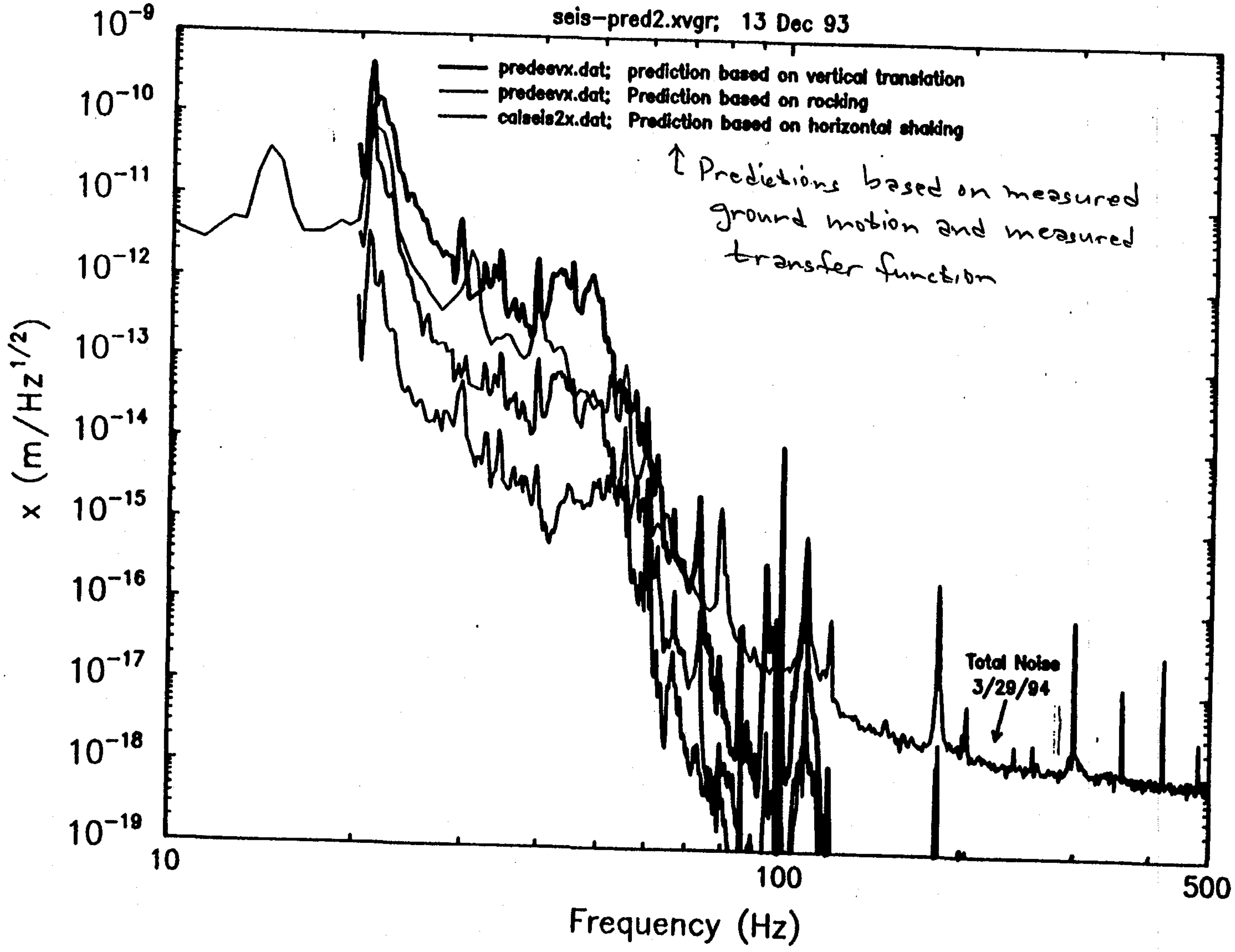


8 Sep 92; wccmp.pro
wc3, htphf

SEISMIC ISOLATION STACK IMPERFECTION:

INTERFEROMETER IS MORE RESPONSIVE TO VERTICAL EXCITATION!

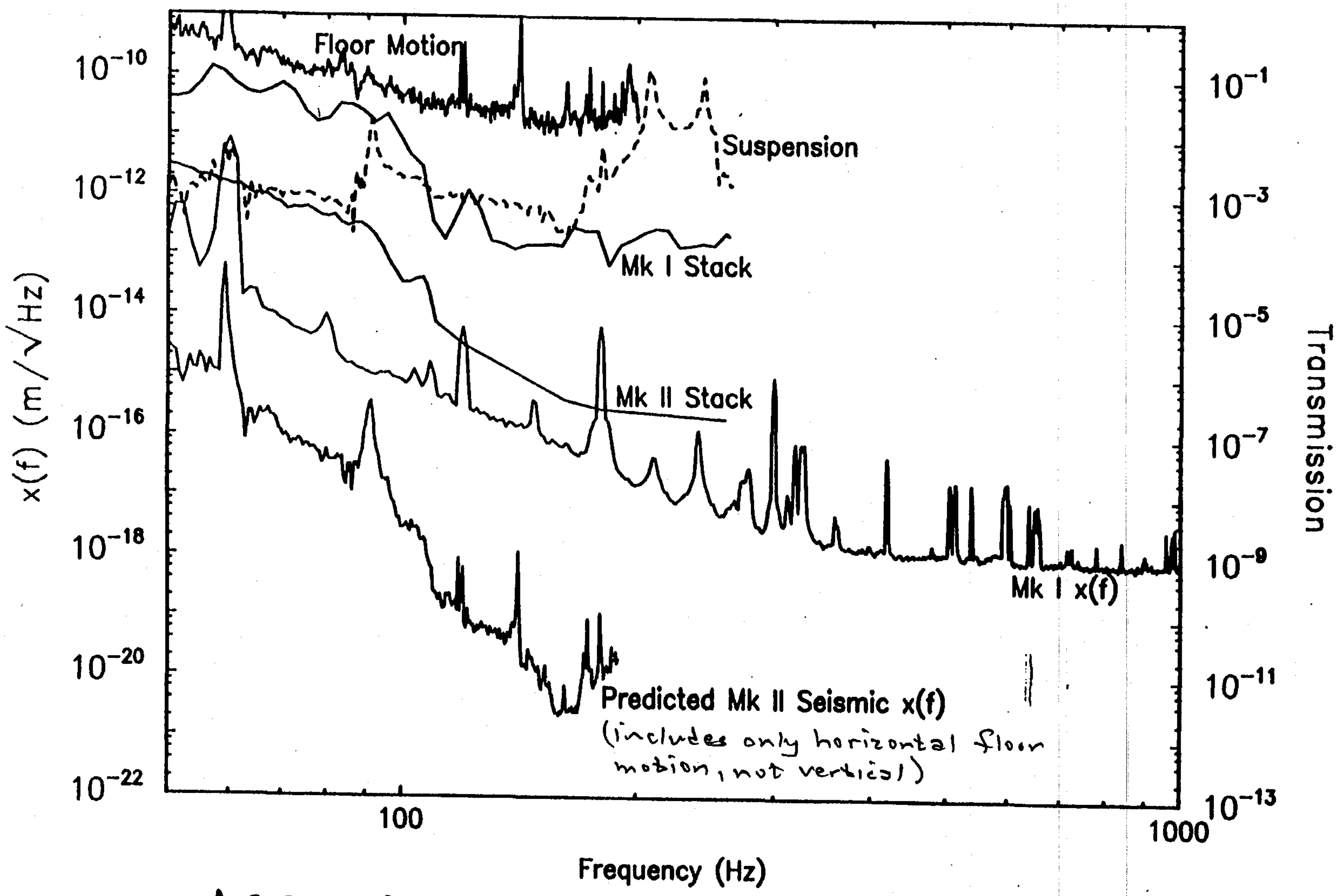
Seismic Feedthrough at East End, Compared to Total Noise



Design Exercise:

SEISMIC ISOLATION DESIGN FOR MARK II:
OLD SUSPENSION, NEW STACK

ISOLATION STACK PERFORMANCE



(includes only horizontal floor motion, not vertical)

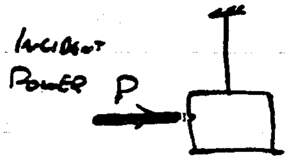
NEXT: NEW SUSPENSIONS,
SOFTER STACKS

Wed May 26 10:10:55 1993

Two Types of Intensity Noise

[Lecture 7, 16-17]

1) Imbalanced radiation pressure fluctuations

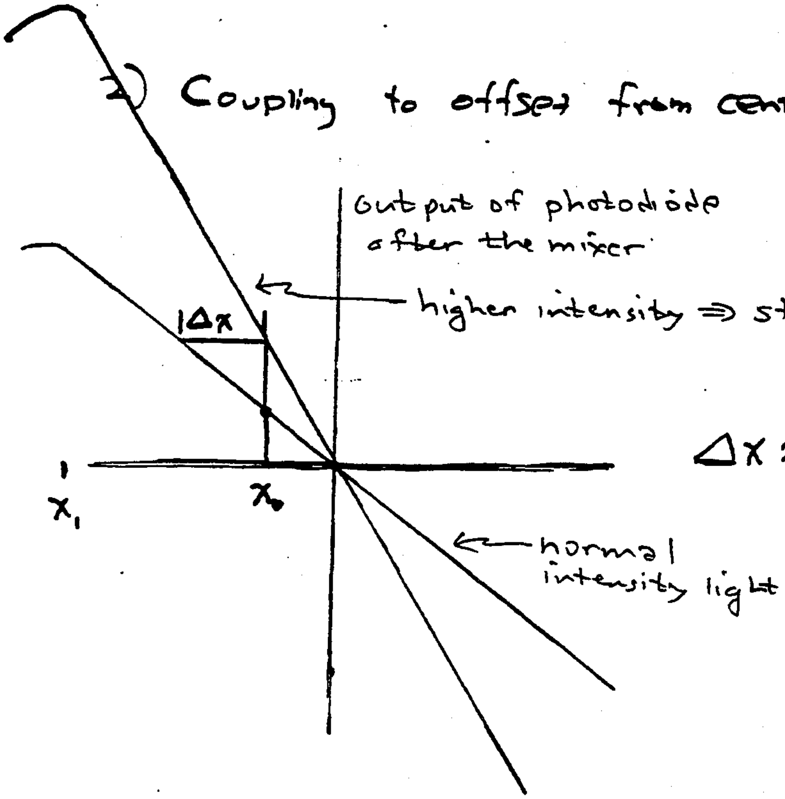


Force $\tilde{F}(f) \propto$ Power $\tilde{P}(f)$

Fluctuation is "fundamental" if Power is stabilized so well that dominant effect is shot noise; the fluctuating force is

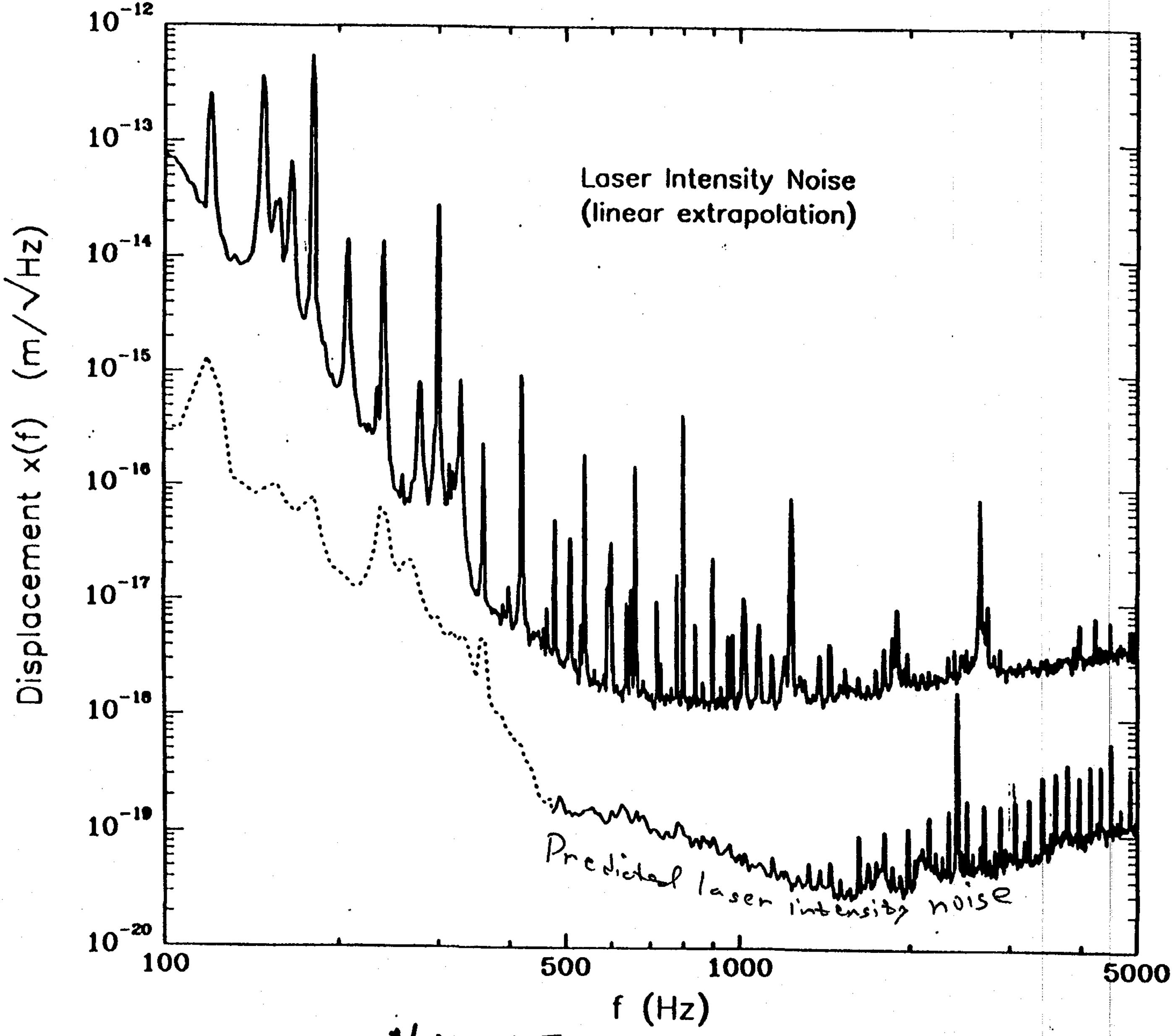
$$\tilde{F}(f) = \sqrt{\frac{\hbar P}{\lambda c}}$$

2) Coupling to offset from center of resonance



$$\Delta x \approx \frac{x_0}{x_1}\,\frac{\Delta I}{I}$$ ← Operating point; offset from dark fringe

MEASURED EFFECT OF INTENSITY NOISE IS SMALL

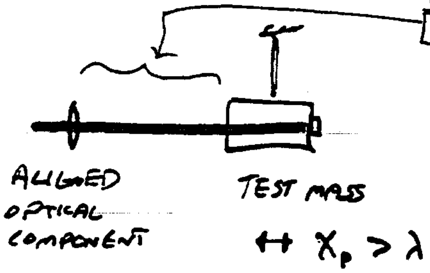


NOTE: THIS IS AN OLD NOISE SPECTRUM

UPCONVERSION BY SCATTERING

[LECTURE 15]

Mechanical vibrations at low frequency & large amplitude produce high-frequency noise



Unintended resonator

$$\text{Maximum fringes/sec} = f_p$$

peak velocity of motion of components of unintended resonator

$$= \frac{V_p}{\lambda/2} = \frac{2\pi X_p f_0}{\lambda/2}$$

e.g. $X_p = 10\lambda$, $f_0 = 1$ Hz $\Rightarrow$ $f_p \simeq 100$ Hz

("UPCONVERSION")

REDUCE EFFECT BY DAMPING RESONANCES (ESPECIALLY EFFECTIVE IF $X_p < \frac{\lambda}{2}$ — NO UPCONVERSION)

AND BY OPTICAL ISOLATION

BATCH START

STAPLE
OR
DIVIDER

THE PHYSICS OF LIGO

II. Selected Readings

Materials from a Course taught at Caltech
by members of the LIGO team and others
in Spring 1994

LIGO-T940067-00-R

organized and edited by Kip S. Thorne

California Institute of Technology

1994

Not to be Reproduced, Because of Copyright Issues

PREFACE

In the spring term of 1994, I organized a course at Caltech on *The Physics of LIGO* (i.e., the physics of the Laser Interferometer Gravitational Wave Observatory). The course consisted of eighteen 1.5-hour-long tutorial lectures, delivered by members of the LIGO team and others, and it was aimed at advanced undergraduates and graduate students in physics, in applied physics, and in engineering and applied sciences, and also at interested postdoctoral fellows, research staff, and faculty.

In my mind the course had several purposes: (i) It used LIGO as a vehicle for teaching students about the physics and technology of high-precision physical experiments. (ii) It served as a tutorial on the physics of LIGO for scientists and engineers, who had joined the LIGO team in the preceding year in preparation for the beginning of LIGO's construction. (iii) It served as an introduction to the science and technology of LIGO for other members of the Caltech community: In spring 1994, LIGO was just beginning to emerge from two years of controversy on the Caltech campus, and a number of faculty and staff wanted to learn in detail about the LIGO team's interferometer R&D, so they could form opinions of their own about whether the Project was well conceived and its interferometer development was being well executed. (It is my impression, in retrospect, that most and perhaps all of the faculty and staff who attended the course regularly emerged with a positive view LIGO.)

The lectures were delivered in Room 107 Downs on Wednesdays from 1:00 to 2:30 PM and Fridays from 10:30AM to noon. The audience typically consisted of about 5 undergraduates, 10 graduate students, 5 postdoctoral fellows, 8 professors, and 15 members of the LIGO team—and, for some lectures, rather more than this, especially more professors. The audience was mostly from Physics and Engineering, but a smattering of other disciplines was represented (including even an occasional social scientist). The undergraduates and some of the graduate students took the course for credit under the rubric of Physics 103.

These two Volumes contain the materials distributed at the lectures, augmented occasionally in Volume I by lecture notes that Malik Rakhmanov (the grader) or I have written, describing the lecture. More specifically:

Volume I contains (i) copies of the transparencies used in each lecture, or—in the case of lectures not based on transparencies—notes on the lecture prepared by Rakhmanov or me; and (ii) lists of references and sets of exercises prepared by me and/or the lecturers.

Volume II contains copies of the most important of the references that the lecturers chose to accompany their lectures. Some references are extracted from textbooks or technical monographs, others are from the original scientific literature, and a few are preprints of papers not yet published. Because we have not sought, from the publishers of these references, permission for widespread duplication and distribution, only a few copies of

Volume II are being made; and Volume II carries an admonition on its cover page that it should not be reproduced.

For these volumes I have given sequential capital-letter labels (A, B, C, ... Z, AA, BB, ... YY) to all the readings that appear in Volume II, and have revised the reference lists in Volumes I and II to reflect this labeling. References not included as readings are now labeled with lower-case letters (a, b, c, ... z).

These Volumes will be of value not only as a historical record, but also as a reference source for members of the LIGO team and others, and as an aid for people who did not attend the lectures and who want to begin learning about LIGO. For example, people who join the LIGO team during the next several years may find these volumes helpful in getting oriented. (To those who joined the team during the summer or early autumn of 1994, I apologize that I have been so slow in putting these volumes together.)

I thank the lecturers for the extensive time, energy, and enthusiasm that they put into this course. No single person could possibly have delivered this set of lectures, especially not I! I also thank Robbie Vogt and Stan Whitcomb who, as Director and Deputy Director of LIGO, encouraged me in 1993 to organize this course and encouraged the members of the LIGO team to help me make it a reality. Finally, for their enthusiastic backing of this effort, I thank the entire LIGO team, Barry Barish (LIGO PI), Tom Everhart (the Caltech President), Paul Jennings (the Provost), Charles Peck (the Chair of Physics, Mathematics, and Astronomy), and a number of Caltech faculty members.

Kip S. Thorne
Caltech
20 October 1994

CONTENTS OF VOLUME II

Note: For each lecture, this volume contains the list of references prepared by the lecturer with comments about the relevance of each reference, followed by a copy of each of the most important references. In the contents below, we list the copied references for each lecture. The copied references are labeled sequentially by capital letters (A, B, C, ... , Z, AA, BB, ... YY). The other references are labeled sequentially by lower-case letters.

J. B. J. Meers, "Recycling in laser-interferometric gravitational-wave detectors," *Phys. Rev. D*, **38**, 2317–2326.

K. B. J. Meers, *Physics Letters A*, "The frequency response of interferometric gravitational wave detectors," *Physics Letters A*, **142**, 465 (1989).

L. B. J. Meers and R. W. P. Drever, "Doubly-resonant signal recycling for interferometric gravitational-wave detectors." (preprint)

M. J. Mizuno, K. A. Strain, P. G. Nelson, J. M. Chen, R. Schilling, A. Rudiger, W. Winkler and K. Danzman, "Resonant sideband extraction: a new configuration for interferometric gravitational wave detectors," *Phys. Lett. A*, **175**, 273–276 (1993).

N. R. W. P. Drever, "Interferometric Detectors of Gravitational Radiation," in *Gravitational Radiation*, N. Deruelle and T. Piran, eds. (North Holland, 1983).

6. **Overview of a real interferometer** *by Stanley E. Whitcomb* [15 April]

O. D. Shoemaker, R. Schilling, L. Schnupp, W. Winkler, K. Maischberger, A. Rudiger, "Noise behavior of the Garching 30-meter prototype gravitational-wave interferometer," *Physical Review D,* **38**, 423-432 (1988).

P. Benjamin C. Kuo, *Automatic Control Systems* (Prentice-Hall), "Chapter 1. Introduction."

7. **Lasers and input optics—I** *by Robert E. Spero* [20 April]

Q. A. Rüdiger, R. Schilling, L. Schnupp, W. Winkler, H. Billing and K. Maischberger, "A mode selector to suppress fluctuations in laser beam geometry," *Optica Acta*, **28**, 641–658 (1981).

R. A. Yariv, *Optical Electronics* (Saunders College Publishing, 1991), "Chapter 10. Noise in Optical Detection and Generation."

S. A. Abramovici and Z. Vager, "Comparison between active- and passive-cavity interferometers," *Phys. Rev. A*, **33**, 3181 (1986).

T. J. Gea-Banacloche, "Passive versus active interferometers: Why cavity losses make them equivalent," *Phys. Rev. A*, **35**, 2518 (1987).

U. T.M. Niebauer, R. Schilling, K. Danzmann, A. Rudiger, W. Winkler, "Nonstationary Shot Noise and its Effect on the Sensitivity of Interferometers *Phys. Rev A* **43**, 5022–5029 (1991).

V. P.H. Roll, R. Krotkov, and R.H. Dicke, "The equivalence of inertial and passive gravitational mass," *Ann. Phys* **26**, 442 (1964); pages 466–470.

8. **Lasers and input optics—II** *by Alex Abramovici* [22 April]

W. A. Rüdiger, R. Schilling, L. Schnupp, W. Winkler, H. Billing and K. Maischberger, "A mode selector to suppress fluctuations in laser beam geometry," *Optica Acta*, **28**, 641–658 (1981).

X. W. Koechner, *Solid-State Laser Engineering* (Springer Verlag, Berlin, 1988), "Chapter 1. Introduction."

9. **Optical elements** *by Rick L. Savage* [27 April]

Y. W. Winkler, K. Danzmann, A. Rüdiger and R. Schilling, "Optical Problems in Interfereometric Gravitational Wave Antennas," in *The Sixth Marcel Grossmann*

Meeting, eds. H. Sato and T. Nakamura (World Scientific, Singapore, 1991), pp. 176–191.

Z. D. Malacara, *Optical Shop Testing* (John Wiley and Sons, New York, 1978), section 1.2, "Fizeau Interferometer," pp. 19–37.

AA. H. A. Macleod, *Thin-Film Optical Filters,* 2nd edition (Adam Hilger Ltd., Bristol, 1986), "Introduction," pp. 1–10.

10. Control systems for test-mass position and orientation *by Seiji Kawamura* [29 April]

BB. S. Kawamura and M. E. Zucker, "Mirror orientation noise in a Fabray-Perot interferometer gravitational wave detector," *Applied Optics*, in press.

CC. M. Stephens, P. Saulson, and J. Kovalik, "A double pendulum vibration isolation system for a laser interferometric gravitational wave antenna," *Rev. Sci. Instrum.*, **62**, 924–932 (1991).

DD. R. C. Dorf, *Modern Control Systems* 5th editon (Addison-Wesley, 1989): §§7.1 and 7.2 of Chapter 7, "Frequency Response Methods;" and §§8.1–8.4 of Chapter 8, "Stability in the Frequency Domain."

11. Optical topology for the locking and control of an interferometer, and signal extraction *by Martin W. Regehr* [4 May]

EE. P. W. Milonni and J. H. Eberly, *Lasers* (Wiley, New York, 1988): §§12.9 "AM Locking" and 12.10 "FM Locking," pp. 385–390.

FF. C. N. Man, D. Shoemaker, M. Pham Tu and D. Dewey, "External modulation technique for sensitive interferometric detection of displacements," *Physics Letters A*, **148**, 8–16.

GG. John H. Mɔore, Christopher C. Davis, and Michael A. Coplan, *Building Scientific Apparatus: A Practical Guide to Design and Construction* (Addison-Wesley, 1983), Sec. 6.8.3 "The lock-in amplifier and gated integrator or boxcar," (pp. 435–437).

HH. Paul Horowitz and Winfield Hill, *The Art of Electronics* (Cambridge University Press, Cambridge, 1980), Sec. 14.15 "Lock-in detection" (pp. 628–631) and an earlier section to which it refers, Sec. 9.29 "PLL components, Phase detector" (pp. 429–430).

12. Seismic isolation *by Lisa A. Sievers* [6 May]

II. Leonard Meirovitch, *Elements of Vibration Analysis* (McGraw-Hill, 1986), pp. 39–57.

JJ. R. del Fabbro, A. di Virgilio, A. Giazotto, H. Kautzky, V. Montelatici, and D. Passuello, "Three-dimensional seismic super-attenuator for low frequency gravitational wave detection," *Physics Letters A*, **124**, 253–257 (1987).

KK. C. A. Cantley, J. Hough, and N. A. Robertson, "Vibration isolation stacks for gravitational wave detectors—Finite element analysis," *Rev. Sci. Instrum.*, **63**, 2210–2219 (1992).

LL. M. Stephens, P. Saulson, and J. Kovalik, "A double pendulum vibration isolation system for a laser interferometric gravitational wave antenna," *Rev. Sci. Instrum.*, **62**, 924–932 (1991).

MM. L. Ju, D. G. Blair, H. Peng, and F. van Kann, "High dynamic range measurements of an all metal isolator using a sapphire transducer," *Mass. Sci. Technol.*, **3**, 463–470 (1992).

13&14. **Test masses and suspensions and their thermal noise** *by Aaron Gillespie* [11 May and 13 May]

NN. H. B. Callen and T. A. Welton, "Irreversibility and generalized noise," *Phys. Rev.*, **83**, 34–40 (1951).

OO. Peter R. Saulson, "Thermal noise in mechanical experiments," *Phys. Rev. D*, **42**, 2437–2445 (1990).

PP. Aaron Gillespie and Frederick Raab, "Thermal noise in mechanical experiments," *Phys. Rev. D*, **42**, 2437–2445 (1990).

QQ. Aaron Gillespie and Frederick Raab, "Thermally excited vibrations of the mirrors of a laser interferometer gravitational wave detector," unpublished (1994).

RR. Aaron Gillespie and Frederick Raab, "Suspension losses in the pendula of laser interferometer gravitational wave detectors," *Phys. Lett. A*, in press (1994).

15. **Light scattering and its control** *by Kip S. Thorne* [18 May, 1st half]

SS. J. M. Elson, H. E. Bennett, and J. M. Bennett, "Scattering from Optical Surfaces," in *Applied Optical Engineering*, Vol. VII (Academic Press 1979), Chapter 7, pp. 191–243.

16. **Squeezed light and its potential use in LIGO** *by H. Jeff Kimble* [18 May 2nd half, and 20 May]

TT. C. M. Caves, "Quantum mechanical noise in an interferometer," *Phys. Rev. D*, **23**, 1693–1708 (1981).

UU. D. F. Walls, "Squeezed states of light," *Nature*, **306**, 141–146 (1983).

VV. M. Xiao, L. A. Wu, and H. J. Kimble, "Precision measurement beyond the shot-noise limit," *Phys. Rev. Lett.*, **59**, 278–281 (1987).

17. **The physics of vacuum systems, and the LIGO vacuum system** *by Jordan Camp* [25 May]

WW. J. Moore, C. Davis, M. Coplan, *Building Scientific Apparatus* (Addison-Wesley, 1983), Chapter 3. "Vacuum technology."

18. **The 40 meter prototype interferometer as an example of many of the issues studied in this course** *by Robert E. Spero* [27 May]

XX. Rainer Weiss, "Electromagnetically coupled broadband gravitational antenna," *Quart. Prog. Rep. Res. Lab. Electron. M.I.T.* **105**, 54 (1972).

YY. Robert L. Forward, "Wideband laser-interferometer gravitational-radiation experiment," *Phys. Rev. D* **17**(2), 379–390 (1977).

CONTENTS OF VOLUME I

Note: The contents of each lecture are described below in outline form (though the lecture typically does not follow the outline sequentially). For each lecture, this volume contains: (i) a list of references prepared by the lecturer and broken down into "Assigned Reading" and "Supplementary Reading," with comments about the relevance of each reference;* (ii) a set of exercises that the reader is invited to try, as a tool to better understanding;† and (iii) the transparencies from which the lecture was delivered and/or notes on the lecture prepared by Kip, or by Malik Rakhmanov.

* Those references labeled by capital letters (A, B, C, ... Z, AA, BB, ... YY) are reproduced in Volume II; those labeled by lower-case letters are not

† The students who took this course for credit were required to work many of these exercises or do supplementary reading and write essays about it.

4. **Idealized theory of interferometers—I** *by Kip S. Thorne* [8 April]
   a. Gaussian beams and their manipulation
   b. beam splitters and mirrors
   c. simple delay-line interferometers
   d. Fabry-Perot cavities
   e. simple Fabry-Perot interferometers

5. **Idealized theory of interferometers—II** *by Ronald W. P. Drever* [13 April]
   a. Power recycled interferometers
   b. Resonantly recycled interferometers
   c. Dual (or signal) recycled interferometers
   d. Doubly resonant signal recycled interferometer
   e. Resonant sideband extraction

6. **Overview of a real interferometer** *by Stanley E. Whitcomb* [15 April]
   a. what is in a real interferometer
   b. survey of potential noise sources
   c. scaling of noise sources
   d. introduction to control systems
   e. diagnostic techniques for real interferometers

7. **Lasers and input optics—I** *by Robert E. Spero* [20 April]
   a. Fabry-Perot cavities as displacement sensors
   b. shot noise in photodetection, signal-to-noise ratio
   c. effect of mirror losses; equivalence of active and passive cavities
   d. phase modulation to sense optical phase and eliminate sensitivity to laser intensity fluctuations; sideband analysis; reflection ("Pound-Drever") locking
   e. experimental demonstration of shot noise limited sensitivity.
   f. non-recombined and recombined optical configurations; sensitivity vs. storage time, visibility, modulation waveform
   g. optimization of optical and modulation parameters

8. **Lasers and input optics—II** *by Alex Abramovici* [22 April]
   a. general requirements on light for LIGO interferometers
   b. configuration of the light source (laser, mode cleaner, other components and subsystems)
   c. Argon ion laser; its single-frequency operation and frequency prestabilization
   d. beam jitter in terms of mode superposition
   e. mode cleaner and mode matching
   f. Nd:YAG laser and frequency doubling

9. **Optical elements** *by Rick L. Savage* [27 April]
   a. overview of LIGO's requirements for mirrors, beam splitters, phase modulators, photodiodes, pick offs, etc.
   b. detailed requirements for test-mass optics:
      general requirements—total losses, reflectivity, radius of curvature, etc.
      mirror surface imperfections and how they influence the interferometer;
      contamination-induced mirror heating.
   c. the LIGO core optics pathfinder program:
      mirror substrates (mechanical quality factors, polishing, Zernike polynomials, measurement of polished surfaces);
      mirror coatings.

10. **Control systems for test-mass position and orientation** *by Seiji Kawamura* [29 April]
    a. test-mass suspension systems
    b. test-mass position and orientation damping
    c. transfer function of a pendulum
    d. sensors and actuators
    e. test mass orientation noise in a Fabry-Perot interferometer
    f. noise from the control system

11. **Optical topology for the locking and control of an interferometer, and signal extraction** *by Martin W. Regehr* [4 May]
    a. overview and explanation of the modulation methods used to extract the gravitational wave signal
    b. methods of extracting the auxiliary signals necessary for locking an interferometer
    c. analysis of multivariable control systems, with examples

12. **Seismic isolation** *by Lisa A. Sievers* [6 May]
    a. seismic background: its origin and spectrum
    b. isolation stacks: basic theory, design issues, chosen design and performance
    c. isolation via pendulum suspension; compound pendulum
    d. active isolation systems

13&14. **Test masses and suspensions and their thermal noise** *by Aaron Gillespie* [11 May and 13 May]
    a. key issues in elasticity theory
    b. fluctuation-dissipation theorm
    c. causes of losses in materials
    d. frequency dependence of noise
    e. suspension noise
    f. violin-mode noise
    g. internal-mode noise
    h. "excess" (non-Gaussian) noise
    i. choices of materials

15. **Light scattering and its control** *by Kip S. Thorne* [18 May, 1st half]
    a. how scattered light can imitate a gravitational wave; the magnitude of the danger
    b. control of scattered light by baffles and by choice of materials
    c. the chosen LIGO baffle design

16. **Squeezed light and its potential use in LIGO** *by H. Jeff Kimble* [18 May 2nd half, and 20 May]
    a. theory of squeezing
    b. practical methods of squeezing
    c. present state of the art
    d. use of squeezed vacuum state in interferometers
    e. methods to beat the standard quantum limit

17. **The physics of vacuum systems, and the LIGO vacuum system** *by Jordan Camp* [25 May]
    a. basic physics and engineering of vacuum systems
    b. noise in an interferometer due to residual gas
    c. LIGO vacuum specifications
    d. LIGO's special low-hydrogen steel
    e. outgasing and pumping strategy
    f. construction of the vacuum system

18. **The 40 meter prototype interferometer as an example of many of the issues studied in this course** *by Robert E. Spero* [27 May]
    The following abstract gives the flavor of how Spero approached this topic:
    Building gravity wave detectors like the 40 m interferometer or LIGO proceeds in two steps: constructing an array of test masses that is free from external disturbances and other sources of displacement noise, and devising a sensitive readout of the relative positions of these masses. The noise sources that constrain sensitivity can be classified as *fundamental*, meaning they were expected, ultimately, to limit the detector's sensitivity and their limits were estimated (around 1971) before the first detectors were built, and *technical*, meaning that, whether they initially were thought of or not, they are unlikely to place ultimate limits on sensitivity. Since the required sensitivity is many orders of magnitude greater than anything previously achieved, one might worry about a third class of noise sources: unanticipated fundamental phenomena revealed in the course of the R&D, which will limit the ultimate sensitivities. Luckily, no such phenomena have been discovered. The 40 m interferometer has been invaluable at sorting out which sources of noise (both fundamental and technical) are the most important in the short run and the long, and in guiding the design of LIGO. Our current understanding of the effects of imperfections in phenomena such as phase modulation, intensity stabilization, mechanical servos, and lock acquisition is due largely to investigations conducted on the 40 m and similar experimental interferometers.

Note: A tour of the 40 meter prototype interferometer was taken twice outside lecture hours: once early in the term, largely for impressionistic purposes; once at the end of the course, following up on the last lecture. A tour of the LIGO Optics laboratory in the basement of West Bridge was also taken twice: once in the middle of the term focusing on issues discussed in the term's first half; once at the end of the term, focusing on issues from the term's second half.

BATCH
START

Lecture 1

STAPLE
OR
DIVIDER

REFA

Three hundred years of gravitation

EDITED BY

S.W.HAWKING

Lucasian Professor of Mathematics, University of Cambridge

W.ISRAEL

Professor of Physics, University of Alberta, and
Senior Fellow, Canadian Institute for Advanced Research

CAMBRIDGE UNIVERSITY PRESS
Cambridge
New York New Rochelle Melbourne Sydney

1987

9

Gravitational radiation

KIP S. THORNE†

9.1 Introduction

9.1.1 The motivations for gravitational-wave research

The discovery of cosmic radio waves in the 1930s and their detailed study in the 40s, 50s, and 60s created a revolution in our view of the universe (Kellerman and Sheets, 1983; Sullivan, 1982, 1984). Previously the universe, as viewed by light, was regarded as serene and quiescent – dominated by stars and planets that wheel smoothly in their orbits, shining steadily and evolving (with few exceptions) on timescales of millions or billions of years. By contrast, the universe as viewed by radio waves was violent: galaxies in collision, jets ejected from galactic nuclei, quasars with luminosities far greater than our galaxy varying on timescales of hours, pulsars with gigantic radio beams rotating at several or many rotations per second; these were the typical strong radio emitters.

The radio revolution was so spectacular because the information carried by radio waves is so different from that carried by light. A factor 10^7 difference in wavelength meant the difference between photons predominantly thermal in origin (light) and photons predominantly nonthermal (radio), the difference between the bremsstrahlung and atomic transitions of stellar and planetary atmospheres on one hand, and the synchrotron radiation of intergalactic magnetized plasmas on the other.

As different as cosmic radio and optical radiations may be, their differences pale by comparison with those of electromagnetic waves and gravitational waves: cosmic gravitational waves should be emitted by, and carry detailed information about, coherent bulk motions of matter (e.g., collapsing stellar cores) or coherent vibrations of spacetime curvature (e.g.,

† Supported in part by the National Science Foundation (AST-8514911).

black holes). By contrast, cosmic electromagnetic waves are usually incoherent superpositions of emission from individual atoms, molecules, and charged particles. Gravitational waves are emitted most strongly in regions of spacetime where gravity is relativistic and where the velocities of bulk motion are near the speed of light. But electromagnetic waves come almost entirely from weak-gravity, low-velocity regions, since strong-gravity regions tend to be obscured by surrounding matter. Gravitational waves pass through surrounding matter with impunity, by contrast with electromagnetic waves which are easily absorbed and scattered, and even by contrast with neutrinos which, although they easily penetrate normal matter, should scatter thousands of times while leaving the core of a supernova.

These differences make it likely that, if cosmic gravitational waves can be detected and studied, they will create a revolution in our view of the universe comparable to or greater than that which resulted from the discovery of radio waves.

It might be argued that we are now so sophisticated and complete in our electromagnetically based (radio, millimeter, infrared, optical, ultraviolet, X-ray, gamma-ray, cosmic ray) understanding of the universe, compared to the optically based astronomers of the 1930s and 1940s, that a gravitational-wave revolution will be far less spectacular than was the radio revolution. It seems unlikely to me that we are so sophisticated. I am painfully aware of our lack of sophistication when I contemplate the sorry state of present estimates of the gravity waves bathing the earth (Section 9.4 below): for each type of gravity-wave source that has been studied, with the exception of binary stars and their coalescences, either (i) the strength of the source's waves for a given distance from earth is uncertain by several orders of magnitude; or (ii) the rate of occurrence of that type of source, and thus the distance to the nearest one, is uncertain by several orders of magnitude; or (iii) the very existence of the source is uncertain.

Although these uncertainties make us unhappy when we try to plan for the design and construction of gravitational-wave detectors, we will be rewarded with great surprises when gravity waves are ultimately detected and studied: the waves will give us extensive information about the universe that we are unlikely ever to obtain in any other way.

Detailed studies of cosmic gravitational waves are also likely to yield experimental tests of fundamental laws of physics which cannot be tested in any other way.

The first discovery of gravitational waves would directly verify the predictions of general relativity, and other relativistic theories of gravity, that such waves should exist. (There has already been an indirect verification, in the form of the observed inspiral of the binary pulsar due to gravitational–radiation reaction; Weisberg and Taylor, 1984; Taylor, 1987.)

By comparing the arrival times of the first bursts of light and gravitational waves from a distant supernova, one could verify general relativity's prediction that electromagnetic and gravitational waves propagate with the same speed – i.e., that they couple to the static gravity (spacetime curvature) of our Galaxy and other galaxies in the same way. For a supernova in the Virgo cluster of galaxies (15 Mpc distant), first detected optically one day after the light curve starts to rise, the electromagnetic and gravitational speeds could be checked to be the same to within a fractional accuracy (1 light day)/(15 Mpc)$=5\times 10^{-11}$.

By measuring the polarization properties of the gravitational waves, one could verify general relativity's prediction that the waves are transverse and traceless – and thus are the classical consequences of spin-two gravitons (Eardley, Lee and Lightman, 1973; Eardley *et al.*, 1973).

By comparing the detailed wave forms of observed gravitational wave bursts with those predicted for the coalescence of black-hole binaries (which will be computed by numerical relativity in the next few years, see Section 9.3.3(e) below), one could verify that certain bursts are indeed produced by black-hole coalescences – and, as a consequence, verify unequivocally the existence of black holes and general relativity's predictions of their behavior in highly dynamical circumstances. Such verifications would constitute by far the strongest test ever of Einstein's laws of gravity.

9.1.2 *A brief history of gravitational-wave research*

Einstein (1916) laid the foundations of gravitational-wave theory within months after his final formulation of general relativity – restricting himself to weak (linearized) waves emitted by bodies with negligible self-gravity and propagating through flat, empty spacetime. During the next few years Einstein (1918), Weyl (1922), and Eddington (1924) elaborated on Einstein's initial work so that by the mid-1920s the linearized theory of gravitational waves was fully understood. However, it was clear – at least to Eddington (1924) – that for sources with significant self-gravity (e.g. binary systems), the linearized analysis was invalid.

Landau and Lifshitz (1941) gave the first fairly satisfactory treatment of the emission of waves by self-gravitating systems; but a series of failed attempts to analyze radiation reaction in such systems in the late 1940s and 1950s (pp. 73 and 74 of Damour, 1983) shook physicists' faith in the ability of the waves to carry off energy, and even in the correctness of the Landau–Lifshitz formula for the emitted wave field. It required a clever thought experiment by Bondi (1957) to restore faith in the energy of the waves, and a series of beautiful and rigorous studies of the asymptotic properties of the waves at infinity by Bondi and collaborators (Bondi, 1960; Bondi, van der Burg, and Metzner, 1962; Sachs, 1962, 1963; Penrose, 1963*a*,*b*) and of the propagation of short-wavelength waves through a curved background spacetime by Isaacson (1968*a*,*b*) to restore faith that the fundamental theory of gravitational waves is soundly based.

The experimental search for cosmic gravitational waves was initiated by Joseph Weber (1960) at a time when almost nothing was known about possible cosmic sources and when nobody else had the vision to see that there were technological possibilities of ultimate success. After a decade of effort, Weber (1969) announced to the world tentative evidence that his resonant-bar gravity-wave detectors – one near Washington, DC, the other near Chicago – were being excited simultaneously by gravitational waves. There followed a six-year period of excitement and feverish effort as 15 other research groups around the world tried to construct and operate similar bar detectors (Tyson and Giffard, 1978; Amaldi and Pizella, 1979; de Sabbata and Weber, 1977; Weber, 1986, and references therein). Sadly, even with markedly improved sensitivities, these efforts gave no convincing evidence that gravity waves were actually being seen.

In parallel with this experimental effort, astrophysicists worldwide struggled through the early 1970s to milk, from electromagnetic observations of the universe and from fundamental theory, as much information as possible about the characteristics of the gravity waves that might be bathing the earth. By the mid-1970s a fuzzy but helpful picture had begun to emerge: While the sensitivities of the detectors to kilohertz-frequency bursts arriving, say, three times per year had improved during the early 70s from dimensionless amplitude $h_{3/\mathrm{yr}} \sim 1 \times 10^{-15}$ to $h_{3/\mathrm{yr}} \sim 3 \times 10^{-16}$ (a factor 10 improvement in energy flux), it seemed highly unlikely that such bursts bathing the earth would exceed $h_{3/\mathrm{yr}} \sim 1 \times 10^{-16}$; a reasonable probability of success would require $h_{3/\mathrm{yr}} \sim 10^{-20}$ or better; and a high probability would require $h_{3/\mathrm{yr}} \sim 10^{-21}$ to $h_{3/\mathrm{yr}} \sim 10^{-22}$ (Smarr, ed., 1979; Fig. 9.4 below). Although these estimates were discouraging, the theoretical efforts that produced them were making clear the enormous potential payoff that could follow the successful detection of gravity waves.

Fortunately, the experimental efforts of the early 1970s had pointed the way toward major possible detector improvements; and, consequently, although most of the first-generation experimental groups became discouraged and dropped out, a handful of highly talented groups continued onward into the 1980s with a second-generation effort involving major technological changes · such as cooling the bars to liquid-helium temperatures, changing bar materials, switching from passive to active transducers, and even developing completely new types of detectors, most notably laser-interferometer gravity-wave detectors (called 'beam' detectors in this chapter). These second-generation efforts have reached fruition in the last few years: bars with kilohertz burst sensitivities $h_{3/yr} \sim 10^{-17}$ (30 times higher in amplitude than the first-generation and 1000 higher in energy) are now collecting data in coordinated searches (Section 9.5.2(d) below); and small-scale beam detectors with $h_{3/yr} \sim 5 \times 10^{-17}$ are now operating (Section 9.5.3(d) below) as prototypes for full-scale detectors with projected ultimate sensitivities in the 10^{-22} region (Section 9.5.3(g) below). The regime of possible success has been reached, and the regimes of reasonably probable success and highly probable success look reachable – though only with vigorous continuing efforts and the expenditure of non-trivial sums of money.

In parallel with these 1980s' second-generation efforts, theorists have redoubled their struggle to firm up our understanding of the waves bathing the earth, but with only modest results: the problem of knowing what kinds of sources actually occur, and how frequently, is hampered by the paucity of electromagnetic information; and, as a result, the apparent recent improvements in our knowledge (Section 9.4 below) might be little more than changes of fashion. On the other hand, given a specific scenario for how a postulated source behaves, theorists have become far more adept than before – thanks not least to supercomputers – at computing the details of the gravitational waves it should emit (Section 9.3.3 below). As a consequence, when waves are ultimately detected, the prospects have become reasonable for deciphering from them the details of their sources.

While the present, 1987, gravity-wave searches might bring success, it is not likely they will. Thus, we must anticipate a continued vigorous effort at technology development during the coming years, with the prospects of success improving significantly at each step along the way. In parallel, we must anticipate a continuing major effort by relativity theorists to refine their ability to decipher the source behaviors corresponding to postulated gravitational-wave forms, and a continuing effort by astrophysicists to give better guidance as to what kinds of sources actually exist and in what numbers. The efforts must be great; but jointly they are likely to give an extremely valuable payoff.

9.1.3 *Overview of this chapter*

This chapter reviews all aspects of gravitational-wave research – experimental, theoretical relativity, and theoretical astrophysics. This review is intended to be readable by physicists (including advanced students) who are not specialists in general relativity, in experimental gravity, or in astrophysics.

A word of warning: in preparing this review I have *not* done a thorough literature search (I lacked the necessary energy!); nor have I cited all major original references of which I am aware (that would have made the reference list even longer than it is!). However, I have attempted to present all significant ideas and issues with which I am familiar, citing wherever possible the earliest occurrence of the idea or issue and one or more recent, thorough discussions of it.

This review is divided into four major parts: the physical and mathematical description of gravitational waves (Section 9.2), the generation and propagation of gravitational waves (Section 9.3), astrophysical sources of gravitational waves (Section 9.4), and the detection of gravitational waves (Section 9.5).

The physical and mathematical description of gravitational waves (Section 9.2) is presented in a form that compactifies and updates the corresponding material in the textbook that I coauthored fourteen years ago (Misner, Thorne, and Wheeler, 1973; cited henceforth as MTW). Emphasis focuses on the 'shortwave approximation' as a tool for defining waves mathematically (Section 9.2.1), and on measurements in the proper reference frame of an observer as a tool for defining waves physically (Section 9.2.2). A special 'TT coordinate system' is then introduced (Section 9.2.3) for use in analyzing systems large compared to a wavelength of the waves; and the energy, momentum, and quantization of gravitational waves are discussed (Section 9.2.4).

The theory of the generation and propagation of gravitational waves (Section 9.3) is far more sophisticated today than when MTW was written: we now understand how, in realistic situations, to split the problem of wave generation off from that of wave propagation, and how to mesh generation and propagation together using the technique of 'matched asymptotic expansions' (Section 9.3.1). We also understand more deeply the most elementary of all ways to compute wave generation, the 'quadrupole formalism', and its relationship to radiation reaction in the emitting system (Section 9.3.2). Unfortunately, the most strongly emitting systems should violate the mathematical approximations that underlie the quadrupole formalism and thus can be analyzed only in rough order of magnitude using it; for more accurate analyses one must use a more sophisticated wave-generation formalism. A catalog of more sophisticated formalisms is given in Section 9.3.3 – which is an updated version of a review I wrote ten years ago (Thorne, 1977). The theory of the propagation of the waves from source to earth is sketched in Section 9.3.4, and various wave-propagation effects (absorption, scattering, dispersion, tails, gravitational focusing, diffraction, parametric amplification by background curvature, non-linear coupling of waves to themselves, and generation of background curvature by the waves' energy and momentum) are described in Section 9.3.5 – which with 9.3.4 is a shortened version of my recent (Thorne, 1983), long review of wave propagation. Section 9.3 concludes with very brief descriptions of elegant mathematical work on idealized wave-propagation situations: the asymptotic structure of waves propagating toward 'future null infinity' in an asymptotically flat spacetime (Section 9.3.6), and exact, analytic solutions to the Einstein equations for wave generation and propagation (Section 9.3.7).

It is nearly a decade since a detailed and thorough review has been written of astrophysical sources and detectors for gravitational waves (Smarr, ed., 1979; Douglass and Braginsky, 1979); and in the intervening time these topics have changed enormously. (For recent reviews of a number of subtopics see the chapters in Deruelle and Piran, 1983.) Sections 9.4 and 9.5 attempt a comprehensive review in a manner that closely ties the sources to the detection efforts. In these sections the sources and the detection strategies are split up into three categories: those for gravitational-wave 'bursts' (Section 9.4.1 and Fig. 9.4); those for periodic gravitational waves (Section 9.4.2 and Fig. 9.6); and those for a stochastic gravitational-wave background (Section 9.4.3 and Fig. 9.7). All previous reviews have been cavalier about factors of 2 in the definitions of wave strengths and detector sensitivities. This review tries to standardize the definitions, including factors of 2. The standardization is based on signal-to-noise-ratio analyses that are given at the beginnings of Sections 9.4.1, 9.4.2, and 9.4.3. For each source that looks favorable for wave detection, Section 9.4 gives a description of our current state of knowledge of the source and gives current estimates of the wave strengths and other wave characteristics.

The burst sources treated in Section 9.4.1 include supernovae (collapse of a normal stellar core to form a neutron star, subsection c), the collapse of a star or star cluster to form a black hole (subsection d), the inspiral and coalescence of compact binaries (neutron stars and black holes, subsection e), and the fall of stars and small holes into supermassive holes (subsection f). Because our knowledge of sources is so poor, it is useful to estimate how strong the strongest wave bursts bathing the earth could be without violating our cherished beliefs about the laws of physics and the nature of the universe; this is done in subsection g. The periodic sources treated in Section 9.4.2 include rotating neutron stars (rigidly rotating pulsars, and neutron stars spun up by accretion until they encounter a radiation-reaction-driven instability, subsection b), and binary stars (including unevolved binaries, WUMa stars, white-dwarf binaries, and neutron-star binaries, subsection c). The stochastic sources in Section 9.4.3 include large numbers of binary stars whose waves superpose stochastically (subsection b), pre-galactic, Population III stars (subsection c), the big-bang singularity in which the universe began – with subsequent parametric amplification of its waves by background curvature in inflationary and other scenarios (subsection d), phase transitions in the subsequent but still early universe (subsection e), and cosmic strings produced by phase transitions (subsection f). Present estimates of the strengths of the waves from all these sources are shown in Figs. 9.4 (burst), 9.6 (periodic), and 9.7 (stochastic) along with the sensitivities of present and proposed detectors.

The detectors in Section 9.5 are divided into those that operate in the high-frequency regime, $f \gtrsim 10$ Hz (Sections 9.5.2, 9.5.3, and 9.5.4), those for low frequencies, 10 Hz $\gtrsim f \gtrsim 10^{-5}$ Hz (Section 9.5.5), and those for very low frequencies, $f \lesssim 10^{-5}$ Hz (Section 9.5.6). The high-frequency detectors are all earth-based; but because of seismic and gravity-gradient noise, the low- and very-low-frequency detectors must be space-based. Sections 9.5.2 and 9.5.3 describe in great detail the earth-based, high-frequency bar and beam detectors which have been under development for many years and show great promise for the future. Section 9.5.4 describes briefly other types of earth-based, high-frequency detectors. Section 9.5.5 describes low-frequency detectors including doppler tracking of spacecraft (subsection a), beam detectors in space which hold great promise for the turn of the century (subsection b), the normal modes of the earth and sun (subsections c and d), the vibrations of blocks of the earth's crust (subsection e), and the earth-orbiting skyhook (subsection f). Section 9.5.6 describes very-low-frequency detectors including the timing of pulsars (neutron-star rotations), which recently has placed interesting observational limits on a stochastic background (subsection a), the timing of orbital motions of binary and planetary systems (subsection b), and astronomical observations of anisotropies in the temperature of the cosmic microwave radiation (subsection c), deviations from the Hubble flow (subsection d), and products of primordial nucleosynthesis (subsection d).

9.1.4 Notation and conventions

Throughout this chapter, unless otherwise stated, we shall assume that general relativity correctly describes classical gravitational waves. Our notation will be that of Misner, Thorne, and Wheeler (1973) – cited as MTW throughout – including, e.g., the use of Greek indices for spacetime (running from 0 to 3) and Latin for space (running from 1 to 3), the use of commas for partial derivatives and semicolons for covariant derivatives, the use of the Einstein summation convention, and the use of geometrized units in which Newton's gravitation constant G and the speed of light c are set equal to unity. Sometimes, particularly when discussing gravitational-wave detectors, we shall restore the Gs and cs to the equations and use cgs units.

9.2 The physical and mathematical description of a gravitational wave

9.2.1 Shortwave approximation

General relativistic gravitational waves are ripples in the curvature of spacetime that propagate with the speed of light. Because gravity is nonlinear, it is not possible in a fully precise manner to separate the contributions of gravitational waves to the curvature from the contributions of the earth, the sun, the galaxy, or anything else; and since such a separation underlies the very concept of a gravitational wave, this means that gravitational waves are not precisely defined entities.

On the other hand, in realistic astrophysical situations the lengthscale $\lambdabar$ on which the waves vary (their reduced wavelength, $\lambdabar = \lambda/2\pi$) is very short compared to the lengthscales $\mathscr{L}$ on which all other important curvatures vary; and this difference in lengthscale makes possible a high-accuracy, but approximate, split of the Riemann curvature tensor $R_{\alpha\beta\gamma\delta}$ into a 'background curvature' $R^{\mathrm{B}}_{\alpha\beta\gamma\delta}$ plus a contribution $R^{\mathrm{GW}}_{\alpha\beta\gamma\delta}$ due to gravitational waves: the background $R^{\mathrm{B}}_{\alpha\beta\gamma\delta}$ is the average of $R_{\alpha\beta\gamma\delta}$ over several wavelengths

$$R^{\mathrm{B}}_{\alpha\beta\gamma\delta} \equiv \langle R_{\alpha\beta\gamma\delta} \rangle; \tag{1a}$$

and the waves' curvature $R^{\mathrm{GW}}_{\alpha\beta\gamma\delta}$ is the rapidly varying difference

$$R^{\mathrm{GW}}_{\alpha\beta\gamma\delta} \equiv R_{\alpha\beta\gamma\delta} - R^{\mathrm{B}}_{\alpha\beta\gamma\delta}. \tag{1b}$$

Heuristically, $R^{\mathrm{B}}_{\alpha\beta\gamma\delta}$ is like the large-scale (10 cm) curvature of an orange, while $R^{\mathrm{GW}}_{\alpha\beta\gamma\delta}$ is like the fine-scale granulation of the skin of the orange (lengthscale a few millimeters).

This method of defining a gravitational wave, introduced into general relativity by Wheeler (1955) and Power and Wheeler (1957), is a special case of a standard technique in mathematical physics variously called 'shortwave approximation' or 'two-timing' or 'two-lengthscale expansion' or 'two-variable expansion' (see e.g. Chapter 3 of Cole, 1968); and it is intimately connected to the 'WKB approximation'. There is an elegant shortwave-approximation formalism for gravitational-wave theory due largely to Brill and Hartle (1964) and to Isaacson (1968*a*,*b*). Not surprisingly the formalism reveals that general relativistic gravitational waves propagate through vacuum in essentially the same manner as light – with the same speed, with the same changes of amplitude due to curvature of the wavefronts, with the same diffraction effects when focussed by a gravitational lens, etc. Because that formalism has been reviewed extensively elsewhere (e.g. Thorne, 1983), I shall not delve into it here, except for a brief description in Section 9.3.4 and an enumeration and brief discussion of some of its predictions in Section 9.3.5 below.

9.2.2 Measurements in the proper reference frame of an observer

In general relativity the Riemann curvature tensor is defined, operationally, by the relative accelerations it produces between adjacent particles (e.g. the 'equation of geodesic deviation', Sections 8.7 and 11.3 of MTW). Correspondingly, a gravitational wave can be defined operationally in the following way:

Consider an observer (freely falling or accelerated, it doesn't matter so long as the acceleration is slowly varying). Let the observer carry with herself a small Cartesian latticework of measuring rods and synchronized clocks (a 'proper reference frame' in the sense of Section 13.6 of MTW, with spatial coordinates x^j that measure proper distance along orthogonal axes). She is to measure the 'force of gravity' F_j that acts on a particle of mass m, momentarily at rest at location x^j. For example, she might let the particle fall freely, measure its acceleration g_j in her proper reference frame, and multiply it by the particle's mass m to get the force $F_j = mg_j$. Alternatively, she might measure the force required to hold the particle fixed in the coordinate grid of her proper reference frame and equate F_j to minus that force. Of course, this is not different in any way from what she would do were she a Newtonian physicist rather than a relativistic physicist. The difference lies in how she analyzes and thinks about the force of gravity F_j. As a relativist, she recognizes (cf. Box 37.1 of MTW) that F_j is made up of a nearly steady, nearly position-independent component caused by her own failure to fall freely, plus a component proportional to the particle's Cartesian coordinate position x^j (relative acceleration of particle and origin of coordinates) which is caused by spacetime curvature $R_{\alpha\beta\gamma\delta}$. This latter contribution,

$$F_j = -mR_{j0k0}x^k \tag{2}$$

(where the index 0 denotes a component along her time basis vector), she splits up into a piece that changes slowly in time (background curvature contribution) plus a piece that is rapidly varying. She makes certain that there are no rapidly moving or rapidly changing nearby sources of gravity to account for the rapid variations; if there are none, then she can attribute the rapidly varying component of the force to gravitational waves

$$F_j^{\mathrm{GW}} = -mR^{\mathrm{GW}}_{j0k0}x^k. \tag{3}$$

It is conventional to use, as the primary entity for describing a gravitational wave, not the Riemann curvature tensor $R^{\mathrm{GW}}_{\alpha\beta\gamma\delta}$ which has dimensions 1/time^2 (or 1/length^2), but rather a dimensionless 'gravitational-wave field' h^{TT}_{jk}. In terms of the force-producing components of the Riemann tensor $R^{\mathrm{GW}}{}_{j0k0}$ and proper time t as measured by our observer, h^{TT}_{jk} is defined by (cf. Section 2.3 of Thorne 1983)

$$\frac{\partial^2 h^{\mathrm{TT}}_{jk}}{\partial t^2} \equiv -2R^{\mathrm{GW}}_{j0k0}. \tag{4}$$

The convenience of this gravitational-wave field lies in its simple relationship to displacements produced by the waves: if, for simplicity, the observer is freely falling and keeps the axes of her coordinate grid tied to gyroscopes and is in a region of spacetime where gravitational waves are the only source of spacetime curvature, then the waves will produce tiny oscillatory changes δx^j in the position of a test particle relative to the origin of her coordinate grid; and these changes will satisfy the equation of motion ('equation of geodesic deviation')

$$m\frac{\mathrm{d}^2\,\delta x^j}{\mathrm{d}t^2} = F_j^{\mathrm{GW}} = -mR^{\mathrm{GW}}_{j0k0}x^k = \frac{1}{2}\,m\,\frac{\partial^2 h^{\mathrm{TT}}_{jk}}{\partial t^2}\,x^k. \tag{5}$$

Because any realistic wave is so weak that the oscillatory changes δx^j are miniscule compared to the distance of the particle from the origin, x^k can be regarded as essentially constant on the right-hand side, and equation (5) can then be integrated easily to give

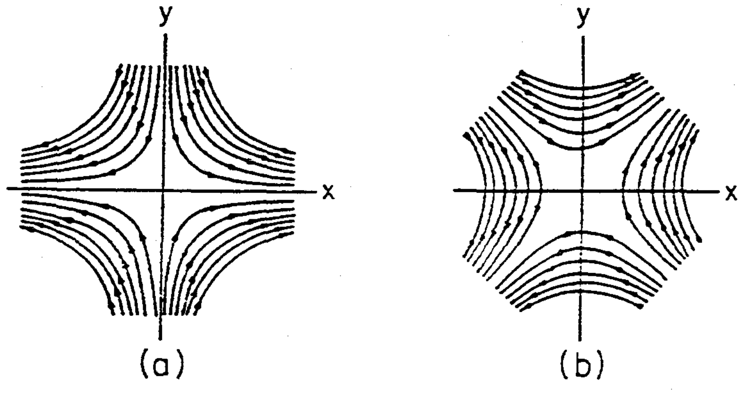
$$\delta x^j = \frac{1}{2} h_{jk}^{\mathrm{TT}} x^k. \tag{6}$$

Thus, aside from a factor $\frac{1}{2}$, h_{jk}^{TT} plays the role of the 'dimensionless strain of space', or the 'time-integrated shear of space': it is the ratio of the wave-induced displacement of a free particle relative to the origin, to its orginal displacement from the origin. (In the above equations and below it does not matter whether a spatial index is up or down, since the spatial coordinates are Cartesian.)

The superscript TT on the gravitational-wave field is to remind us that, according to general relativity, the field is 'transverse and traceless'. More specifically: the Einstein field equations guarantee that $R_{\alpha\beta\gamma\delta}^{\mathrm{GW}}$ and hence also h_{jk}^{TT} propagate with the speed of light. Since the sources of cosmic gravity waves are very far away, the waves look very nearly planar as they pass through the observer's proper reference frame. If we orient the x, y, z spatial axes so the waves propagate in the z direction, then the 'transversality' of the waves means that the only non-zero components of the wave field are h_{xx}^{TT}, $h_{xy}^{\mathrm{TT}} = h_{yx}^{\mathrm{TT}}$, and h_{yy}^{TT}; and the 'trace-free' property means that $h_{xx}^{\mathrm{TT}} = -h_{yy}^{\mathrm{TT}}$. Thus, the gravitational waves, like electromagnetic waves, have only two independent components – two polarization states.

Because of their TT nature, gravitational waves produce a quadrupolar, divergence-free force field (equation (5) and Fig. 9.1). This force field has two components corresponding to the two polarization states of the waves: the

Fig. 9.1. Lines of force for gravitational waves (equations (5) and (7)): (*a*) with '+' polarization, $m\delta\ddot{x} = \frac{1}{2}\ddot{h}_+ x$ and $m\delta\ddot{y} = -\frac{1}{2}\ddot{h}_+ y$; and (*b*) with '×' polarization, $m\delta\ddot{x} = \frac{1}{2}\ddot{h}_\times y$ and $m\delta\ddot{y} = \frac{1}{2}\ddot{h}_\times x$ – where dots denote time derivatives.



quantity

$$h_+ \equiv h^{TT}_{xx} = -h^{TT}_{yy} \tag{7a}$$

produces a force field with the orientation of a '+' sign, while

$$h_\times \equiv h^{TT}_{xy} = h^{TT}_{yx} \tag{7b}$$

produces one with the orientation of a '×' sign. Thus, h_+ and $h_\times$ are called the 'plus' and 'cross' (or + and ×) gravity-wave amplitudes. From these amplitudes and the polarization tensors $e^+_{xx} = -e^+_{yy} = 1$, $e^\times_{xy} = e^\times_{yx} = 1$ (all other components zero), one can reconstruct the full wave field

$$h^{TT}_{jk} = h_+ e^+_{jk} + h_\times e^\times_{jk}. \tag{7c}$$

It is straightforward to show that, if one rotates the x and y axes in the transverse plane through an angle $\Delta\psi$, the gravity-wave amplitudes are changed to

$$\begin{aligned} h^{\text{new}}_+ &= h^{\text{old}}_+ \cos 2\Delta\psi + h^{\text{old}}_\times \sin 2\Delta\psi, \\ h^{\text{new}}_\times &= -h^{\text{old}}_+ \sin 2\Delta\psi + h^{\text{old}}_\times \cos 2\Delta\psi. \end{aligned} \tag{7d}$$

The quadrupolar symmetry of the force field, together with the tenets of canonical field theory, tells us that general relativistic gravitational waves must be associated with quanta of spin two ('gravitons'). The spin is always the ratio of 360 degrees to the angle, about the propagation direction, through which one must rotate an instantaneous field to make it return to its original state – the 'return angle'. For electromagnetic waves the return angle is 360 degrees and the spin is one; for general relativistic gravitational waves the return angle is 180 degrees and the spin is two. In other relativistic theories of gravity the wave field has other symmetries and therefore other spins for its quanta – and in most relativistic theories, by contrast with general relativity, the symmetries are not even frame-invariant (local Lorentz-invariant); and, as a result, the waves of those theories cannot be incorporated into canonical field theory and cannot be quantized by canonical techniques. For details see Eardley, Lee and Lightman (1973), and Eardley *et al.* (1973).

The above definition of the gravitational-wave field h^{TT}_{jk} relies on a specific choice of reference frame. It turns out that, if one pursues this definition in two different reference frames related by a boost in some arbitrary direction, and if one chooses the spatial axes of the two frames so they are unrotated relative to each other, then the instantaneous wave fields h^{TT}_{jk} (and correspondingly h_+ and $h_\times$) will be the same. Stated more precisely – but in a language I shall not explain – the general-relativistic gravitational-wave field h^{TT}_{jk} has 'boost-weight zero' (it is a scalar under boosts); but, as discussed above, it has 'spin-weight two' (it behaves like a spin-two field under rotations). For further details see Section 2.3.2 of Thorne (1983), and for the mathematics that goes along with the concepts of 'spin-weight' and 'boost-weight', see Geroch, Held and Penrose (1973).

9.2.3 *TT coordinate system*

The above description of the force law and force fields produced by a gravitational wave is just the leading order in a power series expansion in distance $r=(\delta_{jk}x^jx^k)^{\frac{1}{2}}$ from the origin of the observer's proper reference frame. The higher-order fractional corrections to the forces are of order $(r/\bar{\lambda})^2$ and, correspondingly, the proper reference frame's coordinates x^j fail to measure proper distance by fractional amounts of order $h^{\mathrm{TT}}_{jk}(r/\bar{\lambda})^2$ (see, e.g. Zhang, 1986, for discussion in the case of a freely falling observer). Thus, the above description of gravity-wave forces is accurate only if the region of interest is spatially small compared to a reduced wavelength. When the region is large, an alternative description is needed.

The nicest alternative description makes use of a 'TT coordinate system' (Section 35.4 of MTW), i.e. a coordinate system which is nearly Minkowski throughout the spacetime region of interest, and in which the contribution of the waves to the deviations from the Minkowski metric is embodied in the same h^{TT}_{jk} as we introduced above.

Because the TT coordinates must be nearly Minkowski, they cannot cover a spacetime region that is too large: their extent in both time and space must be far smaller than the background radius of curvature $\mathscr{R}_{\mathrm{B}}\sim|R^{\mathrm{B}}_{\alpha\beta\gamma\delta}|^{-\frac{1}{2}}$. In typical situations, this limitation is very mild compared to the limitation on the size of an observer's proper reference frame: one can stretch TT coordinates over any region small compared to the Hubble distance (cutting out holes in the vicinities of black holes and neutron stars), but if the waves have a frequency of a kilohertz, then any proper reference frame used to study their forces must be small compared to $\bar{\lambda}\sim 50$ km.

In a TT coordinate system the spacetime metric coefficients take the form

$$g_{\alpha\beta}=\eta_{\alpha\beta}+h^{\mathrm{B}}_{\alpha\beta}+h^{\mathrm{TT}}_{\alpha\beta}, \tag{8}$$

where $\eta_{\alpha\beta}$ are the Minkowski metric coefficients (diagonal -1, $+1$, $+1$, $+1$), $h^{\mathrm{B}}_{\alpha\beta}$ is the background metric perturbation which varies on a long lengthscale $\mathscr{L}$, and $h^{\mathrm{TT}}_{\alpha\beta}$ is the gravity-wave metric perturbation which varies on the short lengthscale $\bar{\lambda}$. The time-time and space-time components of the gravity-wave perturbation h^{TT}_{00} and $h^{\mathrm{TT}}_{0j}=h^{\mathrm{TT}}_{j0}$ vanish, and the space-space components h^{TT}_{jk} are the same as the gravity-wave field that would be computed from the Riemann tensor (equation (4)) in the proper reference frame of any observer who is nearly at rest in the TT coordinate system. For proofs and discussions, see e.g. Sections 35.4, 37.1, and 37.2 of MTW.

Whereas physics in a proper reference frame can be formulated in the Newtonian language of three-dimensional forces, including gravitational forces, physics in a TT coordinate system must be formulated in the relativistic language of geodesic motion and vanishing divergence of the stress-energy tensor; cf. Section 9.5.1 below.

9.2.4 Energy, momentum, and quantization of gravitational waves

In the 1940s and early 1950s a controversy raged over whether or not gravitational waves can carry energy. The controversy was ultimately resolved by Herman Bondi (1957) using a simple thought experiment: place several beads on a rough stick and let a gravitational wave pass. The above description of the forces produced by the waves (which was first understood fully by Bondi and by Felix Pirani 1956 and their colleagues) guarantees that – if the stick is not *too* rough – the wave will push the beads back and forth on the stick, heating it. Surely, if the wave can heat a stick, it must carry energy.

It was not until the late 1960s that a fully satisfactory mathematical description of the energy in a gravity wave was devised: Richard Isaacson (1968*a*,*b*), using the shortwave approximation which he had developed in detail on the basis of cruder earlier work of Wheeler (1955), Power and Wheeler (1957) and Brill and Hartle (1964), introduced a stress-energy tensor $T^{\mathrm{GW}}_{\alpha\beta}$ for gravitational waves. This stress-energy tensor, like the background curvature $R^{\mathrm{B}}_{\alpha\beta\gamma\delta}$, is smooth on the lengthscale $\lambdabar$; it is obtained, in fact, by averaging the squared gradient of the wave field over several wavelengths:

$$T^{\mathrm{GW}}_{\alpha\beta}=\frac{1}{32\pi}\sum_{i,j}\langle h^{\mathrm{TT}}_{ij,\alpha}h^{\mathrm{TT}}_{ij,\beta}\rangle, \tag{9}$$

where $\langle\cdots\rangle$ means 'average over several wavelengths'. (For a pedagogical derivation and discussion see Sections 35.7–35.15 of MTW; for a beautiful rederivation by the method of averaged Lagrangians see MacCallum and Taub, 1973.) If the waves are propagating in the *z* direction, this stress-energy tensor takes the standard form for a bundle of zero-rest-mass particles (gravitons) moving at the speed of light in the *z* direction:

$$T^{\mathrm{GW}}_{00}=-T^{\mathrm{GW}}_{0z}=-T^{\mathrm{GW}}_{z0}=T^{\mathrm{GW}}_{zz}=\frac{1}{16\pi}\langle(\partial h_{+}/\partial t)^2+(\partial h_{\times}/\partial t)^2\rangle. \tag{10}$$

In order of magnitude, restoring the factors of G and c, the energy flux in the waves if they have frequency $f = c/2\pi\lambda\!\!\!^{-}$ is

$$-T^{\rm GW}_{0z} \simeq \frac{\pi}{4}\frac{c^3}{G} f^2 \langle h_+^2 + h_\times^2 \rangle = 320\,\frac{\rm erg}{\rm cm^2\,s}\left(\frac{f}{1\ \rm kHz}\right)^2 \left\langle \frac{h_+^2 + h_\times^2}{(10^{-21})^2} \right\rangle. \qquad (11)$$

The numbers in this equation correspond to a strongly emitting supernova in the Virgo cluster of galaxies, where there are several supernovae per year. Contrast this huge gravity-wave energy flux with the peak electromagnetic flux at the height of the supernova, $\sim 10^{-9}$ erg cm^{-2} s^{-1}; but note that the gravity waves should last for only a few milliseconds, while the strong electromagnetic output lasts for days.

Corresponding to the huge energy flux (11) in an astrophysically interesting gravitational wave is a huge occupation number for the quantum states of the gravitational-wave field: it is not hard to show that for the above supernova burst only a handful of quantum states are occupied; and they each contain $n \sim 10^{75}$ gravitons (equations (6)–(8) of Thorne *et al.*, 1979). This means that the waves behave exceedingly classically; quantum-mechanical corrections to the classical theory have fractional magnitude $1/\sqrt{n} \sim 10^{-37}$. (Although the full quantization of the gravitational field is exceedingly difficult and not yet fully under control, the quantization of weak gravitational waves propagating through a smooth background spacetime – equivalent to weak waves in flat spacetime – has been well understood for decades; see, e.g., the most elementary aspects of Feynman, 1963; Dewitt, 1967*a*,*b*.)

Isaacson's stress-energy tensor (9) for gravitational waves has the same properties and plays the same role as the stress-energy tensor for any other field or form of matter in the background spacetime. For example, $T^{\rm GW}_{\alpha\beta}$ generates background curvature through the Einstein field equations (averaged over several wavelengths of the waves); also $T^{\rm GW}_{\alpha\beta}$ has vanishing divergence (conservation of gravity-wave energy and momentum) in spacetime regions where the waves are not being generated, absorbed, or scattered. For full details see Isaacson (1968*b*) or Section 35.15 of MTW.

9.3 The generation and propagation of gravitational waves

9.3.1 Wave propagation split off from wave generation

Turn, now, to the generation of gravitational waves and their propagation from their source to the earth. Mathematically, the wave-generation problem and the wave-propagation problem are each difficult – though for very different reasons, so that to handle the difficulties requires two very different sets of mathematical tools and approximations. For ease of analysis, then, it is important to split the propagation problem off from the generation problem and treat them separately. This is accomplished by dividing the space around the source into three regions (Section III of Thorne, 1980*b*): a 'wave-generation region' at distances from the source $r \lesssim r_I =$ ('Inner radius'); a 'local wave zone' at distances $r_I \lesssim r \lesssim r_O =$ ('Outer radius'); and a 'distant wave zone' at distances $r \gtrsim r_O$. The theory of wave generation is developed with one set of mathematical tools in the wave-generation region and the local wave zone, i.e. at distances $r \lesssim r_O$; the theory of propagation to earth is developed with the other set of tools in the local wave zone and the distant wave zone, i.e. at distances $r \gtrsim r_I$; and the two theories are matched together in their domain of overlap, the local wave zone $r_I \lesssim r \lesssim r_O$.

The inner radius r_I is far enough out to be in the wave zone, $r_I \gg \bar{\lambda}$; far enough to be in a region where the source's gravity is weak, $r_I \gg 2M \equiv$ (Schwarzschild radius of source) $= 2 \times$ (mass of source); and far enough to be outside the source, $r_I \gg L \equiv$ (size of source). The outer radius r_O is far enough beyond the inner radius to leave many wavelengths in the local wave zone, $r_O - r_I \gg \bar{\lambda}$; but not so far that the gravitational redshift can produce a significant net phase shift during propagation through the local wave zone, $\delta\phi = (M/\bar{\lambda})\ln(r_O/r_I) \ll 1$; and not so far that the background curvature of the external universe can significantly affect the propagation, $r_O - r_I \ll \mathscr{R}_B \equiv |R^B_{\alpha\beta\gamma\delta}|^{-\frac{1}{2}} \equiv$ (radius of curvature of background spacetime). These choices of r_I and r_O permit one to ignore, in the local wave zone $r_I \lesssim r \lesssim r_O$, the background curvature both of the source and of the external universe; i.e. they permit one to regard the waves in the local wave zone as propagating through flat spacetime. This greatly simplifies calculations.

For a given, astrophysically interesting source the wave-generation task consists of computing with reasonable accuracy the dynamical behavior of the source's gravitational field in the wave-generation region $r \lesssim r_I$, and further computing how that dynamical curvature develops into outward propagating waves in the local wave zone, $r_I \lesssim r \lesssim r_O$. Once this is done, the theorist can switch tasks and mathematical formalisms from wave generation to wave propagation. The wave-propagation task takes as input the propagating waves in the local wave zone and carries them outward (typically using the shortwave approximation and formalism) through the universe from source to earth.

As an addendum to the wave-generation task, one often computes the back-reaction effects of the wave emission on the source, i.e. the 'radiation reaction'.

9.3.2 The quadrupole formalism for wave generation and radiation reaction

Of all the techniques for computing wave generation, one, the 'quadrupole formalism', is especially important because it is highly accurate for many sources and is accurate in order of magnitude for most. The quadrupole formalism was derived originally by Einstein (1916, 1918) for sources with negligible self-gravity and slow internal motions. Later, in a series of steps by Landau and Lifshitz (1941), Fock (1959), Ipser (1971) and Thorne (1980*b*, Sections VII and XII), it became clear that the quadrupole formalism requires for high accuracy no constraints whatsoever on the strength of the source's internal gravity; all that is required is slow motion – more specifically, that the source's size L be small compared to the reduced wavelength $\lambdabar$ of the waves it emits. In other words, the quadrupole formalism is to gravity-wave generation what the 'poor-antenna approximation' (dipole formalism) is to radio-wave generation. Moreover, as in the radio-wave problem, the quadrupole formalism typically is accurate to within factors of order 2 even for sources with sizes of order a reduced wavelength $\lambdabar$ (cf. equation (4.8) of Thorne, 1980*b*); and since very few astrophysical systems are larger than a reduced wavelength of the waves they emit, this justifies the use of the quadrupole formalism for order-of-magnitude astrophysical estimates in almost all situations.

The quadrupole formalism writes the gravitational-wave field in the source's local wave zone (where the background curvature can be ignored) in the following simple form:

$$h_{jk}^{\mathrm{TT}} = \frac{2}{r}\frac{\partial^2}{\partial t^2}\,[\mathscr{I}_{jk}(t-r)]^{\mathrm{TT}}. \qquad (12)$$

Here r is the distance to the source's center, t is proper time as measured by an observer at rest with respect to the source, $t-r$ is retarded time, the superscript TT means 'algebraically project out and keep only the part that is transverse to the (radial) direction of propagation and is traceless' (Box 35.1 of MTW), and $\mathscr{I}_{jk}(t-r)$ is the source's mass quadrupole moment evaluated at the retarded time $t-r$.

The meaning of 'mass quadrupole moment' is well known when the source has weak internal gravity and small internal stresses, so Newtonian gravity is a good approximation to general relativity inside and near the source. Then $\mathscr{I}_{jk}$ is the symmetric, trace-free (STF) part of the second moment of the source's mass density ρ, as computed in a Cartesian coordinate system centered on the source:

$$\mathscr{I}_{jk}(t)=\left[\int\rho(t)x^jx^k\,\mathrm{d}^3x\right]^{\mathrm{STF}}=\int\rho(t)\left[x^jx^k-\frac{1}{3}r^2\delta_{jk}\right]\mathrm{d}^3x; \qquad (13a)$$

equivalently, it is the coefficient of the $1/r^3$ part of the source's Newtonian gravitational potential

$$\Phi=-\frac{M}{r}-\frac{3}{2}\frac{\mathscr{I}_{jk}(t)x^jx^k}{r^5}-\frac{5}{2}\frac{\mathscr{I}_{ijk}x^ix^jx^k}{r^7}+\cdots. \qquad (13b)$$

If the source has strong internal gravity, one can no longer express its mass quadrupole moment as the simple integral (13a). However, so long as the source has slow internal motions ($L\ll\lambdabar$), there will be a region of space far enough from the source to be in vacuum ($r>L$) and far enough for gravity to be weak ($r\gg 2M$), yet near enough for retardation and wave behavior to be unimportant ($r\ll\lambdabar$). In this 'weak-field, vacuum, near-zone' gravity can be described with high accuracy as Newtonian; and the Newtonian potential is $\Phi\cong-\frac{1}{2}(g_{00}+1)$, where g_{00} is the time-time part of the metric in a coordinate system that is as Minkowski as possible throughout the weak-field, vacuum near zone. The mass quadrupole moment can then be read off this Newtonian potential using the standard formula (13b). For further discussion see the review in Section 3 of Thorne (1983) or the original treatment in Part Two of Thorne (1980*b*).

For order-of-magnitude calculations of h_{jk}^{TT} (equation (12)), one can approximate the TT part of the second time derivative of the mass quadrupole moment by that portion of the source's internal kinetic energy which is associated with non-spherical motions, $E_{\mathrm{kin}}^{\mathrm{ns}}$; and, unless the source is at very large cosmological redshifts $z\gg 1$, one can propagate the waves to earth as though the intervening spacetime were flat – with the result that the local wave-zone formula for the waves, equation (12), is valid also at earth. The result is the simple order-of-magnitude formula

$$h\sim\frac{E_{\mathrm{kin}}^{\mathrm{ns}}}{r} \qquad (14)$$

for the magnitude h of the gravity-wave field h_{jk}^{TT} at earth.

From the 'exact' quadrupole formula (12) for the wave field in the local wave zone and Isaacson's formula (9) for the stress-energy tensor of the waves, one can compute the fluxes of energy and angular momentum carried by the waves. By integrating those fluxes over a sphere surrounding the source in the local wave zone, one obtains for the rates of emission of energy (Einstein, 1916, 1918) and angular momentum (Peters, 1964)

$$\frac{\mathrm{d}E^{\mathrm{GW}}}{\mathrm{d}t}=\frac{1}{5}\sum_{j,k}\left\langle\left(\frac{\mathrm{d}^3\mathscr{I}_{jk}}{\mathrm{d}t^3}\right)^2\right\rangle, \tag{15}$$

$$\frac{\mathrm{d}J_i^{\mathrm{GW}}}{\mathrm{d}t}=\frac{2}{5}\sum_{j,k,l}\varepsilon_{ijk}\left\langle\frac{\partial^2\mathscr{I}_{jl}}{\partial t^2}\frac{\partial^3\mathscr{I}_{lk}}{\partial t^3}\right\rangle. \tag{16}$$

The linear momentum carried off by the waves vanishes when one computes it by the quadrupole formalism; but when one includes higher-order corrections to the field emitted by slow-motion sources, one finds for the rate of emission of linear momentum (first derived by Papapetrou, 1962, 1971; for a more modern derivation in the notation of this chapter see Section IV.C of Thorne, 1980*b*)

$$\frac{\mathrm{d}P_i^{\mathrm{GW}}}{\mathrm{d}t}=\frac{2}{63}\sum_{j,k}\left\langle\frac{\partial^3\mathscr{I}_{jk}}{\partial t^3}\frac{\partial^4\mathscr{I}_{jki}}{\partial t^4}\right\rangle+\frac{16}{45}\sum_{j,k,a}\varepsilon_{ijk}\left\langle\frac{\partial^3\mathscr{I}_{ja}}{\partial t^3}\frac{\partial^3\mathscr{S}_{ka}}{\partial t^3}\right\rangle. \tag{17}$$

Here $\mathscr{I}_{ijk}$ is the source's 'mass octupole moment' and $\mathscr{S}_{ij}$ is its 'current quadrupole moment' (gravitational analog of magnetic quadrupole moment). For sources with weak internal gravity and stresses (nearly Newtonian sources), these moments are computable from the simple volume integrals

$$\mathscr{I}_{ijk}=\left(\int\rho x^i x^j x^k\,\mathrm{d}^3x\right)^{\mathrm{STF}}, \tag{18a}$$

$$\mathscr{S}_{ij}=\left(\int\rho\varepsilon_{ipq}x^p v^q x^j\,\mathrm{d}^3x\right)^{\mathrm{STF}}, \tag{18b}$$

where, as in equation (15a), STF means 'make it symmetric and trace-free', i.e. 'symmetrize on all free indices and remove the traces on all pairs of free indices'. Independently of the strengths of the internal gravity and stresses, the moments can be read off the Newtonian potential $\Phi\cong-\frac{1}{2}(g_{00}+1)$ (equation (13b)) and off the 'gravitomagnetic potential' $\beta_i\equiv g_{0i}$ in the source's weak-field near zone:

$$\beta_i=-2\varepsilon_{ipq}\frac{J^p x^q}{r^3}-4\varepsilon_{ipq}\frac{\mathscr{S}_{pa}x^q x^a}{r^5}-\cdots. \tag{19}$$

In equation (19) J^p, the moment in the leading, dipolar term, is the source's angular momentum. For further details, discussions, and derivations see Thorne (1983) or Thorne (1980*b*). For a discussion of the gravitomagnetic potential see, e.g., Chapter 3 of Thorne, Price and Macdonald (1986).

The laws of conservation of energy, angular momentum, and linear momentum imply that radiation reaction should deplete the source's energy, angular momentum, and linear momentum at just the right rates to compensate for the losses (15), (16), and (17); and a detailed analysis of the radiation reaction forces reveals that this is so (Peres, 1960). Particularly convenient in analyzing the radiation reaction in a source with weak self-gravity is a Newtonian-type radiation-reaction potential (Burke, 1969; Thorne, 1969; Chandrasekhar and Esposito, 1970; Section 36.8 of MTW).

Different physicists feel comfortable with different levels of rigor. In recent years these differences have shown up strongly and publicly in a controversy over derivations of the quadrupole wave-generation formula (12) and the formula for the energy sapped from a source by radiation reaction (negative of equation (15)). Many physicists – myself among them – were quite satisfied with derivations at the level of rigor, e.g., of Landau and Lifshitz (1941) and Peres (1960). Others (e.g. Ehlers *et al.*, 1976) felt that those early derivations were inadequately rigorous and, correspondingly, that the quadrupole formulae were suspect for sources with non-negligible self-gravity. The controversy was heightened by the fact that there were mathematical errors in some (but not all) of the early derivations (see Walker and Will, 1980 and Section 3.4.2 of Thorne, 1983 for discussions).

The controversy was still raging in the early 1980s; see, e.g. Ashtekar (1983). However, during the mid 1980s it has largely subsided. There are now many new derivations of the quadrupole formulae, with much improved rigor, and they all produce the same, standard results; see, e.g., Anderson *et al.* (1982), Blanchet and Damour (1984); Christodoulou and Schmidt (1979); Isaacson, Welling and Winicour (1984); and for reviews see Will (1986), Schutz (1986*a*) and Damour (1987).

Of particular interest is radiation reaction in the binary pulsar PSR1913 + 16, which should cause the binary system's two neutron stars to spiral together slowly with a consequent gradual decrease in their orbital period. Because the duration of observations of the pulsar is short (12 years), the cumulative effects of radiation reaction on the orbit during those observations are 100 times smaller than post-Newtonian effects; and, consequently, the detailed effects of the radiation reaction could not be fully understood until the orbital equations were fully under control up through post-post-Newtonian order. Damour and Deruelle (1986) have now brought the orbital equations fully under control; and there is now a beautiful agreement between those equations – including the quadrupolar radiation reaction – and the observational data. For a detailed discussion see Chapter 6 of this book.

9.3.3 A catalog of formalisms for computing wave generation

The order-of-magnitude formula $h \sim E^{\mathrm{ns}}_{\mathrm{kin}}/r$ shows that of all sources at fixed distance r, the strongest emitters will be those with the largest non-spherical kinetic energies, i.e. those with the largest internal masses and with internal velocities approaching the speed of light. Thus, the strongest emitters are likely to violate the slow-motion assumption which underlies the quadrupole formalism, and will require for accurate analysis either higher-order corrections to the quadrupole formalism, or wave-generation formalisms that do not entail any slow-motion assumption.

There are a number of other wave-generation formalisms which can be applied to such sources. This section is a catalog of them, with references to detailed presentations and applications. For an out-of-date but unified presentation of most of these formalisms see Thorne (1977).

To be tractable with a minimum of numerical computation, a wave-generation formalism must break the extreme non-linearity of the Einstein field equations by imposing a power-series expansion in some small quantity and keeping only the lowest, linear order or the lowest few orders. Wave-generation formalisms can be classified according to their choice of the small expansion parameter. *Slow-motion formalisms* (of which the quadrupole formalism is an example) expand in $L/\lambdabar =$ (size of source)/(reduced wavelength of waves); subsection (a) below. *Post-Minkowski formalisms* expand in the strength of the gravitational field inside the source, i.e. in the magnitude of the deviations of the metric coefficients from their Minkowski values; subsection (b). *Post-Newtonian formalisms* expand simultaneously in $L/\lambdabar$ and the strength of the internal gravitational field; subsection (c). *Perturbation formalisms* expand in the deviations of the metric from its form for some non-radiative, astrophysical system – e.g. from the Kerr metric for a rotating black hole, or from the metric for an equilibrium, rotating, relativistic stellar model, or from the Friedman–Robertson–Walker metric for a homogeneous, isotropic 'big-bang'; subsection (d).

Of all astrophysical sources, the very strongest emitters will entail gravitationally induced large-amplitude, high-velocity, non-spherical internal motions – e.g. the inspiral and coalescence of a binary black hole or binary neutron-star system. For such sources there is no small parameter in which one can expand. The only way to compute the full details of the wave field emitted by such sources is by the techniques of *numerical relativity*: the numerical solution of the full Einstein field equations on a supercomputer; subsection (e).

(a) *Slow-motion formalisms*

As for electromagnetic waves, so also for gravitational waves, slow-motion expansions give rise automatically to multipolar expansions: the electromagnetic vector potential is a sum of an electric dipolar term (typical magnitude $(Q/r)(L/\lambdabar)$ where Q is the charge of the source), plus an electric quadrupole and a magnetic dipole [magnitude $(Q/r)(L/\lambdabar)^2$], plus an electric octupole and a magnetic quadrupole [magnitude $(Q/r)(L/\lambdabar)^3$], etc. Similarly, the gravitational-wave field h_{jk}^{TT} is the sum of a mass-quadrupole term $[\sim(\partial^2/\partial t^2)\mathscr{I}_{jk}/r \sim (M/r)(L/\lambdabar)^2]$, plus a mass octupole $[\sim(\partial^3/\partial t^3)\mathscr{I}_{ijk}/r \sim (M/r)(L/\lambdabar)^3]$ and a current quadrupole $[\sim(\partial^2/\partial t^2)\mathscr{S}_{ij}/r \sim (M/r)(L/\lambdabar)^3]$, etc. In each case, electromagnetic and gravitational, there are two families of moments involved; and in each case as one goes to higher orders in $L/\lambdabar$ one is driven to include higher-order moments of the source.

Early foundations for these expansions in the gravitational case will be found in Bonnor (1959) and Pirani (1964); full mathematical details will be found in Thorne (1980*b*); and major improvements and elucidations will be found in Blanchet and Damour (1986), Blanchet (1987*a*) and Damour (1987).

A good example which shows how, as the source's internal motions are speeded up, the higher moments gradually become more and more important, is the gravitational bremsstrahlung radiation emitted when two stars fly past each other with some high initial velocity; see Kovacs and Thorne (1978) and Turner and Will (1978) for full details.

In some slow-motion systems the mass quadrupole contribution may be suppressed, leaving the current quadrupole or the mass octupole to dominate. A good example is the torsional oscillation of a neutron star, in which the motions are slow because the shear modulus is weak, and the mass moments vanish because of parity considerations leaving the current quadrupole to dominate the radiation; for details see Schumaker and Thorne (1983). Another example is the Chandrasekhar (1970)–Friedman–Schutz (1978) (CFS) instability in neutron stars, which preferentially excites mass octupole ($l=3$) or hexadecapole ($l=4$) or $l=5$ modes of pulsation causing them to radiate more strongly than quadrupole; see Section 4.3(b) below.

(b) *Post-Minkowski formalisms*

Post-Minkowski wave-generation formalisms are sometimes called 'post-linear' because they entail expanding in the strength of the gravitational field beyond linear order; and they are sometimes called 'fast-motion' to contrast them with slow-motion formalisms.

There is a systematic way to take a post-Minkowski wave-generation formalism that is accurate to a given order in the strength of the source's internal gravity, and iterate it to obtain a formalism of higher accuracy (Thorne and Kovacs, 1975; Thorne, 1977).

The wave-generation formalism that is accurate to first post-Minkowski order (first order in the strength of internal gravity) is 'Linearized theory', i.e. the linear approximation to general relativity. Linearized theory is discussed in most textbooks, e.g. Chapter 18 and Sections 35.1–35.6 of MTW. Halpern and Desbrandes (1969) and, independently, Press (1977) have derived a particularly useful Linearized wave-generation formula for systems with sizes L large compared to a reduced wavelength $\lambdabar$. Examples of gravity-wave generation that have been analyzed by Linearized theory are: (i) the coherent (but painfully slow) transformation of electromagnetic waves into gravitational waves (first considered by Gertsenshtein, 1962, subsequent work reviewed in Section 4.1 of Grishchuk and Polnarev, 1980); and (ii) the waves emitted by the explosion of a non-spherical nuclear bomb (Wheeler, 1962; Wood *et al.*, 1970).

Linearized theory is completely ignorant of the source's internal gravity; it can correctly predict the emitted waves only if the source's motions are governed by non-gravitational forces – typically by electric or magnetic forces. For systems with significant but weak internal gravity (e.g. stellar pulsations, binary systems, and high-speed stellar encounters), one must use a wave-generation formalism accurate to the next, 'post-post-Minkowski' or 'post-Linear' order. For the details of such a formalism, see, e.g., Thorne and Kovacs (1975), Crowley and Thorne (1977); and for its application to high-speed stellar encounters (gravitational bremsstrahlung radiation) see Kovacs and Thorne (1977, 1978). For a recent review see Westpfahl (1985).

Thus far nobody has developed a post^2-Linear wave-generation formalism in detail – i.e. a formalism accurate to post^3-Minkowski order. There has been no great need for such a formalism, and the post-Linear formalism is sufficiently hard to work with in practice (cf. Kovacs and Thorne, 1977) that it is not clear whether post^2-Linear would be significantly easier than full-blown numerical relativity.

(c) *Post-Newtonian formalisms*

Post-Newtonian approximations to general relativity make the assumption – in accord with the virial theorem for gravitationally bound systems – that inside the source the deviations of the metric from Minkowski (i.e. the dimensionless strength of the source's gravity) have a magnitude ε of order $(L/\lambda)^2$; and accordingly they expand the Einstein field equations simultaneously in ε (post-Minkowski expansion) and L/λ (slow-motion expansion). See, e.g., Burke (1979) for a review of the method.

The lowest-order wave generation formalism that results from this expansion is called 'Newtonian' because it computes the evolution of the source using Newton's laws of gravity and mechanics, then evaluates the source's time-evolving quadrupole moment using the standard Newtonian volume integral (13a), then inserts that quadrupole moment into the standard quadrupole wave-generation formula (12). It is this Newtonian version of the quadrupole formalism that has been especially controversial (see the end of Section 3.2 above) but is now almost universally agreed to be highly reliable.

There have been a large number of important wave-generation calculations with this formalism. Some examples are: (i) the waves emitted by binary systems in Newtonian, elliptical orbits (Peters and Mathews, 1963); (ii) the waves emitted by a variety of models of stars that collapse to form neutron stars (Saenz and Shapiro, 1978, 1981); and (iii) the waves emitted in the head-on collision of two compact stars (Gilden and Shapiro, 1984).

The Newtonian wave-generation formalism starts losing accuracy when the source's internal gravity becomes too strong ($\varepsilon \sim 0.05$) and its internal velocities too high ($v \sim 0.2$) (Turner and Will, 1978) – e.g. in the late stages of the spiraling together of a neutron-star binary system. In such a situation it is useful to include the next higher-order corrections (one order higher in the strength of gravity ε, two higher in the speed L/λ). The result is the post-Newtonian formalism (Epstein and Wagoner, 1975; Wagoner, 1977; Tsvetkov, 1984). Examples of calculations that have been performed with the post-Newtonian formalism are the radiation from a system of bodies whose sizes are all small compared to their separations (Wagoner and Will, 1976), gravitational bremsstrahlung at moderate velocities (Turner and Will, 1978), and the radiation emitted by a slowly rotating star that collapses to a neutron star (Turner and Wagoner, 1979).

The foundations for a post^2-Newtonian wave-generation formalism have also been worked out (Section V.E. of Thorne, 1980*b*); but it has never been developed in full detail or applied to any sources. Such a formalism – by contrast with the post^2-Linear – would likely be far more tractable than full-blown numerical relativity; so it may one day prove useful in tying down the gravitational waves from large-amplitude processes involving neutron stars.

(d) *Perturbation formalisms*

Much can be learned about gravitational waves from neutron stars and black holes by studying weak, non-radial perturbations around their equilibrium structures. The theory of non-radial perturbations of non-rotating relativistic stars was developed by Campolattaro, Detweiler, Ipser, Price and Thorne (see pp. 195–201 of Thorne 1978 for a review and references); and extensions of the theory have been given by Schumaker and Thorne (1983), Detweiler and Lindblom (1985) and Finn (1986). Among its recent applications are computations by Lindblom and Detweiler (1983) of the normal modes of neutron stars with a variety of equations of state. For rotating stars the corresponding theory is due largely to Chandrasekhar, Friedman and Schutz (reviewed through 1977 on pp. 201–8 of Thorne 1978 and reviewed more recently by Schutz 1987). The most important recent applications are studies of the waves emitted by the 'Chandrasekhar–Friedman–Schutz instability' in rapidly rotating neutron stars; see Section 9.4.2(b) below.

For non-rotating (Schwarzschild) black holes the perturbation theory is due to Regge and Wheeler (1957) with major subsequent contributions by Bardeen, Chandrasekhar, Detweiler, Edelstein, Moncrief, Press, Vishveshwara and Zerilli; see pp. 180–8 of Thorne (1978) for a review. Recent applications include a definitive numerical evaluation of the eigenfrequencies and damping times of a Schwarzschild hole's quasinormal modes (Leaver, 1985, 1986*a*), and of the Green's function for arbitrary perturbations of a Schwarzschild hole (Leaver, 1986*b*). Finally, for rotating (Kerr) black holes the theory is due to Teukolsky (1972, 1973), Teukolsky and Press (1974), Wald (1973), Chrzanowski (1975), Cohen and Kegeles (1975), Chandrasekhar and Detweiler (1976), Detweiler (1977), Chandrasekhar (1983) and Sasaki and Nakamura (1982); and recent applications include the evaluation of the eigenfrequencies and damping times of a Kerr hole's quasinormal modes (Detweiler, 1980; Leaver, 1985, 1986*a*), and evaluations of the gravitational waves emitted when a compact body orbits a Kerr hole (Detweiler, 1978), scatters gravitationally off a Kerr hole (Kojima and Nakamura, 1984*b*), or plunges into a Kerr hole (Detweiler and Szedenits, 1979; Kojima and Nakamura, 1984*a*).

Primordial gravitational waves (waves created in or near the big-bang) cannot be analyzed by splitting space into a wave-generation region plus wave zones (Section 9.3.1), because the transition from wave generation to wave propagation is a temporal rather than a spatial one: the transition occurs when the size of the cosmological horizon expands to become much larger than a wavelength, thereby unfreezing a set of frozen-in initial perturbations. The only way, today, to analyze primordial waves is by a perturbation formalism somewhat akin to that used for stars and black holes: the unperturbed configuration is typically a non-radiative, Friedman–Robertson–Walker cosmological model, and the perturbations are studied by linearizing the Einstein equations – or a quantized variant of them – around that model. The resulting theory is due to Lifshitz (1946); see also Section 7.3 of Zel'dovich and Novikov (1983), and references cited in Section 9.4.3(d) below.

(e) *Numerical relativity*

Numerical solution of the full Einstein equations is the only way, today, to study wave generation in the strongest and most interesting of all gravity-wave sources: those with high internal velocities, strong internal gravity, and large deviations from a non-radiating spacetime. During the past decade several dozen researchers have worked vigorously to develop the field of numerical relativity. The results of this effort are: an analytic formulation of the initial value problem and the dynamical evolution problem for the Einstein field equations, in a form that facilitates numerical solutions (York, 1983, and references cited therein); for axisymmetric systems, viable ways to slice spacetime and choose the spatial coordinates so as to avoid pathologies and to compute the emitted gravitational waves efficiently (Smarr and York, 1978; Piran, 1983; Bardeen and Piran, 1983; Sasaki, 1984; Stewart and Friedrich, 1983; Isaacson, Welling and Winicour, 1983; Gomez *et al.*, 1986; Anderson and Hobill, 1986; Evans and Abrahams, 1987); and several good computer codes for evolving axisymmetric systems and computing their waves (Nakamura, 1983; Stark and Piran, 1986; Piran and Stark, 1986; Evans, 1986).

Among the problems that have been studied successfully with the codes thus far are the gravitational radiation produced by a head-on collision of two Schwarzschild black holes (Smarr, 1977*a*), by the collapse of a rotating star to form a Kerr black hole (Stark and Piran, 1986), and by the vibrations of a neutron star (Evans, 1986).

The experts in numerical relativity are now beginning to move on from axisymmetric systems, which have two non-trivial spatial dimensions and one non-trivial time, to asymmetric, generic systems with three non-trivial spatial dimensions and one time (e.g. Nakamura, 1987). This full '3 + 1' effort will require using the world's largest supercomputers, and will require new techniques for slicing spacetime into space plus time, for choosing the spatial coordinates, and for differencing the Einstein equations. The effort may absorb almost as many person-years as the development of gravitational-wave detectors; but it will be well worthwhile: the payoffs will include the ability to compute in detail the waveforms from the strongest gravity-wave sources in the universe, such as the spiraling together and coalescence of two black holes – waveforms that will be crucial to the interpretation of gravity-wave observations and to their use for strong-field, highly dynamical tests of general relativity.

One should not be misled into believing that numerical relativity will be the totally dominant tool for realistic gravitational-wave calculations in the coming years. On the contrary, we can expect a healthy interaction between numerical and analytical techniques; for discussion see Schutz (1986*c*).

9.3.4 *Wave propagation in the real universe*

Since I have recently written a detailed review of the theory of gravitational wave propagation (Section 2 of Thorne, 1983), I shall only sketch the main points briefly.

No matter how strong a source of gravity waves may be, once its waves are fully formed (once they have reached a location, e.g. in the local wave zone, where the inhomogeneity lengthscale $\mathscr{L}$ of the background curvature is large compared to their wavelength $\lambdabar$), they will have dimensionless amplitudes h small compared to unity. This can be seen as follows:

The stress-energy tensor of the gravitational waves, $T^{\mathrm{GW}}_{\alpha\beta}$ (equation (9)), acts as a source for the background curvature through the Einstein equations $G^{\mathrm{B}}_{\alpha\beta}=8\pi(T^{\mathrm{GW}}_{\alpha\beta}+T^{\mathrm{other}}_{\alpha\beta})$. Since the background Einstein tensor $G^{\mathrm{B}}_{\alpha\beta}$ has magnitude less than or of order $1/\mathscr{R}^2_{\mathrm{B}}$ (where $\mathscr{R}_{\mathrm{B}}$ is the background radius of curvature as defined from the Riemann tensor), and since the magnitude of the gravity-wave stress-energy tensor (9) is $h^2/\lambdabar^2$, the Einstein equations imply that

$$h \lesssim \lambdabar/\mathscr{R}_{\mathrm{B}}. \tag{20}$$

Since the inhomogeneity lengthscale $\mathscr{L}$ of the background curvature is always less than or of order the radius of curvature $\mathscr{R}_{\mathrm{B}}$, equation (20) implies the claimed result:

$$h \lesssim \lambdabar/\mathscr{R}_{\mathrm{B}} \lesssim \lambdabar/\mathscr{L} \ll 1. \tag{21}$$

This permits one to study the subsequent propagation of the waves, once they are fully formed, using a linearized approximation to the Einstein field equations (Sections 35.13 and 35.14 of MTW): (i) one introduces a field $\bar{h}_{\alpha\beta}$ (which is actually the trace-reversed contribution of the waves to the spacetime metric in a suitable gauge). This field is defined to be equal to $h^{\mathrm{TT}}_{\alpha\beta}$ in the region from which the waves are propagating (the local wave zone in the case of isolated sources; the very early universe in the case of primordial waves). One then evolves the field $\bar{h}_{\alpha\beta}$ out into the surrounding universe and to earth using the curved-spacetime wave equation (equation (35.64) of MTW)

$$\bar{h}_{\alpha\beta|\mu}{}^{|\mu} + g^{\mathrm{B}}_{\alpha\beta}\bar{h}^{\mu\nu}{}_{|\nu\mu} - 2\bar{h}_{\mu(\alpha}{}^{|\mu}{}_{|\beta)} + 2R^{\mathrm{B}}_{\mu\alpha\nu\beta}\bar{h}^{\mu\nu} - 2R^{\mathrm{B}}_{\mu(\alpha}\bar{h}_{\beta)}{}^{\mu} = -16\pi\delta T_{\alpha\beta}. \tag{22}$$

Here $g^{\mathrm{B}}_{\alpha\beta}$ is the background metric, the subscript and superscript $|$ denote covariant derivatives with respect to the background metric, $R^{\mathrm{B}}_{\alpha\beta}$ and $R^{\mathrm{B}}_{\alpha\beta\gamma\delta}$ are the Ricci and Riemann curvature tensors of the background; and $\delta T_{\alpha\beta}$ is the perturbation in the non-gravitational stress-energy tensor produced by the trace-reversed metric perturbation $\bar{h}_{\alpha\beta}$ itself.

Although the field $\bar{h}_{\alpha\beta}$ initially is wavelike and thus has $\lambda\!\!\!^{-} \ll \mathscr{L} \lesssim \mathscr{R}_{\mathrm{B}}$, it might propagate into regions where the background has very short lengthscales, $\mathscr{L} \lesssim \lambda\!\!\!^{-}$. (For example, the waves produced by the Crab pulsar, with $\lambda\!\!\!^{-} \sim 1000$ km may propagate through a massive white dwarf with $\mathscr{L} \sim 1000$ km and $\mathscr{R}_{\mathrm{B}} \sim 30\,000$ km, or even through a neutron star with $\mathscr{L} \sim 10$ km and $\mathscr{R}_{\mathrm{B}} \sim 30$ km). If this happens, one need not worry. The wave equation (22), because it depends for its validity only on the weakness of the field $\bar{h}^{\alpha\beta}$ and not on the shortwave assumption, remains valid and carries the field through the region of short background lengthscale (where, strictly speaking, it is no longer a gravitational wave), and thence onward into regions of long lengthscale (where it once again is a gravitational wave).

In those regions where $\bar{h}_{\alpha\beta}$ is a wave, i.e. has $\lambda\!\!\!^{-} \ll \mathscr{L}$, one can compute from it the gravitational-wave field h^{TT}_{jk} by a very simple prescription: introduce the proper reference frame of a specific observer; and in that frame discard the time-time and time-space parts of $\bar{h}_{\alpha\beta}$, and algebraically project out from the space-space parts those pieces that are transverse to the propagation direction and are trace-free. The result will be h^{TT}_{jk}. For further discussion and justifications see Box 35.1 of MTW and Section 2.4.2 of Thorne (1983).

As I shall discuss below, the effects of the wave-stimulated stress-energy perturbations $\delta T_{\alpha\beta}$ are never large enough to be astrophysically important, so one can ignore them in propagation calculations. Moreover, in regions (nearly everywhere) where $\bar{\lambda} \ll \mathscr{R}_B$, the contributions of the background curvature tensors $R^B_{\mu\alpha\nu\beta}$ and $R^B_{\mu\alpha}$ to the wave equation can be ignored, and one can specialize the gauge so as to make $\bar{h}_{\alpha\beta}$ divergence free. The wave equation (22) then assumes the simplest form imaginable:

$$\Box \bar{h}_{\alpha\beta} \equiv \bar{h}_{\alpha\beta|\mu}{}^{\mu} = 0. \tag{23}$$

When, in addition, the radius of curvature of the wave fronts is large compared to $\bar{\lambda}$ (as is true everywhere except near the very rare focal points of gravitational lenses), the wave equation (23) can be solved easily by the techniques of geometric optics (Isaacson, 1968*a*; Exercise 35.15 of MTW; Section 2.5 of Thorne, 1983): the field $\bar{h}^{\alpha\beta}$ propagates along null rays; its polarization is parallel transported along the rays; and its amplitude, like the amplitude of light, varies along each ray as 1/(the radius of curvature of the wave front).

Because good gravitational lenses are so rare in the real universe, and because regions (black holes and neutron stars) with very strong curvature ($\mathscr{R}_B \lesssim \bar{\lambda}$) are so rare and so small, the waves from almost every source will propagate to earth via pure geometric optics. Moreover, because the universe is almost globally Lorentz (flat) in its background geometry on lengthscales small compared to the Hubble distance, for sources at distances much less than Hubble (at cosmological redshifts $z \ll 1$), this geometric optics propagation will produce a simple $1/r$ falloff of amplitude and will preserve the wave form (the dependence on retarded time) and the polarization. More specifically, it will produce a gravitational-wave field in TT coordinates with the simple form

$$h^{TT}_{jk} = \frac{A^{TT}_{jk}(t-r, \theta, \phi)}{r}, \tag{24}$$

where r is distance to the source and θ, ϕ are spherical polar angles centered on the source. For a slow-motion source the function A^{TT}_{jk} is just twice the TT part of the second time derivative of the source's quadrupole moment, as one sees trivially by matching (24) onto (12) in the source's local wave zone.

For a source at a large cosmological redshift $z \gtrsim 1$, if one approximates the background spacetime geometry by that of a Friedmann–Robertson–Walker cosmological model, the geometric-optics propagation produces the same effects for gravity waves as for light: (i) the magnitude of h^{TT}_{jk} falls off with the same $1/R$ behavior as for light, where R is $(1/2\pi) \times$ (circumference of a sphere passing through the earth and centered on the source, at the time the waves reach earth) (equations (29.28)–(29.33) of MTW); (ii) the polarization, like that of light in vacuum, is parallel transported radially from source to earth; and (iii) the time dependence of the wave form is unchanged by propagation, except for a frequency-independent redshift $f_{\text{received}}/f_{\text{emitted}} = 1/(1+z)$. For further details and derivations see Section 2.5.4 of Thorne (1983) or Section 7.2 of Thorne (1977).

9.3.5 *A catalog of wave-propagation effects*

In principle gravitational waves can experience almost all the familiar peculiarities of propagation that electromagnetic waves experience. Here I shall enumerate those that have been studied, mention briefly their importance or unimportance, and give references for further detail.

(a) *Absorption, scattering and dispersion by matter and electromagnetic fields*

Gravitational waves are so weakly absorbed by matter (Section 2.4.3 of Thorne, 1983) that absorption is astrophysically important only near the Planck era of the big-bang (Section 7.2 of Zel'dovich and Novikov, 1983). For experimenters, however, the tiny absorption that should occur in a gravity wave detector is very important; see Section 9.5.2(b) below.

The scattering of gravitational waves by matter is also so weak that it is never astrophysically important, except near the Planck era. In the analysis of resonant gravity wave detectors, however, scattering is conceptually important: the method of detailed balance (an idealized calculation in which a resonant detector is driven by monochromatic waves into vibrations of such unrealistically high amplitude that it reradiates at the same rate as it absorbs – i.e. it scatters the waves strongly) is a powerful way of computing the cross-section of a resonant detector; see Section 37.7 of MTW.

Coherent scattering by a medium produces dispersion – i.e. frequency-dependent propagation speeds. Although dispersion is very important for electromagnetic waves, it is never of any importance in the real universe for gravitational waves (Section 2.4.3 of Thorne, 1983). It is instructive, however, to imagine and study theoretically an unrealistic form of matter ('respondium') with such strong dispersion that it actually reflects gravitational waves (Press, 1979).

For detailed analyses of absorption, scattering, and/or dispersion by specific kinds of matter see the following references and the references cited therein: black holes, Matzner *et al.* (1985); De Logi and Kovacs (1977); neutron stars and other stars, Linet (1984); elementary particles, Section 7.2 of Zel'dovich and Novikov (1983); a magnetized plasma, Macedo and Nelson (1983); a uniform medium, Esposito (1971*a*,*b*), Papadopoulos and Esposito (1985), Szekeres (1971), Section 4.2 of Grishchuk and Polnarev

(1980), Section 2.4.3 of Thorne (1980); and a medium with boundaries, Dyson (1969), Carter and Quintana (1977).

Because gravitational and electromagnetic waves should propagate with the same speed, they can interact in a coherent way (Gertsenshtein, 1962). The interaction is so weak, however, that a substantial transformation of one into the other requires propagation over a distance of order the radius of curvature of the background spacetime which their own energy density produces. Thus, such coherent interaction is not likely ever to be important in the real universe – except possibly in gravity-wave detectors; see Section 9.5.4(a) below. For a review of the extensive literature on this subject see Grishchuk and Polnarev (1980). For a brief pedagogical discussion see Section 17.9 of Zel'dovich and Novikov (1983).

(b) *Scattering by background curvature, and tails of waves*

In regions where the background radius of curvature is comparable to or shorter than the reduced wavelength, $\mathscr{R}_B \lesssim \lambda\!\!\!^{-}$, the background strongly scatters the waves. This is very important in some sources of waves, as the waves are trying to form; for example, it is responsible for the normal-mode vibrations of black holes (Press, 1971), and it leads to the formation of 'tails' of the waves in a source's near zone (Price, 1972*a*,*b*; Thorne, 1972; Cunningham, Price and Moncrief, 1979) and to radiative tails in the wave zone (Leaver, 1986*b*; Blanchet, 1987) – which, however, are not likely ever to be observationally important.

(c) *Gravitational focusing*

Lumps of background curvature associated with black holes, stars, star clusters, and galaxies will focus gravitational waves in precisely the same manner as they focus electromagnetic waves; and just as this focusing is observationally important for the light and radio waves from a few very distant quasars, so it might also be important for very distant discrete sources of gravitational waves. Focusing by the sun, in the case of waves with sufficiently short wavelength can be significant, but not at earth; the focal point lies farther out in the solar system, near the orbit of Jupiter (Cyranski and Lubkin, 1974).

(d) *Diffraction*

Near the focal point of a gravitational lens the waves cease to propagate along null rays and begin to diffract, thereby lessening the strength of the focusing. The analysis of this is no different for gravitational waves than for electromagnetic or scalar waves, since polarization plays no important role. Diffraction causes focusing to be significant only when the lens has a gravitational radius $2M$ that is larger than or of order the waves' reduced wavelength $\lambdabar$. For an order of magnitude discussion see, e.g. Section 2.6.1 of Thorne (1983); for full details see Bontz and Haugan (1981).

(e) *Parametric amplification by background curvature*

In regions of a dynamical spacetime (e.g. the expanding universe) in which the characteristic wavelength $\lambdabar$ of gravitational waves is larger than or comparable to the background radius of curvature $\mathscr{R}_B$, $\lambdabar \gtrsim \mathscr{R}_B$, the waves can be parametrically amplified by interaction with the dynamical background (Grishchuk, 1974, 1975*a*,*b*, 1977; Grishchuk and Polnarev, 1980). Viewed quantum mechanically, the interaction causes stimulated emission of new gravitons. This effect may well have enabled the expansion of the universe to amplify vacuum fluctuations from the big-bang singularity (from the Planck time) into a strong, stochastic background of gravitational waves today; see Section 9.4.3(d) below.

(f) *Non-linear coupling of the waves to themselves* (*frequency doubling, etc.*)

Because general relativistic gravity is non-linear, there is a non-linear coupling of gravitational waves to themselves; and in principle this leads to such non-linear conversion processes as frequency doubling. However, in practice these effects are not important in regions where the waves *are* waves (where $\lambdabar \ll \mathscr{L}$). This is because, in such regions, the dimensionless amplitude of the waves is very small compared to unity (equation (21) above). For a more detailed discussion see Section 2 of Thorne (1985).

(g) *Generation of background curvature by the waves*

The generation of background curvature by the stress-energy of the waves (Isaacson, 1968*b*, MTW Section 35.15) is important in cosmological models in any epoch when the waves are sufficiently strong that their energy density is comparable to that of matter; see, e.g. Hu (1978) and Chapter 17 of Zel'dovich and Novikov (1983). It is also important in a 'geon' – i.e. a bundle of gravitational waves that is held together by its own gravitational pull on itself (Wheeler, 1962; pp. 409–38 of Wheeler, 1964; Brill and Hartle, 1964). But geons surely do not exist in the real universe; they are only theoretical entities, useful for exploring issues in fundamental physics.

Wave-produced background curvature is also important in the idealized situation where one plane-fronted wave gets focused by passing through another ('plane-wave collision'): the focusing itself is produced by the wave-generated background curvature. Moreover, if the waves are precisely planar, a spacetime singularity forms at the focal plane (Khan and Penrose, 1971; Szekeres, 1972; Nutku and Halil, 1977; Tipler, 1980; Matzner and Tipler, 1984; Chandrasekhar and Xanthopoulos, 1986; Yurtsever, 1987*a*), and the generation of background curvature plays a key role in the singularity. In the more realistic case (which, however, almost certainly does not occur in the real universe except conceivably near the big-bang), in which the waves are almost planar but die out slowly at large transverse distances, if the transverse size is sufficiently large compared to the initial wave amplitude, then the focusing probably still drives the amplitude up far enough – before diffraction can act – to make background curvature generation become strong and force a singularity to form (Yurtsever, 1987*b*).

9.3.6 Wave propagation in an idealized, asymptotically flat universe

The split of wave propagation and wave generation into two separate calculations is worrisome to those physicists who seek the highest levels of rigor. Unfortunately, in analyses of waves in the real universe, where the background spacetime is complex, the split is essential. Progress cannot be made without it. However, the split can be avoided to some extent in an idealized 'universe', where the source of interest resides alone in an otherwise empty and asymptotically flat spacetime.

In such an idealized universe, and only there, it has been possible to treat wave generation and wave propagation simultaneously, with a single, elegant formalism that spans the weak-gravity portions of the wave-generation region, and all of the local wave zone and distant wave zone. In my opinion, the nicest version of this unified formalism is a variant of the post-Minkowski formalism due to Blanchet and Damour (1986) and Blanchet (1987*a*). See Damour (1986) for a review of this and of earlier work by others. This formalism is especially nice because, although it entails an expansion in the strength of the gravitational field, the expansion has been carried out to all orders, and a number of beautiful theorems have been proved about it.

Also of great interest, elegance, and beauty, in an idealized asymptotically flat universe, are expansions of the spacetime curvature along the outgoing light cone in inverse powers of the distance to the source (Bondi, 1960; Bondi, van der Burg and Metzner, 1962; Sachs, 1962, 1963; Penrose, 1963*a*,*b*; Newman and Penrose, 1965). Such expansions, carefully formulated and combined with conformal transformations that bring 'infinity' in to finite locations, reveal asymptotic structures of asymptotically flat spacetime that illuminate the properties of gravitational radiation. For recent reviews and references see Newman and Tod (1980), Walker (1983), Schmidt (1979, 1986), Ashtekar (1984).

9.3.7 Exact, analytic solutions for wave generation and propagation

Those who seek high rigor (and even less rigorous people like me) have also found pleasure in the few existing exact, analytic solutions of the Einstein field equations that describe sources which radiate into asymptotically flat spacetime (e.g. Bicak, 1968; Bicak, Hoenselaers and Schmidt, 1983; Bicak, 1985; and references therein); and in idealized exact solutions for cylindrical waves (Einstein and Rosen, 1936; Weber and Wheeler, 1957) and for planar waves (e.g. Rosen, 1937; Bondi, Pirani and Robinson, 1959; Ehlers and Kundt, 1962; Sections 35.9–35.12 of MTW). Also pleasing for its rigor is the extreme limit of geometric optics, where the wavelength becomes so short that the radiation is compacted into a *gravitational shock wave* (e.g. Pirani, 1957; Papapetrou, 1977 and references therein).

9.4 Astrophysical sources of gravitational waves

In the real universe it is useful to divide the anticipated gravitational waves into three classes, according to their temporal behaviors: *bursts*, which last for only a few cycles, or at most for times short compared to a typical observing run; *periodic waves*, which are superpositions of sinusoids with frequencies f_i that are more or less constant over times long compared to an observing run; and *stochastic waves*, which fluctuate stochastically and last for a time long compared to an observing run.

In this section I shall describe the present knowledge and speculations about various sources of gravitational waves. Attention will be restricted to those sources which are most promising for detection by present or planned detectors, with Section 9.4.1 focusing on burst sources, 9.4.2 on periodic sources, and 9.4.3 on stochastic sources.

As was emphasized in Section 9.1.1, for each source, with the single exception of binary stars and their coalescences (Section 9.4.1(e) below), either (i) the strength of the source's waves for a given distance from earth is uncertain by several orders of magnitude; or (ii) the rate of occurrence of that type of source, and thus the distance to the nearest one, is uncertain by several orders of magnitude; or (iii) the very existence of the source is uncertain. Despite these uncertainties, it is important in the wave-detection effort to estimate as best one can, for each source, the strengths of the waves bathing the earth. Such estimates will be stated below, with references; and they are collected together below in Figs. 9.6 (burst sources), 9.7 (periodic sources) and 9.6 (stochastic sources).

9.4.1 Burst sources

(a) *Bursts with and without memory*

As Braginsky and Grishchuk (1985) have recently emphasized, gravitational-wave bursts can be subdivided into two classes: *normal bursts*, in which h_{jk}^{TT} begins zero before the burst and returns to zero afterward, and *bursts with memory*, in which h_{jk}^{TT} (by convention) begins zero before the burst and then settles down into a non-zero, constant value $\Delta h_{jk}^{\mathrm{TT}}$ (the burst's 'memory') after the burst is over.

Physically, a burst with memory arises whenever the source, before the burst or afterward or both, consists of several free bodies that are moving with uniform velocities relative to each other. For example, the explosion of a star into several pieces will produce a burst with memory, as will the collision of two freely moving (not binary) stars or black holes, or the gravitational scattering of a star by a black hole; but the birth of a black hole in non-spherical stellar collapse will produce a burst without memory. In all cases, the 'memory' is the change in the total '$1/r$' coulomb-type gravitational field of the source (Braginsky and Thorne, 1987). For example, in the quadrupole approximation, when one takes account of the fact that one can add a time-independent constant to h_{jk}^{TT} ('gauge change' that moves stuff between the background field and the wave field) so as to make h_{jk}^{TT} zero initially, equations (12) and (13a) give

$$\Delta h_{jk}^{\mathrm{TT}} = \Delta \sum_A \left(\frac{4m_A v_A^j v_A^k}{r} \right)^{\mathrm{STF}}. \tag{25}$$

Here the summation is over the free bodies in the system, m_A and v_A^j are the mass and velocity of body A, and Δ denotes the change from before the burst is emitted to afterward.

As is discussed by Braginsky and Thorne (1987), the memory part of a burst, $\Delta h_{jk}^{\mathrm{TT}}$, can be studied by any detector (with adequate sensitivity) that operates at a frequency lower than the burst's characteristic frequency, $f \lesssim f_c$. Put equivalently, one can think of a burst with memory as having a signal that extends down to all frequencies below f_c (cf. the 'zero-frequency limit' discussed by Smarr, 1977*b* and by Bontz and Price, 1979).

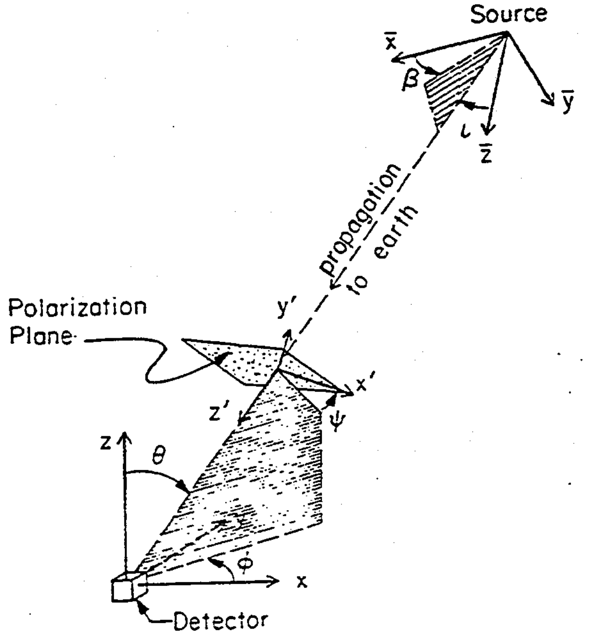
Current prejudice suggests that the strongest of burst sources (and thus the most interesting) may produce normal bursts rather than bursts with memory; but this prejudice could perfectly well be wrong. In accord with this prejudice, the remainder of this section will focus on normal bursts.

(b) *Characterization of the waves from a normal burst source and the noise in a detector searching for them*

Burst sources are best characterized by their full wave forms $h_{jk}^{TT}(t)$. However, when comparing with detector sensitivities, it is helpful to have a more compact characterization. Past discussions have used a loosely defined 'characteristic amplitude' h_c and 'characteristic frequency' f_c. However, the factors of order 3 that are glossed over by the loose definitions are beginning to be important in the planning of gravity-wave searches, especially in the case of the inspiral and coalescence of binary neutron stars (subsection (e) below); and I therefore shall be quite careful in my definitions of h_c and f_c and subsequently in my corresponding discussions of detector sensitivities.

As an aid to defining h_c and f_c with care, I show in Fig. 9.2 a diagram of the emission, propagation, and absorption of the wave. The source has

Fig. 9.2. The angles $\iota, \beta, \psi, \theta, \phi$ which characterize the emission, propagation and detection of a gravitational wave.



preferred local Cartesian axes $(\bar{x}, \bar{y}, \bar{z})$ with respect to which its internal structure is especially simple. We shall denote by $(\imath, \beta)$ the direction toward earth (spherical polar angles) relative to those axes. The detector, similarly, has preferred local Cartesian axes (x, y, z) with respect to which its internal structure is especially simple; and we shall denote by (θ, ϕ) the direction toward the source (spherical polar angles) relative to those detector axes. The waves themselves, as they pass through the detector, are most simply described in a third set of Cartesian axes (x', y', z') with origin at the detector's center of mass, z' axis along the waves' propagation direction, and x' and y' axes so oriented in the polarization plane as to make the wave forms $h_+ = h^{\mathrm{TT}}_{x'x'} = -h^{\mathrm{TT}}_{y'y'}$, and $h_\times = h^{\mathrm{TT}}_{x'y'} = h^{\mathrm{TT}}_{y'x'}$, especially simple. We shall denote by ψ the angle between the x' polarization axis and the $\phi = 0$ plane.

From a model for the source one can compute the wave forms $h_+(t - z'; \imath, \beta)$ and $h_\times(t - z'; \imath, \beta)$ arriving at the detector. If the detector is small compared to a reduced wavelength, as we shall assume throughout Section 9.4 and in Sections 9.5.1–9.5.3 but not 9.5.4–9.5.6, then it will feel a simple linear combination of h_+ and $h_\times$; i.e. it will detect the wave form

$$h(t) = F_+(\theta, \phi, \psi) h_+(t; \imath, \beta) + F_\times(\theta, \phi, \psi) h_\times(t; \imath, \beta). \qquad (26)$$

Here F_+ and $F_\times$ are detector beam-pattern functions (to be discussed in Sections 9.5.2 and 9.5.3 below), which depend on the direction of the source (θ, ϕ) and the orientation ψ of the polarization axes relative to the detector's orientation (Fig. 9.2) and which have values in the range $0 \leqslant |F_A| \leqslant 1$. There will be some special choice of θ, ϕ, ψ for which $F_+ = 1$ and $F_\times = 0$; we shall call this the 'optimum source direction and polarization' for the + mode of the waves.

In Section 9.5 we shall characterize the noise in a detector by a frequency-dependent spectral density $S_h(f)$ (with dimensions Hz^{-1}) defined as follows: if a precisely sinusoidal gravitational wave with known phase α, known frequency $f > 0$ and unknown rms amplitude h_0,

$$h = (2)^{\frac{1}{2}} h_0 \cos(2\pi f t + \alpha), \qquad \alpha = \text{const.}, \qquad (27\text{a})$$

impinges on the detector, and if the experimenters seek to detect the wave by Fourier analyzing the detector output with a bandwidth Δf (integration time $\hat{\tau} = 1/\Delta f$), then the amplitude signal-to-noise ratio will be

$$\frac{S}{N} = \frac{h_0}{[S_h(f)\Delta f]^{\frac{1}{2}}}. \qquad (27\text{b})$$

In much of the gravitational-wave literature this $S_h(f)$ is denoted $[h(f)]^2$; i.e. $h(f) \equiv [S_h(f)]^{\frac{1}{2}}$ (with dimensions $\mathrm{Hz}^{-\frac{1}{2}}$).

In this section we are interested not in periodic gravitational waves, but rather in bursts – i.e. in waves with complicated time dependences $h(t)$ and with durations short compared to the experimenters' observation time. We shall assume that the wave form $h(t)$ is known, and that our only question is whether the wave is really present or not – with some postulated starting time t_o. (Later we shall discuss the statistical consequences of an unknown starting time.) There is a well-known optimal strategy for searching for such a wave in the presence of noise with known spectral density:

(i) One constructs a *Wiener optimal filter* $K(t)$, that function of time whose Fourier transform $\tilde{K}(f)$ is the same as the transform $\tilde{h}(f)$ of the signal $h(t)$, weighted by $1/S_h(f)$ (so that noisy frequencies are suppressed):

$$\tilde{K}(f) \equiv \frac{\tilde{h}(f)}{S_h(f)}, \qquad \tilde{h}(f) \equiv \int_{-\infty}^{+\infty} h(t)\, e^{i2\pi ft}\, dt, \tag{28a}$$

$$K(t) = \int_{-\infty}^{+\infty} \tilde{K}(f)\, e^{-i2\pi ft}\, df. \tag{28b}$$

(ii) One then takes the output of the detector, which includes noise and possibly signal; from it one computes the gravitational wave form $h_{\text{output}}(t)$ that would have been required to produce the observed output if there were no noise present; and one then computes a quantity

$$W = \int_{-\infty}^{+\infty} K(t-t_o) h_{\text{output}}(t)\, dt. \tag{28c}$$

(iii) This quantity will have root-mean-square contribution N from noise; and if the signal was actually present with starting time t_o, it will have a signal contribution S (i.e. $W = N + S$), with squared signal-to-noise ratio

$$\frac{S^2}{N^2} = \int_0^{\infty} \frac{2|\tilde{h}(f)|^2}{S_h(f)}\, df. \tag{29}$$

See Wiener (1949), Sections 25–27 of Wainstein and Zubakov (1962), Kafka (1977), and Michelson and Taber (1984) for proofs and discussion. The integral in (29) is from 0 to ∞ rather than $-\infty$ to $+\infty$ and there is a factor 2 present because $S_h(f)$ is a 'one-sided spectral density' (defined only for $f \geqslant 0$; negative frequencies are folded into positive). The squared signal-to-noise ratio (29) will be the basis for our definitions of h_c and f_c.

Expression (29) is the squared signal-to-noise ratio for a specific ('fiducial') source at some specific distance r_o from earth. Suppose that space is uniformly filled with sources, all identical to this fiducial source but with random directions (θ, ϕ), orientations (ι, β), and polarization angles ψ. Suppose, further, that inside the source's distance r_o there is, on average, one burst each D_o days. Then what, on average, will be the squared signal-to-noise ratio $(S^2/N^2)_{\text{strongest}}$ of the strongest burst that occurs each D_o days? One might expect the answer to be the fiducial S^2/N^2 of (29), averaged over all the angles θ, ϕ, ψ, ι, β. Not so. There is a statistical preference for directions and polarizations that give larger values of S^2/N^2, because they can be seen out to greater distances where the event rate is greater. This effect gives, assuming the event rate goes up as r^3 and h^{TT}_{jk} goes down as $1/r$, $(S^2/N^2)_{\text{strongest}}=\langle S^3/N^3\rangle^{\frac{2}{3}}$ where $\langle\cdots\rangle$ denotes an average over randomly distributed angles. Now, the $\frac{2}{3}$ power is a pain to deal with in subsequent calculations, so we shall switch to a straightforward angular average $\langle S^2/N^2\rangle$ and to compensate we shall insert a multiplicative factor $\frac{3}{2}$, which is (approximately) the ratio of $\langle S^3/N^3\rangle^{\frac{2}{3}}$ to $\langle S^2/N^2\rangle$ if both source and detector have quadrupole beam patterns ('slow-motion, poor-antenna regime'). This, together with $\langle F^2_+\rangle=\langle F^2_\times\rangle$ and $\langle F_+F_\times\rangle=0$ (true for any quadrupole-beam-pattern gravity-wave detector), permits us to write $(S^2/N^2)_{\text{strongest}}$ in the following approximate form, where we omit the subscript 'strongest' for ease of notation and where $\langle|\tilde{h}_+|^2+|\tilde{h}_\times|^2\rangle$ is averaged over source angles (ι, β):

$$\frac{S^2}{N^2}\cong 3\langle F^2_+(\theta,\phi,\psi)\rangle\int_0^\infty\frac{\langle|\tilde{h}_+|^2+|\tilde{h}_\times|^2\rangle}{S_h(f)}\,\mathrm{d}f. \tag{30}$$

We shall now specialize, for the remainder of this subsection, to gravity-wave searches with broad-band detectors, i.e. detectors for which the noise $S_h(f)$ is small over a band of frequencies $f_{\min}\lesssim f\lesssim f_{\max}$, with $f_{\max}\gtrsim 2f_{\min}$. Narrow-band detectors will be treated separately in Section 9.5.2(b) below. For broad-band detectors equation (30) motivates us to define the characteristic frequency and amplitude of our fiducial source, at distance r_o, by

$$f_c\equiv\left[\int_0^\infty\frac{\langle|\tilde{h}_+|^2+|\tilde{h}_\times|^2\rangle}{S_h(f)}\,\mathrm{d}f\right]^{-1}\left[\int_0^\infty\frac{\langle|\tilde{h}_+|^2+|\tilde{h}_\times|^2\rangle}{S_h(f)}\,f\,\mathrm{d}f\right], \tag{31a}$$

$$h_c\equiv\left[3\int_0^\infty\frac{S_h(f_c)}{S_h(f)}\langle|\tilde{h}_+|^2+|\tilde{h}_\times|^2\rangle f\,\mathrm{d}f\right]^{\frac{1}{2}}, \tag{31b}$$

where the average is over randomly distributed source orientation angles ι, β. Similarly, we shall define a characteristic detector noise amplitude at frequency f by

$$h_n(f)\equiv\frac{[fS_h(f)]^{\frac{1}{2}}}{\langle F^2_+(\theta,\phi,\psi)\rangle^{\frac{1}{2}}}. \tag{32}$$

In terms of these quantities equation (30) reads

$$\frac{S}{N}=\frac{h_c}{h_n(f_c)}. \qquad (33)$$

This is the S/N of the strongest burst that arrives at earth, on average, at the same rate as bursts occur inside the fiducial source's distance r_o.

In Fig. 9.4 below we characterize detector burst sensitivities not by $h_n(f)$, but rather by a quantity $h_{3/yr}$ that answers the following question: *What is the characteristic strength $h_c=h_{3/yr}$ of a source with sufficiently large S/N (equation (33)) that, if it is seen three times per year (once each 10^7 seconds) by two identical detectors operating in coincidence, we can be 90% confident the detectors are not just seeing their own noise.* The use of two detectors permits (we shall presume) the elimination of non-Gaussian noise. Then the Gaussian distribution of the amplitude noise, together with the use of two detectors and the fact that in $\hat{\tau}=10^7$ s the experimenters must try roughly $2\pi f_c\hat{\tau}\simeq f_c/10^{-8}$ Hz starting times t_o for their Wiener optimal filter, implies that we require $S/N\simeq[\ln(f_c/10^{-8}\text{ Hz})]^{\frac{1}{2}}$; and, correspondingly,

$$h_{3/yr}\simeq\left[\ln\left(\frac{f_c}{10^{-8}\text{ Hz}}\right)\right]^{\frac{1}{2}}\frac{[f_cS_h(f_c)]^{\frac{1}{2}}}{\langle F_+^2(\theta,\phi,\psi)\rangle^{\frac{1}{2}}}\simeq(3\text{–}5)h_n(f_c). \qquad (34)$$

Here the range 3–5 corresponds to the range of frequencies of interest for most burst searches, 10^{-4} Hz–10^{+4} Hz. Note that because the events being sought are so far out on the tail of the Gaussian probability distribution, changing by a factor 10 or 100 the number of trial starting times t_o, or asking for 99% or 99.9% confidence rather than 90%, would have a negligible effect on this $h_{3/yr}$.

It is often useful to rewrite the f_c and h_c of equations (31) in terms of the energy flux per unit frequency $\mathrm{d}E_{GW}/\mathrm{d}A\,\mathrm{d}f$ carried past the detector by the waves. From equation (10) for the waves' stress-energy tensor together with Parseval's theorem we infer that

$$\frac{\mathrm{d}E_{GW}}{\mathrm{d}A\,\mathrm{d}f}=\frac{\pi}{2}f^2(|\tilde{h}_+(f)|^2+|\tilde{h}_\times(f)|^2), \qquad (35)$$

where the extra factor 2 comes from folding negative frequencies into positive (so $f>0$). When this quantity is averaged over all directions ι, β, it gives $(4\pi r_o^2)^{-1}\,\mathrm{d}E_{GW}/\mathrm{d}f$, where r_c is the distance to the source; and consequently equations (31) become

$$f_c=\left[\int_{-\infty}^{\infty}\frac{\mathrm{d}E_{GW}/\mathrm{d}f}{fS_h(f)}\,\mathrm{d}\ln f\right]^{-1}\left[\int_{-\infty}^{\infty}\frac{\mathrm{d}E_{GW}/\mathrm{d}f}{S_h(f)}\,\mathrm{d}\ln f\right], \qquad (36a)$$

$$h_c = \left[\int_{-\infty}^{\infty} \frac{S_h(f_c)}{S_h(f)} \frac{3}{2\pi^2 r_o^2} \frac{\mathrm{d}E_{\mathrm{GW}}}{\mathrm{d}f} \,\mathrm{d}\ln f\right]^{\frac{1}{2}}. \tag{36b}$$

For most burst sources (e.g. supernovae) the wave form is so uncertain that a careful calculation of h_c is unjustified. In such cases it is useful to reexpress h_c (equation (36b)), approximately, in terms of the total energy ΔE_{GW} radiated:

$$h_c \simeq \left(\frac{3}{2\pi^2} \frac{\Delta E_{\mathrm{GW}}/f_c}{r_o^2}\right)^{\frac{1}{2}} = 2.7 \times 10^{-20} \left(\frac{\Delta E_{\mathrm{GW}}}{M_{\odot}c^2}\right)^{\frac{1}{2}} \left(\frac{1\,\mathrm{kHz}}{f_c}\right)^{\frac{1}{2}} \left(\frac{10\,\mathrm{Mpc}}{r_o}\right), \tag{37}$$

where $M_{\odot}$ is the mass of the sun and 10 Mpc is the distance to the center of the Virgo cluster of galaxies (assuming a Hubble constant of 100 km s^{-1} Mpc^{-1}).

We turn now to a discussion of specific burst sources, and their characteristic frequencies f_c and wave strengths h_c.

(c) *Supernovae (collapse to neutron star)*

Supernovae of 'type II' are believed, with a high level of confidence, to be created by the gravitational collapse, to a neutron-star state, of the cores of massive, highly evolved stars. Supernovae of 'type I', by contrast, are thought to result from nuclear explosions of white dwarfs that are accreting mass from close companions – explosions in which the stellar core probably does not, but might, collapse to a neutron-star state (Woosley and Weaver, 1986; Evans, Iben and Smarr, 1987, and references therein). In addition to these two types of optically observed supernovae, there may well be stellar collapses to a neutron star that produce little optical display ('optically silent supernovae').

The rate of occurrence of supernovae of types I and II is fairly well determined observationally: in our Galaxy roughly one type I each 40 years and one type II each 40 years; out to the distance of the center of the Virgo Cluster of Galaxies (10 Mpc) several type I and several type II per year; rate increasing roughly as (distance)^3 near and beyond Virgo; see, e.g. Tammann (1981). Thus, to have an interesting event rate one must have adequate sensitivity to reach the center of Virgo, about 10 Mpc. Optically silent supernovae could be more frequent, since statistics on normal stars massive enough to form neutron stars when they die, permit a neutron-star birth rate that could be as high as one every four years in our galaxy if mass loss in the late stages of stellar evolution is smaller than normally thought. Alternatively, it is conceivable that optically silent supernovae never occur.

The strengths of the waves from a supernova depend crucially on the degree of non-sphericity in the stellar collapse that triggers it, and somewhat on the speed of collapse – i.e. on whether the collapse is nearly free-fall ('cold collapse') or is more gentle due to the resistance of thermal pressure ('hot collapse'). Perfectly spherical collapse will produce no waves; highly non-spherical collapse will produce strong waves. Little is known about the degree of non-sphericity in type II (which are surely due to stellar collapse), but current prejudice suggests that the typical type II might be quite spherical and thus poorly radiating. If type I are due to explosion of an accreting white dwarf and, contrary to current thought, the explosion is accompanied by collapse of the stellar core to a neutron star, then the white dwarf might be rapidly rotating due to the accretion, and centrifugal forces might then cause it to collapse very non-spherically and radiate strongly.

In the mid-1970s there was a swing of fashion from believing that the collapse is cold and fast to believing it is hot and slow (e.g. Wilson, 1974; Schramm and Arnett, 1975). Some astrophysicists were aghast at the consequence of this swing: for example, in the highly non-spherical but axisymmetric collapse models of Saenz and Shapiro (1978, their tables 1–4) the total energy radiated as gravitational waves was reduced from $\Delta E_{\mathrm{GW}}/M_{\odot}c^2 \sim 6\times10^{-3}$ for cold and fast to $\sim 1\times10^{-5}$ for hot and slow. This boded ill for attempts to detect gravitational waves, the astrophysicists thought. However, closer scrutiny revealed relatively little change in the prospects for detection: the total energy radiated ΔE_{GW} is a rather poor indicator of detectability. Much more relevant is the amplitude signal-to-noise ratio $S/N = h_c/h_n(f_c)$ (equation (33)). The planned LIGO beam detectors, when optimized for frequency f_c, have $h_n(f_c) \propto f_c$ (Fig. 9.4 and equation (125a)); and this together with equation (37) gives

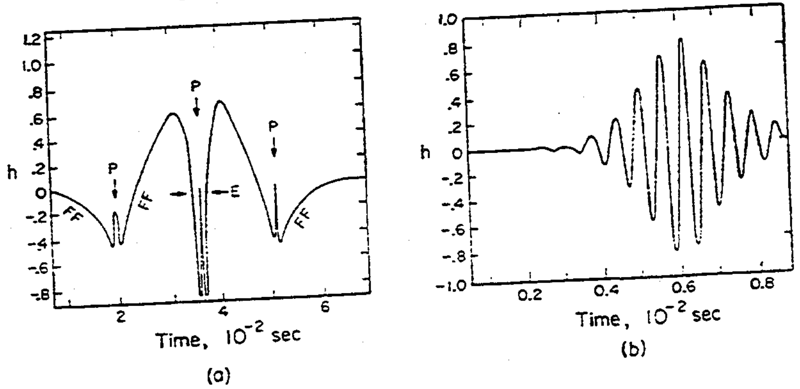
$$S/N \propto (\Delta E_{\mathrm{GW}}/f_c^3)^{\frac{1}{2}}. \tag{38}$$

Although the new fashion (hot and slow) corresponded to a reduction in ΔE_{GW} by a factor 600, the characteristic frequency of the waves also went down (from $\simeq 3000$ Hz to $\simeq 500$ Hz; see Saenz and Shapiro, 1978); and, correspondingly, S/N was reduced by less than a factor 2.

This illustrates the importance of thinking in terms of h_c, f_c, and $S/N = h_c/h_n(f_c)$ rather than in terms of energy radiated, when evaluating the strengths of gravitational waves.

Corresponding to our poor knowledge of the strengths of the waves from supernovae is a similar poor knowledge of the wave forms. It seems unlikely that theorists will be able to firm up their predictions of the waveforms before observers detect and study them. Thus, it is best to think of the computed wave forms as giving us a tool for translating future observations into an understanding of what is happening in the stellar core. As an example consider Fig. 9.3, which shows the wave forms computed by Saenz and Shapiro (1978, 1981), using the quadrupole formalism, for two different stellar collapses. In each case – and in general, when one thinks the quadrupole formalism may be accurate – one can invert equation (12) to see how the source's quadrupole moment is behaving. The left curve (Saenz and Shapiro, 1978) shows several epochs labeled FF in which $h_+(t)$ varies approximately as $|t-t_0|^{-\frac{2}{3}}$, corresponding to free-fall motion; and these free-fall epochs are separated by three brief periods with sharply reversed peaks (labeled 'P' in the diagram) corresponding to a sharp acceleration in the direction opposite to the free fall. Considering the timescales of hundredths of a second for the free-fall epoch (characteristic of late stages of collapse to a neutron star) and ~0.5 milliseconds for the peak epochs (characteristic of neutron-star pulsation), the natural and correct interpretation is that these waves are from collapse to a neutron star in which the stellar core bounced sharply three times. The fact that the three sharp peaks are all in the same direction (up, not down) indicates that the sharp bounces were all along the same axis. Surely the other axis that projects on our sky should have bounced as well, or at least stopped its collapse; so there should be at least one sharp peak in the down direction. Indeed there is; it is superposed on the central up peak (region labeled E in the diagram). The natural and correct interpretation is that the star was centrifugally flattened by rotation; its pole collapsed fast and bounced three times (up peaks P) while its equator collapsed more slowly and bounced once (down peak E).

Fig. 9.3. Wave forms produced by two very different scenarios for the collapse of a normal star to form a neutron star. Wave form (*a*) is from Saenz and Shapiro (1978); (*b*) is from Saenz and Shapiro (1981).

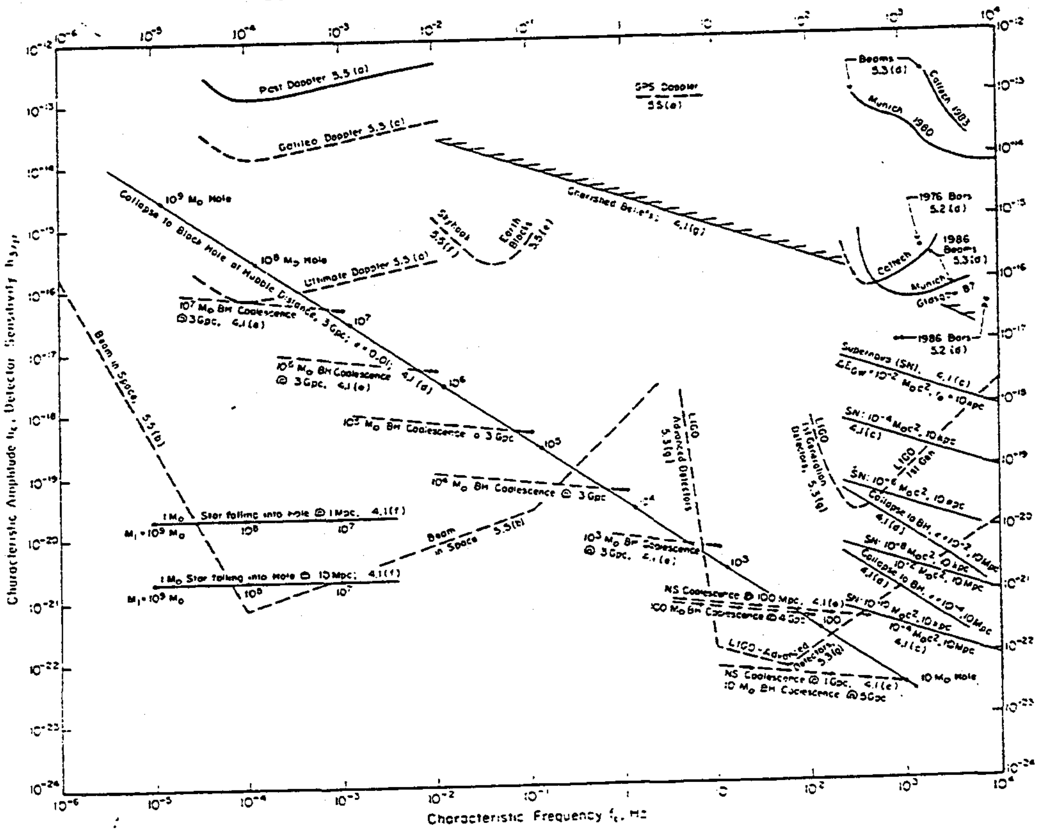


The very different right-hand curve in Fig. 9.3 (Saenz and Shapiro, 1981) implies a quadrupole moment that somehow is driven into sinusoidal oscillations which initially increase in amplitude and then die out. Again, the fact that the period (~ 0.6 ms) is that of a neutron-star pulsation suggests that something is triggering, then damping, such a pulsation. If this wave form was seen roughly one day before an optical supernova was found in the Virgo cluster, one would infer that the waves were from the supernova and one could deduce that the quadrupole-moment oscillations are so large in amplitude that they must have absorbed, say, 10^{-4} of the collapse energy. The natural explanation might be parametric amplification of quadrupole neutron star pulsations by a bouncing stellar collapse, followed by hydrodynamic damping – the process that gave rise to this computed wave form.

A large amount of effort has gone into model calculations of gravitational collapse to a neutron star and the waves it emits; but the effort has not produced a consensus by any means! For a detailed review of the literature up to 1982 see Eardley (1983); for an update on that review see Müller (1984). In the case of rapidly rotating collapse, where the emission should be strongest but the event rate is totally unknown (most collapses *could* be slowly rotating), there are three radically different scenarios and corresponding wave characteristics: (i) the star may remain axisymmetric throughout the collapse. In this case the best current 'wisdom' (but by no means a consensus) comes from calculations by Müller (1982) and is pessimistic. Those calculations predict the strongest emission to come in two different spectral regions: $f_c \sim 1000$ Hz where $\Delta E_{GW} \sim 1 \times 10^{-7} M_\odot c^2$ and $h_c \sim 1 \times 10^{-23}$ (10 Mpc/r_o) due to the initial collapse and bounce; and $f_c \sim 10^4$ Hz where $\Delta E_{GW} \sim 10^{-6} M_\odot c^2$ and $h_c \sim 1 \times 10^{-23}$ (10 Mpc/r_o) due to pulsations of the newly formed neutron star. (ii) The star may become unstable to an '$m=2$ bar-mode' deformation so it rotates end-over-end like an American football. In this case the best current 'wisdom' is more optimistic: the calculations of Ipser and Managan (1984) predict a highly monochromatic emission at $f \sim 1000$ Hz, lasting for ~ 30 cycles and producing $\Delta E_{GW} \sim 3 \times 10^{-4} M_\odot c^2$ and $h_c \sim 5 \times 10^{-22}$ (10 Mpc/r_o). (iii) The collapsing star may become so strongly unstable to non-axisymmetric perturbations that on the way down it breaks up into two or more discrete lumps. Very little is known about this possibility, and radiation reaction from the $m=2$ mode (case ii) might prevent it from occurring at all (Ipser, 1986). Eardley (1983) argues that if it does occur, it may produce quite strong waves: $\Delta E_{GW} \sim$ (a few) $\times 10^{-2} M_\odot c^2$ at $f_c \sim 1000$ Hz, corresponding to $h_c \sim 4 \times 10^{-21}$ (10 Mpc/r_o).

Because this best wisdom is so insecure, Fig. 9.4 shows wave strengths based not on these specific models, but rather on the general equation (37) for several possible values of ΔE_{GW} and r_o, and for the entire range of characteristic frequencies that have shown up in model calculations, 200 Hz $\lesssim f_c \lesssim$ 10 000 Hz. Note that detectability depends strongly on whether the waves come off at low frequencies or high: a factor 8 reduction in ΔE_{GW} can be compensated by a factor 2 reduction in f_c.

Fig. 9.4. The characteristic amplitudes h_c (equation (31b)) and frequencies f_c (equation (31a)) of gravitational waves from several postulated *burst sources* (thin curves), and the sensitivities $h_{3/yr}$ of several existing and planned detectors (thick curves and circles) ($h_{3/yr}$ is the amplitude h_c of the weakest source that can be detected three times per year with 90% confidence by two identical detectors operating in coincidence). The abbreviations BH, NS and SN are used for black hole, neutron star and supernova. The sources are discussed in detail in the indicated subsections of Section 9.4.1, and the detectors in the indicated subsections of Section 9.5.



We note in passing that neutron-star pulsations might be excited not only as part of the star's birth throes, but also as a consequence of a sudden strain release or phase transition in an old neutron star (Thorne, 1978; Ramaty *et al.*, 1980; Haensel, Zdunik and Schaeffer, 1986). Since there are 10^8–10^9 old neutron stars in our galaxy, it is conceivable – though not highly likely – that our galaxy could produce an interesting event rate. To be detectable by the planned LIGOs, the waves would need to have $h_c \gtrsim 10^{-21}$ at $f_c \sim 3000$ Hz corresponding to

$$\Delta E_{\rm GW} \gtrsim 7 \times 10^{45}\ {\rm erg} \times (10\ {\rm kpc}/r_o)^2. \tag{39}$$

(d) *Collapse of a star or star cluster to form a black hole*

As with collapse to form a neutron star, so also for collapse to a black hole, the strengths of the waves produced are highly sensitive to the degree of non-sphericity, and the typical degree of non-sphericity is unknown. Equally unknown in the black-hole case, by contrast with the neutron star, is the frequency of occurrence of such collapses:

It is very likely that black holes exist in our universe with masses throughout the range $2\,{\rm M}_\odot \lesssim M \lesssim 10^{10}\,{\rm M}_\odot$ (see Chapter 8 of this book). The holes of lowest mass can only form by direct collapse of a star. Those of higher mass, however, can form by many routes (direct collapse; gradual growth from a small hole by accretion; collision and coalescence of smaller holes; . . .). For discussions of the routes that might occur in a dense galactic nucleus see Blandford (1979) and Rees (1983). Which routes actually occur and how often are almost totally unknown. However, roughly known upper limits on the birth rates of the smallest and the largest holes give – under the assumption that all births are by direct collapse – corresponding upper limits on the rate at which gravity-wave bursts from such collapses hit the earth. For holes of a few solar masses the birth rate under reasonable assumptions (Section 8.3.3 of Chapter 8) should not exceed $\sim\frac{1}{3}$ the birth rate of neutron stars; and correspondingly, at the distance of the Virgo cluster it should not exceed ~ 1/year. Bethe (1986) argues that the rate may actually be of this magnitude. At the other end of the spectrum, holes with masses $M \gtrsim 10^6\,{\rm M}_\odot$ probably occur only in galactic nuclei; and over its lifetime each galactic nucleus might give birth, at maximum, to only a few such holes. Correspondingly (Thorne and Braginsky, 1976; Blandford, 1979), the maximum rate of collapse-births of supermassive holes, $M \gtrsim 10^6\,{\rm M}_\odot$, is a few per year throughout the observable universe – i.e. out to the Hubble distance. It is fashionable to believe that the actual rate is much less than this upper limit (e.g. Rees, 1983).

In one respect collapse to a black hole is better understood than collapse to a neutron star: the final object is far simpler, and correspondingly the waves from its vibrations, if they are triggered by the collapse, are far better understood. Detailed calculations suggest, in fact, that black-hole vibrations are rather easy to trigger (e.g. Detweiler, 1977) and that when they are triggered, the most slowly damped one or two quadrupole modes will dominate. Thus, while the details of the initial burst of waves may depend on unknown details of the collapse, the late-time behavior will have a well-established damped oscillatory form from which one can read off the mass of the hole with excellent accuracy and its angular momentum with modest accuracy (Detweiler, 1980; Leaver, 1985, 1986*a*; Stark and Piran, 1986; Piran and Stark, 1986).

As a specific example, Fig. 9.5 shows the waves produced by a specific model of a collapsing, rotating star as computed using numerical relativity techniques by Stark and Piran (1986). The solid curve is the computed waveform, and the dashed curve is a fit to it using a mixture of the two most weakly dampled quadrupole modes of a non-rotating hole. The fact that the fit is so good shows (i) that the final ringdown is, indeed, due to the black-hole vibrations; and (ii) that the hole was not rotating extremely rapidly, i.e. it had $(1-a/M)\gtrsim 0.3$, where a is the specific-angular momentum and M the mass. If the hole had had $1-a/M<0.3$, the 'Q' of the hole's oscillations would have been noticeably larger, and the ringdown of the waves would have been noticeably slower. (For details of the normal modes of black holes and the frequencies and damping times of the waves they should produce see Detweiler, 1980, and Leaver, 1985, 1986*a*.)

If collapse to a black hole radiates with an efficiency $\Delta E/Mc^2\equiv\varepsilon$ and the hole is at a distance r_o and has a mass M, then the characteristic frequency and amplitude of its waves will be (Stark and Piran, 1986; Piran and Stark, 1986; and equation (37) above)

$$f_c\cong\frac{1}{5\pi M}=(1.3\times 10^4\ \mathrm{Hz})\left(\frac{M_\odot}{M}\right), \tag{40a}$$

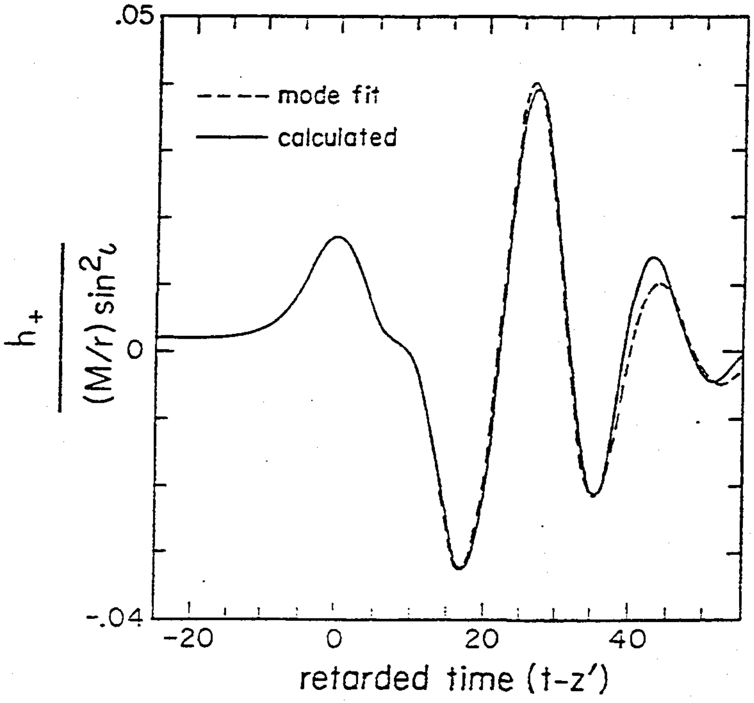
$$\begin{aligned} h_c &\cong\left(\frac{15}{2\pi}\varepsilon\right)^{\frac{1}{2}}\frac{M}{r_o}=7\times 10^{-22}\left(\frac{\varepsilon}{0.01}\right)^{\frac{1}{2}}\left(\frac{M}{M_\odot}\right)\left(\frac{10\ \mathrm{Mpc}}{r_o}\right)\\ &=1.0\times 10^{-20}\left(\frac{\varepsilon}{0.01}\right)^{\frac{1}{2}}\left(\frac{10^3\ \mathrm{Hz}}{f_c}\right)\left(\frac{10\ \mathrm{Mpc}}{r_o}\right). \end{aligned} \tag{40b}$$

If the collapse is axisymmetric, then the efficiency ε probably does not exceed 7×10^{-4} (Stark and Piran, 1986). However, in the non-axisymmetric case (e.g. formation of an elongated configuration due to rapid rotation, or bifurcation into one or more lumps during collapse (the 'collapse, pursuit, and plunge' scenario of Ruffini and Wheeler, 1971)), the efficiency might be in the range 0.01–0.1 (see, e.g., Eardley, 1983, and Rees, 1983). The source characteristics (40) are shown in Fig. 9.4 for black-hole births at the Hubble distance and at the distance of Virgo, with efficiencies of $\varepsilon = 10^{-2}$ and 10^{-4}.

(e) *Coalescence of compact binaries (neutron stars and black holes)*

Since a large fraction of all stars are in close binary systems, the dead remnants of stellar evolution may contain a significant number of binary

Fig. 9.5. The gravitational wave form produced by the gravitational collapse of an axisymmetric rotating star to produce a Kerr black hole, as computed by Piran and Stark (1986) using numerical relativity techniques. The star and the black hole it forms both have $J/M^2 \equiv a/M = 0.63$ (where J is angular momentum and M is mass). The dashed curve is a fit, to the wave form, of a superposition of the waves from the two most slowly damped quadrupolar normal modes of a non-rotating hole with mass M, $h_+ \sim \mathrm{Real}\{A_1 e^{-i\omega_1 t} + A_2 e^{-i\omega_2 t}\}$ with $\omega_1 = (0.374 - 0.089i)/M$, $\omega_2 = (0.348 - 0.274i)/M$. The fitting amplitudes are $A_1 = -0.9 - 1.1i$, $A_2 = 0.9 + 1.4i$. Thus, these two modes are roughly equally excited by the collapse.



systems whose components are neutron stars or black holes, and are close enough together to be driven into coalescence by gravitational radiation reaction in a time less than the age of the universe. The binary pulsar PSR 1913 + 16 is an example of such a system; it will coalesce 3.5×10^8 years from now.

As the two bodies in a compact binary spiral together, they emit periodic gravitational waves with a frequency that sweeps upward toward a maximum,

$$f_{\max}\simeq 1\ \text{kHz for neutron stars,} \tag{41a}$$

$$f_{\max}\simeq\frac{10\ \text{kHz}}{M_1/M_\odot}\ \text{for holes with the larger having mass } M_1. \tag{41b}$$

The wave form during the nearly Newtonian part of the frequency sweep, $f\ll f_{\max}$, is easily computed from the quadrupole formalism (Section 9.3.2). The post-Newtonian corrections to this waveform will become more and more important as f rises toward $f_{\max}$; they are given in Wagoner and Will (1976); see also Gal'tsov, Matiukhin and Petukhov (1980). Ultimately, near $f_{\max}$, higher-order corrections or full non-linear relativity are needed to get the wave form reasonably accurately. The final, coalescence stage will be especially interesting and complex in the case of a neutron-star binary, and may be quite sensitive to the masses of the two stars; and as with supernovae, we might not understand reliably what to expect until gravitational-wave observations show us. For a first, preliminary theoretical effort at understanding, see Clark and Eardley (1977). For black holes, by contrast, numerical relativity is likely to give us, within the next five years or so, a detailed and highly reliable picture of the final coalescence and the wave forms it produces, including the dependence on the hole's masses and angular momenta. Comparison of the predicted wave forms with observed ones will constitute the strongest test ever of general relativity. (The wave forms for the astrophysically unlikely cases of head-on collisions of two identical non-rotating holes or neutron stars have already been evaluated by numerical relativity; see Smarr, 1977*a* for black holes, and Gilden and Shapiro, 1984 for neutron stars.)

Because the binary system spends far more time in the early, low-frequency part of the sweep than in the later, high-frequency part or in the final coalescence, and because planned gravity wave detectors have less amplitude noise at low frequencies, $f\sim100$ Hz, than at high, $f\gg100$ Hz (cf. Fig. 9.4), it will be easier for detectors to see the Newtonian regime of the sweep than the post-Newtonian regime or the final coalescence – except in the case of black-hole binaries with $M_1 \sim 100$–$1000\, M_\odot$ or $M_1 \gtrsim 10^6\, M_\odot$ (Fig. 9.4). In the Newtonian regime, if we orient the polarization axes $\vec{e}_{x'}$ and $\vec{e}_{y'}$ along the major and minor axes of the projection of the orbital plane on the sky, then the wave form will be

$$h_+ = 2(1+\cos^2 \imath)(\mu/r)(\pi M f)^{\frac{2}{3}} \cos(2\pi f t), \tag{42a}$$

$$h_\times = \pm 4 \cos \imath (\mu/r)(\pi M f)^{\frac{2}{3}} \sin(2\pi f t). \tag{42b}$$

Here it is assumed that the orbit is circular because radiation reaction long ago will have forced circularization (Peters and Mathews, 1963); $\imath$ is the angle of inclination of the orbit to the line of sight; M and μ are the total and reduced masses

$$M = M_1 + M_2, \qquad \mu = M_1 M_2 / M; \tag{42c}$$

and f, the frequency of the waves (equal to twice the orbital frequency), is given as a function of time by (MTW equation (36.17))

$$f = \frac{1}{\pi}\left[\frac{5}{256}\frac{1}{\mu M^{\frac{2}{3}}}\frac{1}{(t_o - t)}\right]^{\frac{3}{8}}. \tag{42d}$$

The most promising detectors for coalescing neutron-star binaries and low-mass black-hole binaries are beam detectors in the planned multi-kilometer LIGOs. As we shall see in Section 9.5.3(e), a beam detector can be operated in several different optical configurations. The optimum configuration for searching for coalescing binaries is likely to be one with *light recycling*, for which the spectral density of shot noise (the dominant noise above some 'seismic cutoff' frequency f_s) will have the form

$$S_h(f) = \text{const} \times f_k[1 + (f/f_k)^2] \quad \text{at } f > f_s. \tag{43a}$$

Here f_k is a 'knee frequency' which the experimenters can adjust by changing the reflectivities of certain mirrors in their detectors; see equation (117c) and Fig. 9.13 below, and associated discussion. The constant in equation (43a) is independent of the choice of f_k. At frequencies below the 'seismic cutoff' f_s seismic noise is likely to come on very strong; accordingly, we shall make the approximation

$$S_h(f) = \infty \quad \text{for } f < f_s. \tag{43b}$$

By Fourier-transforming the wave forms (42), squaring, and averaging over the source orientation angle $\imath$, we obtain

$$\langle |\tilde{h}_+|^2 + |\tilde{h}_\times|^2 \rangle = \frac{\pi}{12}\left(\frac{\mu}{r}\right)^2 \frac{M^3}{\mu} \frac{1}{(\pi M f)^{\frac{7}{3}}}. \tag{44}$$

By inserting this source strength (44) and the detector noise (43) into equation (30) and maximizing the resulting signal-to-noise ratio with respect to the knee frequency f_k, we find that the experimenter will do best to choose

$$f_k = 1.44 f_s. \tag{45}$$

For smaller choices of f_k there is not a wide enough frequency band between the seismic cutoff and the knee to take optimal advantage of the broad-band nature of the signal. For larger values of f_k the experimenter loses because the height $S_h(f)$ of the 'noise floor' at $f_s < f \lesssim f_k$ (equation (43a)) is proportional to f_k. With this choice of knee, equations (31a) and (43)–(45) give for the characteristic frequency $f_c = 0.909 f_k$. Below (equation (125a)) we shall characterize the sensitivites of beam detectors in full scale LIGOs, when searching for bursts, by the noise amplitude h_n *at the knee*; and, correspondingly, we here shall set f_c (which after all is somewhat arbitrary) to f_k rather than $0.909 f_k$:

$$f_c = f_k = 1.44 f_s. \tag{46a}$$

Equations (31b) and (43)–(45) then give for the characteristic amplitude of the waves from inspiraling binaries

$$h_c = 0.237 \frac{\mu^{\frac{1}{2}} M^{\frac{1}{3}}}{r_o f_c^{\frac{1}{6}}} = 4.1 \times 10^{-22} \left(\frac{\mu}{M_\odot}\right)^{\frac{1}{2}} \left(\frac{M}{M_\odot}\right)^{\frac{1}{3}} \left(\frac{100\,\text{Mpc}}{r}\right) \left(\frac{100\,\text{Hz}}{f_c}\right)^{\frac{1}{6}}. \tag{46b}$$

The characteristic amplitude (46b) is enhanced over the actual rms wave strength $\langle h_+^2(t_c) + h_\times^2(t_c) \rangle^{\frac{1}{2}}$ the waves have at the time $t = t_c$ when they sweep through frequency f_c – enhanced by very nearly the square root of the number of periods, $n = (f^2/\dot{f})_{f=f_c} = (5/96\pi)(M/\mu)(\pi M f_c)^{-\frac{5}{3}}$, that the binary spends in the vicinity of the frequency f_c. This $\sqrt{n} \simeq 28(\mu/M_\odot)^{-\frac{1}{2}}(M/M_\odot)^{-\frac{1}{3}}(f_c/100\,\text{Hz})^{-\frac{5}{6}}$ enhancement corresponds to the enhancement in effective signal that the experimenters will achieve by optimal signal processing in their search for these frequency-sweeping bursts.

From a study of the waveform (42) using broad-band detectors at several widely spaced locations on the earth, one can deduce the following information: (i) the direction to the source (which comes from phase differences in the signal at different detectors in different locations); (ii) the inclination of the orbit to the line of sight (which comes from the amplitudes in the two different polarization modes); (iii) the direction the stars move in their orbit (which comes from the + or − sign in equation (42b)); (iv) the combination $(\mu^3 M^2)^{\frac{1}{5}}$ of the reduced and total masses; and (v) the distance r to the source. If the mass combination $(\mu^3 M^2)^{\frac{1}{5}}$ is $\lesssim 1.5 M_\odot$, one can be fairly sure the binary was made of neutron stars; if it is much larger, one can be fairly sure that at least one of the bodies was a massive black hole.

Especially intriguing is the possibility (Schutz, 1986*b*) that, in the case of neutron stars the coalescence will produce electromagnetic emission (e.g. due to an explosion of the less massive star, Blinnikov *et al.*, 1984) that is strong enough to be detected at earth and thereby to pin down the source's location with far higher precision than can be obtained from the gravitational waves. In this case, a redshift will probably be obtainable from optical observations; and that redshift together with the gravitational-wave-determined distance r will give a value for the Hubble constant. Schutz (1986*b*), from a detailed study of the expected noise in future beam detectors, concludes that the prospects are good thereby to obtain a significantly better value for the Hubble constant than we now have. Even in the absence of electromagnetic signals from the coalescence, it may prove possible by statistical means to determine the Hubble constant using combined data for a number of coalescences; see Schutz (1986*b*) for details.

Clark, van den Heuvel and Sutantyo (1979) have estimated, from neutron-star observations in our own galaxy, that to see three coalescences of neutron-star binaries per year one must look out to a distance of 100^{+100}_{-40} Mpc, where the quoted uncertainties are at the 90% confidence level. Correspondingly, in Fig. 9.4 are shown the characteristic amplitude and frequency, for a range of values of the seismic cutoff f_s and corresponding values of $f_c = 1.44 f_s$, produced by the coalescence of two 1.4 solar mass neutron stars at 100 Mpc (estimated event rate about three per year) and at $\frac{1}{3}$ the Hubble distance (event rate about ten per day). Because much has been learned observationally about the statistics of neutron-star binaries since these estimates of event rates were made, a careful restudy of the estimates is much needed.

As Fig. 9.4 shows, future earth based beam detectors may be able to see black-hole coalescences throughout the universe, so long as the more massive of the two holes does not exceed $1000\,M_\odot/(1+z)$, where z is the hole's cosmological redshift. The coalescence rate for black-hole binaries of a few solar masses could be of order that for neutron-star binaries (a few per year at 100 Mpc), or might well be far lower. Particularly intriguing is a scenario, suggested as very plausible by Shapiro and Teukolsky (1985) and Quinlan and Shapiro (1987), in which a large fraction of galactic nuclei create, at some phase of their evolution, a dense cluster of neutron stars and small-mass black holes which – on a timescale of only a few years – form tight binaries that coalesce, with the coalesced holes then forming new tight binaries that coalesce, . . . until the cluster goes unstable and collapses to form a single large hole. This scenario suggests that in typical years the earth might be hit by a number of spiraling wave bursts from coalescing binaries of masses $3\,M_\odot$–$1000\,M_\odot$ at the Hubble distance, $z \sim 1$. Also intriguing but much less likely is a scenario discussed by Bond and Carr (1984) in which a sizable fraction of the mass of the universe is in black-hole binaries with masses ~ 100–$1000\,M_\odot$ for which the coalescence rate could be several per year in the local group of galaxies (distance ~ 1 Mpc).

One or more black hole binaries of any mass up to $\sim 10^8\,M_\odot$ *might* have formed in the nuclei of a reasonable fraction of all galaxies during the past life of the universe, leading to event rates $\gtrsim 1$/year out to the Hubble distance (cf. Fig. 1 of Rees, 1983) – or they might never form. There is actually observational evidence for supermassive black-hole binaries formed by the coalescence of galactic nuclei (Begelman, Blandford and Rees, 1980; Rees, 1983); but the rate of such events probably does not exceed $\frac{1}{100}$ years out to the Hubble distance (Rees, 1983).

(f) *The fall of stars and small holes into supermassive holes*

The supermassive ($M_1 \gtrsim 10^5\,M_\odot$) black holes thought to inhabit the nuclei of galaxies might typically grow by accretion on timescales as short as 10^8 years; see, e.g. Section 8.6 of Blandford and Thorne (1979). When such a hole grows larger than $10^9\,M_\odot$, normal stars can pass near or plunge through its horizon without being torn apart tidally, and the number of stars that so scatter or plunge could well be of order one per year or more (e.g. Dymnikova, Popov and Zentsova, 1982). For smaller supermassive holes, scattering or plunging normal stars will be tidally disrupted, reducing the strength of their waves; but the reduction will not be great, at least in the case of radial infall, unless the hole is below $10^6\,M_\odot$ (Nakamura and Sasaki, 1981; Haugan, Shapiro and Wasserman, 1982). For any hole, neutron stars and satellite holes can scatter or plunge through without enough disruption to strongly suppress their radiation; but the event rate (per supermassive hole) will typically be well below one per year.

The wave forms emitted when a star or small hole is scattered by or plunges into a supermassive hole have been evaluated with high precision using perturbation formalisms; see, e.g. Detweiler and Szedenits (1979), Kojima and Nakamura (1984*a*). The characteristic frequency and amplitude for typical (non-head-on) impact parameters are

$$f_c \cong \frac{1}{20 M_1} = 10^{-4}\,\mathrm{Hz}\left(\frac{10^8\,M_\odot}{M_1}\right), \tag{47a}$$

$$h_c \cong \frac{M_2}{2r_o} = 2\times 10^{-21}\left(\frac{M_2}{M_\odot}\right)\left(\frac{10\ \text{Mpc}}{r_c}\right), \tag{47b}$$

where M_1 is the mass of the large hole. M_2 is that of the infalling body, and $r_o \sim 10$ Mpc might give a reasonable event rate since there are ~ 100 galaxies as massive as or more massive than our own inside this distance, including M87 for which observational data suggest a central black hole of mass $M_1 \sim 4\times 10^9\ M_\odot$ (Section 8.3.1.4 of Chapter 8). These h_c and f_c are plotted in Fig. 9.4 for several interesting sets of parameters. It is conceivable that such plunge bursts will be seen by beam detectors in space, if and when they are flown.

(g) *Cherished beliefs*

The above discussion makes clear how uncertain is our electromagnetically based knowledge of gravitational-wave sources. Correspondingly, it seems very likely that when gravitational waves are finally seen, they will come predominantly from sources we have not thought of or we have underestimated; and it seems quite possible that the waves will be stronger than the above estimates suggest.

In light of this, it is interesting to ask the following question: How strong could be the strongest bursts that strike the earth on average three times per year without violating our 'cherished beliefs' about the laws of physics and the universe? Zimmermann and Thorne (1980) have enumerated a set of cherished beliefs, including (i) that general relativity is correct, (ii) that we do not live in a special time or place in the universe, (iii) that there are no enormous primordial bursts, (iv) that no single, coherently radiating object in our galaxy has a mass exceeding $10^8\ M_\odot$, and (v) that the strongest bursts were not beamed by their sources into solid angles $\ll \pi$. From these cherished beliefs they have derived the upper limit on the 3/year burst strength which is shown in Fig. 9.4. That limit could actually be achieved at frequencies $f \gtrsim 30$ Hz by an unlikely but not implausible scenario in which a large fraction of the mass of the universe was cycled, long ago, through a pre-galactic population of massive stars ('Population III' stars), leaving much of the universe's mass in compact binary systems that inhabit the halos of galaxies. If the mean lifetime of these binaries against spiraling together and coalescing is of order the age of the universe, then such coalescences in the halo of our own Milky Way galaxy would give bursts in the frequency domain $f \gtrsim 30$ Hz at the cherished belief level. For further discussion see Zimmermann and Thorne (1980), and Bond and Carr (1984).

While this scenario is unlikely, it is not totally implausible, and it serves to remind us that our best estimates of the waves bathing the earth could be grossly pessimistic.

9.4.2 Periodic sources

(a) *Characterization of the waves from periodic sources and the noise in a Detector searching for them*

The gravitational waves from a periodic source will be characterized by a discrete set of frequencies, and the waves at a given frequency will typically be right-hand or left-hand elliptically polarized; i.e. for some suitable choice of polarization axes $\vec{e}_{x'}$, $\vec{e}_{y'}$, they will have the form (similar to that of a decaying binary, equation (42), but with constant frequency)

$$h_+(t)=h_{o+}\cos 2\pi ft, \qquad h_\times(t)=\pm h_{o\times}\sin 2\pi ft. \tag{48}$$

If one wishes to quantify in a precise and standard manner all the properties of these waves, including the orientations of the 'preferred' x' and y' polarization axes, one might best do so using 'Stokes Parameters' analogous to those used in electromagnetic theory (Section 15 of Chandrasekhar, 1950). However, in this chapter we shall be concerned only with the frequencies f of the waves, and at a given emitted frequency, with a suitably defined characteristic amplitude h_c and a corresponding noise amplitude h_n in a detector searching for the waves.

As an aid in defining h_c and h_n, consider the following situation (analog of that for burst sources in Section 9.4.1(a)): a theorist tells us the frequency f, the phase, and the amplitudes $h_{o+}(\iota, \beta, r)$ and $h_{o\times}(\iota, \beta, r)$ to be expected from a specific model for a source, with orientation angles, ι, β and distance r. Suppose, further, that this type of source is distributed randomly throughout the universe and that the mean number of sources inside the distance r_o is n_o. What, then, will be (on average) the signal-to-noise ratio for the n_oth brightest source that a detector, broad-band or narrow, will see in a search (at known frequency and phase) lasting a time $\hat{\tau}$? By virtue of equations (27) and by analogy with equation (33) for burst sources, the answer turns out to be

$$\frac{S}{N}\cong\frac{h_c}{h_n(f)}. \tag{49}$$

Here

$$h_c\equiv(2/3)^{\frac{1}{2}}|h_{o+}(\iota, \beta, r_o)|^2+|h_{o\times}(\iota, \beta, r_o)|^2\rangle^{\frac{1}{2}} \tag{50}$$

(with $\langle\cdots\rangle$ denoting an average over ι and β) is the characteristic amplitude of the periodic source (analog of (31b) for a burst source), and

$$h_n(f) \equiv \frac{[S_h(f)/\hat{\tau}]^{\frac{1}{2}}}{\langle |F_-(\theta, \phi, \psi)|^2 \rangle^{\frac{1}{2}}} \tag{51}$$

is the noise amplitude. (The h_c of equation (50) is $(\frac{4}{3})^{\frac{1}{2}}$ larger than one naively would expect from equation (27). This factor $(\frac{4}{3})^{\frac{1}{2}}$ is an approximate correction for the fact that the angle averages in S/N should not be over squares but, rather, over squares associated with the rotation of the earth during data collection, then over $(\text{squares})^{\frac{1}{2}}$ covering the rest of the sky and the orientation of the source, followed by a $\frac{1}{3}$ power after averaging; cf. the discussion preceding equation (30).) Correspondingly, if experimenters wish to be 90% confident of having seen that n_0th brightest source after $\frac{1}{3}$ year of search, then S/N must exceed $1.655 \simeq 1.7$ (Gaussian probability distribution), and h_c must exceed

$$h_{3/\mathrm{yr}} = 1.7 h_n = \frac{1.7}{\langle |F_+|^2 \rangle^{\frac{1}{2}}} [S_h(f) \times 10^{-7}\ \mathrm{Hz}]^{\frac{1}{2}} \text{ if } f \text{ and phase are known.} \tag{52a}$$

In the case that theory and electromagnetic observation have failed to tell us in advance the phase and frequency of the source, except to within Δf, the experimenters must try $\sim f/\Delta f$ values of the frequency, and correspondingly the Gaussian statistics of the noise will produce

$$h_{3/\mathrm{yr}} \simeq [2 \ln(f/\Delta f)]^{\frac{1}{2}} h_n(f) = \frac{[2 \ln(f/\Delta f)]^{\frac{1}{2}}}{\langle F_+^2 \rangle^{\frac{1}{2}}} [S_h(f) \times 10^{-7}\ \mathrm{Hz}]^{\frac{1}{2}}$$

$$\text{if } f \text{ is known only to within } \Delta f \sim f. \tag{52b}$$

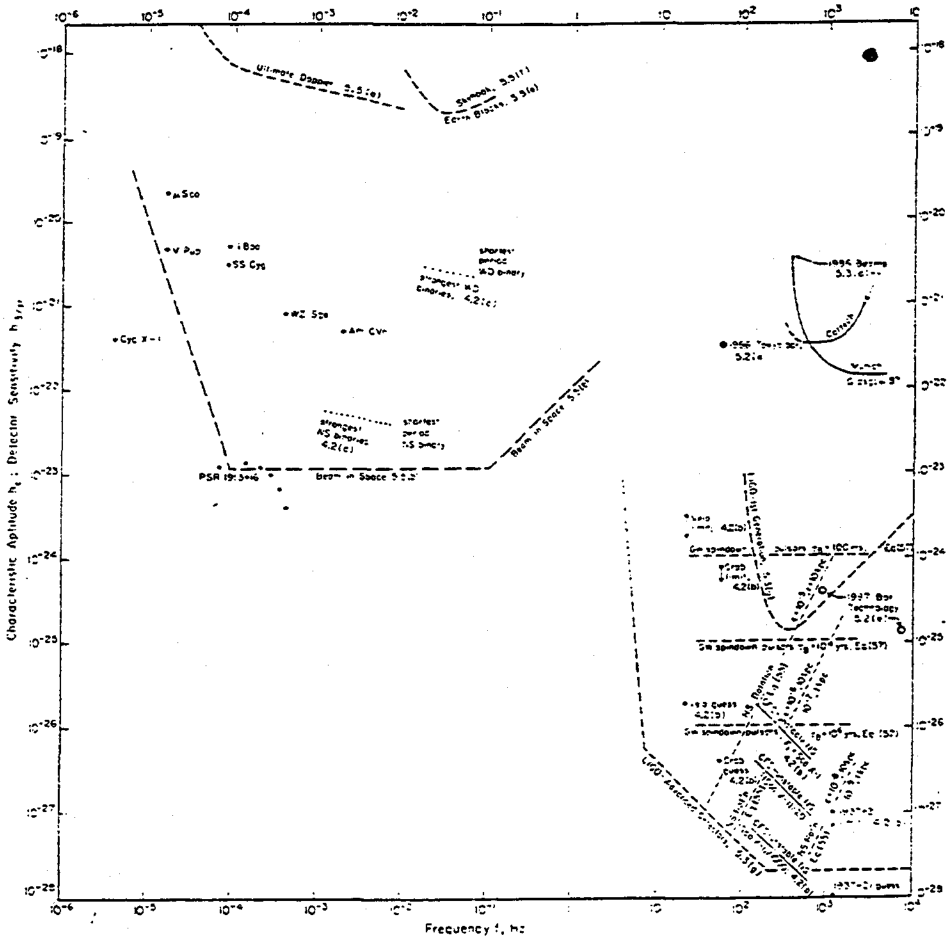
Fig. 9.6 shows h_c, for several postulated types of source (to be discussed in the following subsections), and correspondingly $h_{3/\mathrm{yr}}$ with f and the phase known, for several types of detectors (to be discussed in Section 9.5 below).

(b) *Rotating neutron stars*

A rotating neutron star (e.g. a pulsar) will emit gravitational waves at several frequencies as a result of deviations from symmetry around its rotation axis (deviations from 'axisymmetry'). The larger are those deviations and the more rapidly they rotate, the stronger will be the radiation.

Deviations from axisymmetry could arise in several ways: (i) The star's solid crust (well-established) or a solid core (not so well established) could support deformations that are residual remnants of the star's past history – a history that might be quite complex, including star quakes in which the crust or core cracks and deforms in much the manner of the solid earth in an earthquake; for detailed discussions see, e.g. Pandharipande, Pines and

Fig. 9.6. The characteristic amplitudes h_c (equation (50)) and frequencies f of waves from several postulated *periodic sources* (thin curves), and the sensitivities $h_{3\,\mathrm{yr}}$ of several existing and planned detectors (thick curves and circles) ($h_{3\,\mathrm{yr}}$ is the amplitude h_c of the weakest source detectable with 90 % confidence in a $\frac{1}{3}$ yr $= 10^7$ s integration if the frequency and phase of the source are known in advance; equation (52a)). The sources shown in the high-frequency region, $f \gtrsim 10$ Hz, are all special cases of rotating, nonaxisymmetric neutron stars (Section 9.4.2(b)). The steeply sloping dotted lines labeled NS Rotation refer to rigidly rotating neutron stars with moment of inertial $I_{zz} = 10^{45}$ g cm^{-2}, and with various ellipticities ε and distances r labeled on the lines (equation (55)). The sources in the low-frequency region, $f \lesssim 0.1$ Hz, are all binary star systems in our galaxy (Section 9.4.2(c)): several specific, known binaries, which are indicated by name (μ Sco, V Pup, . . .); the strongest six spectral lines from the famous binary pulsar PSR1913 + 16; and the estimated strengths of the strongest white-dward ('WD') and neutron-star ('NS') binaries in our galaxy. The detectors are discussed in detail in the indicated subsections of Section 9.5.



Smith (1976) and references therein. In old pulsars that have been spun up by accretion to near-millisecond rotation rates, theory and observational data suggest that the crust and core are quite well annealed into a nearly axisymmetric shape (Alpar and Pines. 1985); but in neutron stars that are only tens or hundreds or thousands of years old, it might well be otherwise (e.g. Zimmermann, 1978). (ii) The star's internal magnetic field, if sufficiently strong, could produce sufficient magnetic pressure to distort the star significantly (Zimmermann, 1978; Gal'tsov, Tsvetkov and Tsirulev, 1984). However, 'sufficiently strong' means, in the case, e.g., of the Crab Pulsar, ten times stronger than the star's measured surface field. (iii) If the star is rotating more rapidly than a critical rotation period, $P_{\text{crit}} \cong 0.7\text{–}1.7$ ms (which depends on the star's structure and its temperature-dependent viscosity), then an instability driven by gravitational radiation reaction ('Chandrasekhar (1970)–Friedman–Schutz' (1978), or 'CFS' instability) will create and maintain significantly strong hydrodynamic waves in the star's surface layers and mantle, propagating in the opposite direction to the star's rotation; and these will radiate strongly. For detailed discussions see Wagoner (1984), Lindblom (1986, 1987), Friedman, Ipser and Parker (1986), Schutz (1987), Cutler and Lindblom (1987).

At present we are extremely ignorant of the degree of asymmetry in rotating neutron stars, and accordingly we are ignorant of the strengths of the periodic waves to be expected from them. Pessimists will note that there is no observational evidence in any observed pulsar for sufficient non-axisymmetry to produce interestingly strong waves. Pessimists will point, especially, to the extremely small slowdown rate of the 1.6 ms pulsar PSR1937+21, which implies such weak radiation reaction that the characteristic amplitude at earth cannot exceed 1×10^{-27} (Fig. 9.6), and the star's non-axisymmetric ellipticity cannot exceed 3×10^{-9}.

Optimists will also point to PSR1937+21 and some other millisecond pulsars, and note a reasonably likely scenario for their origin (van den Heuvel, 1984; Wagoner, 1984): that they were spun up long ago by accretion from a binary companion until they hit the CFS instability, that they remained just beyond the instability point for awhile, with the spinup torque of accretion being counterbalanced by gravitational radiation reaction, and that the accretion stopped long ago leaving the stars plenty of time to anneal and settle down into their presently observed, highly axisymmetric states. Given the extreme observational difficulty of finding by electromagnetic means evidence for rapid neutron-star rotation (see, e.g., Section IV of Reynolds and Stinebring (1984) for searches in the radio), it may well be that there are a number of accreting neutron stars in our galaxy now in the CFS regime, radiating strong gravitational waves. For such a star the energy being radiated in gravitational waves and that being radiated as accretion-induced X-rays will both be proportional to the accretion rate; and consequently the characteristic amplitude of the gravitational waves at earth will be proportional to the square root of the X-ray flux arriving at earth, F_X:

$$h_c \simeq 2\times 10^{-27}\left(\frac{300\ \mathrm{Hz}}{f}\right)^{\frac{1}{2}}\left(\frac{F_X}{10^{-8}\ \mathrm{erg\ cm^{-2}\ s^{-1}}}\right)^{\frac{1}{2}} \tag{53}$$

(Wagoner, 1984). The frequency f of the waves will be $f = lv_p/(2\pi R)$ where R is the star's radius, $l = 3$ or 4 or 5 is the spherical-harmonic order of the hydrodynamic wave, and v_p is the pattern speed of the wave as seen in the inertial frame of distant observers. The X-ray flux $F_X \cong 10^{-8}$ erg cm^{-2} s^{-1} is $\frac{1}{20}$ that of Sco X-1, the brightest quasi-steady source in the sky and itself a candidate for a CFS-unstable object. As Fig. 9.6 shows, stars with X-ray fluxes as low as $\frac{1}{1000}$ Sco X-1 could be interestingly strong sources of gravitational waves. In such a star the density waves in the surface layers should modulate the emitted X-rays, but the sensitivities of past X-ray telescopes have been too poor to detect such rapid and weak modulations. There is an interesting proposal (Wood *et al.*, 1986) for a new, more sensitive X-ray telescope designed to search for such modulations in Sco X-1 and other, weaker X-ray sources. Such a telescope, operated in coordination with gravitational-wave detectors, might one day give a wealth of new information about neutron stars.

Even the most rapidly rotating of neutron stars will be smaller than a reduced wavelength of its emitted gravitational waves and thus can be described with reasonable accuracy by a slow-motion formalism (Ipser, 1971). If the emitter is CFS density waves, the multipoles involved are $l = 3$, 4, or 5. If the emitter is solidly supported or magnetic-field supported deformations, the waves should be largely quadrupolar, $l = 2$. Let us focus now on the latter case, i.e. on a star with rotation period large enough that the CFS instability does not act. Although gravity inside the star is significantly non-Newtonian, one can still define for the star a moment-of-inertia tensor I_{jk} equal to the ratio of its angular momentum to its angular velocity (both being vectors defined in the weak-gravity region well outside the star; see Thorne and Gürsel, 1983). If the star rotates about a principal axis of this moment-of-inertia tensor, i.e. if its angular momentum and angular velocity are parallel, then it will not precess, and (assuming a rotation period sufficiently long that it is CFS stable), its gravitational waves will be emitted at twice the rotation frequency. From the quadrupole variant of the slow-motion formalism we can then compute that the waves will have the standard periodic form of equation (48) with amplitudes

$$h_{o+}=2(1+\cos^2\iota)\,\frac{(\mathcal{I}_{\hat{x}\hat{x}}-\mathcal{I}_{\hat{y}\hat{y}})(\pi f)^2}{r},$$

$$h_{o\times}=4\cos\iota\,\frac{(\mathcal{I}_{\hat{x}\hat{x}}-\mathcal{I}_{\hat{y}\hat{y}})(\pi f)^2}{r}. \tag{54}$$

Here ι is the angle between the neutron star's rotation axis and the line of sight from the earth, and $\mathcal{I}_{\hat{x}\hat{x}}$ and $\mathcal{I}_{\hat{y}\hat{y}}$ are the components of the star's quadrupole moment along the principal axes in its equatorial plane. The characteristic amplitude of these waves (equation (50)) is

$$h_c=8\pi^2\left(\frac{2}{15}\right)^{\frac{1}{2}}\frac{\varepsilon I_{\hat{z}\hat{z}}f^2}{r}=7.7\times10^{-20}\varepsilon\left(\frac{I_{\hat{z}\hat{z}}}{10^{45}\text{ g cm}^2}\right)\left(\frac{f}{1\text{ kHz}}\right)^2\left(\frac{10\text{ kpc}}{r}\right), \tag{55}$$

where $I_{\hat{z}\hat{z}}$ is the moment of inertia of the star about its rotation axis (so $-\frac{1}{2}I_{\hat{z}\hat{z}}(\pi f)^2$ is its rotational energy), and

$$\varepsilon\equiv\frac{\mathcal{I}_{\hat{x}\hat{x}}-\mathcal{I}_{\hat{y}\hat{y}}}{I_{\hat{z}\hat{z}}} \tag{56}$$

is its 'gravitational ellipticity' in the equatorial plane. All neutron stars for which masses have been measured have M near $1.4M_\odot$; and depending on the equation of state these masses correspond to $3\times10^{44}\text{ g cm}^2\lesssim I_{\hat{z}\hat{z}}\lesssim 3\times 10^{45}\text{ g cm}^2$. The likely values of the ellipticity ε are far less clear.

The observed slow-down rates of the Crab ($f=60$ Hz, $r=2$ kpc), Vela ($f=22$ Hz, $r=500$ pc), and PSR1937+21 ($f=1.25$ kHz, $r=5$ kpc) pulsars, if due to gravitational radiation reaction (possible but not likely), correspond to $\varepsilon\simeq6\times10^{-4}$, 4×10^{-3}, and 3×10^{-9} respectively; and to $h_c\simeq 8\times10^{-25}$, 3×10^{-24}, and 1×10^{-27}. Zimmermann (1978) argues that reasonable values for the Crab and Vela are $\varepsilon\sim3\times10^{-6}$ and 3×10^{-5} corresponding to $h_c\sim4\times10^{-27}$ and 2×10^{-26}. Alpar and Pines (1985) suggest reasonable values for PSR1937+21 (which is old and well annealed) in the range $\varepsilon\sim4\times10^{-10}$–$1\times10^{-11}$ corresponding to $h_c\sim1\times10^{-28}$.

Blandford (1984) points out that, if there is a population of young pulsars (not yet discovered) that are spinning down by gravitational radiation reaction on a spin-down timescale τ_{GW}, then (i) the nearest will be at a distance $r\simeq R_G(\tau_B/\tau_{GW})^{\frac{1}{2}}$ (assumed $\leqslant\tau_{GW}$) is the mean time between births of these pulsars in our galaxy and R_G is the radius of the galaxy's disk;

(ii) the flux of gravitational-wave energy at earth from the nearest such pulsar $(3/64\pi)(2\pi f)^2 h_c^2$, will be equal to $[I_{\hat{z}\hat{z}}(\pi f)^2 \tau_{GW}](4\pi r^2)^{-1}$; and (iii) as a consequence, the characteristic amplitude of the waves from the nearest one will be

$$h_c \simeq \left[\frac{4}{3}\frac{I_{\hat{z}\hat{z}}}{r^2 \tau_{GW}}\right]^{\frac{1}{2}} \simeq \left[\frac{4}{3}\frac{I_{\hat{z}\hat{z}}}{R_G^2 \tau_B}\right]^{\frac{1}{2}} \sim 1.1 \times 10^{-25}\left(\frac{10^4 \text{ years}}{\tau_B}\right)^{\frac{1}{2}}, \qquad (57)$$

independently of its frequency and ellipticity.

Fig. 9.6 shows values of h_c and f corresponding to some of the above possibilities.

If the star does not rotate about a principal axis of its moment of inertia, then it will precess. When one idealizes the star's interior as rigid, then although the interior gravity is significantly non-Newtonian, the precession is still described by the classic equations of Euler (see Thorne and Gürsel, 1983, for a proof); and the resulting waves will have a form that typically will entail significant spectral components at three frequencies: twice the rotation frequency, and the rotation frequency plus and minus the precession frequency (Zimmermann, 1980; Zimmermann and Szedenits, 1979). In reality, pliability of the neutron-star material will cause significant deviations from rigid rotation; but these three frequencies may still be dominant. If the star is idealized as a fluid body deformed by the pressure of an off-axis internal magnetic field, then the star does not precess and the radiation is emitted at the rotation frequency and twice the rotation frequency (Gal'tsov, Tsvetkov and Tsirulev, 1984).

If and when gravitational waves from rotating neutron stars are detected, they may carry a wealth of information about the star's structure and dynamics in the amplitudes and relative phasings of their various spectral components. Especially interesting may be the evolution of the various spectral components after a star quake; together with electromagnetic timing of the post-quake rotation, these may give us new insights into the coupling of the solid crust or core to the fluid mantle.

(c) *Binary stars*

Ordinary binary star systems are the most reliably understood of all sources of gravitational waves. From the measured mass and orbital parameters of a binary and its estimated distance, one can compute with confidence the details of its waves. Unfortunately, ordinary binaries have orbital periods no shorter than about an hour and, correspondingly, gravitational-wave frequencies $f \lesssim 10^{-3}$ Hz. The shortest known binary of all is a white-dwarf/

neutron-star system with orbital period 11 minutes and gravitational-wave frequency $f \simeq 3 \times 10^{-3}$ Hz (Priedhorsky, Stella and White, 1986). Because of seismic noise, detectors in earth laboratories cannot hope to see waves of such low frequency. However, beam detectors in space, tentatively planned for the turn of the century, should see them with relative ease (cf. Section 9.5.5(b) below).

Detailed formulas for the waves from a binary star, including the effects of the eccentricity and inclination of its orbit, have been derived from the quadrupole formalism by Peters and Mathews (1963) and Wahlquist (1987). See also Wagoner and Will (1976), and Gal'tsov, Matiukhin and Petukhov (1980) for post-Newtonian corrections. By virtue of the eccentricity of the orbit, waves will be emitted in equally spaced 'spectral lines' at twice the orbital frequency and harmonics thereof. For eccentricity $\varepsilon \lesssim 0.2$ the line at $f = 2f_{\text{orb}}$ is dominant; for $\varepsilon \simeq 0.5$ the lines at $f/f_{\text{orb}} \simeq 2$ through 8 are all strong; for $\varepsilon \simeq 0.7$ the lines at $f/f_{\text{orb}} \simeq 4$ through 20 are all strong. In the low-eccentricity case $\varepsilon \lesssim 0.2$ the waves have the form (42) with $f = 2f_{\text{orb}}$, which corresponds to a characteristic amplitude (equation (50))

$$\begin{aligned} h_c &= 8\left(\frac{2}{15}\right)^{\frac{1}{2}} \frac{\mu}{r} (\pi M f)^{\frac{2}{3}} \\ &= 8.7 \times 10^{-21} \left(\frac{\mu}{M_\odot}\right)\left(\frac{M}{M_\odot}\right)^{\frac{2}{3}}\left(\frac{100\ \text{pc}}{r}\right)\left(\frac{f}{10^{-3}\ \text{Hz}}\right)^{\frac{2}{3}}, \qquad (58) \end{aligned}$$

where μ is the reduced mass and M the total mass of the system. This amplitude is plotted in Fig. 9.6 for a few of the most strongly radiating known binaries. For lists of the most strongly radiating binaries and their characteristics see Braginsky (1965) and Douglass and Braginsky (1979).

White-dwarf and neutron-star binaries should also be important emitters – and they should extend to higher frequencies than ordinary binaries; but there is a paucity of observational data on them and the example of shortest known period has f only 3×10^{-3} Hz (see above). From the data that do exist (e.g. Iben and Tutukov, 1984), Lipunov and Postnov (1986), Lipunov, Postnov and Prokhorov (1987), and Hils *et al.* (1987) have estimated the characteristic amplitudes of the strongest white-dwarf and neutron-star binaries; see Fig. 9.6. For a very detailed treatment of the white-dwarf case see Evans, Iben and Smarr (1987). The highest frequency to be expected for any white-dwarf binary in our galaxy is 0.06 Hz since mass transfer from the less massive star to the more massive begins at or before this frequency; the highest for any neutron-star binary is 0.007 Hz since any binary of higher frequency than this would coalesce in a time less than the mean interval between coalescences, $\sim 10^4$ years.

Gravitational radiation reaction plays an important role in driving the evolution of close binary systems; see Paczynski and Sienkiewicz (1981) for details.

9.4.3 Stochastic sources

(a) *Characterization of stochastic gravitational waves*

It is useful to think about stochastic gravitational waves in terms of traveling-wave normal modes of the gravitational field. As with the electromagnetic field, there are two modes (because of two polarization states) for each volume $(2\pi\hbar)^3$ in phase space; and correspondingly, one can easily show, the energy density per unit logarithmic interval of frequency divided by the critical energy density $\rho_{\rm crit} \sim 10^{-8}$ erg cm^{-3} to close the universe is

$$\Omega_{\rm GW}(f) \equiv \frac{\mathrm{d}E^{\rm GW}/\mathrm{d}^3x\,\mathrm{d}\ln f}{\rho_{\rm crit}} \sim \frac{\bar{n}}{10^{37}}\left(\frac{f}{1\ \mathrm{kHz}}\right)^3. \tag{59}$$

Here $\bar{n}$ is the average number of quanta in all modes with frequencies of order f. Below we shall see that $\Omega_{\rm GW}(f)$ is likely to be $\gtrsim 10^{-14}$ at all frequencies of interest, and correspondingly the mean number of quanta in each mode is likely to be $\gtrsim 10^{20}$ – so large that a classical treatment is in order.

In TT coordinates the metric perturbation due to stochastic gravitational waves, evaluated at any chosen location x^i, will be a sum over contributions of all the modes of the field

$$h^{\rm TT}_{jk}(t, x^i) = \sum_K h^{\rm TT}_{Kjk}(t, x^i). \tag{60}$$

Here the index K labels modes of the field. For stochastic gravitational waves, the wave field $h^{\rm TT}_{Kjk}$ associated with mode K can be regarded as a 'random process' (i.e. a stochastically fluctuating function of time t); and the total field $h^{\rm TT}_{jk}$ at location x^i is the sum over all of the modes' random processes.

The field of a chosen mode K can be expressed as

$$h^{\rm TT}_{Kjk} = h_K(t, x^i)e^K_{jk}, \tag{61}$$

where $h_K(t)$ is its scalar wave function and e^K_{jk} is its constant polarization vector, so normalized that $e^K_{jk}e^K_{jk} = 2$ in Cartesian coordinates; cf. equation (7c). Then h_K is a scalar random process in time (at fixed x^i) and its statistical properties can be characterized by a spectral density $S_{h_K}(f)$.

I shall assume that the modes are defined in such a way that there is no significant correlation between their wave fields $h^{\rm TT}_{Kjk}$. As a result, when one averages the stress-energy tensor (9) of the waves at x^i over a sufficiently long time, all cross-terms between different modes get washed out and one obtains for the time-averaged specific intensity I_f at location x^i

$$I_f(t, x^i)\,\Delta\Omega \equiv \frac{\mathrm{d}E}{\mathrm{d}A\,\mathrm{d}t\,\mathrm{d}f\,\mathrm{d}\Omega}\,\Delta\Omega = \sum_{K \text{ in } \Delta\Omega} \frac{\pi f^2}{4} S_{h_K}. \tag{62}$$

Here E denotes energy, A denotes area, Ω denotes solid angle, and the sum is over all modes K with propagation directions in the infinitesimal solid angle $\Delta\Omega$. The total energy per unit logarithmic interval (used in defining $\Omega_{\mathrm{GW}}(f)$ above) can be expressed in terms of the specific intensity in the standard way

$$\Omega_{\mathrm{GW}}(f)\rho_{\mathrm{crit}} = \frac{\mathrm{d}E}{\mathrm{d}^3x\,\mathrm{d}\ln f} = \int f I_f\,\mathrm{d}\Omega = \sum_K \frac{\pi f^3}{4} S_{h_K}, \tag{63}$$

where the integral is over the entire sphere and the sum is over all modes.

As for burst and periodic waves, so also for stochastic, we shall introduce a single characteristic amplitude h_c that is tied to a specific experimental situation: the experimenters use two identical gravity-wave receivers (broad-band or narrow), separated by a distance $\ll \lambda\!\!\!^{-} = c/2\pi f$, to search for *isotropic*, stochastic waves in the neighborhood of frequency f. The search is performed by a standard technique (Bendat, 1958; Section 9 of Drever, 1983): the outputs, $h_1(t)$ and $h_2(t)$ of detectors 1 and 2 are passed through identical filters which admit only Fourier components in a bandwidth $\Delta f \lesssim f$ centered on frequencies $\pm f$. The filtered outputs $w_1(t)$ and $w_2(t)$ are then multiplied together and integrated for a time $\hat{\tau}$ to get a single number $W = \int_{t_0}^{t_0+\hat{\tau}} w_1(t) w_2(t)$. This number will consist of a signal due to the identical stochastic backgrounds contained in $w_1(t)$ and $w_2(t)$, and a Gaussian noise due to the independent noises in the two detectors (each with the same spectral density $S_n(f)$). It turns out (e.g. Chapter 7 of Bendat 1958) that the ratio of the signal to the root-mean-square noise is

$$\frac{S}{N} = \frac{h_c(f)}{h_n(f)}, \tag{64}$$

where

$$\begin{aligned} h_c(f) &\equiv \left[\sum_K f S_{h_K}(f)\right]^{\frac{1}{2}} = \left[\frac{4}{\pi f}\int I_f\,\mathrm{d}\Omega\right]^{\frac{1}{2}} = \left[\frac{4}{\pi f^2}\,\Omega_{\mathrm{GW}}(f)\rho_{\mathrm{crit}}\right]^{\frac{1}{2}} \\ &= 1.3\times 10^{-18}\left(\frac{\rho_{\mathrm{crit}}}{1.7\times 10^{-8}\ \mathrm{erg\ cm^{-3}}}\right)^{\frac{1}{2}}\left(\frac{1\ \mathrm{Hz}}{f}\right)[\Omega_{\mathrm{GW}}(f)]^{\frac{1}{2}} \end{aligned} \tag{65}$$

is the characteristic amplitude of the isotropic, stochastic waves, and

$$h_n(f) = \frac{1}{(\frac{1}{2}\hat{\tau}\Delta f)^{\frac{1}{2}}}\,\frac{[f S_n(f)]^{\frac{1}{2}}}{\langle F_+^2\rangle^{\frac{1}{2}}} \tag{66}$$

is the characteristic noise amplitude of the detectors. (Note: $\rho_{\mathrm{crit}}/1.7\times$

10^{-8} erg cm$^{-3}=(H_o/100$ km s^{-1} Mpc$^{-1})^2$ where H_o is the Hubble constant.) Correspondingly, if the experimenters wish to be 90% confident of having seen the stochastic background during a search of duration $\hat{\tau}=\frac{1}{3}$ year, the Gaussian probability distribution for S/N requires $S/N=1.7$, which in turn means that h_c must exceed the noise level

$$h_{3/\mathrm{yr}}(f)=1.7h_n(f)=2.0\left(\frac{\Delta f}{10^{-7}\,\mathrm{Hz}}\right)^{-\frac{1}{4}}\frac{[fS_h(f)]^{\frac{1}{2}}}{\langle F_+^2\rangle^{\frac{1}{2}}}. \tag{67}$$

The characteristic amplitudes h_c of various possible stochastic sources, and the noise levels $h_{3/\mathrm{yr}}$ of various detectors are shown in Fig. 9.7.

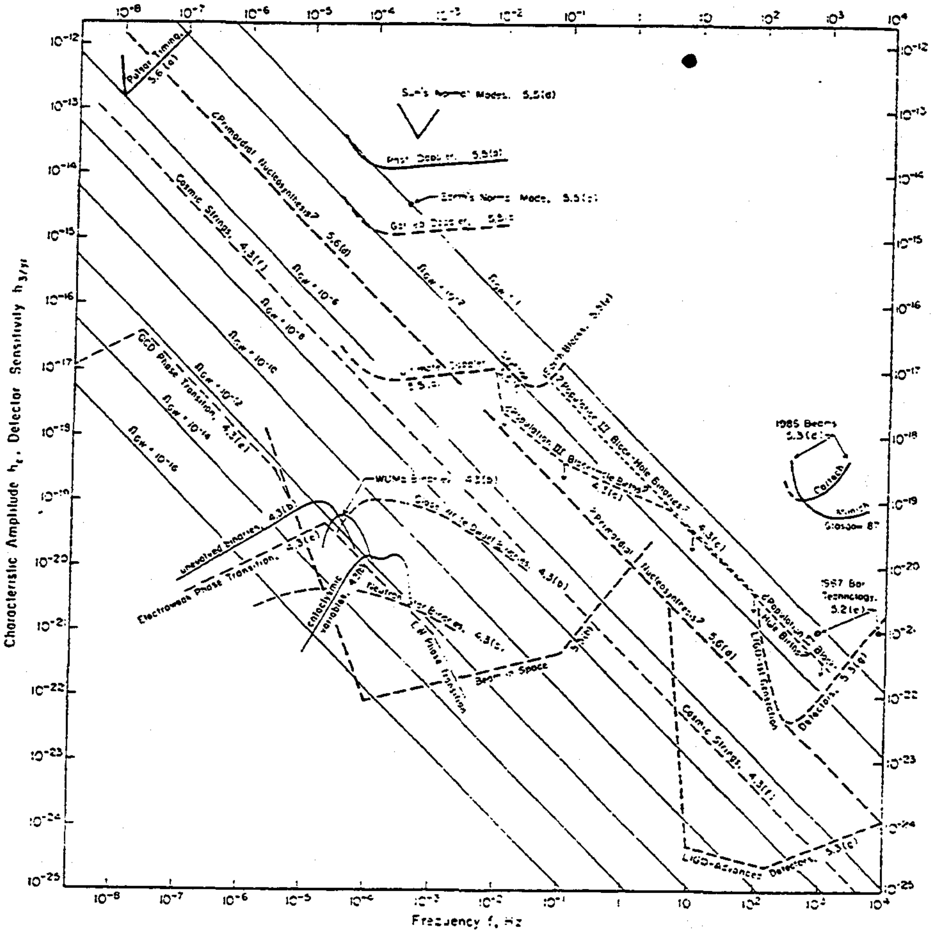
(b) *Binary stars*

So many binary stars in our galaxy and in other galaxies radiate in the frequency region $f\lesssim 0.03$ Hz that they should superpose to produce a strong stochastic background. Lipunov and Postnov (1986), Lipunov, Postnov and Prokhorov (1987) and Hils *et al.* (1987) have made careful calculations of the characteristic amplitude of this stochastic background as a function of frequency; the results of Hils *et al.* are shown in Fig. 9.7 for the contribution of our own galaxy (which should be concentrated in the galactic plane). The contributions of all other galaxies (which should be isotropic) should be down from those of our own galaxy by $(h_c)_{\mathrm{other}}/(h_c)_{\mathrm{us}}\sim 0.15$.

The binary stochastic background in Fig. 9.7 is broken up into contributions from various types of binaries. Those shown as solid curves (unevolved binaries, WUMa stars (first discussed by Mironovskii, 1966), and cataclysmic variables (white-dwarf/normal-star systems)) are rather firmly based on optical studies of the statistics of these types of stars, and thus are rather reliable. Those shown dashed (close white-dwarf binaries (see Evans, Iben and Smarr, 1987, and the above references), and neutron-star binaries) are based on so little observational data and so much theory that they are highly uncertain.

The binary background presents a serious potential obstacle to searches for other kinds of waves in the frequency band 0.03 Hz$\lesssim f\lesssim 10^{-5}$ Hz where space-based beam detectors will operate. A broad-band burst can be seen above this background only if it has $h_{c\,\mathrm{burst}}>h_{c\,\mathrm{background}}$ – which means, e.g., that a $1M_\odot$ star falling into a supermassive black hole in the Virgo cluster will not be discernible unless the hole's mass is $M<3\times 10^5\,M_\odot$ (cf. Figs. 9.4 and 9.7). A periodic source can be seen, after an integration time $\hat{\tau}$, only if it has $h_{c\,\mathrm{periodic}}>(f\hat{\tau})^{-\frac{1}{2}}h_{c\,\mathrm{background}}$ – which means, e.g., that if the close white-dwarf binaries are as numerous as estimated, the binaries ι Boo and SS Cyg will be discernible only after integration times of $\hat{\tau}>10^7$ s. For further discussion see Evans, Iben and Smarr (1987) and Hils *et al.* (1987).

Fig. 9.7. The characteristic amplitudes h_c (equation (65)) and frequencies f of waves from several postulated *stochastic sources* (thin curves), and the sensitivities $h_{3\,\mathrm{yr}}$ of several existing and planned detectors (thick curves and circles) ($h_{3/\mathrm{yr}}$ is the amplitude of the weakest source that can be detected with 90% confidence in a $\frac{1}{3}$ yr = 10^7 s integration). The very thin diagonal lines indicate values of h_c corresponding to constant $\Omega_{GW}(f)$ = (gravity wave energy in bandwidth $\Delta f = f$) (energy to close the universe if $H_0 = 100$ km s^{-1} Mpc^{-1}). The sources shown as solid curves (background from various types of binary stars in our galaxy: Section 9.4.3(c)) are rather firm predictions. The sources shown as dashed and dotted curves are much less firm; see Sections 9.4.3(c)–(f) for discussion of specific sources. Not shown are primordial waves from the big-bang (Section 4.3(d)), which could have Ω_{GW} as large as 1 or as small as 10^{-14} or less according to various plausible scenarios – but which are limited by observations as discussed in Section 9.5.6. The detectors are discussed in detail in the indicated subsections of Section 9.5.



(c) *Population III stars*

If there was a pre-galactic population of massive stars ('Population III stars'; Carr, 1986), the violent events that terminated their lives (supernovae and collapse to black holes) might have produced gravitational waves that we today would see as isotropic and stochastic. Whereas existing binary stars are a firm source of stochastic background, these Population III stars are a highly speculative one. Carr (1980) has derived an upper limit on the characteristic amplitude that could have been produced by the deaths of such stars under any reasonable scenario; it is shown in Fig. 9.7 (upper short-dashed curve) along with the maximum characteristic amplitude that could come from the remnants of these stars if they became black-hole binaries that decay by radiation reaction on a timescale $\tau_{GW} \simeq 10^{10}$ years (Bond and Carr, 1984).

(d) *Primordial gravitational waves*

Photons coming from the big bang last scattered off matter at a cosmological redshift $z \sim 1000$ when the universe was roughly one million years old; and neutrinos last scattered at $z \sim 10^{10}$ when it was about 0.1 s old. An order-of-magnitude calculation shows that gravitons, by contrast, last scattered at roughly the Planck time, i.e. during the first 10^{-43} seconds when spacetime was quantized and the laws of physics were exceedingly different from today (Section 7.2 of Zel'dovich and Novikov, 1983). (An unlikely exception occurs if, at the epoch when the waves' reduced wavelength was $\lambda\!\!\!^{-} \sim$ (horizon size), much of the universe's energy density was in relativistic particles with mean free paths of order $\lambda\!\!\!^{-}$; then non-negligible absorption may occur; see Vishniac (1982).) Thus, in studying primordial gravitational waves (waves created in the big-bang), one usually can ignore their subsequent interactions with matter.

Not so for their subsequent interactions with the background spacetime curvature of the universe. Grishchuk (1974, 1975*a*,*b*, 1977) has shown that, as the primordial perturbations that give rise to present-day waves 'come inside the cosmological horizon' – and also before they enter the horizon – they can be parametrically amplified by their interaction with the dynamical background spacetime curvature; in other words, they can trigger further graviton creation. In this way exceedingly small initial fluctuations can be amplified into an interestingly strong stochastic background today.

Just how much stochastic background is produced depends crucially on ill-understood aspects of the initial singularity and on the equation-of-state-dependent and vacuum-dependent expansion rate in the very early universe. Some otherwise plausible models produce so much, $\Omega_{GW}(f) \gg 1$, as to be in violent conflict with the observed current state of the universe (e.g. p. 621 of Zel'dovich and Novikov, 1983). Other, equally plausible models can produce so little, $\Omega_{GW}(f) \ll 10^{-14}$, that there is no hope of detecting the waves in the foreseeable future.

In currently fashionable inflationary models of the universe vacuum fluctuations which initially are smaller than the horizon ($\bar{\lambda} \ll \mathscr{R}_B =$ (background radius of curvature)) are driven outside the horizon ($\bar{\lambda} \gg \mathscr{R}_B$) by the inflationary expansion. While outside the horizon, they are 'frozen' with constant amplitude h. Much later, after inflation ends, non-inflationary expansion brings them back inside the horizon. The number of quanta in each mode before entering and after leaving the horizon is

$$n \sim \left[\frac{1}{16\pi}\left(\frac{h}{\bar{\lambda}}\right)^2\right](2\pi\bar{\lambda})^3\left(\frac{1}{\hbar/\bar{\lambda}}\right) \sim \frac{\pi}{4}\frac{(h\bar{\lambda})^2}{\hbar},$$

where the first factor is the waves' energy density, the second is the volume occupied by each mode, and the third is 1/(energy of one graviton). Before entering the horizon $n \simeq \frac{1}{2}$; so the above relation says that upon leaving it

$$n_{\text{out}} \simeq n_{\text{enter}} \frac{\bar{\lambda}_{\text{leave}}}{\bar{\lambda}_{\text{enter}}} = \frac{1}{2}\frac{a_{\text{leave}}}{a_{\text{enter}}},$$

where a is the expansion factor of the universe. Thus, the epoch of amplitude freezing is actually an epoch of parametric amplification (stimulated creation of new gravitons); and the total number of gravitons created depends on the total amount of expansion that occurs while the waves are outside the horizon. (There will be additional parametric amplification as the waves emerge from the horizon, $\bar{\lambda} \sim \mathscr{R}_B$, but in inflationary models that is generally small compared to the amplification during freezing, $\bar{\lambda} \gg \mathscr{R}_B$.) The total amount of inflationary expansion differs from one inflationary model to another; and correspondingly, the models can give Ω_{GW} as large as unity or Ω_{GW} too small ($\ll 10^{-14}$) for there to be hope of detecting the waves.

For discussions of the influence of the equation of state in the early universe on the spectrum of the amplified waves, see Grishchuk (1977) and Fig. 4 of Grishchuk and Polnarev (1980). For calculations of the waves produced by specific inflationary scenarios see Starobinsky (1979), Rubakov, Sazhin and Veryaskin (1982), Abbott and Wise (1984), Halliwell and Hawking (1985), Mijic, Morris and Suen (1986) and references therein. Because the range of possible strengths of primordial waves is so great, we do not bother to show it in Fig. 9.7 – aside from indicating the values of h_c corresponding to various values of $\Omega_{GW}(f)$.

(e) *Phase transitions*

During the early expansion of the universe, there may have been first-order phase transitions associated with QCD interactions and with Electroweak interactions. In each of these phase transitions the original phase would be supercooled, by the cosmological expansion, below the equilibrium temperature of the new phase. Bubbles of the new phase would then nucleate at isolated locations and expand at near-light velocity until they have compressed the original phase enough for the two phases to coexist in equilibrium. As Witten (1984) has pointed out, and Hogan (1986) has analyzed in detail, this 'cavitation' should have produced gravitational waves in two ways: (i) directly from expanding bubbles and the subsequent sound waves they generate, and (ii) subsequently from the inhomogeneities associated with the two co-existing phases (large-scale density inhomogeneities and corresponding inhomogeneities in the Hubble expansion rate). The resulting gravitational waves should possess a spectrum that peaks at wavelengths which were of order the horizon size when the cavitation occurred. Those wavelengths correspond to frequencies today $f_{\max} \sim (2 \times 10^{-7}\ \text{Hz})(kT/1\ \text{GeV})$, where T is the temperature of the phase transition. Hogan's (1986) predicted spectra, shown in Fig. 9.7, thus peak at $f_{\max} \sim 2 \times 10^{-8}$ Hz (QCD, $T \sim 100$ MeV) and $f_{\max} \sim 2 \times 10^{-5}$ Hz (Electroweak, $T \sim 100$ Gev). The $f^{-\frac{1}{2}}$ shape of the spectra at frequencies $f > f_{\max}$ is a firm prediction, but the shape f^{-1} at $f < f_{\max}$ is not (it could well be f^{-p} with $p > 1$). The amplitude shown is a reasonable upper limit, unless the phase transition is unusually catastrophic with very strong supercooling.

(f) *Cosmic strings*

Long before the QCD and Electroweak phase transitions – i.e. nearer the initial singularity – there may have been a phase transition associated with the grand-unified interactions, and that transition may have created cosmic strings – one-dimensional 'defects' in the vacuum with mass per unit length estimated to be $\mu \sim 10^{-6}$ and with tension equal to mass per unit length (Zel'dovich, 1980; Vilenkin, 1981*a*). As the universe's horizon expands to uncover the stochastic inhomogeneities in a string's shape, those inhomogeneities should begin to vibrate with speeds up to the speed of light. By self-intersection of the string, closed loops should form; and those loops could have acted as seeds for the condensation of galaxies and galaxy clusters (Zel'dovich, 1980; Vilenkin, 1981*a*; Turok and Brandenberger, 1986; Sato, 1987).

An unavoidable byproduct of this model for galaxy formation is huge amounts of stochastic gravitational waves produced by the vibrations of the closed loops (Vilenkin, 1981*b*). Detailed calculations by Vachaspati and Vilenkin (1985) (confirming earlier, less accurate calculations by many others) predict that, if the strings are not superconducting (for the superconducting case see Ostriker *et al.*, 1986),

$$\Omega_{\rm GW}(f) \sim 10^{-7}\left(\frac{\mu}{10^{-6}}\right)^{\frac{1}{2}} \quad \text{for all } f \gtrsim 10^{-8}\ \text{Hz}\left(\frac{10^{-6}}{\mu}\right). \qquad (68)$$

(For the spectrum at lower frequencies see Hogan and Rees, 1984.) If μ is far less than 10^{-6} (i.e. if gravity-wave observations constrain $\Omega_{\rm GW}(f)$ to be $\ll 10^{-7}$), then the non-superconducting cosmic-string theory of galaxy formation will face severe difficulties. Fig. 9.7 shows the predicted waves (68). From that diagram and the corresponding discussions in Section 9.5 it is clear that several different observational techniques have the prospect of placing cosmic string theory in jeopardy – or, hopefully, of discovering string-produced waves. (The apparent disproof of $\Omega_{\rm GW} \sim 10^{-7}$ coming from 5°-scale anisotropy of the cosmic microwave radiation (Fig. 9.7 and Section 9.5.6(c)) does not in fact constrain cosmic strings, since this observational limit is sensitive only to waves that were present and had (reduced wavelength) ~ (horizon size) at the epoch of recombination – before the strings that produce this wavelength began to vibrate and radiate.)

9.5 Detection of gravitational waves

9.5.1 *Methods of analyzing gravitational wave detectors*

When analyzing the performance of a gravitational wave detector, it is important to pay attention to the size L of the detector compared with a reduced wavelength $\lambda\!\!\!^{-}$ of the waves it seeks.

If $L \ll \lambda\!\!\!^{-}$ then the detector can be contained entirely in the proper reference frame of its center, and the analysis can be performed using non-relativistic concepts augmented by the quadrupolar gravity-wave force field (3), (5). If one prefers, of course, one instead can analyze the detector in TT coordinates using general relativistic concepts and the spacetime metric (8). The two analyses are guaranteed to give the same predictions for the detector's performance, unless errors are made. However, errors are much more likely in the TT analysis than in the proper-reference-frame analysis, because our physical intuition about how experimental apparatus behaves is proper-reference-frame based rather than TT-coordinate based. As an example, we intuitively assume that if a microwave cavity is rigid, its walls will reside at fixed coordinate locations x^i. This remains true in the detector's proper reference frame (aside from fractional changes of order $(L^2/\lambda\!\!\!^{-2})h$, which are truly negligible if the detector is small and which the proper-reference-frame analysis ignores). But it is not true in TT coordinates; there the coordinate locations of a rigid wall are disturbed by fractional amounts of order h, which are crucial to analyses of microwave-cavity-based gravity wave detectors.

Thus, for small detectors, $L \ll \lambda\!\!\!^{-}$, the proper-reference-frame analysis is much to be preferred.

For large detectors, $L \gtrsim \lambda\!\!\!^{-}$, one cannot introduce a proper reference frame that covers the entire detector. Such detectors can only be analyzed using general relativistic concepts in TT coordinates (usually the best) or in some other suitable coordinate system (rarely as good).

9.5.2 *Resonant bar detectors*

More effort has been put into resonant bars than into any other type of gravity-wave detector. Weber's original detectors were of the resonant-bar type; all but one of the first-generation (pre-1977) earth-based detectors were of this type; and eight of the world's twelve research groups now building and operating earth-based detectors are working with bars. Of the current bar efforts three are in the United States (the University of Maryland (Weber, 1986), Stanford University (Boughn *et al.*, 1982; Michelson, 1983) and Louisiana State University (Hamilton *et al.*, 1986)); two are in Europe (the University of Rome with its detector sited at CERN (Amaldi *et al.*, 1984) and Moscow University (Braginsky, 1983)); and three are in the Far East (The University of Western Australia in Perth (Blair, 1983), Tokyo University (Owa *et al.*, 1986) and Guangzhou, China (Hu *et al.*, 1986)). The improvements in resonant-bar sensitivities since Weber's first detector have been a factor of roughly 200 in amplitude, corresponding to 40 000 in energy; and significant further improvements are yet to come.

(a) *How a resonant-bar detector works*

Schematically (Fig. 9.8), a resonant-bar detector consists of a large, heavy, solid bar whose mechanical oscillations are driven by gravitational waves, a transducer that converts information about the bar's oscillations into an electrical signal, an amplifier for the electrical signal, and a recording system. The transducer and amplifier together are sometimes called the sensor.

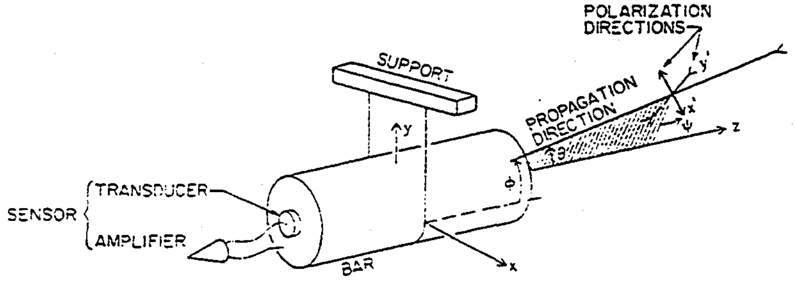
The transducer typically is mounted on one end of the bar (though other mountings are sometimes used), and it produces an output voltage or current proportional to the displacement $x(t)$ of the bar's end from equilibrium. Although $x(t)$ is a sum of contributions from all the $\sim 10^{29}$ normal modes of the bar, the transducer's output is filtered by the amplifier so that only the contribution of the bar's fundamental normal mode is passed on through. This is accomplished by a band-pass filter centered on the frequency f_0 of the fundamental mode, with bandwidth Δf somewhat smaller than the difference $f_1 - f_0$ between the bar's fundamental and its first harmonic. Thus, in effect, it is the fundamental mode of the bar that acts as the gravity-wave detector; and all the other normal modes are almost irrelevant.

Since the fundamental mode involves the relative in and out motion of the bar's left and right ends with just one node (at the bar's center), it corresponds to a standing sound wave with wavelength twice the length of the bar. Correspondingly, the bar's length must be

$$L \simeq \tfrac{1}{2} v_s / f_0, \tag{69}$$

where v_s is the speed of sound in the bar. Typical solid materials have longitudinal sound speeds of order 5 km s^{-1}; astrophysics suggests (Section 9.4) that 1 kHz is a reasonable frequency to search for gravitational waves; and correspondingly the lengths of typical resonant bar detectors are about 2 m and their masses are several tonnes. Notice that equation (69) gives for the ratio of the length of the bar to the reduced wavelength of the

Fig. 9.8. Schematic diagram of a *resonant-bar detector* for gravitational waves. The angles (θ, ϕ, ψ) characterizing the propagation and polarization directions of the waves relative to the detector are a specialization of the angles (θ, ϕ, ψ) shown in Fig. 9.2.



gravitational waves, $\lambda\!\!\!^{-} = c/2\pi f_0$,

$$L/\lambda\!\!\!^{-} \simeq \pi v_s/c \simeq 5\times 10^{-5}. \tag{70}$$

This justifies, with high accuracy, the use of a Newtonian-language, proper-reference-frame viewpoint in analyses of bar detectors. I shall adopt that viewpoint in this chapter.

The contribution of the fundamental mode to the displacement $x(t)$ of the bar's end can be expressed in the standard harmonic-oscillator form

$$x(t) = \mathrm{Real}[X(t)\,\mathrm{e}^{-\mathrm{i}2\pi f_0 t}], \tag{71}$$

where

$$X(t) = X_1 + \mathrm{i}X_2 \tag{72}$$

is the mode's complex amplitude and f_0 is its eigenfrequency. It is actually the quantity $X(t)$ that the sensor monitors. To the extent that one can ignore forces from the fundamental mode's environment (e.g. very weak coupling to the other modes), only gravitational waves will produce time changes of $X(t)$. When waves hit, they drive $X(t)$ to evolve in a complicated, time-dependent way, and that evolution in principle can be deconvolved to reveal some details of the waveform $h_{jk}^{\mathrm{TT}}(t)$.

Unfortunately, noise in the sensor is typically so severe that to control it the experimenter must use a bandwidth Δf that is far smaller than the frequency f_0 (typically $\Delta f/f_0 \sim 0.01$ in present detectors). Correspondingly, the sensor averages $X(t)$ for a time $\hat{\tau} \simeq 1/\Delta f$ that is long compared to the period $P_0 = 1/f_0$ of the gravitational waves being sought, before passing $X(t)$ on to the recording system. This means that for typical gravitational-wave bursts (e.g. from supernovae), which have durations of only a few times P_0, all that can be monitored is the total change ΔX in the complex amplitude from before the wave arrives until after it has passed. Only uncommonly long bursts, those lasting for more than $f_0/\Delta f \sim 100$ cycles, can be monitored in greater detail. In the future, however, there is hope of bringing the sensor noise under better control and thereby opening up the bandwidth to $\Delta f \simeq 0.2 f_0$ to permit detailed monitoring of much shorter bursts (Michelson and Taber, 1984). (For a description of several first-generation broad-band bars see Figure 2 of Drever, 1977 and associated text, and references therein.)

(b) *The sensitivity of bar detectors to short bursts*

A bar detector couples to the field (equation (26))

$$h(t) = F_+(\theta, \phi, \psi)h_+(t; \iota, \beta) + F_\times(\theta, \phi, \psi)h_\times(t; \iota, \beta), \tag{73}$$

where, if the bar is axially symmetric and the direction (θ, ϕ) and polarization angle ψ are defined as in Fig. 9.8, the beam-pattern factors are

$$F_{+} = \sin^2 \theta \cos 2\psi, \qquad F_{\times} = \sin^2 \theta \sin 2\psi. \tag{74}$$

(See Chapter 37 of MTW for the key elements of a derivation.) We shall presume, throughout this subsection, that $h(t)$ is a burst of such short duration $\Delta t \lesssim \hat{\tau} = 1/\Delta f$ that the optimal way to search for it is to measure the mean square change $|\Delta X|^2$ it produces in the fundamental mode's complex amplitude. In this case the general formula (29) for the ratio S^2/N^2 of the burst's squared signal to the mean-square Gaussian noise in the detector can be reduced to the simple form (see, e.g., Giffard, 1976; Pallotino and Pizella, 1981; Michelson and Taber, 1984)

$$\frac{S^2}{N^2} = \frac{\frac{1}{2}M_{\text{eff}}(2\pi f_0)|\Delta X|^2}{kT_n} \tag{75}$$

Here k is Boltzmann's constant, T_n is a 'noise temperature' which characterizes the overall noise in the detector, and M_{eff} is an 'effective mass' associated with the fundamental mode, so defined that $\frac{1}{2}M_{\text{eff}}|X|^2(2\pi f_0)^2$ is the total energy in the mode when it is vibrating with complex amplitude X. (Since X is actually the amplitude of motion of the end of the bar, for a bar that has uniform cross-section and is long compared to its diameter, the effective mass is $M_{\text{eff}} = \frac{1}{2}(1+v^2)M$ where v is the Poisson ratio of the bar's material and M is the bar's mass.)

Because the net wave-induced change ΔX in complex amplitude is independent of the mode's initial complex amplitude, the numerator of equation (75) is the energy that the wave would have deposited in the mode if the mode had been initially unexcited. This deposited energy can conveniently be expressed in terms of the cross-section $\sigma_0(f)$ that the mode would present to the wave if the wave had hit it from an optimal direction (broadside, $\theta = \pi/2$) and with an optimal polarization (+ mode with $\psi = 0$):

$$\tfrac{1}{2}M_{\text{eff}}(2\pi f_0)^2|\Delta X|^2 = \int_0^{\infty} \frac{\pi}{2} f^2 |\tilde{h}(f)|^2 \sigma_0(f)\, \mathrm{d}f. \tag{76}$$

Here $\tilde{h}(f)$ is the Fourier transform of $h(t)$ (equation (73)), and for an optimal direction and polarization $(\pi/2)f^2|\tilde{h}(f)|^2$ would be the energy per unit area per unit frequency ($f \geqslant 0$) carried by the waves. Because $\sigma_0(f)$ is extremely sharply peaked around the resonant frequency f_0 (Section 37.5 of MTW), we can rewrite (76) and thence (75) in the form

$$\frac{S^2}{N^2} = \frac{\pi}{2} f_0^2 |\tilde{h}(f_0)|^2 \frac{\int \sigma_0\, \mathrm{d}f}{kT_n}. \tag{77}$$

This is a special narrow-band-detector version of the general equation (29) for arbitrary detectors. Correspondingly, equation (30) for the strongest burst seen, on average, at the same rate as bursts occur inside the distance to our source, becomes

$$\frac{S^2}{N^2} \cong \frac{3}{2}\frac{\pi}{2} f_0^2 \langle |\tilde{h}_+(f_0)|^2 + |\tilde{h}_\times(f_0)|^2 \rangle \langle F_+^2 \rangle \frac{\int \sigma_0 \, \mathrm{d}f}{kT_\mathrm{n}}, \tag{78}$$

where, for the F_+ and $F_\times$ of (74),

$$\langle F_+^2 \rangle = \langle F_\times^2 \rangle = \frac{4}{15} = 0.267. \tag{79}$$

Equation (78) motivates us to define

$$h_\mathrm{c} \equiv (3)^{\frac{1}{2}} f_0 \langle |\tilde{h}_+(f_0)|^2 + |\tilde{h}_\times(f_0)|^2 \rangle^{\frac{1}{2}} \tag{80}$$

as the *characteristic amplitude of the burst* (analog of (31b) for broad-band detectors) and

$$h_\mathrm{n} \equiv \left[\frac{15}{\pi} \frac{kT_\mathrm{n}}{\int \sigma_0 \, \mathrm{d}f}\right]^{\frac{1}{2}} \cong 2.2 \left[\frac{G}{c^3} \frac{kT_\mathrm{n}}{\int \sigma_0 \, \mathrm{d}f}\right]^{\frac{1}{2}} \tag{81}$$

as the *detector's characteristic noise amplitude* (analog of (32) for a broad-band detector). Correspondingly, the characteristic amplitude $h_\mathrm{c} = h_{3/\mathrm{yr}}$ that a source must have to be detectable with 90% confidence in a search lasting $\frac{1}{3}$ year is

$$\begin{aligned} h_{3/\mathrm{yr}} &\cong 5h_\mathrm{n} \cong 11 \left[\frac{G}{c^3} \frac{kT_\mathrm{n}}{\int \sigma_0 \, \mathrm{d}f}\right]^{\frac{1}{2}} \\ &= 2.0 \times 10^{-16} \left[\frac{T_\mathrm{n}/1K}{\int \sigma_0 \, \mathrm{d}f / 10^{-21} \ \mathrm{cm}^2 \ \mathrm{Hz}}\right]^{\frac{1}{2}}. \end{aligned} \tag{82}$$

For a broad-band burst that is peaked near the detector's resonant frequency f_0 (so the f_c of equation (31a) is approximately f_0) and that lasts for a time not much longer than $\Delta t = 1/f_0$, the narrow-band characteristic amplitude (80) will be roughly equal to the broad-band characteristic amplitude (31b). For such bursts, and only for such bursts, it makes sense to plot a bar detector's narrow-band $h_{3/\mathrm{yr}}$ on the same graph (Fig. 9.4) as a broad-band detector's $h_{3/\mathrm{yr}}$. Inspiraling binaries do not belong to this class of bursts, so their detection by bars must be discussed separately from Fig. 9.4.

(c) *How to optimize the sensitivities of bar detectors*

Equation (82) shows that to optimize the sensitivity of a bar detector to short, broad-band bursts, one must achieve the largest possible frequency-integrated cross section $\int \sigma_0\, \mathrm{d}f$ and the smallest possible noise temperature T_n.

The integrated cross section $\int \sigma_0\, \mathrm{d}f$ can be computed by analyzing the response of the bar's fundamental mode to the force field (5) produced by a sinusoidal wave with optimal direction and polarization, and by then integrating over the wave's frequency. See MTW Box 37.4 for details. For a cylindrical bar with length L somewhat greater than radius R (the usual situation), the result depends only on the bar's mass M and internal sound velocity $v_s = (E/\rho)^{\frac{1}{2}}$ with $E =$ (Young's modulus) and $\rho =$ (density)(Paik and Wagoner, 1976):

$$\int \sigma_0\, \mathrm{d}f = \frac{8}{\pi}\frac{GMv_s^2}{c^3}\left[1 + \tfrac{1}{2}\nu(1-2\nu)(\pi R/L)^2 + \cdots\right]$$

$$= 1.6 \times 10^{-21}\,\mathrm{cm}^2\,\mathrm{Hz}\left(\frac{M}{10^3\,\mathrm{kg}}\right)\left(\frac{v_s}{5\,\mathrm{km\,s^{-1}}}\right)^2. \qquad (83)$$

(Here ν is the Poisson ratio of the bar's material, and only the leading shape-dependent corrections are given.) Thus, it is desirable to build detectors that are as massive and have as high sound speed as possible.

The noise temperature T_n is determined by a combination of noise in the sensor and thermal noise in the bar.

The *thermal noise in the bar* is caused by weak coupling of the bar's fundamental mode to its environment – its $\sim 10^{29}$ other modes, the wire or cable or prongs that suspend the bar, and the residual gas in the vacuum chamber (Braginsky, Mitrofanov and Panov, 1985). If, as is normal, this environment is thermalized at some physical temperature T_b (subscript 'b' for bar or for thermal bath), then these couplings cause the mode's amplitude to execute a random walk (Brownian motion) in the domain $|X| \lesssim X_{\mathrm{th}}$ corresponding to an energy kT_b:

$$X_{\mathrm{th}} = \left[\frac{2kT_b}{M_{\mathrm{eff}}(2\pi f_0)^2}\right]^{\frac{1}{2}}. \qquad (84)$$

The fluctuation-dissipation theorem states that the timescale on which this random walk produces changes of order X_{th} is the same as the timescale $\tau^* = Q/\pi f_0$ for large-amplitude vibrations to be damped frictionally. (Here Q is the quality factor of the mode's large-amplitude oscillations: the number of radians of oscillation required for the energy to damp by a factor e.) Consequently, the mean square Brownian change in the mode's amplitude during the sensor's averaging time $\hat{\tau} = 1 \cdot \Delta f$ is

$$|\Delta X_{\rm th}|^2 = X_{\rm th}^2 \frac{\hat{\tau}}{\tau^*} = X_{\rm th}^2 \frac{\pi}{Q} \frac{f_0}{\Delta f} \tag{85}$$

corresponding to

$$\tfrac{1}{2} M_{\rm eff}(2\pi f_0)^2 |\Delta X_{\rm th}|^2 = \pi \frac{kT_{\rm b}}{Q} \frac{f_0}{\Delta f}. \tag{86}$$

The details of the *sensor noise* depend on a variety of factors, including: the detailed structure of the transducer (for a review of many transducer structures see Section 4.1.5 of Amaldi and Pizella, 1979), the strength of coupling of the transducer to the bar (characterized by a dimensionless coupling constant β that is roughly equal to the number of cycles of oscillation required for all the fundamental mode's energy to be fed into the transducer); the impedance mismatch between the transducer and the amplifier; the noise temperature $T_{\rm a}$ of the amplifier (converted to an equivalent noise temperature $(f_0/f_{\rm a})T_{\rm a}$ at the bar's frequency f_0 if the amplifier operates at a different frequency $f_{\rm a}$ than the bar); and the sensor bandwidth Δf. Although the details vary from one sensor to another (see, e.g., Weiss, 1978; Pallotino and Pizella, 1980; Blair, 1983; Michelson and Taber, 1984), the spirit of the details is captured by the following approximate expression for the mean square change $|\Delta X_{\rm sensor}|^2$ that the experimenter would infer from the sensor's output if (i) only the sensor noise were present, (ii) back-action forces of the sensor on the bar were negligible (see subsection (f) below), and (iii) impedances were properly matched:

$$\tfrac{1}{2} M_{\rm eff}(2\pi f_0)^2 |\Delta X_{\rm sensor}|^2 \simeq \frac{kT_{\rm a}(f_0/f_{\rm a})}{\beta} \frac{\Delta f}{f_0}. \tag{87}$$

The sum of $\frac{1}{2} M_{\rm eff}(2\pi f_0)^2 |\Delta X_{\rm sensor}|^2$ (equation 87)) and $\frac{1}{2} M_{\rm eff}(2\pi f_0)^2 |\Delta X_{\rm th}|^2$ (equation (86)) is the mean square noise N^2 that appears in the denominator of equation (75); i.e., by definition, the detector's noise energy $kT_{\rm n}$ is this sum. If the experimenters choose too large a bandwidth Δf, then the sensor noise (87) will become inordinately large, producing a large $T_{\rm n}$ and masking the gravity-wave signal. If they choose too small a bandwidth, then the bar's thermal noise (86) will become inordinately large. There is an optimal bandwidth, typically $\Delta f \sim 10$ Hz in present detectors (corresponding to the averaging time $\hat{\tau} \sim 0.1$ s) at which roughly half the noise is from the sensor and half is from thermal motion of the bar. When the bandwidth is chosen optimally these two noise sources together produce a detector noise temperature

$$T_{\mathrm{n}} \simeq \left[\frac{\pi}{\beta}\frac{T_{\mathrm{b}}}{Q}\left(T_{\mathrm{a}}\frac{f_0}{f_{\mathrm{a}}}\right)\right]^{\frac{1}{2}}. \tag{88}$$

In practice, the experimenters choose a bar early in their experiments, thereby fixing the integrated cross section $\int \sigma_{\mathrm{o}}\,\mathrm{d}f$; and they then struggle for many years to develop a sensor and its coupling to the bar, and a thermally cold environment, that will minimize the noise temperature T_{n}. Maximizing $\int \sigma_{\mathrm{o}}\,\mathrm{d}f$ is achieved by maximizing the bar's mass and its velocity of sound (and, to a small extent, optimizing its shape subject to other experimental constraints such as available cryostats.) Minimizing T_{n} is achieved according to equation (88) by (i) maximizing the fundamental mode's quality factor Q (i.e. minimizing its coupling to the rest of the world), (ii) cooling the bar to as low a physical temperature T_{b} as possible, (iii) maximizing the strength β of coupling of the transducer to the bar, (iv) using an amplifier with as low a 'noise number' $kT_{\mathrm{a}}/(2\pi\hbar f_{\mathrm{a}})$ as possible (the Heisenberg uncertainty principle limits the noise number to be $\gtrsim 1$; Weber, 1959; Heffner, 1962; Caves, 1982), and (v) struggling to get good impedance matching of the transducer and amplifier (a requirement for equation (88) to be valid).

(d) *Parameters of first- and second-generation bar detectors*

The first-generation bar detectors (pre-1977) were all made of aluminum, weighed roughly 1.5 tonnes, and had $f_0 \simeq 1.6$ kHz and $Q \sim 10^5$; they all operated at room temperature, $T_{\mathrm{b}} \cong 300$ K; and most used piezo-electric transducers – i.e. crystals glued to the bar which produce small voltages when squeezed. For most the coupling of the transducer to the bar was weak, $\beta \lesssim 10^{-4}$; but those in Britain achieved strong coupling, $\beta \simeq 0.2$ and hence wide bandwidth $\Delta f/f_0 \sim 1$ at the price of reducing the bar's Q from $Q \sim 10^5$ to $Q \sim 2000$. The best amplifiers that could be impedance matched to the piezo-electric transducers had rather large noise numbers. For these first-generation bars the integrated cross-sections were $\int \sigma_{\mathrm{o}}\,\mathrm{d}f \simeq 2 \times 10^{-21}\,\mathrm{cm}^2\,\mathrm{Hz}$, and the lowest detector noise temperatures were $T_{\mathrm{n}} \simeq 4$ K corresponding to a minimum detectable burst amplitude with $\frac{1}{3}$ year of observation $h_{3/\mathrm{yr}} \cong 3 \times 10^{-16}$. Despite great effort in the early 70s by excellent experimenters, there was great room for improvement. (For a thorough review of the first-generation experiments see Amaldi and Pizella, 1979; for other reviews, see Drever, 1977, Tyson and Giffard, 1978 and Weber, 1986.)

In moving into the second generation, almost all the groups cooled their bars to liquid helium temperatures ($T_{\mathrm{b}} = 1.5$–4 K rather than 300 K).

Several of the groups (Maryland, Stanford, LSU) constructed massive bars ($M \cong 2$–5 tonnes) from a new alloy of aluminum that the Tokyo group discovered has a Q 10 times higher than the old one (5×10^7 vs 5×10^6 at 10^4 K temperature; Suzuki, Tsubono and Hirakawa, 1978). Perth and Moscow, by contrast, chose to use exotic bar materials. Perth used a 1.5 tonne Niobium bar with $Q = 2 \times 10^8$, while Moscow initially used a 10 kg sapphire bar with $Q = 4 \times 10^9$, and later when difficulty with cracking of the sapphire occurred, Moscow switched to a 10 kg silicon crystal bar with $Q = 2 \times 10^9$ – silicon and sapphire because of their very large Qs, an advantage bought at the price of low mass.

In the second generation each of the groups deemphasized piezo-electric transducers and developed some variant of one or another of two new transducer concepts: Moscow, LSU, Perth and Tokyo developed *parametric transducers* in which the bar's vibrations with frequency $f_0 \sim 10^3$ Hz modulate the capacitance in a microwave cavity (or rf circuit), thereby moving microwave photons from the frequency $f_D \sim 10^9$ Hz, at which the cavity is driven, into side bands at $f = f_D \pm f_0$, at which the amplifier measures the signal. The number of photons moved is proportional to the amplitude of vibration of the bar. Stanford and Maryland developed *resonant transducers* in which a mechanical diaphragm, with a mechanical resonant frequency very nearly that of the bar's fundamental mode, is attached to the bar's end. The vibration energy is quickly transferred back and forth between the bar and the diaphragm, with the diaphragm's displacement amplitude amplified over that of the bar by $|X_{\text{diaphragm}}|/|X_{\text{bar}}| \sim [(\text{bar mass})/(\text{diaphragm mass})]^{\frac{1}{2}}$. The vibrating diaphragm modulates the inductance of a superconducting circuit, and a SQUID (superconducting quantum interference device) amplifier is used to monitor the current in the circuit. Rome developed a similar resonant transducer with the diaphragm replaced by a toad stool which was one plate of a capacitor and with a FET transducer used to read out oscillations of the toad stool's voltage.

At present (late 1986) Stanford, Rome and LSU are all on the air with systems that have resonant frequencies $f_0 \cong 900$ Hz, cross-sections $\int \sigma_0 \, df \cong (4 \text{ to } 8) \times 10^{-21}$ cm^2 Hz, and noise temperatures $T_n \cong (0.01$–$0.04)$ K. Correspondingly the weakest burst they can detect with $\frac{1}{3}$ year of observation is $h_{3/\text{yr}} \cong 1.0 \times 10^{-17}$. Maryland and Perth are both likely to be on the air with similar characteristics before this book is published: and all five groups are hoping to push their noise levels down to $h_{3/\text{yr}} \sim 2 \times 10^{-18}$ within several years by further straightforward refinements. The Guangzhou group, having only entered the field very recently, is still at room temperature, but with a sensitivity $h_{3/\text{yr}} \simeq 1.6 \times 10^{-16}$, two times better than that of any of the first-generation room-temperature bars (Hu *et al.* 1986). The Moscow group with its small silicon and sapphire bars operates at a much higher frequency than any of the other groups, $f_0 \cong$ 8 kHz; by the time this book is published they will likely be on the air with $h_{3/\text{yr}} \sim 4 \times 10^{-17}$. The Tokyo group, on the other hand, has chosen a much lower frequency than the others, $f_0 \cong 60$ Hz and is now operating a narrow-band search for periodic waves from the Crab pulsar (see below). These sensitivities are shown in Fig. 9.4, along with those of other kinds of detectors. For details of the present detector configurations and near-term plans, see the gravitational-wave articles in the proceedings of recent conferences (Ruffini, ed., 1986; MacCallum, ed., 1987).

(e) *Sensitivies of bar detectors to periodic and stochastic waves*

Although I have discussed the detectors' performances entirely in the context of searching for short bursts of gravitational waves, bar detectors can be used also to search for periodic gravitational waves and a stochastic background at frequencies within the bandwidth Δf of their sensors.

In a search for periodic gravitational waves, the experimenters will typically use a bar with eigenfrequency f_0 slightly different from the expected frequency f of the waves. The waves then will drive the fundamental mode's complex amplitude X into oscillations with the beat frequency $f - f_0$. In searching for these oscillations the experimenters integrate for a long time; and correspondingly they can turn down the coupling strength β of the transducer to the bar, and/or narrow the bandwidth Δf, until the sensor's noise becomes negligible compared to the thermal noise of the bar (see, e.g., Michelson and Taber, 1981, 1984). (The maximum resulting bandwidth with the present Stanford detector would be $\Delta f \simeq 0.5$ Hz). When this is done, the general formulas of Section 9.4.2(a) are valid, with $S_h(f)$ the spectral density of the bar's thermal noise, converted into an equivalent gravity-wave spectral density

$$S_h(f) = \frac{4kT_\text{b}}{\int \sigma_\text{o}\, \mathrm{d}f} \frac{1}{f_0 Q}, \qquad (89)$$

and with the beam-factor averages having the values (79). In particular, the detector's characteristic noise amplitude (equation (51)) is

$$h_\text{n} = \left[15 \frac{G}{c^3} \frac{kT_\text{b}}{\int \sigma_\text{o}\, \mathrm{d}f} \frac{1}{Q} \frac{1}{f_0 \hat{\tau}} \right]^{\frac{1}{2}}, \qquad (90)$$

and the brightest source that can be seen with 90% confidence in $\hat{\tau} = \frac{1}{3}$ year of integration is

$$h_{3/\mathrm{yr}} = 1.7h_\mathrm{n} = 3.9 \times 10^{-25}\left(\frac{T_\mathrm{b}}{1\,\mathrm{K}}\right)^{\frac{1}{2}}\left(\frac{10^{-21}\,\mathrm{cm}^2\,\mathrm{Hz}}{\int \sigma_\mathrm{o}\,\mathrm{d}f}\right)^{\frac{1}{2}}\left(\frac{1000\,\mathrm{Hz}}{f_0}\right)^{\frac{3}{2}}\left(\frac{10^7}{Q}\right)^{\frac{1}{2}}. \tag{91}$$

The Tokyo group is currently carrying out a search for gravitational waves from the Crab pulsar using the above technique (Owa *et al.*, 1986). Their 74 kg, cryogenically cooled antenna has $Q=2.1\times 10^7$, $T_\mathrm{b}=4$ K, $f_0=60$ Hz, and $\int \sigma_\mathrm{o}\,\mathrm{d}f \simeq 2.2\times 10^{-27}\,\mathrm{cm}^2\,\mathrm{Hz}$ (so low because for noncylindrical antennas with frequency f_0 lowered by special shaping, $\int \sigma_\mathrm{o}\,\mathrm{d}f \propto f_0^2$; Hirakawa *et al.*, 1976). Correspondingly it has $h_{3/\mathrm{yr}} \sim 3\times 10^{-22}$. No other present bars are optimized for periodic waves, since there are no known sources in their frequency bands ($\sim$900 Hz and $\sim$8 kHz). However, with the technology of present burst-optimized bars it should be possible to achieve the thermal-noise-limited sensitivity (91) with $T_\mathrm{b}=4$ K, $\int \sigma_\mathrm{o}\,\mathrm{d}f \simeq 8\times 10^{-21}\,\mathrm{cm}^2$ Hz, $f_0 \simeq 900$ Hz, and $Q \simeq 5\times 10^6$ corresponding to $h_{3/\mathrm{yr}} \simeq 4\times 10^{-25}$ (Stanford, LSU, Rome, Maryland); and $T_\mathrm{b}=4$ K, $\int \sigma_\mathrm{o}\mathrm{d}f \simeq 2\times 10^{-23}\,\mathrm{cm}^2$ Hz, $f_0 \simeq 8$ kHz, and $Q \simeq 2\times 10^9$ corresponding to $h_{3/\mathrm{yr}} \simeq 1.4\times 10^{-25}$ (Moscow). See Fig. 9.6.

When searching for stochastic background it is desirable to open up the bandwidth Δf until the sensor noise becomes almost as large as the bar's thermal noise ($\Delta f \simeq 0.5$ Hz for the present Stanford bar). With $S_h(f)$ then given (approximately) by the thermal-noise spectral density (equation (89)), the noise amplitude h_n and the 90%-conficence $\frac{1}{3}$-year sensitivity $h_{3/\mathrm{yr}}$ for stochastic background become (equations (66) and (67))

$$h_\mathrm{n}(f) \simeq \left[15\frac{G}{c^3}\frac{kT_\mathrm{b}}{\int \sigma_\mathrm{o}\,\mathrm{d}f}\frac{1}{Q}\frac{1}{\sqrt{(\frac{1}{2}\hat{\tau}\,\Delta f)}}\right]^{\frac{1}{2}}, \tag{92}$$

$$h_{3/\mathrm{yr}} = 1.7h_\mathrm{n} \cong 8\times 10^{-22}\left(\frac{T_\mathrm{b}}{1\,\mathrm{K}}\right)^{\frac{1}{2}}\left(\frac{10^{-21}\,\mathrm{cm}^2\,\mathrm{Hz}}{\int \sigma_\mathrm{o}\,\mathrm{d}f}\right)^{\frac{1}{2}}\left(\frac{10^7}{Q}\right)^{\frac{1}{2}}\left(\frac{1\,\mathrm{Hz}}{\Delta f}\right)^{\frac{1}{4}}. \tag{93}$$

This $h_{3/\mathrm{yr}}$ is shown in Fig. 9.7 for the parameters of 1987 bar technology (those given in the last paragraph, plus $\Delta f \simeq 1$ Hz). For details of searches for stochastic background that were carried out using first-generation bar detectors, see Hough *et al.* (1975) and Hirakawa and Narihara (1975). For discussions of sensitivity that are more detailed and sophisticated than the above sketch, see Hirakawa, Owa and Iso (1985), Weiss (1979) and references therein.

(f) *Quantum limit and quantum non-demolition*

Quantum mechanics constrains the sensitivity that can be achieved by bar detectors using the present kinds of sensors: the fundamental mode of a bar, being highly decoupled from the rest of the world, can be regarded as a simple harmonic oscillator with mass $M_{\rm eff}$ and angular frequency $\omega_0 = 2\pi f_0$. As such, it is subject to the laws of quantum mechanics for oscillators: its generalized position x and momentum p must be regarded as hermitian operators that fail to commute, $[x, p] = i\hbar$; and, correspondingly, the real and imaginary parts of its complex amplitude,

$$X_1 = x \cos \omega_0 t - \left(\frac{p}{M_{\rm eff}\omega_0}\right) \sin \omega_0 t,$$

$$X_2 = x \sin \omega_0 t + \left(\frac{p}{M_{\rm eff}\omega_0}\right) \cos \omega_0 t, \tag{94}$$

are non-commuting hermitian operators with commutator

$$[X_1, X_2] = \mathrm{i}\frac{\hbar}{M_{\rm eff}\omega_0}. \tag{95}$$

From this commutation relation we infer, via the Heisenberg uncertainty principle, an absolute limit on the variances of X_1 and X_2 in any quantum mechanical state (Thorne *et al.*, 1978):

$$\Delta X_1 \, \Delta X_2 \geqslant \frac{\hbar}{2M_{\rm eff}\omega_0}. \tag{96}$$

The principles of quantum mechanics guarantee that one can never measure X_1 and X_2 simultaneously with a precision that violates this uncertainty principle. In fact, it turns out (Caves, 1982; Yamamoto and Haus, 1986) that even the most ideal of measuring systems will introduce additional noise equal to (96), thereby producing a lower bound

$$\Delta X_1 \, \Delta X_2 \geqslant \frac{\hbar}{M_{\rm eff}\omega_0} \tag{97}$$

on the products of the rms noise in the measured values of X_1 and X_2.

The sensors used in the present generation of bar detectors measure X_1 and X_2 simultaneously with equal accuracies; and, correspondingly, in searches for short bursts they are subject to the 'standard quantum limit' (Braginsky and Vorontsov, 1974; Giffard, 1976)

$$T_{\rm n} = \tfrac{1}{2} M_{\rm eff}\omega_0^2[(\Delta X_1)^2 + (\Delta X_2)^2]/k \geqslant \hbar\omega_0/k = 4.8 \times 10^{-8}\,\mathrm{K}\left(\frac{f_0}{1000\,\mathrm{Hz}}\right). \tag{98}$$

Notice that this standard quantum limit places the severe constraint

$$h_{3/\mathrm{yr}} \gtrsim 4.4 \times 10^{-20} \left(\frac{f_0}{1000\ \mathrm{Hz}} \right)^{\frac{1}{2}} \left(\frac{10^{-21}\ \mathrm{cm}^2\ \mathrm{Hz}}{\int \sigma_0\, \mathrm{d}f} \right)^{\frac{1}{2}} \tag{99}$$

on the detector's burst sensitivity (82). It is fairly likely, though far from certain, that the strongest kilohertz-frequency bursts striking the earth three times per year have characteristic amplitudes $h_c < 10^{-20}$ (see Section 9.4.1); and, correspondingly, it may turn out to be crucial for bar detectors of the future to circumvent the standard quantum limit (98), (99).

The uncertainty principle (96) suggests a promising method for circumventing the standard quantum limit (Thorne *et al.*, 1978; Braginsky, Vorontsov and Khalili, 1978): one should devise a new kind of sensor that measures X_1 with high accuracy, while giving up accuracy on X_2. Such sensors, called 'back-action-evading sensors' (a special case of 'quantum non-demolition sensors'), are now under development in a number of laboratories (Braginsky, 1983; Bocko and Johnson, 1984; Oelfke, 1983; Blair, 1982); and they may make possible bar sensitivities in the 1990s that will beat the standard quantum limit by modest factors. For a detailed review of quantum non-demolition measurements – i.e. measurements that do not change the quantum state of the system being measured – see Caves (1983).

Although a back-action-evading sensor gives up accuracy on one of the wave's two quadrature components, that accuracy can be regained by looking at the same wave with two different detectors: on one detector, with complex amplitude $X = X_1 + iX_2$, measure X_1 with high accuracy and X_2 with poor; on the other, with complex amplitude $Y = Y_1 + iY_2$, measure Y_2 with high accuracy and Y_1 with poor. From X_1 infer the detailed evolution of one of the wave's two quadrature components; from Y_2 infer the evolution of the other. In this way, in principle, the quantum mechanical properties of the detector can be completely circumvented and the only constraints of principle on the accuracy of measurement are associated with quantization of the waves themselves. For further discussion and details see the reviews by Caves *et al.* (1980); Caves (1983); Braginsky, Vorontsov and Thorne (1980).

It is worth noting that a back-action-evasion measurement, ideally performed, should drive the bar's fundamental mode into a 'squeezed state' (Hollenhorst, 1979). Squeezed states have been studied extensively in recent years in the context of quantum optics (see, e.g. Schumaker, 1986; and Walls, 1983); and we shall return to them in Section 9.5.3(f) below when discussing beam detectors.

(g) *Looking toward the future*

It may be that bar detectors in the distant future will find their greatest applications at much lower frequencies than are now common: at lower frequencies the bar can be much longer and more massive, and correspondingly more sensitive; and good bandwidth, $\Delta f \sim f_0$ will correspond to a longer averaging time and thus will be easier to achieve. For discussion, see Michelson (1986).

Although individual bar detectors in the foreseeable future will be limited to moderately small bandwidths $\Delta f/f \lesssim 0.2$, and therefore will be able to acquire only very limited information about the wave form $h^{TT}_{jk}(t)$ of a gravity-wave burst, once waves are being detected in profusion it may be possible cheaply to replicate the detectors with a variety of sizes and hence a variety of fundamental-mode frequencies, thereby creating a 'xylophone' of networked detectors with a large overall bandwidth (Michelson and Taber, 1984).

9.5.3 Beam detectors

(a) *A brief history of beam-detector research*

The germ of the idea of a laser-interferometer gravitational-wave detector ('beam detector') can be found in Pirani (1956); but – so far as I am aware – the first explicit suggestion of such a detector was made by Gertsenshtein and Pustovoit (1962). In the mid-1960s Joseph Weber, unaware of the Gertsenshtein–Pustovoit work, reinvented the idea but left it lying in his laboratory notebook unpublished and unpursued. In 1970 Rainer Weiss at MIT, unaware of Gertsenshtein–Pustovoit or Weber, reinvented the idea and carried out a detailed design and feasibility study (Weiss, 1972) in which many of the techniques now being used were conceived. Unfortunately, Weiss was unable to obtain funding to push forward with a significant experimental effort.

Robert Forward at Hughes Research Laboratories in Malibu, California, having learned the concept of the beam detector from Weber (his former thesis advisor), was motivated indirectly by Weiss in 1971 to construct a prototype detector with funding from Hughes. By 1972 Forward and his colleagues at Hughes were operating the world's first prototype beam detector – an instrument that demonstrated the idea could really work, and that was remarkably sensitive considering the modest effort put into it: $[S_h(f)]^{\frac{1}{2}} \simeq 2 \times 10^{-16}$ Hz$^{-\frac{1}{2}}$ between 2500 Hz and 25000 Hz, corresponding to $h_{3/\mathrm{yr}} \simeq 1 \times 10^{-13}$ for 2500 Hz bursts (Moss, Miller and Forward, 1971;

Forward and Moss, 1972; Forward, 1978). Regrettably. Forward could not obtain funds to move from this first prototype to a more sophisticated instrument; so his project was shut down.

With the completion of the first generation of bar detectors in 1975, each experimental group that decided to stay in the field looked carefully at a variety of possibilities for sensitivity improvement. While most groups decided to stick with bars, two switched to beam detectors: Munich, led by H. Billing, and Glasgow, led by Ronald Drever with Jim Hough second in command. The Munich group was strongly influenced by a proposal to develop beam detectors that Weiss had submitted to NSF, and that NSF had refused to fund; and so Munich pushed forward (Winkler, 1977) along the lines that Weiss had hoped to follow, using a Michelson interferometer design (see below). The Glasgow group first built a small Michelson interferometer (Drever *et al.*, 1977), then switched in 1977 to a new Fabry–Perot design invented by Drever (Drever *et al.*, 1980).

In 1979 Caltech managed to attract Drever away from Glasgow (part-time at first, full-time later), leaving Hough as the Glasgow leader. At Caltech Drever started up a beam-detector project; and NSF, finally recognizing that beam detectors were worth funding, agreed to support both Weiss at MIT to develop his original idea of a Michelson system and Drever at Caltech to develop his Fabry–Perot system. More recently, in 1983, Alain Brillet initiated a beam-detector effort in Orsay, France (Brillet and Tourrenc, 1983; Brillet, 1985).

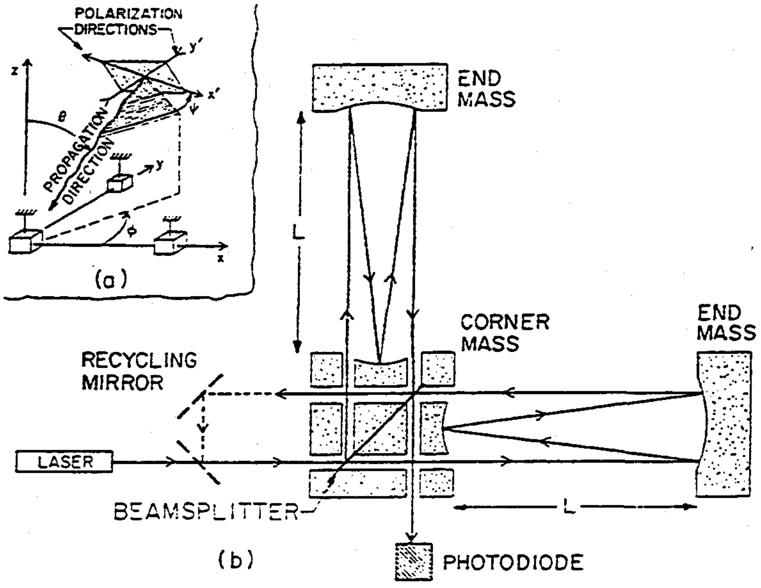
Munich, Glasgow, Caltech and MIT all now have working beam detectors with amplitude sensitivities ~2000 times better than that of Forward's first prototype but ~5 times worse than the best bars. These detectors are small-scale (1–40 m) prototypes for the full-scale (several kilometer) beam detectors that will be required for real success. Design and costing studies are now underway for the full-scale systems (called 'Laser Interferometer Gravity Wave Observatories' or LIGOs); for details of these studies, see Linsay *et al.* (1983), Drever *et al.* (1985), Maischberger *et al.* (1985), Winkler *et al.* (1986), Hough *et al.* (1986). There is hope that full-scale LIGOs will be constructed in the late 1980s and early 1990s and will be operating in the mid- to late-1990s with sensitivities in the region where gravity waves are expected.

(b) *How a beam detector works*

The current and planned beam detectors are designed to operate at frequencies below 10 kHz because astrophysical arguments suggest that the waves will be weak above this frequency (see page 158 of Thorne, 1978); and they have their best sensitivities at frequencies below 1 kHz. Correspondingly, the waves they seek all have reduced wavelengths $\lambda\!\!\!^{-} > 5$ km and most have $\lambda\!\!\!^{-} > 50$ km. Since the planned detectors all have sizes $L \leqslant 4$ km, the condition $L \ll \lambda\!\!\!^{-}$ for use of a 'proper-reference-frame analysis' is satisfied, though only marginally in extreme cases. I shall use such an analysis in the discussion below. For an outline of the alternative, TT analysis, see, e.g., Exercise 37.6 of MTW.

A beam *detector* consists of one or more *receivers* that are operated simultaneously, with cross-correlated outputs – the cross-correlation, as usual, being the key to removing spurious, non-Gaussian noise. A simple version of a Michelson-type receiver is shown in Fig. 9.9, three-dimensionally in part (*a*) and as seen from above in part (*b*). (Ignore for the moment the propagation and polarization pieces of part (*a*).) The receiver consists of three masses which hang on wires from overhead supports and swing like pendula. The masses are arranged at the ends and corner of a right-angle L. When a gravitational wave propagates vertically through the receiver with polarization axes along the L ('+' polarization), its

Fig. 9.9. Schematic diagram of a *Michelson-type beam receiver* for gravitational waves (part (*b*)), and of the waves' propagation and polarization angles (θ, ϕ, ψ) relative to the receiver (part (*a*); cf. Fig. 9.2).



quadrupolar force field (5) pushes together the masses on one arm of the L while pushing apart the masses on the other arm. In the next half-cycle of the wave, the directions of the pushes are reversed. Since the waves being sought have frequencies f far above the 1 Hz swinging frequency of the pendula, the pendular restoring forces have no opportunity to make themselves felt: the masses respond to the gravity-wave pushes as though they were free. With the origin of the proper reference frame placed on the central mass, the central mass is left unaffected while the end masses oscillate longitudinally with displacements

$$\delta x(t)=\tfrac{1}{2}Lh_{+}(t) \quad \text{for mass on } x \text{ axis}, \tag{100a}$$

$$\delta y(t)=-\tfrac{1}{2}Lh_{+}(t) \quad \text{for mass on } y \text{ axis} \tag{100b}$$

(equation (6)). Here L is the (approximately equal) length of each arm. Correspondingly, there is an oscillation in the difference $l(t)$ of the arm lengths, $\delta l(t)=\delta x(t)-\delta y(t)$, given by

$$\delta l(t)=h_{+}(t)L. \tag{101}$$

It is straightforward to show that in the more general case of a wave which impinges from a direction (θ, ϕ) on the sky with polarization axes rotated at an angle ψ relative to the constant-ϕ plane (Fig. 9.9(a)), the difference in arm lengths l oscillates as

$$\delta l(t)=h(t)L, \tag{102}$$

where $h(t)$ has the standard form (26)

$$h(t)=F_{+}(\theta, \phi, \psi)h_{+}(t; \iota, \beta)+F_{\times}(\theta, \phi, \psi)h_{\times}(t; \iota, \beta), \tag{103}$$

with beam-pattern factors (cf. Forward, 1978; Rudenko and Sazhin, 1980; Estabrook, 1985; Schutz and Tinto, 1987)

$$F_{+}(\theta, \phi, \psi)=\tfrac{1}{2}(1+\cos^2\theta)\cos 2\phi\cos 2\psi-\cos\theta\sin 2\phi\sin 2\psi, \tag{104a}$$

$$F_{\times}(\theta, \phi, \psi)=\tfrac{1}{2}(1+\cos^2\theta)\cos 2\phi\sin 2\psi+\cos\theta\sin 2\phi\cos 2\psi. \tag{104b}$$

The difference of arm lengths $l(t)$ is monitored by Michelson interferometry: a beam splitter and two mirrors are attached to the corner mass as shown in Fig. 9.9(b), and one mirror is attached to each end mass. A laser beam shines through a hole in the corner mass and onto the beam splitter, which directs half the beam toward each end mass. The mirrors on the end masses reflect the beams back toward the corner-mass mirrors, which in turn reflect the beams back to the end masses, which reflect the beams back through holes in the corner-mass mirrors and onto the beam splitter where they are recombined. Part of the recombined beam goes out one side of the beam splitter toward the laser (ignore for now the 'recycling mirror' in (*b*), it is absent in the simple version of the receiver being discussed here); the other part of the recombined beam goes out the other side toward a photodetector. Oscillations in the arm-length difference $\delta l(t)$ produce oscillations in the relative phases of the recombining light, and thence oscillations in the fraction of the light which goes to the photodetector versus that which goes back toward the laser. The photodetector, by monitoring the oscillations in received intensity, in effect is monitoring the oscillations $\delta l(t)$ of arm-length difference and thence the gravity-wave oscillations $h(t)$.

In practice the laser beams are made to bounce back and forth in the arms not just twice as shown in Fig. 9.9(*b*), but rather a large number of round-trip times B, making B distinct spots on each end mirror. In the simple case that the gravity-wave-induced arm-length difference $\delta l = Lh$ does not change much during these many round trips (see subsection (e) below for the case of large change), the bouncing light beam will build up during its B trips a total phase delay

$$\Delta\Phi = \frac{2B\,\delta l}{\bar{\lambda}_e} = \frac{2BL}{\bar{\lambda}_e}\,h \qquad (105)$$

where $\bar{\lambda}_e = \lambda_e/2\pi$ is the reduced wavelength of the light ($\bar{\lambda}_e = 0.0818$ microns for the light from the argon ion lasers currently being used). This phase delay can be monitored, by the photodetector, with a precision $\Delta\Phi = 1/(N_\gamma \eta)^{\frac{1}{2}}$, where N_γ is the total number of photons that the laser puts out during the time $\hat{\tau}$ over which the photodetector intensity is averaged, and η is the photon counting efficiency of the photodetector ($\eta \sim 0.4$–0.9). When searching for a gravity-wave burst with characteristic frequency f, it is optimal to average the photodetector intensity for half a gravity-wave period, $\hat{\tau} \cong 1/2f$; and correspondingly the phase delay can be inferred with a photon-counting-noise ('shot-noise') precision

$$\Delta\Phi_{\text{shot}} = \frac{1}{(N_\gamma \eta)^{\frac{1}{2}}} \simeq \left(\frac{\hbar c/\bar{\lambda}_e}{I_o \eta (1/2f)}\right)^{\frac{1}{2}}, \qquad (106)$$

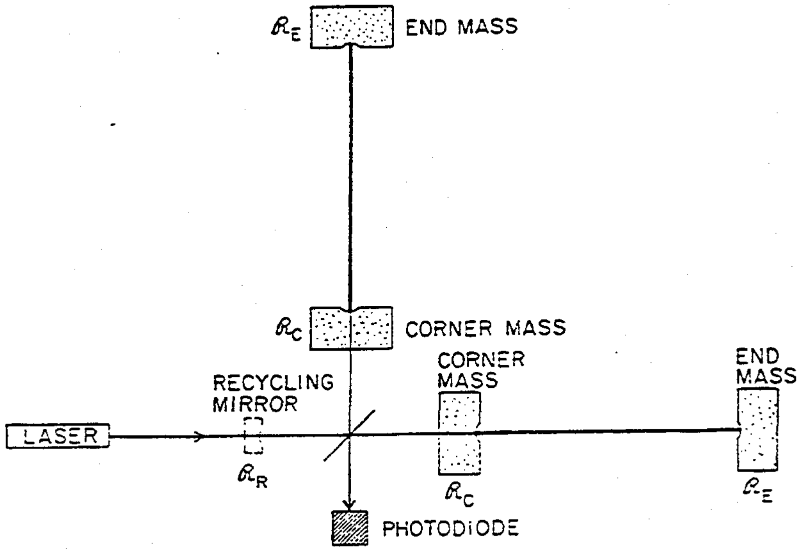
where I_o is the laser output power. By comparing equations (105) and (106) we obtain a rough estimate of the amplitude of a gravity-wave burst that produces a signal of the same strength as the rms shot noise

$$\begin{aligned} h_{\text{shot}} &\cong \left[\frac{2\hbar c \lambda\!\!\!^{-}_{e}}{I_{o}\eta}\frac{f}{(2BL)^{2}}\right]^{\frac{1}{2}} \\ &\cong 7.2\times 10^{-21}\,\frac{50}{B}\,\frac{1\,\text{km}}{L}\left(\frac{\lambda\!\!\!^{-}_{e}}{0.082\,\mu\text{m}}\right)^{\frac{1}{2}}\left(\frac{10\ \text{Watts}}{I_{o}\eta}\right)^{\frac{1}{2}}\left(\frac{f}{1000\ \text{Hz}}\right)^{\frac{1}{2}}. \end{aligned} \qquad (107)$$

This formula and these numbers give some indication of the potential sensitivities of beam receivers.

Fabry–Perot beam receivers have essentially the same potential sensitivities as Michelson receivers. Fig. 9.10 shows a Fabry–Perot receiver viewed from above. Here, by contrast with Fig. 9.9, the corner mass has been broken into three separate pieces, one carrying each mirror and one carrying the beam splitter. This breaking up of the corner mass, initiated in Munich, reduces spurious forces on the mirrors; since 1985 it has been standard in all beam receivers. In the Fabry–Perot system of Fig. 9.10 each arm is operated as a resonant Fabry–Perot cavity: light from the laser is split at the beam splitter and enters the two cavities through the backs of the corner masses' partially transmitting mirrors. The two arms are arranged to have equilibrium lengths very nearly equal to a half-integral multiple of the wavelength $2\pi\lambda\!\!\!^{-}_{e}$ of the laser light. Consequently, the entering light finds itself in resonance with a mode of each cavity; and it gets resonantly

Fig. 9.10. Schematic diagram of a *Fabry–Perot-type beam receiver* for gravitational waves.



trapped in the cavity, building up to high intensity before exiting back toward the beam splitter.

Slight changes in the length of each cavity drive the cavity slightly off resonance and thereby produce sharp changes in the phase of the exiting light. Consequently, when the exiting light beams from the two cavities recombine at the beam splitter, their relative phase is highly sensitive to slight modulations δl of the two cavities' length difference l; and correspondingly the intensity of light onto the photodetector is highly sensitive to δl. If the corner mirrors have a probability for reflecting photons $\mathscr{R}_C$ and a transmission probability $1-\mathscr{R}_C$ (and no scattering), and if the end mirrors reflect much more efficiently, then this Fabry–Perot sensitivity is described by the same formulas (105)–(107) as for a Michelson receiver, with the number of round-trips B in each Michelson arm replaced by $4/(1-\mathscr{R}_C)$:

$$B \rightarrow 4/(1-\mathscr{R}_C). \tag{108}$$

It is also possible (and, in fact, is current practice) to operate a Fabry–Perot receiver in an alternative mode where, instead of recombining and interfering the beams, the laser's frequency is locked to the eigenfrequency of one arm and the difference between the laser's frequency and that of the other arm is the gravity-wave signal; see, e.g. Hough *et al.* (1983) or Spero (1986*a*) for details. This mode of operation is technically easier than beam recombination, but the ease is bought at the price of some debilitation in the ultimately achievable shot noise.

(c) *Noise in beam detectors*

Photon shot noise is but one of many noise sources that plague beam detectors. Almost always, thus far, the other noise sources are so strong that the effects of shot noise are lost amidst them. Typically the experimenters struggle for a long time to reduce the other noises sufficiently that shot noise shows up; then they improve the shot noise somewhat by increasing the laser power; then they begin a long struggle once again with the other noise sources. In this way the overall gravity-wave amplitude noise at kilohertz frequencies has been reduced during the period 1980–86 by a factor ~ 1000 (10^6-fold improvement in energy noise).

Among the most serious of the other noise sources are the following: amplitude and phase fluctuations in the laser output; imperfect matching of the laser beam into the Fabry–Perot cavities or the Micheison mirror system (imperfections in beam direction, beam shape, and beam wavelength); imperfect matching of the wave fronts of the recombining beams at the beam splitter; fluctuations in the index of refraction of the gas inside the arms (even though the arms are inside vacuum pipes, residual gas can cause problems); scattering of light from one part of the optical system into another; thermal noise (thermal vibrations) in the end and corner masses and in the wires that suspend them; imperfect alignment of the mirrors; and seismic and acoustic noise from the outside world, which cause vibrations of the wires from which the masses hang. Almost all of these effects produce a displacement noise δl that is independent of the arm length L; and, correspondingly, their effects on gravity-wave amplitude sensitivities scale as $h \propto 1/L$. It is this scaling that motivates the move from short prototypes to long LIGOs.

A beam receiver is intrinsically broad band: its output, the intensity of the light into the photodetector $I_{\rm pd}(t)$, is averaged (by filtering) over the shortest timescale, $\hat{\tau} \sim 10^{-4}$ s, that gravity waves are likely to contain, and then is recorded for future analysis. The future analysis can include searching in the data for signals of all frequencies $f \lesssim 1/\hat{\tau} \sim 10^4$ Hz, and for signals with complicated time dependences that embody a broad range of frequencies. Correspondingly, a beam receiver's net noise depends on the frequency f at which one studies the recorded output.

More specifically, the noise (excluding non-Gaussian, spurious events which are removed by coincident operation of two independent receivers) is characterized by the spectral density $S_l(f)$ of the output-inferred armlength difference l; or, equally well, by the spectral density $S_h(f)$ of the gravity wave $h(t)$ (equation (103)). Because the effect of the gravity wave on the arm-length difference is $\delta l(t) = Lh(t)$, $S_h(f)$ is also called the spectral density of strain, and these two spectral densities are related by

$$S_h(f) = (1/L^2)S_l(f). \tag{109}$$

It is $S_h(f)$ or its square root, or $S_l(f)$ or its square root, that experimenters generally quote when discussing the overall performance of their beam detectors.

In Sections 9.4.1(b), 9.4.2(a) and 9.4.3(a) we derived expressions in terms of $S_h(f)$ for the minimum-amplitude signal $h_{3/\rm yr}$ that can be detected with 90% confidence in a $\frac{1}{3}$ year search using a detector composed of two identical cross-correlated receivers. Those expressions involve averages over the receiver's beam-pattern factors. For the L-shaped beam receivers of Figs. 9.9 and 9.10, with beam-pattern factors (104), the appropriate averages are

$$\langle F_+^2 \rangle = \langle F_\times^2 \rangle = \tfrac{1}{5}, \tag{110}$$

and, correspondingly,

$$h_{3/\mathrm{yr}}(f_c)=11[f_c S_h(f_c)]^{\frac{1}{2}} \text{ for bursts (equation (34))}, \qquad (111)$$

$$h_{3/\mathrm{yr}}(f)=3.8[S_h(f)\times 10^{-7}\ \mathrm{Hz}]^{\frac{1}{2}}$$

$$\text{for periodic waves (equation (52a))}, \qquad (112)$$

$$h_{3/\mathrm{yr}}(f)=4.5\left(\frac{\Delta f}{10^{-7}\ \mathrm{Hz}}\right)^{-\frac{1}{4}}[fS_h(f)]^{\frac{1}{2}}$$

$$\text{for stochastic waves (equation (67))}. \qquad (113)$$

Here in the burst equation the numerical factor in equation (34) has been taken to be 5 corresponding to a characteristic frequency $f_c \sim 10$ Hz to 10^4 Hz – the largest band that earth-based receivers are likely ever to operate in; and in the periodic equation it is assumed that the frequency and phase are known in advance (equation (52a) rather than (52b)). Expressions (111)–(113) are the bases of the beam-detector noise amplitudes shown in Figs. 9.4, 9.6 and 9.7.

When the two receivers that make up a detector do not lie in the same plane (because of curvature of the earth between them) or do not have their arms parallel, their joint sensitivity is somewhat debilitated. Detailed analyses of this have been carried out by Whitcomb and Saulson (1984) and by Schutz and Tinto (1986).

(d) *The noise in the present prototypes*

At present the MIT group is developing Michelson receivers using a prototype with $L=1.5$ m; the Munich group is developing Michelsons using $L=30$ m (Shoemaker *et al.*, 1986); the Glasgow group is developing Fabry–Perot receivers using $L=10$ m; and the Caltech group is developing Fabry–Perots using $L=40$ m. Although the arm lengths and optics of these four prototypes are very different, their displacement sensitivities are roughly the same – and have been so throughout the past six years, during which all have improved roughly in step by about three orders of magnitude. Throughout these improvements Munich has maintained a slight (factor 2 or 3 in amplitude) lead over the other groups. (*Note added in proof*. Since this was written Glasgow has forged ahead of Munich by a factor 3 in displacement sensitivity, making them equal in h sensitivity.) Fig. 9.11 shows the spectral density of strain noise for the Munich prototype as of February 1986 (Schilling, 1986). The noise is due, predominantly, to seismic noise (inadequate isolation) below 250 Hz, probably seismic noise between 250 and 1000 Hz, photon shot noise between 1000 and 6000 Hz, and thermal noise in the mirrors and pockels cells above 6000 Hz. In Figs. 9.4, 9.6 and 9.7 (upper right) are shown the sensitivities $h_{3/\mathrm{yr}}$ (equations (111)–(113)) for the Munich and Caltech prototypes in 1986 (Schilling, 1986; Spero, 1986*b*) and Glasgow in early 1987. As an illustration of the sensitivity progress, Fig. 9.4 also shows $h_{3/\mathrm{yr}}$ for bursts in the Munich and Caltech prototypes a few months after each was first turned on (1980 and 1983).

(e) *Spectral density of shot noise for simple, recycling and resonating receivers*

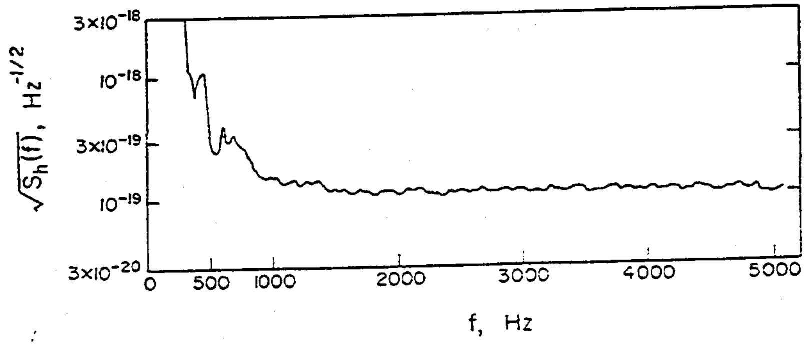
In the present prototypes it is advantageous to store the light in the arms as long as possible, thereby building up the largest possible phase shift and gravity-wave sensitivity; cf. the B-dependence in equations (105) and (107). However, in a kilometer-scale LIGO one easily can store the light longer than half a gravity-wave period, i.e. for more than $B=75(1\ \mathrm{km}/L)\times(1000\ \mathrm{Hz}/f)$ round-trip traverses of an arm. Such long storage is self-defeating: the phase shift built up so laboriously during the first half-period of the wave gets removed during the second half-period because $h(t)$ reverses sign.

This shows up clearly in the spectral densities of shot noise $S_h(f)$ for the idealized Michelson and Fabry–Perot receivers of Figs. 9.9 and 9.10 (still without the 'recycling mirrors' in place). Assuming as above that the Michelson mirrors have negligible losses during B round trips in each arm, and the Fabry–Perot end mirrors have negligible transmission compared to the corner mirrors

$$1-\mathscr{R}_{\mathrm{E}} \ll 1-\mathscr{R}_{\mathrm{C}} \equiv 4/B \tag{114}$$

(cf. equation (108)), the spectral densities of shot noise are (cf. Gürsel *et al.*,

Fig. 9.11. Square root of spectral density of noise $[S_h(f)]^{\frac{1}{2}}$ plotted against frequency f for the Munich Michelson-type beam detector with 30 m arms, as of February 1986.

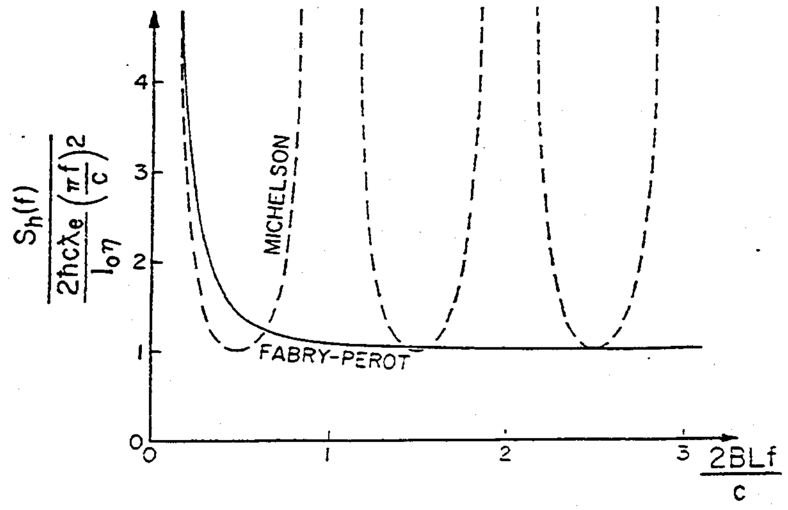


1983; Meers, 1983; Brillet and Meers, 1987)

$$S_h(f) = \frac{2\hbar c\lambdabar_e}{I_o\eta}\left(\frac{1}{2BL}\right)^2 \times \begin{cases} 1+(2\pi BLf/c)^2 & \text{for Fabry–Perot} \quad (115a) \\ \left[\dfrac{2\pi BLf/c}{\sin(2\pi BLf/c)}\right]^2 & \text{for Michelson.} \quad (115b) \end{cases}$$

These spectral densities are shown in Fig. 9.12 as a function of $2BLf/c=$ (light storage time)/(gravity-wave period), with (gravity-wave period) = $1/f$ held fixed. When (storage time)/(period) $\ll 1$ the two receivers have the same shot noise, and that shot noise improves with increasing storage time as $S_h(f) \propto (\text{storage time})^{-2}$. When (storage time)/(period) $\gtrsim 1$ the Fabry–Perot shot noise stops improving, while the Michelson shot noise undergoes a series of oscillations with minima equal to the Fabry–Perot shot noise. Note that the minima of the Michelson oscillations occur when $2BLf/c=$ (light-storage time)/(gravity-wave period) = 0.5, 1.5, 2.5, . . . , while their

Fig. 9.12. The improvement of photon shot noise with increasing number of bounces of light in a simple beam receiver. The receiver is of the Fabry–Perot (Fig. 9.10) or Michelson (Fig. 9.9) type with recycling mirrors absent, with $1-\mathscr{R}_E \ll 1-\mathscr{R}_C = 4/B$ in the Fabry–Perot case and with B round trips in each arm in the Michelson case. Plotted vertically is spectral density of shot noise in units proportional to f^2 but independent of the light-storage time $2BL/c$. Plotted horizontally is (light-storage time)/(gravity-wave period) = $(2BL/c)f$. The Fabry–Perot shot noise is given by the solid curve (equation (115a)); the Michelson by the dashed curve (equation (115b)).



maxima ($S_h = \infty$) occur when the light is stored for an integral number of gravity-wave periods. This is just what we would expect from the fact that $h(t)$ reverses sign every half-period, thereby removing during a second half-period the signal put onto the light during a first half-period. In the Fabry–Perot receiver there are no oscillations of $S_h(f)$ because different photons experience different storage times – i.e. because of the probabilistic nature of the reflectivity $\mathscr{R}_{\rm C}$.

Drever (1983) has devised a method for improving the sensitivity of either a Michelson or a Fabry–Perot when mirror reflectivities permit storing light for much longer than a half-period. The basic idea is to extract the light after a half-period, when further storage is self-defeating, and then reinsert it into the cavity along with and in phase with new laser light. More specifically, for the Michelson of Fig. 9.9(*b*), one adjusts the relative arm lengths so that very little of the recombined light emerges from the beam splitter toward the photodiode where the gravity-wave signal is read out (a mode of operation that optimizes the sensitivity, it turns out). Then most of the recombined light emerges toward the laser; and 'recycling mirrors' are inserted to direct that emergent light back into the beam splitter along with fresh laser light. For the Fabry–Perot of Fig. 9.10 the same effect is achieved with a single recycling mirror.

As an aid in quantifying the recycling-induced improvement in shot noise, consider the (realistic) situation in which the technology of mirror coatings limits the mirror reflectivities to some maximum value $\mathscr{R}_{\rm max}$ (0.9999 at present; perhaps 0.999 99 a few years from now). It obviously is optimal to place at the end of each arm a mirror with this maximum, so $\mathscr{R}_{\rm E} = \mathscr{R}_{\rm max}$. Consider, for concreteness, the Fabry–Perot receiver of Fig. 9.10. If the corner mirrors are chosen also to have the maximum reflectivity, $\mathscr{R}_{\rm C} = \mathscr{R}_{\rm E}$, then essentially all the unused light leaks out the end mirrors and nothing is gained by recycling. In this case of a simple, non-recycling Fabry–Perot with $\mathscr{R}_{\rm C} = \mathscr{R}_{\rm E} = \mathscr{R}_{\rm max}$ the spectral density of shot noise is (cf. Gürsel *et al.*, 1983; Meers, 1983; Brillet and Meers, 1987)

$$S_h(f) = S_{\rm o}[1 + (f/f_{\rm o})^2], \tag{116a}$$

where

$$\begin{aligned} S_{\rm o} &= \frac{2\hbar c \bar{\lambda}_{\rm e}}{I_{\rm o}\eta}\left(\frac{1-\mathscr{R}_{\rm E}}{2L}\right)^2 \\ &= \left[\frac{3.6\times 10^{-25}}{({\rm Hz})^{\frac{1}{2}}}\left(\frac{\bar{\lambda}_{\rm e}}{0.0818\ \mu{\rm m}}\right)^{\frac{1}{2}}\left(\frac{100\ {\rm W}}{I_{\rm o}\eta}\right)^{\frac{1}{2}}\left(\frac{1-\mathscr{R}_{\rm E}}{10^{-4}}\right)\left(\frac{1\ {\rm km}}{L}\right)\right]^2, \end{aligned} \tag{116b}$$

$$f_o = \frac{(1-\mathscr{R}_E)c}{4\pi L} = 2.4\ \text{Hz}\left(\frac{1-\mathscr{R}_E}{10^{-4}}\right)\left(\frac{1\ \text{km}}{L}\right). \tag{116c}$$

(Note that, because so much light is lost out the end mirrors, this noise is worse than the $\mathscr{R}_C \gtrsim \mathscr{R}_E$ limit of equation (115a) – worse by a factor 4 at frequencies $f \gg f_o$ and by a factor 16 at $f \ll f_o$.) This non-recycled ('simple') Fabry–Perot noise is shown as a solid curve in Fig. 9.13.

An experimenter who wishes to improve on this noise level by recycling must choose a frequency $f_k \gg f_o$ near which the noise level is to be minimized. The minimal noise level near $f = f_k$ will then be achieved by setting

$$1-\mathscr{R}_C = 8\pi L f_k/c, \tag{117a}$$

so the effective storage time in each arm is $2BL/c = 8Lc^{-1}/(1-\mathscr{R}_C) = (\pi f_k)^{-1} = (1/\pi) \times$ (period of a gravity wave with the optimal frequency f_k). Minimal noise also requires a special choice for the reflectivity $\mathscr{R}_R$ of the recycling mirror (Fig. 9.10)

$$1-\mathscr{R}_R = \frac{4(1-\mathscr{R}_E)}{(1-\mathscr{R}_C)}. \tag{117b}$$

With these choices of reflectivity, the light-recycling Fabry–Perot receiver of Fig. 9.10 has shot noise (cf. Gürsel *et al.*, 1983; Meers, 1983; Brillet and Meers, 1987)

$$S_h(f) = \frac{f_k}{2f_o} S_o\left[1+\left(\frac{f}{f_k}\right)^2\right]. \tag{117c}$$

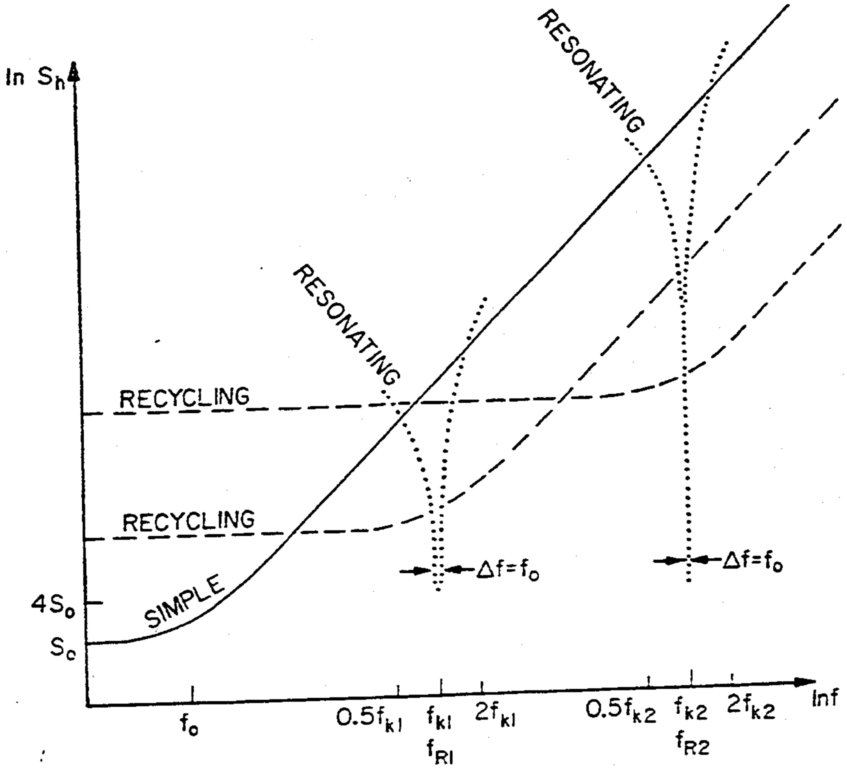
This noise is depicted in Fig. 9.13 for two choices of f_k (dashed curves). Note that this noise has a 'knee' at the frequency f_k; for this reason f_k is called the *knee frequency*. Note further that recycling produces an overall improvement in $S_h(f)$, at frequencies $f \gtrsim f_k$, by a factor $f_o/2f_k$ relative to a non-recycled Fabry–Perot with equal reflectivities for corner and end mirrors (equations (116)) and an improvement by $2f_o/f_k$ relative to the very best non-recycled Fabry–Perot – one with $1-\mathscr{R}_{\max} = 1-\mathscr{R}_E \ll 1-\mathscr{R}_C \ll 8\pi L f_k/c$ (equation (115a)). This improvement at $f \gtrsim f_k$ is bought at the price of worsened noise at $f \lesssim (f_k f_o/2)^{\frac{1}{2}}$. Note further that the improvement factor, over an optimal non-recycled Fabry–Perot at $f \gtrsim f_k$, is

$$\frac{S_h^{\text{recycled}}}{S_h^{\text{non-recycled}}} = \frac{2f_o}{f_k} = \frac{1-\mathscr{R}_E}{1-\mathscr{R}_C}, \tag{118}$$

which is 1/(the mean number of times that the light can be recycled before it is lost by leakage through the high-reflectivity end mirrors). This is just what physical intuition should suggest.

For the recycled Michelson receiver of Fig. 9.9, recycling produces essentially the same $S_h(f)$ as for the Fabry–Perot of Fig. 9.10 at frequencies $f \lesssim f_k$. However, above the knee frequency the recycled Michelson exhibits a

Fig. 9.13. Spectral density of shot noise for Fabry–Perot beam receivers operated in various optical configurations. The vertical and horizontal scales are both logarithmic. The solid curve is for a simple Fabry–Perot with all mirrors – corner and end – having the maximum achievable reflectivity, $\mathscr{R}_C = \mathscr{R}_E = \mathscr{R}_{max}$ (equations (116)). The dashed curves are for a Fabry–Perot with light recycling (equation (117c)), in which the end mirrors have the maximum achievable reflectivity $\mathscr{R}_E = \mathscr{R}_{max}$, the corner mirrors are adjusted to produce the desired knee frequency (f_{k1} and f_{k2} for the two curves shown; equation (117a)), and the recycling mirror is adjusted to minimize the noise at the knee frequency (equation (117b)). The dotted curves are for a Fabry–Perot with light resonating (equation (119c); Fig. 9.14) in which the end mirrors have the maximum achievable reflectivity $\mathscr{R}_E = \mathscr{R}_{max}$; the corner mirrors are adjusted to produce the desired resonant frequency for gravity waves (f_{R1} and f_{R2} for the two curves shown; equation (119a)); and the recycling mirror is adjusted to minimize the noise at the resonant frequency (equation (119b)).

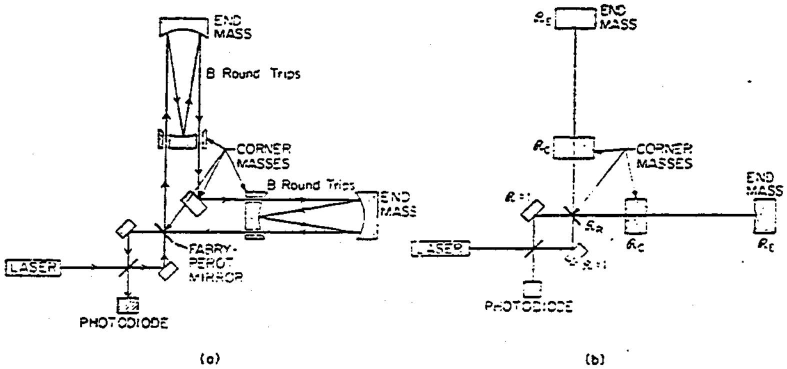


sequence of noise peaks and troughs analogous to those for a non-recycled Michelson (equation (115b) and Fig. 9.12).

Of all optical configurations yet invented, recycling is the best when one is searching for broad-band gravitational waves (waves with $\Delta f \gtrsim f$). However, when one is searching for narrow-band waves (waves with $\Delta f \ll f$, e.g. the periodic waves from a pulsar), recycling gives worse noise than a configuration called *light resonating* (invented by Drever, 1983), which is shown in Fig. 9.14. As usual, the method is conceptually simplest in the Michelson case but is described by the simplest formulas in the Fabry–Perot case.

The basic idea, as shown for a Michelson in Fig. 9.14(a), is to store the light in an arm for one half-cycle of the gravity-wave frequency of interest $f_{\rm R}$ (i.e. for $B = c/4f_{\rm R}L$ round trips); then, instead of extracting it through the beam splitter and reinjecting it, simply move it into the other arm using a high-reflectivity, Fabry–Perot-type mirror – i.e. exchange the light between the arms. Then during the next half-cycle, with arms interchanged and the sign of the gravity wave reversed, each piece of light will experience a phase shift in the same direction as during the first half-cycle. At the end of the second half-cycle, interchange the light in the two arms, and repeat as many times as mirror losses will permit. The arrows in Fig. 9.14(a) show the light, as a result of these exchanges, circulating around the interferometer in a clockwise direction, but there is an equal amount of light circulating in a counterclockwise direction. The light circulating one way experiences a continual increase in phase shift throughout its circulation; the light

Fig. 9.14. Schematic diagrams of optical configurations that could be used for *resonating light* in Michelson-type (a) and Fabry–Perot-type (b) beam receivers (Dreaver, 1983).



irculating oppositely experiences a continual decrease in phase shift; and hese oppositely circulating beams, upon re-emerging through the back of he Fabry–Perot-type mirror and recombining at the beam splitter, produce . much enhanced interference signal in the photodetector.

For the Fabry–Perot receiver of Fig. 9.14(*b*) a single 'resonating mirror', lenoted $\mathscr{R}_R$, produces the resonant interchange of the light between the two ırms: the lengths of the cavities are so adjusted as to produce in each cavity a esonance at a frequency f_e very near the laser frequency; the reflectivities $\mathscr{R}_C$ of the corner mirrors, and the path lengths between them and the esonating mirror, are then so adjusted as to produce two resonant modes of he coupled cavities at light frequencies $f_e \pm f_R/2$ – which requires

$$1-\mathscr{R}_C=4\pi L f_R/c; \qquad (119a)$$

he laser frequency is then adjusted so that it drives the mode at $f_e-f_R/2$; ınd the gravity waves with frequency f_R, by wiggling the end mirrors, then ıpconvert photons into the mode at $f_e+f_R/2$. In order to maximize the esulting gravity-wave-induced signal at the photodetector, the reflectivity $\mathscr{R}_R$ of the resonating mirror must be adjusted to

$$1-\mathscr{R}_R=2\frac{1-\mathscr{R}_E}{1-\mathscr{R}_C}=2\frac{f_o}{f_R}. \qquad (119b)$$

Vith these optimizations the spectral density of shot noise near f_R is (cf. Gürsel *et al.*, 1983; Meers, 1983; Vinet, 1986; Brillet and Meers, 1987)

$$S_h(f)=4S_o\left[1+\left(\frac{f-f_R}{f_o}\right)^2\right] \quad \text{for } |f-f_R| \ll f_o. \qquad (119c)$$

ls is seen in Fig. 9.13 (dotted curves), this represents an enormous mprovement in noise over either a simple Fabry–Perot or a recycling abry–Perot. However, this improvement is bought at the price of an :normous narrow-banding of the receiver response: at frequencies $f \sim f_R/2$ ınd $f \sim 2f_R$ the noise is comparable to that in a simple Fabry–Perot; and at $f \ll f_R/2$ and $f \gg 2f_R$ it is far worse.

For an optimally configured, resonating Michelson the sensitivy near esonance, $|f-f_R| \ll f_o$, is similar to that of a Fabry–Perot (equation (119c)).

Neither recycling nor resonating is potentially useful in present ›rototypes because present mirror reflectivities limit the prototypes to short :torage times. However, in the planned full-scale LIGOs where 100-fold onger arms permit 100-fold greater storage times, recycling and resonating ›oth give promise of substantial improvements in beam-detector :ensitivities. The first attempts to implement recycling on a prototype were carried out by the Munich group in 1986, with moderate success. Resonating has not yet been attempted.

(f) *Quantum limit, quantum non-demolition, and squeezed vacuum*

A beam receiver is similar to a 'Heisenberg microscope', in which a photon is used to measure the position of a particle. The analog of the Heisenberg photon is the laser beam and the analogs of the Heisenberg particle are the end and corner masses. Just as the Heisenberg photon kicks the particle it is trying to measure thereby enforcing the Heisenberg uncertainty principle, $\Delta x\,\Delta p \geqslant \hbar/2$, so fluctuations in the beam's light pressure kick the end and corner masses thereby enforcing a quantum limit on the sensitivity of the beam receiver. As I shall discuss below, these quantum fluctuations have their ultimate source in the vacuum fluctuations of quantum electrodynamics (Caves, 1980, 1981).

Caves (1987*b*) has used Feynman path-integral techniques (Caves, 1986, 1987*a*) to derive a pseudospectral density of displacement noise that characterizes the quantum limit of any harmonic oscillator (such as a beam receiver's swinging masses) when it is studied at frequencies high above the resonant frequency f_0, using techniques that produce the minimum possible noise:

$$S_x(f) = \frac{2\hbar}{m(2\pi f)^2}, \tag{120}$$

where m is the oscillator's mass and x is its displacement from equilibrium. For a beam receiver, whose swinging masses have resonant frequencies $f_0 \sim 1\,\text{Hz}$ far below the region of interest, this pseudospectral density for their displacements can be translated into a pseudospectral density for the gravity-wave field $h(t)$ that one reads off the receiver's output:

$$S_h(f) = \frac{8\hbar}{m(2\pi f)^2 L^2}. \tag{121}$$

Here each of the masses on which a mirror is attached is assumed to have the same mass m. This pseudospectral density can be used, in the same manner as an ordinary classical spectral density (subsection (c) above), to compute the standard quantum limit on the performance of a beam detector in various situations.

As for a bar detector, so also for a beam detector, there should be methods of circumventing this standard quantum limit. Unfortunately nobody has yet found a remotely practical way of doing so, except in one special case: Unruh (1982) has invented a method, which Caves' path-integral formalism says should work (Caves, 1987*b*), for beating the standard quantum limit when performing a narrow-band measurement of periodic gravitational waves. The key idea (which was invented independently and earlier in the classical domain by Gordienko, Gusev and Rudenko, 1977) is to place a spring between each mirror and the companion mass on which it rides, thereby turning the mirror and companion into a two-mode system. By setting the ratio of the spring frequency over the gravity-wave frequency equal to the square root of the mass of the big companion over the mass of the little mirror, one can force the laser beam's fluctuating light pressure to act on the motion of the big mass and not the mirror, while the interferometer reads out the mirror's motion. The result is an improved immunity of the receiver to Heisenberg's quantum noise over a narrow frequency band. While this scheme is clever and conceptually satisfactory, whether it can be implemented in a practical manner is not yet clear.

Although, for broad-band measurements with beam receivers, there as yet is no known way of beating the standard quantum limit, Caves (1981) has invented a clever scheme for getting closer to the quantum limit when inadequate laser power causes the shot noise to exceed it. The key to Caves' idea is his discovery that the ultimate source of the photon shot noise and the fluctuating radiation-pressure noise is *not*, as people previously thought, fluctuations in the laser output. Rather, it is fluctuations in the quantum-electrodynamic vacuum state ('vacuum fluctuations') that enter the beam splitter at right-angles to the incoming laser light and superpose on the laser light as it heads toward the two arms. As Caves has shown, by 'squeezing the vacuum' (i.e. reducing the vacuum fluctuations in the $\cos[(ct-x)/\lambda\!\!\!^{-}_e]$ part of the light while increasing them in the $\sin[(ct-x)/\lambda\!\!\!^{-}_e]$ part; achievable in principle by sending the vacuum through a pumped, non-linear medium), one can reduce the beam receiver's shot noise at the expense of increasing the fluctuations in its light-pressure noise. The net result is the same as if one were using a laser of higher power: in the typical case, where the actual power is too low to permit achieving the quantum limit, one improves the sensitivity and moves toward the quantum limit.

Recently Caves (1987*c*) has elucidated the ultimate sensitivities achievable when one combines his squeezed-vacuum technique with light recycling and resonating. He finds that because, in a resonating system, noise associated with losses in the end mirrors is as important on resonance as vacuum fluctuations entering the beam splitter, squeezing of the vacuum is not useful. By contrast, when combined with light recycling, squeezed-vacuum techniques can reduce $S_h(f)$, over a broad band of frequencies $\Delta f \gtrsim f$, to S_o

or to the quantum limit, whichever is larger:

$$S_h(f) = \max \begin{cases} S_\circ = \dfrac{2\hbar c \lambdabar_e}{I_\circ \eta}\left(\dfrac{1-\mathscr{R}_E}{2L}\right)^2 \\ \dfrac{8\hbar}{m(2\pi f)^2 L^2} \end{cases} \tag{122}$$

Squeezed states of light have been produced by experimenters recently (Slusher *et al.*, 1985; Wu *et al.*, 1986), triggering great excitement in the field of non-linear quantum optics. This achievement has triggered an effort by the Munich group to implement Caves' squeezed-vacuum technique – an effort that is likely to succeed within the next few years.

(g) *Anticipated sensitivities of full-scale LIGO receivers*

The present Munich burst sensitivity $h_{3/\mathrm{yr}} \cong 4 \times 10^{-17}$ at $f \sim 1000$ Hz, when scaled from the prototype arm length of 30 m to the planned Munich LIGO arm length of 3 km, would become $h_{3/\mathrm{yr}} \cong 4 \times 10^{-19}$. Although scaling up is not straightforward (mirror diameters must be increased a factor ten, for example), the experimenters are quite confident of doing somewhat better than this scaled-up number by the early 1990s when the first LIGOs start operating.

The upper thick, dashed curves in Figs. 9.4, 9.6 and 9.7 (earlier in this chapter) show possible burst, periodic and stochastic sensitivities (based on a $\frac{1}{3}$-year search time and 90% confidence levels), for a first generation of detectors in the full-scale LIGOs. These curves correspond, at frequencies above ~500 Hz, to receivers that are shot-noise-limited, without recycling or resonating, with Argon-ion laser light $\lambdabar_e = 0.0818\ \mu$m, with (laser power) × (photodetector efficiency) $= I_\circ \eta = 10$ W, and with light storage times of a half-gravity-wave period or longer, so [cf. equation (115) and Fig. 9.12]

$$[S_h(f)]^{\frac{1}{2}} = \left[\frac{2\hbar c \lambdabar_e}{I_\circ \eta}\left(\frac{\pi f}{c}\right)^2\right]^{\frac{1}{2}} = (2.4 \times 10^{-22}\ \mathrm{Hz}^{-\frac{1}{2}})\left(\frac{f}{1000\ \mathrm{Hz}}\right); \tag{123a}$$

and, correspondingly (equations (111)–(113))

$$h_{3/\mathrm{yr}} = 11(f_c S_h)^{\frac{1}{2}} = 8 \times 10^{-20}\left(\frac{f_c}{1000\ \mathrm{Hz}}\right)^{\frac{3}{2}} \quad \text{for bursts,} \tag{123b}$$

$$h_{3/\mathrm{yr}} = 3.8(S_h \times 10^{-7}\ \mathrm{Hz})^{\frac{1}{2}} = 2.9 \times 10^{-25}\left(\frac{f}{1000\ \mathrm{Hz}}\right) \quad \text{for periodic sources,} \tag{123c}$$

$$h_{3/\mathrm{yr}}=4.5\left(\frac{f}{10^{-7}\,\mathrm{Hz}}\right)^{-\frac{1}{4}}(fS_h)^{\frac{1}{2}} \qquad (123c)$$

$$=1.1\times10^{-22}\left(\frac{f}{1000\,\mathrm{Hz}}\right)^{\frac{5}{4}} \quad \text{for stochastic sources,} \qquad (123c)$$

where the search for stochastic waves is presumed to use a bandwidth $\Delta f=f$. Below ~ 500 Hz it is presumed that seismic noise debilitates the performance of these first-generation LIGO receivers.

During the years following the first gravity-wave searches in the LIGOs, the experimenters plan to run a sequence of detectors with ever improving sensitivities, pushing the sensitivities ever downward, and improving the seismic isolation so the detectors can operate at ever decreasing minimum frequencies. A reasonable goal by the end of the 1990s is to approach the lower-most thick-dashed curves in Figs. 9.4, 9.6 and 9.7. These curves correspond to *advanced detectors* with the following characteristics:

$L=$(arm length)$=4$ km,
$\lambda\!\!\!^{-}_{e}=0.0818\ \mu$m (Argon-ion laser light),
$I_o\eta=$(laser power) × (photodetector sensitivity)$=100$ W,
$\mathscr{R}_E=\mathscr{R}_{\max}=$(maximum mirror reflectivity)$=0.9999$,
$m=$(mirror mass)$=1000$ kg,
photon shot noise dominant at $100\ \mathrm{Hz}\lesssim f\lesssim 10^4$ Hz,
quantum-limit noise dominant at $10\ \mathrm{Hz}\lesssim f\lesssim 100$ Hz,
seismic noise dominant at $f\lesssim 10$ Hz. (124)

It is presumed that light recycling is used for burst searches, with the knee frequency adjusted to equal the frequency of interest, so the photon shot noise (equations (111), (117c) and (116b,c) with $f_k=f_c$) is

$$h_{3/\mathrm{yr}}=11(f_cS_h)^{\frac{1}{2}}=11(S_of_c^2/f_o)^{\frac{1}{2}}=11\left[\frac{2\pi\hbar\lambda\!\!\!^{-}_{e}}{I_o\eta}\frac{(1-\mathscr{R}_E)}{L}f_c^2\right]^{\frac{1}{2}}$$

$$=1.3\times10^{-21}\left(\frac{f_c}{1\,\mathrm{kHz}}\right) \quad \text{for bursts;} \qquad (125a)$$

and it is presumed that light resonating is used for periodic and stochastic searches, with the resonant frequency adjusted to equal the frequency of interest, so the photon shot noise (equations (112), (113), (119c) and (116b,c) with $f_R=f$) is

$$h_{3/\mathrm{yr}}=3.8(4S_o\times10^{-7}\,\mathrm{Hz})^{\frac{1}{2}}=3.8\left[\frac{2\hbar c\lambda\!\!\!^{-}_{e}}{I_o\eta}\times10^{-7}\,\mathrm{Hz}\right]^{\frac{1}{2}}\left(\frac{1-\mathscr{R}_E}{L}\right)$$

$$=2.1\times10^{-28} \quad \text{for periodic waves,} \qquad (125b)$$

$$h_{3/\mathrm{yr}}=4.5\left(\frac{f_{\mathrm{o}}}{10^{-7}\,\mathrm{Hz}}\right)^{-\frac{1}{4}}(4S_{\mathrm{o}}f)^{\frac{1}{2}}=4.5\left[\frac{2\hbar c\lambda\!\!\!^{-}_{\mathrm{e}}}{I_{\mathrm{o}}\eta}f\right]^{\frac{1}{2}}\left[\frac{4\pi}{c}\left(\frac{1-\mathscr{R}_{\mathrm{E}}}{L}\right)^{3}\times 10^{-7}\,\mathrm{Hz}\right]^{\frac{1}{4}}$$

$$=5.2\times 10^{-25}\left(\frac{f}{1\,\mathrm{kHz}}\right)^{\frac{1}{2}}\quad\text{for stochastic waves.}\qquad(125\mathrm{c})$$

Here it is assumed that the stochastic search is restricted to the narrow bandwidth $\Delta f=f_{\mathrm{o}}$ over which the resonating receiver has good performance. The quantum-limit noise, which exceeds these shot noise levels at $f\lesssim 100$ Hz, is given by (equations (121) and (111)–(113))

$$h_{3/\mathrm{yr}}=11\left[\frac{2}{\pi^2}\frac{\hbar}{mL^2}\frac{1}{f}\right]^{\frac{1}{2}}=1.3\times 10^{-23}\left(\frac{1000\,\mathrm{Hz}}{f}\right)^{\frac{1}{2}}\quad\text{for bursts,}\qquad(126\mathrm{a})$$

$$h_{3/\mathrm{yr}}=3.8\left[\frac{2}{\pi^2}\frac{\hbar}{mL^2}\frac{10^{-7}\,\mathrm{Hz}}{f^2}\right]^{\frac{1}{2}}$$

$$=4.4\times 10^{-29}\left(\frac{1000\,\mathrm{Hz}}{f}\right)\quad\text{for periodic waves,}\qquad(126\mathrm{b})$$

$$h_{3/\mathrm{yr}}=4.5\left[\frac{f_{\mathrm{o}}(S_{\mathrm{QL}}/S_{\mathrm{o}})^{\frac{1}{2}}}{10^{-7}\,\mathrm{Hz}}\right]^{-\frac{1}{4}}\left[\frac{4}{3}fS_{\mathrm{QL}}\right]^{\frac{1}{2}}$$

$$=4.5\left(\frac{4}{3\pi}\right)^{\frac{1}{2}}\left(\frac{4\hbar}{mL^2}\right)^{\frac{3}{8}}\left[\frac{\hbar\lambda\!\!\!^{-}_{\mathrm{e}}}{I_{\mathrm{o}}\eta c}\left(\frac{10^{-7}\,\mathrm{Hz}}{f}\right)^{2}\right]^{\frac{1}{8}}$$

$$=1.5\times 10^{-25}\left(\frac{1000\,\mathrm{Hz}}{f}\right)^{\frac{1}{4}}\quad\text{for stochastic waves.}\qquad(126\mathrm{c})$$

Here for stochastic waves S_{QL} is the quantum-limit spectral density (equation (121)), and the chosen bandwidth $\Delta f=f_{\mathrm{o}}(S_{\mathrm{QL}}/S_{\mathrm{o}})^{\frac{1}{2}}$ is that which makes shot noise (equation (119c) integrated over Δf, with $f_{\mathrm{o}}<\Delta f<f_{\mathrm{R}}$) equal to $\frac{1}{3}$ the quantum noise and minimizes their sum. At frequencies $f\lesssim 10$ Hz, seismic noise is presumed to strongly debilitate the receiver performances (cf. Saulson, 1984).

Figs. 9.4, 9.6 and 9.7 show, together with the above receiver sensitivities, the characteristic strengths of the gravity waves from a variety of sources that were discussed in Section 9.4 above, and the present and projected sensitivities of other kinds of detectors. As these diagrams suggest, the prospects for successful detection of gravity waves with the planned LIGOs are high: at the sensitivities of the 'advanced detectors' one could see the coalescence of neutron-star binaries at $\frac{1}{2}$ the Hubble distance where the event rate is estimated to be several per day, the coalescence of $10M_{\odot}$ black-hole binaries throughout the universe, supernovae in our galaxy if they put out $10^{-9}M_{\odot}$ of energy near 1000 Hz frequency and in the Virgo cluster if they put out $10^{-3}M_{\odot}$, millisecond pulsars whose frequencies are known in advance from electromagnetic measurements if they have ellipticities $10^{-8.5}$ or larger, CFS-unstable neutron stars with X-ray luminosities as small as $\frac{1}{1000}$ that of Sco X-1, and a stochastic background at $f \sim 30$ Hz with Ω_{GW} as small as 10^{-10}. It seems likely to me that such sensitivities are more than adequate for success; waves are likely to be detected at more modest sensitivities than these – somewhere between those of the 'first-generation' detectors and those of the 'advanced detectors' (Figs. 9.4, 9.6 and 9.7).

(h) *Ideas for other types of beam detectors*

The Michelson and Fabry–Perot configurations, on which almost all experimental work to date has focussed, are not the only possible types of beam receivers. A number of others have been conceived of, but have not been pursued vigorously. Since the LIGOs are intended to house several different receivers simultaneously and to have lifetimes of $\gtrsim 20$ years, during which a number of generations of receivers will be built and operated, some of these other configurations might one day operate in a LIGO. These other configurations include: (i) A *frequency-tagged interferometer* or *Michelson with overlapping beams* (Drever and Weiss, 1983), in which the light beam in each arm is made to shift in frequency with each round-trip pass, so that the beams of successive passes can overlap each other but not interfere with each other; such an interferometer is basically a Michelson but with small, Fabry–Perot-size mirrors. (ii) *Active interferometers* (Bagaev *et al.*, 1981; Brillet and Tourrenc, 1983), in which an active medium (atoms or molecules with transitions near the frequency of the laser light) resides in the arms (or arm – there might be only one) of the interferometer. (iii) *Spectroscopic detectors* (Nesterikhin, Rautian and Smirnov, 1978; Brillet and Tourrenc, 1983; Borde *et al.*, 1983), in which laser spectroscopy is used to monitor the frequency shifts produced by the gravitational waves.

Although in principle these alternative types of beam detectors can achieve sensitivities comparable to those of the standard (Michelson and Fabry–Perot) types, practical issues make the active and spectroscopic detectors look substantially less promising (Brillet and Tourrenc, 1983); and the frequency-tagged detector has not yet been pursued in sufficient depth to venture a tentative verdict.

9.5.4 *Other types of earth-laboratory detectors*

In addition to bar detectors and beam detectors, a number of other types of gravity-wave detectors that could operate in an earth-bound laboratory have been conceived of. Although none of these has looked sufficiently promising to justify substantial experimental effort, some are question marks (because they have not been pursued far enough for a clear verdict) rather than discards. Because I have not looked at most of them in enough detail, I shall not venture to sort out the question marks from the total rejects.

(a) *Electromagnetically coupled detectors*

One type of transducer used on bar detectors (e.g. by the Moscow and Perth groups) is a re-entrant microwave-cavity resonator whose capacitance is modulated by the bar's vibrations (Section 10 of Braginsky, Mitrofanov and Panov, 1985; Blair, 1983). There is an obvious similarity of this to a Fabry–Perot beam receiver in which each arm is an optical cavity with length modulated by the motions of the swinging masses. In fact, one can imagine a continuous sequence of detector configurations that leads from one to the other, by splitting the bar into pieces and gradually moving them apart, with the microwave cavity gradually being distorted and expanded until it becomes the optical cavity of the Fabry–Perot (Caves, 1978).

This argument leads to the recognition that bar detectors and beam detectors are but two examples of a large class of possible 'electromagnetically coupled detectors', in which gravitational waves drive the motions of masses, and electromagnetic fields measure those motions; or, when the detector gets larger than a reduced wavelength, gravity waves drive vibrations of both the electromagnetic fields and the masses, which are coupled together.

Other earth-laboratory-scale electromagnetically coupled configurations, besides resonant bars and optical beams, that have been considered theoretically and show some promise but have not been pursued in a serious experimental way, include: (i) large microwave cavities in which wall motion, driven by gravitational waves with $\lambdabar \gg L$, upconverts microwave quanta from one mode to another mode of slightly higher frequency (Pegoraro, Picasso and Radicati, 1978; Caves, 1979); (ii) optical or microwave cavities with $L \lesssim \lambdabar$, intended for detection of high-frequency waves $f \gg 10^4$ Hz, in which the gravitational waves interact directly with the resonating electromagnetic field to move quanta from one mode into another, or interact directly with a DC electric or magnetic field to create quanta at the gravity-wave frequency (Braginsky *et al.*, 1973); (iii) as a specific, much studied example of such a cavity: an optical or microwave ring resonator in which circularly polarized gravitational waves propagating orthogonal to the resonator plane resonate with the circulating electromagnetic field, producing a linearly growing phase shift in the field. This scheme, proposed by Braginsky and Menskii (1971), was incorrectly analyzed by them and by me (Box 37.6 of MTW) – a source of some embarrassment. My error was in thinking I could cover the detector with a single proper reference frame, when its diameter L is larger than the reduced wavelength $\lambdabar$ of the gravitational waves it sees. My incorrect conclusion, and that of Braginsky and Menskii (1971) was a quadratically growing phase shift, $\Delta\Phi \propto h(ct/\lambdabar)(ct/\lambdabar_e)$. The correct conclusion (Linet and Tourrenc, 1976) is a linearly growing phase shift, $\Delta\Phi \propto h(ct/\lambdabar_e)$, which makes this scheme no better in principle than a standard beam detector. WARNING: The literature is full of similarly incorrect analyses of the various detectors described in this section, 9.5.4; (iv) an optical cavity filled with an isotropic medium, or an optical fiber, in which gravity-wave-produced strains induce optical effects such as birefringence (Iacopini *et al.*, 1979; Vinet, 1985); (v) detectors using the Mossbauer effect (Kauffman, 1970).

For a general review of electromagnetically coupled detectors, see Grishchuk (1983), and for general analyses of them, see Tourrenc and Grossiord (1974), Tourrenc (1978) and Teissier du Cros (1985).

Analyses of such detectors sometimes produce wildly overoptimistic conclusions because they overlook the mundane issue of thermal noise in the mechanical parts of the detector (e.g. the walls of an electromagnetic cavity). For an example of an analysis that *does* take thermal noise into account properly, see Caves (1979).

(b) *Superfluid interferometers and superconducting circuits*

Anandan (1981), Chiao (1982) and Anandan and Chiao (1982) have suggested that *if* it is possible to construct a superfluid weak link, analogous to the superconducting weak link (Josephson junction) on which SQUIDS are based, then such a link could form the basis for a superfluid ring interferometer that would be sensitive to gravitational waves. Unfortunately, such weak links do not yet exist. Schrader (1984) and independently Anandan (1985, 1986) have suggested a gravity-wave detector based on the direct interaction between the gravitational wave and the magnetic field in superconducting solenoids, with the resulting current change monitored by a SQUID.

9.5.5 Low-frequency detectors (10–10^{-5} Hz)

As one goes to lower and lower frequencies it becomes harder and harder to isolate a gravity-wave detector in an earth-bound laboratory from seismic and acoustic vibrations and from fluctuating gravity gradients due to people, animals, trucks, etc. Ultimately, somewhere around 10 Hz, isolation will become impossibly difficult (see, e.g., Saulson, 1984). Thus, to operate in the 'low-frequency region' $10\,\mathrm{Hz} \gg f \gtrsim 10^{-5}\,\mathrm{Hz}$ will require putting detectors in space, or using normal modes of the earth or the sun as the detectors.

A number of space-borne or earth- or sun-normal-mode detectors have been conceived of for operation at low frequencies; and several have been constructed and have produced interesting limits on cosmic gravitational waves. In this section I shall describe these briefly.

(a) *Doppler tracking of spacecraft*

At periods between a few minutes and a few hours the best present gravity wave detector is the doppler tracking of spacecraft.

In doppler tracking a highly stable clock on earth ('master oscillator') is used to control the frequency of a monochromatic radio wave, which is transmitted from earth to the spacecraft. This 'uplink' radio wave is received by the spacecraft and 'transponded' back to the earth; i.e. the uplink wave is used by the spacecraft to control, in a phase-coherent way, the frequency of the 'downlink' radio wave that it transmits back to earth. When the down link radio wave is received at earth, its frequency is compared with that of the master oscillator; and from that comparison a doppler shift is read out.

When gravitational radiation sweeps through the solar system – most interestingly with a reduced wavelength of order the earth-spacecraft distance so it must be analyzed by TT methods – it perturbs the earth, the spacecraft, and the propagating radio wave. The net result is to produce fluctuations in the measured doppler shift with magnitude $\delta v/v \sim h^{\mathrm{TT}}_{jk}$. Each feature in the gravitational waveform $h^{\mathrm{TT}}_{jk}(t)$ shows up three times in the doppler shift, in a manner that can be regarded as due to interaction of the radio wave with the gravity wave at the events of emission from earth, transponding by spacecraft, and reception at earth (Estabrook and Wahlquist, 1975). This triplet structure can be used to help distinguish the effects of the gravitational wave from noise in the doppler data.

The use of doppler data for gravity-wave searches was first proposed by Braginsky and Gertsenshtein (1967), and was first pursued with preexisting data by Anderson (1971). The response of doppler tracking to a gravity wave was worked out by Davies (1974) for special cases and by Estabrook and Wahlquist (1975) for the general case. Experimental feasibility and noise thresholds were first discussed by Wahlquist *et al.* (1977). Experimental results have been obtained with the Viking spacecraft (Armstrong *et al.*, 1979), the Voyager spacecraft (Hellings *et al.*, 1981), Pioneer 10 (Anderson *et al.*, 1984), and Pioneer 11 (Armstrong *et al.*, 1987). A future experiment will entail coordinated observations with the Galileo and Ulysses spacecraft (Estabrook, 1987).

In all past experiments the most serious noise source was fluctuations in the interplanetary plasma (solar wind), which cause fluctuations in the dispersion of the radio waves and thence fluctuations in the doppler shift. These fluctuations have limited the sensitivity under optimal circumstances to $[fS_h(f)]^{\frac{1}{2}} \sim 3 \times 10^{-14}$ in the band $10^{-4}\,\mathrm{Hz} \lesssim f \lesssim 10^{-2}\,\mathrm{Hz}$. Fortunately, dispersion in an ionized plasma becomes less serious when one moves to higher radio-wave frequencies – a move that is also desired to produce higher bit rates for information transfer between the spacecraft and earth. Some future spacecraft, including the Galileo mission, will use X-band radio signals (10 GHz frequency) on both the uplink and the downlink, rather than S-band (2 GHz) up and X-band down as in the past; and correspondingly their gravity-wave sensitivities should improve to $[fS_h(f)]^{\frac{1}{2}} \sim 3 \times 10^{-15}$. In fact, on such spacecraft the ionized-plasma dispersion will be sufficiently low that fluctuations in dispersion due to water vapor in the earth's atmosphere may become the most serious noise source (Armstrong and Sramek, 1982). Plans for monitoring the water vapor as a means of reducing this noise are being developed (Resch *et al.*, 1984). In the more distant future, doppler tracking may be improved by using dual-frequency X-band/K-band (30 GHz) tracking to reduce and monitor the ionized plasma dispersion, by installing a highly stable clock on board the spacecraft (Smarr *et al.*, 1983) and/or by moving the earth-based antenna into earth orbit. These improvements might ultimately produce $[fS_h(f)]^{\frac{1}{2}}$ as low as 10^{-17} (Estabrook, 1987).

Figs. 9.4, 9.6 and 9.7 show the sensitivities $h_{3\,\mathrm{yr}}$ corresponding to these noise levels (equations (34), (52a) and (67) with $\Delta f = f$).

At frequencies $f \sim 3$ Hz, far above the regime 10^{-2}–10^{-4} Hz of normal doppler tracking, but below the regime of earth-based detectors, one might hope to use doppler tracking of spacecraft in earth orbit (distances $\sim c/4f \sim 25\,000$ km). The best such spacecraft are those of the Global Positioning System (GPS). Unfortunately, even with the best projected improvements of the GPS we cannot anticipate $[fS_h(f)]^{\frac{1}{2}}$ better than $\sim 10^{-14}$ corresponding to a burst sensitivity $h_{3\,\mathrm{yr}} \sim 10^{-13}$; see Fig. 9.4 and see Hansen, Chiu and Chao (1986) for a detailed study.

(b) *Beam detectors in space*

Several conceptual designs for optical beam detectors in space were suggested by Weiss *et al.* (1976) and are discussed in Weiss (1979). More recently Faller and Bender (1981), Bender *et al.* (1984) and Faller *et al.* (1985) have been carrying out a detailed design and feasibility study for such detectors; and it is possible they will fly sometime around the turn of the century. As now conceived (Faller *et al.*, 1985), such a detector might consist of three 'drag-free' satellites, one at the corner and two at the ends of an L. The satellites might be in the same solar orbit as the earth–moon system, but far from the earth and moon; and their arm length might be $L \sim 1$ million km. Each satellite might carry its own 100 mW laser, with the end lasers phase locked to the light that arrives from the corner laser. With 50 cm diameter telescopes on each satellite to focus the laser beams, this could result in a one-pass ($B=1$) Michelson interferometer system with shot-noise-limited sensitivity $[S_h(f)]^{\frac{1}{2}} = 10^{-20}\,\mathrm{Hz}^{-\frac{1}{2}}$ over the range $10^{-1}\,\mathrm{Hz} \gtrsim f \gtrsim 10^{-4}\,\mathrm{Hz}$. Above $10^{-1}\,\mathrm{Hz}$ the sensitivity would be degraded because of too-long an arm length ($L > \lambda\!\!\!^{-}$). Below 10^{-4} Hz stochastic non-gravitational forces might degrade the sensitivity. This projected sensitivity translates into the amplitude levels $h_{3/\mathrm{yr}}$ (equations (34), (52a) and (67) with $\Delta f = f$) shown in Figs. 9.4, 9.6 and 9.7.

(c) *Earth's normal modes*

Forward *et al.* (1961) and Weber (1967) pioneered the use of the earth's quadrupolar normal modes as 'resonant-bar' gravity-wave detectors. More recently Boughn and Kuhn (1984) have given a careful formulation of the method for inferring, from earth-normal-mode observations, limits on any stochastic gravitational waves that might be exciting the earth. Using their method on data from the IDA seismic network, Boughn and Kuhn (1987) have derived a slightly stronger limit on stochastic background than the early Weber (1967) limit: they find $\Omega_{\mathrm{GW}}(f) \lesssim 1$ at $f = 0.31$ mHz (Fig. 9.7). They expect, further, that a larger amount of data may permit this limit to be lowered to $\Omega_{\mathrm{GW}} \lesssim 0.1$.

(d) *Sun's normal modes*

Walgate (1983) stimulated people to think about the sun's quadrupolar normal modes as gravity-wave detectors, when he speculated (falsely; see, e.g., Kuhn and Bough, 1984, and Anderson *et al.*, 1984) that an observed mode was being excited by monochromatic gravitational waves from the binary star Geminga.

Boughn and Kuhn (1984) have carried out a detailed analysis of the gravity-wave implications of observed solar pulsations. From the data of Isaac (1981) on low-order p- and g-modes they find $\Omega_{GW}(f) \lesssim 100$ at $f \simeq 3 \times 10^{-4}$ Hz; and they argue that better observational data may produce orders of magnitude improvement, bringing $\Omega_{GW}(f)$ well below unity.

(e) *Vibrations of blocks of the earth's crust*

Braginsky *et al.* (1985) have recently pointed out that blocks of the earth's crust with sizes 50 km $\lesssim L \lesssim$ 70 km could function as resonant-bar detectors for waves of frequency $f \simeq 0.03$ Hz. They propose monitoring such a block with an array of seismic stations and cross-correlating the data to pull out the seismic motions associated with the block's quadrupolar modes. Their calculations suggest a possible sensitivity $[fS_h(f)]^{\frac{1}{2}} \sim 2 \times 10^{-17}$ corresponding to $h_{3/\mathrm{yr}} \sim 2 \times 10^{-16}$ for bursts.

(f) *Skyhook*

Braginsky and Thorne (1985) have suggested an earth-orbiting 'skyhook' gravity-wave detector that would operate in the 0.1–0.01 Hz region with sensitivity $[fS_h(f)]^{\frac{1}{2}} \sim 3 \times 10^{-17}$, which is much better than present or near-future doppler tracking and roughly comparable to that hoped for from blocks of the earth's crust (see Figs. 9.4, 9.6 and 9.7). The skyhook would consist of two masses, one on each end of a long thin cable with a spring at its center. As it orbits the earth, the cable would be stretched radially by the earth's tidal gravitational field. Gravitational waves would pull the masses apart and push them together in an oscillatory fashion; their motion would be transmitted to the spring by the cable; and a sensor would monitor the spring's resulting motion.

If it ever flies, the skyhook's role will be to provide, with a simple and inexpensive device, a moderate-sensitivity coverage of the 0.1–0.01 Hz region during the epoch before far more sensitive beam detectors are built and installed in space.

9.5.6 *Very-low-frequency detectors (frequencies below 10^{-5} Hz)*

At frequencies below about 10^{-5} Hz the only sources of gravitational waves are probably stochastic background from the early universe (Sections 9.4.3(d,e,f)); and the best detectors involve use of distant astronomical bodies.

(a) *Pulsar timing*

Remarkably good limits on very-low-frequency gravitational waves have come from the timing of pulsars (Taylor, 1987). The basic idea for pulsar timing as a gravity-wave detector (Sazhin, 1978; Detweiler, 1979) is this: the rotation of the pulsar's underlying neutron star is a highly stable clock, and pulsar timing compares that clock's ticking rate with the ticking of the very best atomic clocks in earth-bound laboratories. When gravity waves sweep over the pulsar, they affect the ticking rate of the pulsar's clock relative to clocks elsewhere in the universe, including earth; and those effects show up as fluctuations in the pulsar timing data. Similarly, when gravity waves sweep over the earth, they affect the ticking rates of all our clocks relative to those elsewhere in the universe; and those effects show up in pulsar timing. To the extent, then, that the pulsar timing data are free of fluctuations, one can infer that gravity waves of a given strength are not sweeping over either the pulsar or the earth. (See Blandford, Narayan and Romani, 1984, for a discussion of how to extract the gravity-wave information from the timing data and the pitfalls one encounters in doing so.) If fluctuations are seen in the timing data and are actually due to gravity waves sweeping over the pulsar, we probably will never be able to confirm that the source is gravity waves rather than fluctuations in the neutron star's rotation. However, if fluctuations are seen and are due to gravity waves sweeping over the earth, we might be able to learn their cause and study the waves' direction of propagation, polarization and wave form by simultaneously timing several pulsars on different parts of the sky.

In response to the Sazhin–Detweiler idea, Hellings and Downs (1983) and Romani and Taylor (1983) used timing data on four especially quiet pulsars (which Downs and Reichley (1983) had tracked for over a decade with the Goldstone antenna of NASA's deep space tracking network) to place a limit $\Omega_{GW}(f) \lesssim 1 \times 10^{-3}(f/10^{-8}\ \mathrm{Hz})^{4}$ (90% confidence) on any stochastic background in the frequency range $4 \times 10^{-9} \lesssim f \lesssim 10^{-7}$ Hz; Bertotti, Carr and Rees (1983) used data from Helfand *et al.* (1980) to obtain a similar limit.

Intrinsic noise in all the pulsars used in these analyses would have made it difficult to improve on these limits at fixed frequency (though, by further observations the region covered could have been extended to lower frequencies; the lowest being of order 1/(total time since the measurements began)). Fortunately, while these analyses were in process, they were put out of business by the discovery (Backer *et al.*, 1982) of a pulsar far quieter than any previously known: the millisecond pulsar PSR 1937+21. From three years of 1937+21 timing data taken with the Arecibo radio telescope, Davis *et al.* (1985) and Taylor (1987) have now placed the limit

$$\Omega_{\mathrm{GW}}(f) \leqslant 1\times 10^{-6}\left(\frac{f}{10^{-8}\,\mathrm{Hz}}\right)^{4} \quad \text{for } f \gtrsim f_{\min} = 10^{-8}\,\mathrm{Hz} \tag{127}$$

on any isotropic stochastic background: see Fig. 9.7.

This level of sensitivity is so good that further progress at fixed frequency is limited by the long-term frequency stability of the world's best atomic clocks. Thus, unless clocks improve, we can expect the coefficient in (127) to improve at best as t^{-1}, while the lower frequency limit $f_{\min}$ decreases as t^{-1} producing $\Omega_{\mathrm{GW}}(f_{\min}) \propto t^{-5}$ (with $t=0$ in 1982). Ultimately, when clocks have improved by one or two orders of magnitude, noise due to interstellar scintillation may become a problem (Armstrong, 1984).

Recently two other quiet, fast pulsars PSR 1855+09 and PSR 1953+29 have been discovered (Segelstein *et al.*, 1986). Together with PSR1937+21 and others that we can hope for, they may one day form a network for gravitational-wave searches and observations. Such a network would alleviate problems with interstellar scintillations and atomic clock fluctuations.

(b) *The timing of orbital motions*

A gravitational wave with period long compared to observation times will produce a gradual secular change in the relative ticking rates of clocks in and out of the wave: $\dot{\omega}/\omega \sim (f/2\pi)h$, where f is the gravity-wave frequency, h is its amplitude, and ω is the ticking rate of the clock in the wave as monitored by any clock that is outside of the wave. For stochastic waves this gives $\dot{\omega}/\omega \sim [\Omega_{\mathrm{GW}}(f)]^{\frac{1}{2}}(10^{10}\,\mathrm{yr})^{-1}$. Because the physical torques on pulsars are so large ($\dot{\omega}/\omega \gg (10^{10}\,\mathrm{yr})^{-1}$), pulsars cannot be useful in searches for waves with periods longer than the observation time. However (Mashhoon, Carr and Hu, 1981; Bertotti, Carr and Rees, 1983), the orbit of the binary pulsar is such a good clock, with such a well-understood slow down, that it can be beat against earth-based clocks to give interesting limits on waves with frequencies $10^{-8}\,\mathrm{Hz} \lesssim f \lesssim 10^{-13}\,\mathrm{Hz}$. (The low-frequency limit, 10^{-13} Hz, corresponds to waves with $\bar{\lambda}$ of order the distance between the earth and the pulsar.) The most recent observational data from Weisberg and Taylor (1984) give the limit

$$\int_{10^{-5}\,\mathrm{Hz}}^{10^{-13}\,\mathrm{Hz}} \Omega_{\mathrm{GW}}(f) f^{-1}\,\mathrm{d}f \leqslant 0.5 H_{100}^{-2}, \tag{128}$$

where H_{100} is the Hubble constant in units of 100 km/s Mpc^{-1}.

Orbital motions are also useful as detectors of waves with frequencies f of order the orbital frequency. The orbiting bodies respond to such waves in the same manner as does a resonant-bar detector. (After all, the orbit is, in a sense, nothing but a two-dimensional oscillator.) The idea of using a lunar or planetary or binary-star orbit in this way as a gravity-wave detector has been proposed or discussed by Braginsky and Gertsenshtein (1967), Anderson (1971), Bertotti (1973), Rudenko (1975) and many others. Mashhoon, Carr and Hu (1981) argue that Viking doppler data on the relative orbits of Mars and Earth place a limit $\Omega_{GW}(f) \lesssim 0.1$ at $f \simeq 3 \times 10^{-8}$ Hz (period of one year), and that laser ranging data on the moon's orbit – which now give $\Omega_{GW}(f) \lesssim 10$ – have the potential in some years to give $\Omega_{GW}(f) \lesssim 0.1$ at $f \simeq 3 \times 10^{-7}$ Hz (period of one month).

(c) *Anisotropies in the temperature of the cosmic microwave radiation*

Large-scale (quadrupolar) anisotropies in the cosmic microwave radiation would be produced by gravitational waves in the vicinity of the earth today (Sachs and Wolfe, 1967; Dautcourt, 1969). Present observational limits on such anisotropies imply $\Omega_{GW}(f) \lesssim 10^{-6}(f/3 \times 10^{-18}\text{ Hz})^2$ for any f, even $f \lesssim 1/(\text{Hubble time})$ (Grishchuk and Zel'dovich, 1978). Small-scale anisotropies (angular scales of order 5 degrees) would have been produced by gravitational waves during the epoch of plasma recombination, when the microwave radiation we now see last interacted with matter. The limit from these is remarkably good: $\Omega_{GW}(f) \lesssim 10^{-13}$ in a narrow frequency window at $f \simeq 10^{-16}$ Hz (Carr, 1980; Zel'dovich and Novikov, 1983; Starobinsky, 1985); but this limit is much less firmly based than others. It presumes we understand thoroughly the propagation of the cosmic microwave radiation during recombination and from the epoch of recombination to the present. Note: this limit does not constrain waves from cosmic strings or any other source that emitted most of its waves later than the epoch of recombination; see Traschen, Turok and Brandenberger (1986) for the effects of strings on the microwave anisotropy.

(d) *Other astronomical observations*

In addition to the timing of pulsars and of orbital motions, and microwave anisotropies, several other astronomical observations can be used as probes of low-frequency gravitational waves: (i) *the observed velocities of galaxies and clusters of galaxies*, i.e. deviations from the Hubble flow (Rees, 1971; Burke, 1975; Dautcourt, 1977). These give $\Omega_{GW}(f) \lesssim (f/3 \times 10^{-17})^2$ for $10^{-17}\text{ Hz} \lesssim f \lesssim 10^{-15}$ Hz (Carr, 1980). (ii) *Peculiarities in primordial nucleosynthesis*, produced by the influence of gravitational wave energy on the expansion rate of the universe during nucleosynthesis (the 'first three minutes') (Schvartzman, 1969). These produce a limit $\Omega_{GW}(f) \lesssim 10^{-4}$ at $f \gtrsim 10^{-10}$ Hz, which is not fully firmly based because it presumes we thoroughly understand conditions in the first few minutes of the universe. Note that this limit is irrelevant for waves emitted since the epoch of nucleosynthesis.

For detailed reviews and references on these and other 'astronomical detectors' of gravity waves, see Dautcourt (1974), Chapter 17 of Zel'dovich and Novikov (1983), and Section 4 of Carr (1980).

9.6 Conclusion

As I look back over this review of gravitational waves, I am struck by the enormous changes in our theoretical understanding – or at least in theoretical fashion – that have occurred over the past 5, 10 and 15 years; and I am impressed even more by the progress that experimenters have made in the quest to invent, design and build detectors of ever greater sensitivity. That the quest ultimately will succeed seems almost assured. The only question is when, and with how much further effort. Five years ago Jerry Ostriker and I made a bet, to wit:

> *Whereas both Jeremiah P. Ostriker and Kip S. Thorne believe that Einstein's equations are valid*
>
> *And both are convinced that these equations predict the existence of gravitational waves*
>
> *And both are confident that Nature will provide what physical law predicts*
>
> *And both have faith that scientists can ultimately observe whatever Nature does supply*
>
> *Nevertheless, they differ on the likely strengths of natural sources and on the probability of a near-future and verifiable detection.*
>
> *Therefore they agree to wager one case of good red wine (JPO to supply French wine, KST to supply California) on the detection of extraterrestrial gravitational waves before the next Millennium (January 1, 2000). KST wins the wager if at least two experimental groups observe phenomena which they agree are gravitational waves. If not, JPO wins.*
>
> *Signed and officially sealed*
> *this sixth day of May 1981*
>
> *Jeremiah P. Ostriker*
> *Kip S. Thorne*

I expect to win – but I won't guarantee it.

Acknowledgments

For helpful comments on the manuscript of this chapter I thank Thibault Damour, Ron Drever, John Armstrong, Peter Bender, Jiři Bičak, David Blair, Roger Blandford, Herman Bondi, Carlton Caves, Frank Estabrook, Charles Evans, Sam Finn, Craig Hogan, Jim Hough, Ed Leaver, Brian Meers, Roger Romani, David Schoemaker, Dan Stinebring, and Kimio Tsubono.

References

Abbott, L. and Wise, M. (1984). *Nuclear Physics B*, **244**, 541.

Alpar, M. A. and Pines, D. (1985). *Nature*, **314**, 334.

Amaldi, E. and Pizella, G. (1979). In *Relativity, Quanta, and Cosmology in the Development of the Scientific Thought of Einstein*, Volume 1, 9.

Amaldi, E., Coccia, E., Cosmelli, C., Ogawa, Y., Pizzella, G., Rapagnani, P., Ricci, F., Bonifazi, P., Castellano, M. G., Vannaroni, G., Bronzoni, F., Carelli, P., Foglietti, V., Cavallari, G., Habel, R., Modena, I. and Pallottino, G. V. (1984). *Il Nuovo Cimento*, **7C**, 338.

Anandan, J. (1981). *Physical Review Letters*, **47**, 463.

Anandan, J. (1985). *Physics Letters*, **105A**, 280.

Anandan, J. (1986). In Ruffini, ed. (1986).

Anandan, J. and Chiao, R. Y. (1982). *General Relativity and Gravitation*, **14**, 515.

Anderson, A. J. (1971). *Nature*, **229**, 547.

Anderson, J. D., Armstrong, J. W., Estabrook, F. B., Hellings, R. W., Law, E. K. and Wahlquist, H. D. (1984). *Nature*, **308**, 158.

Anderson, J. L. and Hobill, D. W. (1986). In *Dynamical Spacetimes and Numerical Relativity*, ed. J. M. Centrella, p. 389. Cambridge University Press.

Anderson, J. L., Kates, R. E., Kegeles, L. S. and Madonna, R. G. (1982). *Physical Review D*, **25**, 2038.

Armstrong, J. W. (1984). *Nature*, **307**, 527.

Armstrong, J. W., Estabrook, F. B. and Wahlquist, H. D. (1987). *Astrophysical Journal* (in press).

Armstrong, J. W. and Sramek, R. A. (1982). *Radio Science*, **17**, 1579.

Armstrong, J. W., Woo, R. and Estabrook, F. B. (1979). *Astrophysical Journal*, **230**, 570.

Ashtekar, A. (1983). In Deruelle and Piran, eds. (1983), p. 421.

Ashtekar, A. (1984). In *General Relativity and Gravitation*, ed. B. Bertotti *et al.*, p. 37. Reidel: Dordrecht.

Backer, D. C., Kulkarni, S. R., Heilis, C., Davis, M. M. and Gross, W. M. (1982). *Nature*, **300**, 615.

Bagaev, S. N., Chebotaev, V. P., Dychkov, A. S. and Goldort, V. G. (1981). *Applied Physics*, **25**, 161.

Bardeen, J. M. and Press, W. H. (1973). *Journal of Mathematical Physics*, **14**, 7.

Bardeen, J. M. and Piran, T. (1983). *Physics Reports*, **96**, 205.

Begelman, M. C., Blandford, R. D. and Rees, M. J. (1980). *Nature*, **287**, 307.

Bender, P. L., Faller, J. E., Hall, J. L., Hils, D. and Vincent, M. A. (1984). Abstract for *Fifth International Laser Ranging Instrumentation Workshop*; Herstmonceaux, 10–14 September, 1984.

Bendat, J. S. (1958). *Principles and Applications of Random Noise Theory*. Wiley: New York.

Bertotti, B. (1973). *Astrophysical Letters*, **14**, 51.

Bertotti, B., Carr, B. J. and Rees, M. J. (1983). *Monthly Notices of the Royal Astronomical Society*, **203**, 945.

Bethe, H. (1986). In *Highlights in Modern Astrophysics*, eds. S. L. Shapiro and S. A. Teukolsky, p. 45. Wiley: New York.

Bicak, J. (1968). *Proceedings of the Royal Society of London A*, **302**, 201.

Bicak, J. (1985). In *Galaxies, Axisymmetric Systems, and Relativity*, ed. M. A. H. MacCallum, Cambridge University Press.

Bicak, J., Hoenselaers, C. and Schmidt, B. G. (1983). *Proceedings of the Royal Society of London A*, **390**, 411.

Blair, D. (1982). *Physics Letters A*, **91**, 197.

Blair, D. G. (1983). In Deruelle and Piran (1983), p. 339.

Blanchet, L. (1987a). *Proceedings of the Royal Society of London A*, **409**, 383.

Blanchet, L. (1987b). Paper in preparation.

Blanchet, L. and Damour, T. (1984). *Physics Letters*, **104A**, 82.

Blanchet, L. and Damour, T. (1986). *Philosophical Transactions of the Royal Society* (in press).

Blandford, R. D. (1979). In Smarr, ed. (1979), p. 191.

Blandford, R. D. (1984). Private communication.

Blandford, R. D., Narayan, R. and Romani, R. W. (1984). *Journal of Astrophysics and Astronomy*, **5**, 369.

Blandford, R. D. and Thorne, K. S. (1979). In *General Relativity: An Einstein Centenary Survey*, eds. S. W. Hawking and W. Israel, Cambridge University Press.

Blinnikov, S. I., Novikov, I. D., Perevodchikova, T. V. and Polnarev, A. G. (1984). *Soviet Astronomy Letters*, **10**, 177.

Bocko, M. F. and Johnson, W. W. (1984). *Physical Review A*, **30**, 2135.

Bond, J. R. and Carr, B. J. (1984). *Monthly Notices of the Royal Astronomical Society*, **207**, 585.

Bondi, H. (1957). *Nature*, **179**, 1072.

Bondi, H. (1960). *Nature*, **186**, 535.

Bondi, H., Pirani, F. A. E. and Robinson, I. (1959). *Proceedings of the Royal Society of London A*, **251**, 519.

Bondi, H., van der Burg, M. G. J. and Metzner, A. W. K. (1962). *Proceedings of the Royal Society of London A*, **269**, 21.

Bonnor, W. B. (1959). *Philosophical Transactions of the Royal Society of London A*, **251**, 233.

Bontz, R. J. and Haugan, M. P. (1981). *Astrophysics and Space Space Science*, **78**, 204.

Bontz, R. J. and Price, R. H. (1979). *Astrophysical Journal*, **228**, 560.

Bordé, C. J., Sharma, J., Tourrenc, P. and Damour, T. (1983). *Journal de Physique-Lettres*, **44**, L983.

Boughn, S. P. and Kuhn, J. R. (1984). *Astrophysical Journal*, **286**, 387.

Boughn, S. P. and Kuhn, J. R. (1987). Paper in preparation.

Boughn, S. P., Fairbank, W. M., Giffard, R. P., Hollenhorst, J. N., Mapoles, E. R., McAshan, M. S., Michelson, P. F., Paik, H. J. and Taber, R. C. (1982). *Astrophysical Journal*, **261**, L19.

Braginsky, V. B. (1965). *Uspekhi Fizicheskikh Nauk*, **86**, 433.

Braginsky, V. B. (1983). In Deruelle and Piran (1983), p. 387.

Braginsky, V. B. and Gertsenshtein, M. E. (1967). *Soviet Physics – JETP Letters*, **5**, 287.

Braginsky, V. B. and Grishchuk, L. P. (1985). *Zhurnal Eksperimentalnoi i Teoreticheskoi Fiziki*, **89**, 744. English translation: *Soviet Physics – JETP*, **62**, 427.

Braginsky, V. B. and Menskii, M. B. (1971). *Soviet Physics – JETP Letters*, **13**, 417.

Braginsky, V. B. and Nazarenko, V. S. (1971). In *Proceedings of the Conference on Experimental Tests of Gravitational Theories, November 11–13, 1970, California Institute of Technology*, ed. R. W. Davies. Jet Propulsion Laboratory Technical Memorandum, 33–499: Pasadena, California.

Braginsky, V. B., Grishchuk, L. P., Doroshkevich, A. G., Zel'dovich, Ya. B., Novikov, I. D. and Sazhin, M. V. (1973). *Soviet Physics – JETP*, **38**, 865.

Braginsky, V. B., Gusev, A. V., Mitrofanov, V. P., Rudenko, V. N. and Yakimov, V. N. (1985). *Uspekhi Fizicheskikh Nauk*, **147**, 422.

Braginsky, V. B., Mitrofanov, V. P. and Panov, V. I. (1985). *Systems with Small Dissipation*. University of Chicago Press: Chicago.

Braginsky, V. B. and Thorne, K. S. (1985). *Nature*, **316**, 610.

Braginsky, V. B. and Thorne, K. S. (1987). *Nature* (in press).

Braginsky, V. B. and Vorontsov, Yu. I. (1974). *Soviet Physics—Uspekhi*, **17**, 644.

Braginsky, V. B., Vorontsov, Yu. I. and Khalili, F. Ya. (1978). *Soviet Physics – JETP Letters*, **27**, 276.

Braginsky, V. B., Vorontsov, Yu. I. and Thorne, K. S. (1980). *Science*, **209**, 547.

Brill, D. R. and Hartle, J. B. (1964). *Phys. Rev. B*, **135**, 271.

Brillet, A. (1985). *Annales de Physique*, **10**, 219.

Brillet, A. and Meers, B. (1987). Paper in preparation.

Brillet, A. and Tourrenc, P. (1983). In Deruelle and Piran (1983).

Burke, W. L. (1969). Unpublished Ph.D. thesis, Caltech.

Burke, W. L. (1975). *Astrophysical Journal*, **196**, 329.

Burke, W. L. (1979). In *Isolated Gravitating Systems in General Relativity*, ed. J. Ehlers, p. 220. North Holland: Amsterdam.

Carr, B. J. (1980). *Astronomy and Astrophysics*, **89**, 6.

Carr, B. J. (1986). In *Inner Space/Outer Space*, ed. E. W. Kolb, University of Chicago Press: Chicago.

Carter, B. and Quintana, H. (1977). *Physical Review D*, **16**, 2928.

Caves, C. M. (1978). Private communication.

Caves, C. M. (1979). *Physics Letters*, **80B**, 323.

Caves, C. M. (1980). *Physical Review Letters*, **45**, 75.

Caves, C. M. (1981). *Physical Review D*, **23**, 1693.

Caves, C. M. (1982). *Physical Review D*, **26**, 1817.

Caves, C. M. (1983). *Foundations of Quantum Mechanics*, ed. S. Kamefuchi *et al.*, p. 195. Physical Society of Japan: Tokyo.

Caves, C. M. (1986). *Physical Review D*, **33**, 1643.

Caves, C. M. (1987*a*). *Physical Review D*, **35**, 1815.

Caves, C. N. (1987*b*). In *Quantum Measurement and Chaos*, ed. E. R. Pike. Plenum: New York.

Caves, C. M. (1987*c*). Paper in preparation.

Caves, C. M., Thorne, K. S., Drever, R. W. P., Sandberg, V. D. and Zimmermann, M. (1980). *Reviews of Modern Physics*, **52**, 341.

Chandrasekhar, S. (1950). *Radiative Transfer*. Oxford University Press: London.

Chandrasekhar, S. (1970). *Physical Review Letters*, **24**, 611.

Chandrasekhar, S. (1975). *Proceedings of the Royal Society of London A*, **343**, 289.

Chandrasekhar, S. (1983). *The Mathematical Theory of Black Holes*. Oxford University Press: Oxford.

Chandrasekhar, S. and Detweiler, S. L. (1975). *Proceedings of the Royal Society of London*, **344**, 441.

Chandrasekhar, S. and Detweiler, S. L. (1976). *Proceedings of the Royal Society of London A*, **350**, 165.

Chandrasekhar, S. and Esposito, F. P. (1970). *Astrophysical Journal*, **160**, 153.

Chandrasekhar, S. and Friedman, J. L. (1972). *Astrophysical Journal*, **176**, 745.

Chandrasekhar, S. and Friedman, J. L. (1973). *Astrophysical Journal*, **181**, 481.

Chandrasekhar, S. and Xanthopoulos, B. C. (1986). *Proceedings of the Royal Society of London*, **408**, 175.

Chiao, R. Y. (1982). *Physical Review B*, **25**, 1655.

Christodoulou, D. and Schmidt, B. G. (1979). *Commun. Math. Phys.*, **68**, 275.

Chrzanowski, P. L. (1975). *Physical Review D*, **11**, 2042.

Clark, J. P. A., van den Heuvel, E. P. J. and Sutantyo, W. (1979). *Astronomy and Astrophysics*, **72**, 120.

Clark, J. P. A. and Eardley, D. M. (1977). *Astrophysical Journal*, **215**, 315.

Cohen, J. M. and Kegeles, L. S. (1975). *Physics Letters*, **54A**, 5.

Cole, J. D. (1968). *Perturbation Methods in Applied Mathematics*. Blaisdell: Waltham, Mass.

Crowley, R. J. and Thorne, K. S. (1977). *Astrophysical Journal*, **215**, 624.

Cunningham, C. T., Price, R. H. and Moncrief, V. (1979). *Astrophysical Journal*, **230**, 870.

Cutler, C. and Lindblom, L. (1987). *Astrophysical Journal* (in press).

Cyranski, J. F. and Lubkin, E. (1974). *Annals of Physics*, **87**, 205.

Damour, T. (1983). In Deruelle and Piran (1983). p. 59.

Damour, T. (1986). In *Proceedings of the Fourth Marcel Grossman Meeting on the Recent Developments of General Relativity*, ed. R. Ruffini. North Holland: Amsterdam

Damour, T. (1987). In *Gravitation in Astrophysics*, eds. B. Carter and J. B. Hartle. Plenum: New York.

Damour, T. and Deruelle, N. (1986). *Annales Institut Henri Poincaré (Physique Théorique)*, **44**, 263.

Dautcourt, G. (1969). *Monthly Notices of the Royal Astronomical Society*, **144**, 255.

Dautcourt, G. (1974). In *Confrontation of Cosmological Theories with Observational Data*, Proceedings of IAU Symposium 63, ed. M. S. Longair, p. 299. Reidel: Dordrecht.

Dautcourt, G. (1977). *Astronomische Nachrichten*, **298**, 81.

Davies, R. W. (1974). In *Colloque Internationaux CNRS No. 220, 'Ondes et Radiations Gravitationelles'*. Institut Henri Poincaré: Paris, p. 33.

Davis, M. M., Taylor, J. H., Weisberg, J. M. and Backer, D. C. (1985). *Nature*, **315**, 547.

De Logi, W. K. and Kovacs, S. J. (1977). *Physical Review D*, **16**, 2331.

Deruelle, N. and Piran, T. (1983). *Gravitational Radiation*. North Holland: Amsterdam.

de Sabbata, V. and Weber, J. eds. (1977). *Topics in Theoretical and Experimental Gravitation Physics*. Plenum: London.

Detweiler, S. L. (1975). *Astrophysical Journal*, **197**, 203.

Detweiler, S. L. (1977). *Proceedings of the Royal Society of London A*, **352**, 381.

Detweiler, S. L. (1978). *Astrophysical Journal*, **225**, 687.

Detweiler, S. L. (1979). *Astrophysical Journal*, **234**, 1100.

Detweiler, S. L. (1980). *Astrophysical Journal*, **239**, 292.

Detweiler, S. L. and Ipser, J. R. (1973). *Astrophysical Journal*, **185**, 685.

Detweiler, S. L. and Lindblom, L. (1985). *Astrophysical Journal*, **292**, 12.

Detweiler, S. L. and Szedenits, E. (1979). *Astrophysical Journal*, **231**, 211.

DeWitt, B. S. (1967*a*). *Phys. Rev.*, **160**, 1113.

DeWitt, B. S. (1967*b*). *Phys. Rev.*, **162**, 1239.

Douglass, D. H. and Braginsky, V. B. (1979). In *General Relativity, an Einstein Centenary Survey*, eds. S. W. Hawking and W. Israel. Cambridge University Press.

Downs, G. S. and Reichley, P. E. (1983). *Astrophysical Journal Supplement Series*, **53**, 169.

Drever, R. W. P. (1977). *Quarterly Journal of the Royal Astronomical Society*, **18**, 9.

Drever, R. W. P. (1983). In Deruelle and Piran (1983). p. 321.

Drever, R. W. P., Hough, J., Edelstein, W., Pugh, J. R. and Martin, W. (1977). In *Gravitazione Sperimentale*, ed. B. Bertotti, p. 365. Accademia Nazionale dei Lincei: Rome.

Drever, R. W. P., Ford, G. M., Hough, J., Kerr, I., Munley, A. J., Pugh, J. R., Robertson, N. A. and Ward, H. (1980). *Proceedings of the Ninth International Conference on General Relativity and Gravitation*, ed. E. Schmutzer, p. 265. VEB Deutscher Verlag der Wissenschaften: Berlin.

Drever, R. W. P. and Weiss, R. (1983). Unpublished work reported briefly in Appendix B.2(c) of Drever *et al.* (1985).

Drever, R. W. P., Weiss, R., Linsay, P. S., Saulson, P. R., Spero, R. and Schutz, B. F. (1985). A Detailed Engineering Design Study and Development and Testing of Components for a Laser Interferometer Gravitational Wave Observatory. Caltech: Pasadena, California, and MIT: Cambridge. Massachusetts (unpublished).

Dymnikova, I. G., Popov, A. K. and Zentsova, A. S. (1982). *Astrophysics and Space Science*, **85**, 231.

Dyson, F. J. (1969). *Astrophysical Journal*, **156**, 529.

Eardley, D. M. (1983). In Deruelle and Piran (1983). p. 257.

Eardley, D. M., Lee, D. L. and Lightman, A. P. (1973). *Physical Review D*, **8**, 3308.

Eardley, D. M., Lee, D. L., Lightman, A. P., Wagoner, R. V. and Will, C. M. (1973). *Physical Review Letters*, **30**, 884.

Eddington, A. S. (1924). *The Mathematical Theory of Relativity*, 2nd edn, Cambridge University Press; see especially supplementary note 8.

Edelstein, L. A. and Vishveshwara, C. V. (1970). *Physical Review D*, **1**, 3514.

Ehlers, J. and Kundt, W. (1962). In *Gravitation: An Introduction to Current Research*, ed. L. Witten. Wiley: New York.

Ehlers, J., Rosenblum, A., Goldberg, J. N. and Havas, P. (1976). *Astrophysical Journal Letters*, **208**, L77.

Einstein, A. (1916). *Preuss. Akad. Wiss. Berlin. Sitzungsberichte der physikalisch-mathematischen Klasse*, p. 688.

Einstein, A. (1918). *Preuss. Akad. Wiss. Berlin, Sitzungsberichte der Physikalisch-mathematischen Klasse*, p. 154.

Einstein, A. and Rosen, N. (1936). *Journal of the Franklin Institute*, **223**, 43.

Epstein, R. and Wagoner, R. V. (1975). *Astrophysical Journal*, **197**, 717.

Esposito, F. P. (1971*a*). *Astrophysical Journal*, **165**, 165.

Esposito, F. P. (1971*b*). *Astrophysical Journal*. **168**, 495.

Estabrook, F. B. (1985). *General Relativity and Gravitation*, **17**, 719.

Estabrook, F. B. (1987). *Acta Astronautica* (in press).

Estabrook, F. B. and Wahlquist, H. D. (1975). *General Relativity and Gravitation*, **6**, 439.

Evans, C. R. (1984). PhD thesis, University of Texas at Austin, unpublished.

Evans, C. R. (1986). In *Dynamical Spacetimes and Numerical Relativity*, ed. J. M. Centrella, p. 3. Cambridge University Press.

Evans, C. R. and Abrahams, A. M. (1987). *Physical Review D* (in preparation).

Evans, C. R., Iben, I. and Smarr, L. (1987). *Astrophysical Journal*, submitted.

Faller, J. E. and Bender, P. L. (1981). Abstract for the *International Conference on Precision Measurements and Fundamental Constants*. 8–12 June 1981.

Faller, J. E., Bender, P. L., Hall, J. L., Hils, D. and Vincent, M. A. (1985). In *Proceedings of the Colloquium 'Kilometric Optical Arrays in Space'*, Cargese (Corsica) 23–5 October 1984. ESA SP-226.

Feynman, R. P. (1963). *Acta Physica Polonica*, **24**, 697: also published in *Proceedings of the 1962 Warsaw Conference on the Theory of Gravitation*. PWN-Editions Scientifiques de Pologne: Warsaw (1964).

Finn, L. S. (1986). *Monthly Notices of the Royal Astronomical Society*, **222**, 393.

Fock, V. A. (1959). *Theory of Space, Time, and Gravitation*, Section 87. Pergamon: London.

Forward, R. L. (1978). *Physical Review D*, **17**, 379.

Forward, R. L. and Moss, G. E. (1972). *Bulletin of the American Physical Society*, **17**, 1183(A).

Forward, R. L., Zipoy, D., Weber, J., Smith, S. and Benioff, H. (1961). *Nature*, **189**, 473.

Friedman, J. L., Ipser, J. R. and Parker, L. (1986). *Astrophysical Journal*, **304**, 115.

Friedman, J. L. and Schutz, B. F. (1975*a*). *Astrophysical Journal (Letters)*, **199**, L157.

Friedman, J. L. and Schutz, B. F. (1975*b*). *Astrophysical Journal*, **200**, 204.

Friedman, J. L. and Schutz, B. F. (1978). *Astrophysical Journal*, **222**, 281.

Futamase, T. and Schutz, B. F. (1983). *Physical Review D*, **28**, 2363.

Gal'tsov, D. V., Matiukhin, A. A. and Petukhov, V. I. (1980). *Physics Letters*, **77A**, 387.

Gal'tsov, D. V., Tsvetkov, V. P. and Tsirulev, A. N. (1984). *Soviet Physics – JETP*, **59**, 472.

Geroch, R., Held, A. and Penrose, R. (1973). *Journal of Mathematical Physics*, **14**, 874.

Gertsenshtein, M. E. (1962). *Soviet Physics – JETP*, **14**, 84.

Gertsenshtein, M. E. and Pustovoit, V. I. (1962). *Soviet Physics – JETP*, **16**, 433.

Gilden, D. L. and Shapiro, S. L. (1984). *Astrophysical Journal*, **287**, 728.

Giffard, R. P. (1976). *Physical Review D*, **14**, 2478.

Gomez, R., Isaacson, R. A., Welling, J. S. and Winicour, J. (1986). In *Dynamical Spacetimes and Numerical Relativity*, ed. J. M. Centrella, p. 236. Cambridge University Press.

Gordienko, N. V., Gusev, A. V. and Rudenko, V. N. (1977). *Vestnik Moskovskovo Universiteta*, **18**, 48.

Grishchuk, L. P. (1974). *Soviet Physics – JETP*, **40**, 409.

Grishchuk, L. P. (1975*a*). *Lettere al Nuovo Cimento*, **12**, 60.

Grishchuk, L. P. (1975*b*). *Soviet Physics – JETP*, **40**, 409.

Grishchuk, L. P. (1977). *Annals of the New York Academy of Sciences*, **302**, 439.

Grishchuk, L. P. (1983). In *Proceedings of the Ninth International Conference on General Relativity and Gravitation*, ed. E. Schmutzer. Cambridge University Press.

Grishchuk, L. P. and Polnarev, A. G. (1980). In *General Relativity and Gravitation*, vol. 2, ed. A. Held, p. 393. Plenum: New York.

Grishchuk, L. P. and Zel'dovich, Ya. B. (1978). *Soviet Astronomy – AJ*, **22**, 125.

Gürsel, Y., Linsay, P., Saulson, P., Spero, R., Weiss, R. and Whitcomb, S. (1983). Unpublished Caltech/MIT manuscript.

Gusev, A. V. and Rudenko, V. N. (1976). *Radiotekhnika i Electronika*, **3**, 1865.

Haensel, P., Zdunik, J. L. and Schaeffer, R. (1986). *Astronomy and Astrophysics*, **160**, 251.

Halliwell, J. J. and Hawking, S. W. (1985). *Physical Review D*, **31**, 1777.

Halpern, L. E. and Desbrandes, R. (1969). *Annales Institut Henri Poincaré*, 9, 309.

Hamilton, W. O., Xu, B.-X., Solomonson, N., Mann, A. G. and Sibley, A. (1986). Performance of the LSU Tuned Bar Gravitational Wave Detector. Unpublished technical memorandum, Louisiana State University.

Hansen, P. M., Chiu, Y. T. and Chao, C.-C. (1986). Aerospace Report No. ATR-86(8421)-2. The Aerospace Corporation: El Segundo, CA.

Hawking, S. W. (1985). *Physics Letters*, 150B, 339.

Haugan, M. P., Shapiro, S. L. and Wasserman, I. (1982). *Astrophysical Journal*, 257, 283.

Heffner, H. (1962). *Proceedings of the IRE*, 50, 1604.

Helfand, D. H., Taylor, J. H., Backus, R. R. and Cordes, J. M. (1980). *Astrophysical Journal*, 237, 206.

Hellings, R. W. and Downs, G. S. (1983). *Astrophysical Journal (Letters)*, 265, L39.

Hellings, R. W., Callahan, P. S., Anderson, J. D. and Moffett, A. T. (1981). *Physical Review D*, 23, 844.

Hils, D. L., Bender, P., Faller, J. E. and Webbink, R. F. (1987). Paper in preparation.

Hirakawa, H. and Narihara, K. (1975). *Physical Review Letters*, 35, 330.

Hirakawa, H., Narihara, K. and Fujimoto, M.-K. (1976). *Journal of the Physical Society of Japan*, 41, 1093.

Hirakawa, H., Owa, S. and Iso, K. (1985). *Journal of the Physical Society of Japan*, 54, 1270.

Hogan, C. J. (1986). *Monthly Notices of the Royal Astronomical Society*, 218, 629.

Hogan, C. J. and Rees, M. J. (1984). *Nature*, 311, 109.

Hollenhorst, J. N. (1979). *Physical Review D*, 19, 1669.

Hough, J., Meers, B. J., Newton, G. P., Robertson, N. A., Ward, H., Schutz, B. F. and Drever, R. W. P. (1986). A British Long Baseline Gravitational Wave Observatory, unpublished report submitted by Glasgow University to the SERC.

Hough, J., Pugh, J. R., Bland, R. and Drever, R. W. P. (1975). *Nature*, 254, 498.

Hough, J., Drever, R. W. P., Munley, A. J., Lee, S.-A., Spero, R., Whitcomb, S. E., Pugh, J., Newton, G., Meers, B., Brooks, E. and Gursel, Y. (1983). In *Quantum Optics, Experimental Gravity, and Measurement Theory*, eds. P. Meystre and M. O. Scully, p. 515. Plenum: New York.

Hu, B. L. (1978). *Physical Review D*, 18, 969.

Hu Enke, Guan Tongren, Yu Bo, Tang Mengxi, Chen Shusen, Zheng Qingzhang, Michelson, P. F., Moskowitz, B. E., McAshan, M. S., Fairbank, W. M. and Bassan, M. (1986). *Chinese Physics Letters* No. 11–12, 1.

Iacopini, E., Picasso, E., Pegoraro, F. and Radicati, L. A. (1979). *Physics Letters*, 73A, 140.

Iben, I. and Tutukov, A. V. (1984). *Astrophysical Journal Supplements*, 54, 335.

Ipser, J. R. (1971). *Astrophysical Journal*, 166, 175.

Ipser, J. R. (1986). Private communication.

Ipser, J. R. and Managan, R. A. (1984). *Astrophysical Journal*, 282, 287.

Ipser, J. R. and Thorne, K. S. (1973). *Astrophysical Journal*, 181, 181.

Isaac, G. R. (1981). *Solar Physics*, 74, 43.

Isaacson, R. A. (1968*a*). *Physical Review*, 166, 1263.

Isaacson, R. A. (1968*b*). *Physical Review*, 166, 1272.

Isaacson, R. A., Welling, J. S. and Winicour, J. (1983). *Journal of Mathematical Physics*, 24, 1824.

Isaacson, R. A., Welling, J. S. and Winicour, J. (1984). *Physical Review Letters*, 53, 1870.

Kafka, P. (1977). In de Sabbata and Weber (1977), p. 161.

Kahn, K. and Penrose, R. (1971). *Nature*, 229, 185.

Kaufmann, W. J. (1970). *Nature*, 227, 157.

Kellerman, K. I. and Sheets, B. (1983). *Serendipitous Discoveries in Radio Astronomy*. National Radio Astronomy Observatory: Green Bank, West Virginia.

Kojima, Y. and Nakamura, T. (1984*a*). *Progress of Theoretical Physics*, 71, 79.

Kojima, Y. and Nakamura, T. (1984*b*). *Progress of Theoretical Physics*, 72, 494.

Kovacs, S. J. and Thorne, K. S. (1977). *Astrophysical Journal*, 217, 252.

Kovacs, S. J. and Thorne, K. S. (1978). *Astrophysical Journal*, 224, 62.

Kuhn, J. R. and Boughn, S. P. (1984). *Nature*, 308, 164.

Landau, L. D. and Lifshitz, E. M. (1941). *Teoriya Polya*. Nauka: Moscow. (English translation of a later edition, *The Classical Theory of Fields*. Addison-Wesley: Cambridge, Mass. (1951).)

Leaver, E. W. (1985). Solutions to a Generalized Spheroidal Wave Equation in Molecular Physics and General Relativity, and an Analysis of the Quasi-Normal Modes of Kerr Black Holes, unpublished PhD thesis. University of Utah: Salt Lake City.

Leaver, E. W. (1986*a*). *Proceedings of the Royal Society of London A*, 402, 285.

Leaver, E. W. (1986*b*). *Physical Review D*, 34, 384.

Lifshitz, E. M. (1946). *Zhurnal Eksperimentalnoi i Teoreticheskoi Fiziki*, 16, 587.

Lindblom, L. (1986). *Astrophysical Journal*, 303, 146.

Lindblom, L. (1987). *Astrophysical Journal* (in press).

Lindblom, L. and Detweiler, S. L. (1983). *Astrophysical Journal*, 53, 73.

Linet, B. (1984). *General Relativity and Gravitation*, 16, 89.

Linet, B. and Tourrenc, P. (1976). *Canadian Journal of Physics*, 54, 1129.

Linsay, P., Saulson, P., Weiss, R. and Whitcomb, S. (1983). A Study of a Long Baseline Gravitational Wave Antenna System; report prepared for The National Science Foundation. MIT: Cambridge, Massachusetts (unpublished).

Lipunov, V. M. and Postnov, K. A. (1986). *Soviet Astronomy Letters*. In press.

Lipunov, V. M., Postnov, K. A. and Prokhorov, X. (1987). Paper in preparation.

Lyamov, V. E. and Rudenko, V. N. (1975). *Soviet Physics – JETP*, 40, 787.

MacCallum, M. A. H., ed. (1987). *Proceedings of the Tenth International Conference on General Relativity and Gravitation*. Cambridge University Press.

MacCallum, M. A. H. and Taub, A. H. (1973). *Communications in Mathematical Physics*, 30, 153.

Macedo, P. G. and Nelson, A. H. (1983). *Physical Review D*, 28, 2382.

Maischberger, K., Rudiger, A., Schilling, R., Schnupp, L., Shoemaker, D. and Winkler, W. (1985). Vorschlag zum Bau eines grossen Laser-Interferometers zur Messung von Gravitationswellen. Max-Planck Institut fur Quantenoptik: Munich (unpublished).

Mashhoon, B., Carr, B. J. and Hu, B. L. (1981). *Astrophysical Journal*, 246, 569.

Matzner, R. A., DeWitt-Morette, C., Nelson, B. and Zhang, T.-R. (1985). *Physical Review D*, 31, 1869.

Matzner, R. A. and Tipler, F. J. (1984). *Physical Review D*, 29, 1575.

Meers, B. (1983). Unpublished PhD thesis, University of Glasgow.

Michelson, P. F. (1983). In Deruelle and Piran (1983), p. 465.

Michelson, P. F. (1986). *Physical Review D*, 34, 2966.

Michelson, P. F. and Taber, R. C. (1981). *Journal of Applied Physics*, 52, 4313.

Michelson, P. F. and Taber, R. C. (1984). *Physical Review D*, 29, 2149.

Mijic, M., Morris, M. and Suen, W.-M. (1986). *Physical Review D*, 34, 2934.

Mironovskii, V. N. (1966). *Soviet Astronomy – AJ*, 9, 752.

Misner, C. W., Thorne, K. S. and Wheeler, J. A. (1973). *Gravitation*. W. H. Freeman & Co.: San Francisco. Cited in text as MTW.

Moncrief, V. (1974). *Annals of Physics*, **88**, 323.

Moss, G. E., Miller, L. R. and Forward, R. L. (1971). *Applied Optics*, **10**, 2495.

Müller, E. (1982). *Astronomy and Astrophysics*, **114**, 53.

Müller, E. (1984). In *Problems of Collapse and Numerical Relativity*, eds. D. Bancel and M. Signore, p. 271. Reidel: Dordrecht.

Nakamura, T. (1983). In Deruelle and Piran (1983).

Nakamura, T. (1986). In *Gravitational Collapse and Relativity*, eds. H. Sato and T. Nakamura, p. 295. World Scientific: Singapore.

Nakamura, T. and Sasaki, M. (1981). *Physics Letters*, **106B**, 69.

Nesterikhin, Y. E., Rautian, S. G. and Smirnov, G. I. (1978). *Soviet Physics – JETP*, **48**, 1.

Newman, E. T. and Penrose, R. (1965). *Physical Review Letters*, **15**, 231.

Newman, E. T. and Tod, K. P. (1980). In *General Relativity and Gravitation, Volume 2*, ed. A. Held, p. 1. Plenum: New York.

Nutku, Y. and Halil, M. (1977). *Physical Review Letters*, **39**, 1379.

Oelfke, W. C. (1983). In *Quantum Optics, Experimental Gravitation and Measurement Theory*, eds. P. Meystre and M. O. Scully, p. 387. Plenum: New York.

Ostriker, J. P., Thompson, C. and Witten, E. (1986). *Physics Letters*, **B180**, 231.

Owa, S., Fujimoto, M.-K., Hirakawa, H., Morimoto, K., Suzuki, T. and Tsubono, K. (1986). In *Proceedings of Fourth Marcel Grossman Meeting on Recent Developments of General Relativity*, ed. R. Ruffini. North Holland: Amsterdam.

Paczynski, B. and Sienkiewicz, R. (1981). *Astrophysical Journal (Letters)*, **248**, L27.

Paik, H. J. and Wagoner, R. V. (1976). *Physical Review D*, **13**, 2694.

Pandharipande, V. R., Pines, D. and Smith, R. A. (1976). *Astrophysical Journal*, **208**, 550.

Pallotino, G. V. and Pizella, G. (1981). *Nuovo Cimento*, **4C**, 237.

Papadopoulos, D. and Esposito, F. P. (1985). *Astrophysical Journal*, **282**, 330.

Papapetrou, A. (1962). *Comptes Rendus Acad. Sci. Paris*, **255**, 1578.

Papapetrou, A. (1971). *Annales Institut Henri Poincaré*, **14**, 79.

Papapetrou, A. (1977). In deSabbata and Weber (1977), p. 83.

Pegoraro, F., Picasso, E. and Radicati, J. (1978). *Journal of Physics A*, **11**, 1949.

Penrose, R. (1963*a*). *Physical Review Letters*, **10**, 66.

Penrose, R. (1963*b*). *Relativity, Groups and Topology*, eds. C. DeWitt and B. DeWitt, p. 565. Gordon and Breach: New York.

Peres, A. (1960). *Nuovo Cimento*, **15**, 351.

Peters, P. C. (1964). *Physical Review*, **136**, B1224.

Peters, P. C. and Mathews, J. (1963). *Physical Review*, **131**, 435.

Piran, T. (1983). In Deruelle and Piran (1983), p. 203.

Piran, T. and Stark, R. F. (1986). In *Dynamical Spacetimes and Numerical Relativity*, ed. J. M. Centrella, p. 40. Cambridge University Press.

Pirani, F. A. E. (1956). *Acta Physica Polonica*, **15**, 389.

Pirani, F. A. E. (1957). *Physical Review*, **105**, 1089.

Pirani, F. A. E. (1964). In *Lectures on General Relativity*, eds. A. Trautman, F. A. E. Pirani and H. Bondi. Prentice-Hall: Englewood Cliffs, NJ.

Power, E. A. and Wheeler, J. A. (1957). *Reviews of Modern Physics*, **29**, 480.

Press, W. H. (1971). *Astrophysical Journal (Letters)*, **170**, L105.

Press, W. H. (1977). *Physical Review D*, **15**, 965.

Press, W. H. (1979). *General Relativity and Gravitation*, **11**, 105.

Price, R. H. (1972*a*). *Physical Review D*, **5**, 2419.

Price, R. H. (1972*b*). *Physical Review D*, **5**, 2439.

Price, R. H. and Thorne, K. S. (1969). *Astrophysical Journal*, **155**, 163.

Priedhorsky, W., Stella, L. and White, N. E., *International Astronomical Union Circular*, No. 4247, 28 August, 1986.

Quinlan, G. and Shapiro, S. L. (1987). *Astrophysical Journal*, in press.

Ramaty, R., Bonazzola, S., Cline, T. L., Kazanas, D. and Meszaros, P. (1980). *Nature*, 287, 122.

Rees, M. J. (1971). *Monthly Notices of the Royal Astronomical Society*, **154**, 187.

Rees, M. J. (1983). In Deruelle and Piran (1983), p. 297.

Regge, T. and Wheeler, J. A. (1957). *Physical Review*, **108**, 1063.

Resch, G. M., Hogg, D. E. and Napier, P. G. (1984). *Radio Science*, **19**, 411.

Reynolds, S. P. and Stinebring, D. R. (1984). *Millisecond Pulsars*, National Radio Astronomy Observatory: Green Bank, West Virginia.

Romani, R. W. and Taylor, J. H. (1983). *Astrophysical Journal (Letters)*, **265**, L35.

Rosen, N. (1937). *Physikalische Zeitschift Sowjetunion*, **12**, 366.

Rubakov, V., Sazhin, M. C. and Veryaskin, A. (1982). *Physics Letters*, **115B**, 189.

Rudenko, V. N. (1975). *Soviet Astronomy – AJ*, **19**, 270.

Rudenko, V. N. and Sazhin, M. V. (1980). *Kvantovaya Elektronika*, **7**, 2344.

Ruffini, R. ed. (1986). *Proceedings of the Fourth Marcel Grossman Meeting on Recent Developments of General Relativity*. North Holland: Amsterdam.

Ruffini, R. and Wheeler, J. A. (1971). In *Proceedings of the Conference on Space Physics*. European Space Research Organization: Paris, p. 45.

Sachs, R. K. (1962). *Proceedings of the Royal Society of London A*, **270**, 103.

Sachs, R. K. (1963). In *Relativity, Groups and Topology*, eds. C. DeWitt and B. DeWitt, p. 521. Gordon and Breach: New York.

Sachs, R. K. and Wolfe, A. M. (1967). *Astrophysical Journal*, **147**, 73.

Saenz, R. A. and Shapiro, S. L. (1978). *Astrophysical Journal*, **221**, 286.

Saenz, R. A. and Shapiro, S. L. (1979). *Astrophysical Journal*, **229**, 1107.

Saenz, R. A. and Shapiro, S. L. (1981). *Astrophysical Journal*, **244**, 1033.

Sasaki, M. (1984). In *Problems of Collapse and Numerical Relativity*, eds. D. Bancel and M. Signore, p. 203. Reidel: Dordrecht.

Sasaki, M. and Nakamura, T. (1982). *Progress of Theoretical Physics*, **67**, 1788.

Sato, H. (1987). Paper in press.

Saulson, P. (1984). *Physical Review D*, **30**, 732.

Saulson, P. (1984). *Physical Review D*, **30**, 732.

Sazhin, M. V. (1978). *Soviet Astronomy – AJ*, **22**, 36.

Schmidt, B. G. (1979). In *Isolated Gravitating Systems*, ed. J. Ehlers, p. 11. North Holland: Amsterdam.

Schmidt, B. G. (1986). Gravitational Radiation Near Spatial and Null Infinity, preprint.

Schilling, R. (1986). Private communication.

Schoemaker, D., Winkler, W., Maischberger, K., Rudiger, A., Schilling, R. and Schnupp, L. (1986). In Ruffini, ed. (1986).

Schrader, R. (1984). *Physics Letters B*, **143B**, 421.

Schramm, D. N. and Arnett, W. D. (1975). *Astrophysical Journal*, **198**, 629.

Schumaker, B. L. (1986). *Physics Reports*, **135**, 318.

Schumaker, B. L. and Thorne, K. S. (1983). *Monthly Notices of the Royal Astronomical Society*, **203**, 457.

Schutz, B. F. (1972). *Astrophysical Journal Supplement Series*, **24**, 343.

Schutz, B. F. (1986*a*). In *Relativity, Supersymmetry, and Cosmology*, ed. O. Bressan, M. Castagnino and V. Hamity, p. 3. World Scientific: Singapore.

Schutz, B. F. (1986*b*). *Nature*, **323**, 310.

Schutz, B. F. (1986*c*). In *Dynamical Spacetimes and Numerical Relativity*, ed. J. Centrella, p. 446. Cambridge University Press.

Schutz. B. F. (1987). In *Gravitation in Astrophysics*. eds. B. Carter and J. Hartle. Plenum: New York.

Schutz, B. F. and Tinto, M. (1987). *Monthly Notices of the Royal Astronomical Society*, 224, 131.

Schvartzman, V. F. (1969). *Soviet Physics – JETP Letters*, 9, 184.

Segelstein, D. J., Rawley, L. A., Stinebring, D. R., Fruchter, A. S. and Taylor, J. H. (1986). *Nature*, 322, 714.

Shapiro, S. L. and Teukolsky, S. A. (1985). *Astrophysical Journal (Letters)*, 292, L41.

Slusher, R. E., Hollberg, L. W., Yurke, B., Mertz, J. C. and Valley, J. F. (1985). *Physical Review Letters*, 55, 2409.

Smarr, L. (1977*a*). *Annals of the New York Academy of Sciences*, 302, 569.

Smarr, L. (1977*b*). *Physical Review D*, 15, 2069.

Smarr, L., ed. (1979). *Sources of Gravitational Radiation*. Cambridge University Press.

Smarr, L. and York, J. W. (1978). *Physical Review D*, 17, 2529.

Smarr, L., Vessot, R. F. C., Lundquist, C. A., Decher, R. and Piran, T. (1983). *General Relativity and Gravitation*, 15, 129.

Spero, R. (1986*a*). In Ruffini, ed. (1986).

Spero, R. (1986*b*). Private communication.

Stark, R. F. and Piran, T. (1985). *Physical Review Letters*, 55, 891 and 56, 97.

Stark, R. F. and Piran, T. (1986). In *Proceedings of the Fourth Marcel Grossman Meeting on Recent Developments of General Relativity*, ed. R. Ruffini (in press).

Starobinsky, A. A. (1979). *Soviet Physics – JETP Letters*, 30, 682.

Starobinsky, A. A. (1985). *Soviet Astronomy Letters*, 11, 133.

Stewart, J. M. and Friedrich, H. (1983). *Proceedings of the Royal Society of London A*, 384, 427.

Suzuki, T., Tsubono, K. and Hirakawa, H. (1978). *Physics Letters A*, 67, 10.

Sullivan, W. T. (1982). *Classics in Radio Astronomy*. Reidel: Dordrecht.

Sullivan, W. T. (1984). *The Early Years of Radio Astronomy*. Cambridge University Press.

Szekeres, P. (1971). *Annals of Physics*, 64, 599.

Tammann, G. A. (1981). In *Supernovae: A Survey of Current Research*, ed. M. J. Rees and R. J. Stoneham, p. 371. Reidel: Dordrecht.

Taylor, J. H. (1987). In *General Relativity and Gravitation*, ed. M. A. H. MacCallum. Cambridge University Press (in press).

Teissier du Cros, F. (1985). *Annales de Physique*, 10, 263.

Taylor, J. H. (1987). In *Proceedings of Eleventh International Conference on General Relativity and Gravitation*. Cambridge University Press (in press).

Teukolsky, S. A. (1972). *Physical Review Letters*, 29, 1114.

Teukolsky, S. A. (1973). *Astrophysical Journal*, 185, 635.

Teukolsky, S. A. and Press, W. H. (1974). *Astrophysical Journal*, 193, 443.

Thorne, K. S. (1969*a*). *Astrophysical Journal*, 158, 1.

Thorne, K. S. (1969*b*). *Astrophysical Journal*, 158, 997.

Thorne, K. S. (1972). In *Magic Without Magic: John Archibald Wheeler*. Appendix A, p. 243. W. H. Freeman: San Francisco.

Thorne, K. S. (1977). In de Sabbata and Weber (1977), p. 1.

Thorne, K. S. (1978). In *Theoretical Principles in Astrophysics and Relativity*. eds. N. R. Lebovitz, W. H. Reid and P. O. Vandervoort, p. 149. University of Chicago Press: Chicago.

Thorne, K. S. (1980*a*). *Reviews of Modern Physics*. 52, 285.

Thorne, K. S. (1980*b*). *Reviews of Modern Physics*. 52, 299.

Thorne, K. S. (1983). Chapter 1 of Deruelle and Piran (1983).

Thorne. K. S. (1985). In *Nonlinear Phenomena in Physics*. ed. F. Claro, p. 280. Springer-Verlag: Berlin.

Thorne, K. S. and Braginsky, V. B. (1976). *Astrophysical Journal (Letters)*, **204**, L1.

Thorne, K. S. and Campolattaro, A. (1967). *Astrophysical Journal*, **149**, 591; and **152**, 673.

Thorne, K. S., Caves, C. M., Sandberg, V. D., Zimmermann, M. and Drever, R. W. P. (1979). In Smarr (1979), p. 49.

Thorne, K. S., Drever, R. W. P., Caves, C. M., Zimmermann, M. and Sandberg, V. D. (1978). *Physical Review Letters*, **40**, 667.

Thorne, K. S. and Gürsel, T. (1983). *Monthly Notices of the Royal Astronomical Society*, **205**, 809.

Thorne, K. S. and Kovacs, S. J. (1975). *Astrophysical Journal*, **200**, 245.

Thorne, K. S., Price, R. H. and Macdonald, D. M., eds. (1986). *Black Holes: The Membrane Paradigm*. Yale University Press: New Haven, Connecticut.

Tipler, F. J. (1980). *Physical Review D*, **22**, 2929.

Tourrenc, P. (1978). *General Relativity and Gravitation*, **9**, 123 and 141.

Tourrenc, P. and Crossiord, J.-L. (1974). *Nuovo Cimento*, **19B**, 105.

Traschen, J., Turok, N. and Brandenberger, R. (1986). *Physical Review D*, **34**, 919.

Tsvetkov, V. P. (1984). *Soviet Astronomy – AJ*, **28**, 394.

Turner, M. and Wagoner, R. V. (1979). In Smarr (1979), p. 383.

Turner, M. and Will, C. M. (1978). *Astrophysical Journal*, **220**, 1107.

Turok, N. and Brandenberger, R. H. (1986). *Physical Review D*, **33**, 2175.

Tyson, J. A. and Giffard, R. P. (1978). *Annual Reviews of Astronomy and Astrophysics*, **16**, 521.

Unruh, W. G. (1982). Unpublished work reported at NATO Advanced Study Institute on Gravitational Radiation, Les Houches, France, June 1982.

Vachaspati, T. and Vilenkin, A. (1985). *Physical Review D*, **31**, 3052.

van den Heuvel, E. P. J. (1984). In Reynolds and Stinebring, ed. (1984).

Vilenkin, A. (1981*a*). *Physical Review D*, **24**, 2082.

Vilenkin, A. (1981*b*). *Physics Letters*, **107B**, 47.

Vinet, J. Y. (1985). *Annales de Physique*, **10**, 253.

Vinet, J. Y. (1986). *Journal de Physique*, **47**, 639.

Vishniac, E. T. (1982). *Astrophysical Journal*, **257**, 456.

Wagoner, R. V. (1977). In *Gravitazione Sperimentale*, ed. B. Bertotti. Accademia Nazionale dei Lincei: Rome.

Wagoner, R. V. (1984). *Astrophysical Journal*, **278**, 345.

Wagoner, R. V. and Will, C. M. (1976). *Astrophysical Journal*, **210**, 764.

Wahlquist, H.-D. (1987). *General Relativity and Gravitation* (in press).

Wahlquist, H. D., Anderson, J. D., Estabrook, F. B. and Thorne, K. S. (1977). In *Gravitazione Sperimentale*, ed. B. Bertotti, p. 335. Accademia Nazionale dei Lincei: Rome.

Wainstein, L. A. and Zubakov, V. D. (1962). *Extraction of Signals from Noise*, Prentice-Hall: London.

Wald, R. M. (1973). *Journal of Mathematical Physics*, **14**, 1453.

Walgate, R. (1983). *Nature*, **305**, 665.

Walker, M. (1983). In *Relativistic Astrophysics and Cosmology*, eds. X. Fustero and E. Verdaguer. World Scientific: Singapore.

Walker, M. and Will, C. M. (1980). *Astrophysical Journal (Letters)*, **242**, L129.

Walls, D. F. (1983). *Nature*, **306**, 141.

Weber, J. (1959). *Reviews of Modern Physics*, **31**, 681 (see especially Section V).

Weber, J. (1960). *Physical Review*, **117**, 306.

Weber, J. (1967). *Physical Review Letters*, **18**, 498.

Weber, J. (1969). *Physical Review Letters*, **22**, 1302.

Weber, J. (1986). In *Proceedings of the Sir Arthur Eddington Centenary Symposium. Volume 3 – Gravitational Radiation and Relativity*, eds. J. Weber and T. M. Karade, p. 1. World Scientific: Singapore.

Weber, J. and Wheeler, J. A. (1957). *Reviews of Modern Physics*, **29**, 509.

Weyl, H. (1922). *Space–Time–Matter*. Methuen: London.

Weisberg, J. M. and Taylor, J. H. (1984). *Physical Review Letters*, **52**, 1348.

Weiss, R. (1972). *Quarterly Progress Report of the Research Laboratory of Electronics of the Massachusetts Institute of Technology*, **105**, 54.

Weiss, R. (1978). In *Sources of Gravitational Radiation*, ed. L. Smarr, p. 7. Cambridge University Press.

Weiss, R. (1979). In Smarr (1979), p. 7.

Weiss, R., Bender, P. L., Misner, C. W. and Pound, R. V. (1976). *Report of the Sub-Panel on Relativity and Gravitation, Management and Operations Working Group for Shuttle Astronomy*. NASA: Washington, DC.

Westpfahl, K. (1985). *Fortschritte der Physik*, **33**, 417.

Wheeler, J. A. (1955). *Physical Review*, **97**, 511.

Wheeler, J. A. (1962). *Geometrodynamics*. Academic Press: New York.

Wheeler, J. A. (1964). In *Relativity, Groups, and Topology*, eds. C. DeWitt and B. DeWitt. Gordon and Breach: New York.

Whitcomb, S. and Saulson, P. (1984). Unpublished research at Caltech and MIT; a few of the results are given in Fig. E.3 of Drever, Weiss *et al.* (1985).

Wiener, N. (1949). *The Extrapolation, Interpolation and Smoothing of Stationary Time Series with Engineering Applications*. Wiley: New York.

Will, C. M. (1986). *Canadian Journal of Physics*, **64**, 140.

Wilson, J. R. (1974). *Physical Review Letters*, **32**, 849.

Winkler, W. (1977). In *Gravitazione Sperimentale*, ed. B. Bertotti. Academia Nazionale dei Lincei: Rome.

Winkler, W., Maischberger, K., Rüdiger, A., Schilling, R., Schnupp, L. and Shoemaker, D. (1986). In Ruffini, ed. (1986).

Witten, E. (1984). *Physical Review D*, **30**, 272.

Wood, K. S., Michelson, P. F., Boynton, P., Yearian, M. R., Gursky, H., Friedman, H. and Dieter, J. (1986). *A Proposal to NASA for an X-Ray Large Array (XLA) for the NASA Space Station*. Stanford: Palo Alto, California.

Wood, L., Zimmermann, G., Nukolls, J. and Chapline, G. (1971). *Bulletin of the American Physical Society*, **16**, 609.

Woosley, S. E. and Weaver, T. A. (1986). *Annual Reviews of Astronomy and Astrophysics*, **24**, 205.

Wu, L.-A., Kimble, H. J., Hall, J. L. and Wu, H. (1986). *Physical Review Letters*, **57**, 2520.

Yamamoto, Y. and Haus, H. A. (1986). *Reviews of Modern Physics*, **58**, 1001.

York, J. W. (1983). In Deruelle and Piran (1983), p. 175.

Yurtsever, U. (1986). *Abstracts of Contributed Papers, 11th International Conference on General Relativity and Gravitation*, Stockholm, Sweden, 6–12 July, 1986, p. 287: University of Stockholm: Stockholm.

Yurtsever, U. (1987*a*). *Physical Review D*, submitted.

Yurtsever, U. (1987*b*). Paper in preparation.

Zel'dovich, Ya. B. (1980). *Monthly Notices of the Royal Astronomical Society*, **192**, 663.

Zel'dovich, Ya. B. and Novikov, I. D. (1983). *Relativistic Astrophysics Vol. 2. The Structure and Evolution of the Universe*. University of Chicago Press: Chicago.

Zhang, X.-H. (1986). *Physical Review D*, **34**, 991.

Zerilli, F. J. (1970). *Physical Review D*, **2**, 2141.

Zimmermann, M. (1978). *Nature*, **271**, 524.

Zimmermann, M. (1980). *Physical Review D*, **21**, 891.

Zimmermann, M. and Szedenits, E. (1979). *Physical Review D*, **20**, 351.

Zimmerman, M. and Thorne, K. S. (1980). In *Essays in General Relativity*, ed. F. J. Tipler, p. 139. Academic Press: New York.

REF B

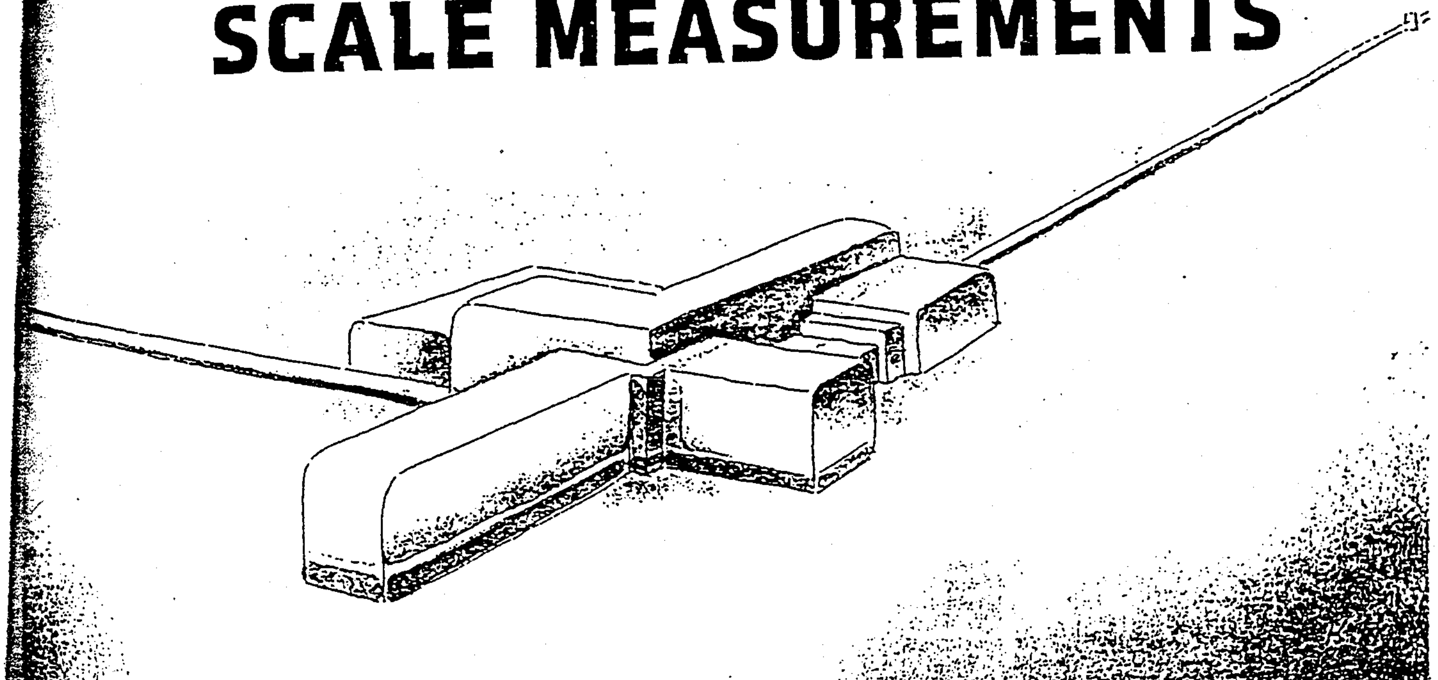
AMERICAN
ASSOCIATION FOR THE
ADVANCEMENT OF
SCIENCE

SCIENCE

17 APRIL 1992
VOL. 256 • PAGES 281–412

$6.00

LARGE SCALE MEASUREMENTS



COVER

Illustration of one site in the proposed Laser Interferometer Gravitational-Wave Observatory (LIGO). The LIGO facilities will consist of two such interferometers located at widely dispersed sites; scientists hope LIGO will be able to detect gravitational waves emanating from collisions of black holes and neutron stars. See page 325. A special section in this issue of *Science* focuses on large scale measurements; see the Editorial and pages 316 to 349. [Illustration: Ruth Sofair Ketler]

LIGO: The Laser Interferometer Gravitational-Wave Observatory

Alex Abramovici, William E. Althouse, Ronald W. P. Drever, Yekta Gürsel, Seiji Kawamura, Frederick J. Raab, David Shoemaker, Lisa Sievers, Robert E. Spero, Kip S. Thorne, Rochus E. Vogt, Rainer Weiss, Stanley E. Whitcomb, Michael E. Zucker

The goal of the Laser Interferometer Gravitational-Wave Observatory (LIGO) Project is to detect and study astrophysical gravitational waves and use data from them for research in physics and astronomy. LIGO will support studies concerning the nature and nonlinear dynamics of gravity, the structures of black holes, and the equation of state of nuclear matter. It will also measure the masses, birth rates, collisions, and distributions of black holes and neutron stars in the universe and probe the cores of supernovae and the very early universe. The technology for LIGO has been developed during the past 20 years. Construction will begin in 1992, and under the present schedule, LIGO's gravitational-wave searches will begin in 1998.

Einstein's general relativity theory describes gravity as due to a curvature of space-time (*1*). When the curvature is weak, it produces the familiar Newtonian gravity that governs the solar system. When the curvature is strong, however, it should behave in a radically different, highly nonlinear way. According to general relativity, the nonlinearity creates black holes (curvature produces curvature without the aid of any matter), governs their structure, and holds them together against disruption (*2*). Inside a black hole, the curvature should nonlinearly amplify itself to produce a space-time singularity (*2*), and near some singularities the nonlinearity should force the curvature to evolve chaotically (*3*). When an object's curvature varies rapidly (for example, because of pulsations, colli-

The authors are the members of the LIGO Science Steering Group. A. Abramovici, W. E. Althouse (Chief Engineer), R. W. P. Drever, S. Kawamura, F. J. Raab, L. Sievers, R. E. Spero, K. S. Thorne, R. E. Vogt (Director), S. E. Whitcomb (Deputy Director), and M. E. Zucker are with the California Institute of Technology, Pasadena, CA 91125. Y. Gürsel is at the Jet Propulsion Laboratory, Pasadena, CA 91109. D. Shoemaker and R. Weiss are at the Massachusetts Institute of Technology, Cambridge, MA 02129.

sions, or rapid orbital motions), it should emanate curvature ripples (gravitational waves) that propagate through the universe at the speed of light and carry the imprint of gravity's quantum-mechanical particle, the graviton.

This description of gravity is almost entirely theoretical and untested. Gravitational waves have never been observed directly [although the orbital decay of neutron star PSR 1913+16 shows indirectly that they exist and carry energy away from binary stars at the rate predicted by general relativity (4)]. Our only observational evidence for the nonlinearity of space-time curvature is the observation of tiny perturbations of planetary and binary star orbits (5), and these teach us nothing about the rich variety of nonlinear curvature phenomena that we expect from experience with other nonlinear systems. Moreover, although astronomers have found strong circumstantial evidence for the existence of black holes and astrophysicists invoke them to help explain astronomical observations, electromagnetic radiation has not yet brought to Earth a clean, unequivocal signature saying "I come from a black hole," and the observations have not been able to be used to test any of the predicted properties of black holes.

LIGO offers an opportunity to bring nonlinear gravity, black holes, and the graviton out of their near isolation as theoretical constructs and into confrontation with experiment. LIGO can verify that these ripples of curvature (gravitational waves) exist. From the force patterns of the waves, LIGO can allow researchers to infer the graviton's spin (6, 7). From the difference in arrival times of electromagnetic- and gravitational-wave bursts from the same distant event, LIGO can allow researchers to infer the difference in the speeds of gravitational waves and light—a difference that must be zero if (as theory predicts) the graviton has zero rest mass (7).

The shapes of a gravitational wave's oscillations (its waveforms; see Fig. 1) carry detailed information about its source, information that LIGO will extract for use in physics research. For example, the waveforms from a small black hole spiraling into a large black hole carry an unequivocal "black hole" signature, a signature that maps out, in detail, a portion of the large hole's space-time geometry (8, 9); by comparing that map with general relativity's predictions, researchers will test highly nonlinear aspects of general relativity. The waveforms from colliding black holes, when compared to those from supercomputer simulations, will give insight into the poorly understood nonlinear dynamics of gravity (10). The waveforms from colliding neutron stars or from neutron stars being torn apart by the tidal gravity of companion black holes may reveal the mass-radius relation for neutron stars, which in turn will give information about the equation of state of nuclear matter (11).

Astrophysical sources have oscillation periods ranging from many hours to less than 1 ms. LIGO is designed to detect only those signals with oscillation periods faster than about 100 ms; slower oscillations are difficult to detect with Earth-based systems because of low-frequency background disturbances. Fortunately, many potential sources have oscillation periods within LIGO's range, including neutron stars, black holes with masses of up to 10^4 solar masses, supernovae cores, and the big bang.

One can best appreciate LIGO's potential for major contributions to astronomy by recalling the history of radio astronomy (12). Before Jansky's discovery of cosmic radio waves in 1932, the universe, as viewed solely through visible light, seemed serene and quiescent, dominated by slowly evolving stars. Radio waves revolutionized this view; they revealed our universe's violent side: pulsars, quasars, active galactic nuclei, and jets that power huge intergalactic clouds of magnetized plasma.

The radio revolution was spectacula because the information carried by radi waves is so different from that carried b light. Light, with its submicrometer wavelength, is emitted mostly by thermally excited atoms in the atmospheres of stars radio waves, with their ten million–fold a large wavelengths, are emitted mostly b high-energy electrons spiraling in the magnetic fields of pulsars, quasars, or jets.

This difference between light and radi waves pales in comparison with the contras between electromagnetic waves and gravitational waves. Electromagnetic astronom usually monitors incoherent superposition of radiation from individual electrons, atoms, or molecules; gravitational waves ar produced most strongly by coherent, bul motions of huge amounts of mass—eithe material mass or the mass-energy of nonlinear space-time curvature (7, 13). Electromagnetic waves are easily absorbed an scattered by matter; gravitational wave travel nearly unchanged through all form and amounts of intervening matter (7).

Compared to electromagnetic telescopes, LIGO is sensitive to very differen aspects of the universe. Electromagneti telescopes study such things as stellar atmospheres, interstellar gas and dust, and primordial gas; LIGO is insensitive to thes LIGO will seek waves from the final inspir and coalescence of binary black holes an neutron stars, the rapidly spinning cores o supernovae, and the first fraction of a second of the big bang (14); to these, electromagnetic telescopes have little or no sensitivity.

These differences produce both uncertainty and great expectations. It is hard t predict, from our present electromagnetic based knowledge, just how sensitive detec

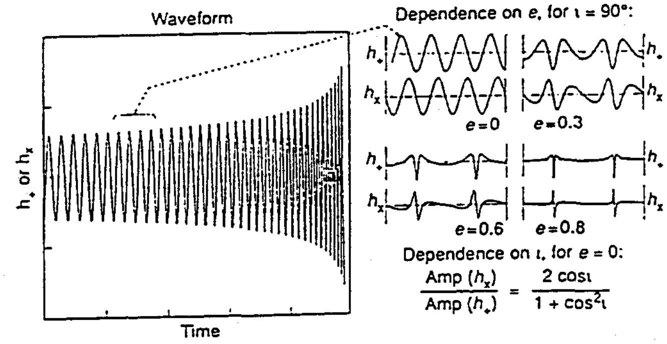


Fig. 1. *An example of gravitational waveforms and the information they carry.* Each gravitational wave has two waveforms, dimensionless functions of time called $h_+(t)$ and $h_\times(t)$. The specific waveforms shown here (*30*) are from the last few minutes or seconds of the spiraling together of a compact binary system (one made of two black holes, two neutron stars, or a black hole and a neutron star). By monitoring these waveforms, LIGO can allow researchers to determine (*30*) the binary's distance from Earth r, the masses of its two bodies or, equivalently, their total mass M and reduced mass μ, and their orbital eccentricity e and orbital inclination to the line of sight ι. To allow the determination of the eccentricity e, LIGO will measure the shapes of the individual waveform oscillations; note the shapes shown on the upper right. For the determination of ι (when $e = 0$ for pedagogic simplicity), LIGO will measure the ratio of the amplitudes, h_+ and $h_\times$; see the formula in the lower right. The parameters r, μ, and M determine (i) the waveforms' absolute amplitudes as they sweep past a frequency f: $h_{\rm amp} \propto \mu M^{2/3} r^{-1} f^{2/3}$; and (ii) the number of cycles $n = f^2(df/dt)^{-1}$ that the waveforms spend near frequency f: $n \propto (\mu M^{2/3} f^{5/3})^{-1}$. From $h_{\rm amp}$ and n, LIGO can be used to determine r and $\mu M^{2/3}$. From $\mu M^{2/3}$, and from late-time post-Newtonian facets of the waveform (*31*) (not shown here) or the frequency at which the inspiral terminates or both, LIGO can be used to deduce the individual values of μ and M (*8*). The simple inspiral waves shown here are modified at late times by post-Newtonian (*31*) and then fully relativistic (*8, 9*) effects and then are followed by much more complicated waveforms from the final collision or tidal disruption of the black holes or neutron stars. It is these final relativistic, collision, and disruption waveforms that will bring LIGO the most interesting information.

tors in LIGO must be to detect their first waves; but once LIGO sees waves, it will bring information about the universe that we have little hope of gaining in any other way. LIGO will teach us about the universe of strongly gravitating objects, such as the masses, birth rates, and spatial distributions of black hole and neutron star binaries in the distant universe, and perhaps about the shapes of rapidly spinning neutron stars in our own galaxy, the spectrum of "cosmic strings," and the first fraction of a second of the universe's expansion (*14*). LIGO may well bring surprises that rival those of radio astronomy.

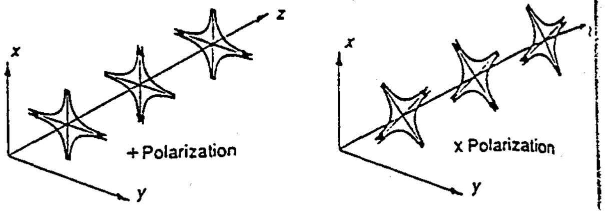


Fig. 2. A general-relativistic gravitational wave propagating in the *z* direction squeezes and stretches the separation of test particles in a plane perpendicular to the *z* axis. The wave acts by a combination of its two polarizations: + ("plus") polarization pushes test particles together along the *x* direction and pushes them apart along the *y* direction when $h_+(t)$ is positive and it reverses the forces when $h_+(t)$ is negative; × ("cross") polarization pushes and pulls test particles, as determined by the sign of $h_\times(t)$, at 45° angles from the *x* and *y* axes.

Laser Interferometer Gravitational-Wave Detectors

According to general relativity theory, every freely moving particle (called a test particle) travels through space-time along a geodesic, a path that is the analog of a straight line in curved space. Just as the curvature of Earth's surface pushes Earth's lines of constant longitude (its geodesics) together as one travels from the equator toward the North or South Pole, so also the curvature of space-time pushes neighboring space-time geodesics together or apart and thereby pushes test particles moving along the geodesics together or apart (*1*).

If the curvature is that of a gravitational wave, then the test particles' relative motion is a sum of contributions from two different polarizations (Fig. 2), each with its own time-evolving waveform: $h_+(t)$ and $h_\times(t)$. This relative motion of test particles is the foundation for several different types of gravitational-wave detectors (*13*, *15*), most notably bar detectors (*16*) and laser interferometer detectors (interferometers for short) (*17*).

An interferometer detector uses four test masses hung by wires near the vertex and ends of an "L" (Fig. 3). The separation L_1 between the two test masses along the first arm is nearly the same as that (L_2) along the second arm, $L_1 = L_2 = L$. At frequencies above their pendular swing frequency (about 1 Hz), the test masses move freely horizontally. A gravitational wave (of the appropriate polarization) incident perpendicular to the plane of the interferometer pushes the masses back and forth relative to each other, stretching one arm while squeezing the other, and thereby changing the arm-length difference $\Delta L \equiv L_1 - L_2$. For other directions of incidence, the fractional difference in arm length caused by the wave, $\Delta L/L$, is equal to a linear combination of the two polarization waveforms (*13*),

$$\frac{\Delta L(t)}{L} = F_+h_+(t) + F_\times h_\times(t) \equiv h(t) \quad (1)$$

The coefficients F_+ and $F_\times$ are of order unity and depend on the direction to the source and the orientation of the detector. We call $h(t)$ the gravitational-wave strain that acts on the detector. Notice that the relative motion of the test masses caused by the wave is proportional to their initial separation, one of the fundamental facts that drives the design of LIGO.

Laser interferometry is used to monitor ΔL, and thence the gravitational-wave strain $h(t) = \Delta L/L$. In LIGO's first interferometers, the interferometry (*18*) will be performed as follows (Fig. 3): One face of each test mass is polished and coated to form a mirror with high reflectivity, low transmissivity, and very low scattering and absorption. The two mirrors along each arm form a Fabry-Perot resonant optical "cavity," which gives the effect of having the light traverse the arm many times. A laser beam shines onto a beam splitter at the vertex of the L, and the splitter directs half of the light along each arm, exciting the two Fabry-Perot cavities. The end mirror of each cavity has much lower transmissivity than the mirror near the vertex, so light from each excited cavity exits through its vertex mirror and back toward the beam splitter. The splitter is adjusted to recombine the two returning beams so that most of the recombined light returns toward the laser and a tiny portion propagates toward the photodetector.

When a gravitational wave changes the length L_1 or L_2 of one of the cavities, it slightly shifts the cavity's resonant frequency relative to the laser frequency and thereby changes the phase of the light in the cavity and the phase of the light that exits from the cavity toward the beam splitter. Because the wave affects the two arms differently, it shifts the relative phases of the light exiting the two cavities and thereby alters their interference at the beam splitter, causing a slight change in the intensity at the photodetector. This change in photodetector signal is proportional to $\Delta L(t)$, and thence to the gravitational-wave strain $h(t)$.

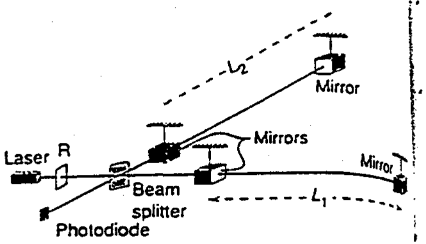


Fig. 3. A schematic view of a LIGO interferometer.

The photodetector signal is highly sensitive to $h(t)$, as the following order-of-magnitude calculation shows. The relative phase change between the light emerging from the two cavities is

$$\Delta\Phi = B\frac{\Delta L}{\lambda} = B\frac{hL}{\lambda} \quad (2)$$

where λ is the light's wavelength and B is the mean number of times the light bounces back and forth in the cavities before exiting (proportional to the cavities' finesse). The phase change $\Delta\Phi$ can be monitored at the photodetector output with a precision that is limited by photon shot noise, that is, by randomness in the arrival times of the photons; the limit is $\Delta\Phi \approx 1/\sqrt{N}$, where N is the number of photons incident on the beam splitter during the time (roughly a gravitational-wave period) that the photodetector's signal is being integrated. Correspondingly, a gravitational-wave strain that would give a signal of the same magnitude as that of the measurement fluctuations is

$$h_{\min} \approx \frac{\lambda}{L}\frac{1}{B}\frac{1}{\sqrt{N}} \quad (3)$$

Actually, $\Delta\Phi$ is proportional to B and $h_{\min}$ is proportional to $1/B$ only if the mean light storage time in the cavities, BL/c, is less than half of a gravitational-wave period (c, speed of light); if B is made larger, there is no further improvement of $h_{\min}$. For example, the limit is $B \approx 400$, assuming 100-Hz waves and an arm length $L = 4$ km. Given the quality of the LIGO mirrors, only 1% of the photons will be lost to scattering and absorption in the 400 bounces, so almost all of the stored light will exit back toward the laser. This light will be recycled (*18*) back into the interferometer by a high-reflectivity mirror placed at the location marked R in Fig. 3. (This makes the entire interferometer a single resonant cavity with arms that are subcavities.) The total power available in the interferometer for making the measurement will then be ~100 times as great as the laser's output power. For a laser output of 60 W, this means that $N \approx 2 \times 10^{20}$ photons are incident on the beam splitter during a 10-ms photodetector integration time. Correspondingly, for the above parameters the minimum detectable wave has a strength (Eq. 3)

$$h_{\min} \approx \left(\frac{0.5\ \mu\text{m}}{4\ \text{km}}\right)\left(\frac{1}{400}\right)$$
$$\left(\frac{1}{\sqrt{2 \times 10^{20}}}\right) \approx 10^{-23} \qquad (4)$$

As shown below in Fig. 10, this sensitivity should be sufficient for the detection of large numbers of gravitational-wave sources and the use of these sources for a rich program of physics and astronomy research.

Several other optical configurations are possible for the interferometer and may be used in future LIGO detectors (*19*). For example, by inserting a light-recycling mirror between the beam splitter and the photodetector, one can greatly improve the interferometer's sensitivity in a narrow frequency band, while degrading it outside that band (*20*).

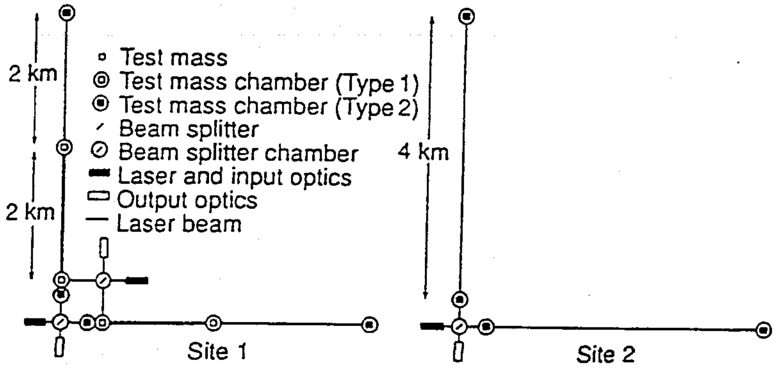


Fig. 4. Schematic layout of the initial LIGO facilities. For each interferometer, the laser beam is conditioned in an input optics chain before it reaches the beam splitter. After passing through the beam splitter, the beams enter the chambers containing the test masses, where they are directed into a 2-km or 4-km interferometer cavity. After leaving the cavity, the beams, now back in the test mass chambers, are directed back to the beam splitter and then into an output optics chain that terminates with the photodetector. At site 1, all elements of the 4-km interferometer lie along the two arms, whereas the beam splitter and input and output optics chains for the 2-km interferometer lie between the two arms.

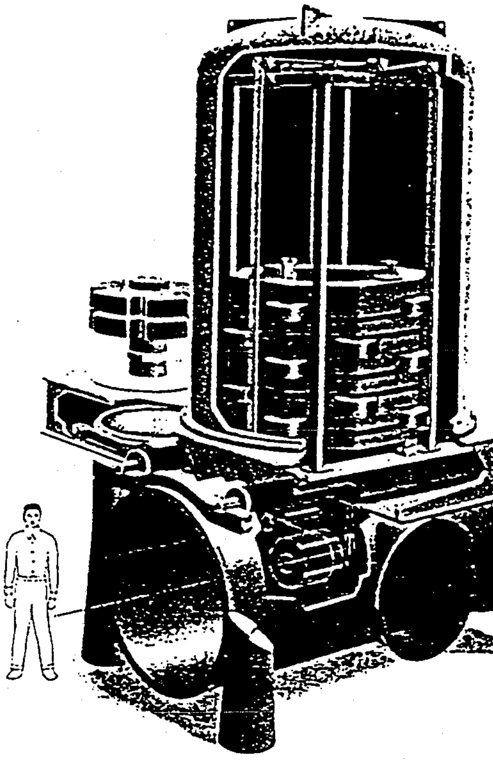


Fig. 5. An artist's conception of a type 1 LIGO test mass chamber (see the symbols in Figs. 4 and 6), that is, the vacuum system module that houses an interferometer's test masses. The vertical cylinder serves as an air lock that can be opened to the outside from above or, with a horizontal gate valve at its base, opened to the main vacuum pipe below. The large assemblage in the air lock is a passive vibration isolation system (a cascaded stack of mechanical filters consisting of masses and elastomer springs), from which are suspended the test mass and a steering mirror that deflects the light beam from the beam splitter to the test mass and back. The upper laser beam is an auxiliary that monitors the separation between test mass suspension points, so feedback can maintain a fixed separation, thereby helping with vibration isolation (*27*). The passive vibration isolation system and its suspended test mass and mirror can be raised into the air lock and the gate valve can be closed to permit modifications without interfering with the main vacuum or with other interferometers.

LIGO

LIGO will be a facility open to the national community and capable of housing many successive generations of interferometers with a variety of optical designs. The principal features of LIGO are dictated by the following considerations:

1) Each interferometer's test masse must be housed in a vacuum to avoic buffeting by air molecules. The optical path must also be in vacuum to prevent fluctuations in the number of air molecules in th beam from causing fluctuations in th light's phase. The most sensitive LIGC interferometers will require a vacuum o 10^{-9} torr.

2) An interferometer's sensitivity improves as its arm length L is increased Achieving sensitivities adequate for th expected waves (Fig. 10 below) require arms several kilometers in length. LIGC has been designed with 4-km-long arms.

3) The vacuum pipe running betwee the test masses of each arm will have diameter of 1.2 m so that it can accommo date multiple detectors as well as auxiliar laser beams required in some advanced detectors (*21*).

4) To firmly distinguish real gravitational waves from "non-Gaussian" bursts c instrumental and environmental noise, th outputs of interferometers at two widel separated sites will have to be correlated Three interferometers will be used: a singl 4-km interferometer at one site and tw interferometers, 4 and 2 km long, sharin the same vacuum system (*22*) at the oth site. If, instead, LIGO were to have only single vacuum system at a single site, nois bursts would probably prevent its interferometers from recording any meaningf data whatsoever.

5) Each interferometer's test mass must be suspended from vibration isolatio systems that protect them from seismic an acoustic vibrations, and the test masse vacuum chambers must be large enough house these isolation systems.

6) Before entering the interferomete the laser light must be conditioned in variety of ways: it must be frequency-sta

lized, phase-modulated, amplitude-stabilized, and spatially shaped to control various spurious noise sources. This conditioning requires a long input optics chain housed in special vacuum chambers. A similar output optics chain must be placed between the interferometer and the photodetector.

Figure 4 shows a schematic layout of LIGO's two sites. At site 1 the two interferometers (one 4 km, the other 2 km) are interleaved in such a way that either can be removed from the vacuum system without interfering with the other and without breaking the main vacuum. Figure 5 shows how this capability is designed in the vacuum chambers housing the test masses.

The two LIGO facilities and their first three-interferometer detector system will be constructed from 1992 through 1996 at a cost of ~$200 million. Subsequent detector systems will cost several million dollars each. The LIGO project is implemented by the "LIGO team," a group of scientists and engineers at the California Institute of Technology (Caltech) and the Massachusetts Institute of Technology, and important contributions are coming from groups at other institutions, including the University of Colorado, Stanford University, and Syracuse University, and from industry.

LIGO's initial configuration is the minimum that can house a three-interferometer detector system capable of detecting the predicted waves and monitoring one of their two waveforms. This initial configuration has been designed to permit an upgrade (presumably after gravitational waves have been detected) into the configuration shown in Fig. 6. The upgraded LIGO can house three independent detector systems that operate simultaneously. These detector systems might be in different stages of development or might be optimized for different types of gravitational waveforms, for example, broadband bursts from black hole collisions or monochromatic waves from pulsars in some chosen narrow-frequency band. From time to time one of the detector systems can be removed and a new one can be inserted in its place, with minimal interference with the other two systems.

Even in this upgraded form, LIGO by itself will not be able to extract all of the information from a gravitational wave [the direction to its source and the two waveforms $h_+(t)$ and $h_\times(t)$]. Full extraction will require combining the outputs of interferometers at three or more widely separated sites, and for all-sky coverage there must be at least four sites (23). LIGO will rely on other nations to provide the third and fourth sites of the network. Vigorous efforts toward doing so are under way in Europe (15) and are being initiated in Japan and Australia (15). The angular resolution of this international network will range from a few arc minutes to a few degrees, depending on the shapes of the waveforms and the signal-to-noise ratio (23). This is comparable to the resolutions of radio telescopes in about 1950, when the first optical identifications of radio sources were being made.

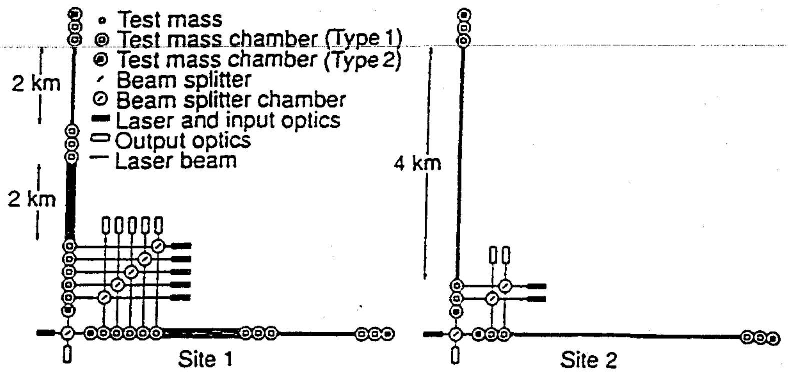


Fig. 6. Schematic layout of the LIGO facilities after a possible future upgrade. Site 1 accommodates three 4-km and three 2-km interferometers without interference, and site 2 accommodates three 4-km interferometers.

LIGO Interferometers and Their Noise

The noise in any interferometer is of two types: Gaussian (noise with a Gaussian probability distribution) and non-Gaussian. Because of the Gaussian distribution's extremely fast falloff with increasing noise amplitude, Gaussian noise is exceedingly unlikely to produce noise bursts larger than a few standard deviations. By contrast, interferometers can show large, non-Gaussian noise bursts several times per hour due, for example, to sudden strain releases in the wires that suspend the test masses. The only sure way to remove such non-Gaussian noise in a LIGO detector system is by correlating the outputs of the system's three interferometers. Once this is done, the system's sensitivity will be governed by the remaining, Gaussian noise.

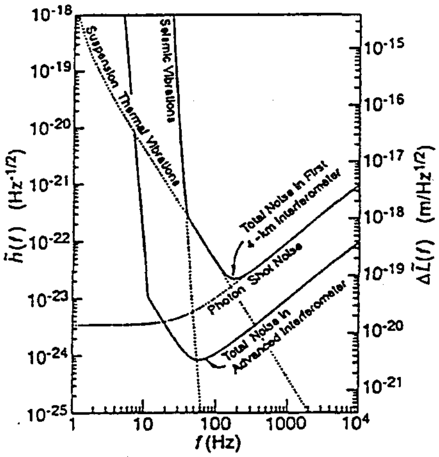


Fig. 7. The expected total noise in each of LIGO's first 4-km interferometers (upper solid curve) and in a more advanced interferometer (lower solid curve). The dashed curves show various contributions to the first interferometer's noise.

The Gaussian noise is characterized by a spectrum $\tilde{h}(f)$ defined as follows. An interferometer's output consists of the true gravitational-wave strain $h(t)$ (Eq. 1) plus the Gaussian noise $h_{\text{noise}}(t)$; $\tilde{h}(f)$ is the square root of the power spectral density of $h_{\text{noise}}(t)$ at frequency f. When the interferometer measures a gravitational-wave burst (such as from an inspiraling black hole binary) that has a strain amplitude h_{amp}, a characteristic (mean) frequency f_c, and a duration of n cycles, the measurement, obtained with optimal filter techniques, will have the signal-to-noise ratio (13)

$$\frac{S}{N} \simeq \frac{h_c}{h_{\text{rms}}} \qquad (5)$$

where

$$h_c \simeq h_{\text{amp}} \sqrt{n} \qquad (6)$$

is called the wave's characteristic amplitude, and

$$h_{\text{rms}} \equiv \sqrt{f_c}\,\tilde{h}(f_c) \qquad (7)$$

is the interferometer's root mean square (rms) noise for a one-cycle-long burst at the source's characteristic frequency f_c (24). [Equations 6 and 7 can be regarded as an approximate definition of the interferometer's noise spectrum $\tilde{h}(f)$.]

The LIGO team has developed tentative design parameters for the first LIGO interferometers. (The design will not be finalized until 1993.) A guiding philosophy for this design is that it should use only current technology. These first interferometers are expected to have a Gaussian noise spectrum $\tilde{h}(f)$ depicted by the upper solid curve in

Fig. 7. The individual contributions to this noise, shown as dashed curves, arise from the following sources:

1) Below ~70 Hz, the total noise (upper solid curve) will be dominated by "seismic noise" (ground vibrations due to the seismic background, to man-made sources such as traffic on roads or railroads, or to wind forces coupled to the ground by trees and buildings), which is transmitted to the test masses through their suspensions. The test masses will be protected from such noise by a passive seismic isolation system (Fig. 5) and the final pendulum suspension. The seismic noise limit is very steep because such a vibration isolation system loses its effectiveness at low frequencies.

2) Between ~70 and ~200 Hz, the noise is predominately due to off-resonance, thermally induced vibrations of the test masses and their suspensions. These thermal vibrations will produce the noise shown in Fig. 7, assuming 10-kg test masses and a 10^7 quality factor for their pendulum suspensions.

3) Above ~200 Hz, photon shot noise will dominate. This is the type of noise discussed in the text preceding Eq. 3. The curve shown in Fig. 7 is based on the assumptions of 2 W of effective laser power [comparable to that of existing argon ion lasers or to frequency-doubled neodymium:yttrium-aluminum-garnet (Nd:YAG) lasers that are being developed by the LIGO team's Stanford collaborators], mirrors with a fractional power loss of 5×10^{-5} per reflection (which has been demonstrated in smaller mirrors, but not yet in the 20-cm size required for LIGO), and a factor of 30 gain in effective laser power by light recycling [comparable with what has been demonstrated in interferometers with fixed mirrors (25)].

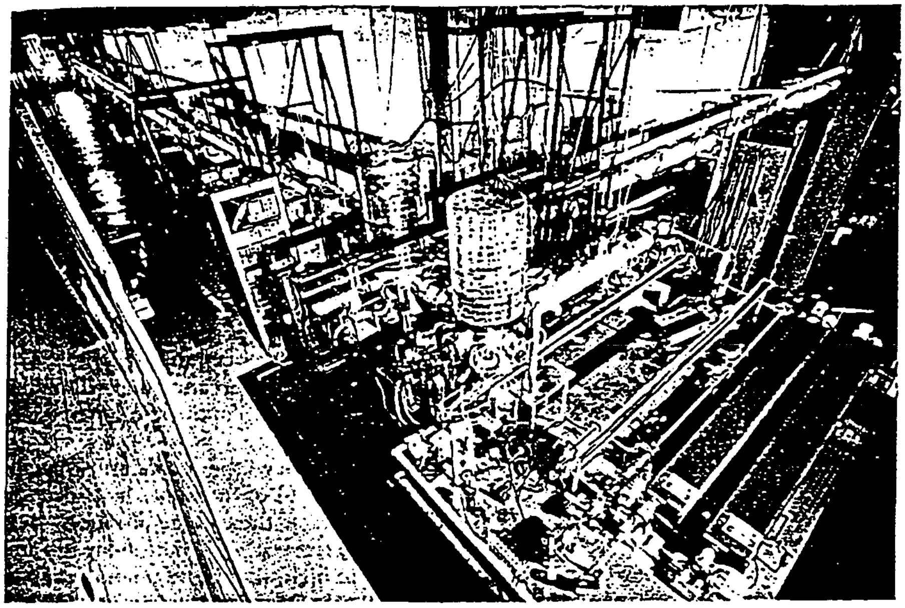


Fig. 8. A photograph of the 40-m prototype interferometer at Caltech.

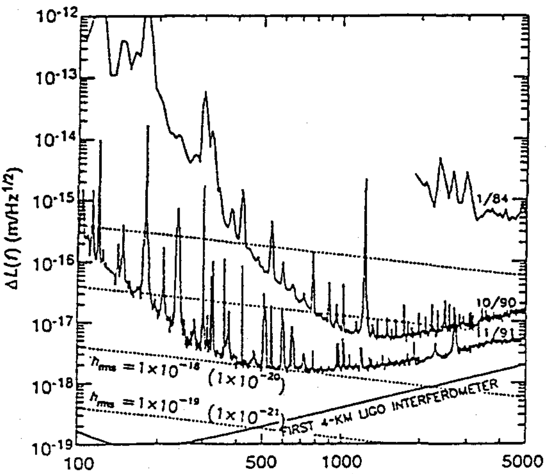


Fig. 9. The displacement noise achieved by 40-m prototype interferometers at various times in the past, and comparison with the estimated displacement noise in the first LIGO interferometers (see Fig. 7 and Eq. 8). The lines of constant h_{rms} were computed from Eqs. 7 and 8 with the prototype arm length, L = 40 m (or, in parentheses, the LIGO arm length, L = 4 km).

A number of other noise sources must be controlled; all are expected to be less important than these three.

Throughout LIGO's operation, its interferometers will be continually improved. The lower solid curve of Fig. 7 depicts the noise in a candidate next-generation advanced interferometer. (The technology and techniques for this and other advanced interferometers are now under development.) This advanced interferometer's noise reduction would be largely achieved in three ways: (i) To reduce the photon shot noise, the effective laser power would be increased to 60 W and the mirror losses would be improved to permit recycling the light 100 rather than 30 times. (The resulting rms shot noise, $h_{rms} = \sqrt{f}\,\tilde{h}(f)$ at f = 100 Hz, is 10^{-23}, in accord with the estimate in Eq. 4.) (ii) To reduce the pendulum suspensions' thermal vibrations, their quality factor would be increased from 10^7 to 10^9 by a change in their material and geometry, and the test masses would be increased from 10 to 1000 kg. (iii) To reduce the seismic vibrations, the isolation stack would be upgraded, and an active isolation system (26, 27) might be used.

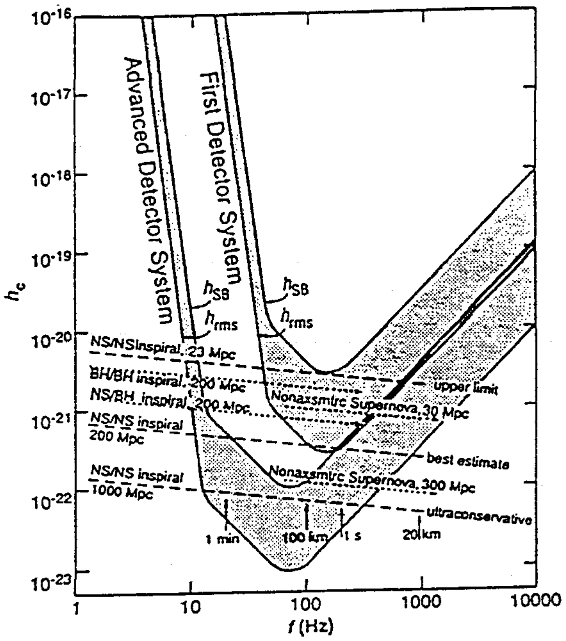
Prototype Interferometer Development

The designs of the first and advanced LIGO interferometers and the confidence that their expected noise levels can be achieved are the result of extensive experiments carried out over the past decade in the LIGO team's laboratories and elsewhere. These experiments have separately tested various performance requirements and interferometer components. The predicted performance of LIGO interferometers must be pieced together from these individual experiments without any unified demonstration, because laboratory-scale interferometers are significantly different from 4-km LIGO interferometers—different because various noise sources scale differently with increasing arm length and changing frequency.

One of the key instruments being used in the development and testing of components and techniques for full-scale LIGO interferometers is a 40-m prototype at Caltech (Fig. 8). Experiments with this prototype (28) have tested a variety of interferometer components and a wide range of their performance measures; Fig. 9 shows measurements of displacement noise, $\Delta\tilde{L}(f)$. [$\Delta\tilde{L}(f)$ is the square root of the power spectral density of ΔL and is related to $\tilde{h}(f)$ by

$$\Delta\tilde{L}(f) = L\tilde{h}(f) \qquad (8)$$

Fig. 10. Comparison between the expected rms noise h_{rms} in LIGO's first and advanced detector systems and the characteristic amplitude h_c of gravitational-wave bursts from several hypothesized sources. NS, neutron star; BH, black hole.



See Eqs. 1 and 7.] The bottom spectrum in Fig. 9 is the prototype's displacement performance in November 1991. Long-term prototype progress is shown by a comparison with the January 1984 spectrum, and recent shorter term progress is shown by a comparison with the October 1990 spectrum. Also shown is the displacement noise expected in LIGO's first 4-km interferometers.

The 1991 spectrum shows a number of peaks superimposed on a smooth background. Some of the low-frequency peaks are due to imperfections in electromechanical servos, and the largest peaks at higher frequency are due to mechanical resonances in an intermediate reference cavity. All of these peaks can be removed with further work, leaving only peaks due to thermally excited vibrational modes of the test masses' suspension wires (the clusters of peaks near 330, 500, and 650 Hz). Even in its present state, the spectrum's baseline level, not the peaks, is the more accurate measure of the detector's displacement sensitivity, because narrow peaks can be filtered out in the data analysis with minimal loss in signal.

The baseline noise is due to a number of different sources. At high frequencies (above ~1 kHz), the noise is dominated by shot noise and is consistent with the level predicted for the current laser power and interferometer optical configuration. At very low frequencies (below ~120 Hz), the observed noise is due to ground vibrations that couple through the suspensions to the test masses. The level of this noise component is somewhat higher than the LIGO goal, in part because of the ground vibration level in the Caltech laboratory, which is a factor of 10 higher than measured levels at remote sites more representative of probable LIGO sites. In the intermediate region, the noise spectrum contains contributions from a number of sources, the relative importance of which varies with frequency. In addition to seismic noise and shot noise, there are significant contributions from thermal noise and from scattered light that recombines and interferes with the main laser beam. All of these noise sources can be reduced with further work. For example, the level of noise due to scattered light is highly dependent on the detailed configuration of the interferometer; because of space constraints within its current vacuum chamber, the 40-m prototype does not contain many of the measures planned for the first LIGO interferometers to control scattered light.

Experiments performed on the 40-m prototype and other special purpose setups have (i) tested and confirmed the photon shot noise formula used to predict the noise spectra of LIGO interferometers; (ii) tested the formula for the noise from residual gas and confirmed the vacuum requirement in the beam tubes; (iii) demonstrated the effective power enhancement due to recycling in a Fabry-Perot interferometer (25); and (iv) allowed the development of a technique for laser stabilization (29) and verified that sufficient frequency stability can be achieved for the first LIGO interferometers.

Taken together, these separate experiments, and others performed in other laboratory facilities, give us confidence that the fundamental noise sources are understood and that the LIGO noise performance goals of Fig. 7 can be achieved. However, a final demonstration of all aspects of the LIGO detectors' performance must await the 1996 installation of the first detector system in the full-scale LIGO facilities.

Comparison of LIGO Sensitivities with Estimated Wave Strengths

LIGO's first detector system might see gravitational waves, and the advanced detectors discussed above are highly likely to see them. Success will probably come between the first-detector level and the advanced level, that is, a few years after LIGO goes into operation.

This prediction is based largely on the best understood of the hypothesized sources: the final, minute-long inspiral of a neutron star binary. Because the 10-km-sized stars are 100 km apart when LIGO is most sensitive to them, they are not yet tearing each other apart, and the details of their wave emission are understood (13, 30, 31). The uncertainty in the waves' strength arises solely from the uncertain distance to the nearest such sources. The observed statistics of binary neutron stars in our own galaxy, extrapolated to include distant galaxies, give a best estimate (32, 33) of 200 Mpc (650 million light years) for the distance to which LIGO must look to see three neutron star inspirals per year. Further analysis (32) of the uncertainties in the data gives an "ultraconservative upper limit" of 1000 Mpc and an "optimistic lower limit" of ~23 Mpc.

Figure 10 shows the characteristic amplitudes h_c (Eq. 6) of the waves from neutron star binaries at these three distances (34). The waves sweep with time from low frequency to high, that is, from left to right in Fig. 10 (see waveform in Fig. 1). As the frequency increases, the waves' amplitude h_{amp} increases; however, the number of cycles $n = f^2(df/dt)^{-1}$ spent near each frequency f decreases, and the characteristic amplitude $h_c = h_{amp}\sqrt{n}$ (which determines the signal-to-noise ratio) decreases slightly (see dashed arrows in Fig. 10). (The vertical arrows along the bottom of Fig. 10 mark the remaining times, 1 min and 1 s, until the final neutron star collision, and the distances, 100 km and 20 km, between the stars.)

Figure 10 allows a comparison of these signal strengths with the expected noise in the first and advanced LIGO detectors. The lower curve of each pair is the rms noise h_{rms} (Eq. 7) in each 4-km interferometer, as computed from the noise estimates of Fig. 7. The upper curve in each pair is the sensitivity to bursts,

$$h_{SB} = 11 h_{rms} \tag{9}$$

that is, the strength h_c that a burst must have—if it arrives only rarely, from a random direction, with an arbitrary polarization, and at a random, unpredictable time—in order for us to be highly confident that it is not due to Gaussian noise (35). [If the neutron stars' final collision produces a strong enough burst of gamma rays for detection at Earth (36), then the gamma burst will dictate the time at which to expect the gravitational-wave burst, and the waves can then be identified with confidence at $h_c = h_{SB}/3$, that is, a factor of 3 below the upper curve in each pair.]

Figure 10 shows that LIGO's first detectors will be about good enough to detect three neutron star inspiral events per year at the "optimistic" level, and the advanced detectors will be about good enough at the "ultraconservative" level. Most likely, the first detection will be somewhere in between.

The first detection can be used for waveform studies (Fig. 1), even if it barely exceeds the detectors' burst sensitivity h_{SB}, because h_{SB} is far above the interferometers' rms noise (35). Each factor of 2 sensitivity improvement thereafter not only will improve the accuracies of waveform studies, but also will increase the rate of observed signals by a factor $2^3 = 8$, because the event rate scales with volume, that is, is proportional to h_{SB}^{-3}. Correspondingly, if three neutron star inspirals occur each year at 200 Mpc (the best estimate), then the rate will be about one event every 2 to 3 days in the advanced detector system.

Figure 10 also shows, at 200 Mpc [the best current estimate for how far LIGO must look to see three such events per year (32, 33)], the waves from a neutron star spiraling into one ~10–solar mass black hole and from two ~10–solar mass black holes spiraling together. Although the three-per-year distance for these sources is much less certain than that for neutron star binaries, the waves at a given distance are stronger—so strong, in fact, that the advanced detector system could see the inspiral of black hole binaries throughout the universe (out to cosmological distances, where the event rate should be ~10,000 times as high as that at the best-estimate distance of 200 Mpc, or several events per hour).

Figure 1 shows how LIGO can allow researchers to infer, from the inspiral waveforms, the details of the binary's orbits and the masses of its black holes or neutron stars. In the last second of the binary's life, its inspiral waveforms gradually give way to collisional, merger, or tidal-disruption waveforms (depending on the nature of the binary), from which one can extract fundamental physics—(i) from the final collision of comparable-mass holes: details of the poorly understood, highly nonlinear dynamics of space-time curvature (10); (ii) from the inspiral of a small hole into a large hole: a detailed, partial map of the large hole's curvature (8, 9); (iii) from the tidal disruption of a neutron star by its companion black hole: the neutron star's radius (which, combined with its measured mass, will give information about the equation of state of nuclear matter); and (iv) from the collision of two neutron stars: their collisional dynamics and perhaps their radii. For two neutron stars, the collisional waveform will be concentrated at frequencies of 500 to 1000 Hz, where the detectors' photon shot noise is severe; so studying the collision will require specialized detectors with enhanced high-frequency sensitivity (and consequently reduced low-frequency performance). For binaries containing two black holes more massive than about ten suns, by contrast, the final waveform may be at low enough frequencies for study with the broadband, advanced detectors of Fig. 10.

Of course, coalescing binaries are not the only potential gravitational-wave sources for LIGO. There are others (14): rotating, slightly nonaxisymmetric neutron stars; collapsing stellar cores (as for example, in supernovae); stochastic waves from vibrating cosmic strings and from the big bang; and, of course, totally unexpected types of sources. However, these other sources are all uncertain in wave strength, event rate, or both.

One example is a nonaxisymmetric supernova. Type II supernovae are known to be triggered by the collapse of a star's core to form a neutron star (37). If the collapse is axially symmetric, then the gravitational waves emitted will be too weak to detect beyond our galaxy and the Magellanic Clouds (38). On the other hand, if the core is spinning sufficiently fast, then centrifugal forces may halt its collapse at radii of several hundred kilometers, and an instability then might [but theory has not yet confirmed this (39)] drive the flattened, spinning core into a nonaxisymmetric shape, so it tumbles like a football turning end over end and emits strong gravitational waves. If it is the gravitational waves, and not hydrodynamic waves, that carry off most of the core's excess angular momentum, thereby permitting it to shrink to neutron star size (~10 km), and if 10% of type II supernovae are triggered by such collapses, then the distance to which one must look to see three such events per year is ~30 Mpc (40) and the wave strength is roughly that shown in Fig. 10 (41). Reducing the fraction of type II supernovae that undergo such collapses by a thousand (to 0.01%) would increase the distance one must look to 300 Mpc. It may well be that such nonaxisymmetric supernovae never occur in nature, or they may occur so rarely that LIGO will never see one. However, there is an observational hint of strong asymmetry in the recent discovery of a neutron star that seems to have been ejected at 1/100th of the speed of light from the center of its supernova explosion (42).

This discussion of nonaxisymmetric supernovae, with all its "ifs" and "mights," illustrates the enormous uncertainties that plague estimates of the gravitational waves from most astrophysical objects. The greatest hope for resolving the uncertainties is to search for the waves, and, when they are found, to study their waveforms.

Conclusions

The fiscal year 1992 National Science Foundation budget contains the first portion of LIGO's ~$200-million construction cost. The selection of the two LIGO sites is now entering its final stage, and the sites should be in hand and the construction begun at the first site by the end of 1992. If future funding is granted at the planned rate, construction at the two sites will be completed in 1995 and 1996, the facilities will have been checked out and begun operation by the end of 1997, and the first detector system will be operational in 1998.

This first detector system may discover gravitational waves. If not, experimenters will press forward with detector improvements (for which development is already under way), leading toward LIGO's advanced detector goals. These improvements are expected to lead to the detection of waves from many sources each year. The scientific community can then begin to harvest the rich information carried by the waves, and an upgrade of LIGO can make it possible for several research groups simultaneously to operate several different detector systems, each optimized for a different type of astrophysical source.

Note added in proof: While this paper was in press, the National Science Foundation selected the two LIGO sites from among 19 proposals. The selected sites are Livingston Louisiana, and Hanford, Washington.

REFERENCES AND NOTES

1. See, for example, C. W. Misner, K. S. Thorne, J. A. Wheeler, *Gravitation* (Freeman, San Francisco 1973).
2. See, for example, I. D. Novikov and V. P. Frolov *Physics of Black Holes* (Kluwer Academic, Dordrecht, Netherlands, 1989), and references therein.
3. See, for example, J. D. Barrow, *Phys. Rep.* 85, (1982).
4. J. H. Taylor and J. M. Weisberg, *Astrophys.* 345, 434 (1989), and references therein.
5. See, for example, C. M. Will, *Theory and Experiment in Gravitational Physics* (Cambridge Uni Press, Cambridge, 1981).

6. D. M. Eardley *et al.*, *Phys. Rev. Lett.* 30, 884 (1973); D. M. Eardley, D. L. Lee, A. P. Lightman, *Phys. Rev. D* 8, 3308 (1973).
7. K. S. Thorne, in *Gravitational Radiation*, N. Deruelle and T. Piran, Eds. (North-Holland, Dordrecht, Netherlands, 1983), pp. 1–54.
8. L. S. Finn, A. Ori, K. S. Thorne, in preparation.
9. S. L. Detweiler, *Astrophys. J.* 225, 687 (1978); ——— and E. Szedenits, *ibid.* 231, 211 (1979); T. Nakamura, K. Oohara, Y. Kojima, *Prog. Theor. Phys.* (*suppl.*) 90, 1 (1987).
10. See, for example, C. R. Evans, L. S. Finn, D. W. Hobill, Eds., *Frontiers in Numerical Relativity* (Cambridge Univ. Press, Cambridge, 1989); G. B. Cook and J. W. York, *Phys. Rev. D* 41, 1077 (1989); J. Bowen, J. Rauber, J. W. York, *Class. Quantum Gravity* 1, 591 (1984); L. Smarr, *Ann. N.Y. Acad. Sci.* 301, 569 (1977).
11. See, for example, S. L. Shapiro and S. A. Teukolsky, *Black Holes, White Dwarfs, and Neutron Stars* (Wiley, New York, 1983).
12. For the history of radio astronomy and the revolution it brought, see K. Kellermann and B. Sheets, Eds., *Serendipitous Discoveries in Radio Astronomy* (National Radio Astronomy Observatory, Green Bank, WV, 1983); W. T. Sullivan, Ed., *Classics in Radio Astronomy* (Kluwer, Dordrecht, Netherlands, 1982); *The Early Years of Radio Astronomy* (Cambridge Univ. Press, Cambridge, 1984).
13. K. S. Thorne, in *300 Years of Gravitation*, S. W. Hawking and W. Israel, Eds. (Cambridge Univ. Press, Cambridge, 1987), pp. 330–458.
14. For discussions and lists of references, see (*13*) and K. S. Thorne, in *Recent Advances in General Relativity*, A. Janis and J. Porter, Eds. (Birkhauser, Boston, 1992), pp. 196–229.
15. D. G. Blair, Ed., *The Detection of Gravitational Waves* (Cambridge Univ. Press, Cambridge, 1991).
16. Bar detectors were invented by J. Weber [*Phys. Rev.* 117, 306 (1960)] and have been developed to a high degree of sophistication since then; see, for example, P. F. Michelson, J. C. Price, R. C. Taber, *Science* 237, 150 (1987); (*15*).
17. The seeds of the idea for laser interferometer gravitational-wave detectors are contained in F. A. E. Pirani, *Acta Phys. Pol.* 15, 389 (1956), in M. E. Gertsenshtein and V. I. Pustovoit, *Sov. Phys. JETP* 16, 433 (1963), and in J. Weber's personal notebooks from the 1960s; the first detailed descriptions of the idea were by G. E. Moss, L. R. Miller, and R. L. Forward [*Appl. Opt.* 10, 2495 (1971)] and by R. Weiss [*Quart. Prog. Rep. Res. Lab. Electron. M.I.T.* 105, 54 (1972)].
18. R. W. P. Drever, in *Gravitational Radiation*, N. Deruelle and T. Piran, Eds. (North-Holland, Dordrecht, Netherlands, 1983), pp. 321–338; R. W. P. Drever *et al.*, in *Quantum Optics, Experimental Gravity, and Measurement Theory*, P. Meystre and M. O. Scully, Eds. (Plenum, New York, 1983), pp. 503–514, and references therein.
19. For example, the two arms can be operated as optical delay lines, with the bouncing light in them making many discrete spots on the mirrors instead of being trapped in a Fabry-Perot cavity [R. Weiss, in (*17*)]. This configuration is now being developed in Europe (*15*).
20. B. J. Meers, *Phys. Rev. D* 38, 2317 (1988); K. A. Strain and B. J. Meers, *Phys. Rev. Lett.* 66, 1391 (1991); and references therein.
21. The many laser beams in each arm may include (i) up to six simultaneous Fabry-Perot cavity beams after an upgrade of LIGO's corner and end stations, (ii) beams connecting the suspension points of the test masses (*27*) as part of seismic isolation systems (see Fig. 5 for an example); and (iii) the many individual beams of a single interferometer with arms that are optical delay lines rather than Fabry-Perot cavities (*19*).
22. Some noise sources will be common to the vacuum-sharing 4- and 2-km interferometers. However, many will not be common, and those that are will typically not mimic a true gravitational wave's 2:1 ratio of signal strengths.
23. Y. Gürsel and M. Tinto, *Phys. Rev. D* 40, 3884 (1990).
24. Similarly, when used in the study of periodic waves (such as from a pulsar) with amplitude h_{amp} and frequency *f*, the interferometer will have the signal-to-noise ratio (*13*)

$$\frac{S}{N}=\frac{h_{amp}\sqrt{\tau_{int}}}{\tilde{h}(f)}$$

where τ_{int} is the total time over which the periodic signal is integrated.
25. P. Fritschel, D. Shoemaker, R. Weiss, *Appl. Opt.*, in press.
26. R. L. Rinker and J. E. Faller, in "Precision instruments and fundamental constants II," B. N. Taylor and W. D. Phillips, Eds., *National Bureau of Standards Spec. Pub.* 617, 411 (1984); N. A. Robertson *et al.*, *J. Phys. E* 15, 1101 (1982); P. R. Saulson, *Rev. Sci. Instrum.* 15, 1315 (1984); N. A. Robertson, in *The Detection of Gravitational Waves*, D. Blair, Ed. (Cambridge Univ. Press, Cambridge, 1990), pp. 353–368.
27. R. W. P. Drever, in *The Detection of Gravitational Waves*, D. Blair, Ed. (Cambridge Univ. Press, Cambridge, 1990), pp. 306–328.
28. A technical description of the prototype and experiments that have been carried out with it will be published soon.
29. R. W. P. Drever *et al.*, *Appl. Phys.* B31, 97 (1983).
30. B. F. Schutz, *Nature* 323, 310 (1986); *Class. Quantum Gravity* 6, 1761 (1989); H. D. Wahlquist, *Gen. Relativ. Gravit.* 19, 1101 (1987).
31. C. W. Lincoln and C. M. Will, *Phys. Rev. D* 42, 1123 (1990).
32. E. S. Phinney, *Astrophys. J.* 380, L17 (1991).
33. R. Narayan *et al.*, *ibid.* 379, L17 (1991).
34. From equation 46b of (*7*), with a $\sqrt{2}$ correction added because of the incorrect use of *ft* rather than $\int(df/dt)dt$ in equation 42 of (*7*), and with neutron star masses of 1.4 solar masses.
35. Because the burst is far out on the tail of the noise's Gaussian probability distribution, h_{SB} is quite insensitive to whether the time between bursts is 1 month or 1 year and to whether by "highly confident" one means 90% or 99%. The factor 11 in Eq. 9 is made up of (i) a factor of $\sqrt{5}$ that is included because, owing to the detectors' quadrupolar beam pattern, waves from a random (typical) direction produce a signal $\sqrt{5}$ times as small as that from an optimal direction and (ii) a factor of 5 from the Gaussian statistics for two identical 4-km interferometers that search for events of duration ~0.01 s that occur a few times per year. See (*13*) for a more careful analysis and discussion; in that reference h_{SB} is called $h_{3/yr}$.
36. P. Haensel, B. Paczynski, P. Amsterdamski, *Astrophys. J.* 375, 209 (1991).
37. See, for example, S. E. Woosley and T. A. Weaver, *Annu. Rev. Astron. Astrophys.* 24, 205 (1986).
38. R. Mönchmeyer, G. Schäfer, E. Müller, R. E. Kates, *Astron. Astrophys.* 246, 417 (1991); L. S. Finn, *Ann. N.Y. Acad. Sci.* 631, 156 (1991).
39. See, for example, the following references for discussion of this so-called "bar mode instability" in various contexts: S. Chandrasekhar, *Ellipsoidal Figures of Equilibrium* (Yale Univ. Press, New Haven, CT, 1969); J. R. Ipser and L. Lindblom, *Phys. Rev. Lett.* 62, 2777 (1989); *Astrophys. J.* 355, 226 (1990).
40. For the rates of supernova explosions, see S. van den Bergh and G. A. Tammann, *Annu. Rev. Astron. Astrophys.* 29, 363 (1991).
41. This wave strength h_c is independent of the strength of the instability and the resulting degree of nonaxisymmetry as long as the gravitational waves are responsible for carrying away the core's angular momentum. For weaker instabilities, the waves' amplitude h_{amp} is smaller and the *number of cycles n* that the waves spend near a frequency *f* is larger; the product $h_c \approx h_{amp}\sqrt{n}$ is constant.
42. D. A. Frail and S. R. Kulkarni, *Nature* 352, 785 (1991).
43. This work was supported in part by NSF grant PHY-8803557.

REF C

Published in Gravitational Radiation, eds. N. Deruelle & T. Piran (North Holland, Amsterdam, 1983) pp 1-57.

OAP-647

THE THEORY OF GRAVITATIONAL RADIATION: AN INTRODUCTORY REVIEW*

KIP S. THORNE

W. K. Kellogg Radiation Laboratory
California Institute of Technology, Pasadena, California 91125
and
Institute for Theoretical Physics
University of California, Santa Barbara, California 93106

ABSTRACT

This is the written version of lectures presented at the Workshop on Gravitational Radiation, Ecole d'Ete de Physique Theorique, Les Houches, France, June 1982.

These lectures are an introduction to and a progress report on the effort to bring gravitational-wave theory into a form suitable for astrophysical studies — a form for use in the future, when waves have been detected and are being interpreted.

Much of the viewpoint embodied in these lectures I have adopted or developed since 1972 when Misner, Wheeler, and I completed Gravitation. If I were rewriting the gravitational-wave parts of Gravitation today, I would do so along the lines of these lectures.

*Supported in part by the National Science Foundation [AST79-22012 and PHY77-27084].

ONE OF THE ORANGE AID PREPRINT SERIES
IN NUCLEAR, ATOMIC & THEORETICAL ASTROPHYSICS

September 1982

TABLE OF CONTENTS

1. INTRODUCTION AND OVERVIEW

1.1 The nature of these lectures

These lectures are an introduction to and a progress report on the effort to bring gravitational-wave theory into a form suitable for astrophysical studies — a form for use in the future, when waves have been detected and are being interpreted. I will not describe all aspects of this effort. Several of the most important aspects will be covered by other lecturers, elsewhere in this volume. These include the computation of waves from models of specific astrophysical sources (lectures of Eardley); the techniques of numerical relativity — our only way of computing waves from high-speed, strong-gravity, large-amplitude sources (lectures of York and Piran); and a full analysis of radiation reaction and other relativistic effects in binary systems such as the binary pulsar — our sole source today of quantitative observational data on the effects of gravitational waves (lectures of Damour).

My own lectures will provide a sort of framework for those of Eardley, York and Piran, and Damour: I shall present the mathematical description of gravitational waves in a form suitable for astrophysical applications (§2); I shall describe a variety of methods for computing the gravitational waves emitted by astrophysical sources (§3); I shall describe methods for analyzing the propagation of waves from their sources, through our lumpy universe, to earth (§2); and I shall describe methods of analyzing the interaction of gravitational waves with earth-based and solar-system-based detectors (§4). Here and there in my lectures I shall sketch derivations of the methods of analysis and of the formulas presented; but in most places I shall simply refer the reader to derivations elsewhere in the literature and/or pose the derivations as exercises for the reader.

1.2 What is a gravitational wave?

A gravitational wave is a ripple in the curvature of spacetime, which propagates with the speed of light (Fig. 1). In the real universe gravitational waves propagate on the back of a large-scale, slowly changing spacetime curvature created by the universe's lumpy, cosmological distribution of matter. The background curvature is characterized, semiquantitatively, by two length scales

$$\mathcal{R} \equiv \begin{pmatrix}\text{radius of curvature}\\ \text{of background spacetime}\end{pmatrix} \equiv \left|\begin{matrix}\text{typical component } R_{\hat\alpha\hat\beta\hat\gamma\hat\delta} \text{ of Rieman}\\ \text{tensor of background in a local Lorentz frame}\end{matrix}\right|^{-\frac{1}{2}} ,$$

$$\mathcal{L} \equiv \begin{pmatrix}\text{inhomogeneity scale of}\\ \text{background curvature}\end{pmatrix} \equiv \begin{pmatrix}\text{length scale on which}\\ R_{\hat\alpha\hat\beta\hat\gamma\hat\delta} \text{ varies}\end{pmatrix} \lesssim \mathcal{R} ; \qquad (1.1)$$

and the gravitational waves are characterized by one length scale

$$ƛ \equiv \begin{pmatrix}\text{reduced wavelength of}\\ \text{gravitational waves}\end{pmatrix} = \frac{1}{2\pi} \times (\text{wavelength } \lambda). \qquad (1.2)$$

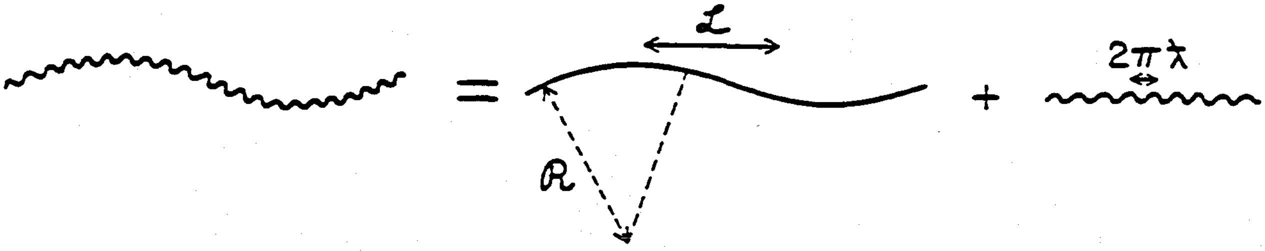


Fig. 1 A heuristic embedding diagram for the decomposition of curve spacetime into a background spacetime plus gravitational waves.

(Of course $\bar{\lambda}$, $\mathcal{L}$, and $\mathfrak{R}$ are not precisely defined; they depend on one's choice of coordinates or reference frame. But in typical astrophysical situations there are preferred frames – e.g. the "asymptotic rest frame" of the source of the waves, or the "mean local rest frame" of nearby galaxies; and these permit $\bar{\lambda}$, $\mathcal{L}$, and $\mathfrak{R}$ to be defined with adequate precision for astrophysical discussion.) The separation of spacetime curvature into a background part $R^{(B)}_{\alpha\beta\gamma\delta}$ and a wave part $R^{(W)}_{\alpha\beta\gamma\delta}$ depends critically on the inequality

$$\bar{\lambda} \ll \mathcal{L} . \tag{1.3}$$

The waves are the part that varies on the lengthscale $\bar{\lambda}$; the background is the part that varies on the scale $\mathcal{L}$; the separation is impossible if $\bar{\lambda} \sim \mathcal{L}$. See Figure 1.

In constructing the theory of gravitational waves one typically expands the equations of general relativity in powers of $\bar{\lambda}/\mathcal{L}$ and $\bar{\lambda}/\mathfrak{R}$. In the real universe these expansions constitute perturbation theory of the background spacetime (these lectures and that of Yvonne Choquet). In an idealized universe consisting of a source surrounded by vacuum (so that $\mathcal{L} \equiv r$ = distance to source) these expansions constitute "asymptotic analyses of spacetime structure near future timelike infinity $\mathscr{I}^{+}$" (lectures of Martin Walker).

1.3 Regions of space around a source of gravitational waves

I shall characterize any source of gravitational waves, semiquantitatively, by the following length scales as measured in the source's "asymptotic rest frame":

$$L \equiv \begin{pmatrix}\text{size of}\\ \text{source}\end{pmatrix} = \begin{pmatrix}\text{radius of region inside which the stress-energy } T^{\alpha\beta}\\ \text{and all black-hole horizons are contained}\end{pmatrix},$$

$$2M \equiv \begin{pmatrix}\text{gravitational}\\ \text{radius of source}\end{pmatrix} = \begin{pmatrix}2\times\text{mass of source in}\\ \text{units where } G = c = 1\end{pmatrix},$$

$$\bar{\lambda} \equiv (\text{reduced wavelength of the waves emitted}), \tag{1.4}$$

$$\left.\begin{array}{l} r_I \equiv (\text{inner radius of local wave zone}) \\ \\ r_O \equiv (\text{outer radius of local wave zone}) \end{array}\right\} (\text{see below}).$$

Corresponding to these length scales, I shall divide space around a source into the following regions (Fig. 2):

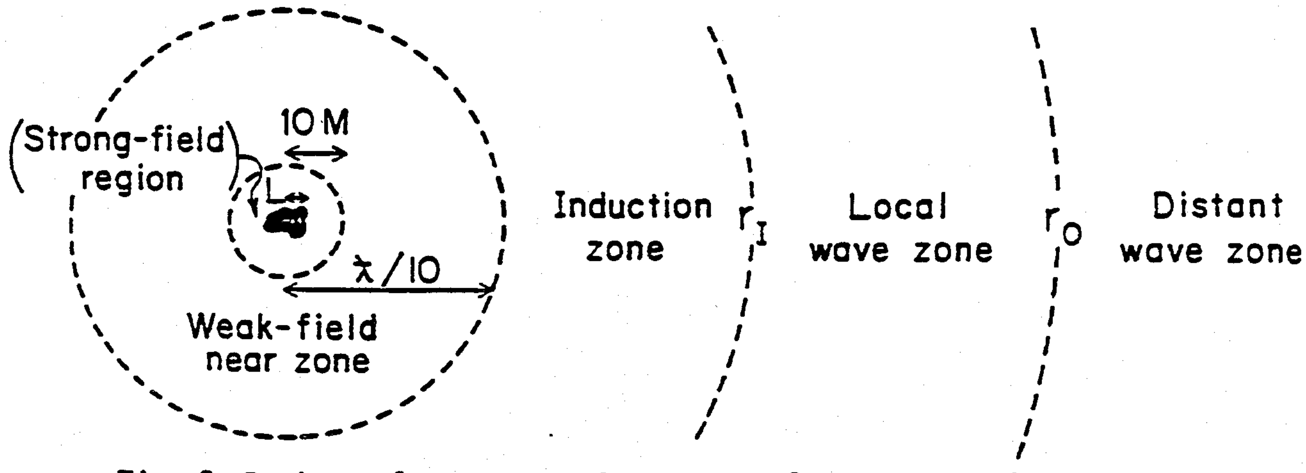


Fig. 2 Regions of space around a source of gravitational waves.

$$\begin{array}{ll} \text{Source:} & r \lesssim L, \\ \text{Strong-field region:} & r \lesssim 10\,M \text{ if } 10\,M \gtrsim L, \\ & \text{typically does not exist if } L \gg 10\,M, \\ \text{Weak-field near zone:} & L < r,\ 10\,M < r < \bar{\lambda}/10, \\ \text{Induction zone:} & L < r,\ \bar{\lambda}/10 < r < r_I \\ \text{Local wave zone:} & r_I < r < r_O \\ \text{Distant wave zone:} & r_O < r. \end{array} \qquad (1.5)$$

Although Figure 2 suggests the lengthscale ordering $L < 10\,M < \bar{\lambda}/10$, no such ordering will be assumed in these lectures. Thus, we might have $\bar{\lambda} \gg L$ and M ("slow-motion source"), or $\bar{\lambda} \ll L$ and M (high-frequency waves from some small piece of a big source; weak field near zone does not exist), or $\bar{\lambda} \sim L$ or M; and we might have $L \gg M$ or perhaps $L \sim M$.

At radius r outside the source ($r > L$) the background curvature due to the source has lengthscales

$$\mathcal{R}_s \simeq (r^3/M)^{\frac{1}{2}} \quad , \quad \mathcal{L}_s \simeq r. \qquad (1.6)$$

Consequently, the dynamically changing part of the curvature can be regarded as "gravitational waves" (i.e. has $\bar{\lambda} \ll \mathcal{L}$) only in the "wave zone" $r \gg \bar{\lambda}$. I split the wave zone up into two parts, the local wave zone and the distant wave zone, so as to facilitate a clean separation of two mathematical problems: the generation of waves by the source, and the propagation of those waves through the lumpy, real universe to earth. The local wave zone ($r_I \lesssim r \lesssim r_O$) will serve as a matching region for the two problems: the theory of wave generation will cover the local wave zone and all regions interior to it; the theory of wave propagation will cover the local wave zone and its exterior.

To facilitate the matching I shall choose r_I and r_O in such a manner that throughout the local wave zone the background curvature can be ignored and the background metric can thus be approximated as that of flat Minkowskii spacetime. More specifically, the inner edge of the local wave zone (r_I) is the location at which one or more of the following effects becomes important: (i) the waves cease to be waves and become a near-zone field, i.e., r becomes $\lesssim \bar{\lambda}$; (ii) the gravitational pull of the source produces a significant red shift, i.e., r becomes $\sim 2M =$ (Schwarzschild radius of source); (iii) the background curvature produced by the source distorts the wave fronts and backscatters the waves significantly, i.e., $(r^3/M)^{1/2}$ becomes $\lesssim \bar{\lambda}$; (iv) the outer limits of the source itself are encountered, i.e., r becomes $\lesssim L =$ (size of source). Thus, the inner edge of the local wave zone is given by

$$r_I = \alpha \times \max\{\bar{\lambda}, 2M, (M\bar{\lambda}^2)^{1/3}, L\}\,, \qquad (1.7)$$
$$\alpha \equiv \left(\begin{array}{c}\text{some suitable number} \\ \text{large compared to unity}\end{array}\right).$$

The outer edge of the local wave zone r_O is the location at which one or more of the following effects becomes important: (i) a significant phase shift has been produced by the "M/r" gravitational field of the source, i.e., $(M/\bar{\lambda}) \times \ln(r/r_I)$ is no longer $\ll \pi$; (ii) the background curvature due to nearby masses or due to the external universe perturbs the propagation of the waves, i.e., r is no longer $\ll \mathcal{R}_u =$ (background radius of curvature of universe). Thus, the outer edge of the local wave zone is given by

$$r_0 = \min[r_I \exp(\lambdabar/\beta M),\ \mathcal{R}_u/\gamma], \qquad (1.8)$$

$$\beta,\gamma = \begin{pmatrix}\text{some suitable numbers}\\ \text{large compared to unity}\end{pmatrix}.$$

Of course, we require that our large numbers α,β,γ be adjusted so that the thickness of the local wave zone is very large compared to the reduced wavelength:

$$r_0 - r_I \gg \lambdabar . \qquad (1.9)$$

In complex situations the location of the local wave zone might not be obvious. Consider, for example, a neutron star passing very near a supermassive black hole. The tidal pull of the hole sets the neutron star into oscillation, and the star's oscillations produce gravitational waves [Mashhoon (1973); Turner (1977)]. If the hole is large enough, or if the star is far enough from it, there may exist a local wave zone around the star which does not also enclose the entire hole. Of greater interest – because more radiation will be produced – is the case where the star is very near the hole and the hole is small enough ($M_h \lesssim 100\ M_\odot$) to produce large-amplitude oscillations, and perhaps even disrupt the star. In this case, before the waves can escape the influence of the star, they get perturbed by the background curvature of the hole. One must then consider the entire star-hole system as the source, and construct a local wave zone that surrounds them both.

* * * * *

Exercise 1. Convince yourself that for all astrophysical sources except the big-bang singularity (e.g., for the neutron-star/black-hole source of the last paragraph) α, β, and γ can be so chosen as to make condition (1.9) true.

1.4 Organization of these lectures: Notation and conventions

Section 2 of these lectures will discuss the propagation of gravitational waves from the local wave zone out through our lumpy universe to the earth. Section 3 will discuss the generation of gravitational waves, including their propagation into the local wave zone where they can be matched onto the propagation theory of Section 2. Section 4 will discuss the detection of gravitational waves on earth and in the solar system.

My notation and conventions are those of Misner, Thorne, and Wheeler (1973) (cited henceforth as "MTW"): I use geometrized units ($c = G = 1$); Greek indices range from 0 to 3 (time and space), Latin indices from 1 to 3 (space only); the metric signature is +2; $\eta_{\alpha\beta} \equiv \text{diag}(-1,1,1,1)$ is the Minkowskii metric; $\epsilon_{\alpha\beta\gamma\delta}$ and ϵ_{ijk} are the spacetime and space Levi-Civita tensors with $\epsilon_{0123} > 0$ in a right-hand-oriented basis; and the signs of the Riemann, Ricci, Einstein, and stress-energy tensors are given by

$$R^{\alpha}{}_{\beta\gamma\delta} = \Gamma^{\alpha}{}_{\beta\delta,\gamma} - \Gamma^{\alpha}{}_{\beta\gamma,\delta} + \text{"}\Gamma\Gamma\text{"} - \text{"}\Gamma\Gamma\text{"}\ , \qquad R_{\alpha\beta} \equiv R^{\mu}{}_{\alpha\mu\beta}\ , \qquad (1.10)$$

$$G_{\alpha\beta} \equiv R_{\alpha\beta} - \tfrac{1}{2} R g_{\alpha\beta} = 8\pi\, T_{\alpha\beta}\ , \qquad T^{00} > 0 \quad .$$

Much of the viewpoint embodied in these lectures I have adopted or developed since 1972 when Misner, Wheeler and I completed MTW. However, many of the new aspects of my viewpoint are contained in my 1975 Erice lectures (Thorne 1977) and/or in a recent Reviews of Modern Physics article (Thorne 1980a; cited henceforth as "RMP").

2. THE PROPAGATION OF GRAVITATIONAL WAVES

2.1 Gravitational waves in metric theories of gravity: Description and propagation speed

Gravitational waves are not unique to Einstein's theory of gravity. They must exist in any theory which incorporates some sort of local Lorentz invariance into its gravitational laws. Many such theories have been invented; see, e.g., Will (1982) for examples and references.

Among the alternative theories of gravity there is a wide class — the so-called "metric theories" — whose members are so similar to general relativity that a discussion of their gravitational waves brings the waves of Einstein's theory into clearer perspective. Thus, I shall initiate my discussion of wave propagation within the framework of an arbitrary metric theory, and then shall specialize to Einstein's general relativity.

A metric theory of gravity is a theory (i) in which gravity is characterized, at least in part, by a 4-dimensional, symmetric spacetime metric $g_{\alpha\beta}$ of signature +2; and (ii) in which the Einstein equivalence principle is satisfied — i.e., all the nongravitational laws of physics take on their standard special relativistic forms in the local Lorentz frames of $g_{\alpha\beta}$ (aside from familiar complications of "curvature coupling"; chapter 16 of MTW).

Examples of metric theories are: general relativity [$g_{\alpha\beta}$ is the sole gravitational field]; the Dicke-Brans-Jordan theory (e.g., Dicke 1964) [contains a scalar gravitational field Φ in addition to $g_{\alpha\beta}$; matter generates Φ via a curved-spacetime wave equation; then Φ and the matter jointly generate $g_{\alpha\beta}$ via Einstein-like field equations]; and Rosen's (1973) theory [a "bimetric" theory with a flat metric $\eta_{\alpha\beta}$ in addition to the physical metric $g_{\alpha\beta}$; the matter generates $g_{\alpha\beta}$ via a flat-spacetime wave equation whose characteristics are null lines of $\eta_{\alpha\beta}$]. See Will (1982) for further details, references, and other examples.

The Einstein equivalence principle guarantees that in any metric theory, as in general relativity, freely falling test particles move along geodesics of $g_{\alpha\beta}$, and that the separation vector ξ^α between two nearby test particles (separation << $\lambda\!\!\!^{-}$) is governed by the equation of geodesic deviation:

$$D^2\xi^\alpha/d\tau^2 + R^\alpha{}_{\beta\gamma\delta}u^\beta\xi^\gamma u^\delta = 0\,. \qquad (2.1a)$$

Here u^α is the 4-velocity of one of the test particles; τ is proper time measured by that particle;

$$D^2\xi^\alpha/d\tau^2 \equiv (\xi^\alpha{}_{;\beta}u^\beta)_{;\gamma}u^\gamma \qquad (2.1b)$$

is the relative acceleration of the particles; and $R^\alpha{}_{\beta\gamma\delta}$ is the Riemann curvature tensor associated with $g_{\alpha\beta}$. Throughout Sections 2 (wave propagation) and 3 (wave generation) I shall use geodesic deviation and the Riemann tensor to characterize the physical effects of gravitational waves. Only in Section 4 (wave detection) will I discuss other physical effects of waves.

The Riemann tensor $R_{\alpha\beta\gamma\delta}$ contains two parts: background curvature and wave curvature

$$R_{\alpha\beta\gamma\delta} = R^{(B)}_{\alpha\beta\gamma\delta} + R^{(W)}_{\alpha\beta\gamma\delta}\,. \qquad (2.2)$$

As discussed in §1.2 $R^{(B)}_{\alpha\beta\gamma\delta}$ varies on a long lengthscale $\mathcal{L}$, while $R^{(W)}_{\alpha\beta\gamma\delta}$ varies on a short lengthscale $\lambda\!\!\!^{-}$. Consequently, if by $\langle\ \rangle$ we denote an average over spacetime regions somewhat larger than $\lambda\!\!\!^{-}$ but much smaller than $\mathcal{L}$ ("Brill-Hartle average";

Exercise 35.14 of MTW), then we can write

$$R^{(B)}_{\alpha\beta\gamma\delta} \equiv \langle R_{\alpha\beta\gamma\delta}\rangle \ , \quad R^{(W)}_{\alpha\beta\gamma\delta} \equiv R_{\alpha\beta\gamma\delta} - \langle R_{\alpha\beta\gamma\delta}\rangle \ . \tag{2.3a}$$

Similarly we can define the background metric, of which $R^{(B)}_{\alpha\beta\gamma\delta}$ is the Riemann tensor, by

$$g^{(B)}_{\alpha\beta} \equiv \langle g_{\alpha\beta}\rangle \ . \tag{2.3b}$$

[For a discussion of delicacies which require the use of "steady coordinates" in the averaging of $g_{\alpha\beta}$ see Isaacson (1968), or more briefly §35.13 of MTW.]

In general relativity and in the Dicke-Brans-Jordan theory gravitational waves propagating through vacuum are governed by the wave equation

$$R^{(W)}_{\alpha\beta\gamma\delta|\mu\nu}\, g^{\mu\nu}_{(B)} = 0 \ , \tag{2.4}$$

whereas in Rosen's theory they are governed by

$$R^{(W)}_{\alpha\beta\gamma\delta,\mu\nu}\, \eta^{\mu\nu} = 0 \ . \tag{2.5}$$

Here "|" denotes covariant derivative with respect to $g^{\mu\nu}_{(B)}$ while "," denotes covariant derivative with respect to the flat matric $\eta^{\mu\nu}$. These equations imply that in general relativity and in Dicke-Brans-Jordan theory gravitational waves propagate through vacuum with precisely the speed of light, $c_{GW} = c_{EM}$, but in Rosen's theory they propagate with a different speed, $c_{GW} \neq c_{EM}$. As a rough rule of thumb, whenever a theory of gravity possesses "prior geometry" such as a flat auxiliary metric (MTW, §17.6), it will have $c_{GW} \neq c_{EM}$; often when there is no prior geometry, $c_{GW} = c_{EM}$.

High-precision experiments to test $c_{GW} = c_{EM}$ will be possible if and when electromagnetic waves and gravitational waves are observed from outbursts in the same distant source. For example, for a supernova in the Virgo cluster of galaxies (about 4×10^7 light years from earth; distance at which several supernovae are seen each year) one can hope to discover the light outburst within one day (of retarded time) after the explosion is triggered by gravitational collapse. If gravitational waves from the collapse are observed, then a test is possible with precision

$$\left|\frac{\Delta c}{c}\right| = \left|\frac{c_{GW} - c_{EM}}{c}\right| \sim \frac{1 \text{ day}}{4 \times 10^7 \text{ yr}} \simeq 1 \times 10^{-10} \ . \tag{2.6}$$

Actually, there already exists strong observational evidence that gravitational waves do not propagate more slowly than light. If they did, then high-energy cosmic rays with speeds v in the range $c_{GW} < v < c_{EM}$ would emit gravitational Cerenkov radiation very efficiently and would be slowed quickly by gravitational radiation reaction to $v = c_{GW}$. Since cosmic rays are actually detected with v as large as $c_{EM} \times (1 - 10^{-18})$, c_{GW} presumably is no smaller than this. (For further details and for a discussion of whether we really understand gravitational Cerenkov radiation in alternative theories of gravity see Caves (1980); also earlier work by Aichelburg, Ecker, and Sexl (1971).

2.2 Plane waves on a flat background in metric theories with $c_{GW} = c_{EM}$

Henceforth I shall restrict attention either to metric theories that have $c_{GW} = c_{EM}$ always (e.g., general relativity and Dicke-Brans-Jordan); or, for theories like Rosen's, to regions of spacetime where c_{GW} happens to equal c_{EM}.

In this section and the next several, I shall make a further restriction to spacetime regions of size $\ll \mathcal{R}$ (but $\gg \bar{\lambda}$). In such regions with good accuracy I can ignore the curvature of the background; i.e., I can and will introduce global Lorentz frames of the background metric, in which

$$g^{(B)}_{\alpha\beta} = \eta_{\alpha\beta} \, . \tag{2.7}$$

Far from their source gravitational waves will have wave fronts with radii of curvature large compared to $\bar{\lambda}$, i.e., they will be locally plane fronted. Thus, with good accuracy I can and shall approximate $R^{(W)}_{\alpha\beta\gamma\delta} = R_{\alpha\beta\gamma\delta}$ as precisely plane fronted; and by correctly orienting my spatial axes I shall make the waves propagate in the $x^3 = z$ direction. Since they propagate with the speed of light, the waves are then functions of $t - z$:

$$R_{\alpha\beta\gamma\delta} = R_{\alpha\beta\gamma\delta}(t-z). \tag{2.8}$$

The analysis of such waves in arbitrary metric theories of gravity, as described below, is due to Eardley, Lee, Lightman, Wagoner, and Will (1973), cited henceforth as ELLWW. For greater detail see Eardley, Lee, and Lightman (1973).

2.2.1 Bianchi identities and dynamical degrees of freedom

Because the Riemann tensor of any metric theory is derivable from a metric $g_{\alpha\beta}$, it must satisfy the Bianchi identities $R_{\alpha\beta[\gamma\delta;\epsilon]} = 0$. For the plane-wave Riemann tensor (2.8) on a flat background (2.7) the total content of the Bianchi identities is

$$\begin{aligned} R_{\alpha\beta 12,0} &= 0 & &\Rightarrow & R_{\alpha\beta 12} &= 0 \, , \\ R_{\alpha\beta 13,0} - R_{\alpha\beta 10,3} &= 0 & &\Rightarrow & R_{\alpha\beta 13} &= -R_{\alpha\beta 10} \, , \\ R_{\alpha\beta 23,0} - R_{\alpha\beta 20,3} &= 0 & &\Rightarrow & R_{\alpha\beta 23} &= -R_{\alpha\beta 20} \, . \end{aligned} \tag{2.9}$$

Recalling the pair-wise symmetry $R_{\mu\nu\alpha\beta} = R_{\alpha\beta\mu\nu}$ we see from (2.9) that any purely spatial pair of indices (12 or 13 or 23) either vanishes or can be converted into a space-time pair (10 or 20 or 30). This means that the six quantities

$$R_{i0j0}(t-z) = R_{j0i0}(t-z) \tag{2.10}$$

are a complete set of independent components of our plane-wave Riemann tensor. All other components of Riemann can be expressed algebraically in terms of these.

In a general metric theory of gravity these six quantities represent six independent degrees of freedom of the gravitational field — i.e., six independent polarizations of a gravitational wave.

In the special case of general relativity a vacuum gravitational wave must have vanishing Ricci tensor

$$R_{\mu\nu} = R^{\alpha}{}_{\mu\alpha\nu} = 0 \tag{2.11}$$

(Einstein field equations). One can show easily that this reduces the number of independent degrees of freedom from six to two:

$$R_{x0x0} = -R_{y0y0} \qquad \text{and} \qquad R_{x0y0} = R_{y0x0} \, . \tag{2.12}$$

* * * * *

Exercise 2. Show that the Bianchi identities for a plane wave on a flat background imply equations (2.9) and that they, in turn, guarantee that R_{i0j0} are a complete set of independent components of the Riemann tensor.

Exercise 3. Show that the vacuum Einstein field equations (2.11) reduce the independent plane-wave components of Riemann to (2.12).

2.2.2 Local inertial frames (side remarks)

In the next section I shall use geodesic deviation to elucidate the physical nature of the six gravity-wave polarizations. But as a foundation for that discussion I must first remind you of the mathematical and physical details of local inertial frames (LIF); see, e.g., §§8.6, 11.6, and 13.6 of MTW.

Physically an LIF is the closest thing to a global inertial frame that an experimenter can construct. The central building block of an LIF is a freely falling test particle ("fiducial particle"), which carries with itself three orthogonally pointing gyroscopes. The experimenter attaches a Cartesian coordinate grid to the gyroscopes. Because of spacetime curvature, this grid cannot be precisely Cartesian; but deviations from Cartesion structure can be made second order in the spatial distance r from the fiducial particle:

$$g_{\alpha\beta} = \eta_{\alpha\beta} + O(r^2 R_{\mu\nu\rho\sigma}) \; . \tag{2.13}$$

From an experimental viewpoint the details of the $O(r^2 R_{\mu\nu\rho\sigma})$ corrections often are unimportant. Those corrections actually produce geodesic deviation, if one calculates geodesics directly from $g_{\alpha\beta}$; but geodesic deviation is more clearly described as the effect of $R_{\mu\nu\rho\sigma}$ in the geodesic deviation equation (2.1a), which now reads for a test particle at spatial location $x^j = \xi^j$ = (separation from fiducial test particle) and, as always in geodesic deviation, with velocity $|dx^j/dt| << 1$:

$$d^2x^j/dt^2 = -R_{j0k0}\, x^k \; . \tag{2.14}$$

It is this "experimenter's version" of geodesic deviation to which I shall appeal in discussing gravitational waves.

* * * * *

Exercise 4. Show that in an LIF with metric (2.13) the fiducial particle (at rest at the spatial origin) moves along a geodesic. Show further that if $\xi^j = x^j$ is the separation vector between the fiducial test particle and another test particle, the equation of geodesic deviation (2.1a) takes on the form (2.14).

Exercise 5. One realization of an LIF is a "Fermi normal coordinate system" obtained by letting the spatial coordinate axes be spacelike geodesics that start out along the directions of the three gyroscopes. Show that in such a coordinate system

$$\begin{aligned} ds^2 = {} & -dt^2(1+R_{0\ell 0m}x^\ell x^m)dt^2 - \tfrac{4}{3} R_{0\ell 0m}x^\ell x^m dt dx^j \\ & +(\delta_{ij} - \tfrac{1}{3} R_{i\ell jm}x^\ell x^m)dx^i dx^j . \end{aligned} \tag{2.15}$$

For details see, e.g., §13.6 of MTW.

Exercise 6. Show that in the de Donder gauge of §3.1.3 below and in the vacuum of general relativity a mathematical realization of an LIF is

$$ds^2 = -dt^2(1+R_{0\ell 0m}x^\ell x^m)dt^2 - \frac{4}{3}R_{0\ell jm}x^\ell x^m dtdx^j$$
$$+\delta_{ij}(1-R_{0\ell 0m}x^\ell x^m)dx^i dx^j . \tag{2.16}$$

(Note: neither this nor (2.15) requires any assumption of a plane-wave Riemann tensor.) For details see, e.g., Hartle and Thorne (1983).

2.2.3 Physical description of plane-wave polarizations

Consider a cloud of test particles surrounding a central, fiducial test particle. Initially the cloud resides in flat spacetime, all its particles are at rest with respect to each other, and its shape is precisely spherical with radius a. Then a gravitational wave hits and deforms the cloud. The deformations can be analyzed using the equation of geodesic deviation only if the cloud is small compared to the inhomogeneity scale of the Riemann tensor, $a \ll ƛ$. Assume this to be the case, and analyze the cloud's deformations in the LIF of the fiducial particle:

$$d^2x^j/dt^2 = -R_{j0k0}(t)x^k. \tag{2.17}$$

Here $x^j(t)$ is the location, in the LIF, of some specific test particle at time t; and $R_{j0k0}(t-z)$ is evaluated at the fiducial particle's location $(x,y,z) = 0$. Consider, in turn, and with the help of Figure 3, the six independent polarizations of the wave (further details in ELLWW):

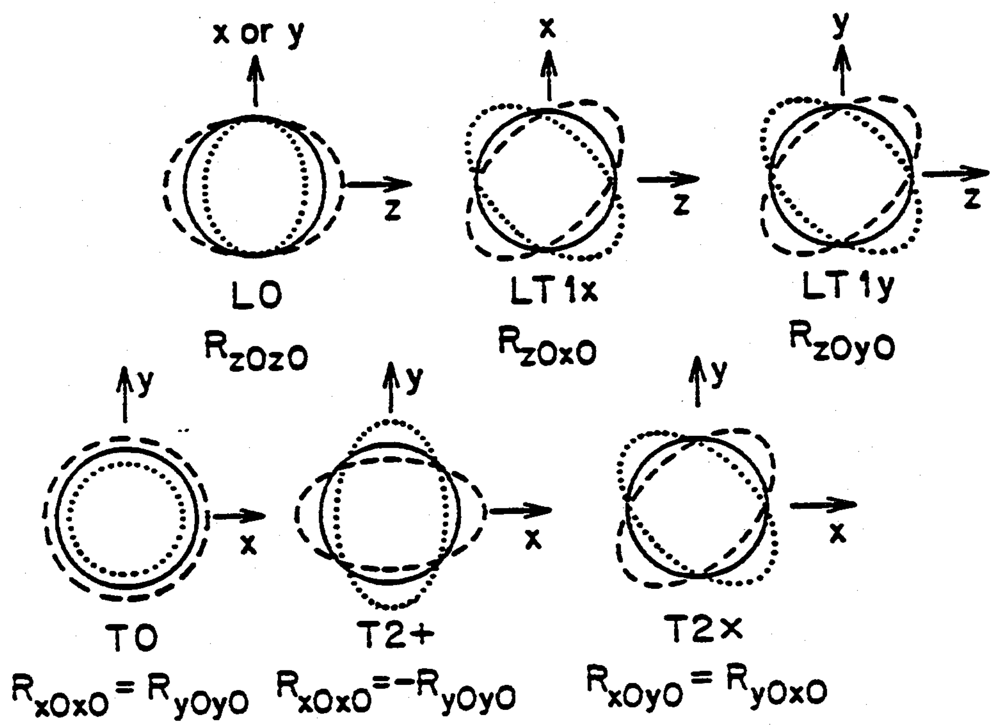


Fig. 3. The deformations of a sphere of test particles produced by gravitational waves with each of six polarizations. As the wave oscillates, the sphere (solid) first gets deformed in the manner shown dashed; then in the manner shown dotted.

R_{z0z0} produces deformations

$$\ddot{x} = 0,\ \ddot{y} = 0,\ \ddot{z} = -R_{z0z0}z. \tag{2.18a}$$

As $R_{z0z0}(t)$ alternately oscillates negative and positive this expands and then squeezes the sphere longitudinally (i.e., along the direction of wave propagation, z), while leaving its transverse cross section unchanged. In this sense the wave is "purely longitudinal". At any moment of time the deformed sphere is invariant under rotations about the propagation direction $\vec{e}_z$. This means, in the language of "canonical field theory", that the wave has "spin zero" (or "helicity zero"). These properties are summarized by saying that the wave is "L0" (longitudinal and spin zero).

$R_{z0x0} = R_{x0z0}$ produces deformations

$$\ddot{x} = -R_{x0z0}z,\quad \ddot{y} = 0,\quad \ddot{z} = -R_{x0z0}x. \tag{2.18b}$$

As $R_{x0z0}(t)$ oscillates this expands and then squeezes the sphere at a 45 degree angle to $\vec{e}_z$ in the "logitudinal-transverse" z-x plane, while leaving it undeformed in the y direction. At any moment the deformed sphere is invariant under a 360° rotation about the propagation direction, a property that it shares with the electric or magnetic field of an electromagnetic wave. Thus, this gravitational wave like an electromagnetic wave has spin one — but whereas the electromagnetic wave is "T1" (transverse, spin one), this gravitational wave is "LT1" (longitudinal-transverse, spin one).

$R_{z0y0} = R_{y0z0}$ produces deformations

$$\ddot{x} = 0,\quad \ddot{y} = -R_{y0z0}z,\quad \ddot{z} = -R_{y0z0}y, \tag{2.18c}$$

and is thus also "LT1". It is the LT1 polarization orthogonal to R_{z0x0}.

A wave with $R_{x0x0} = R_{y0y0}$ and all other R_{j0k0} zero produces deformations

$$\ddot{x} = R_{x0x0}x,\quad \ddot{y} = R_{x0x0}y,\quad \ddot{z} = 0. \tag{2.18d}$$

This wave alternately expands and compresses the sphere in the transverse plane while leaving it transversely circular and leaving it totally unchanged in the longitudinal direction. Thus, this wave is T0 ("transverse, spin-zero").

A wave with $R_{x0x0} = -R_{y0y0}$ and all other R_{j0k0} zero produces deformations

$$\ddot{x} = -R_{x0x0}x,\quad \ddot{y} = +R_{x0x0}y,\quad \ddot{z} = 0. \tag{2.18e}$$

As $R_{x0x0}(t)$ oscillates this wave expands the sphere in the x direction and squeezes it in y, then expands it in y and squeezes it in x. Thus the deformations are purely transverse, and at any moment the deformed sphere is invariant under a 180° rotation about the propagation direction. The "spin" ("helicity") S of any wave which propagates with the speed of light is determined by the angle Θ of rotations about the propagation direction that leave all momentary physical effects of the wave unchanged:

$$S = 360^\circ/\Theta.$$

Thus, this wave has spin S = 2, i.e., it is a T2 ("transverse, spin 2) wave. The orientation of the polarization is identified as "+" (x and y; horizontal and vertical).

$R_{x0y0} = R_{y0x0}$ produces deformations

$$\ddot{x} = R_{x0y0}y, \quad \ddot{y} = R_{x0y0}x, \quad \ddot{z} = 0 \tag{2.18f}$$

which are also T2; but the orientation of the polarization is "X". This is the T2 wave orthogonal to "+".

That the spin 0 waves have just one polarization state while the spin 1 and spin 2 waves each have two orthogonal polarization states is familiar both from quantum mechanics and from canonical, classical field theory. It is a consequence of the fact that the waves propagate with the speed of light — i.e., their quanta have zero rest mass.

This familiar feature is deceptively reassuring. Actually, a nasty surprise awaits us if we try to check the fundamental tenet of canonical field theory that the spin of a wave must be Lorentz invariant. If we begin in one Lorentz frame with a gravitational wave that is pure T0 or pure T2, we will find it to be pure T0 or pure T2 in all other Lorentz frames. However, if we begin with a pure LT1 wave in one frame, we will find a mixture of LT1, T0, and T2 in other frames; and if we begin with pure L0 in one frame, we will find a mixture of L0, LT1, T0, and T2 in other frames. See Eardley, Lee, and Lightman (1973) for proofs.

This means that any metric theory of gravity possessing L0 or LT1 waves violates the tenets of canonical field theory and cannot be quantized by canonical methods, even in the weak-gravity limit. Rosen's theory is an example (in the special case $c_{GW} = c_{EM}$ to which our discussion applies; when $c_{GW} \neq c_{EM}$ there exist eight polarizations, not six [C. M. Caves, private communication]). By contrast, metric theories with purely transverse, speed-of-light gravity waves obey the canonical tenets and are quantizable by canonical means, at least in the weak-gravity limit. Examples are general relativity which has pure T2 waves (Exercise 3 above) and the Dicke-Brans-Jordan theory which has both T2 waves and T0 waves.

* * * * *

Exercise 7. Verify the claims made above about the behavior of polarization states under Lorentz transformations. (See Eardley, Lee, and Lightman 1973 for solution.)

2.3 Plane waves on a flat background in general relativity

2.3.1 The gravitational-wave field h^{TT}_{jk}

Specialize now and henceforth to general relativity. Then a plane wave on a flat background has precisely two orthogonal polarization states: T2+ and T2X (denoted simply "+" and "X" henceforth).

Choose a specific Lorentz frame of the flat background space, and in that frame define a "gravitational-wave field" h^{TT}_{jk} by

$$R_{j0k0}(t-z) = -\tfrac{1}{2}h^{TT}_{jk,tt} \tag{2.19}$$

with $h^{TT}_{jk} = 0$ before any waves ever arrive. Then the waves are fully characterized equally well by R_{j0k0} or by h^{TT}_{jk}. Note that the only nonzero components of h^{TT}_{jk} are

$$h^{TT}_{xx} = -h^{TT}_{yy} \equiv A_+(t-z), \qquad h^{TT}_{xy} = h^{TT}_{yx} = A_\times(t-z), \tag{2.20}$$

A_+ being the "amplitude function" for the + polarization state and $A_\times$ being that for the × state. Note that h^{TT}_{jk} is purely spatial, symmetric, transverse to the propagation direction $\vec{n} = \vec{e}_z$, and also traceless (hence the TT superscript)

$$h^{TT}_{jk} = h^{TT}_{kj}, \qquad h^{TT}_{jk}\, n^k = 0, \qquad \delta^{jk}\, h^{TT}_{jk} = 0. \tag{2.21}$$

Note further that the experimenter's version of the equation of geodesic deviation (eq. 2.17) can be integrated to give

$$\delta x^j = \tfrac{1}{2}\, h^{TT}_{jk}\, x^k \tag{2.22}$$

for the change in location, in an LIF, of a test particle initially at x^k. This equation suggests the common interpretation of h^{TT}_{jk} as a "dimensionless strain of space".

2.3.2 Behavior of h^{TT}_{jk} under Lorentz transformations

A single gravitational wave is described, in different Lorentz frames of the flat background, by different h^{TT}_{jk} fields, each one purely spatial and TT in its own frame. How are these gravitational-wave fields related to each other? I shall state the answer in this section, leaving the proof as an exercise for the reader.

Begin in some fiducial background Lorentz frame, in which h^{TT}_{jk} is described by equations (2.20) above. Define

$$\psi \equiv t - z = \text{"retarded time"}, \tag{2.23a}$$

regard ψ as a scalar field in spacetime (a sort of "phase function" for the wave), and regard the amplitude functions A_+ and $A_\times$ of equations (2.20) as scalar fields which are known functions of ψ. From the scalar field ψ construct the "propagation vector"

$$\vec{k} \equiv -\vec{\nabla}\psi, \quad \text{so} \quad k^0 = k^z = 1, \;\; k^x = k^y = 0 \text{ in fiducial frame.} \tag{2.23b}$$

This propagation vector and the basis vector $\vec{e}_x$, which determines the orientations of the + and × polarization states, together define a "fiducial 2-flat" (plane)

$$\overleftrightarrow{f} \equiv \vec{k} \wedge \vec{e}_x \equiv (\text{2-flat spanned by } \vec{k} \text{ and } \vec{e}_x). \tag{2.23c}$$

Now choose some other background Lorentz frame with 4-velocity $\vec{u}'$. In that frame define basis vectors

$$\vec{e}_{0'} \equiv \vec{u}', \quad \vec{e}_{z'} \equiv \begin{bmatrix}\text{unit vector obtained by projecting } \vec{k} \\ \text{orthogonal to } \vec{u}' \text{ and renormalizing}\end{bmatrix},$$

$$\vec{e}_{x'} \equiv [\text{unit vector lying in } \overleftrightarrow{f} \text{ and orthogonal to } \vec{u}'], \tag{2.24a}$$

$$\vec{e}_{y'} \equiv \begin{bmatrix}\text{unit vector such that } \vec{e}_{0'}, \vec{e}_{x'}, \vec{e}_{y'}, \vec{e}_{z'} \text{ are a right-hand} \\ \text{oriented, orthonormal frame}\end{bmatrix}.$$

Then in this basis the gravitational-wave field has components

$$h'^{TT}_{x'x'} = -h'^{TT}_{y'y'} = A_+(\psi), \quad h'^{TT}_{x'y'} = h'^{TT}_{y'x'} = A_\times(\psi), \tag{2.24b}$$

where A_+ and $A_\times$ are the same scalar fields as describe the waves in the fiducial frame. Note, however, that ψ is not $t'-z'$; rather, it differs from $t'-z'$ by the standard doppler-shift factor:

$$\psi = t-z = (\nu'/\nu)(t'-z'), \tag{2.24c}$$

$$\frac{\nu'}{\nu} = \left[\begin{array}{c}\text{ratio of frequencies of photons, propagating in } \vec{k} \text{ direction,} \\ \text{as measured in the two reference frames}\end{array}\right].$$

Thus, under a Lorentz transformation the gravitational-wave frequencies get doppler shifted just like those of light, but the amplitude functions are left unchanged and the polarization directions change only by a projection that keeps them spatial (eq. 2.24a for $\vec{e}_x'$). The fact, that h^{TT}_{jk} is a "spin-2" quantity with amplitude functions A_+ and $A_\times$ that are Lorentz invariant is embodied in the statement that "$A_+ + iA_\times$ has spin-weight 2 and boost-weight 0" [language of Geroch, Held, and Penrose (1973); $i \equiv \sqrt{-1}$].

It is often convenient to define polarization tensors

$$\overleftrightarrow{e}_+' \equiv \vec{e}_{x'} \otimes \vec{e}_{x'} - \vec{e}_{y'} \otimes \vec{e}_{y'} \quad , \quad \overleftrightarrow{e}_\times' \equiv \vec{e}_{x'} \otimes \vec{e}_{y'} + \vec{e}_{y'} \otimes \vec{e}_{x'} , \tag{2.24d}$$

and to rewrite equations (2.24b) as

$$\overleftrightarrow{h}'^{TT} = A_+(\psi)\overleftrightarrow{e}_+' + A_\times(\psi)\overleftrightarrow{e}_\times' . \tag{2.24e}$$

Another useful relation is

$$R_{\alpha'\beta'\gamma'\delta'} = \tfrac{1}{2}\left(\ddot{h}'^{TT}_{\gamma'\delta'}k_{\beta'}k_{\gamma'} + \ddot{h}'^{TT}_{\beta'\gamma'}k_{\alpha'}k_{\delta'} - \ddot{h}'^{TT}_{\beta'\delta'}k_{\alpha'}k_{\gamma'} - \ddot{h}'^{TT}_{\alpha'\gamma'}k_{\beta'}k_{\delta'}\right), \tag{2.24f}$$

where $^\bullet = d/d\psi$.

* * * * *

Exercise 8. Show that in the fiducial frame of this section, equation (2.19) is equivalent to $R_{j0k0} = -(1/2)\ddot{h}^{TT}_{jk}k_0k_0$ where the dot means $d/d\psi$. Show further that in the fiducial frame the full Riemann tensor is given by equation (2.24f) without the primes. (Hint: use equations 2.9).

Exercise 9. Show that the Riemann tensor of (2.24f) with the primes and with $h'^{TT}_{\alpha'\gamma'}$ given by (2.24a,b) is obtained from that of Exercise 8 (no primes) by a standard Lorentz transformation. Convince yourself that this fully justifies the claimed behavior of h^{TT}_{jk} under a change of frames (eqs. 2.24a,b).

2.3.3 Relationship of h^{TT}_{jk} to Bondi news function

Consider gravitational waves propagating radially outward from a source, and approximate the background as flat. Introduce spherical polar coordinates and denote the associated orthonormal basis vectors by,

$$\vec{e}_{\hat{r}} = \partial/\partial r, \quad \vec{e}_{\hat{\theta}} = r^{-1}\partial/\partial\theta, \quad \vec{e}_{\hat{\varphi}} = (r\sin\theta)^{-1}\partial/\partial\varphi. \tag{2.25a}$$

Let $\vec{e}_{\hat{\theta}}$ be the fiducial direction (analog of $\vec{e}_x$ above) used in defining the polarization base states, so that

$$\overleftrightarrow{e}_+ = \vec{e}_{\hat{\theta}} \otimes \vec{e}_{\hat{\theta}} - \vec{e}_{\hat{\varphi}} \otimes \vec{e}_{\hat{\varphi}} \quad , \quad \overleftrightarrow{e}_\times = \vec{e}_{\hat{\theta}} \otimes \vec{e}_{\hat{\varphi}} + \vec{e}_{\hat{\varphi}} \otimes \vec{e}_{\hat{\theta}} ; \tag{2.25b}$$

$$h^{TT}_{jk} = A_+ \overleftrightarrow{e}_+ + A_\times \overleftrightarrow{e}_\times . \qquad (2.25c)$$

Because the wave fronts are spherical (though very nearly plane on length scales $ƛ << r$), A_+ and $A_\times$ die out as $1/r$ (Exercise 14 below)

$$A_+ = r^{-1} F_+(\psi;\theta,\varphi), \quad A_\times = r^{-1} F_\times(\psi;\theta,\varphi), \quad \psi = t-r. \qquad (2.25d)$$

In gravitational-wave studies near "future timelike infinity" $\mathscr{I}^+$, mathematical physicists often use instead of h^{TT}_{jk} a different description of the waves due to Bondi, van der Burg, and Metzner (1962) and to Sachs (1962): The role of the gravitational-wave amplitude is played by the "Bondi News Function"

$$N \equiv \frac{1}{2} \frac{\partial}{\partial t} (F_+ + i F_\times) = \frac{1}{2} r \frac{\partial}{\partial t} (A_+ + i A_\times), \qquad (2.26a)$$

where $i \equiv \sqrt{-1}$. This complex News function has the advantage of depending only on angles θ, φ, and retarded time ψ ($1/r$ dependence factored out); but for this it pays the price of not being a scalar field. In the language of Geroch, Held, and Penrose, it has "boost weight 2", whereas $A_+ + iA_\times$ has "boost weight 0". It is conventional in the Bondi-Sachs formalism to introduce the complex vector

$$\vec{\bar{m}} = (1/\sqrt{2})\ (\vec{e}_{\hat{\theta}} - i\, \vec{e}_{\hat{\varphi}}). \qquad (2.26b)$$

In terms of N and $\vec{\bar{m}}$ the gravitational-wave field is

$$\frac{\partial}{\partial t} h^{TT}_{jk} = \text{Real} \left\{ \frac{4N}{r} \bar{m}_j \bar{m}_k \right\} . \qquad (2.26c)$$

Several lecturers in this volume will use the Bondi-Sachs formalism (e.g., M. Walker, A. Ashtekar, and R. Isaacson).

2.4 Weak perturbations of curved spacetime in general relativity

2.4.1 Metric perturbations and Einstein field equations

Abandon, now, the approximation that the background spacetime is flat. As a foundation for discussing gravitational waves in curved spacetime, consider the general problem of linear perturbations of a curved background metric:

$$g_{\mu\nu} = g^{(B)}_{\mu\nu} + h_{\mu\nu} . \qquad (2.27)$$

In analyzing the metric perturbations $h_{\mu\nu}$, I shall not make explicit the small dimensionless parameter that underlies the perturbation expansion. It might be $ƛ/\mathscr{L}$ (gravitational wave expansion, §2.4.2 below); it might be the dimensionless amplitude of pulsation of a neutron star, $\delta R/R$ (linear pulsation-theory expansion, Thorne and Campolattaro 1967); it might be the mass ratio m/M of a small body m falling into a Schwarzschild black hole M and generating gravitational waves as it falls (linear perturbations of Schwarzschild geometry; Davis et al. 1971). In the latter two cases, near the star and hole $ƛ/\mathscr{L}$ is not $<< 1$; but nevertheless the linearized equations of this section are valid.

The perturbed Einstein field equations for $h_{\mu\nu}$ are expressed most conveniently in terms of the "trace-reversed" metric perturbation

$$\bar{h}_{\mu\nu} \equiv h_{\mu\nu} - \frac{1}{2} h\, g^{(B)}_{\mu\nu} \quad , \quad h \equiv h_{\mu\nu}\, g^{\mu\nu}_{(B)} . \qquad (2.28)$$

A straightforward calculation (cf. §§35.13 and 35.14 of MTW) gives for the first-order perturbations of the field equations

$$\bar{h}_{\mu\nu|\alpha}{}^{\alpha} + g^{(B)}_{\mu\nu}\,\bar{h}^{\alpha\beta}{}_{|\beta\alpha} - 2\bar{h}_{\alpha(\mu|}{}^{\alpha}{}_{\nu)} + 2R^{(B)}_{\alpha\mu\beta\nu}\,\bar{h}^{\alpha\beta} - 2R^{(B)}_{\alpha(\mu}\,\bar{h}_{\nu)}{}^{\alpha} = -16\pi\delta T_{\mu\nu}. \qquad (2.29)$$

Here a slash "|" denotes covariant derivative with respect to $g^{(B)}_{\mu\nu}$; indices on $\bar{h}_{\alpha\beta}$ are raised and lowered with $g^{(B)}_{\mu\nu}$; $R^{(B)}_{\alpha\beta\gamma\delta}$ and $R^{(B)}_{\alpha\beta}$ are the Riemann and Ricci tensors of the background, and $\delta T_{\mu\nu}$ is the first-order perturbation of the stress-energy tensor.

The first-order perturbed field equation (2.29) can be used to study a wide variety of phenomena, including wave generation (§3.5 below), wave propagation on a curved vacuum background (§2.4.2), absorption and dispersion of waves due to interaction with matter (§2.4.3), and scattering of waves off background curvature and the resulting production of wave tails (§2.4.2).

2.4.2 Wave propagation on a curved vacuum background

Consider gravitational waves of reduced wavelength $\bar{\lambda}$ propagating on a curved vacuum background with radius of curvature $\mathcal{R}$ and inhomogeneity scale $\mathcal{L}$. In keeping with the discussion in §1.2 assume $\bar{\lambda} \ll \mathcal{R}$, but for the moment do not assume $\bar{\lambda} \ll \mathcal{L}$. Then "vacuum" implies $\delta T_{\mu\nu} = 0$ in the field equations (2.29); and $\bar{\lambda} \ll \mathcal{R}$ implies that the terms involving $R^{(B)}_{\alpha\mu\beta\nu}$ and $R^{(B)}_{\alpha\beta}$, which are of size $h/\mathcal{R}^2$, can be neglected compared to the first three terms, which are of size $h/\bar{\lambda}^2$. Simplify the resulting field equations further by an infinitesimal coordinate change ("gauge change")

$$x^{\alpha}_{\text{new}} = x^{\alpha}_{\text{old}} + \xi^{\alpha}, \qquad h^{\text{new}}_{\mu\nu} = h^{\text{old}}_{\mu\nu} - \xi_{\mu|\nu} - \xi_{\nu|\mu} \qquad (2.30)$$

so designed as to make

$$\bar{h}_{\mu}{}^{\alpha}{}_{|\alpha} = 0 \qquad (\text{"Lorentz gauge"}). \qquad (2.31a)$$

(See, e.g., §35.14 of MTW for discussion of such gauge changes.) The first-order field equations (2.29) then become a simple source-free wave equation in curved spacetime:

$$\bar{h}_{\mu\nu|\alpha}{}^{\alpha} = 0. \qquad (2.31b)$$

The Riemann curvature tensor associated with these waves

$$R^{(W)}_{\alpha\mu\beta\nu} = \tfrac{1}{2}\left(h_{\alpha\nu|\mu\beta} + h_{\mu\beta|\nu\alpha} - h_{\mu\nu|\alpha\beta} - h_{\alpha\beta|\mu\nu}\right) \qquad (2.32)$$

will also satisfy the wave equation

$$R^{(W)}_{\alpha\mu\beta\nu|\sigma}{}^{\sigma} = 0 \qquad (2.33)$$

(covariant derivatives "|" commute because $\bar{\lambda} \ll \mathcal{R}$). Note that although we require $\bar{\lambda} \ll \mathcal{L}$ in order to give a clear definition of "wave", we need not place any restriction on $\bar{\lambda}/\mathcal{L}$ in order to derive the wave equation (2.31b).

The Lorentz gauge condition (2.31a) is preserved by any gauge change (2.30) whose generating function ξ_{α}, like $\bar{h}_{\mu\nu}$, satisfies the wave equation $\xi_{\alpha|\mu}{}^{\mu} = 0$. One of the four degrees of freedom in such a gauge change can be used to make $\bar{h}_{\mu\nu}$ trace-free everywhere

$$\bar{h}_{\alpha}{}^{\alpha} = 0 \;\text{ so }\; \bar{h}_{\mu\nu} = h_{\mu\nu} \qquad (\text{"trace-free Lorentz gauge"}) \qquad (2.34)$$

(MTW exercise 35.13); and the other three degrees of freedom can be used locally (in a local inertial frame of the background $g^{(B)}_{\mu\nu}$) to guarantee that

$$h_{0\alpha} = 0, \quad h_{ij} = h^{TT}_{ij} \quad \text{("local TT gauge")}, \tag{2.35}$$

where h^{TT}_{ij} is the gravitational-wave field defined, in the background LIF, by

$$R^{(W)}_{i0j0} \equiv - \tfrac{1}{2} h^{TT}_{ij,00} \; . \tag{2.36}$$

If the background is approximated as flat throughout the wave zone, then one can introduce a global inertial frame of $g^{(B)}_{\mu\nu}$ throughout the wave zone and one can impose the TT gauge globally. However, if the background is curved, a global TT gauge cannot exist (MTW exercise 35.13).

One often knows $h_{\alpha\beta}$ or $\bar{h}_{\alpha\beta}$ in a Lorentz but non-TT gauge and wants to compute its "gauge-invariant part" h^{TT}_{ij} in some LIF of the background. Such a computation is performed most easily by a "TT projection", which is mathematically equivalent to a gauge transformation (MTW Box 35.1): One identifies the propagation direction n_j in the LIF as the direction in which the wave is varying rapidly (on length scale ƛ). One then obtains h^{TT}_{ij} by discarding all parts of h_{ij} or $\bar{h}_{ij}$ along n_j and by then removing the trace:

$$h^{TT}_{ij} = P_{ia}h_{ab}P_{bj} - \tfrac{1}{2} P_{ij}P_{ab}h_{ab} = (\text{same expression with } h_{ab} \rightarrow \bar{h}_{ab}), \tag{2.37}$$

where $P_{ab} = \delta_{ab} + n_a n_b$. WARNING: This projection process gives the correct answer only in an LIF of the background and only if $h_{\mu\nu}$ is in a Lorentz gauge.

* * * * * *

Exercise 10. Show that the infinitesimal coordinate change (2.30) produces the claimed gauge change of $h_{\mu\nu}$. Show further that the Riemann tensor of the waves is correctly given by (2.32) in any gauge, and that this Riemann tensor is invariant under gauge changes (2.30).

Exercise 11. Show that a gauge change with $\xi_{\alpha|\mu}{}^{\mu} = 0$ can be used to make a Lorentz-gauge $h_{\mu\nu}$ trace-free globally (eq. 2.34) and TT locally (eq. 2.35). Show further that the TT projection process (2.37) produces the same result as this gauge transformation.

2.4.3 Absorption and dispersion of waves by matter

When electromagnetic waves propagate through matter (e.g., light through water, radio waves through the interplanetary medium), they are partially absorbed and partially scatter off charges; and the scattered and primary waves superpose in such a way as to change the propagation speed from that of light in vacuum ("Dispersion"). A typical model calculation of this absorption and dispersion involves electrons of charge e, mass m, and number density n, each bound to a "lattice point" by a 3-dimensional, isotropic, damped, harmonic-oscillator force:

$$\ddot{\underset{\sim}{z}} + (1/\tau_*)\dot{\underset{\sim}{z}} + \omega_o^2 \underset{\sim}{z} = (e/m)\underset{\sim}{E} = - (e/m)\dot{\underset{\sim}{A}}, \tag{2.38a}$$

where $\underset{\sim}{A}$ is the vector potential in transverse Lorentz gauge and a dot denotes $\partial/\partial t$. These electrons produce a current density $\underset{\sim}{J} = ne(d\underset{\sim}{z}/dt)$, which enters into Maxwell's equations for wave propagation $\Box \underset{\sim}{A} = -4\pi \underset{\sim}{J}$ to give waves of the form $\underset{\sim}{E} = \underset{\sim}{E}_o \exp(-i\omega t + i\underset{\sim}{k}\cdot\underset{\sim}{x})$ with the dispersion relation (for weak dispersion)

$$\frac{\omega}{k} = (\text{phase speed}) = 1 - \frac{2\pi n e^2/m}{\omega_o{}^2 - \omega^2 - i\omega/\tau_*} \;. \tag{2.38b}$$

This dispersion relation shows both absorption (imaginary part of ω/k) and dispersion (real part), and in real situations either or both can be very large.

When gravitational waves propagate through matter they should also suffer adsorption and dispersion. However, in real astrophysical situations the absorption and dispersion will be totally negligible, as the following model calculation shows. (For previous model calculations similar to this one see Szekeres 1971.)

The best absorbers or scatterers of gravitational waves that man has devised are Weber-type resonant-bar gravitational-wave detectors (§§4.1.2 and 4.1.4). On larger scales, a spherical self-gravitating body such as the earth or a neutron star is also a reasonably good absorber and scatterer (good compared to other kinds of objects such as interstellar gas). Consider, then, an idealized "medium" made of many solid spheres (spheres to avoid anisotropy of response to gravity waves), each of which has quadrupole vibration frequency ω_o, damping time (due to internal friction) τ_*, mass m and radius R. For ease of calculation (and because we only need order of magnitude estimates) ignore the self gravity and mutual gravitational interactions of the spheres, and place the spheres at rest in a flat background spacetime with number per unit volume n. Let h^{TT}_{ij} be the gravitational-wave field and require $\bar{\lambda} > n^{-1/3} > R$. The waves' geodesic deviation force drives each sphere into quadrupolar oscillations with quadrupole moment $\mathcal{I}_{jk}$ satisfying the equation of motion (Exercise 22 in §4.1.4 below)

$$\ddot{\mathcal{I}}_{jk} + (1/\tau_*)\dot{\mathcal{I}}_{jk} + \omega_o{}^2\mathcal{I}_{jk} = (1/5)mR^2\ddot{h}^{TT}_{jk} \tag{2.39a}$$

(analog of the electromagnetic equation 2.38a). As a result of these oscillations each sphere reradiates. The wave equation for h^{TT}_{jk} with these reradiating sources (analog of $\Box\underline{A} = -4\pi\underline{J}$) is

$$\Box h^{TT}_{jk} \equiv \eta^{\alpha\beta} h^{TT}_{jk,\alpha\beta} = -8\pi n\ddot{\mathcal{I}}_{jk} \tag{2.39b}$$

(Exercise 12). By combining equations (2.39a,b) and assuming a wave of the form $h_{jk} \propto \exp(-i\omega t + i\underline{k}\cdot\underline{x})$ we obtain the gravitational-wave dispersion relation

$$\frac{\omega}{k} = (\text{phase speed}) = 1 - \frac{(4\pi/5)nmR^2\omega^2}{\omega_o{}^2 - \omega^2 - i\omega/\tau_*} \;. \tag{2.39c}$$

To see that the absorption and dispersion are negligible, compare the length scale $\ell = |(1 - \omega/k)\omega|^{-1}$ for substantial absorption or for a phase shift of $\sim \pi/2$ with the radius of curvature of spacetime produced by the scatterers (i.e., the maximum size that the scattering region can have without curling itself up into a closed universe), $\mathcal{R} = (nm)^{-1/2}$:

$$\frac{\ell}{\mathcal{R}} = \underbrace{\left|\frac{\omega_o{}^2 - \omega^2 - i\omega/\tau_*}{(4\pi/5)\omega^2}\right|}_{\substack{\gtrsim 1 \text{ off resonance} \\ \sim(1/Q) \text{ on resonance}}} \frac{1}{\underbrace{(nR^3)^{1/2}}_{\lesssim 1}\;\underbrace{(m/R)^{1/2}}_{< 1/\sqrt{2}}\;\underbrace{(\omega R)}_{<1}} \;. \tag{2.40}$$

Here $Q = 1/\omega\tau_*$ is the quality factor of a scatterer, $nR^3 \lesssim 1$ because the scatterers cannot be packed closer together than their own radii, $m/R < 1/2$ because a scatterer cannot be smaller than a black hole of the same mass, and $\omega R = R/\bar{\lambda} < 1$

was required to permit a geodesic-deviation analysis (see above). In the most extreme of idealized universes $\ell/\mathcal{R}$ can be no smaller than unity off resonance (dispersion) and $1/Q$ on resonance (absorption); and such extreme values can be achieved only for neutron stars or black holes ($m/R \sim 1$) packed side by side ($nR^3 \sim 1$) with $R \sim \lambdabar$. In the real universe, $\ell/\mathcal{R}$ will always be $\ggg 1$; i.e., absorption and dispersion will be negligible regardless of what material the waves encounter and regardless of how far they propagate through it.*

For this reason, henceforth in discussing wave propagation through astrophysical matter (e.g., the interior of the Earth or Sun) I shall approximate $\bar{h}_{\mu\nu}|_\alpha{}^\alpha = -16\pi\delta T_{\mu\nu}$ by $\bar{h}_{\mu\nu}|_\alpha{}^\alpha = 0$. The matter will influence wave propagation only through the background curvature it produces (covariant derivative "$|$"), not through any direct scattering or absorption ($\delta T_{\mu\nu}$); see §2.6.1 below.

* * * * *

Exercise 12. For non-self-gravitating matter in flat spacetime and in Lorentz coordinates, show that $T^{\alpha\beta}{}_{,\beta} = 0$ implies $T^{jk} = (1/2)(\rho x^j x^k)_{,00} +$ (perfect spatial divergence), where ρ is mass density. Average this over a lattice of oscillating spheres with number density $n > \lambdabar^{-3}$ to get $T^{jk} = (1/2)n\ddot{I}_{jk}$, where $I_{jk} = \int \rho x^j x^k d^3x$ is the second moment of the mass distribution of each sphere. Passing gravitational waves excite the oscillations in accord with equation (2.39a) (result to be proved in Exercise 22). These oscillations involve no volume changes, so $\ddot{I}_{jk} = \ddot{\mathcal{J}}_{jk} =$ (trace-free part of $\ddot{I}_{jk}$); moreover, equation (2.39a) shows that $\ddot{\mathcal{J}}_{jk}$ is transverse and traceless. Show that this permits TT gauge to be imposed in the field equations (2.29) in the presence of the oscillating, reradiating spheres (usually it can be imposed only outside all sources), and that the resulting field equations reduce to (2.39b). Then derive the gravitational-wave dispersion relation (2.39c) and the estimate (2.40) of the effects of dispersion and absorption.

2.4.4 Scattering of waves off background curvature, and tails of waves

A self-gravitating body of mass m and size R will typically generate gravitational waves with reduced wavelength

$$\lambdabar \sim (R^3/m)^{\frac{1}{2}} \sim \mathcal{R}_s = (\text{radius of curvature of spacetime near source}). \qquad (2.41)$$

If the body has strong self gravity, $m/R \sim 1$ (neutron star or black hole), then $\lambdabar \sim \mathcal{R}_s$ in the innermost parts of the wave zone; and the curvature coupling terms must be retained in the first-order Einstein equations (2.29). These terms cause the waves to scatter off the background curvature; and repetitively backscattered waves superimposing on each other produce a gravity-wave "tail" that lingers near the source long after the primary waves have departed, dying out as $t^{-(2\ell+2)}$ for waves of multipole order ℓ. See, e.g., Price (1972) for a more detailed discussion, and Cunningham, Price, and Moncrief (1978) for an explicit example.

I regard these backscatterings and tails as part of the wave generation problem and as irrelevant to the problem of wave propagation. In fact, I have defined the inner edge of the "local wave zone" to be so located that throughout it, and throughout the wave propagation problem, $\lambdabar \ll \mathcal{R}$ and backscatter and tails can be ignored (eq. 1.7 and associated discussion).

*For description of a physically unrealistic but conceivable material in which dispersion is so strong that it actually reflects gravitational waves see Press (1979).

2.4.5 The stress-energy tensor for gravitational waves

Gravitational waves carry energy and momentum and can exchange them with matter, e.g., with a gravitational-wave detector. Isaacson (1968) (see also §35.15) has quantified this by examining nonlinear corrections to the wave-propagation equation (2.31b). In this section I shall sketch the main ideas of his analysis.

Consider a gravitational wave with $\lambdabar \ll \mathcal{L} \lesssim \mathcal{R}$, and expand the metric of the full spacetime in a perturbation series

$$\underset{1,\mathcal{L}}{g_{\mu\nu}} = g^{(B)}_{\mu\nu} + \underset{\mathcal{A},\lambdabar}{h_{\mu\nu}} + \underset{\mathcal{A}^2,\lambdabar}{j_{\mu\nu}} + \cdots \quad . \tag{2.42a}$$

Below each term I have written the characteristic magnitudes $(1, \mathcal{A}, \mathcal{A}^2)$ of the metric components, and the lengthscales $(\mathcal{L}, \lambdabar)$ on which they vary in the most "steady" of coordinate systems. Note that $j_{\mu\nu}$ is a nonlinear correction to the propagating waves. By inserting this perturbation series into the standard expression (MTW eqs. 8.47-8.49) for the Einstein curvature tensor $G_{\mu\nu}$ in terms of $g_{\mu\nu}$ and its derivatives, and by grouping terms according to their magnitudes and their lengthscales of variation, one obtains

$$G_{\mu\nu} = \underset{\leq 1/\mathcal{R}^2,\mathcal{L}}{G^{(B)}_{\mu\nu}} + \underset{\mathcal{A}/\lambdabar^2,\lambdabar}{G^{(1)}_{\mu\nu}(h)} + \underset{\mathcal{A}^2/\lambdabar^2,\lambdabar}{G^{(2)}_{\mu\nu}(h)} + \underset{\mathcal{A}^2/\lambdabar^2,\lambdabar}{G^{(1)}_{\mu\nu}(j)} + \cdots \quad . \tag{2.42b}$$

Here $G^{(B)}_{\mu\nu}$ is the Einstein tensor of the background metric $g^{(B)}_{\mu\nu}$; $G^{(1)}_{\mu\nu}$(h or j) is the linearized correction to $G_{\mu\nu}$ (MTW eq. 35.58a, trace-reversed); and $G^{(2)}_{\mu\nu}(h)$ is the quadratic correction (MTW eq. 35.58b), trace-reversed).

Isaacson splits the Einstein equations into three parts: a part which varies on scales $\mathcal{L}$ (obtained by averaging, "$\langle\ \rangle$", over a few wavelengths)

$$G^{(B)}_{\mu\nu} = 8\pi\left(T^{(B)}_{\mu\nu} + \langle T^{(2)}_{\mu\nu}\rangle + T^{(W)}_{\mu\nu}\right), \quad T^{(W)}_{\mu\nu} \equiv -(1/8\pi)\,\langle G^{(2)}_{\mu\nu}(h)\rangle ; \tag{2.43a}$$

a part of magnitude $\mathcal{A}/\lambdabar^2$ which varies on scales $\lambdabar$ and averages to zero on larger scales

$$G^{(1)}_{\mu\nu}(h) = 8\pi\, T^{(1)}_{\mu\nu} \iff \bar{h}_{\mu\nu|\alpha}{}^{\alpha} = -16\pi\, T^{(1)}_{\mu\nu} \text{ in Lorentz gauge;} \tag{2.43b}$$

and a part of magnitude $\mathcal{A}^2/\lambdabar^2$ which varies on scales $\lambdabar$ and averages to zero on larger scales

$$G^{(1)}_{\mu\nu}(j) = -G^{(2)}_{\mu\nu}(h) + \langle G^{(2)}_{\mu\nu}(h)\rangle + 8\pi\left(T^{(2)}_{\mu\nu} - \langle T^{(2)}_{\mu\nu}\rangle\right). \tag{2.43c}$$

Here $T^{(B)}_{\mu\nu}$ is the stress-energy tensor of the background; $T^{(1)}_{\mu\nu}$ and $T^{(2)}_{\mu\nu}$ are its first- and second-order perturbations; $T^{(W)}_{\mu\nu} \equiv -(1/8\pi)\langle G^{(2)\mu\nu}_{\mu\nu}(h)\rangle$ is a stress-energy tensor associated with the gravitational waves; and the averaging $\langle\ \rangle$ can be performed in the most naive of manners if the coordinates are sufficiently "steady", but must be performed carefully, by Brill-Hartle techniques (MTW exercise 35.14), if they are not. The "smoothed" field equations (2.43a), together with the contracted Bianchi identities $G^{\mu\nu}_{(B)}{}_{|\nu} \equiv 0$, imply a conservation law for energy and momentum in the presence of gravitational waves:

$$T^{\mu\nu}_{(B)}|_{\nu} + \langle T^{\mu\nu}_{(2)}\rangle|_{\nu} + T^{\mu\nu}_{(W)}|_{\nu} = 0. \tag{2.44}$$

(Here and throughout this section indices are raised and lowered with $g^{(B)}_{\mu\nu}$.)

To understand the physics of the field equations (2.43) and conservation law (2.44), let us reconsider the propagation of waves through a cloud of spherical oscillators (§2.4.3). Equation (2.43b) is the wave equation (2.39b) for h, which we used to calculate the absorption and dispersion of the waves. In this wave equation $T^{(1)}_{\mu\nu}$ is the part of $T_{\mu\nu}$ that is linear in the oscillators' amplitude of motion; in Exercise 12 its spatial part (after averaging over scales $< \bar{\lambda}$) was shown to be $T^{(1)}_{jk} = \frac{1}{2} n \ddot{\mathcal{J}}_{jk}$. Equation (2.43c) describes the generation of nonlinear corrections $j_{\mu\nu}$ to the propagating waves. In Lorentz gauge it takes the explicit form

$$j_{\mu\nu}|_{\alpha}{}^{\alpha} = \begin{pmatrix}\text{source terms quadratic in } h_{\alpha\beta} \text{ and in} \\ \text{the amplitude of oscillator motion}\end{pmatrix}. \tag{2.45}$$

In §2.6.2 we shall see that these nonlinear corrections, like absorption and dispersion, are negligible in realistic astrophysical circumstances. Equation (2.43a) describes the generation of smooth, background curvature by the stress-energy of the gravitational waves $T^{(W)}_{\mu\nu}$ and of the matter. Note that the waves contribute an amount of order $\mathcal{A}^2/\bar{\lambda}^2$ to the background curvature $1/\mathcal{R}^2$, and that therefore $1/\mathcal{R}^2 \gtrsim \mathcal{A}^2/\bar{\lambda}^2$; i.e.,

$$\mathcal{A} \lesssim \bar{\lambda}/\mathcal{R}. \tag{2.46}$$

Since $\bar{\lambda}/\mathcal{R} \ll 1$ for propagating waves, such waves must necessarily have dimensionless amplitudes $\mathcal{A} \ll 1$. If ever $\mathcal{A}$ were to become of order unity, the wave would cease to be separable from the background curvature; the two would become united as a dynamically vibrating spacetime curvature to which the theory of propagating gravitational waves cannot be applied. Equation (2.44) describes the exchange of energy and momentum between matter and waves. In this conservation law $T^{\mu\nu}_{(B)} = mnu^{\mu}u^{\nu}$ is the stress-energy of the unperturbed spheres (number density n, mass m, 4-velocity u^{μ}) and by itself has vanishing divergence. The term $\langle T^{\mu\nu}_{(2)}\rangle$ is the quadratic-order stress-energy associated with the spheres' oscillations, averaged spatially over a few wavelengths and temporally over a few periods of the waves. In an LIF of the oscillators the only nonzero components are the energy density $\langle T^{00}_{(2)}\rangle$ which includes the kinetic energy and the potential energy of oscillation and the thermal energy produced when the oscillations are damped by internal friction [$1/\tau_*$ term in oscillators' equation of motion (2.39a)], and a transverse stress of magnitude comparable to $\langle T^{00}_{(2)}\rangle$. Thus, for our idealized problem the conservation law (2.44) describes the absorption of gravitational-wave energy by the oscillators and the subsequent conversion of oscillation energy into heat.

The gravitational-wave stress-energy tensor $T^{(W)}_{\mu\nu}$ "lives in" the background spacetime and is manipulated using background-spacetime mathematics [e.g., covariant derivative "|" in the conservation law (2.44)]. Because of the averaging $\langle\ \rangle$ in its definition, $T^{(W)}_{\mu\nu}$ gives a well-defined localization of the waves' energy and momentum only on lengthscales somewhat larger than $\bar{\lambda}$ (no way to say whether the energy is in the "crest" of a wave or in its "trough"); no more precise localization of gravitational energy is possible in general relativity. Like $R^{(W)}_{\alpha\beta\gamma\delta}$ and unlike $h_{\alpha\beta}$, the stress-energy tensor $T^{(W)}_{\alpha\beta}$ is gauge invariant. Explicit calculations (Isaacson 1968; MTW §35.15) give

$$\begin{aligned} T^{(W)}_{\mu\nu} &= \frac{1}{32\pi}\left\langle \bar{h}_{\alpha\beta}|_{\mu}\, \bar{h}^{\alpha\beta}|_{\nu} - \tfrac{1}{2}\bar{h}|_{\mu}\bar{h}|_{\nu} - 2\,\bar{h}^{\alpha\beta}|_{\beta}\,\bar{h}_{\alpha(\mu|\nu)}\right\rangle \quad \text{in any gauge} \\ &= \frac{1}{32\pi}\left\langle \bar{h}_{\alpha\beta}|_{\mu}\,\bar{h}^{\alpha\beta}|_{\nu}\right\rangle \quad \text{in trace-free Lorentz gauge} \end{aligned} \tag{2.47}$$

(continued next page)

$$= \frac{1}{32\pi} \left\langle h^{TT}_{jk,\mu}\, h^{TT}_{jk,\nu} \right\rangle \text{ in LIF of any observer.}$$

2.5 Wave propagation in the geometric optics limit

2.5.1 Differential equations of geometric optics

Return now to the explicit problem of the propagation of gravitational waves from the local wave zone of a source out through the lumpy universe toward earth. Throughout the local wave zone, and almost everywhere in the universe, not only will $\lambdabar$ be very small compared to the background radius of curvature $\mathcal{R}$, but also it will be small compared to the scale $\mathcal{L}$ on which the background curvature varies

$$\lambdabar \ll \mathcal{L}. \tag{2.48}$$

Here, as in the above discussion of the waves' stress-energy, I shall assume that $\lambdabar \ll \mathcal{L}$; later (§2.6.1) I shall relax that assumption. Here I shall also assume that $\lambdabar \ll \mathcal{L}^{(W)} \equiv$ (radius of curvature of the wave fronts of the waves or any smaller scale length for transverse variation of the waves).

The assumptions $\lambdabar \ll \mathcal{L}$ and $\lambdabar \ll \mathcal{L}^{(W)}$ permit us to solve for the propagation using the techniques of geometric optics (e.g., MTW Exercise 35.15): Introduce trace-free Lorentz gauge everywhere, and ignore the effects of direct interaction between the propagating waves and matter (negligible absorption and dispersion). Then

$$h^{\alpha}{}_{\alpha} = 0, \qquad h^{\mu\alpha}{}_{|\alpha} = 0, \qquad h_{\mu\nu|\alpha}{}^{\alpha} = 0. \tag{2.49}$$

The solution of these gauge and propagation equations is a rapidly varying function of retarded time ψ and a slowly varying function of the other spacetime coordinates:

$$h_{\mu\nu} = h_{\mu\nu}(\underset{\text{variation on scale } \lambdabar}{\psi},\ \underset{\text{variation on scales } \mathcal{L}^{(W)},\, \mathcal{L},\, \mathcal{R}}{x^{\alpha}}) \tag{2.50}$$

As in the discussion of waves in flat spacetime (§2.3.2), define the propagation vector

$$\vec{k} \equiv - \vec{\nabla}\psi . \tag{2.51a}$$

Then, aside from fractional corrections of order $\lambdabar/\mathcal{L}^{(W)}$, $\lambdabar/\mathcal{L}$, $\lambdabar/\mathcal{R}$ the gauge and field equations (2.49) imply

$$k_{\alpha}k^{\alpha} = 0 \text{ and } k_{\beta|\alpha}k^{\alpha} = 0 \Longleftrightarrow \vec{k} \text{ is tangent to null geodesics ("rays")}, \tag{2.51b}$$

$$h_{\alpha}{}^{\alpha} = 0,\ h_{\alpha\beta}k^{\beta} = 0 \quad \Longleftrightarrow h_{\alpha\beta} \text{ is trace free and orthogonal to } \vec{k}, \tag{2.51c}$$

$$h_{\mu\nu|\alpha}k^{\alpha} = - \tfrac{1}{2}\,(\vec{\nabla}\cdot\vec{k})h_{\mu\nu} \qquad \text{(propagation equation for } h_{\mu\nu}). \tag{2.51d}$$

Had we been analyzing the propagation of electromagnetic waves rather than gravitational, our Lorentz gauge equations for the vector potential would have been

$$A^{\alpha}{}_{|\alpha} = 0, \qquad A_{\mu|\alpha}{}^{\alpha} = 0 \tag{2.52a}$$

(MTW eq. 16.5' with curvature coupling term removed because $\lambdabar \ll \mathcal{R}$) (cf. eq.

2.49); our geometric-optics ansatz would have been

$$A_\mu = A_\mu(\psi;\, x^\alpha) \tag{2.52b}$$

(cf. eq. 2.50); and in the geometric optics limit the gauge and wave equations (2.52a) would have reduced to

$$k_\alpha k^\alpha = 0, \qquad k_{\beta|\alpha}k^\alpha = 0, \qquad A_\alpha k^\alpha = 0, \qquad A_{\mu|\alpha}k^\alpha = -\tfrac{1}{2}(\vec{\nabla}\cdot\vec{k})A_\mu \tag{2.52c}$$

(cf. eqs. 2.51c,d).

* * * * *

Exercise 13. Show that in the geometric optics limit $\lambdabar \ll \mathcal{L} \lesssim \mathcal{R}$ and $\lambdabar \ll \mathcal{L}^{(W)}$, and with the geometric optics ansatz (2.50), the gravitational gauge and propagation equations (2.49) reduce to the geometric optics equations (2.51). Similarly show that for electromagnetic waves (2.52a) reduce to (2.52c).

2.5.2 Solution of geometric optics equations in local wave zone

In the local wave zone of the source introduce (flat-background) spherical coordinates (t,r,θ,φ). The waves propagate radially outward from the source along the null-geodesic rays

$$\psi = t - r,\ \theta,\ \varphi \text{ all constant}, \qquad k^0 = k^r = 1. \tag{2.53a}$$

Throughout the local wave zone introduce transverse basis vectors $\vec{e}_{\hat\theta} = r^{-1}\,\partial/\partial\theta$ and $\vec{e}_{\hat\varphi} = (r\sin\theta)^{-1}\,\partial/\partial\varphi$ and polarization tensors

$$\overleftrightarrow{e}_+ \equiv \vec{e}_{\hat\theta}\otimes\vec{e}_{\hat\theta} - \vec{e}_{\hat\varphi}\otimes\vec{e}_{\hat\varphi}, \qquad \overleftrightarrow{e}_\times \equiv \vec{e}_{\hat\theta}\otimes\vec{e}_{\hat\varphi} + \vec{e}_{\hat\varphi}\otimes\vec{e}_{\hat\theta}\ . \tag{2.53b}$$

Then it turns out (Exercise 14) that in TT gauge the general solution to the gauge and propagation equations (2.51c,d) is

$$\begin{aligned} \overleftrightarrow{h}^{TT} &= A_+(\psi;r,\theta,\varphi)\,\overleftrightarrow{e}_+ + A_\times(\psi;r,\theta,\varphi) \\ A_+ &= r^{-1}F_+(\psi;\theta,\varphi), \qquad A_\times = r^{-1}F_\times(\psi;\theta,\varphi); \end{aligned} \tag{2.53c}$$

and similarly for electromagnetic waves

$$\begin{aligned} \vec{A} &= A_{\hat\theta}(\psi;r,\theta,\varphi)\,\vec{e}_{\hat\theta} + A_{\hat\varphi}(\psi;r,\theta,\varphi)\,\vec{e}_{\hat\varphi} \\ A_{\hat\theta} &= r^{-1}F_{\hat\theta}(\psi;\theta,\varphi), \qquad A_{\hat\varphi} = r^{-1}F_{\hat\varphi}(\psi;\theta,\varphi). \end{aligned} \tag{2.54}$$

Stated in words: In polarization bases $\overleftrightarrow{e}_+$, $\overleftrightarrow{e}_\times$ and $\vec{e}_{\hat\theta}$, $\vec{e}_{\hat\varphi}$ which are parallel transported along the rays, the amplitude functions A_+, $A_\times$ of gravitational waves and $A_{\hat\theta}$, $A_{\hat\varphi}$ of electromagnetic waves die out as $1/r$ but otherwise are constant along the rays.

The precise forms of $F_+(\psi;\theta,\varphi) = rA_+$ and $F_\times(\psi;\theta,\varphi) = rA_\times$ are to be determined by solution of the wave generation problem (§3 below). The local-wave-zone waves (2.53) are then to be used as "starting conditions" for propagation out through the universe.

* * * * *

Exercise 14. Show that equations (2.53) are the general solution of the gravitational geometric optics equations (2.51) specialized to TT gauge, for waves propagating radially outward through the local wave zone of a source. Similarly show that equations (2.54) are the solution of the electromagnetic equations (2.52c).

2.5.3 Solution of geometric optics equations in distant wave zone

Suppose that the wave generation problem has been solved to give h^{TT}_{jk} in the form (2.52) throughout the local wave zone. These waves can then be propagated throughout the rest of the universe (assuming $\lambdabar \ll \mathcal{L}$ and $\lambdabar \ll \mathcal{L}^{(W)}$) using the following constructive method [solution of geometric optics equations (2.51)]:

First extend the radial null rays (2.53a) of the local wave zone out through the universe by solving the geodesic equation. Continue to label each ray by ψ,θ,φ and parametrize it by an affine parameter denoted r (and equal to the radial coordinate in the local wave zone):

$$\text{rays are } (\psi,\theta,\varphi) = \text{const}; \quad \vec{k} \equiv -\vec{\nabla}\psi = d/dr. \tag{2.55a}$$

Next, along each ray parallel propagate the fiducial basis vector $\vec{e}_{\hat{\theta}}$

$$\vec{\nabla}_{\vec{k}}\, \vec{e}_{\hat{\theta}} = 0 \text{ everywhere}, \qquad \vec{e}_{\hat{\theta}} = r^{-1}\, \partial/\partial\theta \text{ in local wave zone.} \tag{2.55b}$$

Together $\vec{e}_{\hat{\theta}}$ and $\vec{k}$ form a fiducial 2-flat $\vec{k} \wedge \vec{e}_{\hat{\theta}}$ to be used below in defining the polarization of the waves. Next, propagate A_+ and $A_\times$ along each ray by solving the ordinary differential

$$(\partial A/\partial r)_{\psi,\theta,\varphi} = -\tfrac{1}{2}(\vec{\nabla}\cdot\vec{k})A, \quad A = [\text{expression (2.53c) in local wave zone}]. \tag{2.55c}$$

The resulting A_+, $A_\times$, and fiducial 2-flat $\vec{k} \wedge \vec{e}_{\hat{\theta}}$ determine the gravitational-wave field in the manner of §2.3.2: At any event in the distant wave zone introduce an observer with 4-velocity $\vec{u}$; introduce an orthonormal basis

$$\vec{e}_0 \equiv \vec{u}, \quad \vec{e}_z \equiv \begin{bmatrix}\text{unit vector obtained by projecting } \vec{k} \\ \text{orthogonal to } \vec{u} \text{ and renormalizing}\end{bmatrix},$$

$$\vec{e}_x \equiv [\text{unit vector lying in } \vec{k} \wedge \vec{e}_{\hat{\theta}} \text{ and orthogonal to } \vec{u}]. \tag{2.55d}$$

$$\vec{e}_y \equiv \begin{bmatrix}\text{unit vector such that } \vec{e}_0, \vec{e}_x, \vec{e}_y, \vec{e}_z \text{ are a right-hand oriented,} \\ \text{orthonormal frame}\end{bmatrix};$$

and introduce corresponding polarization tensors

$$\overleftrightarrow{e}_+ \equiv \vec{e}_x \otimes \vec{e}_x - \vec{e}_y \otimes \vec{e}_y, \qquad \overleftrightarrow{e}_\times \equiv \vec{e}_x \otimes \vec{e}_y + \vec{e}_y \otimes \vec{e}_x. \tag{2.55e}$$

Then the gravitational-wave field in this LIF is

$$\overleftrightarrow{h}^{TT} = A_+\, \overleftrightarrow{e}_+ + A_\times\, \overleftrightarrow{e}_\times; \tag{2.55f}$$

and the Riemann tensor and stress-energy tensor associated with the waves are

$$R^{(W)}_{\alpha\beta\gamma\delta} = \tfrac{1}{2}\,(\ddot{h}^{TT}_{\alpha\delta}\, k_\beta k_\gamma + \ddot{h}^{TT}_{\beta\gamma}\, k_\alpha k_\delta - \ddot{h}^{TT}_{\beta\delta}\, k_\alpha k_\gamma - \ddot{h}^{TT}_{\alpha\gamma}\, k_\beta k_\delta),$$

$$T^{(W)}_{\alpha\beta} = \frac{1}{16\pi}\langle \dot{A}_+{}^2 + \dot{A}_\times{}^2\rangle k_\alpha k_\beta\,, \quad \text{where} \quad \dot{} \equiv \partial/\partial\psi. \tag{2.56}$$

For electromagnetic waves the geometric optics equations (2.52c) have a similar solution. In a basis $\vec{e}_{\hat\theta}$, $\vec{e}_{\hat\varphi}$ obtained by parallel transport (eqs. 2.55b) along the rays (eqs. 2.55a) the components of the vector potential, $A_{\hat\theta}$ and $A_{\hat\varphi}$, satisfy identically the same propagation equation (2.55c) as the gravitational-wave amplitude functions A_+ and $A_\times$. Moreover, an observer with 4-velocity $\vec{u}$ can always put the waves into purely spatial Lorentz gauge (no component of $\vec{A}$ along $\vec{u}$) by a gauge change, which produces

$$\vec{A}^S = A_{\hat\theta}\,\vec{e}_x + A_{\hat\varphi}\,\vec{e}_y \tag{2.57}$$

(analog of eq. 2.55f) with $\vec{e}_x$, $\vec{e}_y$ given by equations (2.55d). The electromagnetic field tensor, and the stress-energy tensor of the waves averaged over several wavelengths (analogs of eqs. 2.56) are

$$F_{\alpha\beta} = \dot{A}^S_\beta\, k_\alpha - \dot{A}^S_\alpha\, k_\beta,$$

$$\langle T_{\alpha\beta}\rangle = \frac{1}{4\pi}\langle A_{\hat\theta}{}^2 + A_{\hat\varphi}{}^2\rangle k_\alpha k_\beta\,. \tag{2.58}$$

*   *   *   *   *   *

Exercise 15. Show that equations (2.55) constitute a solution of the gravitational geometric optics equations (2.51), transformed locally to TT gauge. Show further that this solution joins smoothly onto the local-wave-zone solution (2.53), and that the Riemann tensor and stress-energy tensor of these waves have the form (2.56). Similarly show that if $A_{\hat\theta}$ and $A_{\hat\varphi}$ are propagated via (2.55c), then $A_{\hat\theta}\vec{e}_{\hat\theta} + A_{\hat\varphi}\vec{e}_{\hat\varphi}$ is a solution of the electromagnetic equations (2.52c); (2.57) is this same solution in another gauge; and (2.58) are the field tensor and averaged stress-energy tensor of the waves.

2.5.4 Example: Propagation through a Friedmann universe

As an example of the geometric optics solution for wave propagation consider, as the background spacetime, a closed Friedmann univers with metric

$$g^{(B)}_{\alpha\beta}\, dx^\alpha dx^\beta = a^2(\eta)\,[-d\eta^2 + d\chi^2 + \Sigma^2\,(d\theta^2 + \sin^2\theta\, d\varphi^2)]$$

$$\Sigma = : \ \chi \text{ for } k = 0, \quad \sin\chi \text{ for } k = +1, \quad \sinh\chi \text{ for } k = -1. \tag{2.59}$$

Here k is the curvature parameter (k = 0 for a spatially flat universe, k = +1 for a closed universe, k = -1 for an open universe; see, e.g., chapters 27-29 of MTW). Orient the coordinates so the source of the waves is at $\chi = 0$, and let the source be active (emit waves) at a coordinate time $\eta \simeq \eta_e$ when the expansion factor of the universe is $a = a_e$ ("e" for "emission"). The flat, spherical coordinates of the local wave zone, and the retarded time are

$$t = a_e(\eta-\eta_e), \quad r = a_e\chi, \quad \theta, \quad \varphi; \qquad \psi = t-r = a_e(\eta-\chi-\eta_e); \tag{2.60}$$

and the waves in the local wave zone are described by equations (2.53b,c).

Throughout the wave zone (local and distant) the rays and propagation vector of equations (2.55a) are

$$\eta - \chi = \eta_e + \psi/a_e , \qquad k^{\eta} = k^{\chi} = \left(\frac{\partial\eta}{\partial r}\right)_{\psi,\theta,\varphi} = \left(\frac{\partial\chi}{\partial r}\right)_{\psi,\theta,\varphi} = \frac{a_e}{a^2}; \qquad (2.61a)$$

the parallel-propagated fiducial basis vector (eq. 2.55b) is

$$\vec{e}_{\hat{\theta}} = (1/a\Sigma)\ \partial/\partial\theta ; \qquad (2.61b)$$

and the transported gravity-wave and electromagnetic-wave amplitude functions (eq. 2.55c) are

$$A_J = \frac{F_J(\psi;\theta,\varphi)}{a\Sigma}\ ; \ J = + \text{ or } \times \text{ (gravity)},\ J = \hat{\theta} \text{ or } \hat{\varphi} \text{ (electromagnetism)}. \qquad (2.61c)$$

If we approximate the earth as at rest in the Friedmann coordinate system at χ_o, θ_o, φ_o and we denote the present epoch by $\eta \simeq \eta_o$, $a = a_o$, then the basis vectors (2.53b) of the earth's LIF are

$$\vec{e}_0 = \frac{1}{a_o}\frac{\partial}{\partial\eta}\ , \quad \vec{e}_x = \frac{1}{a_o\Sigma_o}\frac{\partial}{\partial\theta}\ , \quad \vec{e}_y = \frac{1}{a_o\Sigma_o \sin\theta_o}\frac{\partial}{\partial\varphi}\ , \quad \vec{e}_z = \frac{1}{a_o}\frac{\partial}{\partial\chi}\ ; \qquad (2.61d)$$

and the gravitational-wave field as measured at earth (eqs. 2.53b,c) is

$$\overleftrightarrow{h}^{TT} = \frac{1}{a_o\Sigma_o}\left[F_+(\psi;\theta_o,\varphi_o)(\vec{e}_x \otimes \vec{e}_x - \vec{e}_y \otimes \vec{e}_y) + F_\times(\psi;\theta_o,\varphi_o)(\vec{e}_x \otimes \vec{e}_y + \vec{e}_y \otimes \vec{e}_x)\right]. \ (2.61e)$$

The energy density in these waves as measured at earth (eq. 2.56) is

$$T_{00}^{(W)} = \underbrace{\frac{\langle \dot{F}_+{}^2 + \dot{F}_\times{}^2 \rangle}{4 \cdot 4\pi a_o^2 \Sigma_o^2}}_{\uparrow\ \text{(surface area around source today)}} \underbrace{\left(\frac{a_e}{a_o}\right)^2}_{(1+Z)^{-2}}\ , \qquad \dot{} = \partial/\partial\psi\ , \qquad (2.62)$$

where Z is the cosmological redshift of the source. Similarly, for electromagnetic waves

$$\vec{A}^S = (1/a_o\Sigma_o)[F_{\hat{\theta}}(\psi;\theta_o,\varphi_o)\ \vec{e}_x + F_{\hat{\varphi}}(\psi;\theta_o,\varphi_o)\ \vec{e}_y]\ , \qquad (2.63a)$$

$$\langle T_{00} \rangle = \frac{\langle \dot{F}_{\hat{\theta}}{}^2 + \dot{F}_{\hat{\varphi}}{}^2 \rangle}{4\pi a_o^2 \Sigma_o^2}\left(\frac{a_e}{a_o}\right)^2 . \qquad (2.63b)$$

Note that the factor $1/a_o\Sigma_o$, by which the amplitudes of the waves die out as they recede from the source, is given in terms of cosmological parameters by

$$\frac{1}{a_o\Sigma_o} \equiv \frac{1}{R} = \frac{H_o q_o^{\,2}(1+Z)}{-q_o + 1 + q_o Z + (q_o-1)(2q_o Z+1)^{1/2}}$$

$$\approx \frac{H_o}{Z}\,[1 + \tfrac{1}{2}\,(1+q_o)Z + O(Z^2)] \qquad \text{for} \qquad Z \ll 1 \qquad (2.64)$$

$$\approx H_o q_o \qquad \text{for } Z \gg 1 \text{ and } Z \gg 1/q_o$$

(MTW eqs. 29.28-29.33). Here H_o is the Hubble expansion rate; q_o is the deceleration parameter of the universe; Z is the cosmological redshift of the source; and I have assumed zero cosmological constant. For formulas with nonzero cosmological constant see MTW eqs. (29.32).

* * * * *

Exercise 16. Show that for propagation through a Friedmann universe equations (2.55)-(2.58) become (2.59)-(2.63).

2.6 Deviations from geometric optics

I have already discussed in detail several ways that wave propagation can differ from geometric optics: absorption and dispersion by matter (§2.4.3; almost always negligible for gravitational waves), and scattering of waves off background curvature with resulting production of tails (§2.4.4; important primarily near source, but also if waves encounter a sufficiently compact body — e.g., a neutron star or black hole). In this section I shall describe two other nongeometric-optics effects: diffraction and nonlinear interactions of the wave with itself.

2.6.1 Diffraction

As gravitational and electromagnetic waves propagate through the universe, they occasionally encounter regions of enhanced spacetime curvature due to concentrations of matter (galaxies, stars, ...) which produce a breakdown in $\lambdabar \ll \mathcal{L}$ and/or in $\lambdabar \ll \mathcal{L}^{(W)}$ and a resulting breakdown in geometric-optics propagation. Such a breakdown is familiar from light propagation, where it is called "diffraction".

Consider, as an example, the propagation of waves through the neighborhood and interior of the sun (Fig. 4), and ignore absorption and dispersion by direct interaction with matter (justified for gravitational waves, §2.4.3; not justified for electromagnetic waves). As they pass near and through the sun, rays from a distant source are deflected and forced to cross each other; i.e., they are

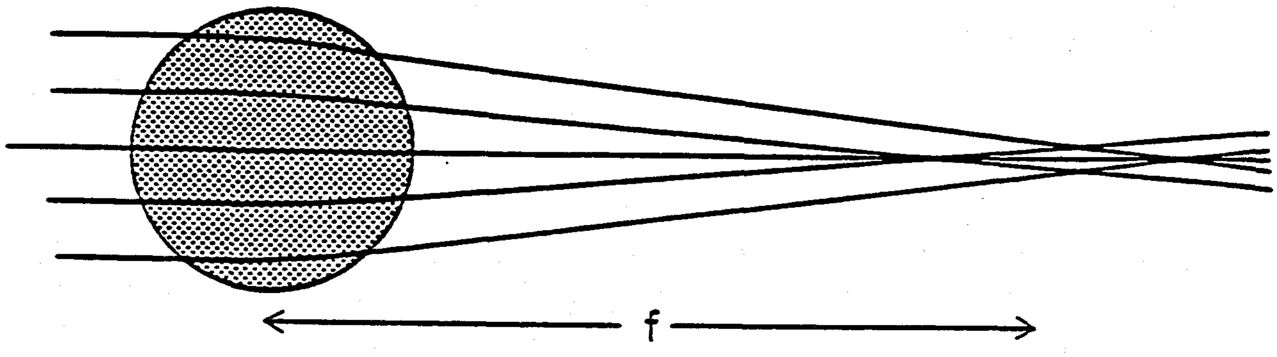


Fig. 4 The rays for geometric-optics wave propagation through the sun.

focussed gravitationally. The dominant source of deflection is the spacetime curvature of the solar core. It produces ray crossing ("caustics") along the optic axis at distances of order (and greater than) the "focal distance"

$$f \sim \frac{\mathcal{L}}{4M/\mathcal{L}} \simeq 20\ \mathrm{AU}\ . \tag{2.65}$$

Here $\mathcal{L} \sim 10^5$ km is the inhomogeneity scale of the solar core, $M \sim 0.3\ M_\odot$ is the mass of the solar core, and the value 20 AU comes from detailed calculations with a detailed solar model (Cyranski and Lubkin 1974).

Geometric optics would predict infinite amplification of the waves at the caustics. However, geometric optics breaks down there because it also predicts $\mathcal{L}^{(W)} \to 0$. To understand the actual behavior of the waves near the caustics, think of the waves which get focussed by the solar core as a single wave packet that has transverse dimension $\Delta y \sim \mathcal{L}$ as it passes through the core. The uncertainty principle for waves $(\Delta y \Delta k_y \gtrsim 1)$ forces this wave packet to spread in a nongeometric optics manner with a spreading angle

$$\theta_s \sim \Delta k_y / k_x \sim \lambdabar/\mathcal{L}\ . \tag{2.66}$$

This spreading is superimposed on the geometric-optics focussing, and it spreads out the highly focussed waves near the caustics over a lateral scale y_s

$$y_s \sim (\lambdabar/\mathcal{L}) f \sim (\lambdabar/4M)\mathcal{L}\ . \tag{2.67}$$

If $y_s \ll \mathcal{L}$ (i.e., if $\lambdabar \ll 4M$) there is substantial focussing: the wave energy density increases near the caustics by a factor $\sim (y_s/\mathcal{L})^2$ and the amplitude increases by $\sim y_s/\mathcal{L} \sim \lambdabar/4M$. The details of this regime are described by the laws of "Fresnel diffraction". On the other hand, if $y_s \gtrsim \mathcal{L}$ (i.e., if $\lambdabar \gtrsim 4M$) there is negligible focussing; and the little focussing that does occur is described by the laws of "Fraunhoffer diffraction". For full details see Bontz and Haugan (1981) and references therein.

For the case of the sun the dividing line between substantial focussing and little focussing is $\lambdabar \sim$ (gravitational radius of sun), i.e., (frequency) $\sim 10^4$ Hz. Since all strong sources of gravitational waves are expected to have $\lambdabar \gtrsim$ (gravitational radius of source) $\gtrsim$ (gravitational radius of Sun), i.e., (frequency) $\lesssim 10^4$ Hz, they all lie in the "little focussing regime" — a conclusion that bodes ill for any efforts to send gravitational-wave detectors on spacecraft to the orbit of Uranus in search of amplified gravitational waves; cf. Sonnabend (1979).

Far beyond the focal region the geometric optics approximation becomes valid again, except for a smearing of lateral structure of the waves over an angular scale $\sim \theta_s$. For example, ray crossing may produce multiple images of a gravitational-wave source in this region; and those images can be computed by geometric optics methods aside from θ_s-smearing.

2.6.2 Nonlinear effects in wave propagation

Once a gravitational wave has entered and passed through the local wave zone, its nonlinear interactions with itself are of no importance. To see this consider the idealized problem of a radially propagating, monochromatic wave in flat spacetime. At linear order, in spherical coordinates write the wave field as

$$h_{\hat{\theta}\hat{\theta}} = -h_{\hat{\varphi}\hat{\varphi}} = A_0(\theta,\varphi)\,\frac{\lambdabar}{r}\cos\left(\frac{t-r}{\lambdabar}\right), \tag{2.68}$$

where hats denote components in an orthonormal, spherical basis. Note that the angular function A_o is the amplitude of the wave in the induction zone, where it is just barely starting to become a wave. For any realistic source $A_o \lesssim 1$; see eq. (3.55b) below.

As these waves propagate, their nonlinear interaction with themselves produces a correction $j_{\mu\nu}$ to $h_{\mu\nu}$. Like $h_{\mu\nu}$, $j_{\mu\nu}$ has the outgoing-wave form

$$j_{\hat\mu\hat\nu} = J_{\hat\mu\hat\nu}(r,\theta,\varphi)\cos[(t-r)/\bar{\lambda} + \text{phase}]. \tag{2.69}$$

Equations (2.43c) and (2.45) describe the growth of this correction as it propagates. They have the form $j_{\mu\nu|\alpha}{}^{\alpha}$ = (source), which for $j_{\mu\nu}$ of the form (2.69) reduces to

$$\frac{1}{r\bar{\lambda}}\frac{\partial}{\partial r}(r\, J_{\hat\mu\hat\nu}) = \alpha\left(\frac{A_o}{r}\right)^2 k_{\hat\mu}k_{\hat\nu} + O\left(\frac{A_o{}^2\bar{\lambda}}{r^3}\right), \tag{2.70}$$

where α is a constant of order unity and $k_{\hat r} = -k_{\hat 0} = 1$ is the propagation vector.

The leading, $1/r^2$ source term in (2.70) produces a rapidly growing correction

$$J_{\hat\mu\hat\nu} = \alpha A_o{}^2\,\frac{\bar{\lambda}}{r}\,\ln\left(\frac{r}{\bar{\lambda}}\right)k_{\hat\mu}k_{\hat\nu}\,; \tag{2.71}$$

but this correction is purely longitudinal, i.e., it has no transverse-traceless part, i.e., it is purely a gauge change. The $1/r^3$ source term in (2.70) produces corrections of negligible size:

$$J_{\hat\mu\hat\nu} \sim A_o{}^2\bar{\lambda}^2/r^2 \ll A_o\bar{\lambda}/r \sim h_{\hat\mu\hat\nu}\,. \tag{2.72}$$

Thus, the effects of nonlinearities are negligible as claimed.

*   *   *   *   *

Exercise 17. Use equations (35.58) of MTW to show that the wave equation (2.43c) for $j_{\mu\nu}$ reduces to (2.70). Show that the solution has the form (2.71), (2.72).

3 THE GENERATION OF GRAVITATIONAL WAVES

Turn now from wave propagation to wave generation. Elsewhere (Thorne 1977) I have given a rather thorough review of the theory of gravitational wave generation, including a variety of computational techniques valid for a variety of types of sources. In these lectures I shall focus almost entirely on computational techniques that involve multipole-moment decompositions. My discussion in large measure will be an overview of a long treatise on "multipole expansions of gravitational radiation" which I published recently in Reviews of Modern Physics (Thorne 1980a; cited henceforth as "RMP").

3.1 Foundations for multipole analyses

3.1.1 Multipole moments of a stationary system in linearized general relativity

I shall motivate my discussion of multipole moments by considering a stationary (time-independent), weakly gravitating system surrounded by vaccum and described using the linearized approximation to general relativity (MTW chapters 18 and 19). In a Cartesion coordinate system and in Lorentz gauge the Einstein field equations and gauge conditions are

$$\nabla^2\bar{h}^{00} = -16\pi\rho, \quad \nabla^2\bar{h}^{0j} = -16\pi\rho v^j, \quad \nabla^2\bar{h}^{jk} = -16\pi T^{jk},$$
$$\bar{h}^{0j}{}_{,j} = 0, \qquad \bar{h}^{ij}{}_{,j} = 0. \tag{3.1}$$

Here ∇^2 is the flat-space Laplacian, $\bar{h}^{\alpha\beta}$ is the trace-reversed metric perturbation, $\rho = T^{00}$ is the source's mass density, v^j is its velocity field, and T^{jk} is its stress tensor. These equations can be solved for $\bar{h}^{\alpha\beta}$ using the usual flat-space Green's function $-(1/4\pi)|\underline{x}-\underline{x}'|^{-1}$; and the resulting integrals can then be expanded in powers of $1/r$. By doing this and by then making gauge changes described in § VIII of RMP, one can bring the <u>external</u> gravitational field into the form

$$\bar{h}^{00} = \frac{4M}{r} + \frac{6}{r^3}\mathscr{I}_{jk}\, n_j n_k + \dots + \frac{4(2\ell-1)!!}{\ell!\, r^{\ell+1}} \underbrace{\mathscr{I}_{a_1\dots a_\ell}}_{\mathscr{I}_{A_\ell}} \underbrace{n_{a_1}\dots n_{a_\ell}}_{N_{A_\ell}} + \dots, \tag{3.2a}$$

$$\bar{h}^{0j} = \frac{2}{r^2}\epsilon_{jka}\, S_k\, n_a + \dots + \frac{4\ell(2\ell-1)!!}{(\ell+1)!\, r^{\ell+1}} \epsilon_{jka_\ell}\, S_{kA_{\ell-1}}\, N_{A_\ell} + \dots, \tag{3.2b}$$

$$\bar{h}^{ij} = 0. \tag{3.2c}$$

Here $r \equiv (\delta_{jk}x^j x^k)^{\frac{1}{2}}$ is radius, $n_j \equiv x^j/r$ is the unit radial vector, ϵ_{ijk} is the Levi-Civita tensor used to form cross products, $(2\ell-1)!!$ is the product $(2\ell-1)\cdot$ $\cdot(2\ell-3)\cdots 1$, shorthand notations have been introduced for strings of indices $a_1\dots a_\ell \equiv A_\ell$ and for products of unit radial vectors $n_{a_1}\cdots n_{A_\ell} = N_{A_\ell}$, and spatial indices are moved up and down with impunity because the spatial coordinates are Cartesian. The "multipole moments" M, $\mathscr{I}_{A_\ell}$, S_{A_ℓ} are given as integrals over the source by

$$M = (\text{mass}) = \int \rho d^3x, \tag{3.3a}$$

$$\mathscr{I}_{a_1\dots a_\ell} = \begin{pmatrix}\text{mass } \ell\text{-pole}\\ \text{moment}\end{pmatrix} = \left[\int(\rho+T^{jj})x^{a_1}\dots x^{a_\ell}d^3x\right]^{\text{STF}}, \tag{3.3b}$$

$$S_{a_1\dots a_\ell} = \begin{pmatrix}\text{current } \ell\text{-pole}\\ \text{moment}\end{pmatrix} = \left[\int(\epsilon_{a_\ell pq}x^p\rho v^q)x^{a_1}\dots x^{a_\ell}d^3x\right]^{\text{STF}}. \tag{3.3c}$$

Here "STF" means "symmetric, trace-free part", i.e., "symmetrize and remove all traces"; cf. equation (2.2) of RMP. Note that the mass moments, which produce Newtonian-type gravitational accelerations $\underline{g} = (1/4)\underline{\nabla}\bar{h}^{00}$, are generated by $\rho+T^{jj}$ = (mass density + trace of stress tensor). For a description of a possible future experiment to verify the role of T^{jj} see §IV.D of Braginsky, Caves, and Thorne (1977).

For any realistic, weakly gravitating astrophysical source $T^{jj} \ll \rho$, so $\mathscr{I}_{a_1\dots a_\ell} = \mathscr{I}_{A_\ell}$ is the STF part of the ℓ'th moment of the mass density; and S_{A_ℓ} is the STF part of the $(\ell-1)$'th moment of the angular momentum density (though I call it the "ℓ'th current moment"). Note that as in electromagnetism, so also here, the external gravitational field is fully characterized by just two families of moments: the "mass moments" $\mathscr{I}_{A_\ell}$ are analogs of electric moments, the "current moments" S_{A_ℓ} are analogs of magnetic moments. In order of magnitude, for a source of mass M, size L, and characteristic internal velocity v,

$$|\mathscr{I}_{A_\ell}| \lesssim ML^\ell, \quad |\mathscr{S}_{A_\ell}| \lesssim MvL^\ell. \tag{3.4}$$

It is remarkable that, by an appropriate adjustment of gauge, $\bar{h}_{ij}$ can be made to vanish identically outside the source; and $\bar{h}_{00}$ is then determined fully by the mass moments while $\bar{h}_{0j}$ is determined fully by the current moments.

Note that the spatial coordinates of equations (3.2) have been "mass-centered" so the mass dipole moment $\mathscr{I}_j$ vanishes. I always mass-center my coordinates, thereby avoiding the issue of arbitrariness in the moments associated with arbitrariness in the origin of coordinates. Note further that the current dipole moment $\mathscr{S}_j$ is precisely the angular momentum of the source.

* * * * *

Exercise 18. Write down the solution of equations (3.1) using the Green's function $-(1/4\pi)|\underset{\sim}{x}-\underset{\sim}{x}'|^{-1}$, expanded in powers of $1/r$. Then specialize the discussion of §VIII of RMP to the stationary case and use its gauge changes to bring $\bar{h}^{\alpha\beta}$ into the form (3.2), (3.3).

3.1.2 Relation of STF tensors to spherical harmonics

The "STF" expansions (3.2) for $\bar{h}^{\alpha\beta}$ are mathematically equivalent to the more familiar expansions in terms of spherical harmonics $Y_{\ell m}(\theta,\varphi)$. The precise relationship between STF expansions and $Y_{\ell m}$ expansions is spelled out in §II of RMP. Here I shall describe only the flavor of that relationship.

Choose a specific value for the spherical-harmonic index ℓ. Then there are $2\ell+1$ linearly independent STF tensors of order ℓ ("STF-ℓ tensors"); and there are $2\ell+1$ linearly independent functions $Y_{\ell m}(\theta,\varphi)$. Moreover, the STF-$\ell$ tensors and the $Y_{\ell m}(\theta,\varphi)$ generate the same irreducible representation of the rotation group. Any scalar function $F(\theta,\varphi)$ can be expanded in two mathematically equivalent forms:

$$\begin{aligned} F(\theta,\varphi) &= \sum_{\ell,m} f_{\ell m}\, Y_{\ell m}(\theta,\varphi) \\ &= \sum_{\ell} \mathscr{F}_{A_\ell}\, N_{A_\ell} . \end{aligned} \tag{3.5}$$

In the first expansion the coefficients $f_{\ell m}$ are constant scalars, and the angular dependence is contained in the harmonics $Y_{\ell m}$. In the second expansion the coefficients $\mathscr{F}_{A_\ell}$ are constant STF-ℓ tensors, and the angular dependence is obtained by contracting the unit vectors $N_{A_\ell} \equiv n_{a_1}\ldots n_{a_\ell}$ into them.

STF expansions were widely used in the nineteenth century, before $Y_{\ell m}(\theta,\varphi)$ came into vogue; see, e.g., Kelvin and Tate (1879); Hobson (1931). In recent years they have been restored to common use by relativity theorists (e.g., Pirani 1964, RMP, Thorne 1981) because they are rather powerful when the spherical harmonics being manipulated are tensorial rather than scalar. In part, this power stems from the fact that the indices of $\mathscr{F}_{A_\ell}$ carry both angular dependence (implicitly) and tensorial component properties (explicitly), and carry them simultaneously. An example is the tensorial harmonic $\epsilon_{pqj}\,\mathscr{B}_{kqA_{\ell-2}}\,n_p\,N_{A_{\ell-2}}$. Harmonics of this form are second-rank tensors (two free indices j and k); they have harmonic order ℓ (ℓ indices on $\mathscr{B}$ implies these generate the same irreducible representation of the rotation group as do $Y_{\ell m}$); and they have parity $\pi = (-1)^{\ell+1}$ ("1" from ϵ_{pqj}, "ℓ" from $\mathscr{B}$).

3.1.3 The Einstein equations in de Donder (harmonic) gauge

In performing multipole decompositions of fully relativistic gravitational fields (the following sections) it is computationally powerful to work in de Donder (harmonic) gauge: Define the "gravitational field" $\bar{h}^{\alpha\beta}$ in terms of the "metric density" $\mathfrak{g}^{\alpha\beta}$ by

$$\mathfrak{g}^{\alpha\beta} \equiv (-g)^{\frac{1}{2}} g^{\alpha\beta} \equiv \eta^{\alpha\beta} - \bar{h}^{\alpha\beta}, \qquad g \equiv \det \|g_{\mu\nu}\| , \tag{3.6}$$

where $\eta^{\alpha\beta}$ is the Minkowski metric, diag(-1,1,1,1); and adjust the coordinates so as to impose the de Donder gauge conditions

$$\mathfrak{g}^{\alpha\beta}{}_{,\beta} = - \bar{h}^{\alpha\beta}{}_{,\beta} = 0. \tag{3.7}$$

Then the Einstein field equations take on the form (Landau and Lifshitz 1962, eq. 100.4; MTW eq. 20.21)

$$\underbrace{\mathfrak{g}^{\mu\nu} \bar{h}^{\alpha\beta}{}_{,\mu\nu}}_{\uparrow\ \text{characteristics: null rays of metric } g_{\alpha\beta}} = -16\pi(-g)(T^{\alpha\beta} + t_{LL}^{\alpha\beta}) - \bar{h}^{\alpha\mu}{}_{,\nu}\, \bar{h}^{\beta\nu}{}_{,\mu} \tag{3.8}$$

or, equivalently

$$\underbrace{\eta^{\mu\nu} \bar{h}^{\alpha\beta}{}_{,\mu\nu}}_{\uparrow\ \text{characteristics: flat-spacetime rays}} = -16\pi(-g)(T^{\alpha\beta} + t_{LL}^{\alpha\beta}) - \bar{h}^{\alpha\mu}{}_{,\nu}\bar{h}^{\beta\nu}{}_{,\mu} + \bar{h}^{\alpha\beta}{}_{,\mu\nu}\bar{h}^{\mu\nu} \equiv W^{\alpha\beta}. \tag{3.8'}$$

Here $t_{LL}^{\alpha\beta}$ is the Landau-Lifshitz pseudotensor (Landau and Lifshitz 1962, eq. 100.7; MTW eq. 20.22) which, in de Donder gauge, can be written as

$$\begin{aligned} 16\pi(-g)t_{LL}^{\alpha\beta} = {} & \tfrac{1}{2}\, \mathfrak{g}^{\alpha\beta}\mathfrak{g}_{\lambda\mu}\, \bar{h}^{\lambda\nu}{}_{,\rho}\bar{h}^{\rho\mu}{}_{,\nu} + \mathfrak{g}_{\lambda\mu}\mathfrak{g}^{\nu\rho}\bar{h}^{\alpha\lambda}{}_{,\nu}\bar{h}^{\beta\mu}{}_{,\rho} \\ & - (\mathfrak{g}^{\alpha\lambda}\mathfrak{g}_{\mu\nu}\bar{h}^{\beta\nu}{}_{,\rho}\bar{h}^{\mu\rho}{}_{,\lambda} + \mathfrak{g}^{\beta\lambda}\mathfrak{g}_{\mu\nu}\bar{h}^{\alpha\nu}{}_{,\rho}\bar{h}^{\mu\rho}{}_{,\lambda}) \\ & + \tfrac{1}{8}\, (2\mathfrak{g}^{\alpha\lambda}\mathfrak{g}^{\beta\mu} - \mathfrak{g}^{\alpha\beta}\mathfrak{g}^{\lambda\mu})(2\mathfrak{g}_{\nu\rho}\mathfrak{g}_{\sigma\tau} - \mathfrak{g}_{\rho\sigma}\mathfrak{g}_{\nu\tau})\bar{h}^{\nu\tau}{}_{,\lambda}\bar{h}^{\rho\sigma}{}_{,\mu} \end{aligned} \tag{3.9}$$

where $\mathfrak{g}_{\alpha\beta} = (-g)^{-\frac{1}{2}} g_{\alpha\beta}$ is the inverse of $\mathfrak{g}^{\alpha\beta}$. The law of local conservation of energy and momentum $T^{\alpha\beta}{}_{;\beta} = 0$ can be written in terms of partial derivations as

$$\left[(-g)(T^{\alpha\beta} + t_{LL}^{\alpha\beta})\right]_{,\beta} = 0 \tag{3.10}$$

(Landau and Lifshitz 1964, eq. 100.8; MTW eq. 20.23b).

The field equations (3.8) and (3.8') can be thought of as wave equations for $\bar{h}^{\alpha\beta}$ with source terms that include "gravitational stress-energy" (nonlinear terms in $\bar{h}^{\mu\nu}$). In the form (3.8) the wave operator is that of curved spacetime; its characteristics are the null rays of the curved spacetime metric $g_{\alpha\beta}$, and none of the source terms involve second derivatives of $\bar{h}^{\mu\nu}$. By contrast the form (3.8') involves a flat-spacetime wave operator; it is obtained from (3.8) by moving the second derivative term $\bar{h}^{\alpha\beta}{}_{,\mu\nu}\bar{h}^{\mu\nu}$ out of the wave operator and into the source.

The form (3.8'), with its flat-spacetime wave operator $\Box \equiv \eta^{\mu\nu}\partial_\mu\partial_\nu$, has great computational advantages over (3.8): It can be solved (formally) for $\bar{h}^{\alpha\beta}$

using a flat-spacetime Green's function, whereas the (formal) solution of (3.8) requires a far more complicated curved-spacetime Green's function (cf. DeWitt and Brehme 1960); and its solution is naturally decomposed into spherical harmonics because spherical harmonics are eigenfunctions of the flat operator $\Box$ but not of the curve-spacetime wave operator (3.8).

On the other hand, the flat operator $\Box$ entails serious dangers: (i) It propagates gravitational waves with the wrong speed, thereby losing at linear order the "Coulomb" phase shift produced by the M/r field of the source, and then trying to correct for this loss at quadratic order with a term that diverges logarithmically in r far from the source. I avoid this danger by restricting my use of $\Box$ to the "wave generation problem", which is formulated entirely at radii $r < r_0$, and by using the correct curved-spacetime wave operator when studying "wave propagation" at radii $r > r_0$ [cf. the paragraph preceding eq. (1.8)]. (ii) The flat-operator field equations (3.8') produce divergences, due to the second-derivative source terms $\bar{h}^{\alpha\beta}{}_{,\mu\nu}\bar{h}^{\mu\nu}$, in calculations of the gravitational interactions of (idealized) point particles; see, e.g., Crowley and Thorne (1977). I avoid this danger in these lectures and in RMP by not using point-particle idealizations.

I believe, and hope, that <u>all</u> of my calculations with the flat-operator field equations (3.8') have been so designed as to avoid these and other pitfalls.

3.1.4 Multipole moments of a fully relativistic, stationary system

The de Donder formulation (3.8') of the Einstein field equations can be used to extend the linearized multipole analysis of §3.1.1 to fully relativistic, stationary systems. Full details are given in §X of RMP. Here I shall sketch the methods and summarize the results.

The key idea of the analysis is to construct, in de Donder gauge, the <u>general</u> external gravitational field of a fully relativistic, stationary (time-independent) system surrounded by vacuum. The de Donder coordinates are chosen to be stationary and asymptotically flat — i.e., to satisfy, in addition to equations (3.6)-(3.10), also

$$\bar{h}^{\alpha\beta}{}_{,0} = 0; \qquad \bar{h}^{\alpha\beta} \propto 1/r \quad \text{as} \quad r \equiv (\delta_{ij}x^i x^j)^{\frac{1}{2}} \to \infty . \tag{3.11}$$

For such a system the gauge conditions (3.7) and vacuum field equations (3.8') are

$$\bar{h}^{\alpha j}{}_{,j} = 0; \; \bar{h}^{\alpha\beta}{}_{,jj} = W^{\alpha\beta} = \left(\begin{array}{l}\text{expression of quadratic order and higher in}\\ \bar{h}^{\mu\nu}\text{ and its spatial derivatives, each term}\\ \text{containing precisely two spatial derivatives}\end{array}\right). \tag{3.12}$$

Here and throughout this section I use the notation of flat-space Cartesian coordinates in which the location of spatial indices, up or down, is of no importance. Equations (3.12) can be solved by a "nonlinearity expansion" in which terms of first order are linear in $\bar{h}^{\alpha\beta}$ (or, equivalently, linear in the gravitation constant G = 1), terms of second order are quadratic, etc. The first-order part of $\bar{h}^{\alpha\beta}$, denoted ${}_1\bar{h}^{\alpha\beta}$, satisfies the linearized equations ${}_1\bar{h}^{\alpha j}{}_{,j} = 0$ and ${}_1\bar{h}^{\alpha\beta}{}_{,jj} = 0$ and thus, with specialization of gauge and mass centering of coordinates, has the general linearized-theory form (3.2):

$$\begin{aligned} {}_1\bar{h}^{00} &= \frac{4M}{r} + \sum_{\ell=2}^{\infty} \frac{4(2\ell-1)!!}{\ell!\,r^{\ell+1}} \mathcal{I}_{A_\ell} N_{A_\ell} , \\ {}_1\bar{h}^{0j} &= \sum_{\ell=1}^{\infty} \frac{4\ell(2\ell-1)!!}{(\ell+1)!\,r^{\ell+1}} \epsilon_{jka_\ell} \mathcal{S}_{kA_{\ell-1}} N_{A_{\ell-1}} , \qquad \bar{h}^{ij} = 0. \end{aligned} \tag{3.13}$$

The quadratic-order part ${}_2\bar{h}^{\alpha\beta}$ satisfies

$$ {}_2\bar{h}^{\alpha j}{}_{,j} = 0, \qquad {}_2\bar{h}^{\alpha\beta}{}_{,jj} = \left(\begin{matrix}\text{quadratic part of } W^{\alpha\beta} \\ \text{constructed from } {}_1\bar{h}^{\mu\nu}\end{matrix}\right) . \tag{3.14} $$

It is straightforward, though tedious, to solve these equations for ${}_2\bar{h}^{\alpha\beta}$ and for higher-order corrections — the kind of task ideally suited for symbolic-manipulation software on a computer; cf. Appendix of Gürsel (1982). The full details are not of interest here, but the spherical-harmonic structure of the solution is of interest. That structure is dictated by the following properties of spherical harmonics: (i) Taking gradients and inverting Laplacians does not change the spherical-harmonic order of a term; and (ii) the product of two harmonics of order ℓ and ℓ' contains pieces of orders $\ell+\ell'$, $\ell+\ell'-1$, ..., $|\ell-\ell'|$. These properties plus the quadratic-order equations (3.14) and linear-order solutions (3.13) imply that the generic term in ${}_2\bar{h}^{\alpha\beta}$ has the form

$$ {}_2\bar{h}^{\alpha\beta} \sim \frac{\mathcal{M}_{A_\ell}}{r^{\ell+1}} \cdot \frac{\mathcal{M}_{B_{\ell'}}}{r^{\ell'+1}} \sim \frac{S_{\ell+\ell'} + S_{\ell+\ell'-1} + \dots + S_{|\ell-\ell'|}}{r^{\ell+\ell'+2}} . \tag{3.15} $$

Here $\mathcal{M}_{A_\ell} = (\mathcal{I}_{A_\ell}$ or $\mathcal{S}_{A_\ell})$, $\mathcal{M}_{B_{\ell'}} = (\mathcal{I}_{B_{\ell'}}$ or $\mathcal{S}_{B_{\ell'}})$, and $S_\ell \equiv$ (something unspecified that has harmonic order ℓ and is independent of r). The key feature of this generic term is that the power $\ell+\ell'+2$ of its radial dependence is larger by a factor 2 than the order of any of its harmonics. By an extension of this argument one sees that in ${}_n\bar{h}^{\alpha\beta}$ the generic term of order $1/r^k$ has harmonics of order $k-n$ and smaller. Thus, the nonlinear parts of the solution add up to give

$$ {}_2\bar{h}^{\alpha\beta} + {}_3\bar{h}^{\alpha\beta} + \dots = \sum_{\ell=1}^{\infty} \frac{1}{r^{\ell+1}} \left(S_{\ell-1} + S_{\ell-2} + \dots + S_0\right). \tag{3.16} $$

By adding these to ${}_1\bar{h}^{\alpha\beta}$ and then computing the corresponding metric from (3.6) one obtains

$$ g_{00} = -1 + 2\frac{M}{r} - 2\frac{M^2}{r^2} + \sum_{\ell=2}^{\infty} \frac{1}{r^{\ell+1}} \left[\frac{2(2\ell-1)!!}{\ell!} \mathcal{I}_{A_\ell} N_{A_\ell} + S_{\ell-1} + \dots + S_0\right], \tag{3.17a} $$

$$ g_{0j} = \sum_{\ell=1}^{\infty} \frac{1}{r^{\ell+1}} \left[-\frac{4\ell(2\ell-1)!!}{(\ell+1)!} \epsilon_{jka_\ell} \mathcal{S}_{kA_{\ell-1}} N_{A_\ell} + S_{\ell-1} + \dots + S_0\right], \tag{3.17b} $$

$$ \begin{aligned} g_{ij} = {} & \delta_{ij}\left(1 + 2\frac{M}{r}\right) + \frac{M^2}{r^2}\left(\delta_{ij} + n_i n_j\right) \\ & + \sum_{\ell=2}^{\infty} \frac{1}{r^{\ell+1}} \left[\frac{2(2\ell-1)!!}{\ell!} \mathcal{I}_{A_\ell} N_{A_\ell} \delta_{ij} + S_{\ell-1} + \dots + S_0\right] . \end{aligned} \tag{3.17c} $$

Note the following features of this general, asymptotically flat, stationary, vacuum metric: (i) As in linearized theory, so also here, the metric is determined fully by two families of moments: the mass moments M, $\mathcal{I}_{ij}$, $\mathcal{I}_{ijk}$, ...; and the current moments $\mathcal{S}_i$, $\mathcal{S}_{ij}$, $\mathcal{S}_{ijk}$, ... (ii) The mass dipole moment vanishes because I have insisted that the coördinates be mass centered. (iii) The moments are constant STF tensors that reside in the asymptotically flat region of spacetime (i.e., rigorously speaking, at spacelike infinity). (iv) In de Donder coordinates the mass ℓ-pole moment $\mathcal{I}_{A_\ell}$ can be "read off" the metric as the $1/r^{\ell+1}$, ℓ-harmonic order part of g_{00}; and the current ℓ-pole moment $\mathcal{S}_{A_\ell}$ can similarly be read off g_{0j}.

It would be very unpleasant if one had to transform a metric to de Donder coordinates in order to compute its multipole moments. Fortunately, there are other ways of computing them. If one only wants to know the moments of order $\ell = 0, 1, 2, \ldots, \ell_{\max}$, it is adequate to find coordinates where the metric has the form

$$g_{\alpha\beta} = \eta_{\alpha\beta} + \sum_{\ell=0}^{\ell_{\max}-1} r^{-(\ell+1)} \left[S_\ell + S_{\ell-1} + \ldots + S_0\right] + O\left[r^{-(\ell_{\max}+1)}\right] , \qquad (3.18)$$

with the $1/r^2$ dipole of g_{00} vanishing. Such coordinates are called "Asymptotically Cartesian and Mass Centered to order $\ell_{\max}$-1" [ACMC − ($\ell_{\max}$-1)]. In them one can read off the first $\ell_{\max}$ moments (both mass and current) by the same prescription as in de Donder coordinates, and one will obtain the same answers as one would in de Donder coordinates (RMP §XI). Alternatively, one can compute the moments by elegant techniques at spacelike infinity, due to Geroch (1970) and Hansen (1974). As Gürsel (1982) has shown, the Geroch-Hansen prescription gives the same moments as the above, aside from normalization:

$$\mathcal{I}_{A_\ell} = \frac{1}{(2\ell-1)!!}\mathcal{M}_{A_\ell} , \qquad \mathcal{S}_{A_\ell} = \frac{(\ell+1)}{2\ell(2\ell-1)!!}\mathcal{J}_{A_\ell} , \qquad (3.19)$$

where $\mathcal{M}_{A_\ell}$ and $\mathcal{J}_{A_\ell}$ are the Geroch-Hansen moments.

3.2 Gravitational wave generation by slow-motion sources: $\lambdabar \gg L \gtrsim M$

3.2.1 Metric in the weak-field near zone

Turn attention now from stationary systems to a system with slowly changing gravitational field:

$$\lambdabar \equiv \begin{pmatrix}\text{timescale}\\ \text{of changes}\end{pmatrix} \gg L \equiv \begin{pmatrix}\text{size of}\\ \text{system}\end{pmatrix} \gtrsim M \equiv \begin{pmatrix}\text{mass of}\\ \text{system}\end{pmatrix} . \qquad (3.20)$$

Such a "slow-motion" system possesses a weak-field near zone (WFNZ)

$$(10M \text{ and } L) < r < \lambdabar/10 \qquad (3.21)$$

(Fig. 2 and associated discussion). In that WFNZ and in de Donder gauge I have developed an algorithm for computing the general "gravitational field" $\bar{h}^{\alpha\beta}$ and the spacetime metric $g_{\alpha\beta}$; see §IX of RMP. That algorithm is based on a simultaneous "nonlinearity expansion" like that used above for stationary systems, and "slow-motion expansion" — i.e., expansion of the time evolution of the metric in powers of $r/\lambdabar$.

At lowest order in $r/\lambdabar$, $\bar{h}^{\alpha\beta}$ and $g_{\alpha\beta}$ are identical to the general stationary solution (eqs. 3.13, 3.16, 3.17); except that now the multipole moments of order $\ell \geq 2$ are slowly changing functions of time t rather than constants. [The $\ell = 0$ and $\ell = 1$ moments, M = (mass) and $\mathcal{S}_j$ = (angular momentum) are forced, by the field equations, to be constant at lowest order in $r/\lambdabar$; but they change due to radiation reaction at orders $(r/\lambdabar)^5$ and higher.] The slow time changes of $\mathcal{I}_{A_\ell}(t)$ and $\mathcal{S}_{A_\ell}(t)$ produce, through the field equations (3.8') and gauge conditions (3.7) and through matching to outgoing waves at $r \gtrsim \lambdabar$, the "motional" corrections of order $r/\lambdabar$, $(r/\lambdabar)^2, \ldots$ to $\bar{h}^{\alpha\beta}$ and $g_{\alpha\beta}$.

3.2.2 Metric in the induction zone and local wave zone

In the inner parts $r \ll \lambdabar$ of the WFNZ the motional corrections are very small but the nonlinear corrections may be large; and $\bar{h}^{\alpha\beta}$, $g_{\alpha\beta}$ are essentially those of a stationary system (eqs. 3.13, 3.16, 3.17) with slowly changing moments. In the outer parts $r \gg L \gtrsim M$ of the WFNZ the nonlinear corrections are very small but as r nears $\lambdabar$ the motional corrections become large. This allows us to ignore nonlinearities when extending the $\bar{h}^{\alpha\beta}$ of the outer part of the WFNZ into the induction zone and local wave zone. In other words, we can compute $\bar{h}^{\alpha\beta}$ in the induction zone and local wave zone by constructing the general outgoing-wave solution of the <u>linearized</u>, time-dependent, vacuum field equations and gauge conditions

$$\eta^{\mu\nu}\,\bar{h}^{\alpha\beta}{}_{,\mu\nu} = 0, \qquad \bar{h}^{\alpha\beta}{}_{,\beta} = 0; \tag{3.22}$$

and by matching that solution onto the $O([r/\lambdabar]^0)$ solution (3.13), (3.16) in the WFNZ. The result is

$$\bar{h}^{00} = \frac{4M}{r} + \sum_{\ell=2}^{\infty} (-1)^{\ell}\,\frac{4}{\ell!}\left[\frac{1}{r}\,\mathcal{I}_{A_\ell}(t-r)\right]_{,A_\ell} + \begin{pmatrix}\text{small nonlinear}\\ \text{terms}\end{pmatrix}, \tag{3.23a}$$

$$\begin{aligned}\bar{h}^{0j} = {} & \frac{2\epsilon_{jpq}\mathcal{S}_p n_q}{r^2} + \sum_{\ell=2}^{\infty} (-1)^{\ell}\,\frac{4\ell}{(\ell+1)!}\left[\frac{1}{r}\,\epsilon_{jpq}\mathcal{S}_{pA_{\ell-1}}(t-r)\right]_{,qA_{\ell-1}} \\ & - \sum_{\ell=2}^{\infty} (-1)^{\ell}\,\frac{4}{\ell!}\left[\frac{1}{r}\,\dot{\mathcal{I}}_{jA_{\ell-1}}(t-r)\right]_{,A_{\ell-1}} + \begin{pmatrix}\text{small nonlinear}\\ \text{terms}\end{pmatrix},\end{aligned} \tag{3.23b}$$

$$\begin{aligned}\bar{h}^{jk} = \sum_{\ell=2}^{\infty} \Bigg\{ & (-1)^{\ell}\,\frac{4}{\ell!}\left[\frac{1}{r}\,\ddot{\mathcal{I}}_{jkA_{\ell-2}}(t-r)\right]_{,A_{\ell-2}} + (-1)^{\ell+1}\,\frac{8\ell}{(\ell+1)!}\times \\ & \times\left[\frac{1}{r}\,\epsilon_{pq(j}\dot{\mathcal{S}}_{k)pA_{\ell-2}}(t-r)\right]_{,qA_{\ell-2}}\Bigg\} + (\text{small nonlinear terms}).\end{aligned} \tag{3.23c}$$

Here dots denote $\partial/\partial t$, and as indicated the moments with $\ell \geq 2$ are to be regarded as functions of $t-r$. The dominant nonlinear corrections to this solution are discussed in §IX of RMP; see also equation (3.25) below. The metric can be computed from (3.23) via equations (3.6), which reduce to

$$g_{\alpha\beta} = \eta_{\alpha\beta} + h_{\alpha\beta} + \text{nonlinearities}, \qquad h_{\alpha\beta} = \bar{h}_{\alpha\beta} - \tfrac{1}{2}\,\bar{h}\eta_{\alpha\beta}\ , \tag{3.24}$$

where indices on $\bar{h}^{\alpha\beta}$ are lowered (as usual) using $\eta_{\mu\nu}$.

The matching of the solution (3.23) onto that of the WFNZ can be done either by elementary techniques, which require care and thought, or by the sophisticated technique of "matched asymptotic expansions" (see the lecture by Kates in this volume; also those of Damour), which do the job with less danger of error.

* * * * *

Exercise 19. Show that (3.23) without nonlinear terms is an exact solution of the linearized field equations and gauge conditions (3.22). Then, at radii $r \ll \lambdabar$, expand (3.23) in powers of $r/\lambdabar$ and show that the leading term, of

$O([r/\bar{\lambda}]^0)$, is identical to the linear part (3.13) of the WFNZ field.

3.2.3 Gravitational-wave field in local wave zone

The gravitational-wave field h^{TT}_{jk} of the local wave zone can be computed from expression (3.23c) by letting the spatial derivatives all act on $\mathcal{I}$ and $\mathcal{S}$ (so as to keep only the $1/r$ part of the field); and by then taking the TT part. The result is

$$h^{TT}_{jk} = \left\{ \sum_{\ell=2}^{\infty} \frac{1}{r} \frac{4}{\ell!} \, {}^{(\ell)}\mathcal{I}_{jkA_{\ell-2}}(t-r)\, N_{A_{\ell-2}} + \sum_{\ell=2}^{\infty} \frac{1}{r} \frac{8\ell}{(\ell+1)!}\, \epsilon_{pq(j}\, {}^{(\ell)}\mathcal{S}_{k)pA_{\ell-2}}(t-r)\, n_q N_{A_{\ell-2}} \right\}^{TT} \left\{ 1 + \underbrace{O\!\left(\frac{M}{\bar{\lambda}} \ln \frac{\bar{\lambda}}{L}\right)}_{\text{effects of nonlinearities}} \right\} . \tag{3.25}$$

Here a prefix superscript (ℓ) means "take ℓ time derivatives"

$$^{(\ell)}\mathcal{I} \equiv (\partial/\partial t)^{\ell}\mathcal{I} \; ; \tag{3.26}$$

and I have indicated the magnitude of the cumulative effects of nonlinearities, integrated up from the inner part of the WFNZ into the local wave zone, which in effect cause the multipole moments of the radiation field to differ slightly from those one would measure in the inner part of the weak-field near zone; see §IX.H of RMP.

Note that the mass quadrupole part of the radiation field (3.25) has the familiar form first derived (in different notation) by Einstein (1918):

$$h^{TT}_{jk} = \frac{2}{r} \ddot{\mathcal{I}}^{TT}_{jk}(t-r). \tag{3.27}$$

For most slow-motion systems these mass quadrupole waves will dominate; but when quadrupole motions are suppressed by special symmetries (e,g., in torsional oscillations of neutron stars, §3.2.7 below), other moments may dominate. Note that, in the absence of suppression due to symmetries, the magnitudes of the various multipole components of the waves are

$$\left(h^{TT}_{jk}\right)_{\text{mass } \ell\text{-pole}} \sim \frac{M}{r}\left(\frac{L}{\bar{\lambda}}\right)^{\ell}, \quad \left(h^{TT}_{jk}\right)_{\text{current } \ell\text{-pole}} \sim \frac{M}{r}\, v \left(\frac{L}{\bar{\lambda}}\right)^{\ell} \tag{3.28}$$

(cf. eq. 3.4). Typically the internal velocity v will be of order $L/\bar{\lambda}$, so

$$\begin{gathered}(\text{current quadrupole waves}) \sim (\text{mass octupole waves}), \\ (\text{current } \ell\text{-pole waves}) \sim (\text{mass } [\ell+1]\text{-pole waves}).\end{gathered} \tag{3.29}$$

This is the same pattern as one sees in electromagnetism, for which "electric" multipoles are the analogs of "mass" multipoles and "magnetic" multipoles are the analogs of "current" multipoles.

3.2.4 Slow-motion method of computing wave generation

Equations (3.17) and (3.25) are the foundation for the slow-motion method of computing gravitational-wave generation (RMP §XII): (i) Analyze the near-zone structure and evolution of any slow-motion ($\bar{\lambda} \gg L \gtrsim M$) system in any convenient coordinate system and by any approximation scheme that gives, with reasonable fractional accuracy, the time evolution of the system's asymmetries. [One attractive approximation scheme is the "instantaneous gravity" method, in which one sets to zero all time derivatives of the metric (but not of the matter variables) when solving the near-zone Einstein field equations; see, e.g., Thorne (1983) and Schumaker and Thorne (1983).] (ii) From the near-zone analysis obtain an approximation to the system's external gravitational field which, at any moment, satisfies the time-independent, vacuum Einstein equations. (iii) Compute the dominant multipole moments of that quasistationary field (moments with largest values of $^{(\ell)}\mathcal{I}_{A_\ell}$ or $^{(\ell)}\mathcal{S}_{A_\ell}$) either by transforming to de Donder or ACMC coordinates and comparing with equations (3.17), or by the methods of Geroch (1970) and Hansen (1974) plus equation (3.19). (iv) Insert those moments into equation (3.25) to obtain the radiation field in the local wave zone.

3.2.5 Example: Rigidly rotating neutron star

As an example, consider the gravitational waves produced by a slowly rotating neutron star (pulsar) idealized as a non-axisymmetric, fully relativistic body which rotates rigidly. Full details are given in Gürsel and Thorne (1983); the main ideas will be sketched here.

The star can rotate rigidly (distance between every pair of neighboring "material particles" forever fixed) only to first order in the angular velocity Ω. At order $(\Omega L)^2$ there is a Lorentz contraction of distances; and as the star's angular velocity precesses that Lorentz contraction changes. Thus, the Gürsel-Thorne analysis, which assumes rigid rotation and works to first order in Ω, has fractional errors of order ΩL.

Gürsel and Thorne show, by a de Donder-gauge analysis of the star's interior [step (i) of "slow-motion wave-generation method"] that to first order in Ω the star's angular momentum $\mathcal{S}_j$ and its angular velocity Ω_j (which are both spatial vectors residing in the nearly flat, weak-field near zone) are proportional to each other

$$\mathcal{S}_j = I_{jk}\Omega_k; \tag{3.30a}$$

and that their ratio, the moment of inertia tensor, is symmetric and rotates with angular velocity Ω_j

$$I_{jk} = I_{kj}, \quad \dot{I}_{jk} = \epsilon_{jpq}\Omega_p I_{qk} + \epsilon_{kpq}\Omega_p I_{qj}. \tag{3.30b}$$

Of course, the angular momentum is conserved (aside from negligible radiation-reaction changes)

$$\dot{\mathcal{S}}_j = 0. \tag{3.30c}$$

Equations (3.30) are identical to the classical Euler equations which govern the precession of a rigidly rotating, nongravitating body (Goldstein 1980). Thus, any fully relativistic, slowly ($\Omega L \ll 1$) and rigidly rotating body undergoes a free precession which is identical to that of a nongravitating body with the same moment of inertia I_{jk} and the same angular momentum $\mathcal{S}_j$. The only influence of relativistic gravity will be through its influence on the values of the components of the moment of inertia tensor I_{jk}; cf. Hartle (1973). Gürsel and Thorne go on to show that the mass moments $\mathcal{I}_{A_\ell}$, which characterize g_{00} in the weak-field near zone, like I_{jk} rotate with angular velocity Ω_j:

$$\dot{\mathcal{I}}_{jk} = \epsilon_{jpq}\Omega_p\mathcal{I}_{qk} + \epsilon_{kpq}\Omega_p\mathcal{I}_{jq}, \quad \text{and similarly for } \mathcal{I}_{A_\ell}. \tag{3.31}$$

The mass quadrupole $\mathcal{I}_{jk}$ will be the dominant source of gravitational waves unless the star has very unexpected symmetries. The waves that it produces are described by the standard quadrupole-moment formula

$$h^{TT}_{jk} = (2/r)\,\ddot{\mathcal{I}}^{TT}_{jk}(t-r). \tag{3.32}$$

Because the precessional equations (3.30) are identical to those of Euler and the waves are given by the same standard quadrupole moment formula (3.32) as one often uses for weakly gravitating systems, one might expect the waves from a fully relativistic, rigidly and slowly rotating body to be the same as those from a weakly gravitating body with the same moment of inertia. However, I doubt that this is so, because I suspect that relativistic gravity destroys the classical relationship

$$\mathcal{I}_{jk} = I_{jk} - \frac{1}{3}\,I_{ii}\,\delta_{jk} \tag{3.33}$$

between the quadrupole moment and the moment of inertia.

Zimmermann and Szedenitz (1979) and Zimmermann (1980) have computed in detail the quadrupole waves from a rigidly and slowly rotating body under the assumption that the classical relationship (3.33) is preserved. They show that the spectrum of the waves is rather rich and contains much detailed information about the star's angular momentum vector and moment of inertia tensor. Future theoretical studies should probe the possible breakdown of the classical relation (3.33) and should quantify deviations from rigid-body rotation due to the finiteness of the shear modulus and bulk modulus of neutron-star matter.

3.2.6 Example: compact binary system

Consider a binary system with stars sufficiently compact that tidal distortions of each other can be ignored (this is frequently true), and with separation between stars that is large compared to gravitational radii. Then general relativistic "near-zone" analyses (e.g., Damour in this volume; or, for the case of two black holes, D'Eath 1975) show that the orbital motions are Keplerian, aside from post-Newtonian corrections of size (gravitational radii)/(separation of stars); and this is true no matter how strong the stars' internal gravity may be. Moreover, the quadrupole moment which one reads off the weak-field, near-zone metric in de Donder gauge (eq. 3.17; Damour in this volume) is the same and evolves the same as that which one would compute for the Kepler problem using Newtonian techniques; and thus the gravitational waves obtained by inserting that quadrupole moment into $h^{TT}_{jk} = (2/r)\ddot{\mathcal{I}}^{TT}_{jk}$ are the same as one would compute for a nearly Newtonian system with the same masses and semimajor axis. For details of those waves see Peters and Mathews (1963).

3.2.7 Example: torsional oscillations of a neutron star

"Glitches" observed in the timing of pulsars are thought to be due to starquakes, i.e., due to ruptures in the crystalline crust of the neutron star. Such ruptures may trigger torsional oscillations of the star with a restoring force, due to the crystal's shear modulus, which is sufficiently small that the oscillations are slow ($\lambdabar \gg L$). In such oscillations its mass-energy density T^{00} remains constant while the momentum density T^{0j} oscillates; and, as a result, the star's mass quadrupole moment $\mathcal{I}_{jk}$ is constant but its current quadrupole moment $\mathcal{S}_{jk}$ oscillates. The resulting gravitational waves are thus current quadrupole rather than mass quadrupole

$$h^{TT}_{jk} = \left[\frac{8}{3r}\,\epsilon_{pq(j}\ddot{\mathcal{S}}_{k)p}(t-r)n_q\right]^{TT}. \tag{3.34}$$

Schumaker and Thorne (1983) have analyzed such torsional oscillations in detail using perturbation theory and have derived, in the "instantaneous gravity approximation", an eigenequation that governs the oscillations and determines the current quadrupole moment $\mathcal{S}_{jk}(t)$ for insertion into the gravity-wave formula (3.34).

3.3 Multipole decomposition of arbitrary waves in the local wave zone

3.3.1 The radiation field

The gravitational waves from any source — slow-motion or fast, weak-gravity or strong — can be decomposed into multipole components in the local wave zone. The multipole moments can be computed as surface integrals of the radiation field (RMP eq. 4.11):

$$ {}^{(\ell)}\mathcal{I}_{A_\ell} = \left[\frac{\ell(\ell-1)(2\ell+1)!!}{2(\ell+1)(\ell+2)} \frac{r}{4\pi} \int h^{TT}_{a_1 a_2} n_{a_3} \dots n_{a_\ell} \, d\Omega \right]^{STF} , \qquad (3.35a) $$

$$ {}^{(\ell)}\mathcal{S}_{A_\ell} = \left[\frac{(\ell-1)(2\ell+1)!!}{4(\ell+2)} \frac{r}{4\pi} \int \epsilon_{a_1 jk} n_j h^{TT}_{k a_2} n_{a_3} \dots n_{a_\ell} \, d\Omega \right]^{STF} ; \qquad (3.35b) $$

and the field can be reconstructed as a sum over the multipole moments — the same sum as we encountered in the theory of slow-motion sources (eq. 3.25)

$$ h^{TT}_{jk} = \sum_{\ell=2}^{\infty} \left\{ \frac{1}{r} \frac{4}{\ell!} \, {}^{(\ell)}\mathcal{I}_{jkA_{\ell-2}}(t-r) N_{A_{\ell-2}} \right. $$
$$ \left. + \frac{1}{r} \frac{8\ell}{(\ell+1)!} \epsilon_{pq(j} \, {}^{(\ell)}\mathcal{S}_{k)pA_{\ell-2}}(t-r) n_q N_{A_{\ell-2}} \right\}^{TT} . \qquad (3.36) $$

For slow-motion sources the lowest few moments will dominate; but for fast-motion sources the radiation may be highly directional and many moments may contribute. See, e.g., Kovács and Thorne (1978) for the example of "gravitational bremsstrahlung radiation", in which the radiation from one particle flying past another at a speed $v \simeq 1$ is beamed forward into a cone of half angle $\sim \gamma^{-1} = (1-v^2)^{1/2} << 1$, and moments $\ell = 2, 3, 4, \dots, \gamma$ all contribute significantly to the waves. In such cases multipole expansions are not very useful.

3.3.2 The energy, momentum, and angular momentum carried by the waves

In the local wave zone the gravitational waves, which have $\ell \geq 2$, coexist with the (nearly) time-independent $\ell = 0$ (mass) and $\ell = 1$ (angular momentum) parts of the source's gravitational field; in TT gauge and neglecting nonlinearities and induction terms the total spacetime metric is

$$ g_{00} = -1 + \underbrace{2M/r}_{\ell=0} , \quad g_{0j} = \underbrace{(-2/r^2)\epsilon_{jk\ell} S_k n_\ell}_{\ell=1} , \quad g_{jk} = (1 + \underbrace{2M/r}_{\ell=0})\delta_{jk} + \underbrace{h^{TT}_{jk}}_{\ell\geq 2} . \qquad (3.37) $$

This metric is written in coordinates that coincide with the asymptotic rest frame of the source; in this frame the source's linear momentum P_j vanishes; i.e., the 4-momentum (a 4-vector residing in the asymptotically flat region) is $\vec{P} = M\partial/\partial t$.

The gravitational waves carry 4-momentum and angular momentum away from the source, thereby causing changes in the asymptotic rest frame, in M, and in S_j. A detailed analysis of those changes is given in chapters 19 and 20 of MTW; here I shall sketch only one variant of the main ideas.

The foundation for the analysis is the quantity

$$H^{\mu\alpha\nu\beta} \equiv \mathfrak{g}^{\mu\nu}\mathfrak{g}^{\alpha\beta} - \mathfrak{g}^{\alpha\nu}\mathfrak{g}^{\mu\beta}, \tag{3.38}$$

where $\mathfrak{g}^{\alpha\beta}$ is the metric density of equations (3.6) and Lorentz gauge is *not* being imposed. In the asymptotic rest frame of the source, where the metric is (3.37) plus nonlinearities and induction terms, the surface integral of $H^{\mu\alpha 0j}{}_{,\alpha}$ plucks out the $1/r$ $\ell = 0$ and $\ell = 1$ parts of $g_{\alpha\beta}$ (which have zero contribution from nonlinearities and induction terms); i.e., it gives the 4-momentum of the source:

$$P^{\mu} = \frac{1}{16\pi}\oint H^{\mu\alpha 0j}{}_{,\alpha}\, d^2S_j = \begin{cases} 0 \text{ for } P^j \\ M \text{ for } P^0, \end{cases} \tag{3.39}$$

where d^2S_j is the surface element computed as though spacetime were flat. The rate of change of the 4-momentum is computed, with the help of Einstein's vacuum field equations in the form $H^{\mu\alpha\nu\beta}{}_{,\alpha\beta} = -H^{\mu\alpha\beta\nu}{}_{,\alpha\beta} = 16\pi(-g)t^{\alpha\beta}_{LL}$ (MTW eq. 20.21):

$$\begin{aligned} \frac{dP^{\mu}}{dt} &= \frac{1}{16\pi}\oint H^{\mu\alpha 0j}{}_{,\alpha 0}\, d^2S_j = \frac{1}{16\pi}\oint \left[H^{\mu\alpha\beta j}{}_{,\alpha\beta} - H^{\mu\alpha ij}{}_{,\alpha i}\right] d^2S_j \\ &= -\oint(-g)t^{\mu j}_{LL}\, d^2S_j - \frac{1}{16\pi}\int \underbrace{H^{\mu\alpha ij}{}_{,\alpha ij}}_{0}\, dx^1dx^2dx^3. \end{aligned} \tag{3.40}$$

The volume integral vanishes by symmetry; and the conversion from surface integral to volume integral requires the topology of the space slices to be Euclidean — which it always can be for astrophysically realistic systems, including those with black holes (see Fig. 5). By averaging equation (3.40) over Δt = (several ƛ) and noting that the average of the Landau-Lifshitz pseudotensor is equal to the Isaacson stress-energy tensor for the gravitational waves (MTW exercise 35.19), we obtain

$$\langle dM/dt\rangle = -\oint T^{0r}_{(W)}\, r^2 d\Omega, \quad \langle dP^j/dt\rangle = -\oint T^{jr}_{(W)}\, r^2 d\Omega. \tag{3.41}$$

When the waves are decomposed into multipoles (eq. 3.36) these integrals give (RMP eqs. 4.16' and 4.20'):

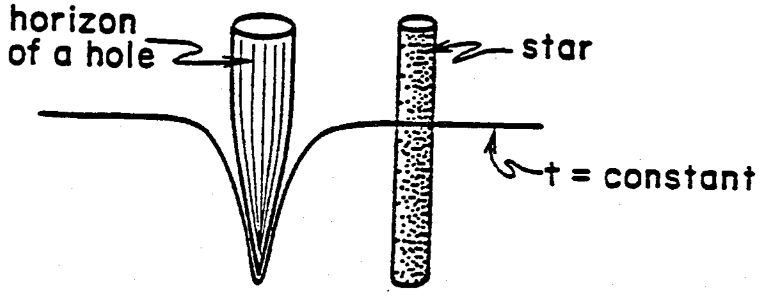


Fig. 5. Spacetime diagram showing how the slices of constant time can be chosen everywhere spacelike and have Euclidean topology even in the presence of a black hole.

$$\left\langle \frac{dM}{dt} \right\rangle = -\sum_{\ell=2}^{\infty} \left\{ \frac{(\ell+1)(\ell+2)}{(\ell-1)\ell\cdot\ell!(2\ell+1)!!} \left\langle {}^{(\ell+1)}\mathcal{I}_{A_\ell} \; {}^{(\ell+1)}\mathcal{I}_{A_\ell} \right\rangle \right. \tag{3.42}$$

$$\left. + \frac{4\ell(\ell+2)}{(\ell-1)\cdot(\ell+1)!(2\ell+1)!!} \left\langle {}^{(\ell+1)}\mathcal{S}_{A_\ell} \; {}^{(\ell+1)}\mathcal{S}_{A_\ell} \right\rangle \right\} ,$$

$$\left\langle \frac{dP_j}{dt} \right\rangle = -\sum_{\ell=2}^{\infty} \left\{ \frac{2(\ell+2)(\ell+3)}{\ell(\ell+1)!(2\ell+3)!!} \left\langle {}^{(\ell+2)}\mathcal{I}_{jA_\ell} \; {}^{(\ell+1)}\mathcal{I}_{A_\ell} \right\rangle \right.$$

$$+ \frac{8(\ell+3)}{(\ell+1)!(2\ell+3)!!} \left\langle {}^{(\ell+2)}\mathcal{S}_{jA_\ell} \; {}^{(\ell+1)}\mathcal{S}_{A_\ell} \right\rangle \tag{3.43}$$

$$\left. + \frac{8(\ell+2)}{(\ell-1)(\ell+1)!(2\ell+1)!!} \epsilon_{jpq} \left\langle {}^{(\ell+1)}\mathcal{I}_{pA_{\ell-1}} \; {}^{(\ell+1)}\mathcal{S}_{qA_{\ell-1}} \right\rangle \right\} .$$

Note that for typical slow-motion sources, with moments of order (3.4) and $v \sim L/\bar{\lambda}$, the mass loss is predominantly due to the mass quadrupole moment beating against itself

$$\langle dM/dt \rangle = -(1/5) \langle \dddot{\mathcal{I}}_{jk} \dddot{\mathcal{I}}_{jk} \rangle \sim M^2L^4/\bar{\lambda}^6 ; \tag{3.44}$$

and the momentum change is due to the mass quadrupole beating against the mass octupole <u>and</u> against the current quadrupole:

$$\left\langle \frac{dP_j}{dt} \right\rangle = -\frac{2}{63} \langle \dddot{\mathcal{I}}_{ab} \ddddot{\mathcal{I}}_{abj} \rangle - \frac{16}{45} \epsilon_{jpq} \langle \dddot{\mathcal{I}}_{pa} \dddot{\mathcal{S}}_{qa} \rangle \sim \frac{M^2L^5}{\bar{\lambda}^7} . \tag{3.45}$$

As the momentum of the source changes, its asymptotic rest frame changes. Since I have formulated my discussion of the external fields of slow-motion sources in mass-centered de Donder coordinates which coincide with the asymptotic rest frame, in applying my equations one must continually readjust the coordinates as time passes.

The intrinsic angular momentum of the source can be computed by a surface integral analogous to (3.39), which picks out the $1/r^2$ dipole part of the metric (3.37):

$$\mathcal{S}_j = \frac{1}{16\pi} \oint \epsilon_{jpq} \left(x^p H^{q\alpha 0i}{}_{,\alpha} + H^{pi0q}\right) d^2S_i . \tag{3.46}$$

By manipulations analogous to (3.40) one can show that

$$d\mathcal{S}_j/dt = -\oint \epsilon_{jpq} x^p (-g) t^{qr}_{LL} r^2 d\Omega . \tag{3.47}$$

By computing the Landau-Lifshitz pseudotensor (MTW eq. 20.22) for the local-wave-zone metric (3.37), inserting it into (3.47), and then averaging over $\Delta t =$ (several $\bar{\lambda}$), we obtain (RMP eq. 4.22)

$$\left\langle \frac{d\mathcal{S}_j}{dt} \right\rangle = \frac{1}{16\pi} \oint \epsilon_{jpq} x^p \left\langle -\left(h^{TT}_{qa} \dot{h}^{TT}_{ab}\right)_{,b} + \tfrac{1}{2} h^{TT}_{ab,q} \dot{h}^{TT}_{ab} \right\rangle r^2 d\Omega . \tag{3.48}$$

When the multipole expansion (3.36) is inserted this becomes (RMP eq. 4.23'):

$$\left\langle\frac{dS_j}{dt}\right\rangle = -\sum_{\ell=2}^{\infty}\left\{\frac{(\ell+1)(\ell+2)}{(\ell-1)\ell!(2\ell+1)!!}\,\epsilon_{jpq}\left\langle {}^{(\ell)}\mathcal{I}_{pA_{\ell-1}}\;{}^{(\ell+1)}\mathcal{I}_{qA_{\ell-1}}\right\rangle\right.$$
$$\left.+\frac{4\ell^2(\ell+2)}{(\ell-1)(\ell+1)!(2\ell+1)!!}\,\epsilon_{jpq}\left\langle {}^{(\ell)}\mathcal{S}_{pA_{\ell-1}}\;{}^{(\ell+1)}\mathcal{S}_{qA_{\ell-1}}\right\rangle\right\}\,. \tag{3.49}$$

For typical slow-motion sources the dominant term is mass quadrupole beating against mass quadrupole:

$$\left\langle\frac{dS_j}{dt}\right\rangle = \frac{2}{5}\,\epsilon_{jpq}\left\langle \ddot{\mathcal{I}}_{pa}\,\dddot{\mathcal{I}}_{qa}\right\rangle \sim \frac{M^2L^4}{ƛ^5}\,. \tag{3.50}$$

The above analysis encounters serious difficulties for a source which changes its asymptotic rest frame significantly in a few gravity-wave periods, i.e., for which $|(dP_j/dt)ƛ| \sim M$. Because the local wave zone, where one constructs the above surface integrals, must have a size $\Delta r \gg ƛ$, the momentum of such a source is not well defined there (it is changing too fast); and thus there is no clean prescription for constructing the source's asymptotic rest frame or for "mass centering" the coordinates in it. As a result, the instantaneous mass M and linear momentum P_j of the source (which depend on the choice of time t) are somewhat ill defined; and the instantaneous angular momentum S_j is even more ill defined because it is sensitive to the mass centering (factor of x^p in eqs. 3.47, 3.48). This difficulty is discussed, using the Bondi-Sachs formulation of gravitational waves at "future null infinity", in a lecture by Ashtekar in this volume.

Fortunately for theorists (unfortunately for experimenters) all realistic astrophysical sources are believed to radiate momentum only weakly

$$|dP_j/dt| \ll M/ƛ \tag{3.51}$$

(see the lectures by Eardley) and thus have asymptotic rest frames that are well enough defined for the above analysis to be well founded.

3.3.3 Order-of-magnitude formulas

For typical slow-motion sources the gravitational-wave amplitude at earth (eq. 3.27 propagated on out to earth through a nearly flat universe) will be

$$h^{TT}_{jk} \simeq \frac{2}{r}\,\ddot{\mathcal{I}}^{TT}_{jk} \sim \frac{M}{r}\left(\frac{L}{ƛ}\right)^2 \sim \frac{G}{c^4}\,\frac{\text{(internal kinetic energy of source)}}{r}$$

$$\sim \text{(Newtonian potential at earth produced by internal kinetic energy of source)}$$

$$\sim 10^{-17} \times \frac{\text{(internal kinetic energy)}}{\text{(total mass-energy of Sun)}} \times \frac{\text{(distance to galactic center)}}{\text{(distance to source)}}\,. \tag{3.52}$$

In using this formula one must include only the internal kinetic energy associated with quadrupolar-type (nonspherical) motions. The total power carried by such sources is expressed most conveniently in terms of the "universal power unit"

$$\mathcal{L}_0 = c^5/G = 1 = 3.63 \times 10^{59}\ \text{erg/sec} = 2.03 \times 10^5\ M_\odot c^2/\text{sec} \tag{3.53}$$

and the source's internal power flow $\mathcal{L}_{int} = (\text{internal kinetic energy})/\bar{\lambda}$:

$$\mathcal{L}_{GW} \simeq \frac{1}{5}\left\langle \dddot{\mathcal{I}}_{jk}\,\dddot{\mathcal{I}}_{jk}\right\rangle \sim \left(\frac{ML^2}{\bar{\lambda}^3}\right)^2 \sim \left(\frac{\mathcal{L}_{int}}{\mathcal{L}_o}\right)^2 \mathcal{L}_o\,. \qquad (3.54)$$

Realistic astrophysical sources — even those with fast, large-amplitude motions and strong internal gravity — are not expected to deviate strongly from these order-of-magnitude formulas; see the lectures of Eardley. Moreover, all calculations to date suggest that no realistic source can radiate away a substantial fraction of its mass more quickly than the light travel time across its gravitational radius; i.e.,

$$\mathcal{L}_{GW} \lesssim M/M = 1 = \mathcal{L}_o \quad \text{for all sources} \qquad (3.55a)$$

(a limit first suggested, so far as I know, by Dyson 1963); and, correspondingly, that the gravity-wave amplitude will always be smaller than

$$h^{TT}_{jk} \lesssim \bar{\lambda}/r \qquad \text{for all sources.} \qquad (3.55b)$$

3.4 Radiation reaction in slow-motion sources

There are three approaches to the theory of gravitational radiation reaction in slow-motion sources, each of which is sufficiently rigorous to make me happy; but none of which is sufficiently rigorous to make the most mathematically careful of my colleagues happy (see, e.g., Ehlers, Rosenblum, Goldberg, and Havas 1976). I shall discuss each of these approaches in turn.

3.4.1 Method of conservation laws

The method of conservation laws is based on equations (3.41) and (3.48), which express the rates of change of the source's mass M, momentum P_j, and angular momentum $\mathcal{S}_j$ in terms of integrals over its gravitational waves, and thence is based on expressions (3.42)-(3.45) and (3.49)-(3.50) for $\dot{M}$ and $\dot{P}_j$ in terms of multipole moments that are computable by "instantaneous-gravity", near-zone analyses. These formulas for $\dot{M}$, $\dot{P}_j$, and $\dot{\mathcal{S}}_j$ ("conservation laws") rely, ultimately, on the vacuum Einstein equations (e.g., through the third equality of equation 3.40).

It is crucial for radiation-reaction theory that the M, P_j, and $\mathcal{S}_j$ of these conservation laws are physically measurable (e.g., by Kepler's laws and the precession of gyroscopes) in the weak-field near zone or the local wave zone, and correspondingly that they are computable in terms of the physical near-zone properties of the gravitating system. For example, in the case of a compact binary system (e.g., the binary pulsar), near-zone analyses give

$$\begin{aligned} M &= m_1 + m_2 - \frac{m_1 m_2}{2a} + M_{pN} + M_{p^2N} \\ \mathcal{S} \equiv |\underset{\sim}{\mathcal{S}}| &= \left[\frac{m_1{}^2 m_2{}^2}{m_1 + m_2}\, a(1-e^2)\right]^{1/2} + \mathcal{S}_{pN} + \mathcal{S}_{p^2N}\,. \end{aligned} \qquad (3.56)$$

Here m_1 and m_2 are the masses of each of the stars as measured by Kepler's laws and as manifest in the stars' external metrics; a and e are the semimajor axis and eccentricity of the orbits at Newtonian order; and M_{pN}, M_{p^2N}, $\mathcal{S}_{pN}$, $\mathcal{S}_{p^2N}$ are post-

Newtonian and post-post Newtonian contributions. Moreover, for such a binary the conservation laws (3.44) and (3.50) reduce to

$$\left\langle\frac{dM}{dt}\right\rangle = -\frac{32}{5}\,\frac{m_1{}^2 m_2{}^2 (m_1+m_2)}{a^5(1-e^2)^{7/2}}\left(1+\frac{73}{24}e^2+\frac{37}{96}e^4\right) ,$$

$$\left\langle\frac{d\mathcal{S}}{dt}\right\rangle = -\frac{32}{5}\,\frac{m_1{}^2 m_2{}^2 (m_1+m_2)^{\frac{1}{2}}}{a^{7/2}(1-e^2)^2}\left(1+\frac{7}{8}e^2\right) \qquad (3.57)$$

(Peters and Mathews 1963, Peters 1964).

Consider the evolution of such a binary over time scales $\Delta t \gg M_{pN}/\langle dM/dt\rangle \simeq$ (10^3 years for the binary pulsar). It is "physically obvious" (or one can show by a careful analysis such as that in Damour's lectures) that m_1 and m_2 are unaffected by the gravity-wave emission. Moreover, over these long time scales the changes of M_{pN}, M_{p^2N}, $\mathcal{S}_{pN}$, and $\mathcal{S}_{p^2N}$ are negligible compared to the much larger M and $\mathcal{S}$ carried off in the radiation. Thus, the changes of M and $\mathcal{S}$ must be fully accounted for by changes of a and e — changes that are fully determined by equations (3.56) and (3.57). And from those changes one can compute the change of orbital period $P = 2\pi[a^3/(m_1+m_2)]^{1/2}$, the most directly measurable quantity.

For evolution of the binary pulsar on time scales $P = 7.75$ hours $\ll \Delta t \ll$ 1000 years (corresponding to current measurements), the same argument gives the same result (which agrees with the measurements), if one makes a "highly plausible" assumption: that $M_{pN} + M_{p^2N}$ and $\mathcal{S}_{pN} + \mathcal{S}_{p^2N}$ are not sharply changing functions of the (nearly conserved) orbital parameters such as a and e, and thus cannot account for any significant piece of the changes in M and $\mathcal{S}$. Of course, one can only feel fully comfortable about this conclusion after detailed pN and p^2N calculations have verified this assumption; see the lectures of Damour.

3.4.2 Radiation reaction potential

For systems which, unlike the binary pulsar, have weak internal gravity as well as slow motions, one can compute radiation reaction using Newtonian gravity augmented by Burke's (1969) radiation-reaction potential:

$$\underset{\sim}{F}_{grav} = -m\underset{\sim}{\nabla}\Phi ; \qquad \Phi = \Phi_{Newton} + \Phi_{react} ;$$

$$\Phi_{react} = \frac{1}{5}\,{}^{(5)}\mathcal{I}_{jk}\,x^j x^k . \qquad (3.58)$$

Here $\underset{\sim}{F}_{grav}$ is the gravitational force that acts on a material element of mass m.

Physically Φ_{react} results from matching the near-zone gravitational field onto outgoing gravitational waves. If the near-zone field were matched onto standing waves Φ_{react} would be zero; if it were matched onto ingoing waves Φ_{react} would change sign. Mathematically one derives the equation of motion (3.58) by constructing the outgoing-wave solution of the Einstein equations in any convenient gauge, matching it onto the near-zone solution, identifying the largest terms in the near-zone metric which are sensitive to outgoing waves versus ingoing waves, and discarding all terms except these sensitive ones and the terms of Newton. By an appropriate change of gauge one then obtains equations of motion of the form (3.58).

I have a confession to make: The derivation along these lines given in §36.11 of MTW is flawed. As Walker and Will (1980) point out, when one works in de Donder gauge (as I did in writing §36.11), one obtains reaction terms of magnitude ${}^{(3)}\mathcal{I}_{jk}$ in the near-zone metric when one matches onto outgoing waves. Although these terms are "pure gauge", i.e., have no direct physical consequences, they produce nonlinear corrections in the final gauge change, corrections which I mistakenly ignored in MTW but which are cancelled by a nonlinear iteration of the Einstein equations that I also mistakenly ignored. The reason for my sloppiness in writing MTW is that I had previously derived the radiation-reaction potential (3.58) working not in de Donder gauge but in "Regge-Wheeler gauge" (Thorne 1969); and in that gauge ${}^{(3)}\mathcal{I}_{jk}$ terms never arise and a final gauge change and nonlinear iteration are not needed. Having been very careful in my Regge-Wheeler-type derivation, I was highly confident of the final result — and, buoyed by this confidence, I became careless when constructing the de Donder-gauge proof in MTW.

Historically, Peres (1960) gave the first correct analysis of gravitational radiation reaction by the technique of identifying the dominant terms sensitive to the outgoing-wave boundary condition. However, Peres did not write his answer in terms of a Newton-type radiation-reaction potential $\Phi_{\rm react}$; that was first done by Burke (1969).

3.4.3 pN, p^2N, and $p^{2.5}$N iteration of the field equations

The method (3.58) of the radiation reaction potential discards all post-Newtonian and post-post-Newtonian effects in the evolution of the system — even though radiation reaction is a somewhat smaller, post$^{2.5}$-Newtonian phenomenon over times of order $\bar\lambda$. The justification is that radiation reaction, unlike pN and p^2N effects, produces secular changes of such quantities as the period of a binary system; and those secular changes can build up over long times $\Delta t \gg \bar\lambda$, becoming ultimately much larger than post-Newtonian order. However, over shorter timescales ($\Delta t \lesssim 1000$ years in the case of the binary pulsar) the cumulative effects do not exceed post-Newtonian order; and thus for full confidence in one's results one should augment the Newtonian forces and $p^{2.5}$N radiation-reaction forces by a full pN and p^2N analysis of the system.

Damour, in his lectures in this volume, describes the history of attempts at such full analyses and presents a beautiful, full analysis of his own for the special case of binary systems with compact (white-dwarf, neutron-star, and black-hole) members.

3.5 Gravitational-wave generation by fast-motion sources: $\bar\lambda \lesssim L$

Elsewhere (Thorne 1977) I have reviewed methods for computing gravitational waves from fast-motion sources. Here I shall give only a brief classification of the various methods and a few recent references where each method is used.

There are three ways to classify methods for computing wave generation: by strength of the source's internal gravity ("weak" if a Newtonian analysis gives $|\Phi| \ll 1$ everywhere; "strong" if $|\Phi| \gtrsim 1$ somewhere); by speed of the source's internal motions ("slow" if $\bar\lambda \gg L$; "fast" if $\bar\lambda \lesssim L$); and by the fractional amount that the source deviates from a nonradiating spacetime ("large deviations" or "small deviations"). Here is a list of frequently used methods of computing wave generation, classified in these three ways:

1. Slow-Motion Method

   - Arbitrary gravity, slow speed, arbitrary deviations.
   - § 3.2 above.

2. Post-Newtonian and Post-Post-Newtonian Wave-Generation Methods

   - Moderate gravity, moderate speed, arbitrary deviations.
   - Epstein and Wagoner (1975); Wagoner and Will (1976); RMP §§V.D,E.

3. Post-Linear (≡ Post-Minkowski) Wave-Generation Method

- Weak gravity, arbitrary speed, arbitrary deviations.
- Kovàcs and Thorne (1978).

4. First-Order Perturbations of Nonradiating Solutions

- Arbitrary gravity, arbitrary speed, small deviations.
- Cunningham, Price, and Moncrief (1978); Schumaker and Thorne (1983); Detweiler (1980); lecture by Nakamura in this volume; eq. (2.29) above.

5. Numerical Relativity

- Arbitrary gravity, arbitrary speed, arbitrary deviations.
- Lectures by York, Piran, Nakamura, and Isaacson in this volume.

The strongest radiators of gravity waves will be those with strong gravity, fast motions, and large deviations (e.g., collisions between two black holes); and such systems can be analyzed quantitively by only one technique: numerical relativity.

The results of wave-generation calculations for various astrophysical sources are reviewed in the lectures of Eardley in this volume.

4 THE DETECTION OF GRAVITATIONAL WAVES

Turn attention now from methods of analyzing the generation of gravitational waves to methods of analyzing their detection. If the size of the detector is small compared to a reduced wavelength, $L \ll \bar{\lambda}$, it can be analyzed in the "proper reference frame" of the detector's center of mass (§4.1). If $L \gtrsim \bar{\lambda}$, the proper reference frame is not a useful concept; and the detector is usually analyzed, instead, using "post-linear" (≡ "post-Minkowski") techniques and some carefully chosen gauge (§4.2).

4.1 Detectors with size $L \ll \bar{\lambda}$

4.1.1 The proper reference frame of an accelerated, rotating laboratory

Most gravity-wave detectors reside in earth-bound laboratories whose walls and floor accelerate relative to local inertial frames ("acceleration of gravity") and rotate relative to local gyroscopes ("Foucault pendulum effect"). In such laboratories the mathematical analog of a LIF is a "proper reference frame" (PRF). A PRF is constructed by choosing a fiducial world line which is usually attached to the detector's center of mass and thus accelerates, by next constructing spatial slices of simultaneity, $t = \text{const}$, which are orthogonal to the fiducial world line and are as flat as the spacetime curvature permits; and by then constructing in each slice of simultaneity a spatial coordinate grid which is as Cartesian as the spacetime curvature permits and is attached to the laboratory walls and thus rotates. The origin of the spatial grid is on the fiducial world line, and the time coordinate t of the slices of simultaneity is equal to proper time along the fiducial world line. Such a PRF is a mathematical realization of the type of coordinate system that a very careful experimental physicist who knows little relativity theory would likely set up in his earth-bound laboratory.

One version of such a PRF is the rotating, accelerating analog of a Fermi normal coordinate system (eq. 2.15); its spacetime metric is (Ni and Zimmermann 1978)

$$ds^2 = -dt^2\left[1 + \underbrace{2\mathbf{a}\cdot\mathbf{x} + (\mathbf{a}\cdot\mathbf{x})^2}_{\text{grav'l redshift}} - \underbrace{(\boldsymbol{\Omega}\times\mathbf{x})^2}_{\text{Lorentz time dilation}} + R_{0\ell 0m}x^\ell x^m\right]$$

$$+ 2dtdx^i\left[\underbrace{\epsilon_{ijk}\Omega^j x^k}_{\text{Sagnac effect}} - \frac{2}{3}R_{0\ell im}x^\ell x^m\right] + dx^i dx^j\left[\delta_{ij} - \frac{1}{3}R_{i\ell jm}x^\ell x^m\right]$$

$$+ O(r^3\, dx^\alpha dx^\beta). \tag{4.1}$$

Here $\mathbf{a}$ is the acceleration of the fiducial world line (minus the local "acceleration of gravity") and $\boldsymbol{\Omega}$ is the angular velocity of the spatial grid, i.e., of the laboratory walls, relative to local gyroscopes. Other versions of a PRF have spatial grids and slices of simultaneity that are bent from those of (4.1) by amounts of the order of the bending enforced by the spacetime curvature: $x^{j'} = x^j + O(r^3/\mathcal{R}^2)$, $t' = t + O(r^3/\mathcal{R}^2)$ where $\mathcal{R} \sim (R_{\alpha\beta\gamma\delta})^{-1/2}$; cf. eq. (2.16). In them g_{00} will be the same as (4.1), but g_{0i} and g_{jk} may be different by amounts of $O(r^2/\mathcal{R}^2)$.

In the PRF (4.1) a test particle acted on by an external force acquires a coordinate acceleration (obtained from the geodesic equation with a force term added)

$$\frac{d^2x^i}{dt^2} = \underbrace{-a^i}_{\substack{\text{"acceleration}\\ \text{of gravity"}}} \quad \underbrace{-2(\boldsymbol{\Omega}\times\mathbf{v})^i}_{\substack{\text{Coriolis}\\ \text{acceleration}}} \quad \underbrace{-[\boldsymbol{\Omega}\times(\boldsymbol{\Omega}\times\mathbf{x})]^i}_{\substack{\text{centrifugal}\\ \text{acceleration}}} \quad \underbrace{-(\dot{\boldsymbol{\Omega}}\times\mathbf{x})^i}_{\substack{\text{effect of}\\ \text{changing } \boldsymbol{\Omega}}} \quad \underbrace{+f^i/m}_{\substack{\text{external}\\ \text{force}}}$$

$$+\left[\begin{matrix}\text{special relativistic corrections to these inertial effects,}\\ \text{equation (20) of Ni and Zimmermann (1978)}\end{matrix}\right]$$

$$\underbrace{-R_{0i0\ell}x^\ell}_{\text{geodesic deviation}} + \underbrace{2R_{ij0\ell}v^j x^\ell}_{\substack{\text{source of "spin-}\\ \text{curvature coupling"}}} + \frac{2}{3}R_{ijk\ell}v^j v^k x^\ell \tag{4.2}$$

$$+ 2v^i R_{0j0\ell}v^j x^\ell + \frac{2}{3}v^i R_{0jk\ell}v^j v^k x^\ell .$$

Here v^j is the coordinate velocity of the particle. The terms on the first line are far bigger than those on the last three lines; they are familiar from nonrelativistic mechanics in a uniform, Newtonian gravitational field.

4.1.2 Examples of detectors

Figure 6 shows three types of gravitational-wave detectors with $L \ll \lambda\!\!\!^{-}$ that are now under construction or look favorable for future construction.

"Weber-type resonant-bar detectors" have been under development for twenty years and are discussed in detail in the lectures of Blair, Braginsky, Hamilton, Michelson, and Pallotino. In such a detector the waves couple to and drive normal modes of oscillation of a mechanical system ("antenna"), usually a solid bar made

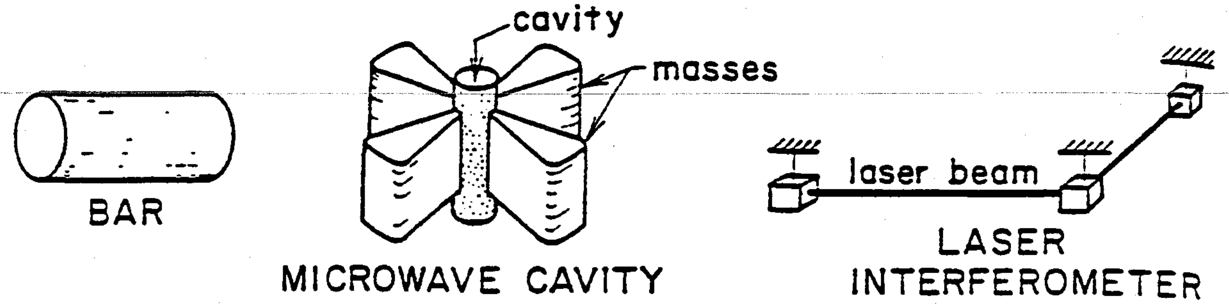


Fig. 6 Examples of gravitational-wave detectors with size L $\ll \lambdabar$.

of aluminum, and those oscillations are monitored by a transducer which is attached to the ends or sides of the bar.

In a "microwave-cavity detector" with L $\ll \lambdabar$ the gravitational waves drive oscillatory deformations of the walls of a microwave cavity; and those wall motions pump microwave quanta from one normal mode of the cavity into another. To enhance the wall deformations, big masses may be attached to the walls at strategic locations. (See, e.g., Braginsky et al. (1974), Caves (1979), Pegoraro and Radicati (1980), Grishchuk and Polnarev (1981), and references therein.) Although the design sensitivities of such detectors are comparable to those of bars, no serious efforts are now under way to develop them.

"Laser interferometer detectors" have been under development for about a decade and are discussed in detail in the lectures of Drever and Brillet. In such a detector three (or more) masses are suspended as pendula from overhead supports and swing back and forth in response to gravitational waves; and their relative motions are monitored by laser interferometry.

4.1.3 Method of analyzing detectors

For me the conceptually clearest way to analyze these three detectors, and any other with L $\ll \lambdabar$, is using the PRF of the detector's center of mass. The gravitational waves enter such an analysis entirely through the Riemann curvature terms of the metric (4.1), which have sizes

$$g^{(W)}_{\alpha\beta} \sim R^{(W)}_{\alpha\beta\gamma\delta}\, L^2 \sim h^{TT}_{jk}\, (L/\lambdabar)^2 \ll h^{TT}_{jk} \quad \text{in PRF.} \tag{4.3}$$

By contrast, in TT gauge the waves would contribute $g^{(W)}_{\alpha\beta} \sim h^{TT}_{jk}$ to the metric. An important consequence is this: In the PRF analysis the direct coupling of the gravitational waves to the detector's electromagnetic field can be ignored; and this is true whether the EM field is in a transducer on the bar, or in a microwave cavity, or in a laser beam. The direct coupling produces terms in Maxwell's equations for the vector potential with size $\delta A/A \sim g^{(W)}_{\mu\nu} \sim h^{TT}_{jk}\,(L/\lambdabar)^2$, which are smaller by $(L/\lambdabar)^2$ than the "indirect" coupling effects

$$\begin{pmatrix}\text{gravity waves deform}\\ \text{or move masses}\end{pmatrix} \rightarrow \frac{\delta L}{L} \sim h^{TT}_{jk} \rightarrow \begin{pmatrix}\text{changes of boundaries}\\ \text{for Maxwell equations}\end{pmatrix} \rightarrow \frac{\delta A}{A} \gtrsim h^{TT}_{jk}. \tag{4.4}$$

By contrast, in TT gauge the direct coupling is not negligible, and one must consider the direct interaction of the gravitational waves with both the electromagnetic field and the mechanical parts of the detector.

For all three detectors in Figure 6 and all other promising ones, the velocities of the mechanical parts of the system relative to the center of mass are $|v^j| \ll 1$. Consequently, in the mechanical equation of motion (4.2) all Riemann curvature terms except $-R_{0i0\ell}x^\ell$ can be ignored. Because $R^{(W)}_{0i0\ell} = -\frac{1}{2}\ddot{h}^{TT}_{i\ell}$ (where a dot means $\partial/\partial t$), the equation of motion for each mass element in the detector becomes

$$\ddot{x}^i = \tfrac{1}{2}\ddot{h}^{TT}_{ij}\,x^j + \begin{pmatrix}\text{all acceleration terms associated with}\\ \text{non-gravitational-wave effects}\end{pmatrix}. \qquad (4.5)$$

In summary, if one knows how to analyze the detector in the absence of gravitational waves, one can take account of the waves by simply adding the driving acceleration $\frac{1}{2}\ddot{h}^{TT}_{jk}\,x^j$ to the equation of motion of the detector's mass elements and by ignoring direct coupling of the waves to the detector's electromagnetic field.

This conclusion is valid even if the detector is large compared to inhomogeneities in the Newtonian gravity of the earth or solar system — e.g., if the detector is a normal mode of the earth itself. Then one cannot use the proper reference frame (4.1), so far as Newtonian-gravity effects are concerned; but one can still use (4.1) so far as gravitational-wave effects are concerned ($g^{(W)}_{\alpha\beta} \sim \ddot{h}^{TT}_{jk}\,L^2$). In other words, one can graft the waves onto a Newtonian analysis by means of

$$\begin{aligned} g^{(W)}_{00} &= -R^{(W)}_{0\ell 0m}\,x^\ell x^m\,, \\ |g^{(W)}_{0i}| \sim |g^{(W)}_{ij}| &\sim R^{(W)}_{\alpha\beta\gamma\delta}\,L^2 \begin{pmatrix}\text{details depend on specific variant of PRF}\\ \text{and are unimportant because } |v^j| \ll 1\end{pmatrix}, \end{aligned} \qquad (4.6)$$

and by means of the resulting equation of motion (4.5).

The gravitational-wave driving acceleration $\frac{1}{2}\ddot{h}^{TT}_{jk}\,x^j$ can be described by a quadrupole-shaped "line-of-force" diagram (Fig. 7; MTW Box 37.2).

4.1.4 Resonant-bar detectors

Consider, as an example, the analysis of a Weber-type resonant-bar gravitational-wave detector (MTW Box 37.4). One begins by computing the normal-mode eigenfrequencies ω_n and eigenfunctions $\underline{u}^{(n)}$ of the antenna ignoring the weak frictional and fluctuational coupling between modes and ignoring external forces (e.g., gravity waves). The resulting eigenfunctions, which are real (not complex) are normalized so that

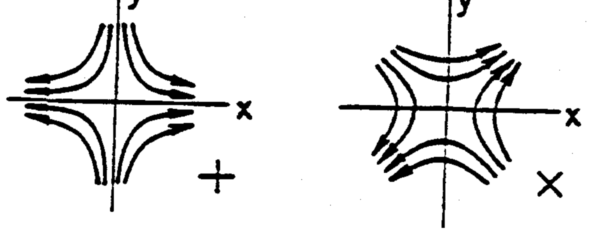


Fig. 7 "Lines of force" for the gravitational-wave acceleration (4.5) in the PRF of a detector. The two drawings correspond to waves with + polarization and with × polarization.

$$\int \rho \underset{\sim}{u}^{(n)} \cdot \underset{\sim}{u}^{(m)} \, d^3x = M\delta_{nm}; \qquad \rho = \text{density}, \quad M = \text{antenna mass}. \tag{4.7}$$

One then expands the vibrational displacement of the bar's material in terms of normal modes

$$\delta \underset{\sim}{x} = \sum_n X^{(n)} \underset{\sim}{u}^{(n)}(\underset{\sim}{x}) \, e^{-i\omega_n t}; \qquad X^{(n)} \equiv \text{"complex amplitude of mode n"}. \tag{4.8}$$

Next one writes down the equation of motion for $\delta \underset{\sim}{x}$ in the presence of gravity waves [force per unit volume equal to $\frac{1}{2}\rho \ddot{h}^{TT}_{jk} x^j$], internal friction, Nyquist forces [i.e., weak fluctuational couplings between normal modes], and coupling to the transducer. When one resolves that full equation of motion into normal modes one obtains

$$\dot{X}^{(n)} = -(2/\tau_n) X^{(n)} + \frac{ie^{i\omega_n t}}{M\omega_n} \int \underset{\sim}{f} \cdot \underset{\sim}{u}^{(n)} \, d^3x , \tag{4.9}$$

where τ_n is the (very long) frictional damping time for mode n and $\underset{\sim}{f}$ is the force per unit volume in the antenna including gravity-wave force, Nyquist forces, and "back-action forces" of the transducer on the antenna. The gravity-wave force, when integrated over the normal mode, $\underset{\sim}{u}^{(n)}$, gives

$$\int \underset{\sim}{f}^{(W)} \cdot \underset{\sim}{u}^{(n)} \, d^3x = \frac{1}{4} \ddot{h}^{TT}_{jk} \, \mathcal{I}^{(n)}_{jk}, \tag{4.10a}$$

$$\begin{aligned} \mathcal{I}^{(n)}_{jk} &= \int \rho (u^j_{(n)} x^k + u^k_{(n)} x^j - \tfrac{2}{3} \delta_{jk} \, \underset{\sim}{u}_{(n)} \cdot \underset{\sim}{x}) \, d^3x \\ &= \left[X^{(n)} e^{-i\omega_n t} \right]^{-1} \times \begin{pmatrix} \text{contribution of mode n to antenna's} \\ \text{quadrupole moment} \end{pmatrix}. \end{aligned} \tag{4.10b}$$

Thus, the same quadrupole moment as would govern emission of gravitational waves by mode n also governs their reception; this is an aspect of the principle of detailed balance (MTW §37.7).

If the antenna is hit by a broad-band burst of gravitational waves with spectral energy flux $\mathcal{F}_\nu$ (ergs cm^{-2} Hz^{-1}) and polarization e_{jk} (normalized so $e_{jk}e_{jk} = 2$)

$$h^{TT}_{jk} = A(t) e_{jk}, \qquad \mathcal{F}_\nu = \frac{\omega^2}{4} \left| \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} A(t) \, e^{i\omega t} dt \right|^2 , \qquad \omega = 2\pi\nu > 0, \tag{4.11}$$

then equations (4.9) and (4.10) give for the wave-induced change of complex amplitude $\delta X^{(n)}$

$$\begin{aligned} \frac{1}{2} M\omega_n{}^2 |\delta X^{(n)}|^2 &= \begin{pmatrix} \text{energy that would be absorbed by antenna mode n} \\ \text{if } X^{(n)} \text{ were zero when the wave burst hit} \end{pmatrix} \\ &= \mathcal{F}_\nu(\omega = \omega_n) \int \sigma_n d\nu. \end{aligned} \tag{4.12}$$

Here $\int \sigma_n d\nu$ is the antenna cross section integrated over frequency

$$\int \sigma_n d\nu = \frac{\pi}{4}\frac{\omega_n^2}{M}\left(\mathcal{I}^{(n)}_{jk}\, e_{jk}\right)^2$$

$$\sim 10^{-21}\ \mathrm{cm}^2\,\mathrm{Hz}\ \text{ for typical antennas with } M \sim 1 \text{ ton, } \omega_n/2\pi \sim 1 \text{ kHz.} \tag{4.13}$$

For further details and discussion see MTW, chapter 37; also the lectures by Blair, Braginsky, Hamilton, Michelson, and Pallotino in this volume.

* * * * *

Exercise 20. Derive equation (4.9) by resolving the equation of motion for δx into normal modes. Show that the driving term due to gravity waves has the form (4.10) and show that $\mathcal{I}^{(n)}_{jk}$ has the claimed relationship to the quadrupole moment.

Exercise 21. Show that (eq. 4.11) correctly represents the spectral energy flux of a gravity wave in the sense that $\int \mathcal{F}_\nu d\nu$ is the energy per unit area that passes the detector. Show that the broad-band burst of gravity waves (4.11) produces the change (4.12) of the antenna's complex amplitude. Show that for typical antennas the frequency-integrated cross section is $\sim 10^{-21}$ cm^2 Hz as claimed.

Exercise 22. Show that for a homogeneous, spherical antenna whose quadrupolar oscillations are being driven by gravity waves, the quadrupole moment $\mathcal{I}_{jk}$ obeys the equation of motion (2.39a).

4.2 Detectors with size $L \gtrsim ƛ$

Examples of gravitational-wave detectors with $L \gtrsim ƛ$ include the Doppler tracking of spacecraft (lectures by Hellings in this volume), and microwave-cavity detectors in which the gravity waves pump microwave quanta from one normal mode to another via direct interaction with the electromagnetic field as well as via deformation of the walls or wave guides which confine the field (Braginsky et al. (1974), Caves (1979), Pegoraro and Radicati (1980), Grishchuk and Polnarev (1981) and references therein).

Because the Riemann tensor of the waves varies significantly over the volume of such a detector, the "proper reference frame" is not a useful tool in analyzing it. Instead, analyses are based on the linearized approximation to general relativity (MTW chapter 18), with the metric $g_{\alpha\beta} = \eta_{\alpha\beta} + h_{\alpha\beta}$ including both Newtonian gravitational potentials Φ and gravitational waves; and the waves are usually treated in TT gauge:

$$g_{00} = -1 - 2\Phi, \qquad g_{0j} = 0, \qquad g_{jk} = \delta_{jk}(1-2\Phi) + h^{TT}_{jk}.$$

The analysis is actually "post-linear" (or "post-Minkowski") in that the "equations of motion" for the matter (sun, planets, detector) are not taken to be $T^{\alpha\beta}{}_{,\beta} = 0$, but rather $T^{\alpha\beta}{}_{;\beta} = 0$ with connection coefficients linear in $h_{\mu\nu}$ and Φ. See, e.g., §2 of Thorne (1977) for a careful discussion of the differences between linear theory and post-linear theory.

For an example of such an analysis see Hellings' treatment, in this volume, of the interaction of gravitational waves with NASA's doppler tracking system. As in that analysis, so in most analyses of detectors with $L \gtrsim ƛ$, direct interaction of the gravitational waves with the electromagnetic field is as important as interaction with the mechanical parts of the detector.

A mathematical trick that sometimes simplifies calculations of the interaction of gravity waves with electromagnetic fields is to write the curved-spacetime

Maxwell equations in a form identical to those for flat spacetime in a moving, anisotropic dielectric medium (e.g., Volkov, Izmest'ev, and Skrotskii 1970). Caves (unpublished) has combined this technique with gauge changes that attach the spatial coordinates onto the cavity walls even when the walls are wiggling, and has thereby produced an elegant and powerful formalism for analyzing microwave cavity detectors with $L \gtrsim \lambdabar$.

5 CONCLUSION

Most of my own research on gravitational-wave theory is motivated by the need to prepare for the day when gravitational waves are detected and astronomers confront the task of extracting astrophysical information from them. The task of preparing for that day is nontrivial. The ideas described in these lectures are a foundation for those preparations, but much further theoretical research is needed — especially the computation of gravitational wave forms $h^{TT}_{jk}(t-r; \theta, \varphi)$ from fast-motion, strong-gravity sources such as black-hole collisions.

We theorists often pay great lip service to a second motivation for our research: to give guidance to experimenters who are designing and constructing gravitational-wave detectors; and experimenters often follow with avid interest and thumping hearts the fluctuations in theoretical predictions of the waves bathing the earth. However, we theorists are far more ignorant than most experimenters imagine. For those strong sources whose wave characteristics are fairly well known (e.g., collisions between black holes), the event rate is uncertain by many orders of magnitude; and for those whose event rates are fairly well known (e.g., supernovae), the wave strengths are uncertain by many orders. Our ignorance has a simple cause: The information carried by electromagnetic waves, which is the foundation of today's theories, is nearly "orthogonal" to the information carried by gravitational waves. As a corollary, when they are ultimately detected, gravitational waves will likely give us a revolutionary new view of the universe; but until they are detected, we theorists can offer little precious advice to our experimental colleagues, and our colleagues should turn a half-deaf ear to our most confident remarks about the characteristics of the waves for which they search.

REFERENCES

Aichelburg, P. C., Ecker, G., and Sexl, R. U. 1971, Nuovo Cimento B, 2, 63.

Bontz, R. J. and Haugan, M. P. 1981, Astrophys. Space Sci., 78, 204.

Bondi, H., van der Burg, M. G. J., and Metzner, A. W. K. 1962, Proc. Roy. Soc., A269, 21.

Braginsky, V. B., Caves, C. M., and Thorne, K. S. 1977, Phys. Rev. D, 15, 2047.

Braginsky, V. B., Grishchuk, L. P., Doroshkevich, A. G., Zel'dovich, Ya. B., Novikov, I. D., and Sazhin, M. V. 1974, Sov. Phys.-JETP, 38, 865.

Burke, W. L. 1969, unpublished Ph.D. thesis, Caltech; see also Phys. Rev. A, 2, 1501 (1970).

Caves, C. M. 1979, Phys. Lett., 80B, 323.

Caves, C. M. 1980, Ann. Phys. (USA), 125, 35.

Crowley, R. J. and Thorne, K. S. 1977, Astrophys. J., 215, 624.

Cunningham, C. T., Price, R. H., and Moncrief, V. 1978, Astrophys. J., 224, 643.

Cyranski, J. F. and Lubkin, E. 1974, Ann. Phys. (USA), 87, 205.

Davis, M., Ruffini, R., Press, W. H., and Price, R. H. 1971, Phys. Rev. Lett., 27, 1466.

D'Eath, P. 1975, Phys. Rev. D, 12, 2183.

Detweiler, S. 1980, Astrophys. J., 239, 292.

DeWitt, B. S. and Brehme, R. W. 1960, Ann. Phys. (USA), 9, 220.

Dicke, R. H. 1964, in Relativity, Groups, and Topology, ed. B. and C. DeWitt (Gordon & Breach, New York).

Dyson, F. 1963, in Interstellar Communication, ed. A. G. W. Cameron (Benjamin, New York).

ELLWW: See Eardley, Lee, Lightman, Wagoner, and Will (1973).

Eardley, D. M., Lee, D. L., and Lightman, A. P. 1973, Phys. Rev. D, 8, 3308.

Eardley, D. M., Lee, D. L., Lightman, A. P., Wagoner, R. V., and Will, C. M. 1973, Phys. Rev. Lett., 30, 884; cited in text as ELLWW.

Ehlers, J., Rosenblum, A., Goldberg, J. N., and Havas, P. 1976, Astrophys. J. (Letters), 208, L77.

Einstein, A. 1918, Preuss. Akad. Wiss. Berlin, Sitzber. 1918, 154.

Epstein, R. and Wagoner, R. V. 1975, Astrophys. J., 197, 717.

Geroch, R. 1970, J. Math. Phys., 11, 2580.

Geroch, R., Held, A., and Penrose, R. 1973, J. Math. Phys., 14, 874.

Goldstein, H. 1980, *Classical Mechanics*, second edition (Addison Wesley, Reading, Mass.), §5-6.

Grishchuk, L. P. and Polnarev, A. G. 1981, chapter 10 of *General Relativity and Gravitation*, Vol. 2, ed. A. Held (Plenum Press, N.Y.).

Gürsel, Y. 1982, J. Gen. Rel. Grav., submitted.

Gürsel, Y. and Thorne, K. S. 1983, Mon. Not. Roy. Astron. Soc., submitted.

Hansen, R. O. 1974, J. Math. Phys., *15*, 46.

Hartle, J. B. 1973, Astrophys. Space Sci., *24*, 385.

Hartle, J. B. and Thorne, K. S. 1983, paper in preparation.

Hobson, E. W. 1931, *The Theory of Spherical and Ellipsoidal Harmonics* (Cambridge U. Press, Cambridge), p. 119. Republished in 1955 by Chelsea Publishing Co., New York.

Isaacson, R. A. 1968, Phys. Rev., *166*, 1263 and 1272.

Kelvin, Lord (Sir William Thomson) and Tate, P. G. 1879, *Treatise on Natural Philosophy*, Appendix B of Chapter 1 of Volume 1 (Cambridge U. Press, Cambridge).

Kovàcs, S. J. and Thorne, K. S. 1978, Astrophys. J., *224*, 62.

MTW: see Misner, Thorne, and Wheeler (1973).

Mashhoon, B. 1973, Astrophys. J. (Letters), *181*, L65.

Misner, C. W., Thorne, K. S., and Wheeler, J. A. 1973, *Gravitation* (Freeman, San Francisco); cited in text as MTW.

Ni, W.-T. and Zimmermann, M. 1978, Phys. Rev. D, *17*, 1473.

Pegoraro, F. and Radicati, L. A. 1980, J. Phys. A, *13*, 2411.

Peres, A. 1960, Nuovo Cimento, *15*, 351.

Peters, P. C. 1964, Phys. Rev., *136*, B1224.

Peters, P. C. and Mathews, J. 1963, Phys. Rev., *131*, 435.

Pirani, F. A. E. 1964, in *Lectures on General Relativity*, ed. A. Trautman, F. A. E. Pirani, and H. Bondi (Prentice Hall, Englewood Cliffs, N.J.).

Press, W. H. 1979, J. Gen. Rel. Grav., *11*, 105.

Price, R. H. 1972, Phys. Rev. D, *5*, 2419.

RMP: see Thorne (1980a).

Rosen, N. 1973, J. Gen. Rel. Grav., *4*, 435; see also N. Rosen, Ann. Phys. (USA), *84*, 455 (1974).

Sachs, R. K. 1962, Proc. Roy. Soc., *A270*, 103.

Schumaker, B. L. and Thorne, K. S. 1983, Mon. Not. Roy. Astron. Soc., in press.

Sonnabend, D. 1979, To the Solar Foci, JPL Publication 79-18 (Jet Propulsion Laboratory, Pasadena).

Szekeres, P. 1971, Ann. Phys., 64, 599.

Thorne, K. S. 1969, Astrophys. J., 158, 997.

Thorne, K. S. 1977, in Topics in Theoretical and Experimental Gravitation Physics, ed. V. De Sabbata and J. Weber (Plenum Press, London).

Thorne, K. S. 1980a, Rev. Mod. Phys., 52, 299; cited in text as RMP.

Thorne, K. S. 1981, Mon. Not. Roy. Astron. Soc., 194, 439.

Thorne, K. S. 1983, Astrophys. J., in preparation.

Thorne, K. S. and Campolattaro, A. 1967, Astrophys. J., 149, 591.

Turner, M. 1977, Astrophys. J., 216, 914.

Volkov, A. M., Izmest'ev, A. A., and Skrotskii, G. V. 1970, Zh. Eksp. Teor. Fiz., 59, 1254 [Sov. Phys.—JETP, 32, 636 (1971)].

Wagoner, R. V. and Will, C. M. 1976, Astrophys. J., 210, 764.

Walker, M. and Will, C. M. 1980, Astrophys. J. (Letters), 242, L129.

Will, C. M. 1982, Theory and Experimentation in Gravitational Physics (Cambridge University Press, New York).

Zimmermann, M. 1980, Phys. Rev. D, 21, 891.

Zimmermann, M. and Szedenits, E. 1979, Phys. Rev. D, 20, 351.

BATCH START

Lecture 2: Random Processes

STAPLE
OR
DIVIDER

LECTURE 2: RANDOM PROCESSES

Lecture by Kip S. Thorne

Assigned Reading:

D. Pages 5-1 through 5-24 of "Chapter 5. Random Processes" from the textbook manuscript *Applications of Classical Physics* by Roger Blandford and Kip Thorne.

Suggested Supplementary Reading:

a. L. A. Wainstein and V. D. Zubakov, *Extraction of Signals from Noise* (Prentice Hall, London, 1962; Dover, New York, 1970). [This wonderful book—a sort of biblical primer on the subject—is long since out of print. Kip will put his personal xerox copy on reserve in Millikan Library for a few weeks, along with the library's only copy.]

Two Suggested Problems from Blandford and Thorne's "Chapter 5, Random Processes":

5.1 *Bandwidths of a finite-Fourier-transform filter and an averaging filter* [page 5-21]

5.2 *Wiener's Optimal Filter* [page 5-22]. This is an especially important exercise, since the optimal filter underlies much of the data analysis to be done in LIGO.

BATCH
START

Lecture 3: Signal Processing

STAPLE
OR
DIVIDER

LECTURE 3: SIGNAL PROCESSING

Lecture by Eanna E. Flanagan

Assigned Reading:

A. "Gravitational Radiation" by Kip S. Thorne, in *300 Years of Gravitation*, eds. S. W. Hawking and W. Israel (Cambridge University Press, 1987), pages 366–371; 385–386; 393–395. [This is the review article handed out last Friday. The assigned sections outline the data processing methods for detecting burst, periodic and stochastic gravitational waves.]

E. "Data Processing, Analysis and Storage for Interferometric Antennas", B.F. Schutz, in "The Detection of Gravitational radiation", edited by D. Blair, (Cambridge 1989) pp. 406–416; 420–422; 428-429; 445–447. [To be handed out on Friday]

Suggested Supplementary Reading:

A. The remainder of Ref. [A] above.

F. "The Last Three Minutes: Issues in Gravitational Wave Measurements of Coalescing Compact Binaries", C. Cutler *et al*, Phys. Rev. Lett. **70**, 2984 (1993). [This is a overview of what is understood to date about the potential for extracting useful information from detected binary inspiral waveforms.]

b. E. S. Phinney, Astrophys. J. **380**, L17 (1991). [This article gives the most up-to-date estimates of the rate of binary neutron star inspirals in the Universe, and discusses in detail the astronomical observations that underlie these estimates.]

c. "Near optimal solution to the inverse problem for gravitational wave bursts", Y. Gursel and M. Tinto, Phys. Rev. D 40, 3884 (1989). [This article describes how best to reconstruct the gravitational waveforms $h_+(t)$ and $h_\times(t)$ for a detected burst of unknown form, from the (noisy) outputs of 3 interferometers.]

d. S. Smith, PhD thesis, Caltech (1987). [A description of the last real search for gravitational waves using data from the 40m prototype interferometer.]

A Few Suggested Problems

1. *The detectability of neutron star – neutron star inspirals at 1000 Mpc by LIGO.*

   a. Suppose that the burst of gravitational waves produced by a neutron star – neutron star inspiral at 1000 Mpc passes through the Earth. Calculate from the following foundations the signal-to-noise ratio obtained after optimal signal processing by one of the two LIGO interferometers: Assume that the "advanced detector" sensitivity benchmark given in Ref. [B] of lecture 1 has been achieved. Approximate this noise curve by the formula

$$S_h(f) = \begin{cases} (h_m^2/f_m)(f/f_m)^2 & f \geq f_m \\ (h_m^2/f_m)(f/f_m)^{-4} & f < f_m, \end{cases}$$

where $h_m = 1.0 \times 10^{-23}$ and $f_m = 70\text{Hz}$. Use the waveform $h(t)$ given in Eqs. (26), (42) and (104) of Ref. [A], also Eq. (29) (with RHS multiplied by a correction factor of 2), and assume that the waves' polarization and the relative orientation of the binary and of the interferometer are such that the signal-to-noise is maximized. Assume both neutron stars have masses of $1.4 M_\odot$. [*Hint:* Use the stationary phase approximation to evaluate the Fourier transform].

b. The current best estimate of the neutron star – neutron star merger rate, inferred from the statistics of observed neutron star binaries in our own galaxy, is that there should be 3 per year within a distance of 200 Mpc (uncertain to within a factor ~ 2 in the distance). Assuming this merger rate, estimate to within a factor ~ 2 the number of inspiral events per year that will produce a signal-to-noise ratio ≥ 6 in each of LIGO's two interferometers (which is roughly the criterion for successful detection), if the advanced detector sensitivity levels are achieved.

2. *The shapes of Wiener optimal filters in the time domain.*

a. For the waveform $h(t)$ and noise spectrum $S_h(f)$ of question 1, numerically calculate and plot the optimal filter $K(t)$. Compare it's shape to that of the original waveform.

b. A chance near-collision of two neutron stars in which their gravitational attraction substantially alters their velocities will produce "gravitational bremsstrahlung" or braking radiation. These waves will have a *memory*: the test masses in a nearby detector would be left with a permanent relative displacement following the waves passage, corresponding to a nonzero final value of $h(t)$. Approximate the memory part of the waveform by a step function, and numerically calculate and plot the optimal filter for the memory $K(t)$. Compare it's shape to that of the original waveform.

3. *Prospects for observing the violent final stages of black-hole – black hole mergers.* One of the aims of LIGO is to measure waves produced by the highly nonlinear dynamics in the final stages of black hole – black hole mergers. Such measurements, if they agree with supercomputer simulations, would convincingly demonstrate the existence of black holes and for the first time experimentally probe general relativity in the highly nonlinear regime. A major effort (called the Grand Challenge project) is currently underway in the numerical relativity community to calculate the waveforms; the task is expected to take several years.

a. Suppose that a pair of rapidly spinning, $15 M_\odot$ black holes coalesce at a cosmological redshift of $z\Delta\lambda/\lambda = 1$. Assume that the final plunge and coalescence of the holes (after the gradual inspiral) radiates 5% of the total mass-energy of the system into gravitational waves, and that this energy is uniformly distributed in frequency between ~ 150 Hz and ~ 350 Hz (the latter being frequency of the lowest quasinormal mode of the final $\sim 30 M_\odot$ black hole). Use Eq. (35) of Ref. [A] to estimate the signal-to-noise ratio in one of the LIGO interferometers after optimal filtering, assuming the noise spectrum of question 1. Note that the effect of redshift on energy flux is exactly the same for gravitational waves as for electromagnetic waves, and also that the energy will be distributed between ~ 75

Hz and ~ 175 Hz as measured at the detector. Assume a Hubble constant of $75\ \mathrm{km\,s^{-1}\,Mpc^{-1}}$ so that the luminosity distance is 4.7 Gpc.

b. What is the signal-to-noise squared per cycle if the waveform contains 10 cycles? What are the prospects for measuring the detailed shape of the waveform?

4. *Phase incoherence effects in searches for a gravitational wave stochastic background* The method outlined in the lecture for searching for a stochastic background by cross correlating the outputs of two interferometers assumed that the both interferometers respond to the same gravitational wave signal $h(t)$. The two LIGO detectors will be approximately parallel, but will be separated by ~ 3000 km. For what range of gravitational wave frequencies will the phase lag between the detectors be small compared to unity for all propagation directions? Will these phase lags necessitate a modification of the search algorithm outlined in the lecture?

from:

R. D. Blandford & K. S. Thorne
Applications of Classical Physics
(1993-94 version of a book we are writing · and are using in Ph136)

5 Random Processes

5.1 Introduction

In this chapter we shall analyze, among others, the following issues:

- What is the time evolution of the distribution function for an ensemble of systems that begins out of statistical equilibrium and is brought into equilibrium through contact with a heat bath?
- How can one characterize the noise that is introduced into experiments or observations by noisy devices such as resistors, amplifiers, etc.?
- What is the influence of such noise on one's ability to detect weak signals?
- What filtering strategies will improve one's ability to extract weak signals from strong noise?
- Frictional damping of a dynamical system generally arises from coupling to many other degrees of freedom (a bath) that can sap the system's energy. What is the connection, if any, between the fluctuating (noise) forces that the bath exerts on the system and its damping influence?

The mathematical foundation for analyzing such issues is the *theory of random processes*; and a portion of that subject is the *theory of stochastic differential equations*. The first two sections of this chapter constitute a quick introduction to the theory of random processes; and subsequent sections then use that theory to analyze the above issues and others. More specifically:

Section 5.2 introduces the concept of a random process and the various probability distributions that describe it, and discusses two special classes of random processes: Markoff processes and Gaussian processes. Section 5.3 introduces two powerful mathematical tools for the analysis of random processes: the correlation function and the spectral density. In Secs. 5.4 and 5.5 we meet the first application of random processes: to noise and its characterization, and to types of signal processing that can be done to extract weak signals from large noise. Finally, in Sec. 5.6 we use the theory of random processes to study the details of how an ensemble of systems, interacting with a bath, evolves into statistical equilibrium. As we shall see, the evolution is governed by a stochastic differential equation called the "Langevin equation," whose solution is described by an evolving probability distribution (the distribution function). As powerful tools in studying the probability's evolution, we develop the fluctuation-dissipation theorem (which characterizes the forces by which the bath interacts with the systems), and the Fokker-Planck equation (which describes how the probability diffuses through phase space).

5.2 Random Processes and their Probability Distributions

Definition of "random process". A (one-dimensional) *random process* is a (scalar) function $y(t)$, where t is usually time, for which the future evolution is not determined uniquely by any set of initial data—or at least by any set that is knowable to you and me. In other words, "random process" is just a fancy phrase that means "unpredictable function". Throughout this chapter we shall insist for simplicity that our random processes y take on a continuum of values ranging over some interval, often but not always $-\infty$ to $+\infty$. The generalization to y's with discrete (e.g., integral) values is straightforward.

Examples of random processes are: (*i*) the total energy $E(t)$ in a cell of gas that is in contact with a heat bath; (*ii*) the temperature $T(t)$ at the corner of Main Street and Center Street in Logan, Utah; (*iii*) the earth-longitude $\phi(t)$ of a specific oxygen molecule in the earth's atmosphere. One can also deal with random processes that are vector or tensor functions of time, but in this chapter's brief introduction we shall refrain from doing so; the generalization to "multidimensional" random processes is straightforward.

Ensembles of random processes. Since the precise time evolution of a random process is not predictable, if one wishes to make predictions one can do so only probablistically. The foundation for probablistic predictions is an *ensemble* of random processes—i.e., a collection of a huge number of random processes each of which behaves in its own, unpredictable way. In the next section we will use the ergodic hypothesis to construct, from a single random process that interests us, a conceptual ensemble whose statistical properties carry information about the time evolution of the interesting process. However, until then we will assume that someone else has given us an ensemble; and we shall develop a probablistic characterization of it.

Probability distributions. An ensemble of random processes is characterized completely by a set of probability distributions p_1, p_2, p_3, ... defined as follows:

$$p_n(y_1,t_1;y_2,t_2;\ldots;y_n,t_n)dy_1dy_2\ldots dy_n \tag{5.1}$$

tells us the probability that a process $y(t)$ drawn at random from the ensemble (*i*) will take on a value between y_1 and y_1+dy_1 at time t_1, and (*ii*) also will take on a value between y_2 and y_2+dy_2 at time t_2, and ..., and (*iii*) also will take on a value between y_n and y_n+dy_n at time t_n. (Note that the subscript n on p_n tells us how many independent values of y appear in p_n.) If we knew the values of all of an ensemble's probability distributions (an infinite number of them!) for all possible choices of their times (an infinite number of choices for each time that appears in each probability distribution) and for all possible values of y (an infinite number of possible values for each time that appears in each probability distribution), then we would have full information about the ensemble's statistical properties. Not surprisingly, it will turn out that, if the ensemble in some sense is in statistical equilibrium, we can compute all its probability distributions from a very small amount of information. But that comes later; first we must develop more formalism.

Ensemble averages. From the probability distributions we can compute ensemble averages (denoted by brackets). For example, the quantity

$$\langle y(t_1)\rangle \equiv \int y_1p_1(y_1,t_1)dy_1 \tag{5.2}$$

is the ensemble-averaged value of y at time t_1. Similarly,

$$\langle y(t_1)y(t_2)\rangle \equiv \int y_1y_2p_2(y_1,t_1;y_2,t_2)dy_1dy_2 \tag{5.3}$$

is the average value of the product $y(t_1)y(t_2)$.

Conditional probabilities. Besides the (absolute) probability distributions p_n, we shall also find useful an infinite series of *conditional* probability distributions P_1, P_2, ..., defined as follows:

$$P_n(y_1, t_1; \ \ldots ; y_{n-1}, t_{n-1} | y_n, t_n) dy_n \tag{5.4}$$

is the probability that *if* $y(t)$ took on the values y_1 at time t_1 and y_2 at time t_2 and ... and y_{n-1} at time t_{n-1}, then it will take on a value between y_n and $y_n + dy_n$ at time t_n.

It should be obvious from the definitions of the probability distributions that

$$p_n(y_1, t_1; \ldots ; y_n, t_n) = p_{n-1}(y_1, t_1; \ldots ; y_{n-1}, t_{n-1}) P_n(y_1, t_1; \ldots ; y_{n-1}, t_{n-1} | y_n, t_n) \, . \tag{5.5}$$

Using this relation, one can compute all the conditional probability distributions P_n from the absolute distributions p_1, p_2, ... Conversely, using this relation recursively, one can build up all the absolute probability distributions p_n from the first one $p_1(y_1, t_1)$ and all the conditional distributions P_2, P_3, ...

Stationary random processes. An ensemble of random processes is said to be *stationary* if and only if its probability distributions p_n depend only on time differences, not on absolute time:

$$p_n(y_1, t_1 + \tau; y_2, t_2 + \tau; \ \ldots ; y_n, t_n + \tau) = p_n(y_1, t_1; y_2, t_2; \ \ldots ; y_n, t_n) \, . \tag{5.6}$$

If this property holds for the absolute probabilities p_n, then Eq. (5.5) guarantees it also will hold for the conditional probabilities P_n.

Colloquially one says that "the random process $y(t)$ is stationary" even though what one really means is that "the ensemble from which the process $y(t)$ comes is stationary". More generally, one often speaks of "a random process $y(t)$" when what one really means is "an ensemble of random processes $\{y(t)\}$".

*Non*stationary random processes arise when one is studying a system whose evolution is influenced by some sort of clock that cares about absolute time. For example, the speeds $v(t)$ of the oxygen molecules in downtown Logan, Utah make up an ensemble of random processes regulated in part by the rotation of the earth and the orbital motion of the earth around the sun; and the influence of these clocks makes $v(t)$ be a nonstationary random process. By contrast, stationary random processes arise in the absence of any regulating clocks. An example is the speeds $v(t)$ of oxygen molecules in a room kept at constant temperature.

Stationarity does *not* mean "no time evolution of probability distributions". For example, suppose one knows that the speed of a specific oxygen molecule vanishes at time t_1, and one is interested in the probability that the molecule will have speed v_2 at time t_2. That probability, $P_2(0, t_1 | v_2, t_2)$ will be sharply peaked around $v_2 = 0$ for small time differences $t_2 - t_1$, and will be Maxwellian for large time differences $t_2 - t_1$ (Fig. 5.1). Despite this evolution, the process is stationary (assuming constant temperature) in that it does not depend on the specific time t_1 at which v happened to vanish, only on the time difference $t_2 - t_1$: $P_2(0, t_1 | v_2, t_2) = P_2(0, 0 | v_2, t_2 - t_1)$.

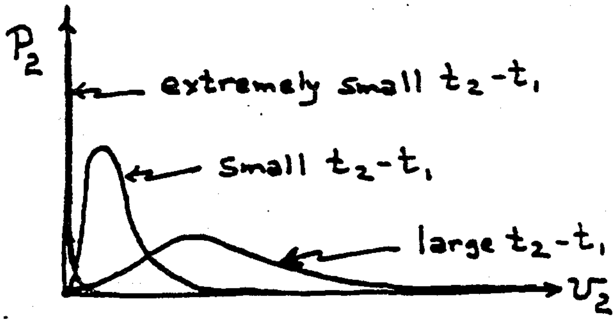


Fig. 5.1 The probability $P_2(0,t_1;v_2,t_2)$ that a molecule which has vanishing speed at time t_1 will have speed v_2 (in a unit interval dv_2) at time t_2. Although the molecular speed is a stationary random process, this probability evolves in time.

Henceforth, throughout this chapter, we shall restrict attention to random processes that are stationary (at least on the timescales of interest to us); and, accordingly, we shall denote

$$p_1(y) \equiv p_1(y,t_1) \tag{5.7}$$

since it does not depend on the time t_1. We shall also denote

$$P_2(y_1|y_2,t) \equiv P_2(y_1,0|y_2,t) \tag{5.8}$$

for the probability that, if a random process begins with the value y_1, then after the lapse of a time t it has the value y_2.

Markoff process. A random process $y(t)$ is said to be *Markoff* (also sometimes called Markovian) if and only if all of its future probabilities are determined by its most recently known value:

$$P_n(y_1,t_1;\ \ldots\ ;y_{n-1},t_{n-1}|y_n,t_n) = P_2(y_{n-1},t_{n-1}|y_n,t_n) \ \text{ for all } t_1 \le t_2 \le\ \ \ldots\ \le t_n\,. \tag{5.9}$$

This relation guarantees that any Markoff process (which, of course, we require to be stationary without saying so) is completely characterized by the probabilities

$$p_1(y) \quad \text{and } P_2(y_1|y_2,t) \equiv \frac{p_2(y_1,0;y_2,t)}{p_1(y_1)}\ ; \tag{5.10}$$

i.e., by one function of one variable and one function of three variables. From these $p_1(y)$ and $P_2(y_1|y_2,t)$ one can reconstruct, using the Markoffian relation (5.9) and the general relation (5.5) between conditional and absolute probabilities, all of the process's distribution functions.

As an example, the x-component of velocity $v_x(t)$ of a dust particle in a room filled with constant-temperature air is Markoff (if we ignore the effects of the floor, ceiling, and walls by making the room be arbitrarily large). By contrast, the position $x(t)$ of the particle is *not* Markoff because the probabilities of future values of x depend not just on the initial value of x, but also on the initial velocity v_x—or, equivalently, the probabilities depend on the values of x at *two* initial, closely spaced times. The pair $\{x(t), v_x(t)\}$ is a two-dimensional Markoff process (but we do not deal with two-dimensional processes in this chapter).

The Smoluchowski equation. Choose three (arbitrary) times t_1, t_2, and t_3 that are ordered, so $t_1 < t_2 < t_3$. Consider an arbitrary random process that begins with a known value y_1 at t_1, and ask for the probability $P_2(y_1|y_3, t_3)$ (per unit y_3) that it will be at y_3 at time t_3. Since the process must go through *some* value y_2 at the intermediate time t_2 (though we don't care what that value is), it must be possible to write the probability to reach y_3 as

$$P_2(y_1, t_1|y_3, t_3) = \int P_2(y_1, t_1|y_2, t_2) P_3(y_1, t_1; y_2, t_2|y_3, t_3) dy_2 \,,$$

where the integration is over all allowed values of y_2. This is not a terribly interesting relation, so we don't even give it an equation number. Much more interesting is its specialization to the case of a Markoff process. In that case $P_3(y_1, t_1; y_2, t_2|y_3, t_3)$ can be replaced by $P_2(y_2, t_2|y_3, t_3) = P_2(y_2, 0|y_3, t_3 - t_2) \equiv P_2(y_2|y_3, t_3 - t_2)$, and the result is an integral equation involving only P_2. Because of stationarity, it is adequate to write that equation for the case $t_1 = 0$:

$$P_2(y_1|y_3, t_3) = \int P_2(y_1|y_2, t_2) P_2(y_2|y_3, t_3 - t_2) dy_2 \,. \tag{5.11}$$

This is the *Smoluchowski equation*; it is valid for any Markoff random process and for times $0 < t_2 < t_3$. We shall discover its power in our derivation of the Fokker Planck equation in Sec. 5.6 below.

Gaussian processes. A random process is said to be Gaussian if and only if *all* of its (absolute) probability distributions are Gaussian, i.e., have the following form:

$$p_n(y_1, t_1; y_2, t_2; \;\ldots\; ; y_n, t_n) = A \exp\left[-\sum_{j=1}^{n} \sum_{k=1}^{n} \alpha_{jk}(y_j - \bar{y})(y_k - \bar{y})\right] , \tag{5.12}$$

where (*i*) A and α_{jk} depend only on the time differences $t_2 - t_1, t_3 - t_1, \ldots, t_n - t_1$; (*ii*) A is a positive normalization constant; (*iii*) $||\alpha_{jk}||$ is a *positive-definite* matrix (otherwise p_n would not be normalizable); and (*iv*) $\bar{y}$ is a constant, which one readily can show is equal to the ensemble average of y,

$$\bar{y} \equiv \langle y \rangle = \int y p_1(y) dy \,. \tag{5.13}$$

Gaussian random processes are very common in physics. For example, the total number of particles $N(t)$ in a gas cell that is in statistical equilibrium with a heat bath is a Gaussian random process [Eqs. (4.57)–(4.60) and associated discussion]. In fact, as we saw in Sec. 4.5, macroscopic variables that characterize huge systems in statistical equilibrium always have Gaussian probability distributions. The underlying reason is that,

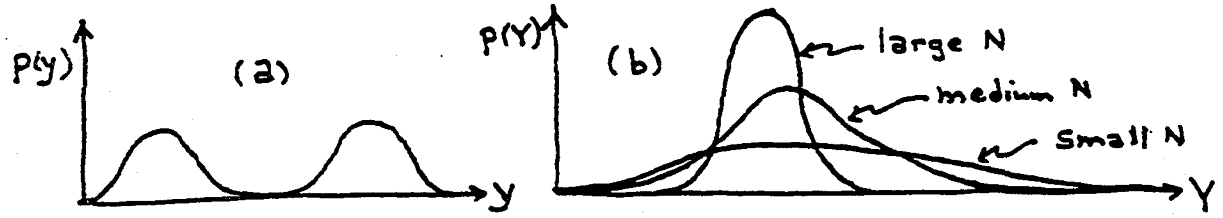


Fig. 5.2 Example of the central limit theorem. The random variable y with the probability distribution $p(y)$ shown in (a) produces, for various values of N, the variable $Y = (y_1 + \ldots + y_N)/N$ with the probability distributions $p(Y)$ shown in (b). In the limit of very large N, $p(Y)$ is a Gaussian.

when a random process is driven by a large number of statistically independent, random influences, its probability distributions become Gaussian. This general fact is a consequence of the "central limit theorem" of probability theory:

Central limit theorem. Let y be a random variable (not necessarily a random process; there need not be any times involved; however, our application is to random processes). Suppose that y is characterized by an *arbitrary* probability distribution $p(y)$ (e.g., that of Fig. 5.2), so the probability of the variable taking on a value between y and $y + dy$ is $p(y)dy$. Denote by $\bar{y}$ and σ_y the mean value of y and its *standard deviation* (the square root of its *variance*)

$$\bar{y} \equiv \langle y \rangle = \int y p(y) dy \, , \quad (\sigma_y)^2 \equiv \langle (y - \bar{y})^2 \rangle = \langle y^2 \rangle - \bar{y}^2 \, . \tag{5.14}$$

Randomly draw from this distribution a large number, N, of values $\{y_1, y_2, \cdots, y_N\}$ and average them to get a number

$$Y \equiv \frac{1}{N} \sum_{i=1}^{N} y_i \, . \tag{5.15}$$

Repeat this many times, and examine the resulting probability distribution for Y. In the limit of arbitrarily large N that distribution will be Gaussian with mean and standard deviation

$$\bar{Y} = \bar{y} \, , \quad \sigma_Y = \frac{\sigma_y}{\sqrt{N}} \, ; \tag{5.16}$$

i.e., it will have the form

$$p(Y) = \frac{1}{\sqrt{2\pi\sigma_Y{}^2}} \exp\left[-\frac{(Y - \bar{Y})^2}{2\sigma_Y{}^2} \right] \tag{5.17}$$

with $\bar{Y}$ and σ_Y given by Eq. (5.16).

The key to proving this theorem is the Fourier transform of the probability distribution. (That Fourier transform is called the distribution's *characteristic function*, but we shall not in this chapter delve into the details of characteristic functions.) Denote the Fourier transform of $p(y)$ by

$$\tilde{p}_y(f) \equiv \int_{-\infty}^{+\infty} e^{i2\pi f y} p(y) dy = \sum_{n=0}^{\infty} \frac{(i2\pi f)^n}{n!} \langle y^n \rangle \, . \tag{5.18}$$

The second expression follows from a power series expansion of the first. Similarly, since a power series expansion analogous to (5.18) must hold for $\tilde{p}_Y(k)$ and since $\langle Y^n \rangle$ can be computed from

$$\begin{aligned} \langle Y^n \rangle &= \langle N^{-n}(y_1 + y_2 + \ \ldots \ + y_N)^n \rangle \\ &= \int N^{-n}(y_1 + \ \ldots \ + y_N)^n p(y_1)...p(y_N) dy_1...dy_N \, , \end{aligned} \tag{5.19}$$

it must be that

$$\begin{aligned} \tilde{p}_Y(f) &= \sum_{n=0}^{\infty} \frac{(i2\pi f)^n}{n!} \langle Y^n \rangle \\ &= \int \exp[i2\pi f N^{-1}(y_1 + \ \ldots \ + y_N)] p(y_1) \ldots p(y_N) dy_1 \ldots dy_n \\ &= [\int e^{i2\pi f y/N} p(y) dy]^N = \left[1 + \frac{i2\pi f \bar{y}}{N} - \frac{(2\pi f)^2 \langle y^2 \rangle}{2N^2} + \mathrm{O}\left(\frac{1}{N^3}\right) \right]^N \\ &= \exp\left[i2\pi f \bar{y} - \frac{(2\pi f)^2 (\langle y^2 \rangle - \bar{y}^2)}{2N} + \mathrm{O}\left(\frac{1}{N^2}\right) \right] . \end{aligned} \tag{5.20}$$

Here the last equality can be obtained by taking the logarithm of the preceding quantity, expanding in powers of $1/N$, and then exponentiating. By inverting the Fourier transform (5.20) and using $(\sigma_y)^2 = \langle y^2 \rangle - \bar{y}^2$, we obtain for $p(Y)$ the Gaussian (5.17). Thus, the central limit theorem is proved.

5.3 Correlation Function, Spectral Density, and Ergodicity

Time averages. Forget, between here and Eq. (5.24), that we have occasionally used $\bar{y}$ to denote the numerical value of an ensemble average, $\langle y \rangle$. Instead, insist that bars denote time averages, so that if $y(t)$ is a random process and F is a function of y, then

$$\bar{F} \equiv \lim_{T \to \infty} \frac{1}{T} \int_{-T/2}^{+T/2} F[y(t)] dt \, . \tag{5.21}$$

Correlation function. Let $y(t)$ be a random process with time average $\bar{y}$. Then the correlation function of $y(t)$ is defined by

$$C_y(\tau) \equiv \overline{[y(t) - \bar{y}][y(t+\tau) - \bar{y}]} \equiv \lim_{T \to \infty} \frac{1}{T} \int_{-T/2}^{+T/2} [y(t) - \bar{y}][y(t+\tau) - \bar{y}] dt \, . \tag{5.22}$$

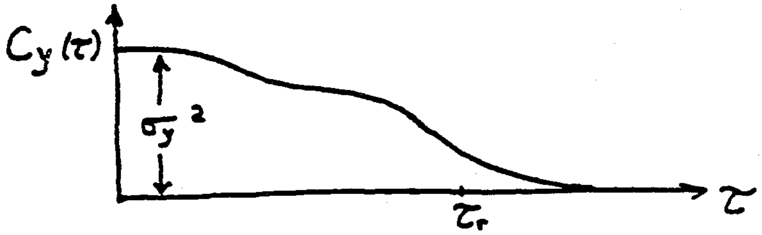


Fig. 5.3 Example of a correlation function that becomes negligible for delay times τ larger than some relaxation time τ_r.

This quantity, as its name suggests, is a measure of the extent to which the values of y at times t and $t+\tau$ tend to be correlated. The quantity τ is sometimes called the *delay time*, and by convention it is taken to be positive. [One can easily see that, if one also defines $C_y(\tau)$ for negative delay times τ by Eq. (5.22), then $C_y(-\tau) = C_y(\tau)$. Thus, nothing is lost by restricting attention to positive delay times.]

Relaxation time. Random processes encountered in physics usually have correlation functions that become negligibly small for all delay times τ that exceed some "relaxation time" τ_r; i.e., they have $C_y(\tau)$ qualitatively like that of Fig. 5.3. *Henceforth we shall restrict attention to random processes with this property.*

Ergodic hypothesis: An ensemble $\mathcal{E}$ of (stationary) random processes will be said to satisfy the ergodic hypothesis if and only if it has the following property: Let $y(t)$ be any random process in the ensemble $\mathcal{E}$. Construct from $y(t)$ a new ensemble $\mathcal{E}'$ whose members are

$$Y^K(t) \equiv y(t+KT) , \tag{5.23}$$

where K runs over all integers, negative and positive, and where T is a time interval large compared to the process's relaxation time, $T \gg \tau_r$. Then $\mathcal{E}'$ has the same probability distributions p_n as $\mathcal{E}$—i.e., $p_n(Y_1, t_1; \;\dots\; ; Y_n, t_n)$ has the same functional form as $p_n(y_1, t_1; \;\dots\; ; y_n, t_n)$—for all times such that $|t_i - t_j| < T$. This is essentially the same ergodic hypothesis as we met in Sec. 3.5.

As in Sec. 3.5, the ergodic hypothesis guarantees that time averages defined using any random process $y(t)$ drawn from the ensemble $\mathcal{E}$ are equal to ensemble averages:

$$\bar{F} \equiv \langle F \rangle , \tag{5.24}$$

where F is any function of y: $F = F(y)$. In this sense, each random process in the ensemble is representative, when viewed over sufficiently long times, of the statistical properties of the entire ensemble—and conversely.

Henceforth we shall restrict attention to ensembles that satisfy the ergodic hypothesis. This, in principle, is a severe restriction. In practice, for a physicist, it is not severe at all. In physics one's objective when introducing ensembles is usually to acquire computational techniques for dealing with a single, or a small number of random processes; and one acquires those techniques by defining one's conceptual ensembles in such a way that they satisfy the ergodic hypothesis.

Because we insist that the ergodic hypothesis be satisfied for all our random processes, the value of the correlation function at zero time delay will be

$$C_y(0) \equiv \overline{(y-\bar{y})^2} = \langle (y-\bar{y})^2 \rangle \,,$$

which by definition is the variance ${\sigma_y}^2$ of y:

$$C_y(0) = {\sigma_y}^2 \,. \tag{5.25}$$

We now turn to some issues which will prepare us for defining the concept of "spectral density".

Fourier transforms. There are several different sets of conventions for the definition of Fourier transforms. In this book we adopt a set which is commonly (but not always) used in the theory of random processes, but which differs from that common in quantum theory. Instead of using the angular frequency ω, we shall use the ordinary frequency $f \equiv \omega/2\pi$; and we shall define the Fourier transform of a function $y(t)$ by

$$\tilde{y}(f) \equiv \int_{-\infty}^{+\infty} y(t) e^{i2\pi f t} dt \,. \tag{5.26}$$

Knowing the Fourier transform $\tilde{y}(f)$, we can invert (5.26) to get $y(t)$ using

$$y(t) \equiv \int_{-\infty}^{+\infty} \tilde{y}(f) e^{-i2\pi f t} df \,. \tag{5.27}$$

Notice that with this set of conventions there are no factors of $1/2\pi$ or $1/\sqrt{2\pi}$ multiplying the integrals. Those factors have been absorbed into the df of (5.27), since $df = d\omega/2\pi$.

Fourier transforms are not useful when dealing with random processes. The reason is that a random process $y(t)$ is generally presumed to go on and on and on forever; and, as a result, its Fourier transform $\tilde{y}(f)$ is divergent. One gets around this problem by crude trickery: (i) From $y(t)$ construct, by truncation, the function

$$y_T(t) \equiv y(t) \;\text{ if }\; -T/2 < t < +T/2 \,, \quad \text{and} \quad y_T(t) \equiv 0 \;\text{ otherwise} \,. \tag{5.28}$$

Then the Fourier transform $\tilde{y}_T(f)$ is finite; and by Parseval's theorem it satisfies

$$\int_{-T/2}^{+T/2} [y(t)]^2 dt = \int_{-\infty}^{+\infty} |\tilde{y}_T(f)|^2 df = 2\int_0^{\infty} |\tilde{y}_T(f)|^2 df \,. \tag{5.29}$$

Here in the last equality we have used the fact that because $y_T(t)$ is real, $\tilde{y}_T^*(f) = \tilde{y}_T(-f)$ where * denotes complex conjugation; and, consequently, the integral from $-\infty$ to 0 of $|\tilde{y}_T(f)|^2$ is the same as the integral from 0 to $+\infty$. Now, the quantities on the two sides of (5.29) diverge in the limit as $T \to \infty$, and it is obvious from the left side that they diverge linearly as T. Correspondingly, the limit

$$\lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{+T/2} [y(t)]^2 dt = \lim_{T\to\infty} \frac{2}{T} \int_0^{\infty} |\tilde{y}_T(f)|^2 df \tag{5.30}$$

is convergent.

Spectral density. These considerations motivate the following definition of the spectral density $G_y(f)$ of the random process $y(t)$:

$$G_y(f) \equiv \lim_{T\to\infty} \frac{2}{T} \left| \int_{-T/2}^{+T/2} [y(t) - \bar{y}] e^{i2\pi f t} dt \right|^2 . \tag{5.31}$$

Notice that the quantity inside the absolute value sign is just $\tilde{y}_T(f)$, but with the mean of y removed before computation of the Fourier transform. (The mean is removed so as to avoid an uninteresting delta function in $G_y(f)$ at zero frequency.) Correspondingly, by virtue of our motivating result (5.30), the spectral density satisfies

$$\int_0^\infty G_y(f)df = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{+T/2} [y(t) - \bar{y}]^2 dt = \overline{(y - \bar{y})^2} = \sigma_y{}^2 . \tag{5.32}$$

In words: The integral of the spectral density of y over all positive frequencies is equal to the variance of y.

By convention, the spectral density is defined only for nonnegative frequencies f. This is because, were we to define it also for negative frequencies, the fact that $y(t)$ is real would imply that $G_y(f) = G_y(-f)$, so the negative frequencies contain no new information. Our insistence that f be positive goes hand in hand with the factor 2 in the $2/T$ of the definition (5.31): that factor 2 in essence folds the negative frequency part over onto the positive frequency part.

Notice that the spectral density has units of y^2 per unit frequency; or, more colloquially (since frequency f is usually measured in Hertz, i.e., cycles per second) its units are y^2/Hz.

The Wiener-Khintchine Theorem says that *for any random process $y(t)$ the correlation function $C_y(\tau)$ and the spectral density $G_y(f)$ are the cosine transforms of each other and thus contain precisely the same information:*

$$C_y(\tau) = \int_0^\infty G_y(f) \cos(2\pi f \tau) df , \tag{5.33}$$

$$G_y(f) = 4 \int_0^\infty C_y(\tau) \cos(2\pi f \tau) d\tau . \tag{5.34}$$

This theorem is readily proved as a consequence of Parseval's theorem: Assume, from the outset, that the mean has been subtracted from $y(t)$ so $\bar{y} = 0$. [This is not really a restriction on the proof, since both the C_y of Eq. (5.22) and the G_y of Eq. (5.31) are insensitive to the mean of y.] Denote by $y_T(t)$ the truncated y of (5.28) and by $\tilde{y}_T(f)$ its Fourier transform. Then the generalization of Parseval's theorem [1]

$$\int_{-\infty}^{+\infty} (gh^* + hg^*)dt = \int_{-\infty}^{+\infty} (\tilde{g}\tilde{h}^* + \tilde{h}\tilde{g}^*)df , \tag{5.35}$$

[1] This follows by subtracting Parseval's theorem for g and for h from Parseval's theorem for $g + h$.

with $g = y_T(t)$, $\tilde{g} = \tilde{y}_T(f)$, $h = y_T(t+\tau)$, and $\tilde{h} = \tilde{y}_T(f)e^{i2\pi f\tau}$, says

$$\begin{aligned} 2\int_{-\infty}^{+\infty} y_T(t)y_T(t+\tau)dt &= \int_{-\infty}^{+\infty} |\tilde{y}_T(f)|^2(e^{-i2\pi f\tau} + e^{i2\pi f\tau})df \\ &= 2\int_{-\infty}^{+\infty} |\tilde{y}_T(f)|^2 \cos(2\pi f\tau)df \, . \end{aligned} \tag{5.36}$$

By dividing by T, taking the limit as $T \to \infty$, converting the second integral into the range $0 \to \infty$, and using Eqs. (5.22) and (5.31), we bring this into the form (5.35). The inverse relation (5.36) follows directly by standard inversion of a cosine transform. *QED*

Doob's Theorem. A large fraction of the random processes that one meets in physics are Gaussian, and many of them are Markoff. As a result, the following remarkable theorem about processes that are both Gaussian and Markoff is quite important: *Any random process $y(t)$ that is both Gaussian and Markoff has the following forms for its correlation function, its spectral density, and the two probability distributions p_1 and P_2 which determine all the others:*

$$C_y(\tau) = {\sigma_y}^2 e^{-\tau/\tau_r} \, , \tag{5.37}$$

$$G_y(f) = \frac{(4/\tau_r){\sigma_y}^2}{(2\pi f)^2 + (1/\tau_r)^2} \, , \tag{5.38}$$

$$p_1(y) = \frac{1}{\sqrt{2\pi{\sigma_y}^2}} \exp\left[-\frac{(y-\bar{y})^2}{2{\sigma_y}^2}\right] \, , \tag{5.39}$$

$$P_2(y_1|y_2,\tau) = \frac{1}{[2\pi(1-e^{-2\tau/\tau_r}){\sigma_y}^2]^{\frac{1}{2}}} \exp\left[-\frac{[y_2 - \bar{y} - e^{-\tau/\tau_r}(y_1-\bar{y})]^2}{2(1-e^{-2\tau/\tau_r}){\sigma_y}^2}\right] \, . \tag{5.40}$$

Here $\bar{y}$ is the process's mean, σ_y is its standard deviation (${\sigma_y}^2$ is its variance), and τ_r is its relaxation time. This result is *Doob's theorem.*[2]

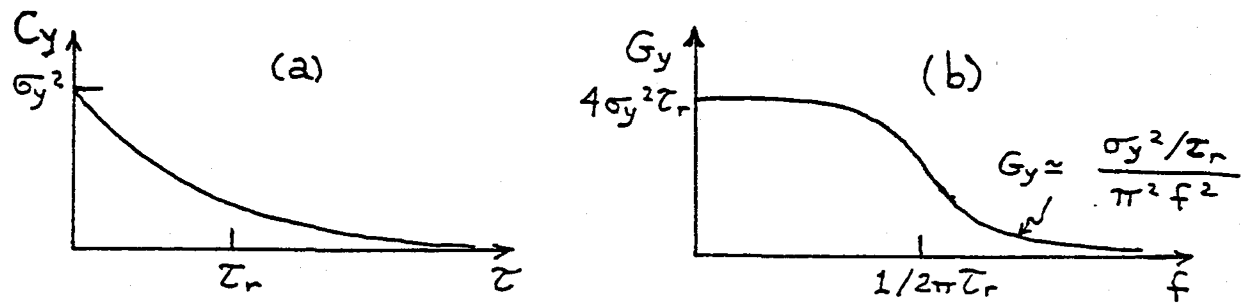


Fig. 5.4 (a) the correlation function (5.37) and spectral density (5.38) for a Gaussian, Markoff process.

[2] It is so named because it was first identified and proved by J. L. Doob (1942).

The correlation function (5.37) and spectral density (5.38) are plotted in Fig. 5.4.

Note the great power of Doob's theorem: Because all of y's probability distributions are computable from p_1 [Eq. (5.39)] and P_2 [Eq. (5.40)], and these are determined by $\bar{y}$, σ_y, and τ_r, this theorem says that *all statistical properties of a Gaussian, Markoff process are determined by just three parameters: its mean* $\bar{y}$*, its variance* ${\sigma_y}^2$*, and its relaxation time* τ_r.

Proof of Doob's Theorem: Let $y(t)$ be Gaussian and Markoff (and, of course, stationary). For ease of notation, set $y_{\rm new} = (y_{\rm old} - \bar{y}_{\rm old})/\sigma_{y_{\rm old}}$, so $\bar{y}_{\rm new} = 0$, $\sigma_{y_{\rm new}} = 1$. If the theorem is true for $y_{\rm new}$, then by the rescalings inherent in the definitions of $C_y(\tau)$, $G_y(f)$, $p_1(y)$, and $P_2(y_1|y_2,\tau)$, it will also be true for $y_{\rm old}$.

Since $y \equiv y_{\rm new}$ is Gaussian, its first two probability distributions must have the following Gaussian forms (these are the most general Gaussians with the required mean $\bar{y} = 0$ and variance ${\sigma_y}^2 = 1$):

$$p_1(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \tag{5.41}$$

$$p_2(y_1,t_1;y_2,t_2) = \frac{1}{\sqrt{(2\pi)^2(1-{C_{21}}^2)}} \exp\left[-\frac{{y_1}^2+{y_2}^2-2C_{21}y_1y_2}{2(1-{C_{21}}^2)}\right] . \tag{5.42}$$

By virtue of the ergodic hypothesis, this p_2 determines the correlation function:

$$C_y(t_2-t_1) \equiv \langle y(t_2)y(t_1)\rangle = \int p_2(y_1,t_1;y_2,t_2)y_1y_2dy_1dy_2 = C_{21} . \tag{5.43}$$

Thus, the constant C_{21} in p_2 is the correlation function. From the general expression (5.5) for conditional probabilities in terms of absolute probabilities we can compute P_2:

$$P_2(y_1,t_1|y_2,t_2) = \frac{1}{\sqrt{2\pi(1-{C_{21}}^2)}} \exp\left[-\frac{(y_2-C_{21}y_1)^2}{2(1-{C_{21}}^2)}\right] . \tag{5.44}$$

We can also use the general expression (5.5) for the relationship between conditional and absolute probabilities to compute p_3:

$$\begin{aligned} p_3(y_1,t_1;y_2,t_2;y_3,t_3) &= p_2(y_1,t_1;y_2,t_2)P_3(y_1,t_1;y_2,t_2|y_3,t_3) \\ &= p_2(y_1,t_1;y_2,t_2)P_2(y_2,t_2|y_3,t_3) \\ &= \frac{1}{\sqrt{(2\pi)^2(1-{C_{21}}^2)}} \exp\left[-\frac{({y_1}^2+{y_2}^2-2C_{21}y_1y_2)}{2(1-{C_{21}}^2)}\right] \\ &\times \frac{1}{\sqrt{2\pi(1-{C_{32}}^2)}} \exp\left[-\frac{(y_3-C_{32}y_2)^2}{2(1-{C_{32}}^2)}\right] . \end{aligned} \tag{5.45}$$

Here the second equality follows from the fact that y is Markoff, and in order that it be valid we insist that $t_1 < t_2 < t_3$. From the explicit form (5.45) of p_3 we can compute

$$C_y(t_3-t_1) \equiv C_{31} \equiv \langle y(t_3)y(t_1)\rangle = \int p_3(y_1,t_1;y_2,t_2;y_3,t_3)y_1y_3dy_1dy_2dy_3 . \tag{5.46}$$

The result is

$$C_{31} = C_{32} C_{21} \,. \tag{5.47}$$

In other words,

$$C_y(t_3 - t_1) = C_y(t_3 - t_2) C_y(t_2 - t_1) \quad \text{for any } t_3 > t_2 > t_1 \,. \tag{5.48}$$

The *unique* solution to this equation, with the "initial condition" that $C_y(0) = \sigma_y{}^2 = 1$, is

$$C_y(\tau) = e^{-\tau/\tau_r} \,, \tag{5.49}$$

where τ_r is a constant (which we identify as the relaxation time; cf. Fig. 5.3). From the Wiener-Khintchine relation (5.34) and this correlation function we obtain

$$G_y(f) = \frac{4/\tau_r}{(2\pi f)^2 + (1/\tau_r)^2} \,. \tag{5.50}$$

Equations (5.50), (5.49), (5.41), and (5.44) are the asserted forms (5.37)–(5.40) of the correlation function, spectral density, and probability distributions in the case of our $y_{\rm new}$ with $\bar{y} = 0$ and $\sigma_y = 1$. From these, by rescaling, we obtain the forms (5.37)–(5.40) for $y_{\rm old}$. Thus, Doob's theorem is proved. *QED*

5.4 Noise and its Types of Spectra

Experimental physicists and engineers encounter random processes in the form of "noise" that is superposed on signals they are trying to measure. *Examples*: (*i*) In radio communication, "static" on the radio is noise. (*ii*) When modulated laser light is used for optical communication, random fluctuations in the arrival times of photons always contaminate the signal; the effects of such fluctuations are called "shot noise" and will be studied below. (*iii*) Even the best of atomic clocks fail to tick with absolutely constant angular frequencies ω; their frequencies fluctuate ever so slightly relative to an ideal clock, and those fluctuations can be regarded as noise.

Sometimes the "signal" that one studies amidst noise is actually itself some very special noise ("one person's signal is another person's noise"). An example is in radio astronomy, where the electric field $E_x(t)$ of the waves from a quasar, in the x-polarization state, is a random process whose spectrum (spectral density) the astronomer attempts to measure. Notice from its definition that the spectral density, $G_{E_x}(f)$ is nothing but the specific intensity, I_ν [Eq. (2.17)], integrated over the solid angle subtended by the source:

$$G_{E_x}(f) = \frac{4\pi}{c} \frac{d\,\text{Energy}}{d\,\text{Area}\, d\,\text{time}\, df} = \frac{4\pi}{c} \int I_\nu d\Omega \,. \tag{5.51}$$

(Here ν and f are just two alternative notations for the same frequency.) It is precisely this $G_{E_x}(f)$ that radio astronomers seek to measure; and they must do so in the presence of noise due to other, nearby radio sources, noise in their radio receivers, and "noise" produced by commercial radio stations.

As an aid to understanding various types of noise, we shall seek an intuitive understanding of the meaning of the spectral density $G_y(f)$: Suppose that we examine the time evolution of a random process $y(t)$ over a specific interval of time Δt. That time evolution will involve fluctuations at various frequencies from $f = \infty$ on down to the lowest frequency for which we can fit at least one period into the time interval studied, i.e., down to $f = 1/\Delta t$. Choose a frequency f in this range, and ask what are the mean square fluctuations in y at that frequency. By definition, they will be

$$[\Delta y(\Delta t, f)]^2 \equiv \lim_{N\to\infty} \frac{2}{N} \sum_{n=-N/2}^{n=+N/2} \left| \frac{1}{\Delta t} \int_{n\Delta t}^{(n+1)\Delta t} y(t) e^{i2\pi f t} dt \right|^2 . \tag{5.52}$$

Here the factor 2 in $2/N$ accounts for our insistence on folding negative frequencies f into positive, and thereby regarding f as nonnegative; i.e., the quantity (5.52) is the mean square fluctuation at frequency $-f$ plus that at $+f$. The phases of the finite Fourier transforms appearing in (5.52) (one transform for each interval of time Δt) will be randomly distributed with respect to each other. As a result, if we add these Fourier transforms and then compute their absolute square rather than computing their absolute squares first and then adding, the new terms we introduce will have random relative phases that cause them to cancel each other. In other words, with vanishing error in the limit $N \to \infty$, we can rewrite (5.52) as

$$[\Delta y(\Delta t, f)]^2 = \lim_{N\to\infty} \frac{2}{N} \left| \sum_{n=-N/2}^{n=+N/2} \frac{1}{\Delta t} \int_{n\Delta t}^{(n+1)\Delta t} y(t) e^{i2\pi f t} dt \right|^2 . \tag{5.53}$$

By defining $T \equiv N\Delta t$ and noting that a constant in $y(t)$ contributes nothing to the Fourier transform at finite (nonzero) frequency f, we can rewrite this expression as

$$[\Delta y(\Delta t, f)]^2 = \lim_{T\to\infty} \frac{2}{T} \left| \int_{-T/2}^{+T/2} (y - \bar{y}) e^{i2\pi f t} dt \right|^2 \frac{1}{\Delta t} = G_y(f) \frac{1}{\Delta t} . \tag{5.54}$$

It is conventional to call the reciprocal of the time Δt on which these fluctuations are studied the *bandwidth* Δf of the study; i.e.,

$$\Delta f \equiv 1/\Delta t \, , \tag{5.55}$$

and correspondingly it is conventional to interpret (5.53) as saying that *the root-mean-square (rms) fluctuations at frequency f and during the time $\Delta t \ge f^{-1}$ are*

$$\Delta y(\Delta t = 1/\Delta f, f) = \sqrt{G_y(f)\Delta f} \, . \tag{5.56}$$

Special noise spectra. Certain spectra have been given special names:

$G_y(f)$ independent of f — white noise spectrum,
$G_y(f) \propto 1/f$ — flicker noise spectrum,
$G_y(f) \propto 1/f^2$ — random walk spectrum.

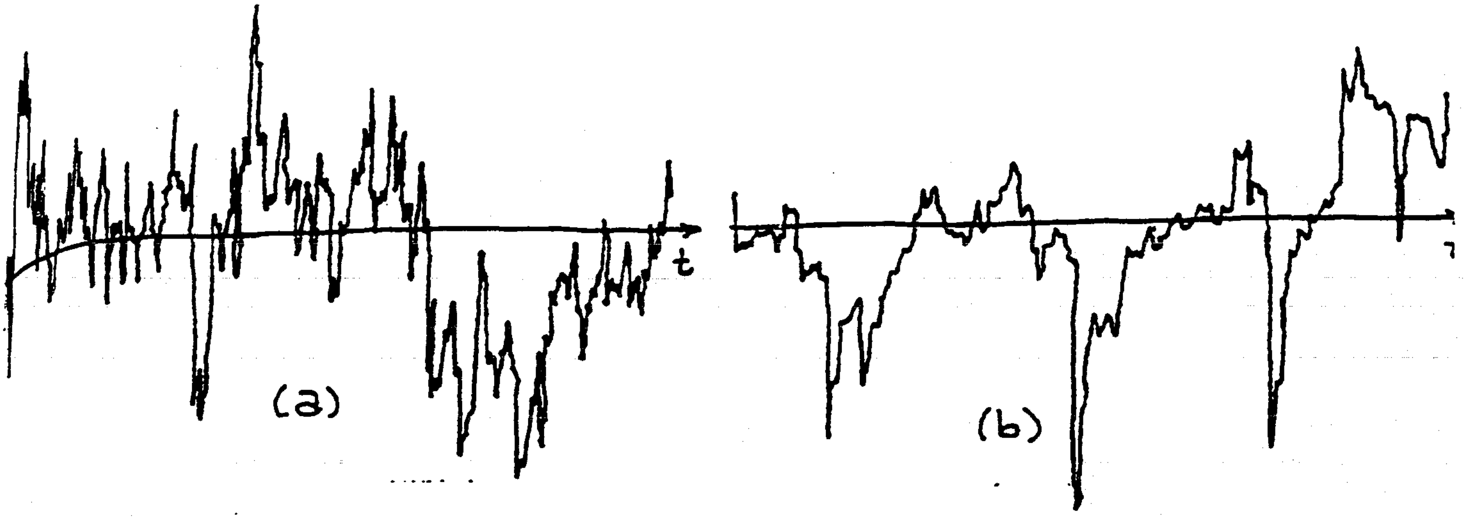


Fig. 5.5 Examples of two random processes that have flicker noise spectra, $G_y(f) \propto 1/f$. [From Press (1978).]

White noise is called "white" because it has equal amounts of "power per unit frequency" G_y at all frequencies, just as white light has roughly equal powers at all light frequencies. Put differently, if $y(t)$ has a white-noise spectrum, then its rms fluctuations over a fixed time interval Δt (i.e., in a fixed bandwidth Δf) are independent of frequency f; i.e., $\Delta y(\Delta t, f) = \sqrt{G_y/\Delta t}$ is independent of f since G_y is independent of f.

Flicker noise gets its name from the fact that, when one looks at the time evolution $y(t)$ of a random process with a flicker-noise spectrum, one sees fluctuations ("flickering") on all timescales, and the rms amplitude of flickering is independent of the timescale one chooses. Stated more precisely, choose any timescale Δt and then choose a frequency $f \sim 3/\Delta t$ so one can fit roughly three periods of oscillation into the chosen timescale. Then the rms amplitude of the fluctuations one observes will be

$$\Delta y(\Delta t, f = 3/\Delta t) = \sqrt{G_y(f) f/3}\,, \tag{5.57}$$

which is a constant independent of f and Δt when the spectrum is that of flicker noise, $G_y \propto 1/f$. Stated differently, flicker noise has the same amount of power in each octave of frequency. Figure 5.5 is an illustration: Both graphs shown there depict random processes with flicker-noise spectra. (The differences between the two graphs will be explained below.) No matter what time interval one chooses, these processes look roughly periodic with one or two or three oscillations in that time interval; and the amplitudes of those oscillations are independent of the chosen time interval.

Random-walk spectra arise when the random process $y(t)$ undergoes a random walk. We shall study an example in Sec. 5.6 below.

Notice that for a Gaussian Markoff process the spectrum (Fig. 5.4) is white at frequencies $f \ll 1/(2\pi\tau_r)$ where τ_r is the relaxation time, and it is random-walk at frequencies $f \gg 1/(2\pi\tau_r)$. This is typical: random processes encountered in the real world tend to

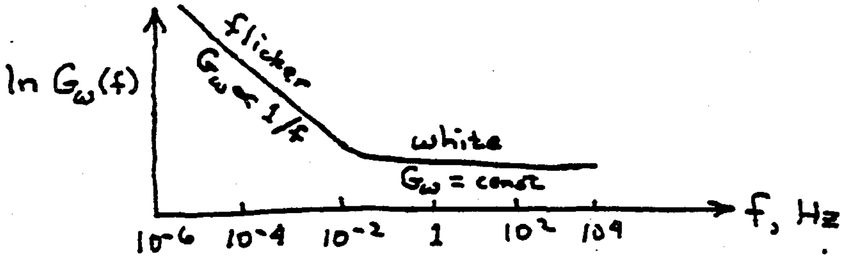


Fig. 5.6 The spectral density of the fluctuations in angular frequency ω of ticking of a Rubidium atomic clock.

have one type of spectrum over one large interval of frequency, then switch to another type over another large interval. The angular frequency of ticking of a Rubidium atomic clock furnishes another example. That angular frequency fluctuates slightly with time, $\omega = \omega(t)$; and those fluctuations have the form shown in Fig. 5.6. At low frequencies $f \lesssim 10^{-2}$ Hz, i.e., over long timescales $\Delta t \gtrsim 100$ sec, ω exhibits flicker noise; and at higher frequencies, i.e., over timescales $\Delta t \lesssim 100$ sec, it exhibits white noise.

In experimental studies of noise, attention focuses very heavily on the spectral density $G_y(f)$ and on quantities that one can compute from it. In the special case of a Gaussian-Markoff process, the spectrum $G_y(f)$ and the mean $\bar{y}$ together contain full information about all statistical properties of the random process. However, most random processes that one encounters are not Markoff (though most *are* Gaussian). (Whenever the spectrum deviates from the special form in Fig. 5.4, one can be sure the process is not Gaussian-Markoff.) Correspondingly, for most processes the spectrum contains only a tiny part of the statistical information required to characterize the process. The two random processes shown in Fig. 5.5 above are a good example. They were constructed on a computer as superpositions of pulses $F(t - t_o)$ with random arrival times t_o and with identical forms

$$F(t) = 0 \quad \text{for } t < 0 \,, \quad F(t) = K/\sqrt{t} \quad \text{for } t > 0 \,. \tag{5.58}$$

The two $y(t)$'s look very different because the first [Fig. 5.5(a)] involves frequent small pulses, while the second [Fig. 5.5(b)] involves less frequent, larger pulses. These differences are obvious to the eye in the time evolutions $y(t)$. However, they do not show up at all in the spectra $G_y(f)$: the spectra are identical; both are of flicker type. Moreover, the differences do not show up in $p_1(y_1)$ or in $p_2(y_1, t_1; y_2 t_2)$ because the two processes are both superpositions of many independent pulses and thus are Gaussian; and for Gaussian processes p_1 and p_2 are determined fully by the mean and the correlation function, or equivalently by the mean and spectral density, which are the same for the two processes. Thus, the differences between the two processes show up only in the probabilities p_n of third order and higher, $n \geq 3$.

5.5 Filters, Signal-to-Noise Ratio, and Shot Noise

Filters. In experimental physics and engineering one often takes a signal $y(t)$ or a random process $y(t)$ and filters it to produce a new function $w(t)$ that is a *linear functional* of $y(t)$:

$$w(t) = \int_{-\infty}^{+\infty} K(t-t')y(t')dt' . \tag{5.59}$$

The quantity $y(t)$ is called the filter's *input*; $K(t-t')$ is the filter's *kernel*, and $w(t)$ is its *output*. We presume throughout this chapter that the kernel depends only on the time difference $t-t'$ and not on absolute time. One says that the filter is *stationary* when this is so; and when it is violated so $K = K(t,t')$ depends on absolute time, the filter is said to be nonstationary. Our restriction to stationary filters goes hand-in-hand with our restriction to stationary random processes, since if $y(t)$ is stationary as we require, and if the filter is stationary as we require, then the filtered process $w(t) = \int_{-\infty}^{+\infty} K(t-t')y(t')dt'$ is stationary.

Some examples of kernels and their filtered outputs are these:

$$\begin{aligned} &K(\tau) = \delta(\tau) \ : \quad w(t) = y(t) \, , \\ &K(\tau) = \delta'(\tau) \ : \quad w(t) = dy/dt \, , \\ &K(\tau) = 0 \text{ for } \tau < 0 \text{ and } 1 \text{ for } \tau > 0 \ : \quad w(t) = \textstyle\int_{-\infty}^{t} y(t')dt' \, . \end{aligned} \tag{5.60}$$

As with any function, a knowledge of the kernel $K(\tau)$ is equivalent to a knowledge of its Fourier transform

$$\tilde{K}(f) \equiv \int_{-\infty}^{+\infty} K(\tau)e^{i2\pi f\tau}d\tau . \tag{5.61}$$

This Fourier transform plays a central role in the theory of filtering (also called the theory of *linear signal processing*): The convolution theorem of Fourier transform theory says that, if $y(t)$ is a function whose Fourier transform $\tilde{y}(f)$ exists (converges), then the Fourier transform of the filter's output $w(t)$ [Eq. (5.59)] is given by

$$\tilde{w}(f) = \tilde{K}(f)\tilde{y}(f) . \tag{5.62}$$

Similarly, by virtue of the definition (5.31) of spectral density in terms of Fourier transforms, if $y(t)$ is a random process with spectral density $G_y(f)$, then the filter's output $w(t)$ will be a random process with spectral density

$$G_w(f) = |\tilde{K}(f)|^2 G_y(f) . \tag{5.63}$$

[Note that, although $\tilde{K}(f)$, like all Fourier transforms, is defined for both positive and negative frequencies, when its modulus is used in (5.63) to compute the the effect of the filter on a spectral density, only positive frequencies are relevant; spectral densities are strictly positive-frequency quantitities.]

The quantity $|\tilde{K}(f)|^2$ that appears in the very important relation (5.63) is most easily computed not by evaluating directly the Fourier transform (5.61) and then squaring, but rather by sending the function $e^{i2\pi ft}$ through the filter and then squaring. To see that this works, notice that the result of sending $e^{i2\pi ft}$ through the filter is

$$\int_{-\infty}^{+\infty} K(t-t')e^{i2\pi ft'}dt' = \tilde{K}^*(f)e^{i2\pi ft} , \tag{5.64}$$

which differs from $\tilde{K}(f)$ by complex conjugation and a change of phase, and which thus has absolute value squared equal to $|\tilde{K}(f)|^2$. For example, If $w(t) = d^n y/dt^n$, then when we send $e^{i2\pi ft}$ through the filter we get $(i2\pi f)^n e^{i2\pi ft}$; and, accordingly, $|\tilde{K}(f)|^2 = (2\pi f)^{2n}$, and $G_w(f) = (2\pi f)^{2n} G_y(f)$.

This last example shows that by differentiating a random process once, one changes its spectral density by a multiplicative factor f^2; for example, one can thereby convert random-walk noise into white noise. Similarly, by integrating a random process once in time (the inverse of differentiating), one multiplies its spectral density by f^{-2}. If one wants, instead, to multiply by f^{-1}, one can achieve that using the filter

$$K(\tau) = 0 \text{ for } \tau < 0\,, \quad K(\tau) = \frac{1}{\sqrt{\tau}} \text{ for } \tau > 0\,; \tag{5.65}$$

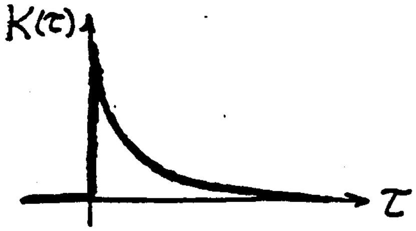


Fig. 5.7 The kernel (5.65) whose filter multiplies the spectral density by a factor $1/f$, thereby converting white noise into flicker noise, and flicker noise into random-walk noise.

see Fig. 5.7. Specifically, it is easy to show, by sending a sinusoid through this filter, that

$$w(t) \equiv \int_{-\infty}^{t} \frac{1}{\sqrt{t-t'}} y(t') dt' \tag{5.66}$$

has

$$G_w(f) = \frac{1}{f} G_y(f)\,. \tag{5.67}$$

Thus, by filtering in this way one can convert white noise into flicker noise, and flicker noise into random-walk noise.

Band-pass filter. In experimental physics and engineering one often meets a random process $Y(t)$ that consists of a sinusoidal signal on which is superposed noise $y(t)$

$$Y(t) = \sqrt{2} Y_s \cos(2\pi f_o t + \delta_o) + y(t)\,. \tag{5.68}$$

We shall assume that the frequency f_o and phase δ_o of the signal are known, and we want to determine the signal's root-mean-square amplitude Y_s. (The factor $\sqrt{2}$ is included in (5.68) because the time average of the square of the cosine is 1/2; and, correspondingly, with the factor $\sqrt{2}$ present, Y_s is the rms signal amplitude.) The noise $y(t)$ is an impediment to the

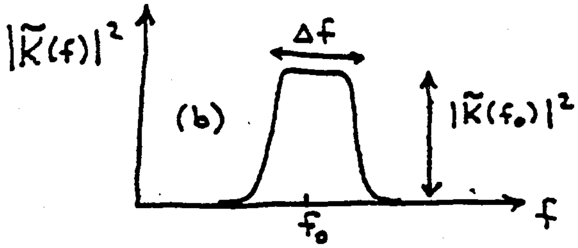


Fig. 5.8 A band-pass filter centered on frequency f_o with bandwidth Δf.

determination of Y_s. To reduce that impediment, we can send $Y(t)$ through a *band-pass filter*, i.e., a filter with a *shape* like that of Fig. 5.8. For such a filter, with central frequency f_o and with bandwidth $\Delta f \ll f_o$, the bandwidth is defined by

$$\Delta f \equiv \frac{\int_0^\infty |\tilde K(f)|^2 df}{|\tilde K(f_o)|^2} \,. \tag{5.69}$$

The output, $W(t)$ of such a filter, when $Y(t)$ is sent in, will have the form

$$W(t) = |\tilde K(f_o)|\, \sqrt{2} Y_s \cos(2\pi f_o t + \delta_1) \;+\; w(t)\,, \tag{5.70}$$

where the first term is the filtered signal and the second is the filtered noise. The output signal's phase δ_1 may be different from the input signal's phase δ_o, but that difference can be evaluated in advance for one's filter and can be taken into account in the measurement of Y_s, and thus it is of no interest to us. Assuming, as we shall, that the input noise $y(t)$ has spectral density G_y which varies negligibly over the small bandwidth of the filter, the filtered noise $\tilde w$ will have spectral density

$$G_w(f) = |\tilde K(f)|^2 G_y(f_o)\,. \tag{5.71}$$

Correspondingly, by virtue of Eq. (5.54) for the rms fluctuations of a random process at various frequencies and on various timescales, $w(t)$ will have the form

$$w(t) = w_o(t) \cos[2\pi f_o t + \phi(t)]\,, \tag{5.72}$$

with an amplitude $w_o(t)$ and phase $\phi(t)$ that fluctuate randomly on timescales $\Delta t \sim 1/\Delta f$, but that are nearly constant on timescales $\Delta t \ll 1/\Delta f$. Here Δf is the bandwidth of the filter, and hence [Eq. (5.71)] the bandwidth within which $G_w(f)$ is concentrated. The filter's net output, $W(t)$, thus consists of a precisely sinusoidal signal at frequency f_o, with known phase δ_1, and with an amplitude that we wish to determine, plus a noise $w(t)$ that is also sinusoidal at frequency f_o but that has amplitude and phase which wander randomly on timescales $\Delta t \sim 1/\Delta f$. The rms output signal is

$$S \equiv |\tilde K(f_o)| Y_s\,, \tag{5.73}$$

[Eq. (5.70)] while the rms output noise is

$$N \equiv \sigma_w = [\int_0^\infty G_w(f) df]^{\frac{1}{2}} = \sqrt{G_y(f_o)}[\int_0^\infty |\tilde K(f)|^2 df]^{\frac{1}{2}} = |\tilde K(f_o)| \sqrt{G_y(f_o) \Delta f}\,, \tag{5.74}$$

where the first integral follows from Eq. (5.32), the second from Eq. (5.71), and the third from the definition (5.69) of the bandwidth Δf. The ratio of the rms signal (5.73) to the rms noise (5.74) after filtering is

$$\frac{S}{N} = \frac{Y_s}{\sqrt{G_y(f_o)\Delta f}} \,. \tag{5.75}$$

Thus, the rms output $S+N$ of the filter is the signal amplitude to within an rms fractional error N/S given by the reciprocal of (5.75). Notice that the narrower the filter's bandwidth, the more accurate will be the measurement of the signal. In practice, of course, one does not know the signal frequency with complete precision in advance, and correspondingly one does not want to make one's filter so narrow that the signal might be lost from it.

A simple example of a band-pass filter is the following *finite-Fourier-transform filter*:

$$w(t) = \int_{t-\Delta t}^{t} \cos[2\pi f_o(t-t')]y(t')dt' \quad \text{where } \Delta t \gg 1/f_o \,. \tag{5.76}$$

In Ex. 5.1 it is shown that this is indeed a band-pass filter, and that the integration time Δt used in the Fourier transform is related to the filter's bandwidth by

$$\Delta f = \frac{1}{\Delta t} \,. \tag{5.77}$$

This is precisely the relation (5.55) that we introduced when discussing the temporal characteristics of a random process; and (setting the filter's "gain" $|\tilde{K}(f_o)|$ to unity), Eq. (5.74) for the rms noise after filtering, rewritten as $N = \sigma_w = \sqrt{G_w(f_o)\Delta f}$, is precisely expression (5.56) for the rms fluctuations in the random process $w(t)$ at frequency f_o and on timescale $\Delta t = 1/\Delta f$.

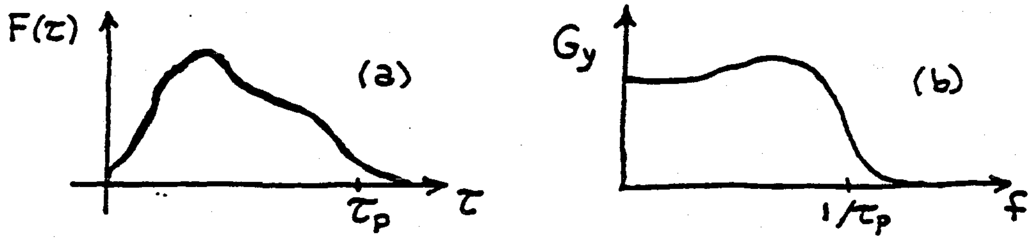


Fig. 5.9 (a) A broad-band pulse that produces shot noise by arriving at random times. (b) The spectral density of the shot noise produced by that pulse.

Shot noise. A specific kind of noise that one frequently meets and frequently wants to filter is *shot noise*. A random process $y(t)$ is said to consist of shot noise if it is a random superposition of a large number of pulses. In this chapter we shall restrict attention to a simple variant of shot noise in which the pulses all have identically the same shape, $F(\tau)$ [e.g., Fig. 5.9(a)]), but their arrival times t_i are random:

$$y(t) = \sum_i F(t-t_i) \,. \tag{5.78}$$

We denote by $\mathcal{R}$ the mean rate of pulse arrivals (the mean number per second). It is straightforward, from the definition (5.31) of spectral density, to see that the spectral density of y is

$$G_y(f) = 2\mathcal{R}|\tilde{F}(f)|^2\,, \tag{5.79}$$

where $\tilde{F}(f)$ is the Fourier transform of $F(\tau)$ [e.g., Fig. 5.9(b)]. Note that, if the pulses are broad-band bursts without much substructure in them [as in Fig. 5.9(a)], then the duration τ_p of the pulse is related to the frequency $f_{\rm max}$ at which the spectral density starts to cut off by $f_{\rm max} \sim 1/\tau_p$; and since the correlation function is the cosine transform of the spectral density, the relaxation time in the correlation function is $\tau_r \sim 1/f_{\rm max} \sim \tau_p$.

In the common (but not universal) case that many pulses are on at once on average, $\mathcal{R}\tau_p \gg 1$, $y(t)$ at any moment of time is the sum of many random processes; and, correspondingly, the central limit theorem guarantees that y is a Gaussian random process. Over time intervals smaller than $\tau_p \sim \tau_r$ the process will not generally be Markoff, because a knowledge of both $y(t_1)$ and $y(t_2)$ gives some rough indication of how many pulses happen to be on and how many new ones turned on during the time interval between t_1 and t_2 and thus are still in their early stages at time t_3; and this knowledge helps one predict $y(t_3)$ with greater confidence than if one knew only $y(t_2)$. In other words, $P_3(y_1, t_1; y_2, t_2|y_3, t_3)$ is not equal to $P_2(y_2, t_2|y_3, t_3)$, which implies non-Markoffian behavior.

On the other hand, if many pulses are on at once, and if one takes a coarse-grained view of time, never examining time intervals as short as τ_p or shorter, then a knowledge of $y(t_1)$ is of no special help in predicting $y(t_2)$, all correlations between different times are lost, and the process is Markoff and (because it is a random superposition of many independent influences) it is also Gaussian; and it thus must have the standard Gaussian-Markoff spectral density (5.38) with vanishing correlation time τ_r—i.e., it must be white. Indeed, it is: The limit of Eq. (5.79) for $f \ll 1/\tau_p$ and the corresponding correlation function are

$$G_y(f) = 2\mathcal{R}|\tilde{F}(0)|^2\,, \quad C_y(\tau) = \mathcal{R}|\tilde{F}(0)|^2\delta(\tau)\,. \tag{5.80}$$

****

EXERCISES

Exercise 5.1, Derivation and Example: Bandwidths of a finite-Fourier-transform filter and an averaging filter

a. If y is a random process with spectral density $G_y(f)$, and $w(t)$ is the output of the finite-Fourier-transform filter (5.76), what is $G_w(f)$?

b. Draw a sketch of the filter function $|\tilde{K}(f)|^2$ for this finite-Fourier-transform filter, and show that its bandwidth is given by (5.77).

c. An "averaging filter" is one which averages its input over some fixed time interval Δt:

$$w(t) \equiv \frac{1}{\Delta t} \int_{t-\Delta t}^{t} y(t')dt' . \qquad (5.81)$$

What is $|\tilde{K}(f)|^2$ for this filter? Draw a sketch of this $|\tilde{K}(f)|^2$.

d. Suppose that $y(t)$ has a spectral density that is very nearly constant at all frequencies $f \lesssim 1/\Delta t$, and that this y is put through the averaging filter (5.81). Show that the rms fluctuations in the averaged output $w(t)$ are

$$\sigma_w = \sqrt{G_y(0)\Delta f} , \qquad (5.82)$$

where Δf, interpretable as the bandwidth of the averaging filter, is

$$\Delta f = \frac{1}{2\Delta t} . \qquad (5.83)$$

(Recall that in our formalism we insist that f be nonnegative.)

Exercise 5.2, Example: Wiener's Optimal Filter

Suppose that you have a noisy receiver of weak signals (a radio telescope, or a gravitational-wave detector, or ...). You are expecting a signal $s(t)$ with finite duration and known form to come in, beginning at a predetermined time $t = 0$, but you are not sure whether it is present or not. If it is present, then your receiver's output will be

$$Y(t) = s(t) + y(t) , \qquad (5.84)$$

where $y(t)$ is the receiver's noise, a random process with spectral density $G_y(f)$ and with zero mean, $\bar{y} = 0$. If it is absent, then $Y(t) = y(t)$. A powerful way to find out whether the signal is present or not is by passing $Y(t)$ through a filter with a carefully chosen kernel $K(t)$. More specifically, compute the number

$$W \equiv \int_{-\infty}^{+\infty} K(t)Y(t)dt . \qquad (5.85)$$

If $K(t)$ is chosen optimally, then W will be maximally sensitive to the signal $s(t)$ and minimally sensitive to the noise $y(t)$; and correspondingly, if W is large you will infer that the signal was present, and if it is small you will infer that the signal was absent. This exercise derives the form of the *optimal filter*, $K(t)$, i.e., the filter that will most effectively discern whether the signal is present or not. As tools in the derivation we use the quantities S and N defined by

$$S \equiv \int_{-\infty}^{+\infty} K(t)s(t)dt , \quad N \equiv \int_{-\infty}^{+\infty} K(t)y(t)dt . \qquad (5.86)$$

Note that S is the filtered signal, N is the filtered noise, and $W = S + N$. Since $K(t)$ and $s(t)$ are precisely defined functions, S is a number; but since $y(t)$ is a random process, the value of N is not predictable, and instead is given by some probability distribution $p_1(N)$. We shall also need the Fourier transform $\tilde{K}(f)$ of the kernel $K(t)$.

a. In the measurement being done one is not filtering a function of time to get a new function of time; rather, one is just computing a number, $W = S + N$. Nevertheless, as an aid in deriving the optimal filter it is helpful to consider the time-dependent output of the filter which results when noise $y(t)$ is fed continuously into it:

$$N(t) \equiv \int_{-\infty}^{+\infty} K(t-t')y(t')dt' \,. \tag{5.87}$$

Show that this random process has a mean squared value

$$\overline{N^2} = \int_0^{\infty} |\tilde{K}(f)|^2 G_y(f) df \,. \tag{5.88}$$

Explain why this quantity is equal to the average of the *number* N^2 computed via (5.86) in an ensemble of many experiments:

$$\overline{N^2} = \langle N^2 \rangle \equiv \int p_1(N) N^2 dN = \int_0^{\infty} |\tilde{K}(f)|^2 G_y(f) df \,. \tag{5.89}$$

b. Show that of all choices of $K(t)$, the one that will give the largest value of

$$\frac{S}{\langle N^2 \rangle^{\frac{1}{2}}} \tag{5.90}$$

is Norbert Wiener's (1949) optimal filter: the $K(t)$ whose Fourier transform $\tilde{K}(f)$ is given by

$$\tilde{K}(f) = \text{const} \times \frac{\tilde{s}(f)}{G_y(f)} \,, \tag{5.91}$$

where $\tilde{s}(f)$ is the Fourier transform of the signal $s(t)$ and $G_y(f)$ is the spectral density of the noise. Note that when the noise is white, so $G_y(f)$ is independent of f, this optimal filter function is just $K(t) = \text{const} \times s(t)$; i.e., one should simply multiply the known signal form into the receiver's output and integrate. On the other hand, when the noise is not white, the optimal filter (5.91) is a distortion of const × $s(t)$ in which frequency components at which the noise is large are suppressed, while frequency components at which the noise is small are enhanced.

Exercise 5.3, Example: Alan Variance of Clocks

Highly stable clocks (e.g., Rubidium clocks or Hydrogen maser clocks) have angular frequencies ω of ticking which tend to wander so much over long time scales that their variances are divergent. More specifically, they typically show flicker noise on long time scales (low frequencies)

$$G_\omega(f) \propto 1/f \quad \text{at low } f\,; \tag{5.92}$$

and correspondingly,

$$\sigma_\omega{}^2 = \int_0^\infty G_\omega(f) df = \infty\,. \tag{5.93}$$

For this reason, clock makers have introduced a special technique for quantifying the frequency fluctuations of their clocks: They define

$$\phi(t) = \int_0^t \omega(t')dt' = (\text{phase})\,, \tag{5.94}$$

$$\Phi_\tau(t) = \frac{[\phi(t+2\tau) - \phi(t+\tau)] - [\phi(t+\tau) - \phi(t)]}{\sqrt{2}\bar{\omega}\tau}\,, \tag{5.95}$$

where $\bar{\omega}$ is the mean frequency. Aside from the $\sqrt{2}$, this is the fractional difference of clock readings for two successive intervals of duration τ. [In practice the measurement of t is made by a clock more accurate than the one being studied; or, if a more accurate clock is not available, by a clock or ensemble of clocks of the same type as is being studied.]

a. Show that the spectral density of $\Phi_\tau(t)$ is related to that of $\omega(t)$ by

$$\begin{aligned} G_{\Phi_\tau}(f) &= \frac{2}{\omega^2}\left[\frac{\cos 2\pi f\tau - 1}{2\pi f\tau}\right]^2 G_\omega(f) \\ &\propto f^2 G_\omega(f) \ \text{ at } f \ll 1/2\pi\tau\,, \\ &\propto f^{-2} G_\omega(f) \ \text{ at } f \gg 1/2\pi\tau\,. \end{aligned} \tag{5.96}$$

Note that $G_{\Phi_\tau}(f)$ is much better behaved (more strongly convergent when integrated) than $G_\omega(f)$, both at low frequencies and at high.

b. The *Alan variance* of the clock is defined as

$$\sigma_\tau{}^2 \equiv [\text{ variance of } \Phi_\tau(t)] = \int_0^\infty G_{\Phi_\tau}(f)df\,. \tag{5.97}$$

Show that

$$\sigma_\tau = \left[\alpha \frac{G_\omega(1/2\tau)}{\bar{\omega}^2}\frac{1}{2\tau}\right]^{\frac{1}{2}}\,, \tag{5.98}$$

where α is a constant of order unity which depends on the spectral shape of $G_\omega(f)$ near $f = 1/2\tau$.

c. Show that if ω has a white-noise spectrum, then the clock stability is better for long averaging times than for short [$\sigma_\tau \propto 1/\sqrt{\tau}$]; that if ω has a flicker-noise spectrum, then the clock stability is independent of averaging time; and if ω has a random-walk spectrum, then the clock stability is better for short averaging times than for long.

****

5.6 The Evolution of a System Interacting with a Heat Bath: Fluctuation-Dissipation Theorem and Fokker-Planck Equation

In this, the last section of the chapter, we use the theory of random processes to study the evolution of a semiclosed system which is interacting weakly with a heat bath. For example, we shall study the details of how an ensemble of such systems moves from a very well known state, with low entropy and with its systems concentrated in a tiny region of phase space, into statistical equilibrium where its entropy is high and its systems are spread out widely over phase space.

As a tool in the analysis of such problems, focus attention on a specific generalized coordinate q of the system being studied, for which the kinetic energy (which appears in the system's Lagrangian) is

$$E_{\rm kinetic} = \frac{1}{2}m\dot{q}^2 \,. \tag{5.99}$$

Here and below the dot denotes a time derivative, and m is the mass associated with that generalized coordinate. The force of the heat bath (i.e., of all the degrees of freedom in that bath) on the generalized coordinate q,

$$\left(m\frac{d^2q}{dt^2}\right)_{\rm bath} = F_{\rm bath} \,, \tag{5.100}$$

is a random process whose mean is a frictional (damping) force proportional to the generalized velocity $\dot{q} \equiv dq/dt$:

$$\bar{F}_{\rm bath} = -H\dot{q} \,, \quad F_{\rm bath} \equiv \bar{F}_{\rm bath} + F' \,, \tag{5.101}$$

Here H is the coefficient of friction—not to be confused with the Hamiltonian or enthalpy, which will play no role here. The fluctuating part F' of $F_{\rm bath}$ is responsible for driving q toward statistical equilibrium. We seek to study the evolution of q under the joint action of the damping and fluctuating forces $-H\dot{q}$ and F'.

Three specific examples, to which we shall return below, are these: (*i*) Our system might be a dust grain with q its x-coordinate and m its mass; and the heat bath might be air molecules at temperature T, which buffet the dust grain. This is a standard version of the problem of Brownian motion. (*ii*) Our system might be an L-C-R circuit (i.e., an electric circuit containing an inductance L, a capacitance C, and a resistance R) with q the total electric charge on the top plate of the capacitor; and the bath in this case would be the many mechanical degrees of freedom in the resistor. For such a circuit the equation of motion is

$$L\ddot{q} + C^{-1}q = F_{\rm bath}(t) = -R\dot{q} + F' \,, \tag{5.102}$$

so the effective mass is the inductance L and the coefficient of friction is the resistance R. (*iii*) The system might be the fundamental mode of a 10 kg sapphire crystal with q its generalized coordinate; and the heat bath might be all the other normal modes of vibration of the crystal, with which the fundamental mode interacts weakly.

The equation of motion for the generalized coordinate $q(t)$ under the joint action of (i) the bath's damping force $-H\dot{q}$, (ii) the bath's fluctuating forces F', and (iii) the system's "internal force" $\mathcal{F}(t)$ will be

$$m\ddot{q} + H\dot{q} = \mathcal{F}(t) + F'(t) . \tag{5.103}$$

The internal force $\mathcal{F}$ is that which one derives from the system's Hamiltonian or Lagrangian in the absence of the heat bath. For the *L-C-R* circuit of Eq. (5.102) that force is $\mathcal{F} = -C^{-1}q$; for the dust particle, if the particle is endowed with a charge e and is in an external electric field with potential $\Phi(t,x,y,z)$, it is $\mathcal{F} = -e\partial\Phi/\partial x$.

Because the equation of motion (5.103) involves a driving force $F'(t)$ that is a random process, one cannot solve it to obtain $q(t)$. Instead, one must solve it in a statistical way to obtain the evolution of q's probability distributions $p_n(q_1,t_1;\dots;q_n,t_n)$. This and other evolution equations which involve random-process driving terms are called, by modern mathematicians, *stochastic differential equations*; and there is an extensive body of mathematical formalism for solving them. In statistical physics the specific stochastic differential equation (5.103) is known as the *Langevin equation.*

Fluctuation-dissipation theorem. Because the damping force $-H\dot{q}$ and the fluctuating force F' both arise from interaction with the same heat bath, there is an intimate connection between them. For example, the stronger the coupling to the bath, the stronger will be the coefficient of friction H and the stronger will be F'. The precise relationship between the "dissipation" embodied in H and the fluctuations embodied in F' is given by the following fluctuation-dissipation theorem (also called Nyquist's theorem): *Let f be a frequency in the range*

$$\frac{1}{\tau_*} \ll f \ll \frac{1}{\tau_r} , \tag{5.104}$$

where τ_r is the (very short) relaxation time for the bath's fluctuating forces F' and where

$$\tau_* \equiv \frac{2m}{H} \tag{5.105}$$

is the timescale for dissipation to change substantially the evolution of the system. In the range of frequencies (5.104) the bath's fluctuating forces have the spectral density

$$\begin{aligned} G_{F'}(f) &= 4H\frac{hf}{e^{hf/kT}-1} \\ &= 4HkT \;\text{ if } kT \gg hf . \end{aligned} \tag{5.106}$$

Here T is the temperature of the bath and h is Planck's constant.

Notice that in the "classical" domain, $kT \gg hf$, the spectral density has a white-noise spectrum; and, in fact, since we are restricting attention to frequencies at which F' has no self correlations ($f^{-1} \gg \tau_r$), F' is Markoff; and since it is produced by interaction with the huge number of degrees of freedom of the bath, F' is also Gaussian. Thus, *in the classical domain F' is a Gaussian, Markoff, white-noise process.* At frequencies $f \gg kT/h$ (quantum domain), by contrast, the fluctuating forces are exponentially suppressed, $G_{F'}(f) = 4Hhfe^{-hf/kT}$, because any degrees of freedom in the bath that possess such high characteristic frequencies have exponentially small probabilities of containing any quanta at all, and thus exponentially small probabilities of producing fluctuating forces on q. Since this quantum-domain $G_{F'}(f)$ does not have the standard Gaussian-Markoff frequency dependence (5.38) *in the quantum domain F' is not a Gaussian-Markoff process.* It presumably fails to be Markoff because of quantum mechanical correlations in the quanta associated with F'.

Proof of the fluctuation-dissipation theorem: In principle we can alter the system's internal restoring force $\mathcal{F}$ without altering its interactions with the heat bath, i.e., without altering H or $G_{F'}(f)$. As an aid in our proof we shall choose $\mathcal{F}$ to be the restoring force of a harmonic oscillator with angular eigenfrequency ω such that $\omega/2\pi$ lies in the range of frequencies (5.104). Then the Langevin equation (5.103) takes the form

$$m\ddot{q} + H\dot{q} + m\omega^2 q = F'(t)\,. \tag{5.107}$$

This equation can be regarded as a filter which produces, from an input $F'(t)$, an output $q(t) = \int_{-\infty}^{+\infty} K(t-t')F'(t')$. The squared Fourier transform $|\tilde{K}(f)|^2$ of this Filter's kernal $K(t-t')$ is readily computed by the standard method [Eq. (5.64) and associated discussion] of inserting a sinusoid into the filter, i.e. into the differential equation, in place of F', then solving for the sinusoidal output q, and then setting $|\tilde{K}|^2 = |q|^2$. The resulting $|\tilde{K}|^2$, is the ratio of the spectral densities of input and output:

$$G_q(f) = |\tilde{K}(f)|^2 G_{F'}(f) = \frac{G_{F'}(f)}{|m[\omega^2 - (2\pi f)^2] + 2\pi i f H|^2}\,. \tag{5.108}$$

The mean energy of the oscillator, averaged over an arbitrarily long timescale, can be computed in either of two ways: (*i*) Because the oscillator is a mode of some boson field and is in statistical equilibrium with a heat bath, its mean occupation number must have the standard Bose-Einstein value $\bar{\eta} = 1/(e^{\hbar\omega/kT} - 1)$, and since each quantum carries an energy $\hbar\omega$, the mean energy is

$$\bar{E} = \frac{\hbar\omega}{e^{\hbar\omega/kT} - 1}\,. \tag{5.109}$$

(*ii*) Because on average half the energy is potential and half kinetic, and the mean potential energy is $\frac{1}{2}m\omega^2\overline{q^2}$, and because the ergodic hypothesis tells us that time averages are the same as ensemble averages, it must be that

$$\bar{E} = 2\frac{1}{2}m\omega^2\langle q^2\rangle = m\omega^2 \int_0^\infty G_q(f)df\,. \tag{5.110}$$

By inserting the spectral density (5.108) and by noting that our restriction of $\omega/2\pi$ to the range (5.104) implies a very sharp resonance in the denominator of the spectral density (5.108), and by performing the frequency integral with the help of the narrowness of the resonance, we obtain

$$\bar{E} = m\omega^2 G_{F'}(f = \omega/2\pi) \times \frac{1}{4m\omega^2 H}\,. \tag{5.111}$$

Equating this to our statistical-equilibrium expression (5.109) for the mean energy, we see that at the frequency $f = \omega/2\pi$ the spectral density $G_{F'}(f)$ has the form (5.106) claimed in the fluctuation-dissipation theorem. Moreover, since $\omega/2\pi$ can be chosen to be any frequency in the range (5.104), the spectral density $G_{F'}(f)$ has the claimed form anywhere in this range. *QED*

One example of the fluctuation-dissipation theorem is the *Johnson noise* in a resistor: In the case of the *L-C-R* circuit of Eq. (5.102) the term $-L\ddot{q}$ is the voltage across the inductor, $C^{-1}q$ is the voltage across the capacitor, $R\dot{q}$ is the dissipative voltage across the resistor, and $F'(t)$ is a fluctuating voltage [more normally denoted $V'(t)$] across the resistor. The fluctuating voltage is called "Johnson noise" and the fluctuation-dissipation relationship $G_V(f) = 4Rhf/(e^{hf/kT} - 1)$ is called *Nyquist's theorem* because J. B. Johnson (1928) discovered the voltage fluctuations $V'(t)$ experimentally and H. Nyquist (1928) derived the fluctuation-dissipation relationship for a resistor in order to explain them. The fluctuation-dissipation theorem as formulated above is a generalization of Nyquist's original theorem to any system with dissipation produced by a heat bath.

Fokker-Planck equation. Turn attention next to the details of how interaction with a heat bath drives an ensemble of simple systems, with one degree of freedom y, into statistical equilibrium. Require, for ease of analysis, that $y(t)$ be Markoff. Thus, for example, y could be the x-velocity v_x of a dust particle that is buffeted by air molecules. However, it could not be the generalized coordinate q or momentum p of a harmonic oscillator (e.g., of the fundamental mode of a sapphire crystal), since neither of them is Markoff. On the other hand, if we had developed the theory of 2-dimensional random processes, y could be the pair (q, p) of the oscillator since that pair is Markoff.

Because $y(t)$ is Markoff, all of its statistical properties are determined by its first absolute probability distribution $p_1(y)$ and its first conditional probability distribution $P_2(y_o|y,t)$. Moreover, because y is interacting with a bath, which keeps producing fluctuating forces that drive it in stochastic ways, y ultimately must reach statistical equilibrium with the bath. This means that at very late times the conditional probability $P_2(y_o|y,t)$ forgets about its initial value y_o and assumes a time-independent form which is the same as $p_1(y)$:

$$\lim_{t\to\infty} P_2(y_o|y,t) = p_1(y) . \tag{5.112}$$

Thus, the conditional probability P_2 by itself contains all the statistical information about the Markoff process $y(t)$.

As a tool in computing the conditional probability distribution $P_2(y_o|y,t)$, we shall derive a differential equation for it, called the *Fokker-Planck equation.* This Fokker-Planck equation has a much wider range of applicability than just to our degree of freedom y interacting with a heat bath. It in fact is valid for any Markoff process. The Fokker-Planck equation says

$$\frac{\partial}{\partial t}P_2 = -\frac{\partial}{\partial y}[A(y)P_2] + \frac{1}{2}\frac{\partial^2}{\partial y^2}[B(y)P_2] . \tag{5.113}$$

Here $P_2 = P_2(y_o|y,t)$ is to be regarded as a function of the variables y and t with y_o fixed;

i.e., (5.113) is to be solved subject to the initial condition

$$P_2(y_o|y,0) = \delta(y - y_o)\,. \tag{5.114}$$

As we shall see later, the Fokker-Planck equation is a diffusion equation for the probability P_2: as time passes the probability diffuses away from its initial location, $y = y_o$, spreading gradually out over a wide range of values of y.

In the Fokker-Planck equation (5.113) the function $A(y)$ produces a motion of the mean away from its initial location, while the function $B(y)$ produces the diffusion of the probability. If one knows in some other way [e.g., by solving the Langevin equation (5.103)] the evolution of P_2 for very short times, from that one can compute the functions $A(y)$ and $B(y)$:

$$A(y) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \int (y' - y) P_2(y|y'\Delta t) dy'\,, \tag{5.115}$$

$$B(y) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \int (y' - y)^2 P_2(y|y', \Delta t) dy'\,. \tag{5.116}$$

Note that the integral (5.115) for $A(y)$ is the mean change $\overline{\Delta y}$ in the value of y that occurs in time Δt, if at the beginning of Δt the value of the process is precisely y; and correspondingly we can write (5.115) in the more suggestive form

$$A(y) = \lim_{\Delta t \to 0} \left(\frac{\overline{\Delta y}}{\Delta t} \right)\,. \tag{5.117}$$

Similarly the integral (5.116) for $B(y)$ is the mean-square change in y, $\overline{(\Delta y)^2}$, if at the beginning of Δt the value of the process is precisely y; and correspondingly, (5.115) can be written

$$B(y) = \lim_{\Delta t \to 0} \left(\frac{\overline{(\Delta y)^2}}{\Delta t} \right)\,. \tag{5.118}$$

It may seem surprising that $\overline{\Delta y}$ and $\overline{(\Delta y)^2}$ can both increase linearly in time for small times [cf. the Δt in the denominators of both (5.117) and (5.118)], thereby both giving rise to finite functions $A(y)$ and $B(y)$. In fact, this is so: The linear evolution of $\overline{\Delta y}$ at small t corresponds to the motion of the mean, i.e., of the peak of the probability distribution; while the linear evolution of $\overline{(\Delta y)^2}$ corresponds to the diffusive spreading of the probability distribution.

Derivation of the Fokker-Planck equation (5.113): Because y is Markoff, it satisfies the Smoluchowski equation (5.11), which we rewrite here with a slight change of notation:

$$P_2(y_o|y, t + \tau) = \int_{-\infty}^{+\infty} P_2(y_o|y - \xi, t) P_2(y - \xi|y - \xi + \xi, \tau) d\xi\,. \tag{5.119}$$

Take τ and ξ to be small, and expand in a Taylor series in τ on the left side of (5.119) and in the ξ of $y - \xi$ on the right side:

$$\begin{aligned} P_2(y_o|y,t) + \sum_{n=1}^{\infty} \frac{1}{n!} \left[\frac{\partial^n}{\partial t^n} P_2(y_o|y,t) \right] \tau^n &= \int_{-\infty}^{+\infty} P_2(y_o|y,t) P_2(y|y+\xi,\tau) d\xi \\ &+ \sum_{n=1}^{\infty} \frac{1}{n!} \int_{-\infty}^{+\infty} (-\xi)^n \frac{\partial^n}{\partial y^n} [P_2(y_o|y,t) P_2(y|y+\xi,\tau)] d\xi\,. \end{aligned} \tag{5.120}$$

In the first integral on the right side the first term is independent of ξ and can be pulled out from under the integral, and the second term then integrates to one; thereby the first integral on the right reduces to $P_2(y_o|y,t)$, which cancels the first term on the left. The result then is

$$\sum_{n=1}^{\infty} \frac{1}{n!}\left[\frac{\partial^n}{\partial t^n} P_2(y_o|y,t)\right]\tau^n \\ = \sum_{n=1}^{\infty} \frac{(-1)^n}{n!} \frac{\partial^n}{\partial y^n}[P_2(y_o|y,t) \int_{-\infty}^{+\infty} \xi^n P_2(y|y+\xi,\tau)d\xi] \,. \tag{5.121}$$

Divide by τ, take the limit $\tau \to 0$, and set $\xi \equiv y' - y$ to obtain

$$\frac{\partial}{\partial t} P_2(y_o|y,t) = \sum_{n=1}^{\infty} \frac{(-1)^n}{n!} \frac{\partial^n}{\partial y^n}[M_n(y) P_2(y_o|y,t)] \,, \tag{5.122}$$

where

$$M_n(y) \equiv \lim_{\Delta t \to 0} \frac{1}{\Delta t} \int (y'-y)^n P_2(y|y',\Delta t) dy' \tag{5.123}$$

is the "n'th moment" of the probability distribution P_2 after time Δt. This is a form of the Fokker-Planck equation that has slightly wider validity than (5.113). Almost always, however, the only nonvanishing functions $M_n(y)$ are $M_1 \equiv A$, which describes the linear motion of the mean, and $M_2 \equiv B$, which describes the linear growth of the variance. Other moments of P_2 grow as higher powers of Δt than the first power, and correspondingly their M_n's vanish. Thus, almost always (and always, so far as we shall be concerned), Eq. (5.122) reduces to the simpler version (5.113) of the Fokker-Planck equation. *QED*

For our applications below it will be true that $p_1(y)$ can be deduced as the limit of $P_2(y_o|y,t)$ for arbitrarily large times t. Occasionally, however, this might not be so. Then, and in general, p_1 can be deduced from the time-independent Fokker-Planck equation:

$$-\frac{\partial}{\partial y}[A(y)p_1(y)] + \frac{1}{2}\frac{\partial^2}{\partial y^2}[B(y)p_1(y)] = 0 \,. \tag{5.124}$$

This equation is a consequence of the following expression for p_1 in terms of P_2,

$$p_1(y) = \int_{-\infty}^{+\infty} p_1(y_o) P_2(y_o|y,t) dy_o \,, \tag{5.125}$$

plus the fact that this p_1 is independent of t despite the presence of t in P_2, plus the Fokker-Planck equation (5.113) for P_2. Notice that, if $P_2(y_o|y,t)$ settles down into a stationary (time-independent) state at large times t, it then satisfies the same time-independent Fokker-Planck equation as $p_1(y)$, which is in accord with the obvious fact that it must then become equal to $p_1(y)$.

Brownian Motion. As an application of the Fokker-Planck equation, we use it in Ex. 5.4 to derive the following description of the evolution into statistical equilibrium of an ensemble of dust particles, all with the same mass m, being buffeted by air molecules:

Denote by $v(t)$ the x-component (or, equally well, the y- or z-component) of velocity of a dust particle. The conditional probability $P_2(v_o|v,t)$ describes the evolution into statistical equilibrium from an initial state, at time $t = 0$, when all the particles in the ensemble have velocity $v = v_o$. We shall restrict attention to time intervals large compared to the extremely small time between collisions with air molecules; i.e., we shall perform a coarse-grain average over some timescale large compared to the mean collision time. Then the fluctuating force $F'(t)$ of the air molecules on the dust particle can be regarded as a Gaussian, Markoff process with white-noise spectral density given by the classical version of the fluctuation-dissipation theorem. Correspondingly, $v(t)$ will also be Gaussian and Markoff, and will satisfy the Fokker-Planck equation (5.113). In Ex. 5.4 we shall use the Fokker-Planck equation to show that the explicit, Gaussian form of the conditional probability $P_2(v_o|v,t)$, which describes evolution into statistical equilibrium, is

$$P_2(v_o|v,t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(v-\bar{v})^2}{2\sigma^2}\right] . \tag{5.126}$$

Here the mean velocity at time t is

$$\bar{v} = v_o e^{-t/\tau_*} \quad \text{with } \tau_* \equiv \frac{m}{H} \tag{5.127}$$

the damping time due to friction; and the variance of the velocity at time t is

$$\sigma^2 = \frac{kT}{m}(1 - e^{-2t/\tau_*}) . \tag{5.128}$$

[*Side remark*: for free masses the damping time is $\tau_* = m/H$ as in (5.127), while for oscillators it is $\tau_* = 2m/H$ as in (5.105), because half the time an oscillator's energy is stored in potential form where it is protected from frictional damping, and thereby the damping time is doubled.] Notice that at very early times the variance (5.128) grows linearly with time (as the Fokker-Planck formalism says it should), and then at very late times it settles down into the standard statistical-equilibrium value:

$$\sigma^2 \simeq \frac{2kT}{m}\frac{t}{\tau_*} \text{ at } t \ll \tau_* , \quad \sigma^2 = \frac{kT}{m} \text{ at } t \gg \tau_* . \tag{5.129}$$

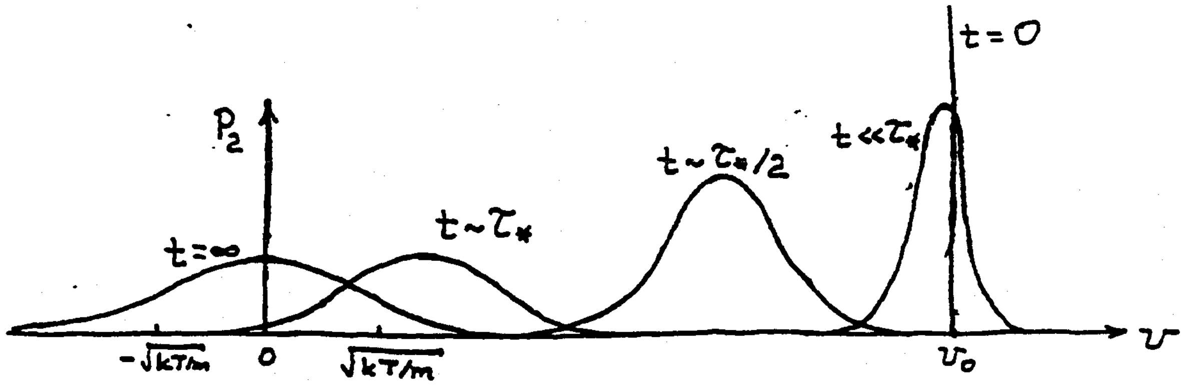


Fig. 5.10 Evolution of a dust particle into statistical equilibrium with thermalized air molecules, as described by the evolving conditional probability distribution $P_2(v_o|v,t)$.

This evolution of $P_2(v_o|v,t)$ is depicted in Fig. 5.10. Notice that, as advertised, it consists of a motion of the mean together with a diffusion of probability from the initial delta function into the standard, statistical-equilibrium, spread-out Gaussian. Correspondingly, there is a gradual loss of information about the initial velocity—the same loss of information as is quantified in the statistical mechanical increase of entropy (Chap. 3). Notice also, as advertised, that at late times $P_2(v_o|v,t)$ settles down into the same distribution as p_1: a Gaussian with zero mean velocity and with variance (i.e., mean square velocity) $\sigma^2 = kT/m$.

Since $v(t)$ is a Gaussian, Markoff process, we can use Doob's theorem (5.37)–(5.40) to read its correlation function and spectral density off its conditional probability distribution (5.126):

$$C_v(\tau) = \frac{kT}{m} e^{-t/\tau_*} \ ,$$
$$G_v(f) = \frac{4kT/m\tau_*}{(2\pi f)^2 + (1/\tau_*)^2} \ .$$

Notice that for frequencies $f \ll 1/\tau_*$, corresponding to such long timescales that initial values have been damped away and only statistical equilibrium shows up, v has a white-noise spectrum. Correspondingly, on long time scales the particle's position x, being the time integral of the velocity v, has a random-walk spectrum:

$$G_x(f) = \frac{4kT\tau_*}{m(2\pi f)^2} \quad \text{for } f \ll 1/\tau_* \ . \tag{5.130}$$

Because the motion of dust particles under the buffeting of air molecules is called a random walk, the $1/f^2$ behavior that $G_x(f)$ exhibits is called the random-walk spectrum. From this random-walk spectrum we can compute the root-mean-square (rms) distance $\sigma_{\Delta x}$ in the x-direction that the dust particle travels in a time interval $\Delta\tau \gg \tau_*$. That $\sigma_{\Delta x}$ is the standard deviation of the random process $\Delta x(t) \equiv x(t+\Delta\tau) - x(t)$. The "filter" that takes $x(t)$ into $\Delta x(t)$ has

$$|\tilde{K}(f)|^2 = |e^{i2\pi f(t+\Delta\tau)} \ - \ e^{i2\pi ft}|^2 = 4\sin^2(\pi f \Delta\tau) \ .$$

Correspondingly, $\Delta x(t)$ has spectral density

$$G_{\Delta x}(f) = |\tilde{K}(f)|^2 G_x(f) = \frac{4kT\tau_*}{m}(\Delta\tau)^2 \left(\frac{\sin(\pi f \Delta\tau)}{\pi f \Delta\tau}\right)^2 \ ; \tag{5.131}$$

and the variance of Δx (i.e., the square of the rms distance traveled) is

$$(\sigma_{\Delta x})^2 = \int_0^\infty G_{\Delta x}(f) df = \frac{2kT{\tau_*}^2}{m}\frac{\Delta\tau}{\tau_*}. \tag{5.132}$$

Thus, during time intervals $\Delta\tau$ the rms distance travelled in the x-direction by the random-walking dust particle is one "mean-free pathlength" [i.e., the mean distance it travels between collisions, i.e., the distance $(2kT/m)^{\frac{1}{2}}\tau_*$ that it would travel during one "damping time" $\tau*$ if it were moving at its rms speed] multiplied by the square root of the mean number of steps taken, $\sqrt{\Delta\tau/\tau_*}$:

$$\sigma_{\Delta x} = \left(\frac{2kT}{m}\right)^{\frac{1}{2}} \tau_* \left(\frac{\Delta\tau}{\tau_*}\right)^{\frac{1}{2}} . \tag{5.133}$$

This "square root of the number of steps taken" behavior is a feature of random walks that one meets time and again in science, engineering, and mathematics.

****

EXERCISES

Exercise 5.4, Derivation and Example: Solution of Fokker-Planck Equation for Brownian motion of a dust particle

a. Write down the explicit form of the Langevin equation [the analog of Eq. (5.103)] for the x-component of velocity $v(t)$ of a dust particle interacting with thermalized air molecules.

b. Suppose that the dust particle has velocity v at time t. By integrating the Langevin equation show that its velocity at time $t + \Delta t$ is $v + \Delta v$ where

$$m\Delta v + Hv\Delta t + \mathrm{O}[(\Delta t)^2] = \int_t^{t+\Delta t} F'(t')dt' . \tag{5.134}$$

Take an ensemble average of this and use $\overline{F'} = 0$ to conclude that the function $A(v)$ appearing in the Fokker-Planck equation (5.113) has the form

$$A(v) \equiv \lim_{\Delta t \to 0} \frac{\overline{\Delta v}}{\Delta t} = -\frac{v}{\tau_*}, \tag{5.135}$$

where $\tau_* = m/H$. Also, from (5.134) show that

$$(\Delta v)^2 = \left[-\frac{v}{\tau_*}\Delta t + \mathrm{O}[(\Delta t)^2] + \frac{1}{m}\int_t^{t+\Delta t} F'(t')dt'\right]^2 . \tag{5.136}$$

Take an ensemble average of this and use $\overline{F'(t_1)F'(t_2)} = C_{F'}(t_2 - t_1)$, together with the Wiener-Khintchine theorem, to evaluate the terms involving $F\prime$ in terms of $G_{F'}$, which in turn is known from the Fluctuation-dissipation theorem. Thereby obtain

$$B(v) = \lim_{\Delta t \to 0} \frac{\overline{(\Delta v)^2}}{\Delta t} = \frac{2HkT}{m^2} . \tag{5.137}$$

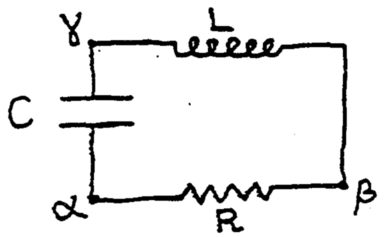


Fig. 5.11 The circuit appearing in Ex. 5.5

Insert these A and B into the Fokker-Planck equation (5.113) for $P_2(v_o|v,t)$ and show that the solution to that equation is (5.126).

Exercise 5.5, Practice: Noise in an L-C-R Circuit

Consider an L-C-R circuit as shown in Fig. 5.11. Recall that this circuit is governed by the differential equation (5.102). Suppose that the resistor has temperature T.

a. A voltmeter measures the potential difference $V_{\alpha\beta}$ between points α and β as a function of time. $V_{\alpha\beta}(t)$ fluctuates stochastically. What is its spectral density?

b. The voltmeter measures the potential difference $V_{\alpha\gamma}$ between points α and γ. What is its spectral density?

c. The voltmeter measures $V_{\beta\gamma}$. What is its spectral density?

d. The voltage $V_{\alpha\beta}$ is averaged from time $t = t_0$ to $t = t_0 + \tau$ giving some average value U_0. The average is measured once again from t_1 to $t_1 + \tau$ giving U_1. A long sequence of such measurements gives an ensemble of numbers $\{U_0, U_1, \ldots, U_n\}$. What are the mean $\bar{U}$ and root mean square deviation $\Delta U \equiv \langle (U - \bar{U})^2 \rangle^{\frac{1}{2}}$ of this ensemble?

Exercise 5.6, Example: Thermal Noise in a Sapphire Crystal

The fundamental mode of vibration of a 10 kg sapphire crystal obeys the harmonic oscillator equation

$$m(\ddot{x} + \frac{2}{\tau_*}\dot{x} + \omega^2 x) = F(t) + F'(t) , \tag{5.138}$$

where x is the displacement of the crystal's end associated with that mode, m, ω, τ_* are the effective mass, angular frequency, and amplitude damping time associated with the mode, $F(t)$ is an external driving force, and $F'(t)$ is the fluctuating force associated with the dissipation that gives rise to τ_*. Assume that $\omega\tau_* \gg 1$.

a. Weak coupling to other modes is responsible for the damping. If the other modes are thermalized at temperature T, what is the spectral density $G_{F'}(f)$ of the fluctuating force F'? What is the spectral density $G_x(f)$ of x?

b. A very weak sinusoidal force drives the fundamental mode precisely on resonance:

$$F = \sqrt{2}F_s \cos\omega t . \tag{5.139}$$

Here F_s is the rms signal. What is the $x(t)$ produced by this signal force?

c. A noiseless sensor monitors this $x(t)$ and feeds it through a narrow-band filter with central frequency $f = \omega/2\pi$ and bandwidth $\Delta f = 1/\hat{\tau}$ (where $\hat{\tau}$ is the averaging time used by the filter). Assume that $\hat{\tau} \gg \tau_*$. What is the rms thermal noise σ_x after filtering? What is the strength F_s of the signal force that produces a signal $x(t) = \sqrt{2}x_s \cos(\omega t + \delta)$ with rms amplitude equal to σ_x? This is the minimum detectable force at the "one-σ level".

d. If the force F is due to a sinusoidal gravitational wave, with dimensionless wave field $h_+(t)$ at the crystal given by $h_+ = \sqrt{2}h_s \cos\omega t$, then $F_s \sim m\omega^2 l h_s$ where l is the length of the crystal. What is the minimum detectable gravitational-wave strength h_s at the one-σ level? Evaluate h_s for the type of detector that Vladimir Braginsky and colleagues have constructed at Moscow University: A 10 kg sapphire crystal with $l \sim 50$ cm, $\omega \sim 30$ kHz, $Q \equiv \omega\tau_*/\pi \simeq 4 \times 10^9$, $T \simeq 4$K, and $\hat{\tau} \simeq 10^7$ seconds. (We shall study gravitational waves in Part VI of this book.)

****

BIBLIOGRAPHY

Random processes are treated in many standard textbooks on statistical physics, e.g. Reif (1965) and Kittel (1958). A standard treatise on signal processing is Wainstein and Zubakov (1965).

Doob, J. L. 1942. "The Brownian movement and stochastic equations" *Annals of Mathematics*, **43**, 351–369.

Johnson, J. B. 1928. "Thermal agitation of electricity in conductors" *Physical Review*, **32**, 97–109.

Kittel, C. 1958. *Elementary Statistical Physics*, New York: Wiley.

Nyquist, H. 1928. "Thermal agitation of electric charge in conductors" *Physical Review*, **32**, 110–113.

Press, William H. 1978. "Flicker noises in astronomy and elsewhere" *Comments on Astrophysics and Space Physics*, **7**, 103–119.

Reif, F. 1965. *Fundamentals of statistical and thermal physics*, New York: McGraw-Hill.

Wainstein, L. A., and Zubakov, V. D. 1965. *Extraction of Signals from Noise*, New York: McGraw-Hill.

Wiener, Norbert 1949. *The Extrapolation, Interpolation, and Smoothing of Stationary Time Series with Engineering Applications*, New York: Wiley.

REF. L

The detection of gravitational waves

Edited by
DAVID G. BLAIR
University of Western Australia

The right of the University of Cambridge to print and sell all manner of books was granted by Henry VIII in 1534. The University has printed and published continuously since 1584.

CAMBRIDGE UNIVERSITY PRESS
Cambridge
New York Port Chester
Melbourne Sydney

Published by the Press Syndicate of the University of Cambridge
The Pitt Building. Trumpington Street. Cambridge CB2 1RP
40 West 20th Street. New York NY 10011-4211. USA
10 Stamford Road. Oakleigh. Melbourne 3166. Australia

© Cambridge University Press 1991

First published 1991

Printed in Great Britain at the University Press. Cambridge

British Library cataloguing in publication data available

Library of Congress cataloguing in publication data available

ISBN 0 521 35278 9 hardback

16
Data processing, analysis, and storage for interferometric antennas

BERNARD F. SCHUTZ

16.1 Introduction

Laser-interferometric gravitational wave antennas face one of the most formidable data handling problems in all of physics. The problem is compounded of several parts: the data will be taken at reasonably high data rates (of the order of 20 kHz of 16 bit data); they may be accompanied by twice as much 'housekeeping' data to ensure that the system is working appropriately; the data will be collected 24 hours a day for many years; the data need to be searched in real time for a variety of rare, weak events of short duration (one second or less); the data need to be searched for pulsar signals; the data from two or more detectors should be cross-correlated with each other; and the data need to be archived in searchable form in case later information makes a re-analysis desirable. One detector might generate 400 Mbytes of data each hour. Even using optical discs or digital magnetic tapes with a capacity of 3 Gbytes, a network of four interferometers would generate almost 5000 discs or tapes per year. The gathering, exchange, analysis, and storage of these data will require international agreements on standards and protocols. The object of all of this effort will of course be to make astronomical observations. Because the detectors are nearly omni-directional, a network of at least three and preferably more detectors will be necessary to reconstruct a gravitational wave event completely, from which the astronomical information can be inferred.

In this chapter I will discuss the mathematical techniques for analysing the data and reconstructing the waves, the technical problems of handling the data, and the possibilities for international cooperation, as they appear in mid-1989. This discussion can only be a snapshot in time, and a personal one at that. The subject is one that can be expected to develop considerably in the next decade. I will orient the discussion toward ground-based interferometers, with the sensitivity and spectral range expected of the instruments that are planned to be built in the next decade. Much of the discussion naturally is equally applicable to present prototypes, but it is important to look ahead towards future detectors so that their data problems can be anticipated in their design. A large part of the section on data analysis also applies to space-based interferometers or to the analysis of ranging data for interplanetary spacecraft, although in these cases the volume of data is much lower because they operate as low-frequency detectors. I will also assume that the interferometers will operate with a bandwidth greater than that of the signal, even when they are configured in a resonant mode. In the extreme narrow-banding case, in which the detectors have a bandwidth smaller than that of the waves, the data analysis problem resembles that for bar detectors, as discussed by Pallottino and Pizzella in chapter 10.

16.1.1 Signals to look for

The likely sources of gravitational radiation are described by David Blair in part I of this book. If a source is strong enough to stand out above the noise in the time-series of data coming off the machine, then simple threshold-crossing criteria can be used to isolate candidate events. If the event is too weak to be seen immediately, it may still be picked up by pattern-matching techniques, but the sensitivity to such events will depend upon how much information we have about the expected waveform. At the present time, we have little idea of what waveform to expect from bursts of radiation from gravitational collapse (supernovae or electromagnetically quiet collapses), so their detectability depends upon their being strong enough to stand up above the broad-band noise. (Future detailed numerical calculations of gravitational collapse may change this, of course.) On the other hand, we have detailed predictions for the waveforms from binary coalescence and from continuous-wave sources such as pulsars; these can be extracted from noisy data by various techniques, such as matched filtering. Pulsars with a known position may be found from the output of a single detector by sampling techniques. An all-sky search for unknown pulsars will be performed at a sensitivity that will ultimately be limited by the available computing power. Cross-correlation techniques between detectors can search for a stochastic background of radiation and detect weak, unpredicted signals.

16.2 Analysis of the data from individual detectors

Bursts and continuous-wave signals can in principle be detected by looking at the output of one instrument. Of course, one must have coincident observations of the same waves in different detectors, for several reasons: to increase one's confidence that the event is real, to improve the signal-to-noise ratio of the detection, and to gain extra information with which to reconstruct the wave. The simplest detection strategy splits into two parts: first find the events in single detectors, then correlate them between detectors. In most cases this is likely to work, but in some cases it will only be possible to detect signals in the first place by cross-correlating the output of different detectors. In this section I will address the problem of finding candidate events in single detectors. Cross-correlation will be treated later.

16.2.1 Finding broad-band bursts

A broad-band burst is an event whose energy is spread across the whole of the bandwidth of the detector, which I will take to be something like 100–5000 Hz (although considerable efforts are now being devoted to techniques for extending the bandwidth down to 40 Hz or less). To be detected it has to compete against all of the detector's noise, and the only way to identify it is to see it across a pre-determined amplitude threshold in the time-series of data coming from the detector. The main burst of radiation from stellar core collapse may be like this. Numerical simulations of axisymmetric collapse (Evans, 1986; Piran and Stark, 1986) reveal, among other things, that after the main burst there is – at least if a black hole is formed – a 'ringdown phase' in which the radiation is dominated by the fundamental quasi-normal mode of the black hole. This phase lends itself to some degree of pattern-recognition, such as that which I will describe for coalescing binaries in the next section. But it is unlikely that ringdown radiation will substantially improve the signal-to-noise ratio of a collapse burst, since it is damped out very quickly. Some simplified models of non-axisymmetric collapse (e.g. Ipser and Managan, 1984) suggest that if angular momentum dominates and non-axisymmetric instabilities deform the collapsing object into a tumbling tri-axial shape, then a considerable part of the radiation will come out at a single slowly changing frequency. If future three-dimensional numerical simulations of collapse bear this out, then this would also be a candidate for pattern-matching. But one must bear in mind that even if we have good predictions of waveforms from simulations, there will be an intrinsic uncertainty due to our complete lack of knowledge of the initial conditions we might expect in a collapse, particularly regarding the angular momentum of the core. So it is not yet clear whether collapses will ever be easier to see than the time–series threshold criteria described next would indicate.

(i) Simple threshold criteria

The idea of setting thresholds is to exclude 'false alarms' – apparent events that are generated by the detector noise. Thresholds are set at a level which will guarantee that any collection of events above the threshold will be free from contamination from false alarms at some level. The 'guarantee' is of course only statistical, and it relies on understanding the noise characteristics of the detector. I will assume here that the noise is Gaussian and white over the observing bandwidth.

This should be a good first approximation, but there are at least two important refinements: first, detector noise is frequency-dependent, and when we consider coalescing binaries this will be important; and second, we must allow for unmodelled sources of noise that will occasionally produce large-amplitude 'events' in individual detectors.

This latter noise can be eliminated by demanding coincident observations in other detectors, provided we assume that it is independent of noise in the other detectors and that it is not Gaussian, in particular that there are fewer low-amplitude noise events for a given number of large-amplitude ones than we would expect of a Gaussian distribution. This implies that the cross-correlated noise between detectors will be dominated by the Gaussian component. These assumptions are usually made in data analysis, but it is important to check them as far as possible in a given set of data.

Thresholds for single detectors Assuming that the noise amplitude n in any sampled point has a Gaussian distribution with zero mean and standard deviation σ, the probability that its absolute value will exceed a threshold T (an event that we call a 'false alarm' relative to the threshold T) is

$$p(|n|>T)=\left(\frac{2}{\pi}\right)^{1/2}\frac{1}{\sigma}\int_T^{\infty}\mathrm{e}^{-n^2/2\sigma^2}\,\mathrm{d}n=\left(\frac{2}{\pi}\right)^{1/2}\left(\frac{\sigma}{T}-\frac{\sigma^3}{T^3}+\cdots\right)\mathrm{e}^{-T^2/2\sigma^2}. \quad (16.1)$$

In the asymptotic approximation given by the second equality, the first term gives 10% accuracy for $T>3.2\sigma$, and the first two terms give similar accuracy for $T>2.5\sigma$. If we want the expected number of false alarms to be one in N_{obs} data points, then we must choose T such that

$$p(|n|>T)=1/N_{\mathrm{obs}}. \quad (16.2)$$

This is a straightforward transcendental equation to solve. For example, if we imagine looking for supernova bursts of a typical duration of 1 ms, then we might be sampling the noise in the output effectively 1000 times per second. (If we want to reconstruct the waveform we might use the data at its raw sampled rate, say 4 kHz; but this would require a larger signal-to-noise ratio than simple detection, for which we could use the data sampled at or averaged over 1 ms intervals.) If we wish no more than one false alarm per year, then we must choose $T=6.6\sigma$.

Thresholds for multiple detectors If we have two detectors, with independent noise but located on the same site, then we can dig deeper into the noise by accepting only *coincidences*, which occur when both detectors simultaneously cross their respective thresholds T_1 and T_2. Given noise levels σ_1 and σ_2, respectively, the criterion for the threshold is

$$p(|n|>T_1)p(|n|>T_2)=1/N_{\mathrm{obs}}, \quad (16.3)$$

For two identical detectors ($\sigma_1=\sigma_2$), each making 1000 observations per second, the threshold T needs to be set at only $4.5\sigma_1$ to give one false alarm per year. Similarly, three identical detectors on the same site require $T=3.6\sigma$ and four can be set at $T=3.0\sigma$. The improvement from two to four detectors is a factor of 1.5 in sensitivity, or a factor of three in the volume of space that can be surveyed, and hence a similar improvement in the expected event rate. This favourable cost/benefit ratio – in this case, a factor of three improvement in event rate for a

Table 16.1. *Thresholds (in units of σ) for various arrays and false-alarm probabilities.*

Number of detectors	False-alarm probability $1/3 \times 10^{10}$	$1/1.5 \times 10^{12}$	$1/6 \times 10^{12}$	$1/3 \times 10^{14}$
1	6.63	7.19	7.37	7.88
2	4.53	4.93	5.06	5.43
3	3.59	3.92	4.03	4.33
4	3.03	3.31	3.41	3.67

factor of two increase in expenditure – is characteristic of networks of gravitational wave detectors, and indeed of any astronomical detector network whose sensitivity is limited by internal noise uncorrelated between instruments. In table 16.1 appropriate thresholds for a number of possible computer arrays and interesting false-alarm probabilities are given. (The last two columns are relevant to coalescing binaries, as discussed later.) The detectors are assumed to be identical. Notice that the thresholds are relatively insensitive to the false-alarm probability, since we are far out on the Gaussian tail. Thresholds are given in units of σ, the r.m.s. noise amplitude.

(ii) Threshold criteria with time delays

I have qualified the discussion of multiple detectors so far by demanding that they be on the same site; the reason is that if they are separated, then allowing for the possible time delay between the arrival of a true signal in different detectors opens up a larger window of time in which noise can masquerade as signal. Suppose that two detectors are separated by such a distance that the maximum time delay between them is W measurement intervals. (For example, Glasgow and California are separated by about 25 ms, which we take to be effectively ±25 measurement intervals for collapse events. This gives a total window size of 50 measurements.) Then in equation (16.3), the appropriate probability to use on the right-hand side is $1/N_{\text{obs}}/W$, since each possible 'event' in one detector must be compared with W possible coincident ones in the other.

In table 16.1, the second and fourth columns of thresholds correspond to false-alarm probabilities that are one-fiftieth of the first and third columns, respectively. For two identical detectors, this 'typical' window $W = 50$ raises the threshold T from 4.53σ to 4.93σ. This is a 9% decrease in sensitivity, or a 29% decrease in the volume of space that can be surveyed.

For three detectors, the situation begins to get more complex: as we will see later, if three detectors see an event that lasts considerably longer than their resolution time, there is a self-consistency check which may be used to reject spurious coincidences. (The check is that three detectors can determine the direction to the source, which must of course remain constant during the event.) For four detectors, even a few resolution times are enough to apply a self-consistency check. In principle, the quantitative effect of these corrections will depend on the signal-to-noise ratio of the event, since strong events can be checked for consistency more rigorously than weak events. But the level of the threshold in turn will determine the minimum signal-to-noise ratio. A full study of this problem has not yet been made, and can probably only be undertaken in the light of a more thorough investigation of the signal-reconstruction problem (see section 16.5).

16.2.2 Extracting coalescing binary signals

Coalescing binaries are good examples of the type of signal that will probably only be seen by applying pattern-matching techniques: the raw amplitude from even the nearest likely source will be below the level of broad-band noise in the detector. Nevertheless, the signal is so predictable that interferometers should be able to see such systems ten times or more as distant as collapsed sources. We will see that the signal depends on two parameters, so when we discuss the coincidence problem from the point of view of pattern-matching, we will have to consider the added uncertainty caused by this.

(i) The coalescing binary waveform

The amplitude of the radiation from a coalescing binary depends on the masses of the stars and the frequency f of the radiation, which together determine how far apart the stars are. It is usual to assume that the stars are in circular orbits. This is a safe assumption if the binary system has existed in its present form long enough for its orbit to have shrunk substantially, since the timescale for the loss of eccentricity, $e/\dot{e}$, is 2/3 of the similar timescale for the decrease of the semimajor axis a. If the binary has only recently been formed, e.g. by tidal capture in a dense star cluster, then more general waveforms can be expected. This complication will not be treated here.

Amplitude The model assumes point particles in a Newtonian orbit, with energy dissipation due to quadrupolar gravitational radiation reaction; corrections to this are discussed briefly below. The radiation amplitude when the radiation frequency is f is given by the function:

$$A_h(f) = 2.6 \times 10^{-23}\left(\frac{\mathcal{M}}{M_\odot}\right)^{5/3}\left(\frac{f}{100\ \mathrm{Hz}}\right)^{2/3}\left(\frac{100\ \mathrm{Mpc}}{r}\right), \qquad (16.4)$$

where $\mathcal{M}$ is what I shall call the *mass parameter* of the binary system, defined for a system consisting of stars of masses m_1 and m_2 by the equation

$$\mathcal{M} = m_1^{3/5} m_2^{3/5}/(m_1 + m_2)^{1/5}, \qquad (16.5)$$

or equivalently by the more transparent formula,

$$\mathcal{M}^{5/3} = \mu M_{\mathrm{T}}^{2/3}, \qquad (16.6)$$

where μ is the usual reduced mass and M_T the total mass of the system. A system consisting of two $1.4M_\odot$ stars has $\mathcal{M} = 1.22M_\odot$.

The numerical value of $A_h(f)$ is actually the *maximum* observable value of the amplitude which one obtains when the system is viewed down the axis of its angular momentum. One must insert angular factors in front of the expression to get the wave amplitude in other directions. If one averages over these angular factors *and* over the angular factors that describe the antenna pattern of an interferometer, one obtains an effective *mean amplitude* only 2/5 of the maximum (Krolak, 1989; Thorne, 1987).

Frequency The binary's orbital period changes as gravitational waves extract energy from the system. The frequency of the radiation is twice the orbital frequency, and its rate of change is

$$\frac{\mathrm{d}f}{\mathrm{d}t} = 13\left(\frac{\mathcal{M}}{M_\odot}\right)^{5/3}\left(\frac{f}{100\ \mathrm{Hz}}\right)^{11/3}\ \mathrm{Hz\ s^{-1}}. \tag{16.7}$$

The maximum wave amplitude we expect, therefore, has the time-dependence

$$h_{\max}(t) = A_h[f(t)]\cos\left(2\pi\int_{t_a}^{t} f(t')\,\mathrm{d}t' + \Phi\right), \tag{16.8}$$

where t_a is an arbitrarily defined 'arrival time', at which the signal reaches the frequency f_a, and Φ is the signal's phase at time t_a. This depends on where in their orbits the stars are when the frequency reaches f_a. The amplitude increases slowly with the frequency-dependence of A_h.

Doing the frequency integral explicitly gives

$$f(t) = 100\ \mathrm{Hz} \times \left[\left(\frac{f_a}{100\ \mathrm{Hz}}\right)^{-8/3} - 0.33\left(\frac{\mathcal{M}}{M_\odot}\right)^{5/3}\left(\frac{t-t_a}{1\ \mathrm{s}}\right)\right]^{-3/8}. \tag{16.9}$$

The phase integral is then

$$2\pi\int_{t_a}^{t} f(t')\,\mathrm{d}t' = 3000\left(\frac{\mathcal{M}}{M_\odot}\right)^{-5/3}$$
$$\times\left\{\left(\frac{f_a}{100\ \mathrm{Hz}}\right)^{-5/3} - \left[\left(\frac{f_a}{100\ \mathrm{Hz}}\right)^{-8/3} - 0.33\left(\frac{\mathcal{M}}{M_\odot}\right)^{5/3}\left(\frac{t-t_a}{1\ \mathrm{s}}\right)\right]^{5/8}\right\}. \tag{16.10}$$

Putting this into equation (16.8) for $h_{\max}(t)$ gives the desired formula, which we will use in the next section.

Notice that coalescence in the two-point-particle model occurs when $f = \infty$. For a system whose radiation is at frequency f, the remaining lifetime until this occurs is

$$T_{\mathrm{coal}}(f) = 3.0\left(\frac{\mathcal{M}}{M_\odot}\right)^{-5/3}\left(\frac{f}{100\ \mathrm{Hz}}\right)^{-8/3}\ \mathrm{s}. \tag{16.11}$$

This is 3/8 of the formal timescale $f/\dot{f}$ deducible from equation (16.7). Of course, for realistic stars the Newtonian point-particle approximation breaks down before this time, but if the stars are neutron stars or solar-mass black holes, corrections need be made only in the last second or less. Corrections due to post-Newtonian effects are the first to become important in this case, followed by tidal and mass-transfer effects. These have been considered in detail by Krolak (1989) and Krolak and Schutz (1987). If at least one of the stars is a white dwarf, tidal corrections will become important when T_{coal} is still 1000 years or so, and f is tens of millihertz; the system would only be observable from space (Evans, Iben and Smarr, 1987).

Fourier transform of the coalescing binary signal We shall need below not only the waveform $h(f)$, but also its Fourier transform. We shall denote the Fourier transform of any function $g(t)$ by $\tilde{g}(f)$, given by

$$\tilde{g}(f) = \int_{-\infty}^{\infty} g(t)\mathrm{e}^{-2\pi i f t}\,\mathrm{d}t. \qquad (16.12)$$

Provided that the frequency of the coalescing binary signal is changing relatively slowly (i.e., that $T_{\mathrm{coal}} \gg 1/f$), the method of stationary phase can be used to approximate the transform of $h_{\max}(t)$, $\tilde{h}_{\max}(f)$ (Dhurandhar, Schutz and Watkins, 1990; Thorne, 1987). We shall only need its magnitude,

$$|\tilde{h}_{\max}(f)| \approx 3.7\times 10^{-24}\left(\frac{\mathcal{M}}{M_{\odot}}\right)^{5/6}\left(\frac{f}{100\,\mathrm{Hz}}\right)^{-7/6}\left(\frac{100\,\mathrm{Mpc}}{r}\right)\mathrm{Hz}^{-1}. \qquad (16.13)$$

This gives good agreement with the results of some numerical integrations performed by Schutz (1986). We shall use it in the following sections.

(ii) The mathematics of matched filtering: finding the signal

Matched filtering is a linear pattern-matching technique designed to extract signals from noise. For references on the theory outlined in this and subsequent sections, the reader may consult a number of books on signal analysis, such as Srinath and Rajasekaran (1979).

Describing the noise To use matched filtering we have first to define some properties of the noise, $n(t)$. We expect that $n(t)$ will be a random variable, and we use angle brackets $\langle\ \rangle$ to denote expectation values of functions of this noise. It is usually more convenient to deal with the noise as a function of frequency, as described by its Fourier transform $\tilde{n}(f)$. We shall assume that the noise has zero mean,

$$\langle n(t)\rangle = \langle \tilde{n}(f)\rangle = 0.$$

We shall also assume that the noise is *stationary*, i.e. that its statistical properties are independent of time. Then the *spectral density* of (amplitude) noise $S(f)$ is defined by the equation

$$\langle \tilde{n}(f)\tilde{n}^{*}(f')\rangle = S(f)\delta(f-f'), \qquad (16.14)$$

where a * denotes complex conjugation. This says two things: (i) the noise at different frequencies is uncorrelated; and (ii) the autocorrelation of the noise at a single frequency has variances $S(f)$, apart from the normalization provided by the delta function, which arises essentially because our formalism assumes that the noise stream is infinite in duration. (Texts on signal processing often define $S(f)$ in terms of a normalized Fourier transform of the autocorrelation function of a discretely sampled time-series of noise $n_j(t)$. The continuous limit of this definition is equivalent to ours.) Since $n(t)$ is real, $S(f)$ is real and an even function of f.

Noise in an interferometer *White* noise has a constant spectrum, which means that $S(f)$ is independent of f. Interferometers have many sources of noise, as described in chapter 11 by W. Winkler in this volume or by Thorne (1987). In this treatment we will consider only two: shot noise, which limits the sensitivity of a detector at most frequencies; and seismic noise, which is idealized as a 'barrier' that makes a lower cutoff on the sensitivity of the detector at a frequency f_s.

The shot noise is intrinsically white (that is, as a noise on the photodetector), but — depending on the configuration of the detector — the detector's sensitivity to gravitational waves depends on frequency, so the relevant noise is the photon white noise divided by the frequency response of the detector (called its *transfer function*). We denote this 'gravitational wave' spectral density by $S_h(f)$. I will assume that the detector is in the standard recycling configuration, so that (allowing for the seismic cutoff) we have

$$S_h(f) = \begin{cases} \frac{1}{2}\sigma_f^2(f_k)[1 + (f/f_k)^2] & \text{for} \quad f > f_s, \\ \infty & \text{for} \quad f < f_s \end{cases} \tag{16.15}$$

Here f_k is the so-called 'knee' frequency, which may be chosen by the experimenter when recycling is implemented, and $\sigma_f(f_k)$ is the standard deviation of the frequency-domain noise at f_k.

In the usual discussions of source strength vs. detector noise (e.g. Thorne, 1987), what is taken to be the detector noise as a function of frequency f is $\sigma_f(f)$, not $[S_h(f)]^{1/2}$, because it is assumed in those discussions that the knee frequency f_k will be optimized by the experimenter for the particular range of frequencies being studied, so that σ_f is representative of the noise that the experimenter would encounter. Later in this section we will see that the optimum value of f_k for observing coalescing binaries is $1.44 f_s$.

The matched filtering theorem Now, the fundamental theorem we need in order to extract the signal from the noise is the matched filtering theorem. If we have a signal $h(t)$ buried in noise $n(t)$, so that the output of our detector is

$$o(t) = h(t) + n(t),$$

and if the Fourier transform of the signal is $\tilde{h}(f)$, then any stationary, linear operation on the output can be expressed as a correlation with a *filter* $q(t)$:

$$c(t) = (o \circ q)(t)$$

$$= \int_{-\infty}^{\infty} o(t')q(t'+t)\,\mathrm{d}t' \qquad (16.16)$$

$$= \int_{-\infty}^{\infty} \tilde{o}(f)\tilde{q}^*(f)\mathrm{e}^{2\pi ift}\,\mathrm{d}f \qquad (16.17)$$

The expectation value of the output $c(t)$ of the filter is the filter's signal,

$$\langle c(t)\rangle = (h \circ q)(t). \qquad (16.18)$$

The noise that passes through the filter is Gaussian if $n(t)$ is Gaussian, and its variance is

$$\langle [c(t) - \langle c(t)\rangle]^2\rangle = \int_{-\infty}^{\infty} S(f)\,|\tilde{q}(f)|^2\,\mathrm{d}f. \qquad (16.19)$$

This gives a 'raw' signal-to-noise ratio of

$$\frac{S}{N}(t) = \frac{(h \circ q)(t)}{\left[\int_{-\infty}^{\infty} S(f)\,|\tilde{q}(f)|^2\,\mathrm{d}f\right]^{1/2}}. \qquad (16.20)$$

The idea of matching the filter to the signal comes from finding the filter $q(t)$ that maximizes this signal-to-noise ratio. It is not difficult to show that the optimal choice of filter for detecting the signal $h(t)$ is

$$\tilde{q}(f) = k\tilde{h}(f)/S_h(f), \qquad (16.21)$$

where k is any constant. With this filter, if the output contains a signal, then $c(t)$ will reach a maximum at a time t that corresponds to the time in the output stream at which the signal reaches the point $t' = 0$ in the waveform $h(t')$. Of course, noise will distort the form of $c(t)$, but the expected amplitude signal-to-noise ratio S/N in $c(t)$ (ratio of maximum value to the standard deviation of the noise) is given by the key equation

$$\left(\frac{S}{N}\right)^2_{\mathrm{opt}} = 2\int_0^{\infty} \frac{|\tilde{h}(f)|^2}{S_h(f)}\,\mathrm{d}f. \qquad (16.22)$$

This is the largest S/N achievable with a linear filter. Moreover, given a waveform $h(t)$ that one wants to look for, and given a seismic cutoff frequency f_s, one can ask what value of the knee frequency f_k one should take in $S_h(f)$ in equation (16.22) to maximize S/N. For coalescing binaries, one can use the explicit expression for $\tilde{h}(f)$ given in equation (16.13) to show that this value, as mentioned earlier, is (Krolak, 1989; Thorne, 1987)

$$(f_{\mathrm{k}})_{\mathrm{opt}} = 1.44 f_s.$$

Thresholds for the detection of coalescing binaries Naturally, in a real experiment one does not know if a signal is present or not. One then uses the size of S/N to decide on the likelihood of the correlation being the result of noise. A widely used criterion is the Neyman–Pearson test of significance (Davis, 1989), based on the *likelihood ratio*, defined as the ratio of the probability that the signal is present to the probability that the signal is absent (false alarm). If the noise is Gaussian, then the Neyman–Pearson 'best' criterion is just to calculate the chance of false alarm in the matched filter given by equation (16.21), exactly as described in section 16.2.1(i) with x/σ replaced by S/N.

Searches for coalescing binaries can therefore be carried out by applying threshold criteria to the correlations produced by filtering. The false-alarm probabilities for detecting a coalescing binary have to be calculated with some care, however, because we must allow for the fact that we have in general to apply many independent filters, for different values of the mass parameter $\mathcal{M}$, and this increases the chance of a false alarm. I will consider the necessary corrections in section 16.2.2(iii) below.

Determining the time-of-arrival of the signal It is important for gravitational wave experiments that, by filtering the data stream, one not only determines the presence of a signal, but one also fixes its 'time-of-arrival', defined as the time $t_{\rm arr}$ at which the signal reaches the $t'=0$ point in the filter $h(t')$. The standard deviation in the measurement of $t_{\rm arr}$ is $\delta t_{\rm arr}$, which is given by an equation similar to equation (16.22) (Dhurandhar, Schutz and Watkins, 1990; Srinath and Rajasekaran, 1979):

$$\frac{1}{\delta t_{\rm arr}^2}=2\int_0^\infty \frac{|\tilde{\dot{h}}(f)|^2}{S_h(f)}\,df = 8\pi^2\int_0^\infty \frac{f^2\,|\tilde{h}(f)|^2}{S_h(f)}\,df, \tag{16.23}$$

where $\tilde{\dot{h}}(f)$ is the Fourier transform of the time derivative of $h(t)$. If either the signal or the detector's sensitivity is narrow-band about a frequency f_0, then a reasonable approximation to equation (16.23) is

$$\delta t_{\rm arr}=\frac{1}{2\pi f_0}\frac{1}{S/N}, \tag{16.24}$$

where S/N is the optimum signal-to-noise ratio as computed from equation (16.22). This is a good approximation as long as S/N is reasonably large compared to unity. If we use equation (16.13) for $\tilde{h}(f)$ then it is not hard to show that, for coalescing binaries (Dhurandhar, Schutz and Watkins, 1990)

$$\delta\tau_{\rm arr}=0.84\left(\frac{100\ \mathrm{Hz}}{f_{\rm s}}\right)\frac{1}{S/N}\,\mathrm{ms}. \tag{16.25}$$

For example, if the signal-to-noise ratio is 7 (the smallest for detection by a single detector) and the seismic limit is 100 Hz, then the timing accuracy would be 0.1 ms. If the signal-to-noise is as high as 30, which could occur a few times per year (see below), then the signal could be timed to 30 μs. Considering that the time it takes the wave to travel from one detector to another will typically be 15–20 ms, this timing accuracy would translate into good directional information. I will explain below how this can be done.

However, in practice it will turn out that these numbers are too optimistic, perhaps by a factor of two. The reason is that one needs to determine other parameters as well from the signal, such as the mass parameter $\mathcal{M}$ and the phase. The errors in these parameters correlate, with the result that $\delta t_{\rm arr}$ is affected by, for example, $\delta\mathcal{M}$. Schutz (1986) has shown numerically that a small change in the mass parameter can masquerade as a displacement in the time-of-arrival of the signal. This effect will have to be quantified before realistic estimates of the timing accuracy can be made.

Another serious source of error in timing has been stressed by Alberto Lobo (private communication). As is apparent in the calculations of Schutz (1986), when a waveform has a frequency that changes only slowly with time, there can be an ambiguity in the identification of the peak in the correlation that gives the correct time-of-arrival. This is because a shift of the filter by one cycle relative to the waveform will not degrade the correlation much if the frequency is roughly constant. Our timing accuracy formula gives in some sense the width of the correlation peak, but the spacing between peaks is much larger, of order $1/f_0$ for coalescing binaries. Unless the signal-to-noise ratio is high enough to permit reliable discrimination between peaks, this may be the dominant timing error. It is possible that cross-correlation between detectors will still be able to give correct time delays, as in section 16.4.2 below, but this remains to be investigated.

It may seem paradoxical that, if detector physicists succeed in lowering the seismic barrier to, say, 50 Hz, the arrival-time-resolution given by equation (16.24) appears to get worse as f_s^{-1}! This is not a real worsening, of course: the increase in S/N due to the lower seismic cutoff (gaining as $f_s^{-7/6}$ if f_k remains optimized to f_s) more than compensates the $1/f_s$ factor, and the timing accuracy improves.

Implications for the sampling rate In practice, one only samples the data stream at a finite rate, not continuously. It is clear from equation (16.22) that one must sample at least as fast as is required to determine $\tilde{h}(f)$ at all frequencies that contribute significantly to the integral for the optimum signal-to-noise ratio: at least twice as fast as the largest required frequency in $\tilde{h}(f)$. For the coalescing binary, whose transform is given approximately by equation (16.13), the power spectrum $|\tilde{h}(f)|^2$ falls off as $f^{-7/3}$, and the recycling shot noise multiplies a further factor of f^{-2} into this. Thus, when f rises to, say, four times f_s, the integrand in equation (16.22) will have fallen off to about 0.005 of its value at f_s. Truncating the integration here should be enough to guarantee that the filter comes within 1% of the optimum signal-to-noise ratio. This would require a sampling rate of $8f_s$, or 800 Hz if we take $f_s = 100$ Hz.

Similar but more stringent requirements apply if one wants good timing. If the sampling rate is smaller than twice the largest frequency at which the integrand in equation (16.23) contributes significantly, then in the numerical calculation the arrival time accuracy will be worse than optimum. This is an important lesson: *in choosing one's sampling speed one should ensure that one can get good accuracy in equation* (16.23), *whose integrand falls off less rapidly with frequency than that of equation* (16.22). If one does sample at an adequate rate, then it is possible to determine the time-of-arrival of a signal to much greater precision than the sampling time, provided the signal-to-noise ratio is much greater than unity. (See, for example, the numerical experiments reported by Gursel and Tinto, 1989.) For a coalescing binary, taking timing accuracy into account does not significantly increase the sampling rate over that required for a good signal-to-noise ratio.

Determining the parameters of the waveform Naively, one might expect that by performing filtering of the incoming data stream with many independent filters, one would just identify the filter that gives the best correlation with the signal and then infer the mass parameter, phase, amplitude, and time-of-arrival from that. It is possible to do better than this, however, using these values as a starting point. This is called non-linear filtering, and there are many possible ways to proceed. For our problem, one of the most attractive is the Kallianpur–Striebel (KS) filter, described by Davis (1989). Rather than reproduce Davis's clear discussion of this method, I will simply refer the reader to his article and to the M.Sc. thesis of Pasetti (1987), which is the first attempt to design a numerical system capable of detecting coalescing binary signals and estimating their parameters. Pasetti gives listings of his computer programs and tests them on simulated data.

(iii) Threshold criteria for filtered signals

Number of filters needed When searching a data stream for coalescing binary signals, we cannot presume ahead of time that we know what the mass parameter $\mathcal{M}$ will be: not all neutron stars may have mass $1.4M_\odot$, and some binaries may contain black holes of mass 15 or $20M_\odot$. We therefore will have to filter the data with a family of filters with $\mathcal{M}$ running through the range, say, $0.25–30M_\odot$.

How many filters should there be? This question has not yet received enough study. The calculations of Dhurandhar, Schutz and Watkins (1991) show that two filters with mass parameters differing by a few per cent have significantly reduced correlation, so the filters in the family should not be more widely spaced than this. However, it is not known whether they should be more closely spaced, to avoid missing weak signals. If we take successive filters to have mass parameters that increase by 1% at each step, then we need about 500 filters to span the range (0.25, 30) in $\mathcal{M}$.

However, there is also another parameter in the filter, equation (16.8): the phase Φ, about which I have so far said little. When the wave arrives at the detector with frequency f_s, so that it is just becoming detectable, its phase may be anything: this depends on the binary's history. Filters with different phases must therefore be used. Inspection of equation (16.8) reveals that the phase is a constant within the cosine term for the duration of the signal. It follows that only two filters with different phases will suffice to determine the phase and amplitude of the signal on the assumption of a given mass parameter. For convenience one might choose $\Phi = 0$ and $\Phi = \pi/2$. This increases the number of filters to about 1000. In section 16.2.2(v) we will look at the computing demands that this filtering makes on the data analysis system. In the present section we shall consider the signal-to-noise implications.

Effective sampling rate First it will be necessary to establish what the filtering equivalent of the sampling rate is, so that we can calculate the probability of, say, one false alarm per year. In our original calculation of the false-alarm probability, the sampling rate told us how many independent data points there were per year, on the assumption of white noise, which meant that each data point was statistically independent, no matter how rapidly samples were taken. In the present case, the output of the filter is the correlation given in equation (16.16). It has noise in it, but the noise is no longer white, having been filtered. The key number that we want here is the 'decorrelation time', defined as the time interval τ_s between successive applications of the filter that will ensure that the outputs of the two filters are statistically independent. The analogue here of the sampling rate in the burst problem is $1/\tau_s$, which I will call the *effective sampling rate*. This is the rate at which successive independent data points arrive from each filter.

To develop a criterion for statistical independence, we consider the autocorrelation function of the filter output when the detector output $o(t)$ is pure noise $n(t)$:

$$a(\tau) = \int_{-\infty}^{\infty} c(t)c(t+\tau)\,\mathrm{d}t. \tag{16.26}$$

We shall take the decorrelation time to be the time τ_s such that $a(\tau)$ is small for all $\tau > \tau_s$. We can learn what this is by noting that it is not hard to show that the Fourier transform of $a(\tau)$ is, when the optimal filter given in equation (16.21) is used,

$$\tilde{a}(f) = \frac{|\tilde{h}(f)|^2}{S_h(f)}. \tag{16.27}$$

For coalescing binaries, we have already discussed some of the properties of this function in section 16.2.2(ii). It is strongly peaked near f_s, and in particular the seismic barrier cuts it off rapidly below f_s. It follows that for times $\tau \gg 1/f_s$ the autocorrelation function is nearly zero: the effective sampling rate is about f_s. To play it safe, we will work with a rate twice this large, or an effective sampling time of 0.005 s. This gives effectively 6×10^9 samples – statistically independent filter outputs – per year.

Thresholds for coalescing binary filters Now, assuming that the noise is Gaussian, the calculation of the false-alarm probability for any size network looks similar to our earlier one in section 16.2.1(ii). What we have to allow for is that there will be some 1000 independent filters, each of which could give a false alarm. Of course, the false alarm occurs only if each detector registers an event in the *same* filter, so it is like doing 1000 independent experiments with no filter at all and a sampling time of 0.005 s, or one experiment with no filter and a sampling time of 5×10^{-6} s. This increases the number of points by a factor of 200 over the number we used in section 16.2.1(i), but this factor makes only a modest difference in the level of the thresholds. For example, for one false alarm per year, and no correction for time-delay windows, the thresholds are: for one detector, 7.4; for two, 5.1; for three, 4.0; and for four, 3.4. For example, the three-detector threshold is 12% higher than for unfiltered data taken at 1 kHz. For further details see table 16.1.

These figures should not be taken as graven in stone: they illustrate the consequences of a particular set of assumptions. A better calculation of the noise properties of the filters is needed, and in any case one will have to ensure that the detector noise really obeys the statistics we have assumed.

(iv) Two ways of looking at the improvement matched filtering brings

The discussion of matched filtering so far has been fairly technical, with the emphasis on making reliable and precise estimates of the achievable signal-to-noise ratios and timing accuracy. In this section I will change the approach and try to develop approximate but instructive ways of looking at the business of matched filtering. The idea is to understand how matched filtering improves the sensitivity of an interferometer beyond its sensitivity to wide-band bursts. We will look at two points of view: comparing the sensitivity of the detector to broad-band and narrow-band signals that have either (i) the same amplitude or (ii) the same total energy.

Improving the visibility of signals of a given amplitude Let us consider two signals of the same amplitude h, one of which is a broad-band burst of radiation centred at f_0 and the other of which is a relatively narrow-band signal with n cycles at roughly the frequency f_1. The signals are observed with different recycling detectors optimized at their respective frequencies, f_0 and f_1, possibly contained in the same detector system, as is envisioned in some present designs. The broad-band signal has

$$\begin{aligned}\left(\frac{S}{N}\right)^2_{\mathrm{bb}} &= 2\int_0^\infty \frac{|\tilde{h}(f)|^2}{S_h(f)}\,\mathrm{d}f. \\ &\approx \frac{2}{\sigma_f^2(f_0)}\int_0^\infty |\tilde{h}(f)|^2\,\mathrm{d}f \\ &\approx \frac{1}{\sigma_f^2(f_0)}\int_{-\infty}^\infty |h(t)|^2\,\mathrm{d}t. \qquad (16.28)\end{aligned}$$

Now, the integrand in equation (16.28) for a burst lasts typically only for a time $1/f_0$, so we have

$$\left(\frac{S}{N}\right)_{\rm bb} \approx \frac{h}{\sigma_f(f_0) f_0^{1/2}}. \tag{16.29}$$

For the narrow-band signal, we obtain again equation (16.28), but with f_0 replaced by f_1. Now, however, the signal lasts n cycles, a time n/f_1. This leads immediately to

$$\left(\frac{S}{N}\right)_{\rm nb} \approx \frac{hn^{1/2}}{\sigma_f(f_1) f_1^{1/2}}. \tag{16.30}$$

Comparing equations (16.29) and (16.30), we see that a narrow-band signal has an advantage of $n^{1/2}$ over a burst of the same amplitude and frequency, provided we have enough understanding of the signal to use matched filtering*.

For the coalescing binary one may approximate n by $f^2/\dot{f}$, and this can be large (of order 200). Coalescing binaries gain further when compared to supernova bursts because of their lower frequency: because σ_f depends on f as $f^{1/2}$, there is a further gain of a factor of f_0/f_1, which can be 7 or so. Therefore, a coalescing binary signal might have something like 100 times the S/N of a supernova burst *of the same amplitude*! This exaggerates somewhat the advantage that coalescing binaries have as a potential source of gravitational waves, since their intrinsic amplitudes may be smaller than those from supernovae, but it does show why they are such interesting sources.

Improving the visibility of signals of a given energy The other way of looking at filtering is in terms of energy. This is very instructive, because it shows 'why' matched filtering works. We have just seen that a narrow-band signal with n cycles has a higher S/N than a broad-band burst of one cycle that has the same amplitude and frequency, by a factor of $n^{1/2}$. But the *energy* in the narrow-band signal is n times that in the burst. This is because the energy flux in a gravitational wave is

$$\mathscr{F}_{\rm gw} \approx \frac{4c^3}{\pi G} h^2 f^2, \tag{16.31}$$

and thus the total energy E in a signal passing through a detector during the time n/f that the burst lasts is given by the proportionality

$$E \propto h^2 f^2 (n/f) = nfh^2.$$

If we solve this expression for nh^2 and put it into equation (16.30), we find

$$\frac{S}{N} \propto \frac{E^{1/2}}{f \sigma_f(f)}. \tag{16.32}$$

* For this reason, plots of burst sensitivity for broadband detectors, such as one finds in Thorne (1987), typically plot the *effective amplitude* $hn^{1/2}$ of a signal, rather than just h. This allows one to compare supernova bursts and coalescing binary signals on the same graph.

Since this is independent of n, it applies to broad-band and narrow-band signals equally. It shows that if two signals send the same total energy through an interferometric detector, and if they have the same frequency, then they will have the same signal-to-noise ratio, again provided we have enough information to do the matched filtering where necessary.

This provides a somewhat more realistic comparison of coalescing binaries and supernovae, since a coalescing binary radiates a substantial amount of energy in gravitational waves, of the order of $0.01M_{\odot}$. This is similar to the energy one might expect from a moderate to strong gravitational collapse. The advantage that coalescing binaries have is that they emit their energy at a lower frequency. The factor of $f\sigma_f \propto f^{3/2}$ in equation (16.32) gives them an advantage of a factor of roughly 20 over a collapse generating the same energy at the same distance. If laser interferometric detectors achieve a broad-band sensitivity of 10^{-22}, as current designs suggest will be possible, then they will be able to see moderate supernovae as far away as 50 Mpc. This volume includes several starburst galaxies, where the supernova rate may be much higher than average. They will therefore also be able to see coalescing binaries at distances approaching 1 Gpc.

(v) The technology of real-time filtering

Basic requirements In this section I will discuss the .technical feasibility of performing matched filtering on a data stream in 'real time', i.e. keeping up with the data as it comes out of a detector. Since coalescing binaries seem to make the most stringent demands, I will take them as fixing the requirements of the computing system. We have seen that we need a data stream sampled at a rate of about 1 kHz in order to obtain the best S/N and timing information, so I will use this data rate to discover the minimum requirements. It is likely that the actual sampling rates used in the experiments will be much higher, but they can easily be filtered down to 1 kHz before being analysed. If the seismic cutoff is 100 Hz, then the duration of the signal, at least until tidal or post-Newtonian effects become important, will be less than 2 s in almost all cases. This means that a filter need have no more than about 2000 2-byte data points.

The quickest way of doing the correlations necessary for filtering is to use fast Fourier transforms (FFTs) to transform the filter and signal, multiply the signal transform by the complex conjugate of the filter transform, and invert the product to find the correlation. The correlation can then be tested for places where it exceeds pre-set thresholds, and the resulting candidate events can be subjected to further analysis later. This further analysis might involve: finding the best value of the mass parameter and phase parameter; filtering with filters matched to the post-Newtonian waveform to find other parameters that could determine the individual masses of the stars; looking for unmodelled effects, such as tides or mass transfer; looking for the final burst of gravitational radiation as the two stars coalesce; and of course processing lists of these events for comparison with the outputs of other detectors. Since the number of significant events is likely to be relatively small, the most demanding aspect of this scenario is likely to be the initial correlation with 1000 coalescing binary filters.

Discrete correlations One way the processing might be done is as follows. The discrete correlation between a data set containing the N values $\{d_j, \ j = 0, \ldots, N-1\}$ and a filter containing the N values $\{h_k, k = 0, \ldots, N-1\}$ is usually given by the *circular* correlation formula:

$$c_k = (d \circ h)_k = \sum_{j=0}^{N-1} d_j h_{j+k}, \qquad k = 0, \ldots, N-1, \tag{16.33}$$

where we extend the filter by making it periodic:

$$h_{j+N} = h_j \, \forall j.$$

The circular correlation formula has a danger, because the data set and filter are not really periodic. In practice, this means that we should make the data set much longer than the (non-zero part of the) filter, so that only when the filter is 'split' between the beginning and the end of the data set does the circular correlation give the wrong answer. Thus, even if each filter requires only $N_h \leq 2000$ points, it is more efficient to split the data set up into segments of length $N \gg N_h$ points, and to use a filter which has formally the same length, but the first $N - N_h$ of whose elements are zero. (I am grateful to Harry Ward for stressing the need to pay attention to this point.) The 'padding' by zeros ensures that the periodicity of h corrupts only the last N_h elements of the correlation. This can be rectified by forgetting these elements and beginning the next data segment N_h elements before the end of the previous one: this overlap ensures that the first N_h elements of the next correlation replace the corrupt elements of the previous one with correct values. Since this procedure involves filtering some parts of the data set twice, it is desirable to make it a small fraction of the set, namely to make N_h small compared to N. This efficiency consideration is, however, balanced by the extra numerical work required to calculate long correlations, increasing as $\ln N$. This arises as follows.

Correlation by FFT The fastest way to do long correlations on a general-purpose computer is to use Fourier transforms (or related Hartley transforms). For a discrete data set $\{d_j, j = 1, \ldots, N-1\}$ the discrete (circular) Fourier transform (DFT) is the set $\{\tilde{d}_k, k = 1, \ldots, N-1\}$ given by

$$\tilde{d}_k = \sum_{j=0}^{N-1} d_j \mathrm{e}^{-2\pi ijk/N}, \tag{16.34}$$

whose inverse is

$$d_j = \frac{1}{N} \sum_{k=0}^{N-1} \tilde{d}_k \mathrm{e}^{2\pi ijk/N}. \tag{16.35}$$

Then the discrete version of the convolution theorem equation (16.17) is as follows. Given the (circular) correlation $\{c_j\}$ of two sets $\{d_j\}$ and $\{h_j\}$ as in equation (16.33), its DFT is

$$\bar{c}_k = (\bar{d}_k)^* \bar{h}_k, \qquad (16.36)$$

where an asterisk denotes complex conjugation.

Fast Fourier transform (FFT) algorithms may require typically $3N \log_2 N$ real floating-point operations (additions and multiplications) to compute the transform of a set of N real elements, provided N is an integer power of two (which can usually be arranged). (I neglect overheads due to integer arithmetic concerned with the index manipulations in such routines and, possibly significantly, memory access overheads.) To compute the correlation of two such sets, then, would require three transforms – two to produce $\bar{d}_k$ and $\bar{h}_k$ and a third to invert the product $\bar{c}_k$ – and the multiplication of the two original transforms, giving a total of $9N \log_2 N + 4N$ real floating-point operations. This is to be compared with the $2N^2 - N$ operations required to calculate the correlation directly from equation (16.33). As long as $N \geq 16$ it will be quicker to use FFTs.

In practice, one would compute once and store the DFT of all M filters, so that in real time the data would have to be transformed only once, and then M products of data and filter calculated and inverse-transformed. This would require $3N(M+1)\log_2 N + 4NM$ floating-point operations.

Optimal length of a data set We must now remind ourselves that in order to achieve the economies of the FFT algorithm, we must use the circular correlation, which has an extra cost associated with the overlaps we are required to take in successive data sets. For a given filter length (say $N_f < N$ non-zero points in the filter time-series), we can reduce the fractional size of these overlaps by making N larger, but this increases the cost of the FFT logarithmically in N. Is there an optimum ratio N_f/N? The total cost of analysing a data set containing a very large number $N_{tot} \gg N$ of elements, split up into segments of length N is

$$N_{\text{fl-pt ops}} = \frac{N_{tot}}{N - N_f}[3N(M+1)\log_2 N + 4NM].$$

We want to minimize this with respect to variations in N holding N_f and M (the number of filters) fixed. It is more convenient to introduce the varible $x = N_f/N$, which measures the fractional overlap of successive data sets. In terms of x the expression is:

$$N_{\text{fl-pt ops}}(x) = \frac{N_{tot}}{1-x}\left[3(M+1)\log_2 \frac{N_f}{x} + 4M\right]. \qquad (16.37)$$

As long as the number of filters M is large, the optimum x will be independent of M: it will depend only on N_f, the 'true' length of the filter. This is illustrated in table 16.2, which gives x and $N_{\text{fl-pt ops}}/N_{tot}M$, the number of floating-point

Table 16.2. *The consequences of various strategies for applying filters of 'true' length N_f, padded out with zeros to a length N, to very long data sets. See text, especially equation (16.37), for details.*

N_f	N	x	$N_{\text{fl-pt ops}}/N_{\text{tot}}M$
1000	2^{11}	0.488	72
1000	2^{12}	0.244	53
1000	2^{13}	0.122	49
1000	2^{14}	0.061	49
1000	2^{15}	0.031	50
2000	2^{12}	0.488	78
2000	2^{13}	0.244	57
2000	2^{14}	0.122	52
2000	2^{15}	0.061	52
2000	2^{16}	0.031	54

operations per data point per filter, as required by various strategies, always taking N_{tot} to be an integer power of two.

If we take N_f to be 2000, then the optimum x is 0.057; if $N_f = 1000$ then the best x is 0.061. But the minimum in $N_{\text{fl-pt ops}}$ is a flat one, and one can increase the value of x quite a bit without compromising speed. This is important, because each stored filter transform must contain N points, so the larger we make x, the smaller will be our core memory requirements. From this it is clear that choosing an overlap between successive data sets of around 25% gives a CPU demand that is only slightly higher than optimum and reduces storage requirements to a minimum.

Demands on computing power Based on these calculations, and assuming a data rate of 1000 2-byte samples per second with a 2 s filter length ($N_f = 2000$), it follows that doing 1000 filters in real time requires a computer capable of 60 Mflops (where 1 Mflop is 10^6 floating-point operations per second), and storage for 1000 filters, each of length 16 kbytes. This is within the capabilities of present-day inexpensive (<\$100k) workstations with add-on array-processors, or of stand-alone arrays of transputers or other fast microprocessors. In five years it should be trivial.

There are many possible ways to speed up the calculation if CPU rates are a problem. It may be that special-purpose digital-signal-processing chips would be faster than general-purpose microprocessors for this problem. It might be possible to do the calculation in block-integer format rather than floating-point, with filters that consist of crude steps rather than accurate representations of the waveform (Dewey, 1986). These should be analysed further. Another possible CPU-saver is described in the next section.

(vi) Smith's interpolation method for coalescing binaries

A different way of looking at coalescing binary signals An alternative strategy for coalescing binaries has been proposed and implemented by Smith (1987). This interesting idea is based upon the following observation: if two coalescing systems of different mass parameters happen to have the same time of coalescence, then their signals' frequencies will remain strictly proportional to one another right up to the moment of coalescence. This follows from the fact that $\mathrm{d}f/\mathrm{d}t$ is proportional to a power of f, so that, as remarked after equation (16.11), there is a constant α independent of the masses such that $T_{\mathrm{coal}} = \alpha f/\dot{f}$. If two signals with present frequencies f_1 and f_2 have the same T_{coal}, then it follows that

$$\frac{\mathrm{d}f_1}{\mathrm{d}f_2} = \frac{\dot{f}_1}{\dot{f}_2} = \frac{f_1}{f_2}.$$

Since if their times to coalescence are equal at one time then they are necessarily equal for all later times, this equation can be integrated to give $f_1/f_2 = \text{const.}$ $\forall t$

Now suppose that the data stream is sampled at constant increments of the phase of signal 1, i.e. it is sampled at a rate that accelerates with the frequency f_1. Then if a Fourier transform is performed on the sampled points, the signal will appear just as pure sinusoid, allowing it to be identified without sophisticated filtering. Moreover, and this is the key point, every other signal with the same time to coalescence will have been sampled at constant increments of its phase as well, since its frequency has been a constant times the first signal's frequency. So signals from any binary coalescing at the same time, no matter what its mass parameter, will be exposed by the single Fourier transform. Thus, one Fourier transform would seem to have done the work of *all 1000* filters!

How much work is required? The situation is not quite that good, however, because a signal with a different coalescence time will not be visible in the transform of the points sampled in the manner just described. Therefore, data must be sampled over again at the increasing rate *ending at each possible time of coalescence of the binary*. If this is done, then every possible signal will be picked up.

One way of implementing this method would be to sample the detector output at a constant rate (e.g. 1000 Hz) and then interpolate to form the data sets that are given to the FFT routine. (Livas, 1987, used this method to search for pulsars in a particular direction.) If we compare this interpolation method with the filtering described earlier, one trades the work of doing 1000 Fourier transforms on a stretch of data for the work of interpolating many times. The actual comparison depends on the number of operations required by the interpolation algorithm, but in general Smith's method with interpolation becomes more attractive as the number of filters one must use increases.

Stroboscopic sampling Another way of implementing Smith's method – and the way she herself used – would be to sample the detector output very fast, say at

10 kHz, and then to extract a data set at a slower rate (perhaps 500–1000 Hz) by selecting from the sampled points those points closest in time to the places one ideally would wish to sample. This is a far faster procedure than interpolating, and it seems to me that it would not necessarily be less accurate than a simple interpolation algorithm. I will call this *stroboscopic sampling*; we will meet it again when we discuss searches for pulsars. I do not know of any detailed theoretical analysis of it; in particular, one would like to understand what it does to the noise background. One also has to be careful about aliasing problems. The idea, at least in astronomy, seems to go back to Horowitz (1969), who devised it for optical searches for pulsars.

Comparison with matched filtering It may well be that for 1000 filters Smith's method will be more efficient than filtering. However, it has at least two significant disadvantages over filtering:

(1) It is restricted only to looking for the Newtonian coalescing binary signal: even any corrections (such as for post-Newtonian effects) will have to be searched for by filtering the sampled data sets, and the sets are essentially useless in searches for other kinds of signals that we may wish to filter from the data.

(2) Signals with the same coalescence time but different mass parameters will enter the observing window (say, $f > 100$ Hz) at different times, and this presents a possible problem that was first pointed out by Harry Ward. If one decides to break the data stream into sets of length, say, 2–3 s, appropriate to coalescing $1.4M_\odot$ neutron stars starting at 100 Hz, then the set will be much too long for a signal from a binary system of two $14M_\odot$ black holes that will coalesce at the same time. The black hole system will have frequency 24 Hz when the data set begins, and will be buried in the low-frequency detector noise. When the data are transformed, this noise will be included in the transform, and the signal-to-noise ratio will accordingly be reduced. The matched filtering method does not suffer from this drawback, since it filters out the low-frequency noise. It might be possible to avoid this problem by pre-filtering the data stream before it is sampled or interpolated, removing the low-frequency noise (and signal).

Given our present uncertainties about sources, my own prejudice is to use filtering because of its inherent flexibility; but Smith's method may become important if filtering places too great demands on the computing system.

16.2.3 Looking for pulsars and other fixed-frequency sources

(i) Why the data-analysis problem is difficult

There are many possible sources of gravitational radiation that essentially radiate at a fixed frequency. Pulsars, unstable accreting neutron stars (the Wagoner mechanism), and the possible long-term spindown of a newly formed neutron star are examples. In some cases, such as nearby known pulsars, we will know ahead of time the frequency to look for and the position of the source. But most continuous sources may have unknown frequencies; indeed they will only be discovered through their gravitational waves. I will first discuss the detection problem for sources of known frequency, and then consider searches for unknown sources. Throughout this discussion, the word 'pulsar' will stand for any continuous source with a stable frequency. The most complete discussion of this problem of which I am aware is the Ph.D. thesis of Livas (1987).

If we were on an observing platform that had a fixed velocity relative to the stars, and therefore to any pulsar we might be looking for, then finding the signal would be just a matter of taking the Fourier transform of the data and looking for a peak at the known frequency. This is a special case of matched filtering, since the Fourier integral is the same as the correlation integral in equation (16.17) with the filter equal to a sinusoid with the frequency of the incoming wave. Therefore, the signal-to-noise ratio for an observation that lasts a time T_{obs} would increase as $T_{\text{obs}}^{1/2}$, just as in equation (16.30). However, the Earth rotates on its axis and moves about the Sun and Moon, and these motions would Doppler-spread the frequency and reduce its visibility against the noise.

How long do we have to look at a source before it becomes necessary to correct for the Earth's motion? If we consider only the Earth's rotation for the moment, then in a time T_{obs} the detector's velocity relative to the source changes by an amount $\Delta v = \Omega_\oplus^2 R_\oplus T_{\text{obs}}$, where $R_\oplus$ is the Earth's radius and $\Omega_\oplus$ its angular velocity of rotation. In a source of frequency f, this produces a change $\Delta f_{\text{Dop}} = vf/c$. But the frequency resolution of an observation is $\Delta f_{\text{obs}} = 2/T_{\text{obs}}$. The Doppler effect begins to be important if $\Delta f_{\text{Dop}} = \Delta f_{\text{obs}}$. Solving this for T_{obs} gives $T_{\max}$, the maximum uncorrected observing time:

$$T_{\max} = \left(\frac{2c}{\Omega_\oplus^2 f R_\oplus}\right)^{1/2} \approx 70\left(\frac{f}{1\ \text{kHz}}\right)^{-1/2} \text{min}. \tag{16.38}$$

Using the same formula for the effects of the Earth's orbit around the Sun gives a time roughly 2.8 times as long. The Earth's motion about the Earth–Moon barycentre also has a significant effect. Since any serious observation is likely to last days or longer, the Doppler effects of all these motions must be removed, even in searches for very low-frequency signals (10 Hz).

(ii) Angular resolution of a pulsar observation

The Doppler corrections one has to apply depend on the location of the source in the sky. Since the spin axis of the Earth is not parallel to orbital angular momentum vectors of its motion about the Sun or Moon, there is no symmetry in the Doppler problem, and every location on the sky needs its own correction.

It is of interest to ask how close two points on the sky may be in order to have the same correction; this is the same as asking what the angular resolution of an observation might be. Let us first imagine for simplicity that our detector participates in only one rotational motion, with angular velocity Ω and radius R. If two sources are separated on the sky by an angle $\Delta\theta$ (in either azimuth or altitude), then the difference between the Doppler corrections for the two sources depends on the *difference* between the changes in the detector's velocities relative to the two sources. For small $\Delta\theta$ this is $\Delta v = \Delta\theta\Omega^2 R T_{\text{obs}}$. Its maximum value is $2\Omega R\Delta\theta$. Using this velocity change, the argument is otherwise identical to that given in the previous section, provided that we keep Δv no larger than $2\Omega R$. The result is that

$$\Delta\theta = T^2_{\max} \max\left(\frac{\Omega^2}{4}, \frac{1}{T^2_{\text{obs}}}\right). \tag{16.39}$$

The dependence of this expression on T_{obs} will be significant when we come to discuss all-sky searches for pulsars in section 16.2.2(v) below, so it is well to remind ourselves how it comes about. There are two factors of T_{obs} because, as T_{obs} increases, (i) our frequency resolution increases, so we are more sensitive to the Doppler effect; and (ii) the Doppler velocity change over the observing period becomes larger.

When looking at a source with a frequency of 1 kHz, then for the Earth's rotation, and an observation lasting longer than half a day, this gives

$$\Delta\theta_{\text{rot}} = 0.02\left(\frac{f}{1\text{ kHz}}\right)^{-1}\text{ rad}, \tag{16.40}$$

which is about half a degree for a millisecond pulsar. The Earth's motion about the Earth–Moon barycentre can have a greater effect, falling to a minimum of 0.002 rad at two weeks. But this is swamped by the effect of the Earth's motion about the Sun, which gives

$$\Delta\theta_{\text{orbit}} = 1\times 10^{-6}\left(\frac{f}{1\text{ kHz}}\right)^{-1}\left(\frac{T_{\text{obs}}}{10^7\text{ s}}\right)^{-2}\text{ rad}, \qquad \text{for } T_{\text{obs}} < 1\times 10^7\text{ s}. \tag{16.41}$$

This reaches a minimum of about 0.2 arcsec for a millisecond pulsar observed for four months. Even at two weeks this motion gives a resolution of 2×10^{-5} rad, much finer than the Earth–Moon motion gives. So the orbital motion of the Earth always dominates the Earth–Moon motion. But it does not dominate the Earth's rotation for short times: up to about 20 hours the limit is given by equation (16.40).

For observations longer than about a day, the Earth's orbital motion therefore affords the better angular resolution, but it also makes the most stringent demands on applying the corrections. In particular, uncertainties in the position of the pulsar being searched, for orbital motion of the pulsar in a binary system, proper motion of the pulsar (e.g., a transverse velocity of 150 km s^{-1} at 100 pc), or unpredicted changes in the period (anything larger than an accumulated fractional change $\Delta f/f$ of $10^{-10}(f/1\text{ kHz})^{-1}$) will all require special techniques to

Special techniques needed for

compensate for the way they spread the frequency out over more than the frequency resolution of the observation.

(iii) The technology of performing long Fourier transforms

We shall see that there are several different strategies one can adopt to search for pulsars, whether known ahead of time or not, but all of them can involve performing Fourier transforms of large data sets. It will help us compare the efficiencies of different strategies if we first look at how this might be done.

If one imagines that the observation lasts 10^7 s with a sampling rate of 1 kHz, then one must perform an FFT with roughly 10^{10} data points. This requires roughly $3N \log_2 N$ operations for $N = 2^{34} = 1.7 \times 10^{10}$. This evaluates to 1.7×10^{12} operations per FFT. Given the 50 Mflops computer we required earlier for filtering for coalescing binaries, this would take about 10 hours. This is not unreasonable: over 200 FFTs could be computed in the time it took to do the observation.

If 1 teraflop
then 10^{12} points
in 5 yr enough
nearly ??

The real difficulty with this is the memory requirement: FFT algorithms require access to the whole data set at once. To achieve these processing speeds, the whole data set would have to be held in fast memory, all 20 Gbytes of it. Unless there is a revolution in fast memory technology, it does not seem likely that this will be possible, at least not at an affordable level. One could imagine being able to store the data on a couple of 10-Gbyte read/write optical discs, and then using a mass-store-FFT algorithm, which uses clever paging of data in and out of store. This would still be very slow, but its exact speed would depend on the computer system.

One method of calculating the Fourier transform would be to split the data set up into M chunks of length L, each chunk being small enough to fit into core. by performing FFTs on data sets of length L it is possible to calculate the contribution of each subset to the total transform. It is not hard to show that the work needed to construct the full transform from these individual sets is about M times the work needed to do it as a single set (see, e.g., Hocking, 1989). With a memory limit of 200 Mbytes and a machine capable of 50 Mflops, it might be possible to do one or two Fourier transforms in the time it takes to do the observation. With the same memory in a machine capable of 1 Gflop, one could do 40 Fourier transforms in the same time. These are big numbers for memory and performance, but they may be within reach of the interferometer projects by the time they go on-line. The numbers become even more tractable if we are looking for a pulsar under 100 Hz: with a data rate of only 100 Hz, say, the work for a given number M of subsets goes down by a factor of about 11. It is clear that it is possible to trade-off memory against CPU speed: the technology of the time will dictate how this trade-off is to be made.

If it proves impossible to compute the full transform exactly, there are approximate methods available, such as to subdivide the full set into M subsets as above, but then only to compute the power spectrum of each subset and to add the power spectra together. This reduces the frequency resolution by a factor of M, with a proportionate decrease in the spatial resolution and in the number of different positions that an observation might need to search. It also reduces the signal-to-noise ratio of the observation. it is likely that techniques developed for radio pulsar searches (Lyne, 1989) will be useful here as well.

(iv) Detecting known pulsars

The earliest example of using a wide-band detector to search for a known pulsar is the experiment of Hough *et al.* (1983), which set an upper limit of $h < 8 \times 10^{-21}$ on radiation from the millisecond pulsar, PSR 1937 + 214. Future interferometers could better this limit by many orders of magnitude, but they will have to do long observing runs (some 10^7 s) to achieve maximum sensitivity. The analysis of the vast amount of data such experiments will generate poses greater problems for analysis than those we addressed for coalescing binaries.

Let us assume that we know the location and frequency of a pulsar, and we wish to detect its radiation. We need to make a correction for the Doppler effects from the known position, or from several contiguous positions if the position is not known accurately enough ahead of time. One might be tempted to approach this problem by filtering, as for coalescing binaries. But because of the computational demands, this is not the best method. Much better is a numerical version of the standard radio technique called *heterodyning**, followed by stroboscopic sampling.

Difficulties with filtering for pulsars Let us consider first why filtering is unsuitable. In this context a filter is just a sinusoidal signal Doppler-shifted to give the expected arrival time of any phase at our detector. If only one rotational motion of our detectors were present, and if the observation were to last several rotation periods, then only points separated in the polar direction would need separate filters: points separated in azimuth have waveforms that are simply shifted in time relative to one another, and so correlating the data in time with only one filter would take care of all such points. This might be useful even for a pulsar of known position, since it might not be known to the accuracy of equations (16.40) and (16.41).

However, our detectors participate in at least *three* rotational motions about different centres, and the observations will probably last only a fraction of a period of the most demanding motion, the solar orbital one. This means that filters lose one of their principal advantages: searching whole data-sets for similar signals arriving at different times.

Filtering requires that at least three FFTs of long data sets must be performed: of the filter, of the sampled data, and of their product to find the correlation.

* I am indebted to Jim Hough and Harry Ward for suggesting this method. The details in this section are based on conversations with them and with Norman MacKenzie, Tim Niebauer, and Roland Schilling.

Even for a well-known source, there will have to be several filters, because the phase of the wave as it arrives will not be predictable, nor will its polarization. The phase of the wave depends on exactly where the radiating 'lump' on the pulsar is. A given detector will respond to the two independent polarizations differently as it moves in orbit around the Sun; the polarization will generally be elliptical, but the proportion of the two independent polarizations and the orientation of the spin axis are unknown. Each of these variables must be filtered for, and each filter needs two more FFTs (the data set needs to be transformed only once). If the source's position and/or frequency are not known accurately, then even more filters will be required, each adding two further FFTs. Given the problems we saw we might have with FFTs, this could be a costly procedure.

Heterodyne detection Suppose the frequency of the pulsar is f_p in the barycentric frame (Solar System rest frame). Then Doppler effects of the Earth's motion plus uncertainties in the pulsar's frequency and its rate of change will require us to look in a narrow range of frequencies $(f_0, f_0 + \Delta f)$ containing f_p. The idea underlying heterodyning is that if the data contain a sinusoidal signal of frequency f,

$$s(t) = \sin(ft + \phi),$$

where ϕ is a possible phase, then if we *multiply* the signal by a 'carrier' sinusoid of frequency f_c in the bandwidth, the result can be written as

$$\sin(f_c t)s(t) = \frac{1}{2}\cos[(f - f_c)t + \phi] + \frac{1}{2}\cos[(f + f_c)t + \phi].$$

We may choose f_c so that the difference frequency $f - f_c$ is within a bandwidth Δf about zero, and yet it contains all the information (amplitude and phase) of the original signal. By filtering the resultant data set down to that bandwidth about the origin, and then re-sampling it at its (now much lower) Nyquist frequency, one can produce a data set containing many fewer points that will still contain all the information in the original band of frequencies. This set will be easier to apply Fourier transforms to than the original.

The saving in size is of order $\Delta f/f$, or 1×10^{-4} for the Doppler broadening due to the Earth's orbital motion. This would reduce the typical data set discussed in the previous section down from 10^{10} points to 10^6. This is of a size that can reasonably be handled on our 50 Mflops computer: an FFT can be done in a matter of seconds, so that complicated filtering and searches for signals become practical without expensive computing machinery.

When one looks at the details of how to implement heterodyning, one has to worry about how the noise is affected and how the procedure can be done with minimum cost. Much more work needs to be done on this question, but two possible implementations might be as follows. The first step in both is to filter the data stream with a narrow band-pass filter that allows only the required bandwidth through. This is to ensure that subsequent steps do not introduce noise (or signals) from other regions of the spectrum into our bandwidth.

In the first implementation, the next step would be to multiply by the heterodyne carrier with frequency $f_c = f_0$, i.e. at the lower edge of the bandwidth. This will ensure that noise from outside the bandwidth is not heterodyned. This allows the band-pass filter to be imperfect, as it must be if it is not to involve prohibitive amounts of computing: it will perhaps need to fall off by a factor of ten within a distance of $\Delta f/2$ of the edges of the band. Then a low-pass filter needs to be applied to get rid of the *sum* frequencies $f_c + f$. The resulting data set is still running at the rate of 10 kHz or so, but all we want is a narrow band, perhaps less than 1 Hz, about zero frequency. By *stroboscopically sampling* (defined earlier) this set at a rate equal to the appropriate Nyquist frequency $(2\Delta f)$ in the barycentric frame for signals arriving from the pulsar's direction, one can produce a data set that is at once small and Doppler-corrected. This sampling involves accepting only one point in every 10^4 or so.

The alternative implementation, which might be even faster, is based on a suggestion of Norman Mackenzie. This is to apply stroboscopic sampling (at a slow rate f_s near the Nyquist rate) directly to the data set after it has been put through the band-pass filter but before heterodyning. This may be thought of as heterodyning by aliasing: what appears in the low-frequency spectrum of the sampled data set is the aliased signal. The aliasing condition is that an original frequency f will appear in the sampled set at a frequency $f - nf_s$, where n is an integer. By choosing n and f_s appropriately, it should be possible to alias the required range of frequencies into a range near zero, without introducing extraneous noise. If the sampling is done at a rate equal to the phase arrival rate for a constant frequency in the barycentric frame at the pulsar's position, it will make all the necessary Doppler corrections automatically. Because this is potentially a very fast method, it deserves more study.

Further refinements can be made. For instance, in the first heterodyning implementation, one should multiply independently by two carriers 90 degrees apart in phase, and then add the resultant difference signals with a similar 90 degree phase shift. This reinforces the signal but adds the two independent quadratures of noise together incoherently, so that the noise is reduced by $\sqrt{2}$ relative to the signal.

Moreover, once a 'slow' data set (near zero frequency) is produced, it may still be necessary to do quite a lot of work on it to extract a pulsar signal. One will have to correct for uncertainties in the pulsar position (and hence in the stroboscopic sampling rate), for changes in the pulsar's intrinsic frequency during the observation period, for possible proper motion or binary motion effects, for the changing orientation of the detector relative to the pulsar direction and so on.

However, regardless of which of the two types of heterodyning implementations turns out to be best, the general principle is clear: if we are only interested in a bandwidth Δf about a frequency f_p, then we should be able to deal with a data set sampled at an effective rate $2\Delta f$ rather than $2f_p$. The resultant savings in computing effort make it possible to contemplate on-line searches for a few selected pulsars with computing resources that are no larger than are needed for filtering for coalescing binaries.

(v) Searching for unknown pulsars

One of the most interesting and important observations that interferometers could make is to discover old nearby pulsars or other continuous wave sources. There may be thousands of spinning neutron stars – old dead pulsars – for each currently active one. The nearest may be only tens of parsecs away. But we would have to conduct an all-sky, all-frequency search to find them. We shall see in this section that the sensitivity we can achieve in such a search is limited by computer technology.

The central problem is the number of independent points on the sky that have to be searched. As we saw in equation (16.39), the angular resolution increases as the square of the observing time, so the number of patches on the sky increases as the fourth power. For observations longer than 20 hours, equation (16.41) implies

$$N_{\text{patches}} = 4\pi/(\Delta\theta)^2 = 1.3 \times 10^{13}\left(\frac{f}{1\ \text{kHz}}\right)^2\left(\frac{T_{\text{obs}}}{10^7\ \text{s}}\right)^4. \qquad (16.42)$$

We will now look at what seems to me to be the most efficient method of searching these patches.

The barycentric Fourier transform The signal from a simple pulsar (i.e. one that does not have added complications like a binary orbit, a rapid spindown, or a large proper motion) would stand out as a strong peak if we were to compute its Fourier transform with respect to the time-of-arrival of the waves at the barycentre of the Solar System, which we take to be a convenient inertial frame. In this section I shall look at the relationship between this transform and the raw-data transform with respect to time at the detector, which relationship depends on the direction we assume for the pulsar. I also look at the relationship between the barycentric transforms of the same signal on two different assumptions for the pulsar position.

We shall need some notation. Let t_d be the time that a given part of the pulsar signal arrives at the detector. Let $t_b(\theta, \phi, t_d)$ be the time that the same signal would arrive at the barycentre if it comes from a pulsar at angular position (θ, ϕ). Let $s_d(t_d)$ be the signal itself at the detector and $s_b(t_b)$ the signal at the barycentre. Note that

$$s_b[t_b(\theta, \phi, t_d)] = s_d(t_d),$$

by definition. The relation between the two timescales is given by

$$t_b = t_d + k(\theta, \phi, t_d), \qquad (16.43)$$

where the function k is slowly varying in time for our problem,

$$\left|\frac{\partial k}{\partial t_{\mathrm{d}}}\right| \ll 1$$

due to the slow velocities that the Earth participates in. The inverse of equation (16.43) is

$$t_{\mathrm{d}} = t_{\mathrm{b}} + g(\theta, \phi, t_{\mathrm{b}}) \tag{16.44}$$

Again the derivative of g is small. From the definition it is evident that

$$g(\theta, \phi, t_{\mathrm{b}}) = -k[\theta, \phi, t_{\mathrm{b}} + g(\theta, \phi, t_{\mathrm{b}})]. \tag{16.45}$$

The exact forms of the functions g and k are complicated, but they need not concern us here.

Now we wish to find the relation between the Fourier transform of s_{b} and that of s_{a} with respect to their respective local times. For a given set of detector data, we have

$$\begin{aligned}\tilde{s}_{\mathrm{b}}(f_{\mathrm{b}}, \theta, \phi) &= \int_{-\infty}^{\infty} s_{\mathrm{b}}[t_{\mathrm{b}}(\theta, \phi)]\mathrm{e}^{-2\pi i f_{\mathrm{b}} t_{\mathrm{b}}}\,\mathrm{d}t_{\mathrm{b}},\\ &= \int_{-\infty}^{\infty} s_{\mathrm{d}}(t_{\mathrm{d}})\mathrm{e}^{-2\pi i f_{\mathrm{b}} t_{\mathrm{b}}}\,\mathrm{d}t_{\mathrm{b}}, \qquad (16.46)\\ &= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} \tilde{s}_{\mathrm{d}}(f_{\mathrm{d}})\mathrm{e}^{2\pi i f_{\mathrm{d}} t_{\mathrm{d}}}\,\mathrm{d}f_{\mathrm{d}}\right]\mathrm{e}^{-2\pi i f_{\mathrm{b}} t_{\mathrm{b}}}\,\mathrm{d}t_{\mathrm{b}},\\ &= \int_{-\infty}^{\infty} \tilde{s}_{\mathrm{d}}(f_{\mathrm{d}})m(\theta, \phi, f_{\mathrm{d}}, f_{\mathrm{b}})\,\mathrm{d}f_{\mathrm{d}}, \qquad (16.47)\end{aligned}$$

where we define

$$m(\theta, \phi, f_{\mathrm{d}}, f_{\mathrm{b}}) = \int_{-\infty}^{\infty} \mathrm{e}^{2\pi i f_{\mathrm{d}} t_{\mathrm{d}}(t_{\mathrm{b}})}\,\mathrm{e}^{-2\pi i f_{\mathrm{b}} t_{\mathrm{b}}}\,\mathrm{d}t_{\mathrm{b}}. \tag{16.48}$$

The inverse of this relation is obtained by a simple permutation of indices:

$$\tilde{s}_{\mathrm{d}}(f_{\mathrm{d}}) = \int_{-\infty}^{\infty} \tilde{s}_{\mathrm{b}}(f_{\mathrm{b}})n(\theta, \phi, f_{\mathrm{d}}, f_{\mathrm{b}})\,\mathrm{d}f_{\mathrm{b}}, \tag{16.49}$$

where the kernel here is

$$n(\theta, \phi, f_{\mathrm{d}}, f_{\mathrm{b}}) = \int_{-\infty}^{\infty} \mathrm{e}^{2\pi i f_{\mathrm{b}} t_{\mathrm{b}}(t_{\mathrm{d}})}\,\mathrm{e}^{-2\pi i f_{\mathrm{d}} t_{\mathrm{d}}}\,\mathrm{d}t_{\mathrm{d}}. \tag{16.50}$$

These equations allow us to find the barycentric transform from the detector transform, and vice versa. In principle, by applying equation (16.47) to the Fourier transform of the detector data one produces a transform in which the signal from a pulsar at a given position should stand out much more strongly. In practice, if one only wants to do this for a few cases, it is much more efficient to use stroboscopic sampling, which effectively computes equation (16.46) by selecting the appropriate values of the integrand. However, when searching the whole sky for pulsars this would involve more work than the method of the next section.

Barycentric transforms for nearby locations If one has computed the barycentric transform $\bar{s}_b$ for some location on the sky, the quickest way to find the transform for a nearby location is to find a direct transformation of $\bar{s}_b$, rather than to start again with s_d or $\bar{s}_d$. In this manner one can compute $\bar{s}_b$ for one location and then 'step' around the sky from there. We derive in this section the appropriate equations.

Consider two locations (θ, ϕ) and (θ', ϕ'). We want $\bar{s}_b$ at (θ', ϕ') in terms of that at (θ, ϕ). From equations (16.47) and (16.49) we have

$$\begin{aligned}\bar{s}_b(f'_b, \theta', \phi') &= \int_{-\infty}^{\infty} \bar{s}_d(f_d) m(\theta', \phi', f_d, f'_b)\, df_d,\\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \bar{s}_b(f_b, \theta, \phi) n(\theta, \phi, f_d, f_b) m(\theta', \phi', f_d, f'_b)\, df_d\, df_b,\\ &= \int_{-\infty}^{\infty} \bar{s}_b(f_b, \theta, \phi) q(\theta', \phi', f'_b; \theta, \phi, f_b)\, df_b, \qquad (16.51)\end{aligned}$$

where we define the 'stepping' kernel q by

$$q(\theta', \phi', f'; \theta, \phi, f) = \int_{-\infty}^{\infty} m(\theta', \phi', f'', f') n(\theta, \phi, f'', f)\, df''. \qquad (16.52)$$

If (θ', ϕ') is close to (θ, ϕ), then the kernel q should be sharply peaked in frequency near $f = f'$. In fact, it is easy to show from the inverse properties that

$$q(\theta, \phi, f'; \theta, \phi, f) = \delta(f - f') \quad \forall \theta, \phi.$$

The peaking of this function is in fact the mathematically precise way of doing the calculation we did roughly earlier, namely seeing how many independent patches on the sky one would have to search. Two angles are independent if q is wider than the frequency resolution of the observation.

The stepping method The way to do an all-sky search uses in fact the converse of the last statement. In order to convert the barycentric Fourier transform for a source at one position to that at another, one must do the integral given in equation (16.51). If the two positions are adjacent patches on the sky, then by definition the function q will be only (at least on average) two frequency bins wide, so that one can produce the barycentric transform for the second patch from that for the first by a calculation taking of order N operations, where N is the number of data points. This can represent a significant saving over doing stroboscopic sampling and an FFT for each patch. This is particularly true for large data sets that exceed the core memory capacity of the computer, because FFT algorithms on such data sets will be very much slower. The present method does not suffer from this drawback because – after the first barycentric transform has been computed – it does not require the whole transform to be held in memory at once. I shall refer to this method as *stepping around the sky.*

Depth of a search as a function of computing power We can now assemble what we know and make an assessment of the computing power required to make a search of a given sensitivity, at least by the method of stepping described here. From equation (16.42) the number of patches on the sky is

$$N_{\text{patches}} = 1.3 \times 10^{13}\left(\frac{f}{1\ \text{kHz}}\right)^2\left(\frac{T_{\text{obs}}}{10^7\ \text{s}}\right)^4.$$

The data set will have a length

$$N_{\text{pts}} = 2 \times 10^{10}\left(\frac{f}{1\ \text{kHz}}\right)\left(\frac{T_{\text{obs}}}{10^7\ \text{s}}\right)\ \text{points},$$

provided we interpret f as the highest observable frequency, so we sample at a rate $2f$. If the stepping operation between adjacent patches requires ten real floating-point operations per data point, then we need to perform

$$N_{\text{fl-pt ops}} = 2.5 \times 10^{24}\left(\frac{f}{1\ \text{kHz}}\right)^3\left(\frac{T_{\text{obs}}}{10^7\ \text{s}}\right)^5$$

floating-point operations to search the whole sky.

In 3 yrs @ 1 tflop
get 10^{20} ops @ $T_{obs} = 10^7$ s
$\Rightarrow T_{obs} \simeq 10^6$ sec $\simeq$ 12 days
gives ~~[illegible]~~ whole sky

In order to do repeatable searches, it must be possible to analyse the data in roughly the time it takes to take it. If the computer speed is called $\mathscr{S}$, measured in floating-point operations performed per second, then the time to perform $N_{\text{fl-pt ops}}$ operations is $N_{\text{fl-pt ops}}/\mathscr{S}$ s. Ignoring overheads due to other factors, we therefore find that the time to analyse the data is

$$T_{\text{anal}} = 2.5 \times 10^{16}\left(\frac{f}{1\ \text{kHz}}\right)^3\left(\frac{T_{\text{obs}}}{10^7\ \text{s}}\right)^5\left(\frac{\mathscr{S}}{100\ \text{Mflops}}\right)^{-1}\ \text{s}.$$

By equating T_{anal} and T_{obs}, we obtain the maximum observation time allowed by a computer of a given speed:

$$T_{\text{max}} = 4.4 \times 10^4\left(\frac{f}{1\ \text{kHz}}\right)^{-3/4}\left(\frac{\mathscr{S}}{100\ \text{Mflops}}\right)^{1/4}\ \text{s.} \simeq 4\times10^5\ \text{s for 1 tflop} \simeq 4\ \text{days} \qquad (16.53)$$

This is about 12 hours for a 100 Mflops computer analysing data for millisecond pulsars (up to 1 kHz). If we lower our sights and try to search for pulsars under 100 Hz (still very interesting), we can run for about three days. Another improvement comes from making a narrow-band search. This is attractive anyway, since narrow-banding enhances the detector's sensitivity in the bandwidth. In a narrow-band search one would use heterodyning to reduce the size of the data set. For a bandwidth B, the analogue of equation (16.53) is

$$T_{\max} = 2.1 \times 10^5 \left(\frac{f}{1\,\mathrm{kHz}}\right)^{-1/2} \left(\frac{B}{2\,\mathrm{Hz}}\right)^{-1/4} \left(\frac{\mathcal{S}}{100\,\mathrm{Mflops}}\right)^{1/4}\ \mathrm{s}. \qquad (16.54)$$

This is better, but still permits only about 2.4 days of observing in a narrow bandwidth at 1 kHz.

The actual figures given here may change with the invention of more efficient algorithms, but what is not likely to change is that the minimum number of operations per patch on the sky scales linearly with the number of data points. This means in turn that the permissible observation time will grow only as the fourth root of the computer speed. Even worse, since the sensitivity one can reach in h scales as the square root of the observation time, the limits on h will scale as the *eighth* root of the computer speed! Changing from a desktop computer capable of 0.1 Mflops to a supercomputer capable of 10 Gflops improves one's limits on h by only a factor of four.

This is the central problem of the all-sky search for pulsars: it is quite possible to run detectors for several months gathering data, and this will probably be done to search for known pulsars, but computing power limits any all-sky, all-frequency search for unknown pulsars to periods of the order of days.

16.3 Combining lists of candidate events from different detectors

Until now I have kept the discussion to the analysis of one detector's data, but it is clear that for the best signal-to-noise ratio and for the extraction of complete astrophysical information, detectors must operate in coincidence. I will consider in this section the simplest method of coordinated observation: exchanging lists of events detected in individual detectors. I have elsewhere (Schutz, 1989) called this the 'threshold mode' of network data analysis, because each detector's criterion for an 'event' is that its amplitude crosses a pre-set threshold.

16.3.1 Threshold mode of data analysis

We have seen in section 16.2.1(ii) how the thresholds can be determined. Once events have been identified by the on-line computer – either in the time-series of data directly or by filtering – it is important that the data from these events be brought together and analysed as quickly as possible. If the event is a supernova, we have considerably less than a day before it might become bright enough to be seen optically, and optical astronomers need to be told of it as quickly as possible. If the event is a coalescing binary, there may be even more urgency: the absence of an envelope around a neutron star means that any radiation emitted may come out with much less delay than in a supernova. Since we know so little about what such events look like, it would be valuable to have optical telescopes and orbiting X-ray telescopes observe the region of the event as quickly as possible.

The rapid exchange of data is certainly possible: with modern computer networks, it would be easy to arrange that the on-line computers could automatically circulate lists of events and associated data periodically, such as every hour. We should bear in mind that, if the threshold is set so that a network would have a four-way false alarm only once per year at a data rate of 1 kHz, then each detector will see a spurious noise-generated event three times per second! It will be impossible to distinguish the real events from the spurious until the lists of events from the various detectors are compared. The initial lists need not contain much data, so links over the usual data networks will be fast enough at this stage.

What sort of data must be exchanged? If the event is seen in a filter, the list should include the amplitude of the event, the parameters of the best-fit filter, and an agreed measure of the time the signal arrived at the detector (such as when a coalescing binary signal reached some fiducial frequency, e.g. 100 Hz). It will probably also be necessary to include calibration data, as the sensitivity of interferometers will probably change from time to time. If the signal has a high signal-to-noise ratio, then it may be desirable to include other information, such as its correlation amplitude with other filters, or even the raw unfiltered data containing the signal. The feasibility of this will depend upon the bandwidth of available communication channels.

If the event is a broad-band burst seen in the time-series, then it will be even more important to exchange the raw data, along with timing and calibration information. If raw-data exchange is impossible, then at least some description of the event will be needed, such as when it first crossed the threshold, when it reaches its maximum, and when it went below threshold.

Once likely coincidences among detectors have been identified, it will then be useful to request the on-line computers to send out more detailed information about the selected candidate coincidences. Since these requests will be rarer, it will not overburden the communications networks to exchange raw data and more complete calibration information for the times in question. If the events then still seem significant, they should be broadcast to other astronomers and analysed more thoroughly at leisure.

16.3.2 Deciding that a gravitational wave has been detected

The question that underlies all of the present article is, how do we decide that a gravitational wave has actually arrived? Various of our topics, such as the construction and use of filters and the setting of appropriate thresholds, are important components of such a decision. What we want to stress in this section is that the laser interferometer community must make sure that it has well-defined criteria for accepting a gravitational wave event as real, and a well-defined procedure for modifying and updating these criteria, *before* it begins observing in earnest.

The first detection of a gravitational wave will be such a momentous event that – if it occurs in an interferometer network – those who operate the network should leave no room for doubt that the event was well above the threshold expected of known noise sources during the time of observation. If criteria are established ahead of time, there can be no question that they have been 'adapted' to the data; conversely, if criteria for gravitational wave detection are formulated after looking at the data, there is always doubt that the events that are then identified really have the significance that might be claimed for them.

In this connection, one should not naively believe that because an unexpected event has a signal-to-noise ratio that would give it a small probability p of arising by chance, then that automatically means that the probability of its being real is correspondingly high. It is very hard to make an accurate calculation of p, since it involves not only the modelled noise but also unmodelled noise and even the circumstances of an experiment. Some Bayesian-type criterion, which involves an *a priori* estimate of the probability that the candidate event would be real, should also be used in such circumstances.

This is not to say that there should be no criteria for accepting unexpected events or unpredicted waveforms. Provided the signal-to-noise ratio of such events is high enough and they have been processed in the same way as all other data, there should be no problem accepting them. But when the signal-to-noise ratio is relatively small and/or the data have been processed in a way that had not been agreed ahead of time, there is considerable danger of accepting false alarms as real.

What can and should be done, however, if unusual events with marginal signal-to-noise ratio are seen, is that new criteria can be adopted to look for them in subsequent data. If they continue to turn up – or if re-analysis of archived data show them – then they can be accepted as real. Similarly, if new theoretical models of gravitational wave sources are evolved, they can be incorporated into the criteria. But the community should not claim detections before this second stage of verification. In particular, if there are marginal and unexpected one-off events apparently associated with rare astronomical phenomena, then it may not be possible to call them real until they have been seen again, however long that may take.

16.4 Using cross-correlation to discover unpredicted sources

The threshold mode of analysis is unsuitable for some sources, such as continuous waves or weak events that we have not predicted well enough ahead of time to construct filters for. In these cases, the 'correlation mode' is appropriate: using cross-correlations between the data streams of different detectors.

Cross-correlation has its own problems, however: its signal-to-noise relations are rather different from filtering, and the different polarizations of different detectors mean that signals in two different detectors from the same gravitational wave may not exactly correlate. In the next section I will give a general discussion of cross-correlation, addressing the behaviour of noise and assuming that the two data streams contain the same signal. One solution to the problem of polarization has been given by Gursel and Tinto (1989). Their approach will be discussed in section 16.4.2.

16.4.1 The mathematics of cross-correlation: enhancing unexpected signals

It is useful to think of cross-correlation as the use of one data stream as a filter to find things in the other data stream. Thus, if the first stream contains a signal that hasn't been predicted, one can still find it in the second. If we adopt this point of view, then we must face two important differences between matched filtering and cross-correlation as a means of enhancing signal-to-noise ratios. These are:

(1) The 'filter' is *noisy*. In fact, in the case of most interest, the signal is below the broad-band noise and the power in the filter is dominated by the noise. If we really had an instrument with an infinite bandwidth, then the noise power would be infinite and we would never see the signal. In practice, we will see below that we must filter the data down to a finite bandwidth before performing the correlation in order to achieve an acceptable signal-to-noise ratio.

(2) The 'filter' also contains the signal we wish to find, of course, but the amplitude of this part of the filter is not known *a priori*: it is the amplitude of the incoming signal. This means that if the incoming signal is reduced by half, the response of the filter to it will go down by a factor of *four*. We shall see that this leads to the biggest difference between matched filtering and cross-correlation when they are applied to long wavetrains: the enhancement of signal-to-noise in cross-correlation increases only as the fourth root of the observing time or the number of cycles in the signal, not as the square root we found in equation (16.30).

If we have two data streams o_1 and o_2 containing the same signal h but independent noise amplitudes n_1 and n_2,

$$o_1(t) = h(t) + n_1(t), \qquad o_2(t) = h(t) + n_2(t), \tag{16.55}$$

their cross-correlation is

$$o_1 \circ o_2 = h \circ h + n_1 \circ h + h \circ n_2 + n_1 \circ n_2. \tag{16.56}$$

The 'signal' is the expectation of this (averaged over both noise amplitudes), which is just $h \circ h$. The variance of the correlation, however, is a problem. The final term contributes

$$\begin{aligned}\langle |n_1 \circ n_2|^2 \rangle &= \left\langle \int \tilde{n}_1(f)\tilde{n}_1^*(f')\tilde{n}_2^*(f)\tilde{n}_2(f')\mathrm{e}^{2\pi \mathrm{i}(f-f')t}\,\mathrm{d}f\,\mathrm{d}f' \right\rangle \\ &= \int S_1(f)S_2(f)\delta(f-f')\delta(f-f')\mathrm{e}^{2\pi \mathrm{i}(f-f')t}\,\mathrm{d}f\,\mathrm{d}f'.\end{aligned}$$

The presence of *two* delta functions in the integrand makes this expression infinite: if we allow all the noise in the detectors to be cross-correlated, then the variance of the correlation will swamp the signal. The solution is (i) to *filter* the output down to a suitable bandwidth B before correlating, and (ii) to perform the correlation only over a finite stretch of data lasting a time T. If we use a superscript F to denote the filtered version of a quantity, then the analogue of $n_1 \circ n_2$ is

$$I_{12}(t) = \int_0^T n_1^{\mathrm{F}}(t')n_2^{\mathrm{F}*}(t'+t)\,\mathrm{d}t'. \tag{16.57}$$

Its variance is

$$\langle |I_{12}(t)|^2 \rangle = \int_0^T \int_0^T \langle n_1^{\mathrm{F}}(y)n_1^{\mathrm{F}*}(y')n_2^{\mathrm{F}}(y+t)n_2^{\mathrm{F}*}(y'+t)\rangle \,\mathrm{d}y\,\mathrm{d}y'. \tag{16.58}$$

The key to evaluating this is the expectation

$$\langle n_1^{\mathrm{F}}(t)n_1^{\mathrm{F}}(t')\rangle = 2\int_{f_1}^{f_2} S_1(f)\cos[2\pi f(t-t')]\,\mathrm{d}f, \tag{16.59}$$

where f_1 and f_2 are the lower and upper limits of the filtered frequency band ($f_2 = f_1 + B$), and where the factor of two arises because negative frequencies must be included in the filtered data as well as positive ones. It is a straightforward calculation to show that, assuming for simplicity that $S_i(f)$ has the constant value σ_{if}^2 over the bandwidth, then for the most important case $2\pi f_1 T \gg 1$ and $2\pi BT \gg 1$,

$$\langle |I_{12}(t)|^2 \rangle \approx 2\sigma_{1f}^2\sigma_{2f}^2 BT. \tag{16.60}$$

This part of the noise is proportional to the bandwidth of the data and the duration of the correlation. The duration will usually be chosen so that the above conditions on B and T are satisfied, for otherwise the experiment would be too brief to detect any signal that fits within the bandwidth B. The remaining contributions to the variance of the cross-correlation come from the second and third terms of equation (16.56) (strictly, from their filtered and finite-time analogues). These are just like equation (16.19), and add to equation (16.60) a term equal to $(\sigma_{1f}^2 + \sigma_{2f}^2)\int_0^T |h^{\mathrm{F}}(t)|^2\,\mathrm{d}t$.

The case of most interest to us is where the 'raw' signal $h^{\mathrm{F}}(t)$ is smaller than the time-series noise in the bandwidth B in each detector, $n_i^{\mathrm{F}}(t)$. Then the variance is dominated by equation (16.60) and we have the following expression for the signal-to-noise ratio of the cross-correlation:

$$\frac{\text{correlation signal}}{\text{correlation noise}} = \frac{\int_0^T |h^{\mathrm{F}}(t)|^2\,\mathrm{d}t}{[2\sigma_{1f}^2\sigma_{2f}^2 BT]^{1/2}}. \tag{16.61}$$

This has considerable resemblance to the filtering signal-to-noise ratio given in equation (16.20), and this justifies and makes precise our notion that cross-correlation can be thought of as using a noisy data stream as the filter. To convert equation (16.20) into equation (16.61), we must (i) replace the filter in the numerator with the signal h^{F} that is in the noisy 'filter', and (ii) replace the filter power in the denominator with the noise power of the noise filter, since we have assumed this power is the largest contributor to the noise.

However, equation (16.61) does not give us the signal-to-noise ratio for the gravitational wave signal, since its numerator is proportional to the *square* of the wave amplitude. This is the effect that we noted at the beginning of this section, that the 'filter' amplitude is proportional to the signal amplitude. A better measure of the amplitude signal-to-noise ratio is the square root of the expression in equation (16.61):

$$\frac{S}{N}=\frac{\left[\int_0^T |h^{\mathrm{F}}(t)|^2\,\mathrm{d}t\right]^{1/2}}{[2\sigma_{1f}^2\sigma_{2f}^2BT]^{1/4}}. \qquad (16.62)$$

There are two cases to consider here: long wavetrains and short pulses.

(i) Long wavetrains

The best signal-to-noise is achieved if we match the observation time T to the duration of the signal or, in the case of pulsars, make T as long as possible. Let us assume for simplicity that the two detectors have the same noise amplitude, and let us denote by R the 'raw' signal-to-noise ratio of the signal (its amplitude relative to the full detector noise in the bandwidth B),

$$R=\frac{h}{(2B\sigma_f^2)^{1/2}}.$$

Then we find

$$\frac{S}{N}\approx\left(\frac{1}{2}BT\right)^{1/4}R. \qquad (16.63)$$

The signal-to-noise ratio increases only as the fourth root of the observation time. If we are looking at, say, the spindown of a newly formed pulsar, lasting 1 s, and we filter to a bandwidth of 1 kHz because we don't know where to look for the signal, then the enhancement factor $(BT/2)^{1/4}$ is about five: short wavetrains are improved, but not dramatically. If we are looking at a pulsar, again in a broad-band search with 1 kHz bandwidth, but in an observation lasting 10^7 s, then the enhancement of signal-to-noise is a factor of about 250. This enhancement could be achieved by the $T_{\mathrm{obs}}^{1/2}$ effect in a single-detector observation lasting only three minutes, for which the data could be trivially analysed. If the single detector is narrow-banded, the time would be even less. Therefore, cross-correlation is not a good way of finding pulsars.

There are other differences between filtering and cross-correlation. Since for signals below the broad-band noise ($R<1$), we do not know where the signal is in the data stream used as a filter, it follows that we cannot determine the time-of-arrival of the signal from the correlation, apart from a relatively crude determination based upon the presence or absence of correlations between given data sets of length T. The correlation also does not tell us the waveform and therefore it cannot determine the true amplitude of the signal. It can, however, determine the time-delays between the arrival of brief events at different detectors.

(ii) Short pulses

Here one would set the bandwidth B equal to that of the pulse; if the pulse has duration roughly $T = 1/B$, and if again the two detectors have the same noise amplitude, then equation (16.62) gives a signal-to-noise ratio that is a factor of roughly $2^{1/4} \approx 1.2$ smaller than the optimum that filtering can achieve. For $TB \approx 1$ our approximations are breaking down, but it is reasonable that using this noisy filter would reduce the signal-to-noise by a factor of order two. Since in this case filtering does not enhance the signal-to-noise ratio, neither does cross-correlation: if a pulse is too weak to be seen above the broad-band (bandwidth B) noise in one detector, if will not be found by cross-correlation.

16.4.2 Cross-correlating differently polarized detectors

A more sophisticated approach to correlation has been devised by Gursel and Tinto (1989) in their approach to the signal-reconstruction problem, which I will describe in detail in section 16.5 below. It works if there are at least three detectors in the network. I shall neglect noise for simplicity in describing the method. If we let θ and ϕ be the angles describing the position of the source on the sky and we use α_i, β_i, and χ_i to represent the latitude, longitude, and orientation of the ith detector, respectively, and if we have some definition of polarization of the waves so that we can describe any wave by its amplitudes h_+ and $h_\times$, then the response $r = \delta l/l$ of the ith detector is a function of the form

$$r_i(t) = E_{+i}(\theta, \phi, \alpha_i, \beta_i, \chi_i)h_+[t - \tau_i(\theta, \phi)] \\ + E_{\times i}(\theta, \phi, \alpha_i, \beta_i, \chi_i)h_\times[t - \tau_i(\theta, \phi)], \quad (16.64)$$

where $\tau_i(\theta, \phi)$ is the time-delay between receiving a wave coming from the direction (θ, ϕ) at some standard location and at the position of the detector. We shall define the 'standard location' by setting $\tau_1 = 0$. We need not be concerned here with the precise form of the functions E_{+i}, $E_{\times i}$, and τ_i, nor with the exact definitions of the various angles.

The response equations of the first two detectors may be solved for h_+ and $h_\times$ and substituted into the response equation for the third to predict its response, for an assumed direction to the source. Let this prediction be $r_{3\text{-pred}}$:

$$r_{3\text{-pred}}(t) = -[D_{23}r_1(t - \tau_3) + D_{31}r_2(t + \tau_2 - \tau_3)]/D_{12}, \quad (16.65)$$

where D_{ij} is the determinant

$$D_{ij} = E_{+i}E_{\times j} - E_{\times i}E_{+j}.$$

If there were no noise in the detectors, then for some choice of angles θ and ϕ there would be exact agreement between $p_{3\text{-pred}}$ and the actual data from detector 3, $r_{3\text{-obs}}$. Given the noise, the best one can do is to find the angles that minimize the squared difference $d(\theta, \phi)$ between the predicted and observed responses during the interval of observation:

$$d(\theta, \phi) = \int_0^T |r_{3\text{-obs}}(t) - r_{3\text{-pred}}(t)|^2 \, \mathrm{d}t. \qquad (16.66)$$

Hidden in the integral for d are the correlation integrals we began with, e.g. $\int r_3(t) r_1(t - \tau_3) \, \mathrm{d}t$. These will normally be the most time-consuming part of the computation of d for various angles, and should usually be done by FFTs. Once the correlations have been computed for all possible time-delays, they may be used to find the minimum of d over all angles; this will determine the position of the source. Notice that if the noise is small, this information can then be substituted back into equation (16.64) for the first two detectors to find $h_+(t)$ and $h_\times(t)$. This reconstructs the signal. But if the source is weaker than the noise, then this substitution will give mostly noise.

The information we have gained about the unpredicted source, even if it is weak, is that it is there: its position is known and its arrival time can be determined roughly by restricting the time-interval over which the correlation integrals are done and finding the interval during which one gets significant correlations.This is enough to alert other astronomers to look for something in the source's position.

The paper by Gursel and Tinto (1989) contains a more sophisticated treatment of the noise than we have described here, allowing for different detectors to have different levels of noise, and constructing almost optimal filters for the signals that weight given detector responses according to where in their antenna pattern the signal seems to be coming from. They also give the results of extensive simulations and estimate the signal-to-noise ratio that will be required to give good predicitons. This paper is an important advance towards a robust solution of the reconstruction problem.

16.4.3 Using cross-correlation to search for a stochastic background

Another very important observation that interferometers will make is to find or set limits upon a background of radiation. This is much easier to do than finding discrete sources of continuous radiation, because there is no direction-finding or frequency-searching to do. This problem has been discussed in detail by Michelson (1987).

The most sensitive search for a background would be with two detectors on the same site, with the same polarization. Current plans for some installations envision more than one interferometer in one vacuum system, which would permit a correlation search. One would have to take care that common external sources of noise are excluded, especially seismic and other ground disturbances, but if this can be done then the two detectors should respond identically to any random waves coming in, and should therefore have the maximum possible correlation for these waves. The correlation can be calculated either by direct multiplication of the sampled data points ($2N$ operations per time delay between the two data sets) or by Fourier transform methods as in section 16.2.3(iii) above. We are only interested in the zero-time-delay value of the correlation, but in order to test the reality of the observed correlation, one would have to compute points at other time delays, where the correlation is expected to fall off. (How rapidly it falls off with increasing time delay depends on the spectrum of the background.) The choice of technique – direct multiplication or Fourier transform – will depend on the number of time-delays one wishes to compute and the capacity of one's computer.

If separated detectors are used, the essential physical point is that two separated detectors will still respond to waves in the same way if the waves have a wavelength λ much longer than the separation between the detectors. Conversely, if the separation between detectors is greater than $\lambda/2\pi$, there is a significant loss of correlation. It is important as well to try to orient the detectors as nearly as possible in the same polarization state. In order to perform a search at 100 Hz, the maximum separation one would like to have is 500 km. This may be achievable within Europe, but it seems most unlikely that detectors in the USA will be built this close together. The data analysis is exactly the same as for two detectors on the same site.

16.5 Reconstructing the signal

The inverse problem is the problem of how to reconstruct the gravitational wave from the observations made by a network of detectors. A single detector produces limited information about the wave; in particular, on its own it cannot give directional information and therefore it cannot say what the intrinsic amplitude is. With three detectors, however, one can reconstruct the wave entirely. In the last two or three years there has been considerable progress in understanding the inverse problem: see Boulanger, le Denmant and Tourrenc (1988), Dhurandhar and Tinto (1988), Gursel and Tinto (1989), and Tinto and Dhurandhar (1989). I will summarize the main ideas as I understand them at present but this is an area in which much more development is likely soon. My thinking in this section has been shaped by conversations with Massimo Tinto and Kip Thorne.

16.5.1 Single bursts seen in several detectors

(i) Unfiltered signals

A gravitational wave is described by two constants – the position angles of its source, (θ, ϕ) – and two functions of time – the amplitudes of the two independent polarizations $h_+(t)$ and $h_\times(t)$. Simple counting arguments give us an idea of how much we can learn from any given number of detectors. I will assume here that we do not have an *a priori* model (filter) for the signal. For signals that stand out above the broad-band noise:

- A single detector gives its response $r(t)$ and nothing else. Nothing exact can be said about the waves unless non-gravitational data can be used, as from optical or neutrino detections of the same event.
- Two detectors yield two responses and one approximate time-delay between the arrival of the wave in one detector and its arrival in the other. Two functions of time and one constant should not be enough to solve the problem, and indeed they are not. The time-delay is only an approximate one, because the two detectors will generally be responding to different linear combinations of $h_+(t)$ and $h_\times(t)$, so there will not be a perfect match between the responses of the two detectors, from which the time-delay must be inferred. The time-delay will confine the source to an error-band about a circle on the sky in a plane perpendicular to the line joining the detectors. The antenna patterns of the detectors can then be used to make some places on this circle more likely than others, but the unknown polarization of the wave will not allow great precision here. If the location of the source can be determined by other means, and if noise is not too large, then the two responses can determine the two amplitudes of the waves.
- Three detectors cross the threshold into precision astronomy, at least when the signals stand out against the broad-band noise. Here we have three functions of time (the responses) and two constants (the time-delays) as data, and this should suffice. As described in section 16.4 above, correlations among the three detectors can pin down the location of the source and, if noise is not too important, the time-dependent amplitudes as well. In this case, there is redundant information in the data that effectively test Einstein's predictions about the polarization of gravitational waves: the waveforms constructed from any pair of detectors should agree with those from the other two pairs to within noise fluctuations.

(ii) Filtered signals

If noise is so important that filtering is necessary, there is a completely different way of doing the counting. A given filter yields only constants as outputs, such as the maximum value of the correlation and the time the signal arrives (i.e. when it best matches the filter). It does not give useful functions of time. We can only assume that the signal's waveform matches the 'best' filter, so instead of two unknown time-dependent amplitudes we will have the response of the filter, the time-of-arrival, and a certain number of parameter constants that distinguish the observed waveform from others in its family.

Let us concentrate on coalescing binaries. The signal from a coalescing binary is an elliptically polarized, roughly sinusoidal waveform. The filters form a two-parameter family, characterized by the mass parameter $\mathcal{M}$ and the phase of the signal Φ, as in equation (16.8). The parameters we want to deduce are: the amplitude h of the signal, the ellipticity e of its polarization ellipse (one minus the ratio of the minor and major axes), an orientation angle ψ of the ellipse on the sky, and the binary's mass parameter $\mathcal{M}$. From these data we can not only determine the distance to the system, but also the inclination angle of the binary orbit to the line of sight (from e) and the orientation of the orbital plane on the sky (ψ).

The mass parameter $\mathcal{M}$ will be determined independently in each detector, and of course they will all agree if the event is real. Each detector in addition contributes the response of the filter, the phase parameter, and the time-of-arrival; these data must be used to deduce the five constants $\{\theta, \phi, h, e, \psi\}$. Here is how various numbers of detectors can use their data*:

- One detector does not have enough data, so it can only make average statements about the amplitude.
- Two detectors provide four useful data: two responses, one phase difference, and one time-delay. (Only the *differences* between the phases and times-of-arrival matter: the phase and time-of-arrival at the first detector are functions of the history of the source.) If the two detectors were identically polarized, the phase difference would necessarily be zero. A non-zero phase difference arises because the two principal polarizations in an elliptically polarized wave are 90° out of phase, so if the detectors respond to different combinations of these two polarizations, they will have different phases. With four data chasing five unknowns, the solution will presumably be a one-dimensional curve on the sky, but the problem has not yet been studied from this perspective.
- Three detectors have seven data: three responses, two phase differences and two time-delays. The two time-delays are sufficient to place the source at either of the intersections of two circles on the sky. For either location, the three responses determine h, e, and ψ. Presumably the phase differences would be consistent only with one of these positions, thereby solving the problem uniquely and incidentally providing the phase differences as a test of general relativity's model for the polarization of gravitational waves.

16.6 Data storage and exchange

Although the amount of data generated by a four-detector network will be huge, I would argue strongly that our present ignorance of gravitational wave sources

* This discussion is very different from previous ones I have given, e.g. Schutz (1989). In these I had not yet appreciated the importance of being able to determine the phase parameter independently of the time-of-arrival. This extra information makes it possible to solve the inverse problem with fewer detectors than I had previously believed.

makes it important that the data should be archived in a form that is relatively unprocessed, and kept for as long a time as possible, certainly for several years. It may be that new and unexpected sources of gravitational waves will be found, which will make it desirable to go over old data and re-filter it. It may also be that new classes of events will be discovered by their electromagnetic radiation, possibly with some considerable delay after the event would have produced gravitational waves, and a retrospective search would be desirable. In any case, we have already seen that it will be important to exchange essentially raw data between sites for cross-correlation searches for unknown events. Once exchanged, it is presumably already in a form in which it can be stored.

16.6.1 Storage requirements

We have seen in the introduction that a network could generate 5000 optical discs or videotapes per year. Data compression techniques and especially the discarding of most of the housekeeping data at times when it merely indicated that the detector was working satisfactorily could reduce this substantially, perhaps by as much as a factor of four. The cost of the storage media is not necessarily trivial. While videotapes are inexpensive, optical discs of large capacity could cost $250k at present prices (which will, hopefully, come down). Added to this is the cost of providing a suitable storage room, personnel to supervise the store, and equipment to make access to the data easy.

16.6.2 Exchanges of data among sites

We have already seen how important it will be to cross-correlate the raw data streams. At a data rate of some 100 kbytes per second, or even at 30 kbytes per second if the data volume is reduced as described above, one would have difficulty using standard international data networks. But these networks are being constantly upgraded, and so in five years the situation may be considerably different: it may be possible, at reasonable cost, to exchange short high-bandwidth bursts of data regularly via optical-fibre-to-satellite-to-optical-fibre routes. Alternatively, a cheaper solution might be to exchange optical discs or videotapes physically, accepting the inevitable delay. If lists of filtered events were exchanged on electronic data networks, then there may be less urgency about exchanging the full data sets.

(i) Protocols, analysis and archiving

It will be clear from our discussion that exchanging and jointly analysing data will require careful planning and coordination among all the groups. Discussions to this end are in a rudimentary stage now, but could soon be formalized more. Besides decisions on compatible hardware, software, data formats and modes of exchange, there are a number of 'political' questions that need to be resolved before observations begin. We are dealing with data that the groups involved have spent literally decades of their scientific careers to be in a position to obtain, and the scientific importance of actual observations of gravitational waves will be momentous. Questions of fairness and proprietary rights to the data could be a source of considerable friction if they are not clearly decided ahead of time. A model for some of these decisions could be the protocols adopted by the GRAVNET network of bar antennas, described elsewhere in this volume. Other models might be international VLBI, or large particle-physics collaborations.

Some of the questions that need to be addressed are:

- how much data needs to be exchanged;
- what groups have the right to see and analyse the data of other groups and what form of acknowledgement they need to give when they use it;
- what powers of veto groups have over the use of their data, for example in publications by other people;
- how long the proprietary veto would last before the data become 'public domain' (the funding agencies will presumably apply pressure to allow ready access to the data by other scientists after some reasonable interval of time);
- how long the data need to be archived.

Given the volume of data and the logistical complications of multi-way exchanges of it, it may be attractive to establish one or more joint data analysis and archiving centres. These could be particularly attractive as sites for any large computers dedicated to the pulsar-search problem. These would collect the data and store it, and perform the cross-correlations that can only be done with the full data sets on hand.

16.7 Conclusions

In this review I have set out what I understand about the data analysis problem as of September, 1989. Evidently, the field is covered very non-uniformly: coalescing binaries have received much more attention than pulsars or stochastic sources so far, and protocols for data exchange are something mainly for the future.

Nevertheless, it is clear that questions of the type we have discussed here will influence in an important way decisions about the detectors: how many there will be, where they will be located, what their orientations will be, what weights one should apply to the various important parameters affecting their sensitivity (e.g., length, seismic isolation, laser power) when deciding how to apportion limited budgets to attain the maximum sensitivity. Other questions that I have not addressed will also be important, particularly choosing the particular recycling configuration most suitable to searching for a given class of sources.

From the present perspective, it seems very likely that in ten years or so a number of large-scale interferometric detectors will be operating with a broad-band sensitivity approaching 10^{-22}. The data should contain plenty of coalescing binaries and at least a few supernovae; but the most exciting thing that we can look forward to is the unexpected: will this sensitivity suffice to discover completely unanticipated sources? The best way to ensure that it does is to make sure that our data-analysis algorithms and data-exchange protocols are adequate to the task: given the enormous efforts being made by the hardware groups to develop the detectors, and the considerable amount of money that will be required to build them, it is important that development of the data-analysis tools not be left too late. Solutions to data-analysis problems must be developed in parallel with detector technology.

Acknowledgements

It is a pleasure to acknowledge the helpfulness of many colleagues, from whom I have learned much about data analysis and many of whom have contributed crucial insights into the problem. They include Mark Davis, Ron Drever, Jim Hough, Andrzej Krolak, Sean Lawrence, Alberto Lobo, Norman MacKenzie, Brian Meers, Eduardo Nahmad-Achar, Tim Niebauer, Norna Robertson, Roland Schilling, Kip Thorne, Massimo Tinto, Philippe Tourrenc, Harry Ward, Rai Weiss, and Sheryl Smith Woodward. I am grateful to Drs Hough, Robertson, Tinto, and Ware for correcting errors in earlier drafts of this chapter. Any remaining errors are, of course, my own.

References

Boulanger, J. L., le Denmant, G. and Tourrenc, Ph. (1988). *Phys. Lett. A* **126,** 213–18.

Davis, M. H. A. (1989). In *Gravitational Wave Data Analysis,* ed. B. F. Schutz, pp. 73–94, Kluwer, Dordrecht.

Dewey, D. (1986). In *Proceedings of the Fourth Marcel Grossmann Meeting on General Relativity,* ed. R. Ruffini, p. 581, Elsevier, Amsterdam.

Dhurandhar, S. V. and Tinto, M. (1988). *Mon. Not. R. Astr. Soc.* **234,** 663–76.

Dhurandhar, S. V., Schutz, B. F. and Watkins, W. J. (1991). In preparation.

Evans, C. R. (1986). In *Dynamical Spacetimes and Numerical Relativity,* ed. J. M. Centrella, pp. 3–39, Cambridge University Press.

Evans, C. R., Iben, Jr., I. and Smarr, L. (1987). *Astrophys. J.* **323,** 129–39.

Gursel, Y. and Tinto, M. (1989). *Phys. Rev. D* **40,** 3884–938.

Hocking, W. K. (1989). *Computers in Physics,* Jan/Feb 1989 issue, pp. 59–65.

Horowitz, P. (1969). *Rev. Sci. Instrum.* **40,** 369–70.

Hough, J., Drever, R. W. P., Ward, F., Munley, A. J., Newton, G. P., Meers, B. J., Hoggan, S. and Kerr, G. A. (1983). *Nature* **303,** 216.

Ipser, J. R. and Managan, R. A. (1984). *Astrophys. J.* **282,** 287.

Krolak, A. (1989). In *Gravitational Wave Data Analysis,* ed. B. F. Schutz, pp. 59–69, Kluwer, Dordrecht.

Krolak, A. and Schutz, B. F. (1987). *Gen. Rel. Gravit.* **19,** 1163–71.

Livas, J. C. (1987). Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, Mass.

Lyne, A. (1989). In *Gravitational Wave Data Analysis*, ed. B. F. Schutz, pp. 95–103, Kluwer, Dordrecht.

Michelson, P. F. (1987). *Mon. Not. R. Astr. Soc.* **227,** 933–41.

Pasetti, A. (1987). M.Sc. thesis, Imperial College, London.

Piran, T. and Stark, R. F. (1986). In *Dynamical Spacetimes and Numerical Relativity*, ed. J. M. Centrella, pp. 40–73, Cambridge University Press.

Schutz, B. F. (1986). In *Gravitational Collapse and Relativity*, H. Sato and T. Nakamura, pp. 350–68, World Scientic, Singapore.

Schutz, B. F. (1989). In *Gravitational Wave Data Analysis*, ed. B. F. Schutz, pp. 315–26, Kluwer, Dordrecht.

Smith, S. (1987). *Phys. Rev. D* **36,** 2901–4.

Srinath, M. D. and Rajasekaran, P. K. (1979). *An Introduction to Statistical Signal Processing with Applications,* Wiley, New York.

Thorne, K. S. (1987). In *300 Years of Gravitation,* eds. S. W. Hawking and W. Israel, pp. 330–458, Cambridge University Press.

Tinto, M. and Dhurandhar, S. V. (1989). *Mon. Not. R. Astr. Soc.* **236,** 621–7.

KEF F

The Last Three Minutes: Issues in Gravitational-Wave Measurements of Coalescing Compact Binaries

Curt Cutler,(1) Theocharis A. Apostolatos,(1) Lars Bildsten,(1) Lee Samuel Finn,(2) Eanna E. Flanagan,(1) Daniel Kennefick,(1) Dragoljub M. Markovic,(1) Amos Ori,(1) Eric Poisson,(1) Gerald Jay Sussman,(1),(a) and Kip S. Thorne(1)

(1) *Theoretical Astrophysics, California Institute of Technology, Pasadena, California 91125*

(2) *Department of Physics and Astronomy, Northwestern University, Evanston, Illinois 60208*

(Received 24 August 1992)

Gravitational-wave interferometers are expected to monitor the last three minutes of inspiral and final coalescence of neutron star and black hole binaries at distances approaching cosmological, where the event rate may be many per year. Because the binary's accumulated orbital phase can be measured to a fractional accuracy $\ll 10^{-3}$ and relativistic effects are large, the wave forms will be far more complex and carry more information than has been expected. Improved wave form modeling is needed as a foundation for extracting the waves' information, but is not necessary for wave detection.

PACS numbers: 04.30.+x, 04.80.+z, 97.60.Jd, 97.60.Lf

A network of gravitational-wave interferometers (the American LIGO [1], the French/Italian VIRGO [2], and possibly others) is expected to be operating by the end of the 1990s. The most promising waves for this network are from the inspiral and coalescence of neutron star (NS) and black hole (BH) binaries [3, 4], with an estimated event rate of $\sim$ (3/year)[distance/(200 Mpc)]^3 [5]. This Letter reports initial results of a new research effort that is changing our understanding of these waves; further details will be given in the authors' papers cited in the references.

A binary's inspiral and coalescence will produce two gravitational wave forms, one for each polarization. By cross correlating the outputs of three or more interferometers, both wave forms can be monitored and the source's direction can be determined to within $\sim$ 1 degree [4, 6].

We shall divide each wave form into two parts: the *inspiral wave form*, emitted before tidal distortions become noticeable ($\lesssim$ 10 cycles before complete disruption or merger [7, 8]), and the *coalescence wave form*, emitted during distortion, disruption, and/or merger.

As the binary, driven by gravitational radiation reaction, spirals together, its *inspiral wave form* sweeps upward in frequency f (it "chirps"). The interferometers will observe the last several thousand cycles of inspiral (from $f \sim 10$ Hz to ~ 1000 Hz), followed by coalescence.

Theoretical calculations of the wave forms are generally made using the post-Newtonian (PN) approximation to general relativity. Previous calculations have focused on the Newtonian-order wave forms [1, 3, 4, 9] and on PN modulations of their amplitude and frequency [10].

We have recently realized that the PN modulations are far less important than PN contributions to the secular growth of the waves' phase $\Phi = 2\pi \int f dt$, which arise largely from PN corrections to radiation reaction [11, 12]. The binary's parameters are determined by integrating the observed (noisy) signal against theoretical templates, and if the signal and template lose phase with each other by as little as one half cycle over the thousands observed as the signal sweeps through the interferometers' band, their overlap integral will be strongly reduced. This sensitivity to phase does *not* mean that accurate templates are needed in searches for the waves (see below). However, once the waves have been found, and if accurate templates are in hand, then from the orbital phasing one can infer each of the system's parameters λ_i to an accuracy of order the change $\Delta\lambda_i$ which alters by unity the number of cycles $\mathcal{N}_{\rm cyc}$ spent in the interferometers' band.

We shall assume (as almost always is the case) that the binary's orbit has been circularized by radiation reaction [10]. Then the only parameters λ_i that can significantly influence the inspiral template's phasing are the bodies' masses, vectorial spin angular momenta, and spin-induced quadrupole moments (which we shall ignore because, even for huge spins, they produce orbital phase shifts no larger than ~ 1 [8]). More specifically, the number of cycles spent in a logarithmic interval of frequency, $d\mathcal{N}_{\rm cyc}/d\ln f = (1/2\pi)(d\Phi/d\ln f)$, is

$$\frac{d\mathcal{N}_{\rm cyc}}{d\ln f} = \frac{5}{96\pi}\frac{1}{\mu M^{2/3}(\pi f)^{5/3}}\left\{1 + \left(\frac{743}{336} + \frac{11}{4}\frac{\mu}{M}\right)x - [4\pi + \text{S.O.}]x^{1.5} + [\text{S.S.}]x^2 + O(x^{2.5})\right\}. \quad (1)$$

Here M is the binary's total mass, μ its reduced mass, and $x \equiv (\pi M f)^{2/3} \simeq M/D$ the PN expansion parameter (with D the bodies' separation and $c = G = 1$). The PN correction [$O(x)$ term] is from [13]. In the P$^{1.5}$N correction [$O(x^{1.5})$ term], the 4π is created by the waves' interaction with the binary's monopolar gravitational field as they propagate from the near zone to the radiation zone [14], and the "S.O." denotes contributions due to spin-orbit coupling [15]. In the P^{2}N correction the "S.S." includes spin-spin coupling effects [15] plus an expression quadratic in μ/M. (For bodies with sizes comparable to their separations, the spin-orbit and spin-spin terms are of PN order; but the compactness of a BH or NS boosts them up to P$^{1.5}$N and P^{2}N, respectively; cf. [15].)

Since the leading-order, Newtonian contribution to

© 1993 The American Physical Society

Eq. (1) gives several thousand cycles in the interferometers' band, to be measurable the higher-order corrections need only be as large as a part in several thousand, even when the signal is only a little above the threshold for detection. It is this that gives measurements based on the waves' phasing their high potential accuracy.

The determination of the binary's masses and spins is made possible by the various frequency dependences in Eq. (1). Analytic [11, 12] and Monte Carlo [11] calculations show that (i) the *chirp mass* $\mathcal{M}_c \equiv \mu^{3/5}M^{2/5}$ [which governs the Newtonian part of (1)] will typically be measured to a *few tenths of a percent*, and (ii) *if* we somehow knew that the spins were small, then the reduced mass μ would be measured to $\sim 1\%$ for NS-NS and NS-BH binaries, and $\sim 3\%$ for BH-BH binaries. (Here and below NS means a $\sim 1.4M_\odot$ neutron star and BH means a $\sim 10M_\odot$ black hole.) These accuracies assume a noise spectrum whose shape is that of "advanced detectors" [1], and the use of two detectors with amplitude signal-to-noise ratios of 9 in each. Unfortunately, the frequency (x) dependences of the various terms in Eq. (1) are not sufficiently different to give a clean separation between μ and the spins. Preliminary estimates [11] in which the S.O. term in Eq. (1) was taken into account but not the S.S. term, suggest that the spin/μ correlation will worsen the typical accuracy of μ by a factor ~ 15, to $\sim 10\%$ for NS-NS, $\sim 15\%$ for NS-BH, and $\sim 70\%$ for BH-BH. These worsened accuracies might be improved significantly by wave form modulations caused by the spin-induced precession of the orbit [16] (see Fig. 1), and even without modulational information, a certain combination of μ and the spins will be determined to a few percent. Much additional theoretical work is needed to firm up the measurement accuracies.

Although the early, 10 Hz–50 Hz part of the wave form produces only $\sim 10\%$ of the signal-to-noise squared, it accounts for $> 80\%$ of the measurable cycles and most of the mass-measurement accuracy. It thus is important that the interferometers achieve good low-frequency performance.

From the above accuracies, we estimate that, when *searching* for inspiral signals in noisy data, one will need $\sim 10^5$ templates to cover the possible values of masses and spins. Since two otherwise identical wave forms displaced by 3 ms will have their overlap integral reduced substantially, $\sim 10^{15}$ templates per year must be integrated against the data. The noise is expected to be Gaussian (thanks to filtering with the templates and cross correlation of detectors), and thus the probability that any particular template will produce a spurious signal-to-noise ratio $S/N > \rho$ in each of two detectors is $[\mathrm{erfc}(\rho/\sqrt{2})]^2$. For $\rho = (5, 6, 7)$, $[\mathrm{erfc}(\rho/\sqrt{2})]^2 \approx (10^{-13}, 10^{-18}, 10^{-23})$. Thus an appropriate threshold for detection is $S/N \geq 6.0$. Notice that this threshold is quite insensitive to our estimate of $\sim 10^{15}$ templates/year, and it was increased by only 10% (from 5.5 to 6.0) by our discovery of the rich influence of μ and spins on the orbital phase. For "advanced-detector sensitivities" [1], NS-NS inspirals will exceed this threshold at distances $\lesssim 1$ Gpc, and BH-BH inspirals at $\lesssim 5$ Gpc.

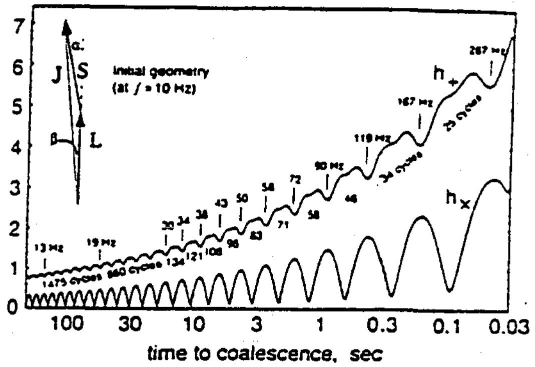


FIG. 1. This figure illustrates the importance of spin-induced orbital precession (the "dragging of inertial frames"), which has been ignored in all previous discussions of gravitational waves from inspiraling binaries. Here a $1M_\odot$ nonspinning NS spirals into a $10M_\odot$, rapidly spinning Kerr black hole (spin parameter $a = 1$). The orbital angular momentum $\mathbf{L}$ is inclined by $\alpha = 11.3$ degrees to the hole's spin angular momentum $\mathbf{S}$, and the two precess around $\mathbf{J} = \mathbf{L} + \mathbf{S}$, whose direction remains fixed in space as $L = |\mathbf{L}|$ shrinks and $S = |\mathbf{S}| = M_{\rm BH}^2 a$ remains constant $(L/S = (\mu/Ma)(\pi M f)^{-1/3} \simeq [(5\ \mathrm{Hz})/f]^{1/3} \lesssim 1)$. The precession of $\mathbf{L}$ around $\mathbf{J}$, with precession angle $\beta \simeq \alpha/(1 + L/S) \simeq 0.11$ to 0.17, modulates the waves by $\delta h \sim (0 \text{ to } 4)\beta \times [\max(h_+, h_\times)]$, where the coefficient depends on (i) the direction to Earth (here out of the paper) and (ii) the orientation of the detector's arms (here horizontal and vertical for h_+, and rotated 45 degrees for $h_\times$). The figure shows the wave forms' modulational envelopes (in arbitrary units, the same for h_+ and $h_\times$), parametrized by the wave frequency f and the number of cycles of *oscillation* between the indicated f's. The total number of *precessions* from f to coalescence is $N_{\rm prec} \simeq (5/64\pi)(Ma/\mu)(\pi M f)^{-2/3} \simeq 20[f/(10\ \mathrm{Hz})]^{-2/3}$. For further details see Ref. [16].

We suspect that, when *searching* for inspiral signals, it will be sufficient to use templates accurate to $\mathrm{P}^{1.5}\mathrm{N}$ order; the omitted, higher-order corrections will be compensated by incorrect estimates of the masses and spins, and the signals will still be found. Better yet will be a set of templates that span the range of expected wave form behaviors in such a way as to minimize the computational time. Such templates need to be developed.

Although highly accurate wave form templates will *not* be needed when searching for waves, they *will* be needed when extracting the waves' information. Making optimal use of the interferometers' data will require general-relativity-based wave form templates whose phasing is correct to within a half cycle or so during the entire frequency sweep from ~ 10 Hz to ~ 1000 Hz. We estimate that $\mathrm{P}^{1.5}\mathrm{N}$-order templates (the most accurate computed to date) will produce systematic errors in μ of order 10%.

By examining an idealized limit ($\mu \ll M$ and vanishing spins), we have found that the PN expansion converges very slowly [14, 17]. We calculated the waves from such a binary to high accuracy using the Teukolsky-Regge-Wheeler [18] black-hole perturbation formalism and then fitting a PN expansion to the results, to obtain [17]

$$\frac{dN_{\rm cyc}}{d\ln f}=\frac{f^2dE/df}{dE/dt}=\frac{5}{96\pi}\frac{M/\mu}{x^{2.5}}\times\frac{1-\frac{3}{2}x-\frac{81}{8}x^2-\frac{675}{16}x^3+\cdots}{1-\frac{1247}{336}x+4\pi x^{1.5}-4.9x^2-38x^{2.5}+135x^3+\cdots}. \tag{2}$$

Here dE/df is half the ratio of the changes of orbital energy dE and orbital frequency $d(f/2)$ as the orbit shrinks, and dE/dt is the power carried off by gravitational waves. The coefficients in the denominator after the 4π were obtained from the fit; all other coefficients are known analytically. The coefficient 4.9 has accuracy ~ 2%; the 38, ~ 10%; and the 135, ~ 50%.

We can use Eq. (3) to estimate how the phase accuracy depends on the order to which the PN expansion is carried. Compare a template constructed using Eq. (3) with a P^{2}N-order template [all $x^{2.5}$ and x^3 terms in the denominator of Eq. (3) omitted]. Adjust the constants of integration so the templates are perfectly in phase at 70 Hz, where the detector has maximum sensitivity [1]. Then the two templates fall out of phase as f decreases below 70 Hz and increases above 70 Hz; for a NS-BH binary the phase difference has grown to half a cycle at 40 Hz and 140 Hz, and to three cycles at 10 Hz and 400 Hz. If the template were computed through P$^{2.5}$N order, there would be no substantial improvement. It is not at all clear how far beyond P$^{2.5}$N the template must be carried to keep its total phase error below a half cycle over the entire range from ~ 10 Hz to ~ 400 Hz, but the slowness of the convergence suggests that new techniques are needed for computing templates to higher accuracies.

One possible new technique is a "weak-reaction expansion." This would be a variant of numerical relativity in which one expands in powers of $1/Q\equiv(dE/dt)(\pi fE)^{-1}$ (a measure of the effects of radiation reaction during one orbit). Because $1/Q\sim x^{2.5}$ (with x the PN parameter) is always $\ll 1$, this expansion might converge quickly. The expansion might begin with a Blackburn-Detweiler [19] type numerical solution of the Einstein equations with standing-wave boundary conditions. From the standing waves, one can read off the gravitational radiation that results when one switches to outgoing waves, and thereby infer the wave forms and inspiral rate to first order in $1/Q$. It is an important challenge to devise a way to iterate to second order.

We turn now from a binary's inspiral wave forms to its *coalescence wave forms*. By the beginning of coalescence, the binary's parameters (masses, spins, geometry) will be known with fair accuracy, and from its masses the nature of its bodies (BH versus NS) should be fairly clear. The coalescence wave forms can then be used to probe issues in the physics of gravity and atomic nuclei. For BH-BH binaries, the issue to be probed (large-amplitude, highly nonlinear vibrations of spacetime curvature) ha been much discussed; see, e.g., [1] and references therein Not previously discussed, however, is the possibility o using NS-BH and NS-NS coalescence to measure the NS mass-radius relation (from which one can infer the equation of state of nuclear matter [20]):

For a NS-BH binary with the BH spinning moderately fast, the NS should disrupt tidally before plunging into the BH. The NS disruption should be quick, as should the final coalescence of a NS-NS binary: within several orbits the binary should get smeared into a roughly axially symmetric configuration, thereby shutting off its waves [7, 8, 21]. The waves' inspiral frequency at the beginning of tidal disruption will be $f_{\rm td}=(\alpha/2\sqrt{2}\pi)\sqrt{M_{\rm NS}/R_{\rm NS}^3}\simeq(1.5\ {\rm kHz})\sqrt{(M_{\rm NS}/1.4M_\odot)[(10\ {\rm km})/R_{\rm NS}]^3}$, where $M_{\rm NS}$ and $R_{\rm NS}$ are the mass and radius of the less massive (or only) neutron star, and α is a factor close to 1 that will depend only weakly on the mass and spin of the companion. Since $M_{\rm NS}$ will be known from the earlier inspiral waves, a measurement of $f_{\rm td}$ will reveal $R_{\rm NS}$.

Shot noise may prevent the LIGO and VIRGO "workhorse," broad-band interferometers from measuring $f_{\rm td}$; but special, narrow-band, "dual recycled" interferometers [22] may do the job. Such an interferometer could give a simple "yes or no" answer as to whether the waves swept through its frequency band $f_0\pm\Delta f$ (and if yes, the spectral density of wave energy in that band), with an amplitude signal-to-noise ratio [23]

$$\frac{S}{N}\simeq 8\frac{[\eta I_0/(100\ {\rm W})]^{1/2}}{(A^2/10^{-5})^{1/2}}\frac{(\mu/1M_\odot)^{1/2}(M/10M_\odot)^{1/3}}{[r/(200\ {\rm Mpc})][f_0/(1\ {\rm kHz})]^{7/6}}. \tag{3}$$

Here η is the photodiode efficiency, I_0 the laser power, and A^2 the fractional light power lost to absorption and scattering in each bounce off a mirror. By collecting such data on a number of binaries, with various choices for the interferometer's frequency f_0, it should be possible to zero in on $f_{\rm td}$ for various types of binaries, and thence on the NS radius-mass relation.

As a foundation for these measurements, theorists should model tidal disruption in NS-BH binaries and the coalescence of NS-NS binaries, to determine the dependence of $f_{\rm td}$ on $R_{\rm NS}$ and the binary's other parameters, and the shape of the waves' spectrum above $f_{\rm td}$. (Such modeling is also driven by the possibility that these coalescences produce observed gamma-ray bursts [24].)

We turn next from a binary's coalescence to the use of its wave forms as cosmological probes, via a variant of a method conceived by Schutz [4]: For a binary at a redshift $z\gtrsim 1$, the inspiral wave forms will be measured as functions of *redshifted* time, and correspondingly they will depend on and reveal the "redshifted masses" $(1+z)M_1$ and $(1+z)M_2$ and the binary's "luminosity distance" r_L. From these, one can infer whether the binary contains a NS. Assuming (as observations suggest) that NS masses in binaries cluster strongly around $1.4M_\odot$, one can estimate $1+z$ as the observed $M_{\rm NS}$ divided by $1.4M_\odot$. From a large number of such measured $(1+z)$'s and r_L's, one can deduce the Universe's redshift-distance relation and thence its Hubble constant $H_0 = h_0 \times 100$ km s^{-1} Mpc^{-1}, density parameter $\Omega_0 = (8\pi/3)(\rho_0/H_0^2)$, and cosmological constant $\Lambda = 3H_0^2\lambda_0$.

We have carried out simulations of such a determination of cosmological parameters [25], assuming (i) the LIGO-VIRGO network with three identical interferometers whose noise spectra have the "advanced detector" shape of Ref. [1], (ii) distance errors $\Delta r_L/r_L \simeq (2/\rho)\times$(a function, ≥ 1, of the binary direction and orientation), and (iii) mass errors $\Delta M_{\rm NS}/M_{\rm NS} \simeq 1.4/\rho$; here ρ is the amplitude signal-to-noise ratio. By contrast with electromagnetic cosmological measurements, which suffer from light absorption and source evolution, this method will suffer just one type of propagation noise (gravitational lensing by mass inhomogeneities), and probably no evolutionary effects (it seems unlikely that the NS mass spectrum will depend on cosmological epoch).

Our simulations [25] show that gravitational lens noise should be negligible compared to detector noise, and they suggest the following one-sigma errors in the cosmological parameters inferred from NS-BH binary measurements. (i) $\Delta h_0 \simeq 0.01N(\hat{\tau}\mathcal{R})^{-1/2}$ for $N \lesssim 3$, where N is the detectors' noise (strain/$\sqrt{\rm Hz}$) in units of the "advanced detector" level shown in Ref. [1], $\mathcal{R}$ is the event rate in units of the best estimate, 100 yr^{-1} Gpc^{-3}, and $\hat{\tau}$ is the observation time in years. (ii) In a "compact" universe ($h_0 = 1$, $\Omega_0 = 1$, $\lambda_0 = 0$), $\Delta\Omega_0 \simeq 0.1N^2(\hat{\tau}\mathcal{R})^{-1/2}$ and $\Delta\lambda_0 \simeq 0.2N^{1.5}(\hat{\tau}\mathcal{R})^{-1/2}$ for $N \lesssim 1$. (iii) In a "spacious" universe ($h_0 = 0.5$, $\Omega_0 = 0.2$, $\lambda_0 = 0$), $\Delta\Omega_0 \simeq N^3(\hat{\tau}\mathcal{R})^{-1/2}$ and $\Delta\lambda_0 \simeq N^{2.5}(\hat{\tau}\mathcal{R})^{-1/2}$ for $N \lesssim 0.75$. For measurements based on NS-NS binaries rather than NS-BH, the fact that one can use the highly accurate chirp mass to deduce the redshift instead of the much less accurate $M_{\rm NS}$ does not significantly compensate for the weaker NS-NS signal strength; consequently, the NS-NS–based errors are approximately as quoted above for NS-BH, but with N replaced by $2N$.

These accuracies suggest that, if event rates are as currently estimated, interesting cosmological measurements, beyond determining h_0, cannot begin until the detector sensitivities reach the "advanced detector level."

Some of the issues discussed above have implications for a possible future space-based interferometer LAGOS, which would operate in the band 0.0001–0.03 Hz [26].

For helpful discussions we thank D. Chernoff, L. Kidder, A. Krolak, E. S. Phinney, B. F. Schutz, C. M. Will, and A. Wiseman. This research was supported in part by NSF Grant No. PHY-9213508, and, in view of its applications to LAGOS, by NASA Grants No. NAGW-2897, No. 2920, and No. 2936, and by a Lee A. Dubridge Fellowship to L. Bildsten, an Alfred P. Sloan Foundation Fellowship to L. S. Finn, and a NSERC Fellowship to E. Poisson.

[(a)] Permanent address: Dept. of Electrical Engineering, MIT, Cambridge, MA 02139.

[1] A. Abramovici *et al.*, Science 256, 325 (1992).

[2] C. Bradaschia *et. al.*, Nucl. Instrum. Methods Phys. Res., Sect. A 289, 518 (1990).

[3] K. S. Thorne, in *300 Years of Gravitation*, edited by S. W. Hawking and W. Israel (Cambridge Univ. Press, Cambridge, 1987), p. 330.

[4] B. F. Schutz, Nature (London) 323, 310 (1986); B. F. Schutz, Classical Quantum Gravity 6, 1761 (1989).

[5] R. Narayan, T. Piran, and A. Shemi, Astrophys. J. 379, L17 (1991); E. S. Phinney Astrophys. J. 380, L17 (1991).

[6] Y. Gursel and M. Tinto, Phys. Rev. D 40, 3884 (1990); P. Jaranowski and A. Krolak (unpublished).

[7] C. Kochanek, Astrophys. J. 398, 234 (1992).

[8] L. Bildsten and C. Cutler, Astrophys. J. 400, 175 (1992).

[9] H. D. Wahlquist, Gen. Relativ. Gravitation 19, 1101 (1987).

[10] C. W. Lincoln and C. M. Will, Phys. Rev. D 42, 1123 (1990).

[11] C. Cutler and E. E. Flanagan (unpublished).

[12] L. S. Finn and D. Chernoff, Phys. Rev. D 47, 2198 (1993).

[13] R. V. Wagoner and C. M. Will, Astrophys. J. 210, 764 (1976); 215, 984 (1977).

[14] E. Poisson, Phys. Rev. D 47, 1497 (1993).

[15] L. E. Kidder, C. M. Will, and A. G. Wiseman, Phys. Rev. D (to be published).

[16] G. J. Sussman, T. A. Apostolatos, C. Cutler, and K. S. Thorne (unpublished).

[17] C. Cutler, L. S. Finn, E. Poisson, and G. J. Sussman, Phys. Rev. D 47, 1511 (1993).

[18] S. A. Teukolsky, Astrophys. J. 185, 635 (1973); T. Regge and J. A. Wheeler, Phys. Rev. 108, 1063 (1957).

[19] J. K. Blackburn and S. L. Detweiler, Phys. Rev. D 46, 2318 (1992), but perhaps with their action principle replaced by a relaxation solution of the elliptic equations.

[20] L. Lindblom, Astrophys. J. 398, 569 (1992).

[21] F. A. Rasio and S. L. Shapiro, Astrophys. J. 401, 226 (1992).

[22] A. Krolak, J. A. Lobo, and B. J. Meers Phys. Rev. D 43, 2470 (1991).

[23] $S/N = (\pi\mu M^3/5r^2)^{1/2}(\pi M f_0)^{-7/6}\mathcal{H}_2[\int_0^\infty df/S_h(f)]^{1/2}$, from Ref. [3] Eqs. (30) [with 3 replaced by 6 due to an error in (29)], (44) [with 12 replaced by 6 due to ft being corrected to $\int f dt$ in (42)], and (110); $\mathcal{H}_2 \simeq 0.8$ is a general relativistic correction for the NS-BH case [L. S. Finn, A. Ori, and K. S. Thorne (unpublished)]. For $\int_0^\infty df/S_h(f)$ we use $(\pi\eta I_0/2\hbar)(l/\lambda)(1/A^2)$, where $l = 4$ km is the interferometer arm length and $\lambda = 0.4$ μm is the laser light wavelength [Eq. (3.7) of Ref. [22]] in the regime $cA^2/l \ll \Delta f \ll f_0$, with $\tau = 2l/c$ and corrections of a multiplicative factor $16/A^2$ and in the denominator $\pi\tau\Delta f \rightarrow 2\pi\tau\Delta f$; A. Krolak (private communication).

[24] B. Paczynski, Astrophys. J. Lett. 308, L43 (1986); J. Goodman, Astrophys. J. Lett. 308, L47 (1986); D. Eichler, M. Livio, T. Piran, and D. N. Schramm, Nature (London) 340, 126 (1989).

[25] D. M. Markovic (to be published).

[26] J. E. Faller, P. L. Bender, J. L. Hall, D. Hils, R. T. Stebbins, and M. A. Vincent, Adv. Space Res. (COSPAR) 9, 107 (1989).

BATCH START

Lecture 4 : IDealized theoRy

STAPLE
OR
DIVIDER

Ph 103c: The Physics of LIGO 8 April 1994

LECTURE 4.
IDEALIZED THEORY OF INTERFEROMETRIC DETECTORS — I.

Lecture by Kip S. Thorne

Assigned Reading:

A. "Gravitational Radiation" by Kip S. Thorne, in *300 Years of Gravitation*, eds. S. W. Hawking and W. Israel (Cambridge University Press, 1987), pages 414–425; ending at beginning of first full paragraph on 425. [This material uses the phrase *beam detector* for an *interferometric gravitational-wave detector*. The principal results quoted in this lecture are derived in the exercises below.]

G. The following portions of "Chapter 7. Diffraction" from the textbook manuscript *Applications of Classical Physics* by Roger Blandford and Kip Thorne: Section 7.2 (pages 7-2 to 7-7), and Section 7.5 (pages 7-20 to 7-27). [This material develops the foundations of the theory of diffraction (Green's theorem and the Helmholtz-Kirchoff formula), explores semi-quantitatively the spreading of a transversely collimated beam of light, develops the formalism of *paraxial Fourier optics* for analyzing quantitatively the propagation of collimated light beams, and uses that formalism to derive the evolution of the cross sectional shape of a Gaussian beam, of the sort used in LIGO.]

Suggested Supplementary Reading:

H. A. E. Siegman, *Lasers* (University Science Books, Mill Valley CA, 1986), chapter 17, "Physical Properties of Gaussian Beams." [This chapter develops in full detail the paraxial-Fourier-optics theory of the manipulation of Gaussian beams by a system of lenses and mirrors, and the shapes of the Gaussian modes of an optical resonator (Fabry-Perot cavity).]

A Few Suggested Problems

1. *Shot Noise.* Reread the discussion of shot noise on pages 5-20 and 5-21 of Blandford and Thorne, *Random Processes* (which was passed out last week). In that discussion let the random process $y(t)$ be the intensity $I(t) = d(\text{energy})/dt$ of a laser beam, and let $F(t)$ be the intensity carried by an individual photon, which has frequency ω.
   (a) Explain why $\tilde{F}(0)$, the Fourier transform of F at zero frequency, is the photon energy $\hbar\omega$.
   (b) Show that the spectral density of I (the "shot-noise spectrum") is

$$G_I(f) = 2\bar{I}\hbar\omega \; , \tag{1}$$

   where $\bar{I}$ is the beam's mean intensity.
   (b) Let $N(t)$ be the number of photons that the beam carries into a photodiode between time t and time $t + \hat{\tau}$ (so $\hat{\tau}$ is the averaging time): $N(t) = \int_t^{t+\hat{\tau}} I(t')dt'$.

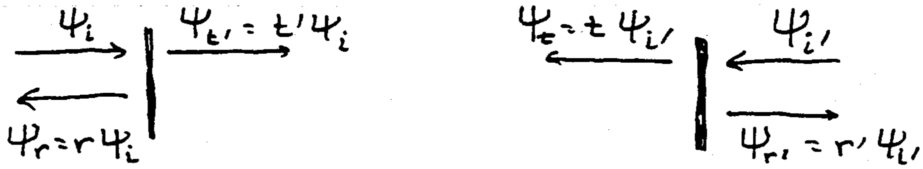
This $N(t)$ is a linear functional of $I(t')$. Use the theory of linear signal processing to derive the spectral density $G_N(f)$ of $N(t)$, and then compute the mean square fluctuations of N: $(\sigma_N)^2 = \int_0^\infty G_N(f)df$. Your result should be $\sigma_N = \sqrt{\bar{N}}$, where $\bar{N}$ is the mean number of photons that arrive in the averaging time $\hat{\tau}$. This is the standard "square-root-of-N" fluctuation in photon arrival for a laser beam.

2. *Reciprocity Relations for a Mirror and a Beam Splitter.* Modern mirrors, beam splitters, and other optical devices are generally made of glass or fused silica (quartz), with dielectric coatings on their surfaces. The coatings consist of alternating layers of materials with different dielectric constants, so the index of refraction n varies periodically. If, for example, the period of n's variations is half a wavelength of the radiation that impinges on the device, then waves reflected from successive dielectric layers build up coherently, producing a large net reflection coefficient. In this exercise we shall derive the reciprocity relations for a mirror of this type, with normally incident radiation. The generalization to radiation incident from other directions, and to other dielectric optical devices is straightforward.

The foundation for the analysis is the wave equation,

$$\left(-\frac{\partial^2}{\partial t^2} + \frac{c^2}{n^2(\mathbf{x})}\nabla^2\right)\psi = 0$$

satisfied by any Cartesian component ψ of the electric field, and the assumption that ψ is precisely monochromatic with angular frequency ω. These imply that the spatial dependence of ψ is governed by the Helmholtz equation with spatially variable wave number $k(\mathbf{x}) = n(\mathbf{x})\omega/c$: $\nabla^2\psi + k^2\psi = 0$.



Let waves $\psi_i e^{ikz}$ impinging perpendicularly (z direction) on the mirror from the "unprimed" side produce reflected and transmitted waves $\psi_r e^{ikz}$ and $\psi_{t'} e^{ikz}$; these waves and their corresponding ψ inside the mirror are one solution ψ_1 of the Helmholtz equation. The complex amplitudes of this solution are related by reflection and transmission coefficients, $\psi_r = r\psi_i$, $\psi_{t'} = t'\psi_i$. Another solution, ψ_2, consists of incident waves from the opposite, "primed" side, $\psi_{i'} e^{-ikz}$ and reflected and transmitted waves $\psi_{r'} e^{+ikz}$, $\psi_t e^{-ikz}$, and the corresponding ψ inside the mirror; and this solution's complex amplitudes are related by $\psi_{r'} = r'\psi_{i'}$, $\psi_t = t\psi_{i'}$.

(a) Show that ψ obeys Green's theorem [Equation (7.3) of Blandford and Thorne] throughout the mirror. Apply Green's theorem, with ψ and ψ_0 chosen to be various pairs of ψ_1, ψ_2, ψ_1^*, ψ_2^* (where the star denotes complex conjugation). Thereby obtain four relationships between r, r', t, and t'.

(b) Show that these relationships can be written in the form

$$r = \sqrt{\mathcal{R}}e^{2i\beta}, \quad r' = -\sqrt{\mathcal{R}}e^{2i\beta'}, \quad t = t' = \sqrt{\mathcal{T}}e^{i(\beta+\beta')} \ ,$$

where β and β' are unconstrained phases, and $\mathcal{R}$ and $\mathcal{T}$, the power reflection and transmission coefficients are related by

$$\mathcal{R} + \mathcal{T} = 1\,, \tag{2}$$

which is just energy conservation.

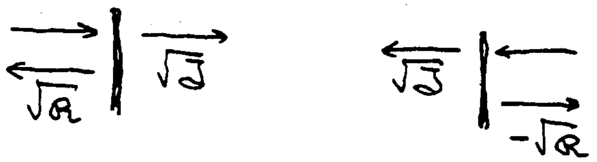
(c) Show that, if one moves the origin of coordinates as seen from the unprimed side by $\delta z = -k\beta$, and moves the origin as seen from the primed side by $\delta z = +k\beta'$, one thereby will make all the reflection and transmission coefficients real:

$$t = t' = \sqrt{\mathcal{T}}, \quad r = -r' = \sqrt{\mathcal{R}}\,. \tag{3}$$

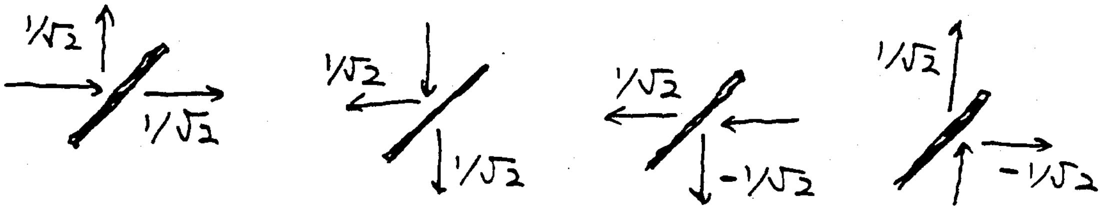
Thus, with an appropriate choice of origin on each side of the mirror, the coefficients can always be made real.

The same is true for the reflection and transmission coefficients of any other optical device made of a lossless, spatially variable dielectric. In particular, for a perfect, 50/50 beam splitter, the transmission coefficient becomes, with appropriate choice of origins, $1/\sqrt{2}$ from each and every one of the four input ports, and the reflection coefficient becomes $+1/\sqrt{2}$ from the input ports on one side of the beam splitter and $-1/\sqrt{2}$ from the input ports on the other side of the beam splitter. These results are summarized by the following figures:

mirror :



beam-splitter



2. *Transfer Function and Photon Shot Noise for a Delay-Line Interferometer* In class, Kip derived the "tranfer function" for a delay-line interferometer in the limiting regime where the waveform $h(t)$ is nearly constant during the time $2BL/c$ that the light is stored in the interferometer arms (during B round trips in an arm whose length is L). His result was

$$I_{PD}(t) = I_1(t) + 2\sqrt{\bar{I}_1 I_0}\,BkLh(t) \tag{4}$$

where I_0 is the mean laser input power entering the beamsplitter, $I_1(t)$ is the (slightly fluctuating because of shot noise) intensity of the light falling onto the photodiode in the absence of a gravitational-wave signal, $\bar{I}_1$ is the mean intensity onto the photodiode, B is the number of round trips in the arms of the interferometer, $k = \omega/c = 2\pi/\lambda_e$ is the light's wave number, L is the arm length, and $h(t)$ is the gravitational waveform. Kip used this and the shot-noise spectral density [Eq. (1) above] to derive the following expression for the shot-noise contribution to the interferometer's gravitational-wave noise output:

$$G_h(f) = \frac{\hbar\omega}{2I_0(BkL)^2}\,. \tag{5}$$

(a) Use the same method of analysis as Kip did in class to derive the transfer function when the gravitational wave is sinusoidal in time with angular frequency $\Omega = 2\pi f$, i.e. when $h(t) = h_o\cos(\Omega t) = h_o\mathrm{Real}(e^{-i\Omega t})$, with a frequency f high enough (gravitational wavelength short enough) that the waveform *can* vary significantly while the light is stored in the arms. Your result should be the same as Eq. (4), with B replaced by

$$B_{\text{eff}} = B\frac{\sin(f/f_0)}{f/f_0}\,, \quad f_0 \equiv \frac{c}{2\pi BL} = \frac{119\text{Hz}}{(B/100)(L/4\text{km})}\,. \tag{6}$$

(b) Show that the shot-noise contribution to $G_h(f)$ has the form (5) with B replaced by B_{eff}.

3. *Transfer Function and Photon Shot Noise for a Fabry-Perot Interferometer.* In class, Kip showed that for a Fabry-Perot interferometer in the regime of slow variations of $h(t)$ the transfer function and photon shot noise have the forms (4) and (5), with B replaced by

$$B_{\text{eff}} = \frac{4}{(1-\mathcal{R})} \tag{7}$$

where $\mathcal{R}$ is the power reflectivity of the interferometer's corner mirrors and where it is assumed that the end mirrors are perfectly reflecting. Show that, if the variations of $h(t)$ are not assumed to be slow, then the transfer function (for monochromatic gravitational waves) has the form (4) and the shot noise contribution to $G_h(f)$ has the form (5), with B replaced by

$$B_{\text{eff}} = \frac{B}{\sqrt{1+(f/f_0)^2}}\,, \tag{8}$$

where f_0 is as in Eq. (6) above.

LECTURE 5.
IDEALIZED THEORY OF INTERFEROMETRIC DETECTORS—II.

Lecture by Ronald W. P. Drever

Assigned Reading:

I. R. W. P. Drever, "Fabry-Perot cavity gravity-wave detectors" by R. W. P. Drever, in *The Detection of Gravitational Waves*, edited by D. G. Blair (Cambridge University Press, 1991), pages 306–317. [This is a qualitative overview of Fabry-Perot gravitational-wave detectors, with emphasis on recycling in the later part (pages 312—317).]

J. B. J. Meers, "Recycling in laser-interferometric gravitational-wave detectors," *Phys. Rev. D*, **38**, 2317–2326. [This is the paper in which Meers introduced his idea of dual recycling and sketched out its features. You are not expected to master all the equations in this paper—which Meers just gives without derivation—but you might try deriving some of the equations as a homework exercise.]

Suggested Supplementary Reading:

K. B. J. Meers, *Physics Letters A*, "The frequency response of interferometric gravitational wave detectors," *Physics Letters A*, **142**, 465 (1989). [In this paper Meers discusses in some detail the frequency responses and sensitivities of various configurations of recycled interferometers.]

L. B. J. Meers and R. W. P. Drever, "Doubly-resonant signal recycling for interferometric gravitational-wave detectors." (preprint) [This paper introduces a new recycling configuration, not considered in previous papers.]

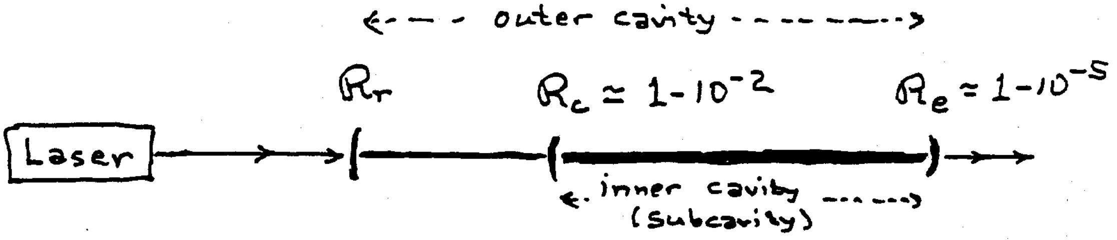
M. J. Mizuno, K. A. Strain, P. G. Nelson, J. M. Chen, R. Schilling, A. Rudiger, W. Winkler and K. Danzman, "Resonant sideband extraction: a new configuration for interferometric gravitational wave detectors," *Phys. Lett. A*, **175**, 273–276 (1993). [This is yet another recycling configuration]

N. R. W. P. Drever, "Interferometric Detectors of Gravitational Radiation," in *Gravitational Radiation*, N. Deruelle and T. Piran, eds. (North Holland, 1983); section 8 (pages 331-335). [This is the article in which Drever first presented in detail his ideas of power recycling and resonant recycling.]

A Few Suggested Problems

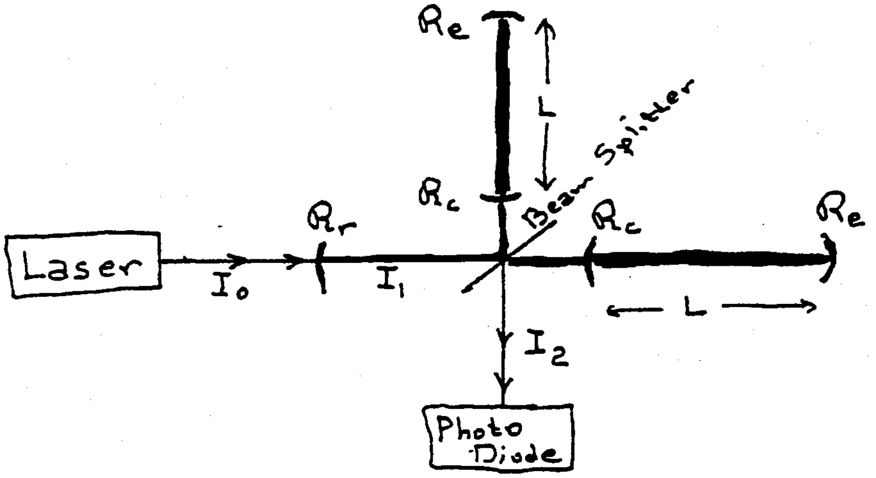
Note: Of all configurations for a recycled interferometer, the only one that is reasonably easy to analyze is power recycling. For this reason, and because this is the type of recycling planned for the first LIGO interferometers, I have chosen to focus solely on power recycling in the following exercises. — Kip.

1. *Simplified Configuration of Nested Cavities that Illustrates Power Recycling:* Consider the configuration of two nested optical cavities shown below:


outer cavity
$\mathcal{R}_r$
$\mathcal{R}_c \simeq 1-10^{-2}$
$\mathcal{R}_e \simeq 1-10^{-5}$
Laser
inner cavity
(subcavity)

All three mirrors are assumed ideal in the sense that they do not scatter or absorb any light; therefore each of them satisfies the reciprocity relations of Assignment 4, Eq. (3). Assume that the power reflectivities of the subcavity are fixed: $\mathcal{R}_e$ is the highest reflectivity the experimenter has available; $\mathcal{R}_c$ is a much lower reflectivity, carefully designed to store the light in the subcavity for a chosen length of time. What reflectivity $\mathcal{R}_r$ should the recycling mirror have in order to maximize the light intensity in the subcavity, when both cavities are operating on resonance? Use physical reasoning to guess the answer before doing the calculation.

2. *Optimization of a Power Recycled Interferometer.* Consider the power-recycled interferometer shown below.



   a. Suppose the interferometer is operated with the photodiode very near a dark fringe, so the light power I_2 is many orders of magnitude less than I_1. As in exercise 1, let $\mathcal{R}_e$ and $\mathcal{R}_c$ be fixed. How should $\mathcal{R}_r$ be chosen to maximize the power in the interferometers' two arms? Guess the answer on physical grounds before doing the calculation.

   b. Again, suppose that I_2 is many orders of magnitude less than I_1. Let a low-frequency gravitational wave (one with $2\pi fBL/c \ll 1$ where $B = 4/(1-\mathcal{R}_c)$ is the effective number of round trips in the arms) impinge on the interferometer. How should $\mathcal{R}_r$ be chosen so as to maximize the gravitational-wave signal to noise ratio in the interferometer? Guess the answer on physical grounds before doing the calculation.

c. Suppose that the mirrors in the two arms are slightly imperfect, and their imperfections cause a mismatching of the phase fronts of the light from the two arms at the beam splitter. As a result, the ratio $I_2/I_1 \equiv \alpha$ has some modest value (e.g. 0.01) instead of being arbitrarily small. In this case, how should $\mathcal{R}_r$ be chosen so as to maximize the signal to noise ratio? Guess the answer on physical grounds before doing the calculation.

3. *Scaling of Photon Shot Noise with Arm Length.* We saw in Kip's lecture that, if one has mirrors of sufficiently high reflectivity and one uses a simple (nonrecycled) interferometer, then the photon shot noise $h_{\rm rms} = \sqrt{fG_h(f)}$ is independent of the interferometer's arm length.

   Suppose, instead, that (i) the highest achievable power reflectivity is $\mathcal{R} = 1-10^{-5}$, (ii) one can do as good a job of phase-front matching at the interferometer as one wishes, so in the above drawing $I_2/I_1 = \alpha$ can be made as small as one wishes, (iii) one has a fixed laser power I_0 (say, 10 Watts) available, (iv) one operates the interferometer in a power-recycled mode, as in the above figure. *Show* that in this case the photon shot noise $h_{\rm rms}$ scales as $1/\sqrt{L}$ in the full LIGO frequency band (a result quoted on page 314 of Ref. I).

   Note: Another example of arm-length scaling is described on page 316 of Ref. I: A resonant-recycled or dual-recycled interferometer looking for periodic gravitational waves, e.g. from a pulsar, has photon shot noise $h_{\rm rms} \propto 1/L$.

LECTURE 6.
OVERVIEW OF A REAL INTERFEROMETER

Lecture by Stanley E. Whitcomb

Assigned Reading:

O. D. Shoemaker, R. Schilling, L. Schnupp, W. Winkler, K. Maischberger, A. Rüdiger, "Noise behavior of the Garching 30-meter prototype gravitational-wave interferometer," *Physical Review D,* **38**, 423-432 (1988). [This is a fairly complete description of a prototype delay-line interferometer, with all the complications of a real device.]

P. The first one or one and a half chapters of any introductory text on servosystems (also called closed loop control systems or servo loops). Don't labor slavishly over the mathematical details, but do try to get a "feel" for how servo loops work. One possibility for this is the first chapter of Benjamin C. Kuo, *Automatic Control Systems* (Prentice-Hall), which is being passed out to the class. Note that this chapter is very qualitative; you might want to dig into other books for more quantitative detail.

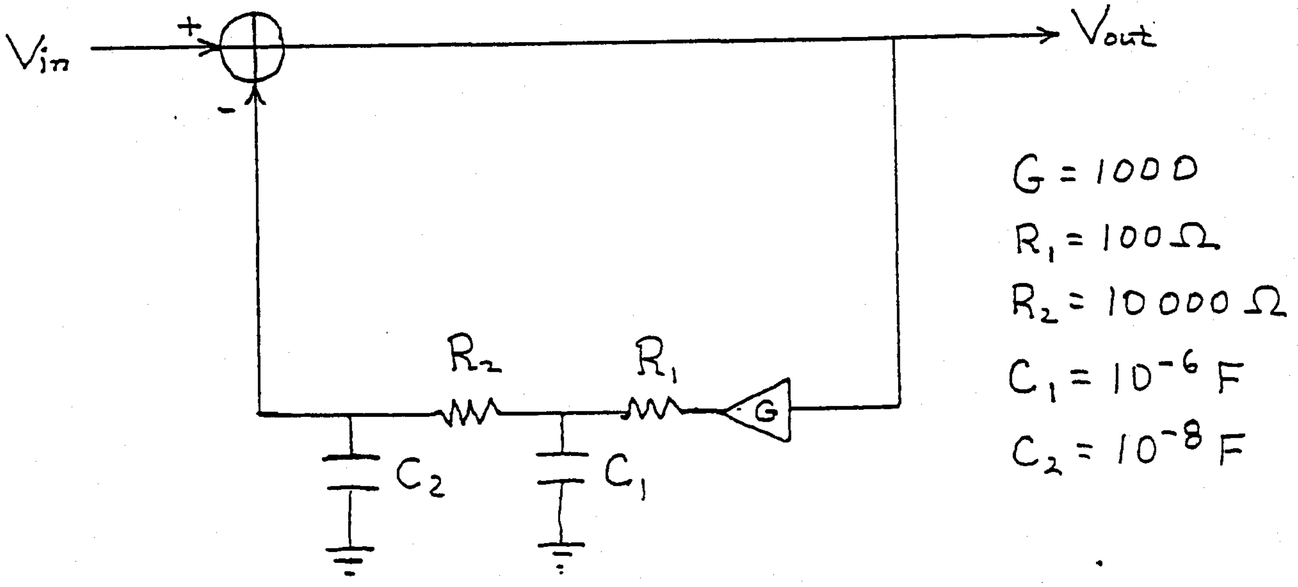
Suggested Supplementary Reading:

Read more deeply into your favorite servo text.

A Few Suggested Problems

1. *Garching Prototype.* There was a discrepancy between the observed and predicted noise spectrum in the Garching prototype (Figure 4 of Reference 1 above). The spectrum's shape is about right, suggesting that maybe the Garching group identified the right noise sources but made a calibration error that produced the numerical disagreement. On the other hand, the disagreement is approximately a factor 3, which seems a large error for a group that is generally regarded as very careful. What do you think about this? What information in the paper might lead you to one conclusion or another about the discrepancy?

2. *Example of a Servo loop.* Consider the following servo loop in an electronic circuit. It is designed to strongly suppress the input voltage V_{in} at frequencies well below some critical frequency ω_o, and pass the voltage signal more or less unchanged at frequencies well above ω_o. For the values of the servo amplifier gain G and the resistances and capacitances shown in the figure, what is the frequency ω_o? Is there any frequency region in which the servo is unstable, in the sense that it strongly amplifies the input voltage signal (i.e., it oscillates with large amplitude when a small amplitude stimulus is applied)? If so, how might you change the circuit to get rid of the unwanted amplification, while maintaining the original goals of voltage suppression well below ω_o and passing the signal unscathed well above ω_o? [*Note* (for theorists who might not know such things): The device symbolized is a voltage amplifier with gain G and it can be regarded as having infinite input impedance and zero output impedance; the device symbolized produces an output that is the difference of its two inputs, and it can be regarded as having infinite input impedances.]



from:
R. D. Blandford & K. S. Thorne
Applications of Classical Physics
(1993-94 version of a book we are writing, and are using in Ph136)

7 Diffraction

7.1 Overview

The previous chapter was devoted to the classical mechanics of wave propagation. We showed how solutions of a classical wave equation could be solved in the short wavelength approximation to yield Hamilton's dynamical equations. We then specialized to stationary media, as we shall continue to do in this chapter. Under these conditions, the frequency of a wave packet is constant. We imported a result from classical mechanics, the principle of stationary action, to show that the true geometrical optics rays were those paths along which the action or the integral of the phase was stationary. Our physical interpretation of this result was that the waves did indeed travel along every path, from some source to a point of observation, where they were added together but they only gave a significant net contribution when they could add in phase, along the true rays. This is, essentially, Huygens' model of wave propagation, or, in modern language, a *path integral*.

Huygens' principle asserted that every point on a wave front acted as a source of secondary waves that combine so that their envelope constitutes the advancing wave front. This principle must be supplemented by two ancillary conditions, that the secondary waves are only formed in the direction of wave propagation and that a 90° phase shift be introduced into the secondary wave. The reason for the former condition is obvious, that for the latter, less so. We shall discuss both together with the formal justification of Huygens' construction below.

In this chapter, we begin our exploration of the "wave mechanics" of optics. This differs increasingly from geometrical optics as the wave frequency decreases. The number of paths that can combine constructively increases and the rays that connect two points become blurred. In quantum mechanics, we recognize this phenomenon as the uncertainty principle and it is just as applicable to photons as to electrons. Solving the wave equation exactly is almost always too hard except in very simple circumstances. As we have emphasized, geometrical optics, like classical mechanics, is one approximate method of solving the fundamental wave equation in the short wavelength limit. We must now develop useful approximate techniques when geometrical optics becomes invalid.

We begin by making a somewhat artificial distinction between phenomena that arise when an effectively infinite number of paths are involved which we call *diffraction* and which we describe in this chapter, and those when a few paths, or, more correctly, a few tight bundles of rays are combined which we term *interference*, and whose discussion we defer to to the next chapter.

In Sec. 7.2, we shall present the Fresnel-Helmholtz-Kirchhoff theory that underlies most elementary discussions of diffraction, and we shall then distinguish between Fraunhofer diffraction (the limiting case when spreading of the wavefront mandated by the uncertainty principle is very important), and Fresnel diffraction (which arises when the diffracting screen is observed from much closer and it is the phase variation across the screen that is the dominant physical effect). In Sec. 7.3, we shall illustrate Fraunhofer diffraction by computing the expected angular resolution of Hubble Space Telescope and in Sec. 7.4, we shall analyze Fresnel diffraction and illustrate it using lunar occultation of radio waves and zone plates. Many contemporary optical devices are regarded as linear systems that transform an input wave signal and produce a linearly related output. Their operation, particularly as image processing devices can be considerably enhanced by processing the signal in the Fourier domain, a procedure known as spatial filtering. We shall illustrate these ideas in Sec. 7.5 using the phase contrast microscope and Gaussian beams. Finally, in Sec. 7.6 we shall analyze the effects of diffraction near a caustic of a wave's phase field, where geometric optics incorrectly predicts a divergent magnification of the wave. As we shall see, diffraction makes the magnification finite.

7.2 Helmholtz- Kirchhoff Integral

In this section, we shall derive a formalism for describing diffraction. We shall restrict our attention to the simplest (and fortunately the most widely useful) case: a scalar wave with field variable ψ of frequency $\omega = ck$ that satisfies the Helmholtz equation

$$\nabla^2\psi + k^2\psi = 0 \tag{7.1}$$

except at boundaries. Generally ψ will represent a real valued physical quantity (although it may, for mathematical convenience, be given a complex representation). This is in contrast to a quantum mechanical wave function satisfying the Schrödinger equation which is an intrinsically complex function. The wave is monochromatic and non-dispersive and the medium is isotropic and homogeneous so that k can be treated as constant. Each of these assumptions can be relaxed with some technical penalty.

The scalar formalism that we shall develop based on Eq. (7.1) is fully valid for weak sound waves in a fluid, e.g. air (Chap. 15). It is also fairly accurate, but not precisely so, for the most widely used application of diffraction theory: the propagation of electromagnetic waves in vacuo or in a medium with homogeneous dielectric constant. In this case ψ can be regarded as one of the Cartesian components of the electric field vector, e.g. E_x (or equally well a Cartesian component of the vector potential or the magnetic field vector). In vacuo or in a homogeneous dielectric medium, Maxwell's equations imply that this $\psi = E_x$ satisfies the scalar wave equation and thence, for fixed frequency, the Helmholtz equation (7.1). However, when the wave hits a boundary of the medium (e.g. the edge of an aperture, or the surface of a mirror or lens), its interaction with the boundary can couple the various components of **E**, thereby invalidating the simple scalar theory we shall develop. Fortunately, this polarizational coupling is usually very weak in the paraxial (small angle) limit,and also under a variety of other circumstances, thereby making our simple scalar formalism quite accurate.

The Helmholtz equation (7.1) is an elliptic, linear, partial differential equation, and it thus permits one to express the value $\psi_\mathcal{P}$ of ψ at any point $\mathcal{P}$ inside some closed surface $\mathcal{S}$ as an integral over $\mathcal{S}$ of some linear combination of ψ and its normal derivative; see Fig. 7.1. To derive such an expression, we first combine the actual wave ψ in the interior of $\mathcal{S}$ with a second solution of the Helmholtz equation, namely

$$\psi_0 = \frac{e^{ikr}}{r}\,. \tag{7.2}$$

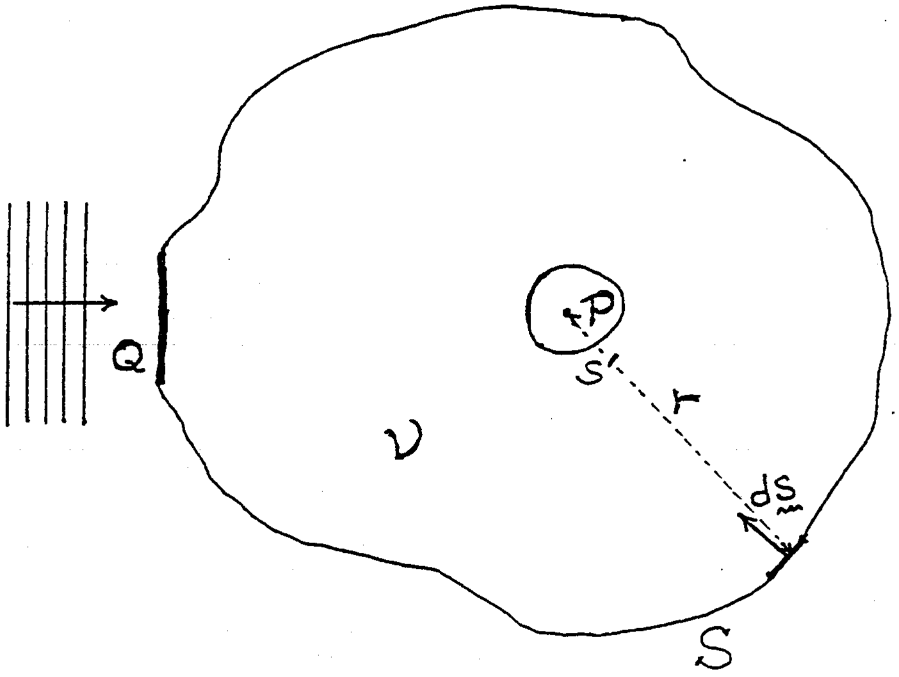


Fig. 7.1 Surface $\mathcal{S}$ for Helmholtz-Kirchhoff Integral. The surface $\mathcal{S}'$ surrounds the observation point $\mathcal{P}$ and $\mathcal{V}$ is the volume bounded by these two surfaces. The aperture $\mathcal{Q}$ is irrelevant to the formulation of the Helmholtz-Kirchoff integral, but appears in subsequent applications.

This is a spherical wave originating from the point $\mathcal{P}$, and r is the distance from $\mathcal{P}$ to the field point. Next, we use Eq. (7.1) to prove Green's theorem:

$$\begin{aligned}\int_{\mathcal{S}+\mathcal{S}'} (\psi \boldsymbol{\nabla}\psi_0 - \psi_0 \boldsymbol{\nabla}\psi)\cdot d\mathbf{S} &= \int_{\mathcal{V}} (\psi\nabla^2\psi_0 - \psi_0\nabla^2\psi)dV \\ &= 0\,. \end{aligned} \tag{7.3}$$

Here we have introduced a second surface, a small sphere $\mathcal{S}'$ of radius r' surrounding $\mathcal{P}$; and $\mathcal{V}$ is the volume between the two surfaces. As we let the radius r' decrease to zero, we find that, $\psi\boldsymbol{\nabla}\psi_0 - \psi_0\boldsymbol{\nabla}\psi \sim -\psi(0)/r'^2 + O(1/r')$ and so the integral over $\mathcal{S}'$ becomes $\sim 4\pi\psi_{\mathcal{P}}$. Rearranging, we obtain

$$\psi_{\mathcal{P}} = \frac{1}{4\pi}\int_{\mathcal{S}} d\mathbf{S}\cdot\left(\psi\nabla\frac{e^{ikr}}{r} - \frac{e^{ikr}}{r}\boldsymbol{\nabla}\psi\right) \tag{7.4}$$

where the element of surface $d\mathbf{S}$ is directed inward as shown in Fig. 7.1.

Equation (7.4), known as the *Helmholtz-Kirchhoff formula*, is the promised expression for the field ψ at some point $\mathcal{P}$ in terms of a linear combination of its value and normal derivative on a bounding surface. The specific combination of ψ and $d\mathbf{S}\cdot\boldsymbol{\nabla}\psi$ that appears in this formula is perfectly immune to contributions from any wave that might originate at $\mathcal{P}$ and pass outward through $\mathcal{S}$ (any "outgoing wave"). The integral thus is influenced only by waves that enter $\mathcal{V}$ through $\mathcal{S}$, propagate through $\mathcal{V}$, and then leave through $\mathcal{S}$. Moreover, if $\mathcal{P}$ is many wavelengths away from the boundary $\mathcal{S}$, then to high accuracy the integral is influenced by the waves ψ only when they are entering through $\mathcal{S}$ (when they are incoming), and not when they are leaving (outgoing). This fact is important for applications, as we shall see.

Diffraction by an Aperture

Next, let us suppose that some aperture $\mathcal{Q}$ of size much larger than a wavelength but much smaller than the distance tó $\mathcal{P}$ is illuminated by a distant wave source (Fig. 7.1). (If the aperture were comparable to a wavelength in size, or if part of it were only a few wavelengths from $\mathcal{P}$, then polarizational coupling effects at the aperture would be large; our assumption avoids this complication.) Let the surface $\mathcal{S}$ pass through $\mathcal{Q}$, and denote by ψ' the wave incident on $\mathcal{Q}$. We assume that the diffracting aperture has a local and linear effect on ψ'. More specifically, we suppose that the wave transmitted through the aperture is given by

$$\psi_{\mathcal{Q}} = t\psi' , \tag{7.5}$$

where t is a complex transmission function that varies over the aperture. In practice, t is usually zero (completely opaque region) or unity (completely transparent region). However t can also represent a variable phase factor when, for example, the aperture comprises a medium of variable thickness and of different refractive index from that of the homogeneous medium outside the aperture.

What this formalism does not allow, though, is that $\psi_{\mathcal{Q}}$ at any point on the aperture be influenced by the wave's interaction with other parts of the aperture. For this reason, not only the aperture, but any structure that it contains must be many wavelengths across. To give a specific example of what might go wrong, suppose that electromagnetic radiation is normally incident upon a wire grid. Surface currents will be induced in the wires by the wave's electric field, and those currents will produce a secondary wave that cancels the primary wave immediately behind each wire, thereby "eclipsing" the wave. However, if the secondary wave from the currents flowing in the next wire is comparable with the first wire's secondary wave, then the transmitted net wave field will get modified in a complex, polarization-dependent manner. Such modifications are negligble if the wires are a number of wavelengths apart.

Let us now use the Helmholtz-Kirchoff formula (7.4) to compute the field at $\mathcal{P}$ due to the wave $\psi_{\mathcal{Q}} = t\psi'$ transmitted through the aperture. Let the surface $\mathcal{S}$ of Fig. 7.1 comprise the aperture $\mathcal{S}$ comprise the aperture $\mathcal{Q}$, a sphere of radius $R \gg r$ centered on $\mathcal{P}$, and the linear extension of the aperture to meet the sphere; and assume that the only incoming waves are those which pass through the aperture. Then, as noted above, when the incoming waves subsequently pass on outward through $\mathcal{S}$, they contribute negligibly to the integral (7.4), so the only contribution is from the aperture itself. [1]

[1] Actually, the incoming waves will diffract around the edge of the aperture onto the back side of the screen that bounds the aperature, i.e. the side facing $\mathcal{P}$; and this diffracted wave will contribute to the Helmholtz-Kirchhoff integral in a polarization-dependent way. However, because the diffracted wave decays along the

On the aperture, because $kr \gg 1$, we can write $\nabla(e^{ikr}/r) \simeq -ik\mathbf{n}e^{ikr}/r$ where $\mathbf{n}$ is a unit vector pointing towards $\mathcal{P}$. Similarly, we write $\nabla\psi \simeq ikt\mathbf{n}'\psi'$, where $\mathbf{n}'$ is a unit vector along the direction of propagation of the incident wave, and where our assumptions have permitted us to ignore the gradient of t. Inserting these gradients into the Helmholtz-Kirchoff formula, we obtain

$$\psi_P = -\frac{ik}{2\pi}\int_{\mathcal{Q}} d\mathbf{S}\cdot\left(\frac{\mathbf{n}+\mathbf{n}'}{2}\right)\frac{e^{ikr}}{r}t\psi' . \tag{7.6}$$

Equation (7.6) can be used to compute the wave from a small aperture at any point in the far field. It has the form of an integral transform of the incident field variable, ψ', where the integral is over the area of the aperture. The kernel of the transform is the product of several factors. There is a factor $1/r$. This guarantees that the flux falls off as the inverse square of the distance to aperture as we might have expected. There is also a phase factor $-ie^{ikr}$ which advances the phase of the wave by an amount equal to the optical path length between the element of the aperture and $\mathcal{P}$ minus $\pi/2$. The amplitude and phase of the wave can also be changed by the transmission function t. Finally there is the geometrical factor $d\hat{\mathbf{S}}\cdot(\mathbf{n}+\mathbf{n}')/2$ which is known as the *obliquity factor.* This factor ensures that the waves from the aperture propagate only forward with respect to the original wave, and not backward (not in the direction $\mathbf{n} = -\mathbf{n}'$). More specifically, this factor prevents the backward propagating secondary wavelets in Huygens construction from reinforcing each other to produce a back-scattered wave. When dealing with paraxial optics (later in this chapter), we can usually set the obliquity factor to unity.

It is instructive to specialize to a point source seen through a small diffracting aperture. If we suppose that the source has unit strength and is located at $\mathcal{P}'$, a distance r' before $\mathcal{Q}$, then $\psi' = -e^{ikr'}/4\pi r'$; and ψ_P can be written in the symmetric form

$$\psi_P = \int\left(\frac{e^{ikr}}{4\pi r}\right) it(\mathbf{k}'+\mathbf{k})\cdot d\mathbf{S}\left(\frac{e^{ikr'}}{4\pi r'}\right) . \tag{7.7}$$

We can think of this expression as the Greens function response at $\mathcal{P}$ to a δ-function source at $\mathcal{P}'$. Alternatively, we can regard it as a *propagator* from $\mathcal{P}'$ to $\mathcal{P}$ by way of the aperture.

****

EXERCISES

Exercise 7.1, Problem: Angular Resolution

screen with an e-folding length of order a wavelength, its contribution will be negligible if the aperture is many wavelengths across and $\mathcal{P}$ is many wavelengths away from the edge of the aperture, as we have assumed.

The neo-impressionist painter George Seurat was a member of the pointillist school. His paintings consisted of an enormous number of closely spaced dots (of size ~ 0.4mm) of pure pigment. The illusion of color mixing was produced only in the eye of the observer. How far from the painting should one stand in order to obtain the desired blending of color?

****

Spreading of the Wavefront

Equation (7.6) [or (7.7)] gives a general prescription for computing the diffraction pattern from an illuminated aperture. It is commonly used in two complementary limits, called "Frauhnhofer" and "Fresnel".

Suppose that the aperture has linear size a and is roughly centered on the geometrical ray from the source point $\mathcal{P}'$ to the field point $\mathcal{P}$. Consider the variations of the phase ϕ of the contributions to $\psi_{\mathcal{P}}$ that come from various places in the aperture. Using elementary trigonometry, we can estimate that locations on the aperture's opposite sides have phases that differ by $\Delta\phi \sim ka^2/2r$. There are two limiting regimes depending on whether the aperture is large or small compared with the so-called *Fresnel length*

$$r_F \equiv \left(\frac{2\pi r}{k}\right)^{1/2} = (\lambda r)^{1/2}. \tag{7.8}$$

(Note that the Fresnel length depends on the distance r of the field point from the aperture.) When $a \ll r_F$, the phase variation $\Delta\phi$ can be ignored; the contributions from different parts of the aperture are essentially in phase with each other. This is the *Fraunhofer* regime. When $a \gg r_F$, the phase variation is of upmost importance in determining the observed intensity pattern $|\psi_{\mathcal{P}}|^2$. This is the *Fresnel* regime; see Fig. 7.2.

We can use an argument familiar, perhaps, from quantum mechanics to deduce the qualitative form of the intensity patterns in these two regimes. For simplicity, let the incoming wave be planar (r' huge) and let it propagate perpendicular to the aperture as shown in Fig. 7.2. Then geometrical optics (photons treated like classical particles) would predict that an opaque screen will cast a sharp shadow; the wave leaves the aperture plane as a beam with a sharp edge. However, wave optics insists that the transverse localization of the wave into a region of size $\Delta x \sim a$ must produce a spread in its transverse wave vector, $\Delta k_x \sim 1/a$ (a momentum uncertainty $\Delta p_x = \hbar\Delta k_x \sim \hbar/a$ in the language of the Heisenberg uncertainty principle). This uncertain transverse wave vector produces, after propagating a distance r, a corresponding uncertainty $(\Delta k_x/k)r \sim r_F^2/a$ in the beam's transverse size; and this uncertainty superposes incoherently on the aperture-induced size a of the beam to produce a net size

$$\begin{aligned}\Delta x &\sim \sqrt{a^2 + (r_F^2/a)^2}\\ &\sim a \quad \text{if} \quad r \ll a^2/\lambda \text{ (Fresnel regime)}\\ &\sim \lambda r/a \quad \text{if} \quad r \gg a^2/\lambda \text{ (Fraunhofer regime).}\end{aligned}$$

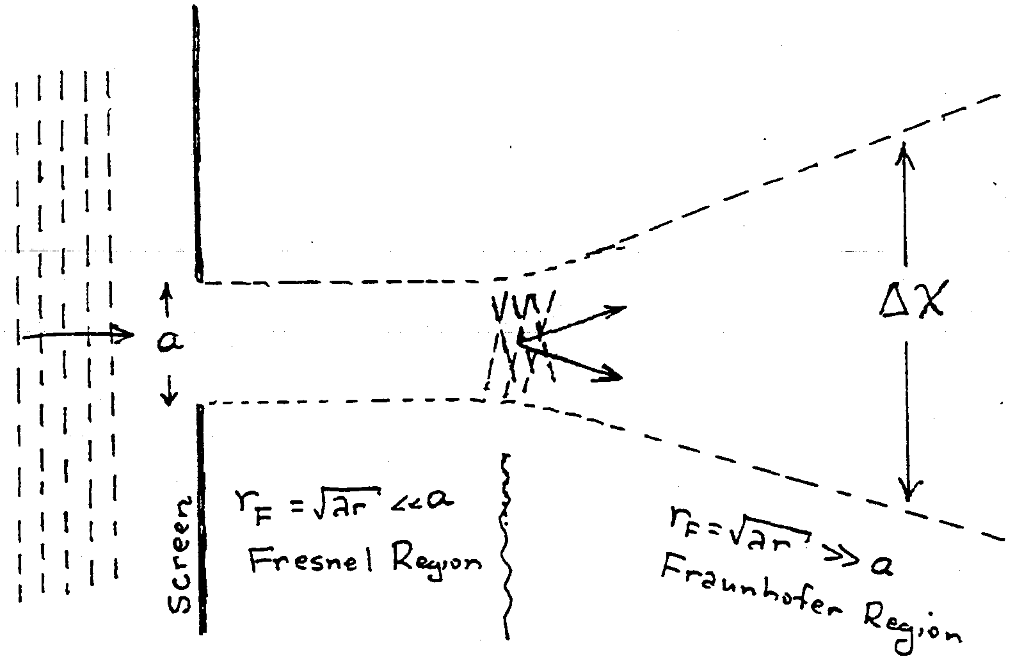

a
Screen
$r_F = \sqrt{\lambda r} \ll a$
Fresnel Region
ΔX
$r_F = \sqrt{\lambda r} \gg a$
Fraunhofer Region

Fig. 7.2 Fraunhofer and Fresnel Diffraction.

In the nearby, Fresnel regime, the aperture creates a beam whose edges will have the same shape and size as the aperture itself, and will be reasonably sharp (but with some oscillatory blurring, associated with the wave-packet spreading, that we shall analyze below). By contrast, in the more distant Fraunhofer regime, wave-front spreading will cause the transverse size of the entire beam to grow linearly with distance; and, as we shall see, the intensity pattern typically will not resemble the aperture at all.

7.3 Fraunhofer Diffraction

Consider the Fraunhofer regime of strong wavefront spreading, $a \ll r_F$, and for simplicity specialize to the case of an incident plane wave with wave vector $\mathbf{k}$ orthogonal to the aperture plane; see Fig. 7.3. Regard the line along $\mathbf{k}$ through the center of the aperture $\mathcal{Q}$ as the "optic axis;" identify points in the aperture by their two-dimensional vectorial separation $\mathbf{x}$ from that axis; identify $\mathcal{P}$ by its distance r from the aperture center and its 2-dimensional transverse separation $r\boldsymbol{\theta}$ from the optic axis; and restrict attention to small-angle diffration $|\theta| \ll 1$. Then the geometrical path length between $\mathcal{P}$ and a point on $\mathcal{Q}$ [denoted r in Eq. (7.6)—note our change of the meaning of r] can be expanded as

$$(r^2 - 2r\mathbf{x}\cdot\boldsymbol{\theta} + x^2)^{1/2} \sim r - \mathbf{x}\cdot\boldsymbol{\theta} + \frac{x^2}{2r} + \ldots \tag{7.9}$$

*cf.*Fig. 7.3. The first term on the right side of this equation just contributes a constant phase to the $\psi_{\mathcal{P}}$ of Eq. (7.6). The third term contributes a phase variation that, by assumption, is $\ll 1$. Therefore, we can retain just the second term and write Eq. (7.6) in

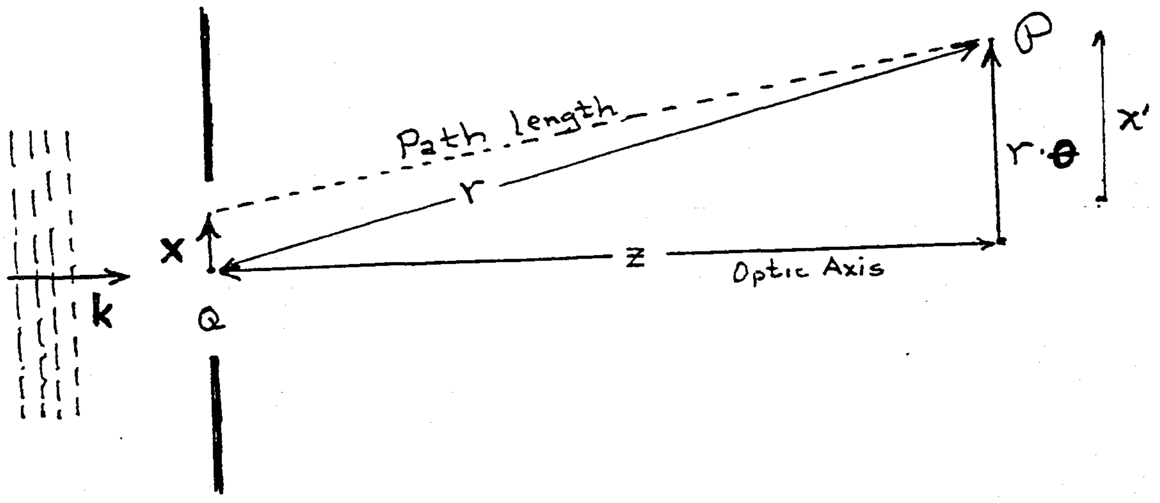


Fig. 7.3 Geometrical interpretation of the path length between a point $\mathcal{Q}$ in the aperture and the point of observation $\mathcal{P}$.

the form

$$\psi_{\mathcal{P}}(\boldsymbol{\theta}) \propto \int d^2x e^{-ik\mathbf{x}\cdot\boldsymbol{\theta}} t(\mathbf{x}) = \tilde{t}(\boldsymbol{\theta}) \,, \tag{7.10}$$

where we have dropped the constant phase factor and constant multiplicative factors. Thus, $\psi_{\mathcal{P}}(\boldsymbol{\theta})$ in the Fraunhofer regime is given by the two-dimensional Fourier transform, denoted $\tilde{t}(\boldsymbol{\theta})$, of the transmission function $t(\mathbf{x})$, with $\mathbf{x}$ made dimensionless in the transform by multiplying by $k = 2\pi/\lambda$.

It is usually uninteresting to normalise Fraunhofer diffraction patterns. Moreover, on those occasions when the absolute value of the observed flux is needed, rather than just the angular shape of the diffraction pattern, it typically can be derived most easily from conservation of the total wave energy. This is why we ignore the proportionality factor in Eq. (7.10).

All of the techniques for handling Fourier transforms that should be familiar from quantum mechanics and elsewhere can be applied to derive Fraunhofer diffraction patterns. In particular, the convolution theorem turns out to be very useful. Let us give an example.

Diffraction Grating

A diffraction grating can be modeled as a finite series of alternating transparent and opaque, long, parallel stripes. Let there be N transparent and opaque stripes each of width $a \gg \lambda$ (Fig. 7.4a), and idealize them as infinitely long so their diffraction pattern is one-dimensional. We shall outline how to use the convolution theorem to derive their Fraunhofer diffraction pattern. The details are left as an exercise for the reader (Ex. 7.2).

First consider a single transparent stripe (slit) of width a centered on $x = 0$, and measure the scalar angle θ from the direction of the incident radiation. This single stripe

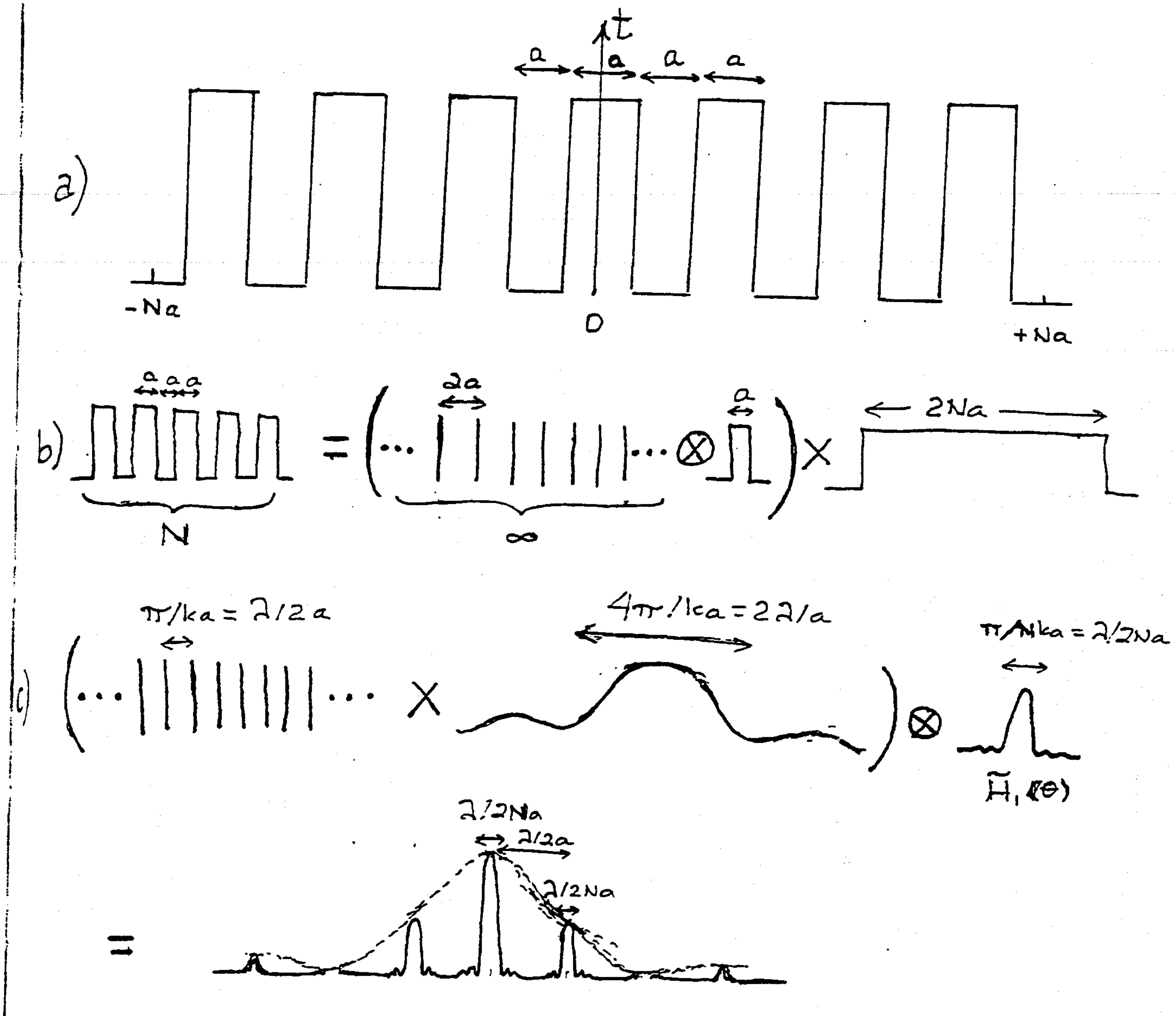


Fig. 7.4 a) Diffraction Grating formed by N alternating transparent and opaque stripes each of width a. b) Decomposition of a finite grating into an infinite series of equally spaced δ-functions that are convolved (the symbol $\otimes$) with the shape of an individual transparent stripe and then multiplied (the symbol $\times$) by a large aperture function covering N such stripes; cf. Eq. (7.14). c) The resulting Fraunhofer diffraction pattern shown schematically as the Fourier transform of a series of delta functions multiplied by the Fourier transform of the large aperture and then convolved with the transform of a single stripe.

has the transmission function

$$t_1(x) = 1 \quad |x| < a/2$$
$$= 0 \quad |x| > a/2\,, \tag{7.11}$$

so its diffraction pattern is

$$\begin{aligned}\psi_P(\theta) &\propto \tilde{t}_1\\ &\propto \int_{-a/2}^{a/2} dx e^{ikx\theta}\\ &\propto \text{sinc}\left(\frac{ka\theta}{2}\right)\,,\end{aligned} \tag{7.12}$$

where $\text{sinc}(x) \equiv \sin(x)/x$.

Now the idealized N-slit grating can be considered as an infinite series of δ-functions with separation $2a$ convolved with the transmission function for a single slit,

$$\int_{-\infty}^{\infty} \left[\sum_{n=-\infty}^{+\infty} \delta(y - 2an)\right] t_1(x-y)dy$$

and then multiplied by the aperture function

$$H(x) = 1 \quad |x| < Na$$
$$= 0 \quad |x| > Na; \tag{7.13}$$

more explicitly,

$$t(x) = \left(\int_{-\infty}^{\infty} \left[\sum_{n=-\infty}^{+\infty} \delta(y - 2an)\right] t_1(x-y)dy\right) H(x)\,, \tag{7.14}$$

which is shown graphically in Fig. 7.4b.

The convolution theorem says that the Fourier transform of a convolution of two functions is the product of the functions' Fourier transforms, and conversely. Let us apply this theorem to expression (7.14) for our transmission grating. The diffraction pattern of the infinite series of δ-functions with spacing $2a$ is itself an infinite series of δ-functions with reciprocal spacing $2\pi/(2ka) = \lambda/2a$ (see the hint in Exercise 7.2). This must be multiplied by the Fourier transform $\tilde{t}_1(\theta)$ of the single slit, and then convolved with the Fourier transform of $H(x)$, $\tilde{H}(\theta) \propto \text{sinc}(Nka\theta)$. The result is shown schematically in Fig. 7.4c. (Each of the transforms is real, so the one-dimensional functions shown in the figure fully embody them.)

The diffracted energy flux is $|\psi_P|^2$, where ψ_P is shown at the bottom of Fig. 7.4. What the grating has done is channel the incident radiation into a few equally spaced beams with directions $\theta = \pi p/ka$, where p is an integer known as the *order* of the grating. Each of these beams has a shape given by $|\tilde{H}(\theta)|^2$: a sharp central peak with half width (distance from center of peak to first null of the intensity) $\lambda/2Na$, followed by a set of *side lobes* whose intensities are $\propto N^{-1}$.

The fact that the deflection angles $\theta = \pi p/ka$ of these beams are proportional to $k^{-1} = \lambda/2\pi$ underlies the use of diffraction gratings for spectroscopy. It is of interest to ask what the wavelength resolution of such an idealized grating might be. If one focuses attention on the p'th order beams at two wavelengths λ and $\delta\lambda$ (which are located at $\theta = p\lambda/2a$ and $p(\lambda + \delta\lambda)/2a$, then one can distinguish the beams from each other when their separation $\delta\theta = p\delta\lambda/2a$ is at least as large as the angular distance $\lambda/2Na$ between the maximum of each beam's diffraction pattern and its first minimum, i.e., when

$$\frac{\lambda}{\delta\lambda} \gtrsim \mathcal{R} \equiv Np\,. \tag{7.15}$$

$\mathcal{R}$ is called the grating's *chromatic resolving power.*

Real gratings are not this simple. Firstly they usually work not by modulating the amplitude of the incident radiation in this simple manner, but instead by modulating the phase. Secondly, the manner in which the phase is modulated is such as to channel most of the incident power into a particular order, a technique known as *blazing.* Thirdly, gratings are often used in reflection rather than transmission. Despite these complications, the principles of a real grating's operation are essentially the same as our idealized grating. Manufactured gratings typically have $N \gtrsim 10,000$, giving a wavelength resolution for visual light that can be as small as $\sim 10\,\text{pm}$, i.e. 10^{-11}m.

Babinet's Principle

We have shown how to compute the Fraunhofer diffraction pattern formed by, for example, a narrow slit. We might also be interested in the pattern from a complementary aperture, a needle of width and length the same as those for the slit. We can derive the needle's pattern by observing that the sum of the waves from the two apertures should equal the wave from a completely unaltered incident wave front. That is to say if we exclude the direction of the incident wave, the field amplitudes diffracted by the two apertures are the negative of each other, and hence the intensities $|\psi|^2$ are the same. Therefore, the Fraunhofer diffraction patterns from the needle and the slit—and indeed from any pair of complementary apertures, e.g., Fig. 7.5—are identical, except in the direction of the incident wave (the "precisely forward" direction).

Hubble Space Telescope

The Hubble Space Telescope, was launched in April 1990 to observe planets, stars and galaxies above the earth's atmosphere. One reason for going into space is to avoid the irregular refractive index variations in the earth's atmosphere, known generically as *seeing*, which degrade the quality of the images. (Another reason is to observe the ultraviolet part of the spectrum.) Seeing typically limits the angular resolution of Earth-bound telescopes at visual wavelengths to $\sim 1''$. We wish to compute how much the angular resolution improves by going into space. As we shall see, the computation is essentially an exercise in Fraunhofer diffraction theory.

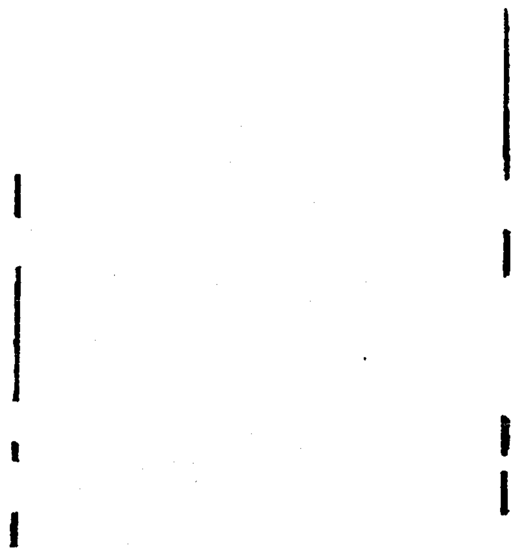


Fig. 7.5 Two complementary apertures used to illustrate Babinet's Principle

The essence of the computation is to idealise the telescope as a circular aperture with diameter equal to the diameter of the primary mirror. Light from this mirror is actually reflected onto a secondary mirror and then follows a complex optical path before being focused onto a variety of detectors. However, this path is irrelevant to the angular resolution. The purpose of the optics is merely to bring the light to a focus close to the mirror, in order to produce an instrument that is compact enough to be launched and to match the sizes of stars' images to the pixel size on the detector. In doing so, however, the optics leaves the angular resolution unchanged; the resolution is the same as if we were simply to observe the light, which passes through the primary mirror's circular aperture, far beyond the mirror, in the Fraunhofer region.

If the telescope aperture were very small, for example a pin hole, then the light from a point source (a very distant star) would create a broad diffraction pattern, and the telescope's angular resolution would be correspondingly poor. As we increase the diameter of the aperture, we still see a diffraction pattern, but its width diminishes.

Using these considerations, we can compute how well the telescope can distinguish nearby stars. We do not expect it to resolve them if they are closer together on the sky than the angular width of the diffraction patttern. Of course, optical imperfections in a real telescope may degrade the image quality even further, but this is the best that we can do, limited only by the uncertainty principle.

The calculation of the Fraunhofer amplitude far from the aperture is straightforward:

$$\begin{aligned} \psi(\theta) &\propto \int d^2x e^{-ik\mathbf{x}\cdot\boldsymbol{\theta}} \\ &\propto \operatorname{jinc}\left(\frac{kD\theta}{2}\right) \end{aligned} \tag{7.16}$$

where D is the diameter of the aperture (i.e., of the telescope's primary mirror) and $\operatorname{jinc}(x) = J_1(x)/x$ with J_1 the Bessel function of order one. The flux from the star observed

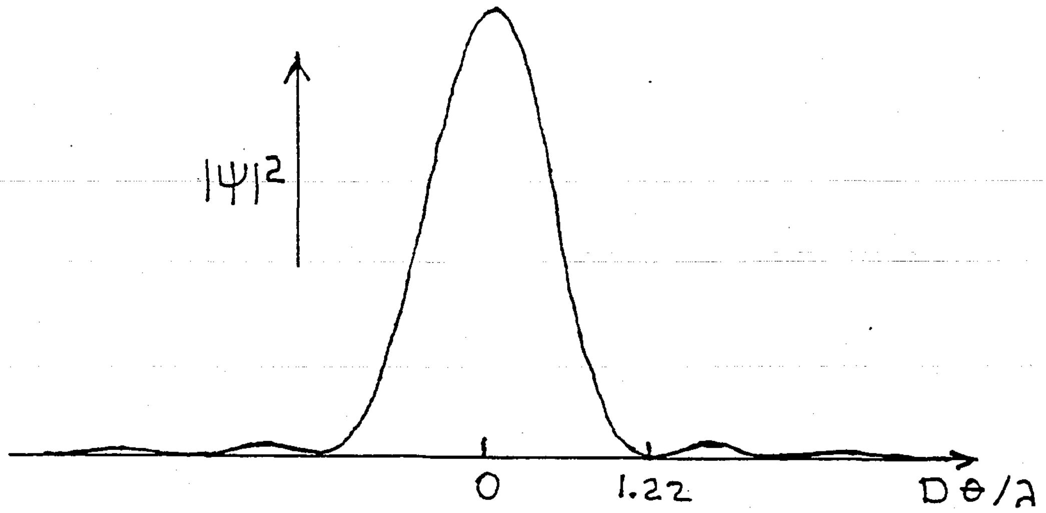


Fig. 7.6 Airy diffraction pattern produced by a circular aperture.

at angle θ is therefore $\propto \mathrm{jinc}^2(kD\theta/2)$. This intensity pattern, known as the *Airy pattern*, is shown in Fig. 7.6. There is a central "Airy disk" surrounded by a circle where the flux vanishes, and then further surrounded by a series of concentric rings whose flux diminishes with radius. Only 16 percent of the total light falls outside the central Airy disk. The radius θ_A of Airy disk, i.e. the radius of the dark circle surrounding it, is determined by the first zero of $J_1(kD\theta/2)$: $\theta_A = 1.22\lambda/D$.

A conventional, though essentially arbitrary, criterion for angular resolution is to say that two point sources can be distinguished if they are separated in angle by more than θ_A. For the Hubble Space Telescope, $D = 2.4$m and $\theta_A \sim 0.04''$ at visual wavelengths, which is over ten times better than is achievable on the ground.

As has been widely publicized, there is a serious problem with Hubble's telescope optics. The hyperboloidal primary mirror was ground to the wrong shape, so rays parallel to the optic axis do not pass through a common focus after reflection off a convex hyperboloidal secondary mirror. This defect, known as *spherical aberration*, creates blurred images. It is hoped that correcting optics will be installed when astronauts next visit the telescope.

****

EXERCISES

Exercise 7.2, Derivation: Diffraction Grating

Use the convolution theorem to carry out the calculation of the Fraunhofer diffraction pattern from the grating shown in Fig. 7.4. [*Hint:* To show that the Fourier transform of the infinite sequence of equally spaced delta functions is a similar sequence of delta functions, perform the Fourier transform to get $\sum_{n=\infty}^{+\infty} e^{i2kan\theta}$ (aside from a multiplicative factor); then use the formulas for a Fourier *series* expansion, and its inverse, for any function that is periodic with period π/ka to show that $\sum_{n=\infty}^{+\infty} e^{i2kan\theta}$ is a sequence of delta functions.]

Exercise 7.3, Problem: Triangular Diffraction Grating

Sketch the Fraunhofer diffraction pattern you would expect to see from a diffraction grating made from three groups of parallel lines aligned at angles of 120° to each other (Fig. 7.7).

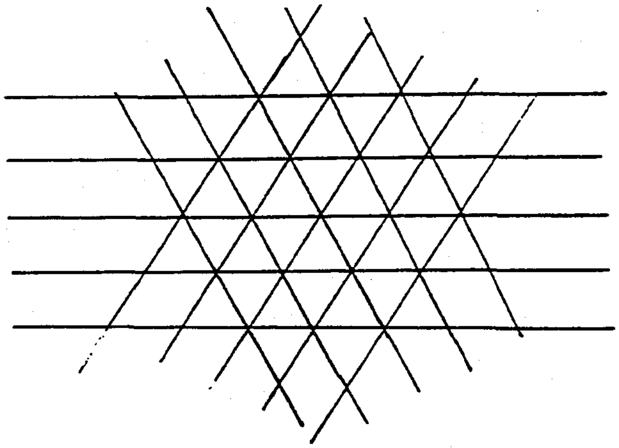


Fig. 7.7 Diffraction grating formed from three groups of parallel lines.

Exercise 7.4, Problem: Light Scattering by Particles

Consider the scattering of light by a particle of size $a \gg 1/k$. One component of the scattered radiation is due to diffraction around the particle. This component is confined to a cone with opening angle $\Delta\theta \sim 1/ka \ll 1$ about the incident wave direction. It contains power $P_S = FA$, where F is the incident energy flux and A is the cross sectional area of the particle perpendicular to the incident wave.

(a) Give a semi-quantitative derivation of $\Delta\theta$ and F_S using Babinet's principle.

(b) Explain why the total "extinction" (absorption plus scattering) cross section is equal to $2A$ independent of the composition of the particle.

****

7.4 Fresnel Diffraction

Now turn to the Fresnel regime, where the aperture is far larger than the Fresnel length r_F and there is a large phase variation over the aperture. As for the Fraunhofer case, we shall specialize to an incoming plane wave with wave vector orthogonal to the aperture plane and to small diffraction angles so that we can ignore the obliquity factor. By contrast with the Frauhofer case, however, we identify $\mathcal{P}$ by its distance z from the aperture plane instead of its distance r from the aperture center, and we use as our integration variable in the aperture $\mathbf{x}' \equiv \mathbf{x} - r\theta$ (*cf.*Fig. 7.3.), thereby writing the dependence of the phase at $\mathcal{P}$ on $\mathbf{x}$ in the form

$$\Delta\phi \equiv k \times [(\text{path length from } \mathbf{x} \text{ to } \mathcal{P}) - z] = \frac{k\mathbf{x}'^2}{2z} \tag{7.17}$$

Let us consider the Fresnel diffraction pattern formed by a simple aperture of arbitrary shape, illuminated by a normally incident plane wave. It is convenient to introduce Cartesian coordinates (x', y') and to define

$$\begin{aligned} s &= \left(\frac{k}{\pi z}\right)^{1/2} x' , \\ t &= \left(\frac{k}{\pi z}\right)^{1/2} y' . \end{aligned} \tag{7.18}$$

We can therefore rewrite Eq. (7.6) (ignoring the obliquity factor) in the form

$$\begin{aligned} \psi_{\mathcal{P}} &= -\frac{ike^{ikz}}{2\pi z}\int_{\mathcal{Q}} dx'dy'e^{i\Delta\phi}\psi_{\mathcal{Q}} \\ &= -\frac{i}{2}\int dse^{i\pi s^2/2}\int dte^{i\pi t^2/2}\psi_{\mathcal{Q}}e^{ikz} . \end{aligned} \tag{7.19}$$

Treating $\psi_{\mathcal{Q}}$ as constant and ignoring a constant phase factor, we find that the two integrals can be expressed in the form $S(s_{max}) - S(s_{min})$ and $S(t_{max}) - S(t_{min})$, so

$$\psi_{\mathcal{P}} = \frac{-i}{2}[S(s_{max}) - S(s_{min})][S(t_{max}) - S(t_{min})]\psi_{\mathcal{Q}}e^{ikz} , \tag{7.20}$$

where the arguments are the limits of integration (which depend on the shape of the aperture and the transverse location of $\mathcal{P}$) and where

$$S(\xi) \equiv \int_0^{\xi} e^{i\pi s^2/2}ds \equiv U(\xi) + iV(\xi) \tag{7.21}$$

with

$$\begin{aligned} U(\xi) &= \int_0^{\xi} ds\cos(\pi s^2/2) , \\ V(\xi) &= \int_0^{\xi} ds\sin(\pi s^2/2) . \end{aligned} \tag{7.22}$$

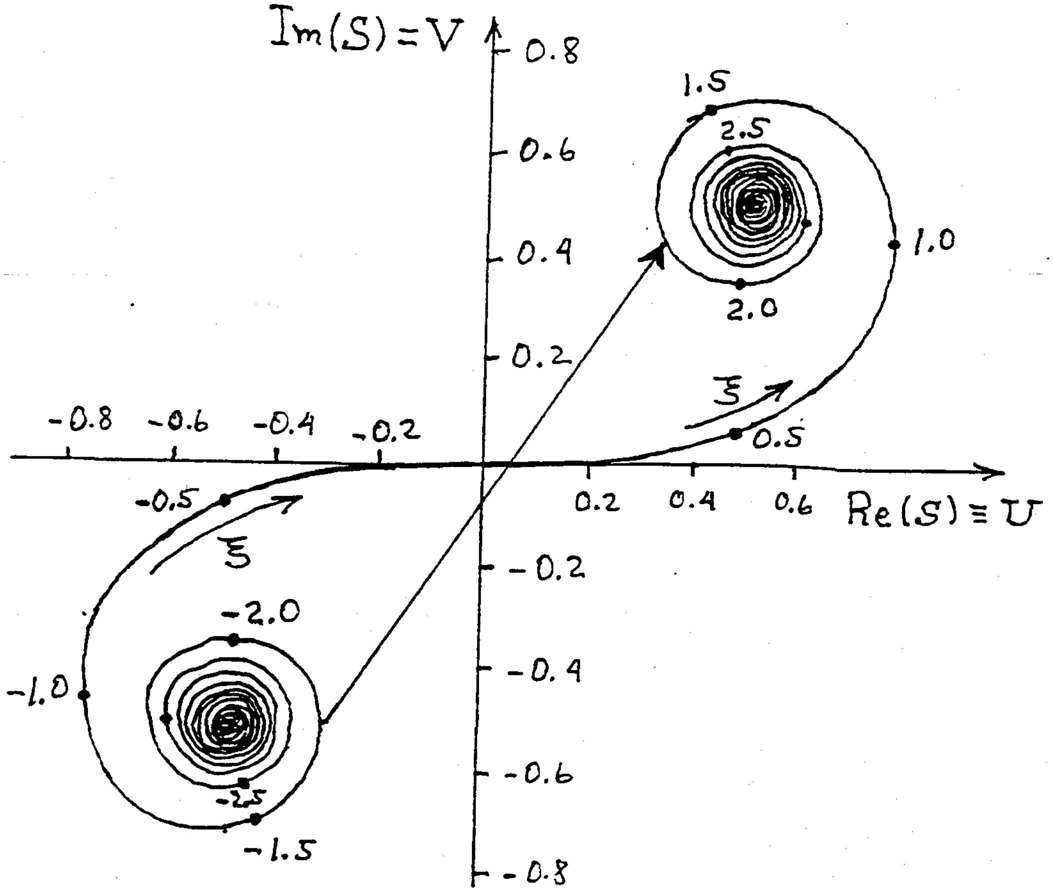


Fig. 7.8 Cornu Spiral.

The real functions $U(\xi), V(\xi)$ are known as Fresnel integrals.

It is convenient to exhibit the Fresnel integrals graphically using a *Cornu Spiral* (Fig. 7.8). This is a graph of the parametric equation $[U(\xi), V(\xi)]$, or equivalently a graph of $S(\xi) = U(\xi) + iV(\xi)$ in the complex plane. The two terms in Eq. (7.21) can be represented in amplitude and phase by arrows in the (U, V) plane reaching from $\xi = s_{min}$ on the Cornu spiral to $\xi = s_{max}$, and from $\xi = t_{min}$ to $\xi = t_{max}$.

The simplest illustration is the totally unobscured, plane wavefront. In this case, the limits of both integrations extend from $-\infty$ to $+\infty$, which as we see in Fig. 7.8 is an arrow of length $2^{1/2}$ and phase $\pi/4$. Therefore, ψ_P is equal to $(2^{1/2}e^{i\pi/4})^2(-i/2)\psi_Q e^{ikz} = \psi_Q e^{ikz}$, as we could have deduced simply by solving the Helmholtz equation (7.1) for a plane wave.

This basic calculation vindicates three procedures that we have already used. Firstly, it illustrates our interpretation of Fermat's principle in geometrical optics. In the limit of short wavelength, the paths that contribute to the wave field are just those along which the phase is stationary to small variations in path. Our present calculation shows that, because of the tightening of the Cornu spiral as one moves toward a large argument, we need only consider those paths that are separated by less than a few Fresnel lengths at Q. (For a laboratory experiment with light and $z \sim 2$m, a Fresnel length is typically $\sim$ 1mm.)

A second, and related, point is that in computing the diffraction pattern from a more complicated aperture, we need only perform the integral (7.6) in the immediate vicinity of the geometrical optics ray. We can ignore the contribution from the extension of the aperture $\mathcal{Q}$ to meet the "sphere at infinity" even when the wave is unobstructed there. The rapid phase variation makes this contribution sum to zero.

Thirdly, in integrating over the whole area of the wave front at $\mathcal{Q}$, we have summed contributions with increasingly large phase differences that add in such a way that the total has a net extra phase of $\pi/2$, relative to the geometrical optics ray. This phase factor cancels exactly the prefactor $-i$ in the Fresnel-Kirchhoff integral, Eq. (7.6). (This phase factor is unimportant in the limit of geometrical optics.)

Lunar Occultation of a Radio Source

The next simplest case of Fresnel diffraction is the pattern formed by a straight edge. To give a specific example consider a cosmologically distant source of radio waves that is occulted by the moon. If we treat the lunar limb as a straight edge, then as it passes in front of the radio source, a changing diffraction pattern will be sampled by a telescope on earth. In this example, we need only consider the integration of Eq. (7.6) in one dimension and can again use the Cornu spiral. Long before the occultation, the complex wave vector will be given by the arrow from $(-1/2,-1/2)$ to $(1/2,1/2)$ multiplied by $(-i/2)^{1/2}$. (The other prefactor of $(-i/2)^{1/2}$ is absorbed in the integration over the perpendicular direction.) The observed wave amplitude is therefore $\psi_Q e^{ikz}$.

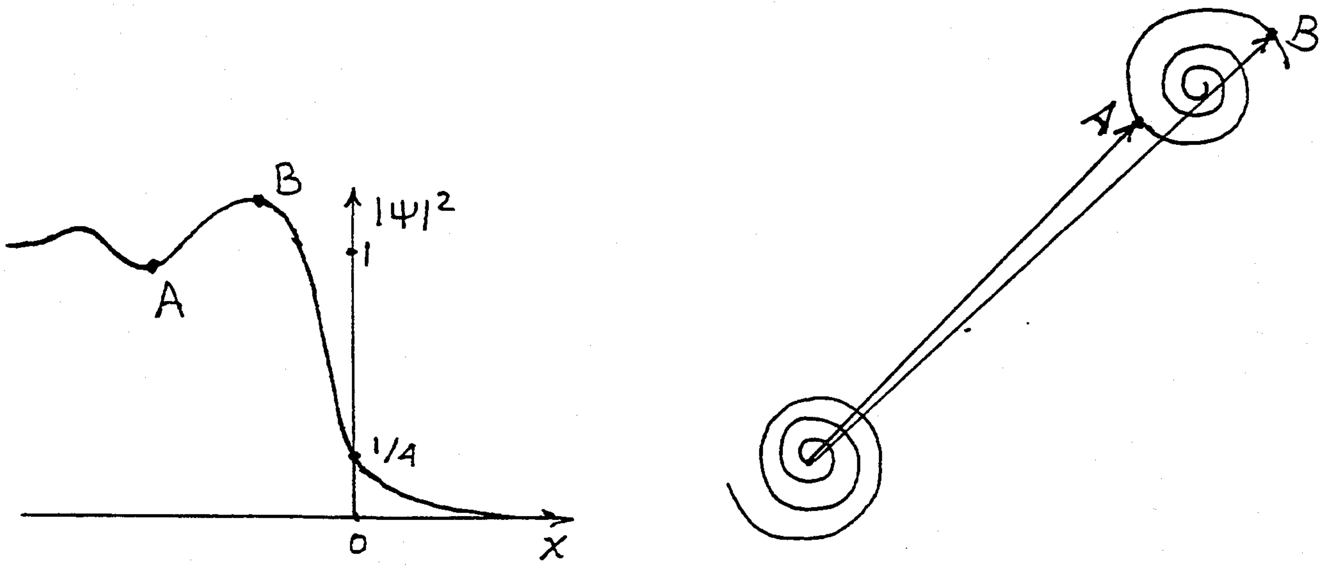


Fig. 7.9 Diffraction pattern formed by a straight edge and graphical interpretation using Cornu Spiral.

Now when the moon starts to occult the radio source, the upper bound on the Fresnel integrals must diminish from $s_{max} = +\infty$, and the complex vector on the Cornu spiral begins to oscillate in length (e.g., from A to B in Fig. 7.9) and in phase. The observed flux will also oscillate, more and more rapidly as geometrical occultation is approached. At the point of geometrical occultation, the complex vector extends from $(-1/2, -1/2)$ to $(0,0)$ and so the observed wave intensity is one quarter the unocculted value. As the occultation proceeds, the length of the complex vector and the observed flux will decrease monotonically to zero, while the phase continues to oscillate.

Historically, diffraction of a radio-source's waves by the moon led to the discovery of quasars—the hyperactive nuclei of distant galaxies. In the early 1960s, a team of British radio observers led by Cyril Hazard knew that the moon would occult a powerful radio source named 3C273, so they set up their telescope to observe the development of the diffraction pattern as the occultation proceded. From the pattern's observed times of ingress (passage into the moon's shadow) and egress (emergence from the moon's shadow), Hazard determined the coordinates of 3C273 on the sky. These coordinates enabled Maarten Schmidt at Palomar to identify the 3C273 optically, and discover (from its optical redshift) that it was surprisingly distant and consequently had an unprecedented luminosity.

In Hazard's occultation measurements, the observing wavelength was $\lambda \sim 0.2$m. Since the moon is roughly $z \sim$400,000km distant, the Fresnel length was about $r_F \sim (\lambda z)^{1/2} \sim$ 10km. The orbital speed of the moon is $u \sim 200$m s^{-1}, so the diffraction pattern took a time $\sim 5r_F/u \sim$ 4min to pass through the telescope.

Circular Apertures

We have shown how the diffraction pattern for a plane wave can be thought of as formed by waves that derive from a patch a few Fresnel lengths in size. This notion can be made quantitatively useful if we re-analyze the unobstructed wave front in circular polar coordinates. Consider a plane wave incident on an aperture $\mathcal{Q}$ and define $\rho \equiv |\mathbf{x}'|/r_F = \frac{1}{2}\sqrt{s^2+t^2}$. Then the phase factor in Eq. (7.19) is $\Delta\phi = \pi\rho^2$ and the the observed wave will thus be given by

$$\begin{aligned}\psi_{\mathcal{P}} &= -i\int_0^\rho 2\pi\rho d\rho e^{i\pi\rho^2}\psi_{\mathcal{Q}}e^{ikz} \\ &= (1-e^{i\pi\rho^2})\psi_{\mathcal{Q}}e^{ikz}\ .\end{aligned} \qquad (7.23)$$

Now, this integral does not appear to converge. We can see what is happening if we sketch an amplitude-and-phase diagram (Fig. 7.10). Adding up the contributions to $\psi_{\mathcal{P}}$ from each annular ring, we see that as we integrate outward from $\rho = 0$, the complex vector has the initial phase retardation of $\pi/2$ but then moves on a semi-circle so that by the time we have integrated out to a radius of r_F, *i.e.*$\rho = 1$, the contribution to the observed wave is $\psi_{\mathcal{P}} = 2\psi_{\mathcal{Q}}$ in phase with the incident wave. Then, when the integration has been extended onward to $2^{1/2}r_F$, $\rho = 2^{1/2}$, the circle has been completed and $\psi_{\mathcal{P}} = 0$! The integral continues on around the same circle as the upper-bound radius is further increased.

Of course, the field must actually have a well-defined value, despite this apparent failure of the integral to converge. To understand how the field becomes well-defined, imagine splitting the aperture $\mathcal{Q}$ up into concentric annular rings, known as *Fresnel half-period*

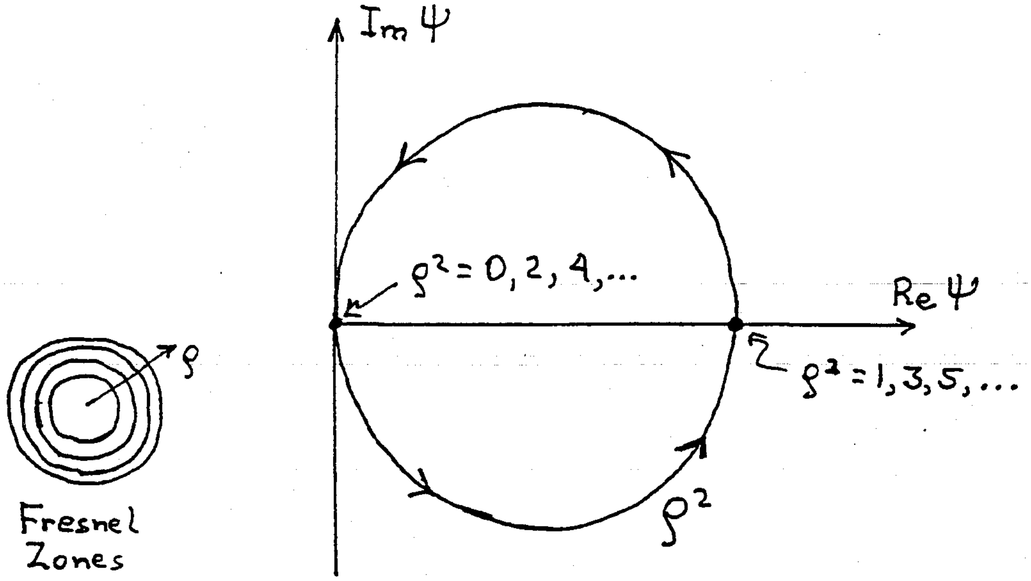


Fig. 7.10 Amplitude-and-phase diagram for an unobstructed plane wave front, decomposed into Fresnel zones.

zones, of radius $n^{1/2} r_F$, where $n = 1, 2, 3 \ldots$. The integral fails to converge because the contribution from each odd-numbered ring cancels that from an adjacent even-numbered ring. However, the thickness of these rings decreases as $n^{-1/2}$, and eventually we must allow for the fact that any physical source of waves will have a finite angular size. The finite size causes different pieces of the source to have their Fresnel rings centered at slightly different points in the aperture plane, and this causes our computation of $\psi_{\mathcal{P}}$ to begin averaging over rings, and the averaging forces the tip of the complex vector to asymptote to the center of the circle in Fig. 7.10. Correspondingly, due to the averaging, the observed intensity asymptoties to $|\psi_{\mathcal{Q}}|^2$.

Although this may not have seemed a particularly wise way to decompose a plane wave front, it does allow a particularly striking experimental verification of our theory of diffraction. Suppose that we fabricate an aperture (called a *zone plate*) in which, for a chosen observation point $\mathcal{P}$ on the optic axis, alternate half-period zones are obscured. Then the wave observed at $\mathcal{P}$ will be the linear sum of several diameters of the circle in Fig. 7.10, and therefore will be far larger than $\psi_{\mathcal{Q}}$. This strong amplification is confined to our chosen spot on the optic axis; most everywhere else, the field's intensity is reduced, thereby conserving energy. Thus, the zone plate behaves like a lens (a "Fresnel lens"). The lens's focal length is $f = kA/2\pi^2$, where A (typically chosen to be a few mm^2 for light) is the area of the first half-period zone.

Zone plates are only good lenses when the radiation is monochromatic, since the focal length is wavelength-dependent, $f \propto \lambda^{-1/2}$. They have the further interesting property that they possess secondary foci, where $3, 5, 7, \ldots$ half-period zones can be observed (Ex. 7.5).

****

EXERCISES

Exercise 7.5, Problem: Zone Plate

(a) Use an amplitude-and-phase diagram to explain why a zone plate has secondary foci at distances of $f/3, f/5, f/7 \ldots$

(b) An opaque, perfectly circular disk of diameter D is placed perpendicular to an incoming plane wave. Show that, at distances r such that $r_F \ll D$, the disk casts a rather sharp shadow, but at the precise center of the shadow there should be a bright spot. How bright?

Exercise 7.6, Problem: Seeing in the atmosphere.

Stars viewed through the atmosphere appear to have angular diameters of order an arc second and to exhibit large amplitude fluctuations of flux with characteristic frequencies that can be as high as 100Hz. Both of these phenomena are a consequence of irregular variations in the refractive index of the atmosphere. An elementary model of this effect consists of a thin phase-changing screen about a km above the ground on which the rms phase variation is $\Delta\phi \gtrsim 1$ and the characteristic spatial scale is a.

(a) Explain why the rays will be irregularly deflected through a scattering angle $\Delta\theta \sim (\lambda/a)\Delta\phi$. Strong intensity variation requires that several rays deriving from points on the screen separated by more than a, combine at each point on the ground. These rays combine to create a diffraction pattern on the ground with scale b.

(b) Show that the Fresnel length in the screen is $\sim (ab)^{1/2}$. Now the time variation arises because winds in the upper atmosphere with speeds $u \sim 30\text{m s}^{-1}$ blow the irregularities and the diffraction pattern past the observer. Use this information to estimate the Fresnel length, r_F, the atmospheric fluctuation scale size a, and the rms phase variation $\Delta\phi$. Do you think the assumptions of this model are well satisfied?

Exercise 7.7, Problem: Spy Satellites

Telescopes can also look down through the same atmospheric irregularities as those discussed in the previous example. In what important respects will the optics differ from that for telescopes looking upward?

****

7.5 Fourier Optics

We have developed a linear theory of wave optics which has allowed us to calculate diffraction patterns in the Fraunhofer and Fresnel limiting regimes. That these calculations agree with laboratory measurements provides some vindication of the theory and the assumptions implicit in it. We now turn to the practical applications of these ideas, specifically to the acquisition and processing of images by instruments operating throughout the electromagnetic spectrum. Although the conceptual framework and mathematical machinery for such image processing were developed over a century ago, it is really only over the past thirty years that these techniques have been widely exploited. Part of the reason is the growing realization of the great similarities between optics and communication theory. In other words, a microscope is simply an image processing device. Moreover, the development of electronic computation has led to enormous strides and computers are now seen as extensions of optical devices. It is a matter of convenience, economics and practicality to decide which parts of the image processing are carried out with mirrors, lenses, etc, and which parts are performed numerically.

We have computed the Fraunhofer diffraction pattern of a circular aperture, for example the mirror or objective lens of a telescope, and have shown how the power diffracted through angles $\theta > \theta_A$ may make it difficult to resolve two nearby point sources. One technique for improving the resolution might be to attenuate the incident radiation at the aperture in such a way that it has a Gaussian profile. Its Fourier transform then would also be a Gaussian, and thus would not exhibit amplitude oscillations ("fringes"). However, such a Gaussian-producing attenuation is difficult in practice, and it turns out—as we shall see–that there are easier options.

Coherent Illumination

In order to discuss the easier options, we must first introduce another distinction, the meaning of which should become clearer in Chap. 8. If the radiation that arrives at the input of an optical system derives from a single source, *e.g.*a point source that has been collimated into a parallel beam by a converging lens, then the radiation is best described by its complex amplitude ψ (as we are doing in this chapter). An example might be a biological specimen on a microscope slide, illuminated externally, for which the phases of the waves leaving different parts of the slide are strongly correlated with each other. This is called *coherent illumination.* If, by contrast, the source is self luminous, with the atoms or molecules in its different parts radiating independently, for example a cluster of stars, then the phases of the radiation from different parts are uncorrelated, and it is the intensity of the radiation, not the complex amplitude, that obeys well-defined (non-probabilistic) evolution laws. This is called *incoherent illumination.* We shall develop Fourier optics for a coherently illuminated source. A parallel theory, with a similar vocabulary can be developed for incoherent sources; and some of the foundations for it will be laid in Chap. 8.

In our treatment of paraxial geometrical optics (Sec. 6.4), we showed how it is possible to regard a group of optical elements as a sequence of linear devices and relate the output rays to the input by linear operators, i.e. matrices. This chapter's theory of diffraction is also linear and so a similar approach can be followed. As in Sec. 6.4, we will restrict attention to small angles relative to some optic axis. We shall describe the wave field at some location z_j along the optic axis by the function $\psi_j(\mathbf{x})$, where $\mathbf{x}$ is a two dimensional vector perpendicular to the optic axis. If we consider a single linear optical device, then we can relate the output field ψ_2 at z_2 to the input, ψ_1 at z_1 using a Greens' function denoted $P_{21}(\mathbf{x}_2, \mathbf{x}_1)$:

$$\psi_2(\mathbf{x}_2) = \int P_{21}(\mathbf{x}_2, \mathbf{x}_1) d^2x_1 \psi_1 \ . \tag{7.24}$$

If ψ_1 were a δ-function, then the output would be simply given by the function P_{21}, after normalization. For this reason, P_{21} is usually known as the *Point Spread Function.* Alternatively, we can think of it as a propagator. If we now combine two optical devices sequentially, so the output of the first system ψ_2 is the input of the second, ψ_3, then the point spread functions combine in the natural manner of any linear propagator to give a total point spread function

$$P_{31}(\mathbf{x}_3, \mathbf{x}_1) = \int P_{32}(\mathbf{x}_3, \mathbf{x}_2) d^2x_2 P_{21}(\mathbf{x}_2, \mathbf{x}_1) \ . \tag{7.25}$$

Point Spread Functions

Just as the simplest matrix for paraxial, geometric-optics propagation is that for free propagation through some distance d, so also the simplest point spread function is that for free propagation. From Eq. (7.19) we see that it is given by

$$P_{21} = \frac{-ik}{2\pi d} e^{ikd} \exp\left(\frac{ik(\mathbf{x}_1 - \mathbf{x}_2)^2}{2d}\right) , \tag{7.26}$$

where $d = z_2 = z_1$ is the distance of propagation along the optic axis. Note that this P_{21} depends upon only on $\mathbf{x}_1 - \mathbf{x}_2$ and not on $\mathbf{x}_1$ or $\mathbf{x}_2$ individually, as it should because there is translational invariance in the $\mathbf{x}_1, \mathbf{x}_2$ planes.

A thin lens adds or subtracts an extra phase $\Delta\phi$ to the wave, and $\Delta\phi$ depends quadratically on distance from the optic axis ($|\mathbf{x}|$), so that the angle of deflection, which is proportional to the gradient of the phase, will depend linearly on $\mathbf{x}$. Correspondingly, the point-spread function for a thin lens is

$$P_{21} = \exp\left(\frac{-ik|\mathbf{x}_1|^2}{2f}\right) \delta(\mathbf{x}_2 - \mathbf{x}_1) \tag{7.27}$$

where f is the focal length, positive for a converging lens and negative for a diverging lens.

Abbé Theory

We can use these two point spread functions to give a wave description of the production of images by a single converging lens, in parallel to the geometric-optics description of Fig. 6.3. We shall do this in two stages. First, we shall propagate the wave from the source plane S a distance u in front of the lens, through the lens L, to its focal plane F a distance f behind the lens (Fig. 7.11). Then we shall propagate the wave a further

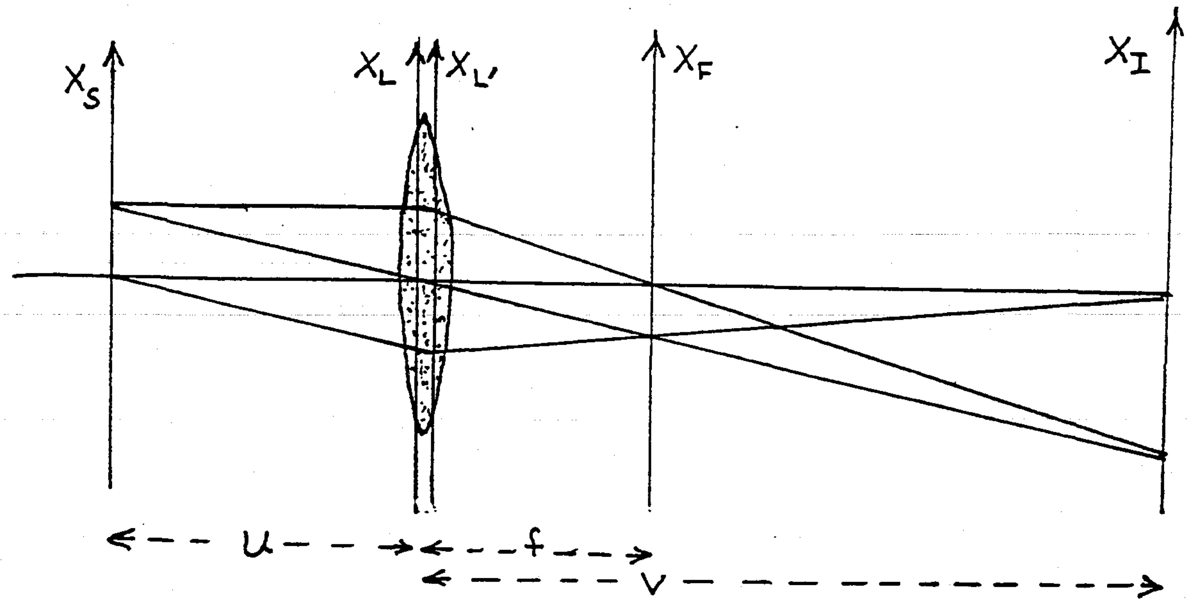


Fig. 7.11 Wave theory of a single converging lens.

distance $v - f$ from the focal plane to the image plane. We know from geometrical optics that $v = fu/(f-u)$ [Eq. (6.41)]. Using equations (7.25), (7.26), (7.27), we obtain

$$\begin{aligned} P_{FS} &= \int P_{FL'} d^2x'_L P_{L'L} d^2x_L P_{LS} \\ &= \int \frac{ik}{2\pi f} e^{ikf} \exp\left(\frac{ik(\mathbf{x}_F - \mathbf{x}'_L)^2}{2f}\right) d^2x_{L'} \delta(\mathbf{x}_{L'} - \mathbf{x}_L) \exp\left(\frac{-ik|\mathbf{x}_L|^2}{2f}\right) d^2x_L \\ &\qquad\qquad \times \frac{-ik}{2\pi u} e^{iku} \exp\left(\frac{ik(\mathbf{x}_S - \mathbf{x}_L)^2}{2u}\right) \\ &= \frac{-ik}{2\pi f} e^{ik(f+u)} \exp\left(-\frac{ikx_F^2}{2(v-f)}\right) \exp\left(-\frac{ik\mathbf{x}_F \cdot \mathbf{x}_S}{f}\right), \end{aligned} \tag{7.28}$$

where we have extended all integrations to $\pm\infty$ and have used the values of the Fresnel integrals at infinity, $S(\pm\infty) = \pm(1+i)/2$ to get the expression on the last line. The wave in the focal plane is given by

$$\begin{aligned} \psi_F(\mathbf{x}_F) &= \int P_{FS} d^2x_S \psi_S(\mathbf{x}_S) \\ &= -\frac{ik}{2\pi f} e^{ik(f+u)} \exp\left(-\frac{ikx_F^2}{2(v-f)}\right) \tilde{\psi}_S(\mathbf{x}_F/f) \end{aligned} \tag{7.29}$$

where

$$\tilde{\psi}_S(\boldsymbol{\theta}) = \int d^2x_S \psi_S(\mathbf{x}_S) e^{-i\boldsymbol{\theta}\cdot\mathbf{x}_S} . \tag{7.30}$$

Thus, we have shown that the field in the back focal plane is, apart from an unimportant phase factor, proportional to the Fourier transform of the field in the source plane. It is therefore the Fraunhofer diffraction pattern of the input wave. That this has to be the case can be understood from Fig. 7.11. The focal plane F is where the converging lens brings parallel rays from the source plane to a focus. By doing so, the lens in effect brings in from "infinity" the Fraunhofer diffraction pattern of the source, and places it into the focal plane.

It now remains to propagate the final distance from the focal plane to the image plane. We do so with the free-propagation point-spread function of Eq. (7.26):

$$\begin{aligned}\psi_I &= \int P_{IF} d^2x_F \psi_F \\ &= -\left(\frac{u}{v}\right) e^{ik(u+v)} \exp\left(\frac{ikx_I^2}{2(v-f)}\right) \psi_S(\mathbf{x}_S = -\mathbf{x}_I u/v) .\end{aligned} \tag{7.31}$$

We have therefore verified that, again ignoring a phase factor, the wave in the image plane is just a magnified version of the wave in the source plane, as we might have expected from geometrical optics. In words, the lens acts by taking the Fourier transform of the source and then takes the Fourier transform again to recover the source structure.

The focal plane is a convenient place to process the image by altering its Fourier transform—a process known as *spatial filtering*. One simple example is the *low-pass filter* in which a small circular aperture or "stop" is introduced into the focal plane, thereby allowing only the low-order spatial Fourier components to be transmitted to the image plane. This will obviously lead to considerable smoothing of the wave. An application is to the output beam from a laser (Chap. 9), which ought to be smooth but has high spatial frequency structure on account of noise and imperfections in the optics. A low-pass filter can be used to clean the beam. In the language of Fourier transforms, if we multiply the transform of the source, in the back focal plane, by a small-diameter circular aperture function, we will thereby convolve the image with a broad Airy-disk smoothing function. Conversely, we can exclude the low spatial frequencies with a high-pass filter, e.g. by placing an opaque circular disk in the focal plane, centered on the optic axis. This will have the effect of accentuating boundaries and discontinuities in the source and can be used to highlight features where the gradient of the brightness is considerable. Another type of filter is used when the image is pixellated and thus has unwanted structure with wavelength equal to pixel size: a narrow range of frequencies centered around this spatial frequency is removed by putting an appropriate filter in the back focal plane.

Phase Contrast Microscopy

"Phase contrast microscopy" is a very useful technique for studying small objects, such as transparent biological specimens, that modify the phase of coherent illuminating light, but not its amplitude. Let us suppose that the phase change in the specimen, $\phi(\mathbf{x})$, is small, as often is the case for biological specimens. We can therefore write the field just after it passes through the specimen as

$$\psi_S(\mathbf{x}) = e^{i\phi(\mathbf{x})} \simeq 1 + i\phi(\mathbf{x}_S) \, ; \tag{7.32}$$

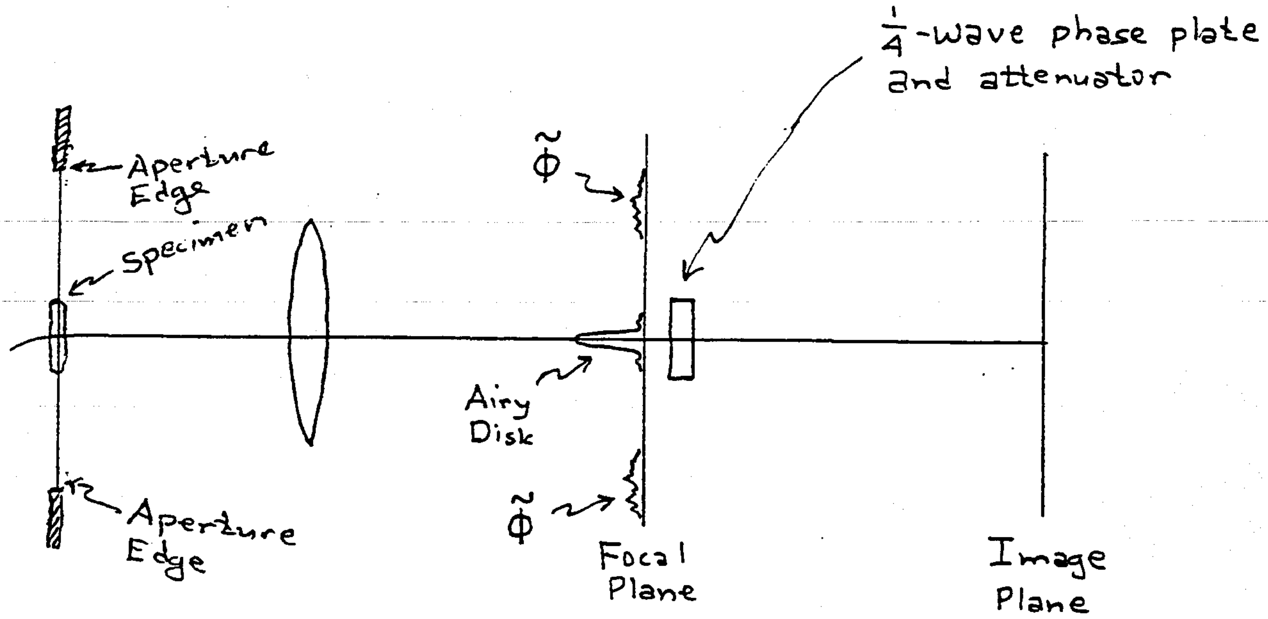


Fig. 7.12 Schematic Phase Contrast Microscope.

see Fig. 7.12. The intensity is not modulated, and therefore the effect of the specimen on the wave is very hard to observe unless one is clever.

The wave in the focal plane is effectively the sum of the Fourier transform of the aperture function, *i.e.*an Airy function of very small width, and the transform of the phase function

$$\psi_F \sim \text{jinc}\left(\frac{kDx_f}{2f}\right) + i\tilde{\phi}\left(\frac{kx_F}{f}\right) . \tag{7.33}$$

If a low pass filter is used to remove the Airy disk completely then the remaining wave in the image plane will be essentially ϕ magnified by v/u. The flux will still be quadratic in the phase and so the contrast in the image will be small. A better technique is to phase shift the Airy disk in the focal plane by $\pm\pi/2$ so that the two terms in Eq. (7.33) are in phase. The intensity variations $[\propto (1 \pm \phi)^2]$ will now be linear in the phase ϕ. An even better procedure is to attenuate the Airy disk until its amplitude is comparable with the rms value of ϕ and also phase shift it by $\pm\pi/2$. This will maximise the contrast in the final image. Analogous techniques are used in communications to inter-convert amplitude-modulated and phase-modulated signals.

Gaussian Beams

The mathematical techniques of Fourier optics enables us to analyze the structure and propagation of light beams that have Gaussian profiles. (Such Gaussian beams are the natural output of ideal lasers, they are the real output of spatially filtered lasers, and they are widely used for optical communications, interferometry and other practical applications. Moreover, they are the closest one can come in the real world of wave optics to the idealization of a geometrical-optics pencil beam.)

Consider a beam that is precisely plane fronted, with a Gaussian profile, at some location z_0 on the optic axis:

$$\psi_0 = \exp\left(\frac{-x^2}{\sigma_0^2}\right). \tag{7.34}$$

The form of this same wave at a distance z further down the optic axis can be computed by folding this ψ_0 into the point spread function (7.26) (with the distance d replaced by z). The result is

$$\psi_z = \frac{1}{(1+z^2/z_0^2)^{1/2}} \exp\left(\frac{-x^2}{\sigma_0^2(1+z^2/z_0^2)}\right) \exp\left[i\left(\frac{kx^2}{2z(1+z_0^2/z^2)} - \tan^{-1}\frac{z}{z_0} + kz\right)\right], \tag{7.35}$$

where

$$z_0 = k\sigma_0^2 = 2\pi\sigma_0^2/\lambda. \tag{7.36}$$

Formula (7.35) for the freely propagating beam is valid for negative z as well as positive. Notice that the beam's energy is concentrated in a region with diameter (beam size)

$$\sigma_z = \sigma_0(1+z^2/z_0^2)^{1/2}. \tag{7.37}$$

This beam size is a minimum at $z = 0$ (the beam's waist), and increases away from there in either direction.

The Gaussian beam's form (7.36) near some arbitrary location z is fully characterized by three parameters: the wavelength $\lambda = 2\pi/k$, the distance z to the waist, and the waist size σ_0 [from which the local beam size σ_z can be computed through Eq. (7.37)]. At location z, the beam's wave fronts (surfaces of constant phase) have radius of curvature $R_z = z(1 + z_0^2/z^2)$. The radius of curvature is infinite at the waist. Near the waist, in the Fresnel region, $(\lambda z)^{1/2} \ll \sigma_0$, it decreases with distance as $R_z \simeq z_0^2/z$. It reaches a minimum value at the boundary between the Fresnel and the Fraunhofer regions, and it then begins to increase as $R_z \propto z$. Correspondingly, in the Fresnel region the beam size is nearly constant, $\sigma_z \simeq \sigma_0$, while in the Fraunhofer region it increases linearly with distance, $\sigma_z \simeq \sigma_0 z/z_0$. These are just the behaviors that one should expect from the uncertainty principle analysis at the end of Sec. 7.2.

It is easy to compute the effects of a thin lens on a Gaussian beam by folding the ψ_z at the lens's location into the lens point spread function (7.27). The result is a phase change that preserves the general Gaussian form of the wave, but alters the distance to the waist and the waist size. Thus, by judicious placement of lenses (or, equally well curved mirrors), and judicious choices of their focal lengths, one can tailor the parameters of a Gaussian beam to fit whatever optical device one is working with. For example, if one wants to send a Gaussian beam into a self-focusing optical fiber, one should place its waist at the entrance to the fiber, and adjust its waist size there to coincide with that of the fiber's Gaussian mode of propagation (the mode analyzed in Ex. 7.8). The beam will then enter the fiber smoothly, and will propagate steadily along the fiber, with the effects of the transversely varying index of refraction continually compensating for the effects of diffraction so as to keep the phase fronts flat and the waist size constant.

****

EXERCISES

Exercise 7.8, Problem: Guided Gaussian Beams

Consider a self-focusing optical fiber discussed in the previous chapter in which the refractive index is

$$n(\mathbf{x}) = n_0(1-\alpha^2 r^2)^{1/2} ,$$

where $r = |\mathbf{x}|$.

(a) Write down the Helmholtz equation in cylindrical polar coordinates and seek an axisymmetric mode for which $\psi = R(r)Z(z)$, where R, Z are functions to be determined and z measures distance along the fiber. In particular show that there exists a mode with a Gaussian radial profile that propagates along the fiber without spreading.

(b) Compute the group and phase velocities along the fiber for this mode.

****

7.6 Diffraction at a Caustic

In Sec. 6.6, we described how caustics can be formed in general in the geometrical optics limit—e.g., on the bottom of a swimming pool when the water's surface is randomly rippled, or behind a gravitational lens. We chose as an example a simple phase changing screen illuminated by a point source and observed from some fixed distance r, and we showed how a pair of images would merge as the transverse distance x of the observer from the caustic decreases to zero. We expanded the phase in a Taylor series $\phi(s,x) = as^3/3 - bxs$, where the coefficients a, b are constant and s is a transverse coordinate in the screen (*cf.* Fig. 6.10). We were then able to show that the magnification of the images diverged $\propto x^{-1/2}$ [Eq. (6.64)], as the caustic was approached. This raised the question of what happens when we take into account the finite wavelength of the wave.

We are now in a position to answer this question. We simply use the Helmholtz-Kirchhoff integral (7.6) to write the expression for the amplitude measured at position x in the form

$$\psi(x) \propto \frac{1}{\lambda r}\int ds e^{i\phi(s,x)} , \tag{7.38}$$

ignoring multiplicative constants and constant phase factors. The phase ϕ varies rapidly with s at large $|s|$, so we can treat the limits of integration as $\pm\infty$. The integral turns out to be the Airy function

$$\int_{-\infty}^{\infty} ds \cos(as^3/3 - bxs) = \frac{2\pi}{a^{1/3}}\mathrm{Ai}(-bx/a^{1/3}) . \tag{7.39}$$

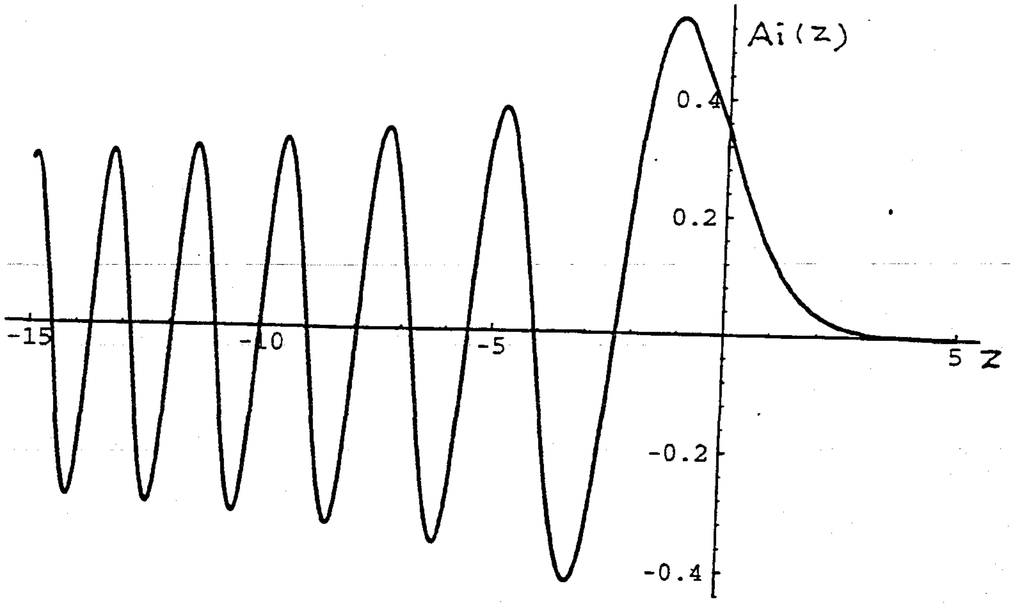


Fig. 7.13 The Airy Function Ai(z) describing diffraction at a caustic. The argument is $z = -bx/a^{1/3}$ where x is distance from the caustic and a, b are constants.

Ai(z) is displayed in Fig. 7.13.

The asymptotic behavior of Ai(z) is

$$\begin{aligned} \mathrm{Ai}(z) &\sim \pi^{-1/2} z^{-1/4} \sin(2z^{3/2}/3 + \pi/4), \quad z \to -\infty \\ &\sim \frac{e^{-2z^{3/2}/3}}{2\pi^{1/2} z^{1/4}}, \quad z \to \infty \, . \end{aligned} \tag{7.40}$$

We see that the amplitude ψ remains finite as the caustic is approached instead of diverging as in the geometrical optics limit. Furthermore, for $x > 0$ (left part of Fig. 7.13), where an observer sees two geometrical optics images, the envelope of ψ diminishes $\propto x^{-1/4}$, so that the intensity $|\psi|^2$ decreases $\propto x^{-1/2}$ just as in the geometrical optics limit. The peak magnification is $\propto a^{-2/3}$. What is actually seen is a series of bands alternating dark and light with spacing calculable using $\Delta(2z^{3/2}/3) = \pi$ or $\Delta x \propto x^{-1/2}$. At sufficient distance from the caustic, it will not be possible to resolve these bands and a uniform illumination of average intensity will be observed. In other words, we have recovered the geometrical optics limit. The scalings derived above, just like the geometrical optics scalings are a universal property of this type of caustic (the simplest caustic of all, the "fold").

There is a helpful analogy, familiar from quantum mechanics. Consider a particle in a harmonic potential well in a very excited state. Its wave function is given in the usual way using Hermite polynomials of large order. Close to the classical turning point, these functions change from being oscillatory to an exponential decay, just like the Airy function (and if we were to expand about the turning point, we would recover Airy functions). What is happening, of course is that the probability density of finding the particle close to its turning point diverges because it is moving very slowly there, and the oscillations are due to interference between waves associated with the particle moving in opposite directions. If we just consider the motions of photons parallel to the aperture then we have essentially the same problem here; the oscillations are associated with interference of the waves associated with the motions of the photons in two beams, one from each of the geometric-optics images. This is our first illustration of the formation of large contrast interference fringes when only a few beams are combined. We shall meet other examples of such interference in the following chapter.

****

EXERCISES

Exercise 7.9, Wavelength scaling at a caustic

Assume that the phase variation introduced at the screen in Fig. 6.9 is non-dispersive so that the $\phi(s, x)$ in Eq. (7.38) is $\phi \propto \lambda^{-1}$. Show that the peak magnification of the interference fringes at the caustic scales with wavelength $\propto \lambda^{-4/3}$. Also show that the spacing of the fringes at a given observing position is proportional to the wavelength.

****

BIBLIOGRAPHY

Some useful references relevant to the topics of this chapter are:

Berry, M. V. & Upstill, C. 1980. "" *Prog. Optics*, **18**, 257.

Born, M. & Wolf, E. 1975. *Principles of Optics*, Oxford: Pergamon.

Goodman, J. W. 19??. *Introduction to Fourier Optics*, New York: McGraw-Hill.

Hecht, E. 1989. *Optics*, New York: Addison Wesley.

Longhurst, R. S. 1973. *Geometrical and Physical Optics*, London: Longmans.

Welford, W. T. 1988. *Optics*, Oxford: Oxford University Press.

REF 1

PROPERTY OF GRAVITATIONAL PHYSICS

LASERS

Anthony E. Siegman

Professor of Electrical Engineering
Stanford University

PROPERTY OF GRAVITATIONAL PHYSICS

PROPERTY OF GRAVITATIONAL PHYSICS

University Science Books
Mill Valley, California

University Science Books
20 Edgehill Road
Mill Valley, CA 94941

Manuscript Editor: Aidan Kelly
Designer: Robert Ishi
Production: Miller/Scheier Associates, Palo Alto, CA
TEXpert: Laura Poplin
Printer and Binder: The Maple-Vail Book Manufacturing Group

Copyright © 1986 by University Science Books
Reproduction or translation of any part of this work
beyond that permitted by Sections 107 or 108 of the
1976 United States Copyright Act without the permission of
the copyright owner is unlawful. Requests for permission
or further information should be addressed to
the Permissions Department, University Science Books.

Library of Congress Catalog Card Number: 86-050346

ISBN 0-935702-11-5

Printed in the United States of America

This manuscript was prepared at Stanford University using the text editing facilities of the Context and Sierra DEC-20 computers and Professor Donald Knuth's TEX typesetting system. Camera-ready copy was printed on an Autologic APS-μ5 phototypesetter.

10 9 8 7 6 5 4 3 2 1

17

PHYSICAL PROPERTIES OF GAUSSIAN BEAMS

The previous chapter developed the analytical tools needed for calculating optical-beam propagation in free space. We also need to have, however, a physical and intuitive understanding of the propagation of real optical beams—an understanding which the next two chapters attempt to develop.

In particular, the Hermite-gaussian or Laguerre-gaussian modes which we introduced in the previous chapter are both mathematically convenient, and also provide very good (though not quite exact) approximations to the transverse modes of stable laser resonators with finite diameter mirrors. Gaussian or quasi gaussian beams are therefore very widely used in analyzing laser problems and related optical systems. A good physical as well as mathematical understanding of gaussian beam properties is particularly important. In this chapter we thus review most of the important physical properties of ideal gaussian optical beams in free space.

17.1 GAUSSIAN BEAM PROPAGATION

We first look in this section at what the analytic expressions for a lowest-order gaussian beam imply physically in terms of aperture transmission, collimated beam distances, far-field angular beam spread, and other practical aspects of gaussian beam propagation.

Analytical Expressions

Let us assume a lowest-order gaussian beam characterized by a spot size w_0 and a planar wavefront $R_0 = \infty$ in the transverse dimension, at a reference plane which for simplicity we take to be $z = 0$. This plane will henceforth be known for obvious reasons as the *beam waist*, as in Figure 17.1.

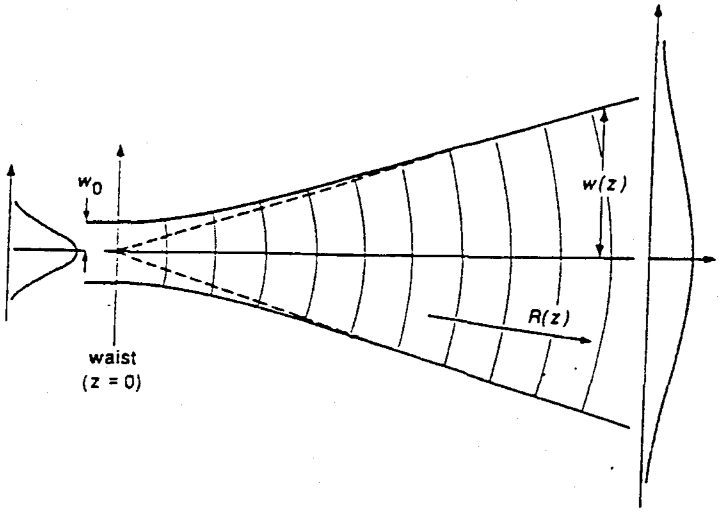


FIGURE 17.1
Notation for a lowest-order gaussian beam diverging away from its waist.

The normalized field pattern of this gaussian beam at any other plane z will then be given by

$$\begin{aligned}\tilde{u}(x,y,z) &= \left(\frac{2}{\pi}\right)^{1/2} \frac{\tilde{q}_0}{w_0\tilde{q}(z)} \exp\left[-jkz - jk\frac{x^2+y^2}{2\tilde{q}(z)}\right] \\ &= \left(\frac{2}{\pi}\right)^{1/2} \frac{\exp[-jkz + j\psi(z)]}{w(z)} \exp\left[-\frac{x^2+y^2}{w^2(z)} - jk\frac{x^2+y^2}{2R(z)}\right],\end{aligned} \tag{1}$$

where the complex radius of curvature $\tilde{q}(z)$ is related to the spot size $w(z)$ and the radius of curvature $R(z)$ at any plane z by the definition

$$\frac{1}{\tilde{q}(z)} \equiv \frac{1}{R(z)} - j\frac{\lambda}{\pi w^2(z)}. \tag{2}$$

In free space this parameter obeys the propagation law

$$\tilde{q}(z) = \tilde{q}_0 + z = z + jz_R, \tag{3}$$

with the initial value

$$\tilde{q}_0 = j\frac{\pi w_0^2}{\lambda} = jz_R. \tag{4}$$

Note that the value of λ in these formulas is always the wavelength of the radiation *in the medium in which the beam is propagating.*

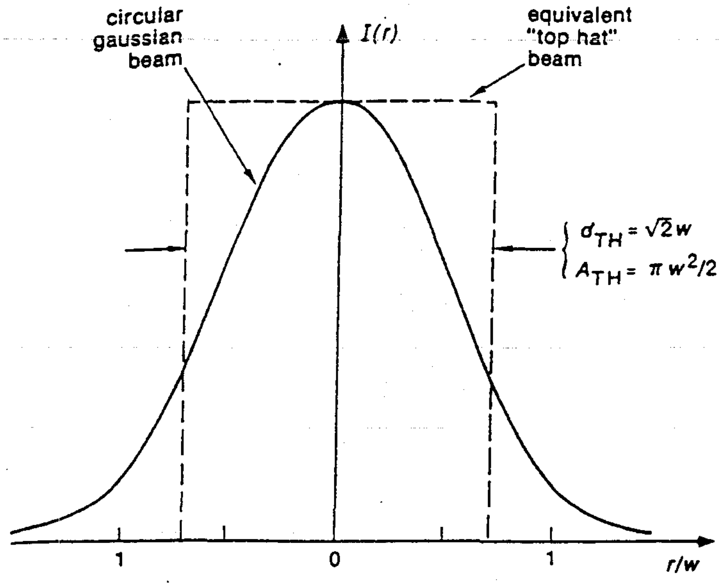


FIGURE 17.2
The equivalent "top hat" radius for a cylindrical gaussian beam.

All the important parameters of this gaussian beam can then be related to the waist spot size w_0 and the ratio z/z_R by the formulas

$$w(z) = w_0\sqrt{1+\left(\frac{z}{z_R}\right)^2},$$

$$R(z) = z + \frac{z_R^2}{z}, \qquad (5)$$

$$\psi(z) = \tan^{-1}\left(\frac{z}{z_R}\right).$$

In other words, the field pattern along the entire gaussian beam is characterized entirely by the single parameter w_0 (or $\tilde{q}_0$, or z_R) at the beam waist, plus the wavelength λ in the medium.

Aperture Transmission

Before exploring the free-space propagation properties of an ideal gaussian beam, we might consider briefly the vignetting effects of the finite apertures that will be present in any real optical system. The intensity of a gaussian beam falls off very rapidly with radius beyond the spot size w. How large must a practical aperture be before its truncation effects on a gaussian beam become negligible?

Suppose we define the total power in an optical beam as $P = \int\int |\tilde{u}|^2\, dA$ where dA integrates over the cross-sectional area. The radial intensity variation of a gaussian beam with spot size w is then given by

$$I(r) = \frac{2P}{\pi w^2}\, e^{-2r^2/w^2}. \qquad (6)$$

The effective diameter and area of a uniform cylindrical beam (a "top hat beam") with the same peak intensity and total power as a cylindrical gaussian beam will

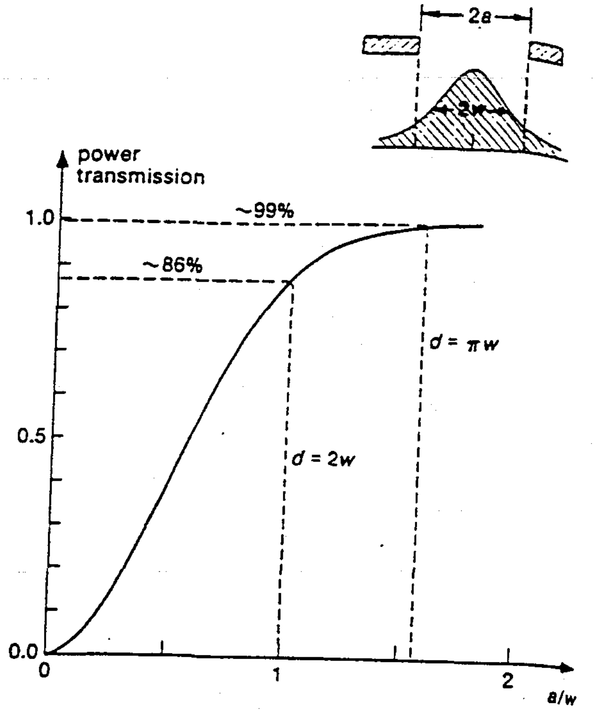


FIGURE 17.3
Power transmission of a cylindrical gaussian beam through a circular aperture.

then be

$$d_{\mathrm{TH}} = \sqrt{2}w \qquad \text{and} \qquad A_{\mathrm{TH}} = \frac{\pi w^2}{2} \tag{7}$$

as shown in Figure 17.2.

An aperture significantly larger than this will be needed, however, to pass a real gaussian beam of spot size w without serious clipping of the beam skirts. The fractional power transfer, for example, for a gaussian beam of spot size w passing through a centered circular aperture of diameter $2a$, as in Figure 17.3, will be given by

$$\text{power transmission} = \frac{2}{\pi w^2}\int_0^a 2\pi r e^{-2r^2/w^2}\,dr = 1 - e^{-2a^2/w^2}. \tag{8}$$

This figure plots this transmission versus aperture radius a normalized to spot size w. An aperture with radius $a = w$ transmits $\approx$ 86% of the total power in the gaussian beam. We will refer to this as the $1/e$ or 86% criterion for aperture size.

A more useful rule of thumb to remember, however, is that an aperture with radius $a = (\pi/2)w$, or diameter $d = \pi w$, will pass just over 99% of the gaussian beam power. We will often use this as a practical design criterion for laser beam apertures, and will refer to it as the "$d = \pi w$" or 99% criterion. (A criterion of $d = 3w$ which gives $\approx$ 98.9% transmission would obviously serve equally well.) Figure 17.4 illustrates just where some of these significant diameters for a gaussian beam will fall on the gaussian beam profile.

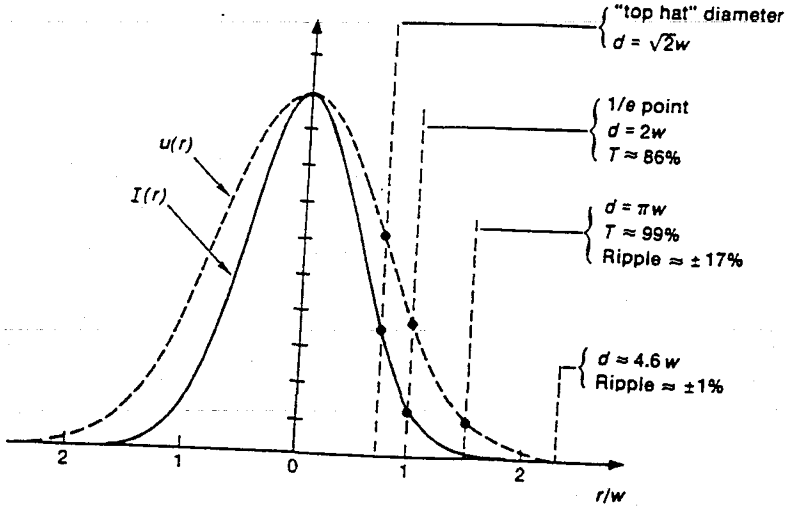


FIGURE 17.4
Significant diameters for hard-edged truncation of a cylindrical gaussian beam. Note that the $d = \pi w$ criterion gives 99% power transmission, but also ±17% intensity ripples and intensity reduction in the near and far fields.

Aperture Diffraction Effects

Optical designers should take note, however, that sharp-edged apertures, especially circular apertures, even though they may cut off only a very small fraction of the total power in an optical beam, will also produce aperture diffraction effects like those shown in Figure 17.5, which will significantly distort the intensity pattern of the transmitted beam in both the near-field (Fresnel) and far-field (Fraunhofer) regions.

We will show in the following chapter, for example, that the diffraction effects on an ideal gaussian beam of a sharp-edged circular aperture even as large as the $d = \pi w$ criterion will cause near-field diffraction ripples with an intensity variation $\Delta I/I \approx \pm 17\%$ in the near field, along with a peak intensity reduction of $\approx 17\%$ on axis in the far field. We have to enlarge the aperture to $d \approx 4.6w$ to get down to ±1% diffraction ripple effects from a sharp-edged circular aperture.

Beam Collimation: The Rayleigh Range and the Confocal Parameter

Another important question is how rapidly an ideal gaussian beam will expand due to diffraction spreading as it propagates away from the waist region or, in practical terms, over how long a distance can we propagate a collimated gaussian beam before it begins to spread significantly?

The variation of the beam spot size $w(z)$ with distance as given by Equation 17.5 is plotted in Figure 17.6 for two different waist spot sizes w_{01} and $w_{02} > w_{01}$, with the transverse scale greatly enlarged. The primary point is that as the input spot size w_0 at the waist is made smaller, the beam expands more rapidly due to diffraction; remains collimated over a shorter distance in the near field; and diverges at a larger beam angle in the far field.

In particular, the distance which the beam travels from the waist before the beam diameter increases by $\sqrt{2}$, or before the beam area doubles, is given simply

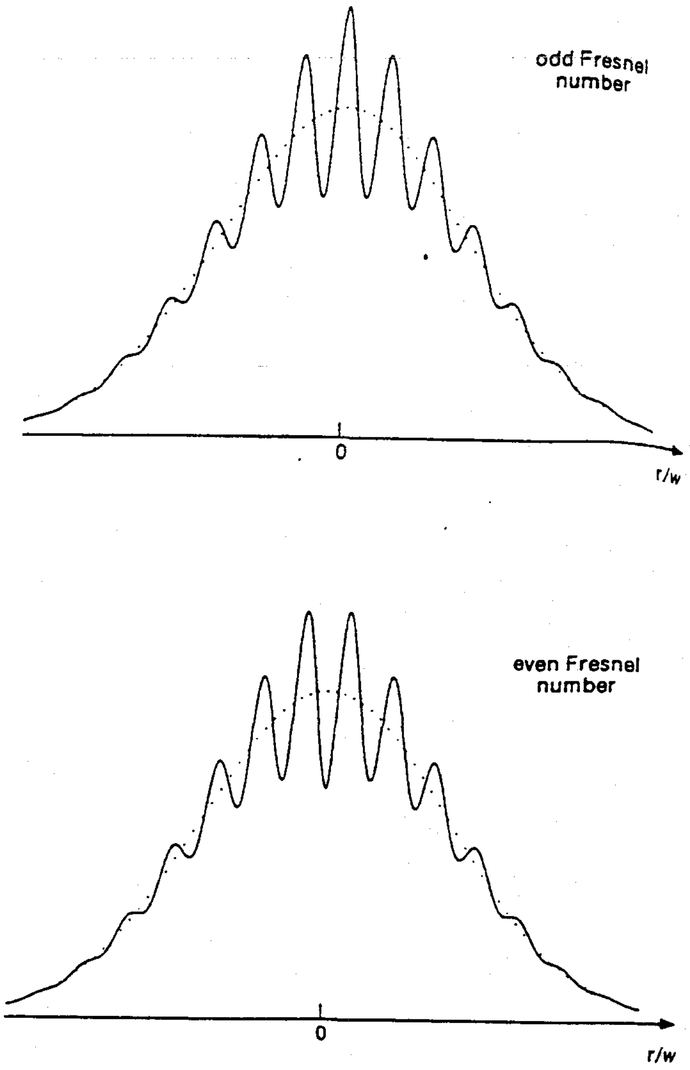


FIGURE 17.5
Near-field Fresnel-diffraction ripples produced by truncation of a gaussian beam.

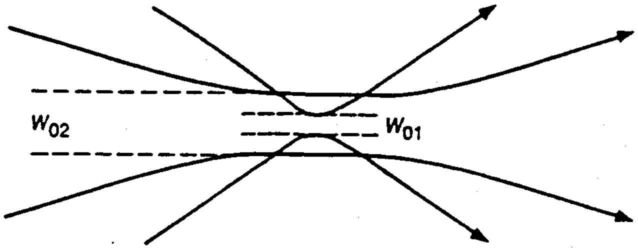


FIGURE 17.6
Diffraction spreading of two gaussian beams with different spot sizes at the waist.

by the parameter

$$z = z_R \equiv \frac{\pi w_0^2}{\lambda} = \text{“Rayleigh range.”} \tag{9}$$

The term *Rayleigh range* is sometimes used in antenna theory to describe the distance $z \approx d^2/\lambda$ that a collimated beam travels from an antenna of aperture diameter d (assuming $d \gg \lambda$) before the beam begins to diverge significantly. We have

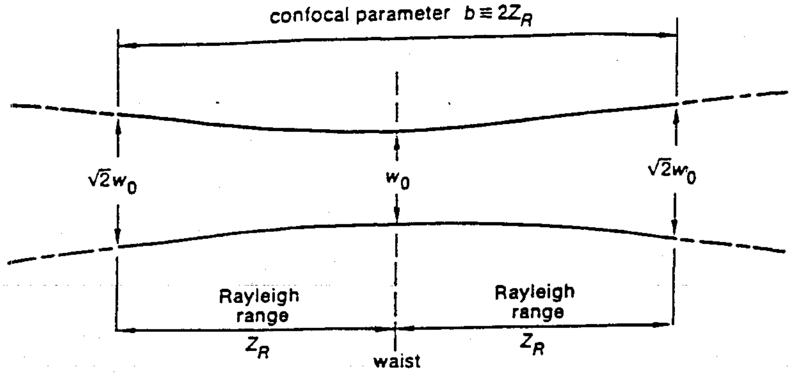


FIGURE 17.7
The collimated waist region of a gaussian beam.

therefore adopted the same term here as a name for the quantity $z_R \equiv \pi w_0^2/\lambda$. The Rayleigh range marks the approximate dividing line between the "near-field" or Fresnel and the "far-field" or Fraunhofer regions for a beam propagating out from a gaussian waist.

To express this same point in another way, if a gaussian beam is focused from an aperture down to a waist and then expands again, the full distance between the $\sqrt{2}w_0$ spot size points is the quantity b given by

$$b = 2z_R = \frac{2\pi w_0^2}{\lambda} = \text{confocal parameter.} \tag{10}$$

This confocal parameter was widely used in earlier writings to characterize gaussian beams. Using the Rayleigh range $z_R \equiv b/2$, as shown in Figure 17.7, seems, however, to give simpler results in most gaussian beam formulas.

Collimated Gaussian Beam Propagation

Over what distance can the collimated waist region of an optical beam then extend, in practical terms? To gain some insight into this question, we might suppose that a gaussian optical beam is to be transmitted from a source aperture of diameter D with a slight initial inward convergence, as shown in Figure 17.8, so that the beam focuses slightly to a waist with spot size w_0 at one Rayleigh range out, and then reexpands to the same diameter D two Rayleigh ranges (or one confocal parameter) out. We will choose the aperture diameter according to the πw or 99% criterion, i.e., we will use $D = \pi \times \sqrt{2}\, w_0$ at each end.

The relation between the collimated beam distance and the transmitting aperture size using this criterion is then

$$\text{collimated range} = 2z_R = \frac{2\pi w_0^2}{\lambda} \approx \frac{D^2}{\pi\lambda}. \tag{11}$$

Some representative numbers for this collimated beam range at two different laser wavelengths are illustrated in Figure 17.8 and in Table 17.1. A visible laser with a 1 cm diameter aperture can project a beam having an effective diameter of a few mm with no significant diffraction spreading over a length of 50 meters or more. Such a beam can be used, for example, as a "weightless string" for alignment on a construction project. With the aid of a simple photocell array,

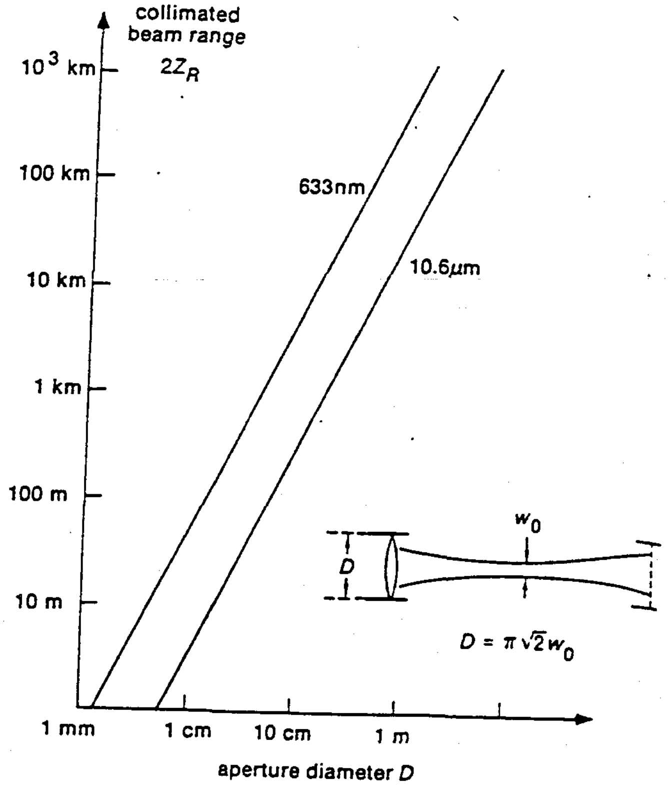


FIGURE 17.8
Collimated gaussian beam ranges versus transmitting aperture diameter D, using the $d = \pi w$ criterion.

TABLE 17.1
Collimated Laser Beam Ranges

Aperture diameter D	*Waist spot size* w_0	*Collimated range,* $2z_r$ (10.6 μm)	*Collimated range,* $2r_R$ (633 nm)
1 cm	2.25 mm	3 m	45 m
10 cm	2.25 cm	300 m	5 km
1 m	22.5 cm	30 km	500 km

the center of such a beam can easily be found to an accuracy of better than $w/20$, or a small fraction of a mm, over the entire distance.

Far-Field Beam Angle: The "Top Hat" Criterion

Suppose we next move out into the far field, where the beam size expands linearly with distance, as in Figure 17.9. At what angle does a gaussian beam spread in the far field, that is, for $z \gg z_R$?

From the gaussian beam equations (17.1-17.5), the $1/e$ spot size $w(z)$ for the field amplitude in the far field for a gaussian beam coming from a waist with spot size w_0 is given by

$$w(z) \approx \frac{w_0 z}{z_R} = \frac{\lambda z}{\pi w_0} \qquad (z \gg z_R), \tag{12}$$

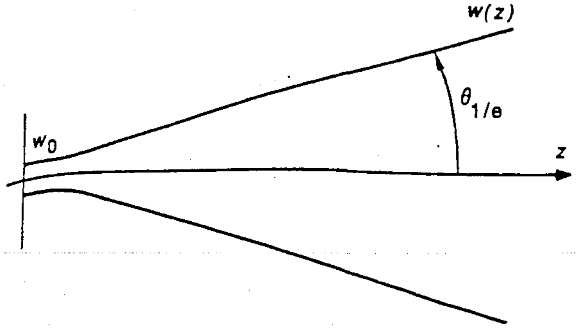


FIGURE 17.9
A gaussian beam spreads with a constant diffraction angle in the far field.

which gives the simple relation

$$w_0 \times w(z) \approx \frac{\lambda z}{\pi} \tag{13}$$

connecting the spot sizes at the waist and in the far field. The far-field angular beam spread for a gaussian beam can then related to the near-field beam size or aperture area in several different ways, depending on how conservative we want to be.

The on-axis beam intensity in the far field, for example, is given by

$$I_{\text{axis}}(z) = \frac{2P}{\pi w^2(z)} \approx \frac{P}{\lambda^2 z^2 / 2\pi w_0^2}. \tag{14}$$

Hence, the on-axis intensity is the same as if the total power P were uniformly distributed over an area $\pi w^2(z)/2 = \lambda^2 z^2/2\pi w_0^2$. The solid angle for an equivalent "top hat" angular distribution in the far field, call it $\Omega_{\text{TH}}(z)$, is thus given by

$$\Omega_{\text{TH}} = \frac{\pi w^2(z)}{2z^2} = \frac{\lambda^2}{2\pi w_0^2}. \tag{15}$$

At the same time, the "equivalent top hat" definition of the source area at the waist is given from Equation 17.7 by $A_{\text{TH}} = \pi w_0^2/2$. The product of these two quantities is thus given by

$$A_{\text{TH}} \times \Omega_{\text{TH}} = \left(\frac{\lambda}{2}\right)^2. \tag{16}$$

The source aperture size (at the waist) and the far-field solid angular spread thus have a product on the order of the wavelength λ squared, although the exact numerical factor will depend on the definitions we choose for the area and the solid angle, as we will see in more detail later.

Far-Field Beam Angle: The 1/e Criterion

Another and perhaps more reasonable definition for the far-field beam angle is to use the $1/e$ or 86% criterion for the beam diameter, so that the far field half-angular spread is defined by the width corresponding to the $1/e$ point for the E field amplitude at large z.

With this definition, the half-angle $\theta_{1/e}$ out to the $1/e$ amplitude points in the far-field beam is given, as shown in Figure 17.9, by

$$\theta_{1/e} = \lim_{z \to \infty} \frac{w(z)}{z} = \frac{\lambda}{\pi w_0}. \tag{17}$$

Twice this angle then gives a full angular spread of

$$2\theta_{1/e} = \frac{2\lambda}{\pi w_0}, \tag{18}$$

which can be interpreted as a more precise formulation, valid for gaussian beams, of the approximate relation $\Delta\theta \approx \lambda/d$ that we gave in Chapter 1. We can then define the gaussian beam solid angle $\Omega_{1/e}$ on this same basis as the circular cone defined by this angular spread, or

$$\Omega_{1/e} = \pi\theta_{1/e}^2 = \frac{\lambda^2}{\pi w_0^2}. \tag{19}$$

This cone will, as noted in the preceding, contain 86% of the total beam power in the far field.

Suppose we use the same $1/e$ criterion to define the effective radius of the input beam at the beam waist (ignoring the fact that an aperture of radius $a = w_0$ at the waist would actually produce some very substantial diffraction effects on the far-field beam pattern). Then the product of the effective source aperture area $A_{1/e} \equiv \pi w_0^2/2$ and the effective far-field solid angle $\pi\theta_{1/e}^2$ using these $1/e$ definitions becomes

$$A_{1/e}\Omega_{1/e} = \pi w_0^2 \times \pi\theta_{1/e}^2 = \lambda^2. \tag{20}$$

This is a precise formulation for gaussian beams of a very general antenna theorem which states that

$$\iint A(\Omega)\, d\Omega = \lambda^2 \tag{21}$$

This theorem says in physical terms that if we measure the effective capture area $A(\Omega)$ of an antenna for plane-wave radiation arriving from a direction specified by the vector angle $\Omega = (\theta, \phi)$, and then integrate these measured areas over all possible arrival angles as specified by $d\Omega$, the result (for a lossless antenna of *any* form) is always just the measurement wavelength λ. This result is valid for any kind of antenna, at radio, microwave or optical wavelengths.

Far-Field Beam Angle: Conservative Criterion

Finally, as a still more conservative way of expressing the same points, we might use the $d = \pi w$ or 99% criterion instead of the $1/e$ criterion to define both the effective source aperture size and the effective far field solid angle. We might then say that a source aperture of diameter $d = \pi w_0$ transmitting a beam of initial spot size w_0 will produce a far-field beam with 99% of its energy within a cone of full angular spread $2\theta_\pi = \pi w(z)/z$. On this basis the source aperture area, call it A_π is $\pi d^2/4$ and the beam far-field solid angle is $\Omega_\pi = \pi\theta_\pi^2$; and these are related by the more conservative criterion

$$A_\pi\Omega_\pi = \left(\frac{\pi}{2}\right)^4 \lambda^2 \approx 6\lambda^2. \tag{22}$$

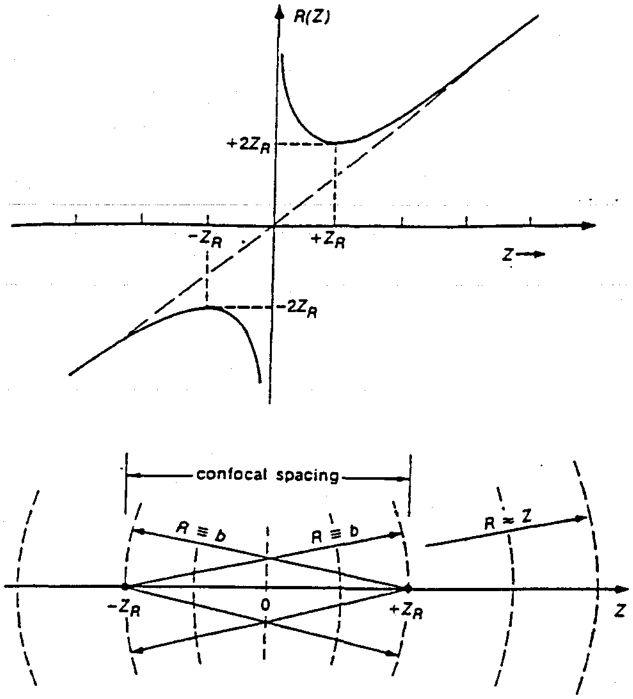


FIGURE 17.10
Radius of curvature for the wavefront of a gaussian beam, versus distance from the waist.

None of the criteria we have introduced here for defining effective aperture size and effective solid angle are divinely ordained, and which of them we use should depend largely on what objective we have in mind.

Wavefront Radius of Curvature

We can next look at how the wavefront curvature of a gaussian beam varies with distance. The radius of curvature $R(z)$ of a gaussian beam has a variation with distance given analytically by

$$R(z) = z + \frac{z_R^2}{z} \approx \begin{cases} \infty & \text{for} \quad z \ll z_R \\ 2z_R & \text{for} \quad z = z_R \\ z & \text{for} \quad z \gg z_R \end{cases} . \tag{23}$$

This is plotted against normalized distance in Figure 17.10(a).

The wavefront is flat or planar right at the waist, corresponding to an infinite radius of curvature or $R(0) = \infty$. As the beam propagates outward, however, the wavefront gradually becomes curved, and the radius of curvature $R(z)$ drops rather rapidly down to finite values (see Figure 17.10). For distances well beyond the Rayleigh range z_R the radius then increases again as $R(z) \approx z$, i.e., the gaussian beam becomes essentially like a spherical wave centered at the beam waist. What this means in physical terms is that the center of curvature of the wavefront starts out at $-\infty$ for a wavefront right at the beam waist, and then moves monotonically inward toward the waist, as the wavefront itself moves outward toward $z \to +\infty$.

Confocal Curvatures

The minimum radius of curvature occurs for the wavefront at a distance from the waist given by $z = z_R$, with the radius value $R = b = 2z_R$. This means that at this point the center of curvature for the wavefront at $z = +z_R$ is located at $z = -z_R$, and vice versa, as illustrated in Figure 17.10.

This particular spacing has a special significance in stable resonator theory. Suppose the curved wavefronts $R(z)$ at $\pm z_R$ are matched exactly by two curved mirrors of radius R and separation $L = R = b = 2z_R$. Since the focal point of a curved mirror of radius R is located at $f = R/2$. the focal points of these two mirrors then coincide exactly at the center of the resonator. The two mirrors are said to form a *symmetric confocal resonator*, thus giving rise to the *confocal parameter* $b \equiv 2z_R \equiv 2\pi w_0^2/\lambda$. Such a resonator has certain particularly interesting mode properties which we will explore later.

REFERENCES

Further discussion of the concept of the "Rayleigh range" can be found in J. F. Ramsay, "Tubular beams from radiating apertures," in *Advances in Microwaves, Vol. 3*, ed. by L. F. Young (Academic Press, New York, 1968), p. 127.

Earlier considerations of the same ideas by Lord Rayleigh (J. W. Strutt) himself can be found in his papers "On images formed with or without reflection or refraction," *Phil. Mag.* 11, 214–218 (1881), and "On pinhole photography," *Phil. Mag.* 31, 87–89 (1891).

Problems for 17.1

1. *Gaussian beam transmission through a square aperture.* Find the power transmission for a gaussian beam through a *square* aperture with sides of length $2a$, in analogy to the circular aperture results given in the text.

2. *Criteria for centering accuracy of a circular aperture.* Suppose a gaussian beam is transmitted through a circular aperture of diameter $d = \pi w$. How critical is the centering of the laser beam axis with respect to the aperture position (or vice versa)? Attempt to evaluate the decrease in beam transmission versus the displacement between beam and aperture centers, using either approximate mathematical methods or computer evaluation.

3. *Setting tolerances on beam collimation and far-field beam angle.* A laser oscillator is designed to give a collimated beam at its output plane with a specified spot size w_0 and a collimated wavefront with radius of curvature $R_0 = \infty$. Due to manufacturing tolerances, however, the actual output wavefront may come out slightly spherical. Suppose we establish as a practical tolerance for good collimation that the far-field beam angle of the laser output beam should not vary by more than 10% from its design value. What is the resulting tolerance on R_0 (or $1/R_0$) at the laser's output plane for a fixed value of w_0? How much wavefront distortion does this represent, expressed in terms of fractional wavelengths of wavefront distortion at the $1/e$ radius of the output beam?

4. *Simulating an annular beam with positive and negative gaussians.* A circularly symmetric beam with a hole in the center can be simulated by superposing

two collimated gaussian beams to give an initial field distribution $\tilde{u}_0(r) = \exp(-r^2/w_1^2) - \exp(-r^2/w_2^2)$ with $w_2 = \beta w_1$ and $0 \leq \beta \leq 1$. Calculate and plot the profile $\tilde{u}(r,z)$ of this beam at different distances z in the near and far fields for different values of β, with distance measured in the normalized coordinate $s \equiv z\lambda/\pi w_1^2 \equiv z/z_{R1}$. How rapidly does the hole in the center of the beam fill in as a function of distance for different values of β?

5. *Locating the center of curvature of a gaussian beam.* The text describes how the radius of curvature $R(z)$ changes as we move out along the z axis away from a gaussian waist at $z = 0$. Give an analytic expression for how the center of curvature of the same wavefront moves.

6. *Beam spot size at long distances.* Suppose a frequency-doubled Nd:YAG laser ($\lambda = 532$ nm) is transmitted through a 1 meter diameter diffraction-limited telescope, using the $d = \pi$ criterion, to illuminate a spot on the face of the moon ($z \approx 384{,}000$ km). What will be the $1/e$ diameter of the spot?

   As a practical matter, because of beam distortions through the atmosphere, except under very exceptional "seeing conditions" the largest aperture that can be diffraction-limited through the Earth's atmosphere has a diameter more like 10 cm. What will be the spot size for this aperture?

17.2 GAUSSIAN BEAM FOCUSING

Besides propagating collimated gaussian beams over long distances, we are often interested in focusing such beams to very small spots, whether for recording data on optical videodisks or tapes, drilling holes in razor blades, or counting cell nuclei in a laser microscope. (Since the standard demonstration of ruby laser intensity in early days was to zap a hole in one or more razor blades with a single laser shot, pulsed laser energies were occasionally quoted in "Gillettes.") What sort of focused spot sizes and intensities can be achieved with a gaussian beam—or for that matter with any reasonably well-formed optical beam?

Focused Spot Sizes

The usual situation where a collimated gaussian beam is strongly focused by a lens of focal length f, as shown in Figure 17.11, can be viewed as simply the far-field beam problem of Figure 17.9 in reverse. The waist region now becomes the focal spot of spot size w_0, whereas the focusing lens can be viewed as being in the far field at $z \approx \pm f$. If $w(f)$ is the gaussian spot size at the lens, we then have the same relationship as Equation 17.13 but with a reverse interpretation, namely,

$$w_0 \times w(f) \approx \frac{f\lambda}{\pi}. \tag{24}$$

What does this expression imply in practical terms?

It seems obvious that in a practical focusing problem, the incident gaussian beam should fill the aperture of the focusing lens to the largest extent possible without a severe loss of power due to the finite aperture of the lens (and also without serious edge diffraction effects). As one reasonable criterion for practical

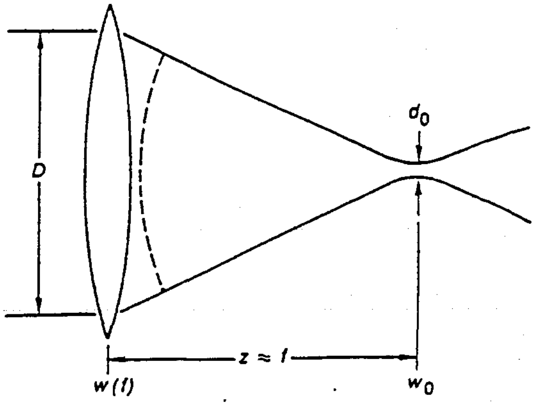


FIGURE 17.11
Focusing of a gaussian beam to a small spot size.

designs, we might adopt the $D = \pi w(f)$ or 99% criterion for the diameter d of the focusing lens, so that we lose < 1% of the incident energy in this lens. At the same time we might adopt the $1/e$ or $d_0 = 2w_0$ criterion for defining the effective diameter d_0 of the focused spot, since this is a diameter which contains 86% of the focused energy, and at the edges of which the focused intensity is already down to $1/e^2 \approx 14\%$ of its peak value. Combining these criteria then gives

$$d_0 \approx \frac{2f\lambda}{D} \tag{25}$$

for the effective diameter of the focused gaussian spot.

The *f-number* of a focusing lens (also called the *relative aperture* or the *speed* of the lens) is defined by

$$f^{\#} \equiv \frac{f}{D}. \tag{26}$$

The focal spot diameter, using the rather arbitrary criteria we have just selected, will then be given by

$$d_0 \approx 2f^{\#}\lambda. \tag{27}$$

As an alternative way of reaching essentially this same conclusion, we can calculate that if a gaussian beam carries total power P and we focus it using a lens of focal length f with the same $D = \pi w$ criterion for the lens diameter, then the peak intensity at the center of the focused spot will be given by

$$I_0 = \frac{2P}{\pi w_0^2} \approx \frac{P}{2(f^{\#}\lambda)^2}. \tag{28}$$

The peak intensity is thus the same as if all the energy were focused into a circle with an area of $2(f^{\#}\lambda)^2$, or a diameter of $(8/\pi)^{1/2}f^{\#}\lambda \approx 1.6f^{\#}\lambda$.

Influence of the Lens f Number and the Lens Fresnel Number

Whatever the choice of definitions, it is evident that an ideal gaussian beam can be focused down to a spot that is roughly one to two optical wavelengths in diameter, multiplied by the f-number of the focusing lens. Note that a long focal length lens, say, an $f/10$ lens, will generally be simple, inexpensive and easy to obtain, with quite small aberration coefficients. Lenses with f-numbers less than 2, and especially with $f^{\#} \leq 1$, on the other hand, generally require complex and expensive multielement designs, and can become very expensive.

Some optics workers also like to characterize a simple lens of diameter $D = 2a$ and focal length f by its *lens Fresnel number* N_f, given by

$$N_f \equiv \frac{a^2}{f\lambda}. \tag{29}$$

In terms of this quantity plus our arbitrary criteria for beam and spot diameters, the focal spot diameter and the lens diameter are then related by

$$\frac{d_0}{D} \approx \frac{1}{2N_f}. \tag{30}$$

Whereas the f-number of a given lens is independent of wavelength, the Fresnel number depends on wavelength. The limitation expressed by Equation 17.30 can become significant particularly for longer wavelengths,, for example, in focusing infrared beams using IR lenses. Strong focusing, down to a spot size much less than the lens diameter, requires a lens with an adequately large Fresnel number N_f.

A crucial condition for accomplishing *strong focusing*, regardless of definitions, is that the incident gaussian beam properly fill the focusing lens aperture, since it is the *gaussian beam diameter* and not the *lens diameter* that is the critical dimension in determining the focal spot size of the gaussian beam.

Depth of Focus

The depth of focus of a gaussian beam is obviously given by the Rayleigh range z_R of the gaussian waist, or perhaps by $2z_R$, depending upon just how we want to define the depth of focus. If we use the latter definition, along with the lens diameter criterion $D = \pi w(f)$, then the depth of focus can be written as

$$\text{depth of focus } = 2z_R \approx 2\pi f^{\#^2}\lambda \approx \frac{\pi}{2}\left(\frac{d_0}{\lambda}\right)^2 \lambda. \tag{31}$$

If the beam is focused down to a spot N wavelengths in diameter, the depth of focus will be $\approx N^2$ wavelengths in length.

All these expressions for focused spot size and depth of focus do of course assume (a) that the gaussian beam entering the lens is more or less collimated, with a planar wavefront, so that the beam focuses approximately at the focal point f; and (b) that the beam is in fact "strongly focused," in the sense that $w_0 \ll w(f)$, or $z_R \ll f$, or $N_f \gg 1$. This latter point is equivalent to saying that the lens is in the far field, as seen looking backward from the waist or the focal point. If either of these assumptions is not entirely valid, corrections must be applied in calculating the exact location and size of the focused spot, as illustrated in several of the Problems following this section.

Focal Spot Deviation

When a collimated beam is focused by an ideal lens, the actual focal spot, meaning the position of minimum spot size and maximum energy density, does

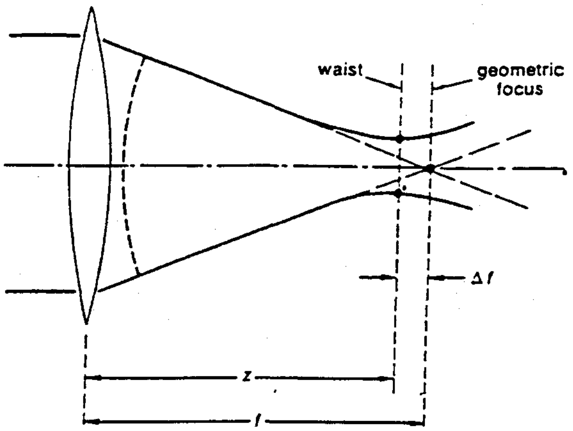


FIGURE 17.12
There is a very small (in practice, negligible) shift in position between the geometrical focus of the lens and the actual waist of the focused gaussian beam.

not in fact occur exactly at the geometrical focus of the lens; but rather is located just slightly *inside* the lens focal length. The amount of this focal spot deviation—which is typically very small—can be easily calculated for a focused gaussian beam from Figure 17.12.

Using the notation of Figure 17.12, we can let the distance from the lens to the beam waist, or the actual focal spot, be z, whereas the focal length of the lens is f. A collimated beam passing through a thin lens of focal length f acquires (by definition) a wavefront radius of curvature equal to f. The wavefront curvature just beyond the thin lens must therefore be given, from the combination of gaussian beam theory and lens theory, by

$$R(z) = z + z_R^2/z = f. \tag{32}$$

The difference between the focal length f and the actual distance z to the waist can then be written as

$$\Delta f \equiv f - z = z_R^2/z \approx z_R^2/f. \tag{33}$$

Since the Rayleigh range z_R of the focused beam is normally much less than the focal length f, the focal deviation is generally much less than the depth of focus (which means that in fact it is really quite negligible). One way of expressing this criterion is

$$\frac{\Delta f}{f} \approx \frac{1}{2N_f^2}. \tag{34}$$

As a practical matter, when adjusting an optical setup one very seldom knows the exact value of the lens focal length, or the exact location of the lens focal point, or the exact degree of collimation of the input beam to sufficiently high accuracy that this focal spot deviation is of any practical significance. We usually find the best focal adjustment in any optical system by small experimental adjustments over a small adjustment range, after the system is assembled.

Summary

Essentially all of the results we have given in this section, for collimated beam length, far-field beam angle, focal spot size, depth of focus, and focal spot deviation, although derived here for a gaussian beam, will apply equally well in fact to any optical beam having a reasonably well-collimated or uniform phase front and a reasonably uniform amplitude profile across the beam. The gaussian beam simply provides a particularly convenient example in which the mathematical expressions for spot size and wavefront curvature are particularly simple. The gaussian beam also has the special characteristic, as distinguished from most other beams, that its transverse profile remains gaussian at every transverse cross section.

Suppose a uniform plane wave is focused to a spot by an ideal thin lens with a circular aperture. The focused spot will then have the form of an Airy disk pattern at the focal plane of the lens; but the spot size and depth of focus of this focused spot will still have essentially the same dependence on lens f-number or Fresnel number as given by Equations 17.24 to 17.31. More detailed calculations will even show that the on-axis intensity will be very slightly higher, and the average beam diameter very slightly smaller, at a transverse cross section just slightly inside the exact focal length; and the position of this plane of tighter focus will be given by the same Equations 17.33 or 17.34 we have just derived for focal spot deviation.

REFERENCES

Some interesting practical methods for measuring small gaussian beam waists can be found in M. B. Schneider and W. W. Webb, "Measurement of submicron laser beam radii," *Appl. Optics* 20, 1382–1388 (April 15 1981); and in D. K. Cohen, B. Little, and F. S. Luecke, "Techniques for measuring 1-μm diam gaussian beams," *Appl. Opt.* **23**, 637–640 (February 15 1984).

If a gaussian beam is focused strongly enough to begin violating the paraxial approximation, we expect deviations from gaussian beam theory to appear particularly strongly in the focal spot region. Some of these deviations have been calculated and plotted in detail by W. H. Carter, "Anomalies in the field of a gaussian beam near focus," *Optics Commun.* 7, 211–218 (March 1973).

Problems for 17.2

1. *Focusing for absolute minimum spot size.* A collimated gaussian beam of fixed spot size w is to be focused to the absolute minimum possible spot size (not necessarily a beam waist) on a work piece, using a single lens located a fixed distance L from the work piece. What should be the exact focal length f of this lens, and what will be the exact spot size of the focused spot? What is the waist size and location of the same focused beam?

2. *Focusing a gaussian beam with an astigmatic lens.* A circular gaussian beam with an initially large spot size w_1 is focused using an astigmatic lens which has different focal lengths f_x and f_y in the x and y transverse coordinates. Develop an analysis for the shape and location of the two different waists in the two transverse directions. Discuss in particular how the axial distance between the two waists will relate to the Rayleigh length for each individual waist if the two focal lengths are similar but differ by, say, 10% to 20% in magnitude.

3. *Focusing into a dielectric medium.* A focusing lens of focal length f is located outside a dielectric sample with a flat front surface, at a distance from the sur-

face that is less than the focal length f. so that the focal spot or gaussian beam waist occurs inside the dielectric sample. What is the spot size at this resulting waist; how does it compare to the spot size that would occur without the dielectric present: and where does it occur with respect to the lens position and the dielectric surface? (Warning! You will have to think through rather carefully what happens to a gaussian beam in passing through a planar dielectric surface.)

17.3 LENS LAWS AND GAUSSIAN MODE MATCHING

A common requirement in laser optical systems is to propagate a gaussian beam through a cascaded sequence of lenses, free-space regions, and other optical elements, as shown in Figure 17.14, perhaps in order to match a gaussian beam coming from a waist with specified spot size w_1 at location z_1 into another laser cavity or interferometer requiring waist spot size w_2 at location z_2. The design steps necessary to accomplish this are usually referred to as *gaussian beam mode matching.*

Such problems if they become at all complicated are probably best handled by the general *ABCD* methods we will introduce later. A quick introduction to elementary gaussian mode matching techniques at this point may, however, be useful.

Lens Laws and Collins Charts

The lens law for purely spherical waves passing through an ideal thin lens of focal length f (Figure 17.13) is

$$\frac{1}{R_2}=\frac{1}{R_1}-\frac{1}{f}. \qquad (35)$$

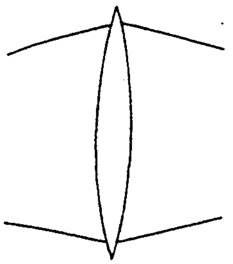
We follow the standard convention in this book of using positive R for *diverging* waves going in the $+z$ direction, and positive f for *converging* or *positive* lenses. A gaussian spherical beam passing through such a thin lens then has its radius of curvature R changed in exactly the same way, whereas its spot size w is unchanged. The lens law for gaussian beams is therefore the direct analog, that is,

$$\frac{1}{\tilde{q}_2}=\frac{1}{\tilde{q}_1}-\frac{1}{f}, \qquad (36)$$

where $\tilde{q}$ is the complex curvature parameter defined in Equation 17.2.

By applying this lens law, plus the propagation rule $\tilde{q}_2=\tilde{q}_1+z_2-z_1$ for a free-space section, we can then propagate a gaussian beam forward or backward through any sequence of thin lenses and spaces. It can be helpful to plot this propagation as a trajectory in the complex $1/\tilde{q}$ plane or, more conveniently, in the complex $j/\tilde{q}$ plane with rectangular coordinates x and y corresponding to $x\equiv\lambda/\pi w^2(z)$ and $y\equiv 1/R(z)$, respectively. From Equation 17.36 the effect of a thin lens is to cause a vertical jump of magnitude $-1/f$ in the $j/\tilde{q}$ plane.

To those familiar with bilateral transformations as used in electrical circuit theory and elsewhere, it will be obvious that the transformation law through a free-space section, as given by $\tilde{q}(z)=\tilde{q}_0+z=z+jz_R$, corresponds to a



$$w_1 = w_2$$

$$\frac{1}{R_2} = \frac{1}{R_1} - \frac{1}{f}$$

$$\frac{1}{q_2} = \frac{1}{q_1} - \frac{1}{f}$$

FIGURE 17.13
Gaussian beam transmitted through an ideal thin lens.

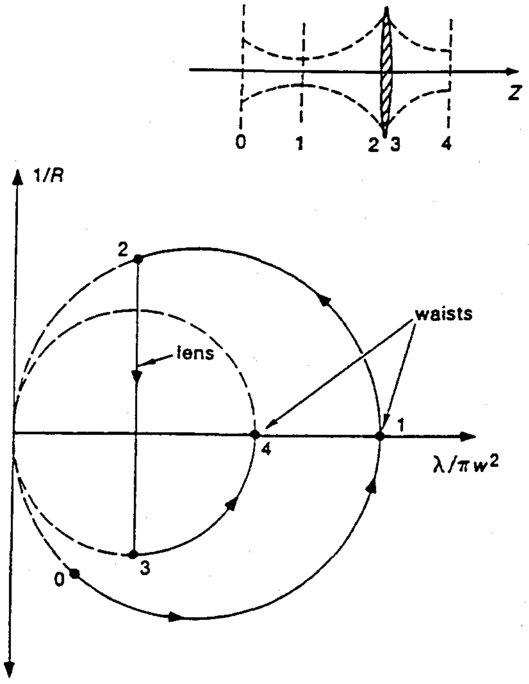


FIGURE 17.14
Gaussian beam propagation through a sequence of optical elements, as diagrammed on a gaussian-beam chart or Collins chart.

transformation around a circular arc in the complex $j/\tilde{q}$ plane, as shown in Figure 17.14. In these so-called *gaussian beam charts* or *Collins charts*—which are very similar in form to the Smith charts of transmission line theory—different gaussian beam waists correspond to different points $x = 1/z_R$, $y = 0$ on the x axis. Free space propagation then corresponds to circular arcs passing through these points and the origin (which corresponds to the far field at $z \rightarrow \infty$); whereas lines of constant z/z_R for different z_R are also circles passing through the origin. Thin lenses are then vertical transitions on the same chart as shown.

Charts of this type may be of some use for visualizing gaussian beam propagation problems or for diagramming solutions. With widespread access to computers, however, their practical uses as calculational tools are negligible.

REFERENCES

The gaussian beam chart or Collins chart is described in S. A. Collins, Jr., "Analysis of optical resonators involving focusing elements," *Appl. Optics* **3**, 1263–1275 (November 1964); and in T. Li, "Dual forms of the gaussian beam chart," *Appl. Optics* **3**, 1315–1317 (November 1964).

Other references on the same topic include J. P. Gordon, "A circle diagram for optical resonators," *Bell Sys. Tech. J.* **43**, 1826–1827 (November 1964); and T. S. Chu, "Geometrical representation of gaussian beam propagation," *Bell. Sys. Tech. J.* **45**, 287–299 (February 1966).

Another graphical approach to gaussian beam propagation and mode matching is given by P. Laures, "Geometrical approach to gaussian beam propagation," *Appl. Opt.* **6**, 747–755 (April 1967).

Problems for 17.3

1. *Thin-lens imaging formulas for gaussian beams.* A gaussian waist with spot size w_1, located a distance L_1 to the left of a thin lens with focal length f, is imaged by that lens into a waist with spot w_2 located a distance L_2 to the right of the lens. Evaluate for this situation (a) the relationship between the "object distance" L_1, the "image distance" L_2, and the focal length f; and (b) the linear magnification $M = w_2/w_1$ between object and image, again in terms of L_1, L_2 and f. Discuss any differences between these gaussian-beam results and the corresponding formulas for purely geometrical optics.

17.4 AXIAL PHASE SHIFTS: THE GUOY EFFECT

The propagation of a gaussian beam also involves a subtle but sometimes important *added phase shift* through the waist region, which we will briefly describe in this section.

Axial Phase Shift

The propagation equation (17.3 or 17.5) for a lowest-order gaussian beam includes both a spot size variation and a cumulative phase shift variation with axial distance z which are given on the optical axis ($x = y = 0$) by the factors

$$\tilde{u}(z) \propto \frac{\tilde{q}_0 e^{-jkz}}{\tilde{q}(z)} = \frac{e^{-jkz}}{1 - jz/z_R} = \frac{\exp[-jkz + j\psi(z)]}{w(z)}. \tag{37}$$

In addition to the free-space or plane-wave phase shift given by the e^{-jkz} term, there is also an added axially-varying phase shift $\psi(z)$ given by

$$\psi(z) = \tan^{-1}\left(\frac{z}{z_R}\right) \tag{38}$$

assuming we measure this added phase shift with respect to the beam waist location.

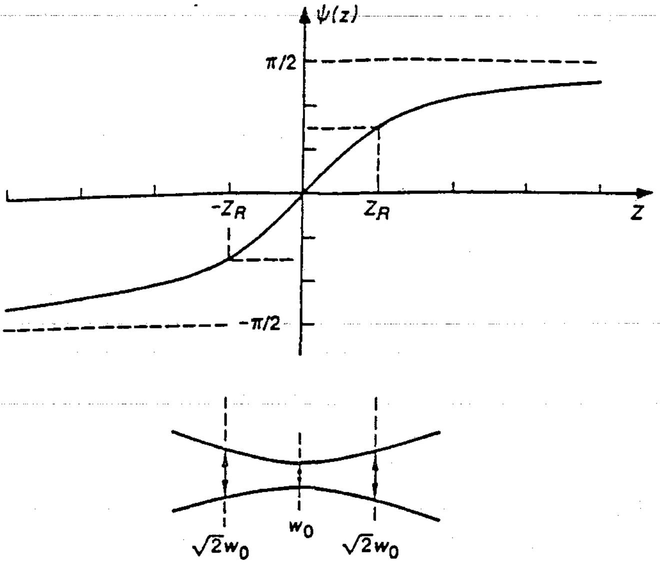


FIGURE 17.15
Guoy phase shift through the waist region of a gaussian beam.

The net effect of this added phase shift $\psi(z)$ for the lowest-order gaussian mode, as plotted in Figure 17.15, is to give an additional cumulative phase shift of $\pm 90°$ on either side of the waist, or a total added phase shift of 180° in passing through the waist, with most of this additional phase shift occurring within one or two Rayleigh ranges on either side of the waist.

This added phase shift means in physical terms that the effective axial propagation constant in the waist region is slightly smaller, i.e., $k_{\text{eq}}(z) = k - \Delta k$, or that the phase velocity and the spacing between phase fronts are slightly larger, i.e., $v_\phi(z) = c + \Delta v$, than for an ideal plane wave. The phase fronts for a gaussian beam passing through a waist will thus shift forward by a total amount of half a wavelength compared to an ideal plane wave, as illustrated in Figure 17.16.

A mathematical understanding of this additional phase shift can be obtained by rewriting the paraxial wave equation (16.7) in the form

$$\frac{\partial \tilde{u}(x,y.z)}{\partial z} = -\frac{j}{2k}\nabla_{xy}^2 \tilde{u}(x,y,z), \tag{39}$$

where ∇_{xy}^2 is the Laplacian in x, y coordinates. The transverse second derivatives of the wave amplitude $\tilde{u}$ thus lead, through the wave equation, to a small but significant additional phase shift per unit length in the axial direction. The resulting increased phase velocity in the axial direction is exactly like the increased phase velocity in a closed waveguide. The transverse derivatives are the largest, and hence the added phase shift term is most significant, within one or two focal depths on either side of a focus, more or less independent of the exact transverse amplitude profile of the focused beam.

The Guoy Effect

This result is in fact simply the gaussian beam version of the *Guoy effect*, which is valid for any kind of optical (or microwave) beam passing through a focal region. This effect, which was first discovered experimentally by Guoy in

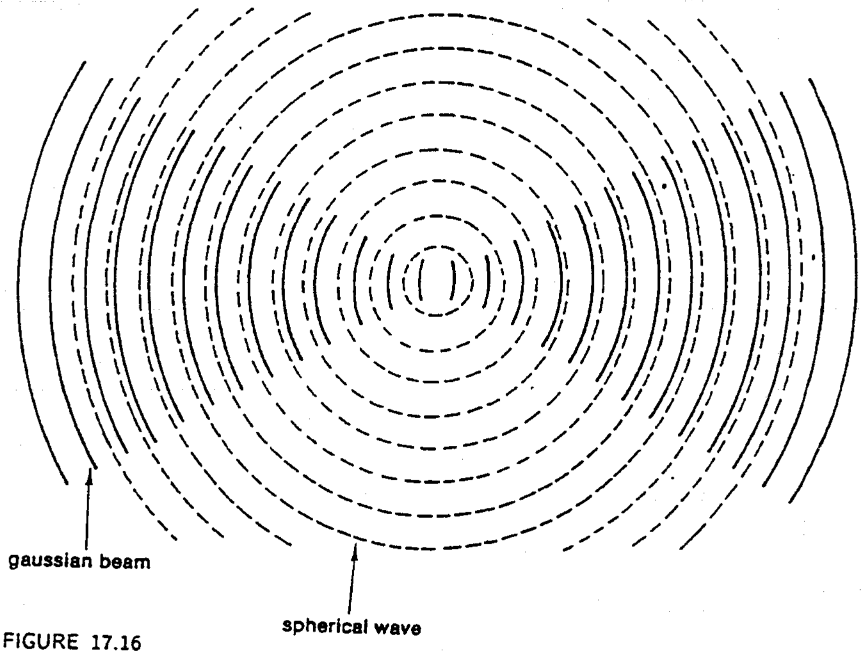


FIGURE 17.16
Alternative picture of the Guoy phase shift through the waist region, as compared to an ideal spherical wave.

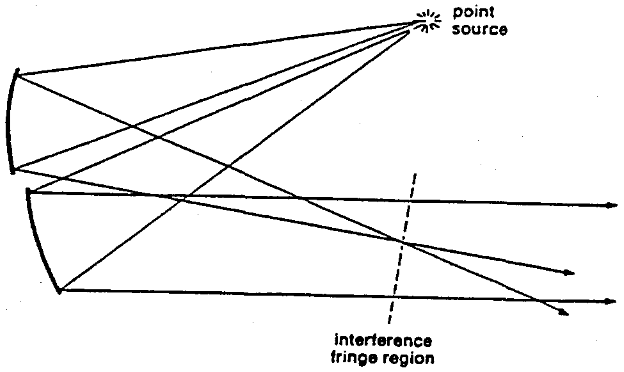


FIGURE 17.17
Experimental apparatus used by Guoy to demonstrate the extra 180° phase shift for an optical beam passing through a focus.

1890, says that a beam with *any* reasonably simple cross section will acquire an extra half-cycle of phase shift in passing through a focal region.

Figure 17.17 shows the simple apparatus employed by Guoy to demonstrate this effect. In the original experiment the light diverging from a small pinhole was reflected into two overlapping beams reflected from both a planar and a curved mirror. Interference effects between the two beams then produced a set of circular interference fringes between the two beams which could be observed at transverse planes near the first image of the pinhole. Guoy noticed that the centermost fringe in this "bulls-eye pattern" changed sign from dark to light (or vice versa) if he observed the fringes at observation planes just before or just after the focal point. This change of sign implied that the focused beam had somehow picked up an extra π phase shift in passing through the focus.

We will see shortly that higher-order transverse modes, because they have more complicated transverse second derivatives in Equation 17.39, have larger

Guoy phase shifts in passing through the waist region. In fact, if the lowest-order gaussian mode has Guoy phase shift $\psi(zz)$ at any plane z, measured relative to the focal point, then an nm-th order Hermite-gaussian mode with the same $\tilde{q}$ parameter will have a Guoy phase shift of $(n+m+1)\times\psi(z)$. These differing phase shifts are directly responsible for the slightly different resonance frequencies and mode beats of different nm-th order transverse modes in stable laser cavities.

The Guoy phase shift also explains the possibly somewhat puzzling 90° phase shift associated with the factor of j in the $j/L\lambda$ constant that occurs as part of the kernel in Huygens' integral (Equations 16.17 or 16.19). The physical interpretation of Huygens' integral considers the Huygens' wavelets as being ideal spherical wavelets diverging from each source point on the wavefront in the input plane, except that there is apparently a 90° phase shift between the incident wavefront and the diverging wavelet. The Guoy effect says that this occurs because each wavelet will acquire exactly 90° of extra phase shift in diverging from its point source or focus to the far field, thus accounting exactly for the j factor in the $j/L\lambda$ term.

REFERENCES

The original references on the Guoy effect can be found in G. Guoy, *Compt. Rendue Acad. Sci. Paris* **110**, 1251–1253 (1890) and *Ann. de Chim. et Phys.* **24**, 145–213 (1981); cf. also F. Reiche, *Ann. Physik* **29**, 65 and 401 (1909).

The Guoy effect occurs equally well with focused microwave or radio-wave beams as well as with optical beams. A useful discussion and illustration of all these situations can be found in C. L. Andrew, *Optics of the Electromagnetic Spectrum* (Prentice-Hall, 1960), pp. 114–118.

Another useful discussion along lines similar to the present section, and with an extensive summary of earlier references, can be found in R. W. Boyd, "Intuitive explanation of the phase anomaly of focused light beams," *J. Opt. Soc. Am.* **70**, 877–880 (July 1980).

For a clear theoretical discussion and graphical illustrations of the Guoy effect near the focal region of a nongaussian (uniform) beam, see E. H. Linfoot and E. Wolf, "Phase distribution near focus in an aberration-free diffraction image," *Proc. Phys. Soc. B* **69**, 823–832 (1956). For theoretical results in the focal region of a gaussian beam, see W. H. Carter, "Anomalies in the field of a gaussian beam near focus," *Optics. Commun.* **7**, 211-218 (March 1983).

For experimental results, see G. W. Farnell, "Measured phase distribution in the image space of a microwave lens," *Canadian J. Phys.* **36**, 935 (1958).

17.5 HIGHER-ORDER GAUSSIAN MODES

Let us now look in somewhat more detail at the higher-order Hermite-gaussian modes we derived in the previous chapter. In doing this we will consider only the "standard" set of higher-order Hermite-gaussians discussed in Section 16.4, since they usually match up most closely with the actual higher-order modes in simple optical resonators (at least in optical resonators which do not have "soft" apertures or radially varying gains or losses).

Higher-Order Hermite-Gaussian Mode Functions

The free-space Hermite-gaussian TEM$_{nm}$ solutions derived in the preceding chapter can be written, in either the x or y transverse dimensions, and with the plane-wave e^{-jkz} phase shift factor included for completeness, in the normalized form

$$\tilde{u}_n(x,z) = \left(\frac{2}{\pi}\right)^{1/4} \left(\frac{1}{2^n n! w_0}\right)^{1/2} \left(\frac{\tilde{q}_0}{\tilde{q}(z)}\right)^{1/2} \left[\frac{\tilde{q}_0}{\tilde{q}_0^*}\frac{\tilde{q}^*(z)}{\tilde{q}(z)}\right]^{n/2} \times H_n\left(\frac{\sqrt{2}x}{w(z)}\right) \exp\left[-jkz - j\frac{kx^2}{2\tilde{q}(z)}\right], \tag{40}$$

where the H_n's are the Hermite polynomials of order n, and the parameters $\tilde{q}(z)$, $w(z)$ and $\psi(z)$ are exactly the same as for the lowest-order gaussian mode as given in Equation 17.5. These same functions can be written in alternative form, emphasing the spot size $w(z)$ and Guoy phase shift $\psi(z)$, in the form

$$\tilde{u}_n(x,z) = \left(\frac{2}{\pi}\right)^{1/4} \left(\frac{\exp[j(2n+1)\psi(z)]}{2^n n! w(z)}\right)^{1/2} \times H_n\left(\frac{\sqrt{2}x}{w(z)}\right) \exp\left[-jkz - j\frac{kx^2}{2R(z)} - \frac{x^2}{w^2(z)}\right], \tag{41}$$

where $\psi(z)$ is still given by $\psi(z) = \tan^{-1}(z/z_R)$.

Note the important point that the higher-order modes, because of their more rapid transverse variation, have a net Guoy phase shift of $(n + 1/2)\psi(z)$ in traveling from the waist to any other plane z, as compared to only $\psi(z)$ for the lowest-order mode. This differential phase shift between Hermite-gaussian modes of different orders is of fundamental importance in explaining, for example, why higher-order transverse modes in a stable laser cavity will have different oscillation frequencies; or how the Hermite-gaussian components that add up to make a uniform rectangular or strip beam in one transverse dimension at an input plane located in the near field (at a beam waist) can add up to give a $(\sin x)/x$ transverse variation for the same beam in the far field.

Hermite-Gaussian Mode Patterns

Figure 17.18 illustrates the transverse amplitude variations for the first six even and odd Hermite-gaussian modes. Note that the first few (unnormalized) Hermite polynomials are given by

$$\begin{aligned} H_0 &= 1 & H_1(x) &= 2x \\ H_2(x) &= 4x^2 - 2 & H_3(x) &= 8x^3 - 12x. \end{aligned} \tag{42}$$

These polynomials obey the recursion relation

$$H_{n+1}(x) = 2xH_n(x) - 2nH_{n-1}(x) \tag{43}$$

which can provide a useful way of calculating the higher-order polynomials in numerical computations.

The Hermite-gaussian beam functions alternate between even and odd symmetry with alternating index n. The n-th order function has n nulls and $n+1$

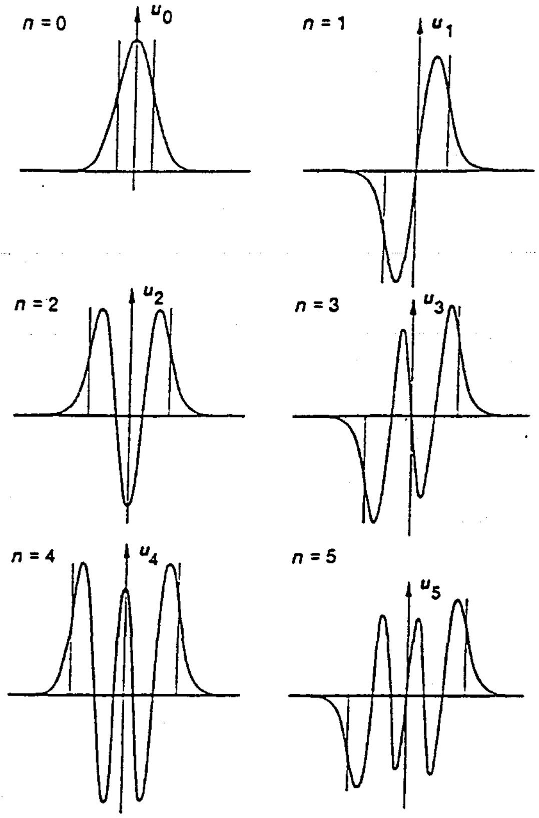


FIGURE 17.18
Amplitude profiles for low-order Hermite-gaussian modes.

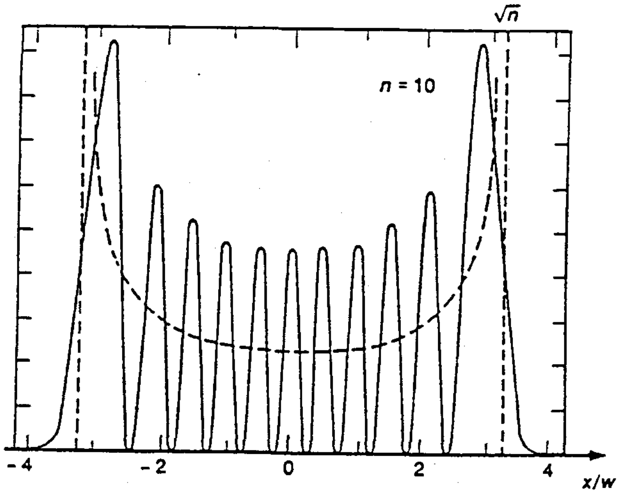


FIGURE 17.19
Intensity profile for the Hermite-gaussian mode pattern with $n = 10$.

peaks. These same Hermite-gaussian functions are also the quantum mechanical eigenfunctions for the linear quantum harmonic oscillator. Figure 17.19 illustrates the intensity variation, or the wave amplitude squared, for the $n = 10$ eigenmode, showing how the wave distribution approaches the classical proba-

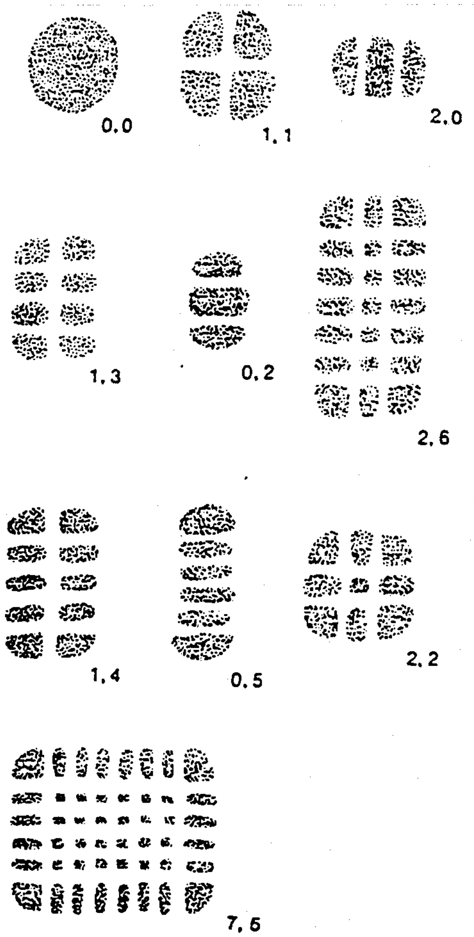


FIGURE 17.20
Transverse mode patterns for Hermite-gaussian modes of various orders.

bility density for a linear harmonic oscillator. It can also be seen that for larger values of n the outermost peaks become noticeably more intense than the inner peaks.

The complete set of Hermite-gaussian transverse modes for a beam in two transverse dimensions can then be written as $\tilde{u}_{nm}(x,y,z) = \tilde{u}_n(x,z) \times \tilde{u}_m(y,z)$, where in the most general situation a different $\tilde{q}(z)$ parameter, and even a different waist location, may apply to the x and the y variations. Figure 17.20 shows how the intensity patterns of various higher-order modes appear if the output beam from a laser oscillating in one of these higher-order modes is projected onto a screen. Note that the Hermite-gaussian functions are everywhere scaled to the spot size w through the arguments x/w and y/w. Hence, *the intensity pattern of any given TEM$_{nm}$ mode changes size but not shape as it propagates forward in z*—a given TEM$_{nm}$ mode looks exactly the same, except for scaling, at every point along the z axis.

The higher-order Laguerre-gaussian mode patterns also described in Section 16.4 (cf. Equation 16.64) are characterized by azimuthal and radial symmetry, rather than by the rectangular symmetry of the Hermite-gaussian modes, as illustrated in Figure 17.21. As explained earlier, most real lasers prefer to oscillate in modes of rectangular rather than cylindrical symmetry, although with very

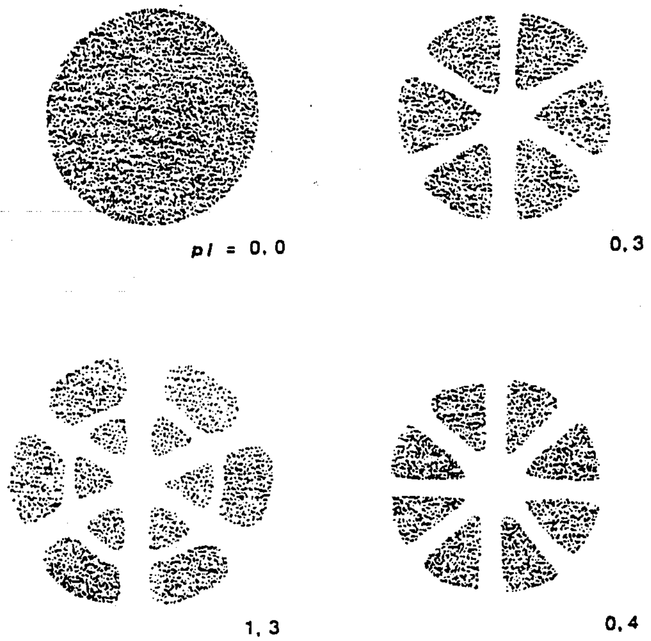


FIGURE 17.21
Transverse mode patterns for Laguerre-gaussian modes of various orders.

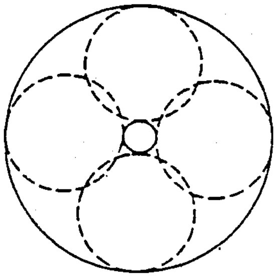


FIGURE 17.22
The "donut" mode is a linear superposition of 10 and 01 Hermite-gaussian modes.

careful adjustment, certain internal-mirror lasers can be made to oscillate in the cylindrical Hermite-gaussian modes.

The "Donut Mode"

In many laser experiments with stable laser resonators, the experimental procedure is to stop down an adjustable circular aperture inside the laser cavity until higher-order mode oscillation is completely suppressed and the laser oscillates only in the desired TEM$_{00}$ mode. For aperture diameters slightly larger than this value, lasers are often observed to produce an output beam in the form of a circularly symmetric ring with a dark spot on axis, as illustrated in Figure 17.22.

This mode, often referred to as the "donut mode," cannot be an $m = 0$ mode, since an $m = 0$ Laguerre-gaussian mode can never have a null on axis. It might be interpreted as a higher-order $\tilde{u}_{pm}(r, \theta)$ Laguerre-gaussian mode with $p = 1$ and an azimuthal variation like $e^{jm\theta}$ with $m \geq p$. In most practical lasers, however,

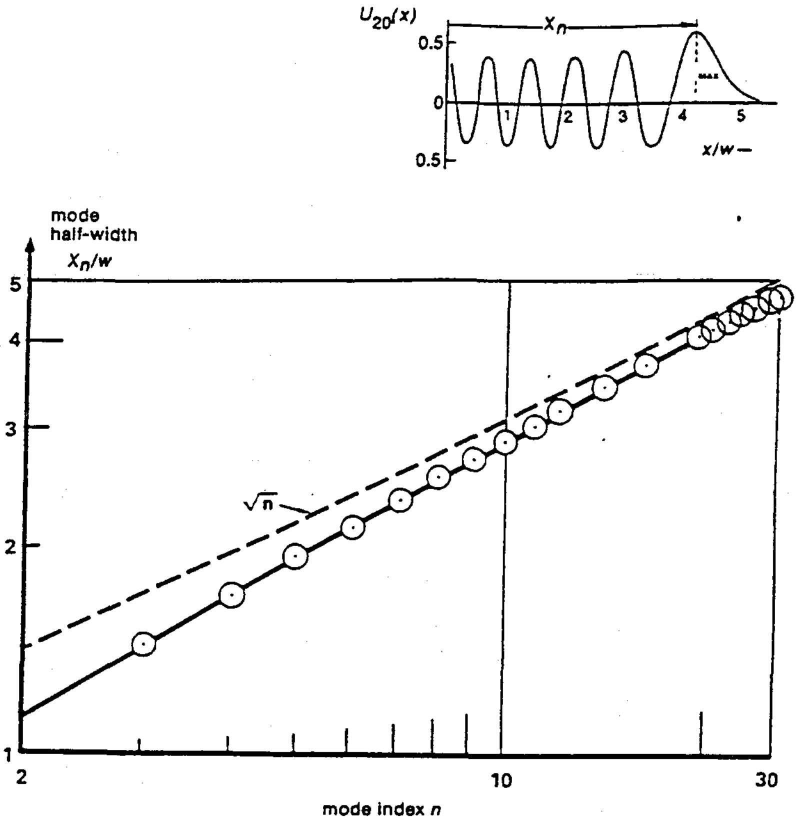


FIGURE 17.23
The outermost peak of an n-th order Hermite-gaussian mode occurs at $x_n \approx \sqrt{n} \times w$. The inset shows the $n = 20$ mode as an example.

this "mode" is more likely to represent a linear combination of the TEM$_{10}$ and TEM$_{01}$ Hermite-gaussian modes oscillating separately and independently, with slightly different oscillation frequencies because of the astigmatism introduced by the Brewster windows in the laser. The time-averaged total power output is then still circularly symmetric about the axis.

Higher-Order Mōde Sizes

It is obvious from inspection as well as from analytical approximations that higher-order Hermite-gaussian or Laguerre-gaussian modes spread out further in diameter as the mode index n (or p) increases. The mode pattern of the $n = 10$ mode function shown in Figure 17.23, for example, spreads out considerably farther than the lowest-order or $n = 0$ gaussian mode. This increase in mode diameter with increasing index n can be put on a quantitative footing as follows.

Let us use the peak of the outermost ripple in the Hermite-gaussian pattern, call its location x_n, as a convenient and fairly realistic measure of the spread or half-width of the Hermite-gaussian function. Numerically calculating and plotting the location of this outmost peak versus the mode index n, as in Figure 17.23—or alternatively, exploring more advanced descriptions of the mathematical properties of the Hermite-gaussian functions—then shows that this width increases with n in approximately the form

$$\text{mode half-width, } x_n \approx \sqrt{n} \times w. \tag{44}$$

In addition, since these higher-order modes have $n/2$ full ripples or periods of approximately equal width across the full width $2\sqrt{n}\,w$ of an n-th order Hermite-gaussian function, the spatial period Λ_n of the quasi-sinusoidal ripples associated with, or describable by, a Hermite-gaussian function of order n and spot size w is given by

$$\text{spatial period, } \Lambda_n \approx \frac{4w}{\sqrt{n}}. \tag{45}$$

Both of these criteria are very reasonable approximations to the mathematical properties of the Hermite-gaussian functions, especially for larger n.

Higher-Order Transverse Mode Aperturing

To illustrate the use of these quantities, suppose that we have an aperture of width or diameter $2a$, corresponding perhaps to a mode control aperture or an end mirror inside a laser cavity; and that we are considering expanding the amplitude distribution across that aperture using a set of Hermite-gaussian modes of spot size w at the plane of the aperture. It is then obvious that only those Hermite-gaussian modes of orders low enough so that $x_n \leq a$, or with indices n less than the value given by

$$n \leq N_{\max} \approx \left(\frac{a}{w}\right)^2 \tag{46}$$

will pass through this aperture, or oscillate inside this cavity with relatively negligible mode losses. Modes with higher mode indices will spill over past the edges of the aperture; and we can expect a rapid increase in energy losses caused by the aperture for all modes with indices larger than this value. (Obviously this criteria is the most accurate for apertures at least several times larger than w, since the sharpness of the outer edge transition becomes increasingly apparent at higher mode numbers.) Larger-diameter lasers often choose to oscillate in multiple transverse modes extending up to and including the highest-order transverse modes that will "fit" inside the laser tube or the laser mirrors according to this criterion, since all of these transverse modes will have comparatively low diffraction losses at the laser tube walls or mirror edges.

Transverse mode-control apertures are often placed inside stable laser cavities in order to attenuate or block higher-order modes from oscillating while producing minimal loss for the lowest-order TEM_{00} modes. A common rule of thumb for the necessary aperture size in low-gain lasers, such as for example He-Ne lasers, is that the mode control aperture should have an aperture size of diameter $2a \approx 3.5$ to $4.0 \times w$, or slightly larger than the $2a = \pi w$ or 99% criterion we introduced at the beginning of this section.

Numerical Hermite-Gaussian Mode Expansions

Suppose we wish to carry out a numerical expansion of some given (or perhaps unknown) function $f(x)$ across an aperture or strip of width $2a$ using a

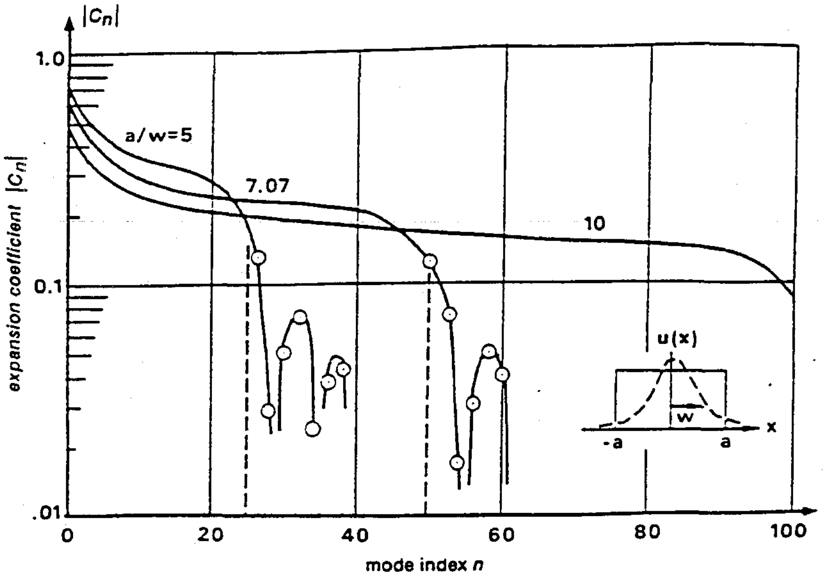


FIGURE 17.24
Expansion coefficient magnitudes $|c_n|$ versus mode index n for expanding a uniform square function of width $2a$ using a Hermite-gaussian basis set, for different choices of the parameter a/w.

Hermite-gaussian basis set in the form

$$f(x) = \sum_{n=0}^{N} c_n \tilde{u}_n(x;w), \qquad -a \le x \le a, \tag{47}$$

where $\tilde{u}_n(x;w)$ refers to an n-th order Hermite-gaussian function characterized by spot size w, and N is the maximum index value to be kept in a finite expansion. Let us explore some of the numerical considerations involved in this expansion, such as the optimum choice of the gaussian spot size w (assuming this to be a free parameter), and the number of terms N that we will need to keep in the summation.

The expansion coefficients for a given function $f(x)$ will be given by the overlap integrals

$$c_n = \int_{-a}^{a} f(x)\, \tilde{u}_n^*(x)\, dx. \tag{48}$$

Figure 17.24 shows, for example, how the expansion coefficient magnitudes $|c_n|$ will decrease in amplitude with increasing mode index n if we expand a simple rectangular function of width $2a$ using Hermite-gaussian basis sets of different fundamental spot size w. The dashed lines represent the values $N_{\max} = (a/w)^2$ in each situation. It is obvious from these plots that the amplitude of the expansion coefficients drops off rapidly in each situation as soon as n increases slightly beyond this value. This fall-off obviously occurs because the Hermite-gaussian modes of order higher than this extend past the edges of the aperture, or the square input function, and hence less and less of the Hermite-gaussian function falls within the overlap integral given in the preceding.

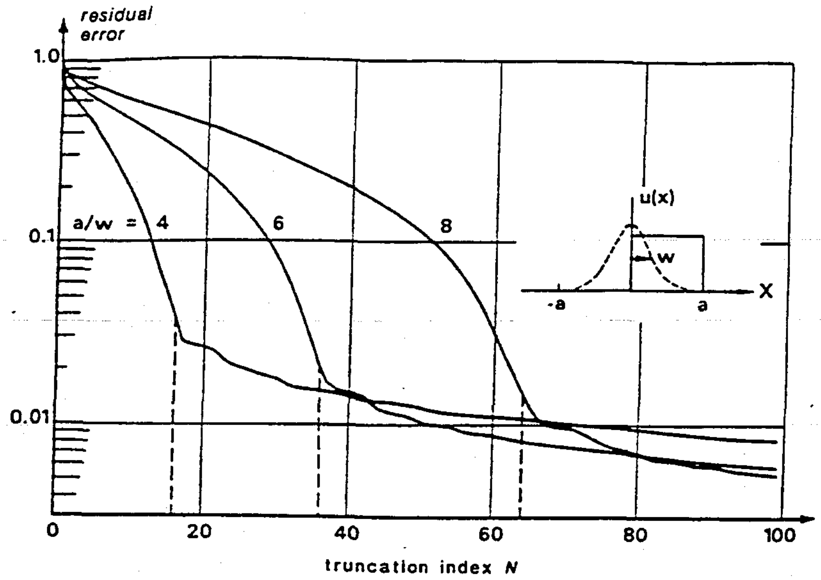


FIGURE 17.25
Residual mean-square error in approximating a half square of width a using a truncated series of Hermite-gaussian functions, plotted versus series truncation index, for different values of a/w.

As a slightly different illustration of the same point, Figure 17.25 shows similar results for the expansion of a half-square (or displaced square) function covering the range $0 \leq x \leq a$ using Hermite-gaussian basis functions. The quantity plotted in this situation is the residual mean-square error in the series approximation to the function $f(x)$ caused by truncating the series expansion at a maximum value N, for different choices of the ratio a/w. Again we see that the maximum error drops rapidly with increasing number of terms in the series expansion, but only up to a rather surprisingly sharp corner at $N \approx N_{\max} = (a/w)^2$. Beyond this value, keeping additional terms only causes a very slow further improvement in the accuracy with which the function is approximated by the finite series.

Spatial Frequency Considerations

Suppose as a more general example that we wish to describe an arbitrary function $f(x)$ across an aperture of width $2a$ with a finite sum of $N+1$ Hermite-gaussian functions $\tilde{u}_n(x;w)$, of arbitrary spot size w, for $0 \leq n \leq N$. How then should we select the spot size w to use in the expansion, and the maximum index N at which the series expansion is to be truncated?

To do this sensibly, we must first calculate (from experimental or other evidence) *what is the maximum spatial frequency or spatial period* Λ *of the fluctuations in the function to be expanded across the interval* $-a \leq x \leq a$? That is, we must pick a value of Λ such that the significant variations in $f(x)$ will be no more rapid than $\approx \cos 2\pi x/\Lambda$ at most.

We must then select values of w and N so that the highest-order Hermite-gaussian functions to be employed will simultaneously satisfy two criteria: they must at least fill the aperture, and they must at least handle the highest spatial variations in the signal. But these are equivalent to the two conditions

$$N \geq N_{\max} \equiv \left(\frac{a}{w}\right)^2 \qquad \text{and} \qquad \Lambda_N \approx \frac{4w}{\sqrt{N}} \leq \Lambda. \tag{49}$$

Satisfying these conditions simultaneously then leads to the spot-size and maximum-index criteria

$$w \leq \sqrt{\frac{a\Lambda}{4}} \qquad \text{and} \qquad N \geq \frac{4a}{\Lambda}. \tag{50}$$

The second of these criteria is obviously a Hermite-gaussian version of the familiar sampling theorem of Fourier transform theory, which says that to describe an arbitrary function which is bandlimited to a spatial frequency $2\pi/\Lambda$, we need at least two sample points per spatial period Λ. In the Hermite-gaussian analog we need $N = 4a/\Lambda$ samples in space across a width of $2a$, or equivalently $N = 4a/\Lambda$ coefficients in a Hermite-gaussian expansion.

REFERENCES

The width and divergence properties of higher-order Hermite-gaussian modes have also been analyzed by W. B. Bridges, "Divergence of high order gaussian modes," *Appl. Optics* 14, 2346–2347 (October 1975). Similar properties for Laguerre-gaussian modes are treated by R. L. Phillips and L. C. Andrews, "Spot size and divergence for Laguerre gaussian beams of any order," *Appl. Optics* 22, 643–644 (March 1 1983).

Some interesting properties of Laguerre-gaussian modes with zero radial order and high azimuthal orders are given by A. H. Paxton, "Propagation of high-order azimuthal Fourier terms of the amplitude distribution of a light beam: a useful feature," *J. Opt. Soc. Am. A* 1, 319–321 (March 1984).

Experimental studies of multimode beam divergence are given by H. M. Lamberton and V. G. Roper, "Beam divergence of a highly multimode CO_2 laser," *J. Phys. E: Sci. Instrum.* 11. 1102–1103 (1978); and J. T. Luxon and D. E. Parker, "Higher-order CO_2 laser beam spot size and depth of focus determined," *Appl. Optics* 20, 1933–1935 (June 1 1981).

More details on astigmatic higher-order modes and their aberrations are given in S. L. Chao and J. M. Forsyth, "Properties of high-order transverse modes in astigmatic laser cavities," *J. Opt. Soc. Am.* 65, 867–875 (August 1975).

An experimental method for analyzing the transverse mode content of laser beams is outlined in M. A. Golub, A. M. Prokhorov, I. N. Sisakyan, and V. A. Soffer, "Synthesis of spatial filters for investigation of the transverse mode composition of coherent radiation," *Sov. J. Quantum Electron.* 12, 1208–1209 (September 1982).

Problems for 17.5

1. *Instantaneous beam profile for the "donut mode."* Suppose you can take a series of instantaneous snapshots of the output beam intensity profile from a laser oscillating in the donut mode, in a situation where there is a finite frequency difference or beat frequency between the TEM_{01} and TEM_{10} modes (perhaps because some Brewster windows make the laser cavity slightly astigmatic and

slightly different in length for x-varying and y-varying modes). What will the time variation of the instantaneous intensity profile look like?

2. *Near-field beam with a central hole.* Consider a near-field beam pattern at the beam waist location which has a central hole in it (in one transverse direction), produced by combining $\tilde{u}_0(x)$ and $\tilde{u}_2(x)$ modes with correct amplitude ratio to produce a null value on axis (i.e., at $x = 0$). What will the far field pattern of this beam look like? Will there still be a hole on axis? Explain.

3. *Recursion relation for expanding a half-rectangle in Hermite-gaussian functions.* If you expand an off-center half-square function of width a into Hermite-gaussian functions, the first two expansion coefficients are

$$c_0 = (2\pi)^{1/4}(w/4a)^{1/2}\mathrm{erf}(a/w)$$
$$c_1 = (2\pi)^{1/4}(w/a)^{1/2}\left[1 - \exp(-a^2/w^2)\right].$$

Using the recursion and differential relations for Hermite-gaussian polynomials, show that all the higher-order expansion coefficients can be obtained from the recursion relation

$$c_n = \left(\frac{n-1}{n}\right)^{1/2} c_{n-2} - \frac{w}{n^{1/2}a}[\tilde{u}_{n-1}(a;w) - \tilde{u}_{n-1}(0;w)],$$

where the $\tilde{u}_n(x;w)$ are the properly normalized Hermite-gaussian functions of spot size w.

17.6 MULTIMODE OPTICAL BEAMS

Lasers that oscillate in multiple higher-order transverse modes are almost always considered as "bad" lasers, since they will have a far-field beam spread cosiderably larger than a well-behaved lowest-order single-transverse-mode laser. The quasi analytic results that we have just obtained for Hermite-gaussian mode expansions can also be applied to give a useful description of multimode or non-diffraction-limited laser beams, in the following fashion.

Description of a Multimode or Non-Diffraction-Limited Beam

Suppose that an oscillating laser emits a reasonably well-collimated but obviously multimode optical beam which occupies a width or diameter $2a$ in the transverse direction at the output from the laser. (By "collimated" we mean simply that any overall spherical curvature of the wavefronts emitted from the laser has been corrected by a suitable collimating lens.)

The far-field angular spread of the multimode beam coming from this laser will then be substantially larger than the value $\Delta\theta \approx \lambda/2a$ that would be characteristic of a more or less diffraction-limited optical beam. From another viewpoint, if the output beam from this laser consists of a mixture of a sizable number of different transverse modes, the wavefront at the laser output is likely to be quite random in character, with considerable spatial incoherence or variation in local amplitude and phase from point to point across the aperture.

How can such a strongly non-diffraction-limited laser beam then be described analytically—especially in situations where little information may be available concerning the detailed mode characteristics of the laser, and where all that is known for certain may be the near-field aperture width and the far-field angular spread of this beam?

Hermite-Gaussian Analysis of Multimode Optical Beams

One useful approach can be to analyze this beam as *if the output fields in* the beam are made up of, or can be analyzed as, a superposition of Hermite-gaussian modes having a characteristic spot size w_0. This assumption might apply quite well, for example, to the output beam from a laser with a stable gaussian resonator, such as we will describe in the following chapter, in which the natural resonator spot size is w_0, but the laser tube diameter or mirror diameter $2a$ is substantially larger than w_0. This laser may then oscillate simultaneously in *multiple transverse modes* which fill the entire diameter $2a$.

More generally, consider an arbitrary, irregular, multimode beam coming from any kind of laser cavity, stable or not; and assume that this beam has sizable fluctuations in amplitude and especially in phase across its diameter $2a$. Regardless of whether the underlying mode structure in this beam is gaussian, we can still use a set of Hermite-gaussian modes to expand the fields. Following the procedure outlined in the previous section, we can first ask what spot size w_0 and number of modes N we would need to choose so that the spatial frequencies and the spatial resolution of the set of Hermite-gaussian modes would be just adequate to describe the most rapid spatial variations across the aperture of that particular beam. We can then use these values to calculate the basis set of Hermite-gaussian functions which we can employ to describe that particular beam with adequate accuracy.

For an aperture of width $2a$, where a is at least a few times larger than w_0, the maximum number of Hermite-gaussian modes that will "fit" within the aperture, or the number of modes that will be needed to describe the fields in the aperture, will then be given by

$$N \approx N_{\text{max}} \equiv (a/w_0)^2. \qquad (51)$$

The corresponding maximum half-angle spread of the overall beam in the far field, using a near-field spot size of w_0 and modes of index running up to $n = N$, will then be

$$\theta_{\text{max}} \approx \sqrt{N} \times \theta_{1/e} = \frac{N^{1/2}\lambda}{\pi w_0} = \frac{a\lambda}{\pi w_0^2}. \qquad (52)$$

If we consider for simplicity a circular aperture of diameter $2a$, then the far field beam will also have a circular cross section of angular diameter $2\theta_{\text{max}}$. The product of the source aperture area $A \equiv \pi a^2$ and the far-field solid angular spread $\Omega \equiv \pi\theta_{\text{max}}^2$ will then be given by

$$A \times \Omega \equiv (\pi a^2) \times (\pi\theta_{\text{max}}^2) \approx (N\lambda)^2. \qquad (53)$$

This product for the multimode or non-diffraction-limited beam is then N^2 times the diffraction-limited value $A \times \Omega = \lambda^2$ we derived earlier for an ideal lowest-order gaussian beam.

"Times Diffraction Limited" (TDL)

An irregular or multimode laser beam which can be described in the fashion leading up to Equation 17.53 is often said to be "N times diffraction limited." That is, its far-field angular spread is $\approx N$ times as large in one dimension (or N^2 times in solid angle) as the diffraction-limited angular spread that would be obtained from a uniphase beam with a reasonably regular amplitude variation filling the same aperture. The quantity N is sometimes referred to as the "times diffraction limited" or "TDL" of the beam. Note that if this same beam is focused to a spot with a suitable lens, the diameter of the focused spot will also be $\approx N$ times the spot size that would be obtained with an ideal diffraction-limited beam.

This argument can also be applied in the reverse direction. That is, given a beam known to be N times diffraction limited (based on experimental data on its initial aperture size and its far-field beam spread), we can treat this beam analytically as if it were made up of a mixture of N^2 Hermite-gaussian modes, with spot size given by $w_0 \approx a/N^{1/2}$, and with mode amplitude coefficients assumed to be approximately equal or perhaps randomly distributed in amplitude. The relatively simple mathematical properties of the Hermite-gaussian modes then make it possible to calculate or at least estimate other properties of this beam that might be of interest (for example, perhaps the amount of harmonic generation it would produce in a given crystal).

It may seem somewhat inconsistent here to employ the Hermite-gaussian modes, which are characteristic of rectangular coordinates, and then compute areas and solid angles assuming circular beams, which might more accurately be described using cylindrical coordinates and Laguerre-gaussian functions. The only excuse is that the Hermite-gaussian properties are perhaps simpler and more familiar than the Laguerre-gaussians, and the right answer comes out by using formulas based on a circular aperture.

REFERENCES

For more discussion of these topics see the following chapter, and also the references at the end of the previous section. See also Z. Karny, S. Lavi, and O. Kafri, "Direct determination of the number of transverse modes of a light beam," *Optics Lett.* 8, 409–411 (July 1983).

For a somewhat different approach to other types of laser beam aberration, see C. B. Hogge, R. R. Butts, and M. Burlakoff, "Characteristics of phase-aberrated nondiffraction-limited laser beams," *Appl. Optics* 13, 1065–1070 (May 1984).

Problems for 17.6

1. *Beam expansion criteria using Laguerre-gaussian functions.* Look up sufficient information concerning the asymptotic properties of Laguerre polynomials and Laguerre-gaussian functions to calculate the number of Laguerre-gaussian functions of given w_0 that will fit within a circular aperture of radius a, and the number of azimuthal orders that will be involved; and then repeat the "TDL" argument of this section working in cylindrical coordinates.

REF. I

The detection of gravitational waves

Edited by

DAVID G. BLAIR

University of Western Australia

CAMBRIDGE UNIVERSITY PRESS

Cambridge

New York   Port Chester

Melbourne   Sydney

Published by the Press Syndicate of the University of Cambridge
The Pitt Building, Trumpington Street, Cambridge CB2 1RP
40 West 20th Street, New York NY 10011-4211, USA
10 Stamford Road, Oakleigh, Melbourne 3166, Australia

© Cambridge University Press 1991

First published 1991

Printed in Great Britain at the University Press, Cambridge

British Library cataloguing in publication data available

Library of Congress cataloguing in publication data available

ISBN 0 521 35278 9 hardback

12

Fabry–Perot cavity gravity-wave detectors

R. W. P. DREVER

12.1 Introduction

The gravitational wave detection technique discussed here is a long-baseline nearly-free-mass technique, devised initially with the aim of obtaining high gravity-wave sensitivity with minimum practicable cost. The distinctive part of the technique is the use as sensors of a pair of optical cavities formed between mirrors attached to test masses defining two perpendicular baselines, illuminated by an external laser source. To introduce the basic concept it may be useful to summarize the train of ideas which led up to it.

Experience and analyses in the early 1970s of resonant-bar gravity-wave detectors indicated that, although it is in principle possible to achieve by this technique the high sensitivity likely to be required for detection of expected astronomical sources, the small energy exchange with the gravitational wave leads to increasingly difficult experimental problems as sensitivity is improved. Alternative techniques using free test masses at large separations, monitored by optical or microwave methods, can sample much larger baselines and make relatively less serious any thermal, seismic, and amplifier noise, as well as the uncertainty-principle quantum limit for the test masses. Measurement of the small relative displacements involved, which might correspond at 1 kHz to strains of order one part in 10^{21} or less in a 1 kHz bandwidth, is however a serious challenge for interferometers of any kind. If a simple Michelson interferometer were used the photon shot noise limit would demand an impracticably high light flux. One way of improving sensitivity was proposed by R. Weiss (1972): the use of an optical delay line to cause the beam in each arm of a Michelson interferometer to pass many times between the test masses, so that the changes in total optical path lengths are increased. This is an effective technique, but it does require large mirrors and vacuum pipes of correspondingly large diameter to accommodate the many folded beams. Further, early experimental work with optical delay lines at the Max Planck Institute in Munich, and with a White cell multireflection system at the University of Glasgow, showed up a scattering problem. Scattering of light from one reflection spot to another can give competing optical paths which are not identical in the two arms. This can give phase noise if the light wavelength fluctuates, unless special precautions are taken.

The 'Fabry–Perot' technique was conceived as a means of minimizing mirror and pipe diameters, and at the same time avoiding the scattering noise problem (Drever *et al.*, 1983a). The test masses incorporate relatively small mirrors, one of which is partially transmitting, arranged to cause the light to reflect many times between the same pair of spots and give a sharp resonance condition analogous to that in the classical Fabry–Perot etalon. The system is then very sensitive to changes in either wavelength or in separation between the mirrors. If two similar cavities are illuminated via a beamsplitter by light from a single source, then the effect of wavelength changes can be discriminated against, and the system made sensitive to differential changes in cavity length which may be caused by gravitational radiation.

An arrangement in which, in principle, this might be done is indicated schematically in figure 12.1. Long optical cavities are formed between suitably curved mirrors attached to each pair of test masses, spanning the baselines between them. Light from a laser passes via a beamsplitter to both cavities, and it

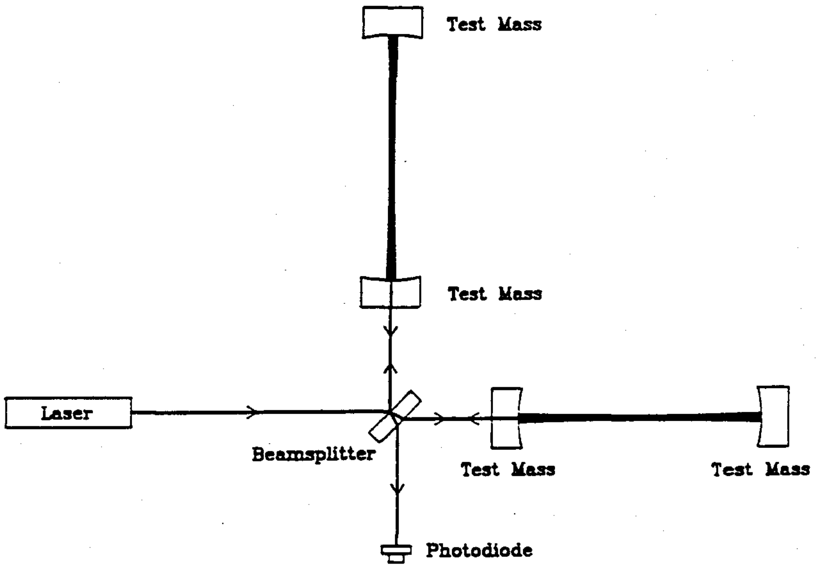


Figure 12.1. Simplified diagram to illustrate the principle of a basic Fabry–Perot cavity gravity-wave detector. With the cavities formed between each pair of test masses in resonance with the laser light, it is arranged that the output beams from the cavities arrive at the photodiode with almost exactly opposite phase to one another, giving near minimum intensity. A small differential change in the length of each cavity gives a change in phase of output light which is magnified by the number of reflections in the cavity, leading to a change in light intensity at the photodiode. (In this and following diagrams the thickness of lines representing light beams is intended to suggest the resonance modes in which light builds up inside the optical cavities, and should not be taken as a real representation of beam diameter or intensity.)

is arranged that resonance takes place in each cavity, giving effectively many superimposed bounces of the light between a single spot on each mirror. A change in length of either cavity causes a change in phase of the light leaving it. The length changes due to gravity waves would give differential phase changes in the two arms, which in principle may be measured with sensitivity similar to that of a delay-line system. However, the diameter required of the mirrors and the space required for the light beam is significantly less for the Fabry–Perot method.

For this technique to be practicable, it is necessary to have ways of obtaining a suitably stable light frequency, and of monitoring small changes in phase in the cavities. Techniques for carrying out these tasks were an important part of the original Fabry–Perot gravity-wave detector concept, and will be discussed below.

It may be noted here that if a sufficiently stable light source were available, a single optical cavity which monitored the distance between one pair of test masses might in principle detect gravity waves. Unfortunately at present there does not seem to be any light source which has adequate frequency stability over the relevant time scales, apart from a laser stabilized to a cavity similar to that used for the test mass measurement itself. The differential use of at least two similar cavities, oriented to be affected differently by the gravitational wave, is necessary in the current Fabry–Perot technique.

Since the basic Fabry–Perot detector system was conceived, several further methods for enhancing sensitivity, and for exploiting the small diameter of the cavity beams to make possible operation of several different interferometers within the same vacuum system, have been proposed (Drever, 1983). The possibility of making multiple interferometers share one set of beam pipes seems likely in the long term to be an important aspect of this optical system.

12.2 Principle of basic interferometer

To achieve high sensitivity in a Fabry–Perot cavity it is desirable to reflect the light back and forth between the mirrors for a time approaching the period of the gravitational wave; so such a cavity will have an extremely narrow optical bandwidth, typically less than 1 kHz. It is clear that the light source used to illuminate the system must have a bandwidth no larger than this, and a key part of the original idea for a Fabry–Perot gravity-wave detector was a new way of stabilizing the frequency of the light from a laser, to a cavity of this type (Drever *et al.*, 1983b). Previous laser stabilizing systems usually involved operating the laser slightly down one side of the resonance peak in the response of a cavity, and then using changes in the intensity of light transmitted by the cavity as a measure of wavelength changes. The concept here involves measurement of phase difference – and not intensity – between light within the cavity and light from the laser, and use of this phase difference as a measure of deviation from resonance. A high gain servo system can use the phase signal to bring the laser into precise

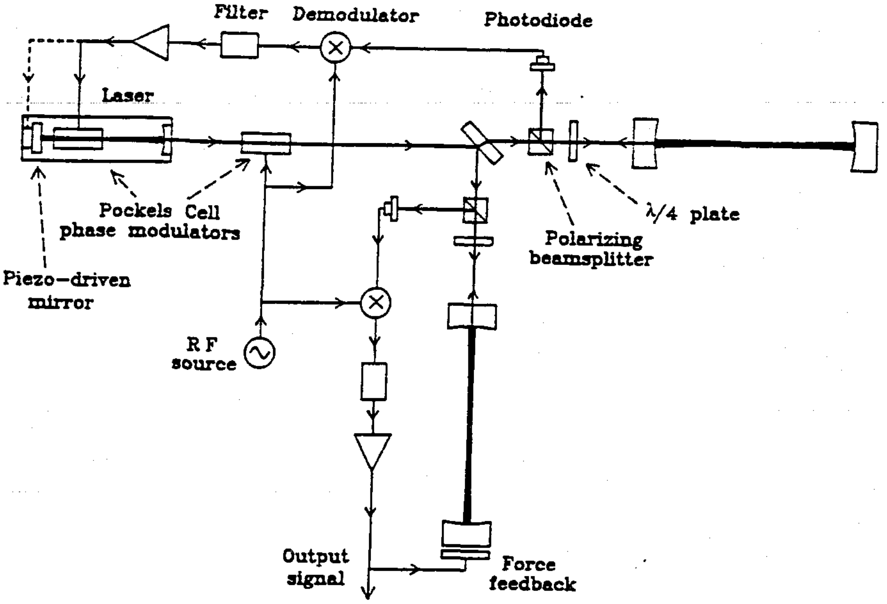


Figure 12.2. A more practical arrangement for a Fabry–Perot cavity gravity-wave detector, in which the laser wavelength is defined by the cavity on the right of the diagram, and the force required to maintain the second (lower) cavity in resonance with this is monitored. The diagram is simplified – auxiliary components are required to condition the beams and control the test masses in a real system.

resonance. A radiofrequency phase modulation technique can be employed to reduce effects of low frequency laser intensity noise; the principle is illustrated in a simplified diagram of one version of a Fabry–Perot gravity-wave interferometer system given in figure 12.2. In this diagram the optical cavity on the right-hand side is used to stabilize the wavelength of the laser. We discuss this part of the system first.

Plane polarized light from the laser is phase modulated by passage through a Pockels cell, at a frequency in the range of 10 to 20 MHz, and passes to a 50% beamsplitter. Half of the light goes to the right-hand optical cavity, through a polarizing beamsplitter and a quarter-wave plate. The axes of the quarter-wave plate are oriented at 45 degrees to the polarization of the input light, so that circularly polarized light enters the cavity. Light coming back from the input mirror of the cavity is circularly polarized in the opposite sense, is transformed into plane polarized light with polarization orthogonal to that of the input beam, and is reflected by the polarizing beamsplitter to the photodetector at the top of the diagram. The light arriving at this photodiode can be considered to have two components: the phase-modulated laser light directly reflected by the input cavity mirror; and light emerging from within the cavity – which has built up over the cavity storage time and thus has had its modulation sidebands removed. If the laser light is precisely in resonance with the cavity these two components have opposite average phase, and the photodiode output has no component at the modulation frequency. If the laser is slightly off resonance, the photodiode gives a signal at the modulation frequency whose amplitude and phase indicate the magnitude and sign of the error. Demodulation of the photodiode signal by a coherent demodulator gives a voltage signal which may be applied to a second Pockels cell within the laser cavity itself, so that the wavelength of the light from the laser is driven closer to the cavity resonance, and the laser becomes locked in wavelength to the cavity.

The bandwidth achievable with this stabilization system is not limited in principle by the storage time of the optical cavity, and a very high loop gain can be obtained with a suitable amplifier system. The wavelength of the laser light is then tightly locked to the spacing between the cavity mirrors, which in the diagram are shown attached to a pair of test masses which form one arm of a gravity-wave detector.

To complete the gravity-wave detector the same technique may be used a second time – but in this case it is employed to lock the length of an optical cavity spanning the second arm of the detector to the wavelength of the now-stabilized laser light. In figure 12.2 the light leaving the beamsplitter in the downwards direction enters the second optical cavity, and, as before, demodulation of the light coming back from the input cavity mirror gives a signal which measures the deviation from resonance. In this case the signal is amplified and used to apply an electrostatic force to the lower test mass, in a direction to bring the cavity closer to resonance. If the loop gains in the servo loops for both cavities are high, then the spacing between the two pairs of mirrors will be locked together, and the electrostatic force required to maintain this condition can become a measure of differential forces induced by gravitational radiation or other phenomena.

The interferometer configuration just described is in principle slightly less sensitive at its photon shot noise limit than the arrangement indicated in figure 12.1, since the output beams from the two cavities are not recombined together at the beamsplitter, and separate measurement of phase is made in each arm instead of a single differential measurement. With available mirrors in a laboratory-scale interferometer the difference is small, although in a large system the recombined arrangement may give a strain sensitivity which is better by a factor of nearly $2\sqrt{2}$, and it facilitates high power operation. In designs for practical versions of Fabry–Perot interferometers with recombined output beams, the laser may be stabilized to one of the main cavities, or to the average of both of them, using only a small part of the returned light; or a separate cavity to which the main arms are indirectly locked may be used as a primary reference for the laser stabilization.

The sensitivity of an interferometer of this type may be limited by many factors, but an important fundamental one comes from the fluctuations in light intensity detected by the photodiode arising from 'photon shot noise', the quantum fluctuations in the number of detected photons. For a given input light power, the photon shot noise sets a limit to the accuracy of measurement of optical phase difference, and thus to change in optical path within the cavities. A small motion of one of the test masses will cause a change in optical path which is increased by the number of times the light traverses the distance between the mirrors forming the cavity. Early in this work it was expected that losses in the mirrors would limit the number of reflections which could be used; and in this regime the sensitivity would improve with the effective number of reflections. However, the development of low-loss mirrors for other applications has led to the potential availability of multilayer dielectric mirrors having reflective losses less than one part in 10^4, and with such mirrors the total time that the light spends within the cavity becomes significant. If the effective storage time of the light is longer than the period, or the time scale, of the gravity wave there may be reversals in the mirror motion, and some resultant cancellation of output signal. The cavity behaves in some respects like an integrator, and as light storage time increases, displacement sensitivity improves up to a limiting value which is approached when the storage time matches the time scale of the gravity wave. In the configuration of figure 12.1, this limiting sensitivity, for detection at unity signal-to-noise ratio of a pulse of strain amplitude h, is given approximately by $h = [(\lambda f^3 \pi \hbar)/(I\eta c)]^{1/2}$.

Here λ is the wavelength of the light; f is the main frequency in the spectrum of the gravity-wave pulse: the amplitude h is assumed to be measured over a bandwidth equal to this; $\hbar$ = Planck's constant/(2π); I = input light power, η = photodiode quantum efficiency; and c = velocity of light. For $f = 1000$ Hz, $\lambda = 514$ nm, $I = 10$ W, and a photodiode of near unity quantum efficiency, this indicates a potential gravity-wave amplitude sensitivity of order 8×10^{-21}.

It may be noted that in this case the shot noise limit to sensitivity is independent of arm length – but in practice limits to sensitivity set by other factors, such as stochastic forces, thermal noise, or the uncertainty-principle quantum limit for the test masses, make a long baseline essential.

The diagrams shown are highly simplified, and in practice it is necessary to take precautions against many phenomena which may disturb sensitive optical measurements, such as lateral and angular motions of the laser beams, fluctuations in laser intensity, and spurious reflections and scattering between components. This makes the real system relatively complex. We will discuss some of these practical issues later; we concentrate on basic principles at this stage.

In an interferometer system in which the outputs from two arms are recombined at a beamsplitter, there are several advantages in adjusting the optical paths so that the output beam to the photodiode is near an intensity minimum – the photodiode is near a 'dark fringe'. It can be shown that if the measurement is limited by photon shot noise, and amplifier noise is unimportant, this gives in principle the maximum possible sensitivity. In practice a high frequency phase modulation technique may be employed to minimize effects of low frequency laser amplitude noise, and a feedback servo used to keep the interferometer centered on an intensity minimum. This arrangement is particularly useful when light power levels in the interferometer are high, for the photodiode has to handle only a small fraction of the total light flux, and a relatively small, high-sensitivity, photodiode may be used. With the photodiode near a dark fringe, most of the light from the interferometer passes out through the other side of the beamsplitter, in a direction towards the laser, and is available for other uses if required.

The potential sensitivity of this system is interesting, but higher performance would be desirable. Some improvement can be obtained by increasing the laser power, but there are practical limits to this. However, some further methods for improving sensitivity of interferometers of this type without increasing input power have been devised, and are likely to be important in large-scale gravity-wave detectors.

12.3 Enhancement of sensitivity by light recycling

When experiments with Fabry–Perot interferometers at Caltech showed that mirrors were potentially available which could give light storage times in laboratory-scale optical cavities comparable with the time scales of signals of interest, and could be expected to give much longer storage times in larger systems, it began to become evident that we might be entering a realm of new possibilities in optical experiments. The amount of light which has to be absorbed by a photodetector to give a sensitive null measurement is in principle very small, and much less than the circulating light power required in the interferometer arms. If scattering and dissipation can be kept small in the whole optical system there may be little loss of light in a period corresponding to the time scale of the gravity wave, and much of the light may still be present at the end of this measurement period. Thus there is a possibility for using most of the light again. In a suitable optical system the light may be 'recycled' many times, so that the light power effective for the measurement may be significantly larger than the output of the laser itself. The basic concepts (Drever, 1983) may be applied to both delay-line Michelson interferometers and to Fabry–Perot interferometers with beam recombining; we discuss only the latter case here.

With the basic interferometer system shown in figure 12.1, the main optical modification to achieve broadband recycling involves the introduction of an additional mirror between the laser and the beamsplitter to return light to the interferometer, with servo control to maintain correct phase of the recycled light. A schematic diagram of the resulting arrangement is given in figure 12.3. The recycling mirror forms what is effectively a large Fabry–Perot cavity encompassing the two main cavities and the beamsplitter, and its position and reflectivity are

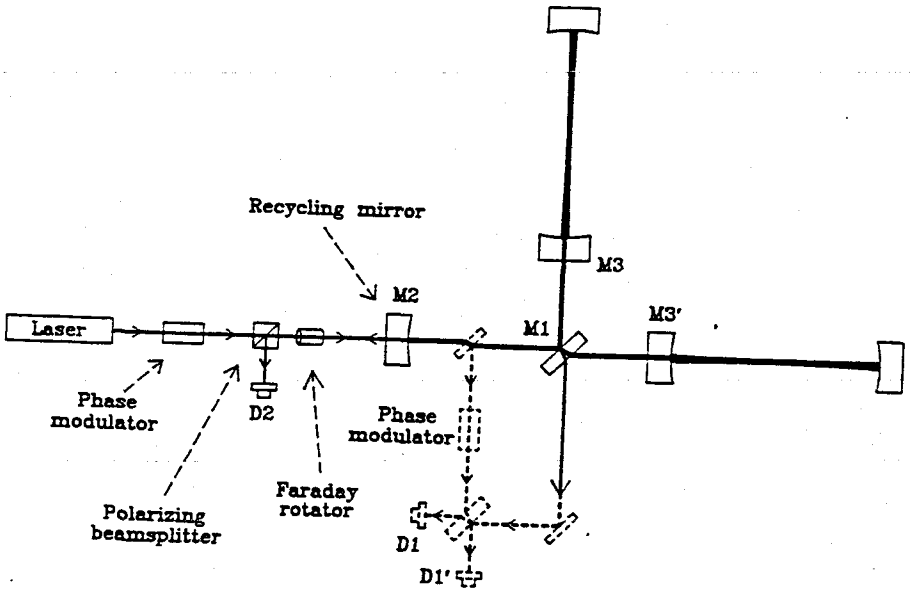


Figure 12.3. Basic arrangement of a Fabry–Perot interferometer with broadband light recycling, using mirror M2 to return light to the system. Effects of laser intensity noise may be reduced by applying high-frequency modulation: one way of doing this is shown by broken lines. A small part of the steady stored light in the recycling system is taken out to a side arm, phase modulated, and added to the differential output of the interferometer at an auxiliary beamsplitter. Phase changes in the main interferometer output can be determined from the signals from photodiodes D1 and D1'.

chosen to give maximum resonance build-up of light in the whole system. The method used to maintain resonance is similar to that described above for holding resonance between the laser in figure 12.2 and the cavity on the right-hand side of that figure. The laser beam is phase modulated at high frequency, and light returning from the recycling mirror is diverted to a photodiode D2, coherently demodulated, and the resulting phase error signal used to adjust either the laser frequency or the position of the recycling mirror. In the former case, the two interferometer cavities are maintained in resonance by separate servo systems using auxiliary photodiodes, not shown, to adjust the cavity lengths by applying feedback forces to the end masses in a way similar to that indicated for the lower cavity in figure 12.2.

For normal operation the reflectivities of the cavity input mirrors for each arm are chosen to give cavity storage times which approximately match the time scale of the gravity waves of interest. Maximum build-up of light is obtained with a transmission for the recycling mirror which suitably matches the total losses in the system.

The number of times that the light can be recycled through the interferometer depends on many factors. These include losses in the mirrors, the beamsplitter, and other optical components; light removed for auxiliary servo systems; and any residual mismatch in the wavefronts of the light from the two main cavities when they are recombined at the beamsplitter. Reduction of wavefront mismatch by use of a specially figured compensating plate in one arm of the interferometer may be practicable.

If it is assumed that all losses are made small except those associated with cavity mirrors of maximum reflectivity R, then the photon shot noise limited sensitivity at unity signal to noise ratio may approach $h=[(\lambda f^3\hbar\{1-R)\}/(LI\eta)]^{1/2}$.

For similar parameters to those used in section 12.2 above, with $R=0.99995$, $L=4$ km, and $I=50$ W, this gives a sensitivity $h=1.2\times10^{-22}$.

This basic type of recycling interferometer seems a very efficient way of achieving a sensitive broadband instrument with minimum light input power. It seems probable that when techniques for using squeezed quantum states of light to reduce photon shot noise become sufficiently developed, these may be applied to essentially the same interferometer design to give further improvement in performance. However, another optical technique, 'resonant recycling', has been devised to obtain further improvement in shot noise limit to sensitivity in a narrow bandwidth (Drever, 1983). In the original version of this scheme, the cavities in the two arms of the interferometer are coupled together by a high-reflectivity mirror in such a way that the normal resonance modes are split, with a splitting which matches the period of the gravity waves of interest. One of the two resonances is then made to match the frequency of light from the laser, while the other matches the frequency of a sideband of this light produced by the mirror motions due to the gravity wave: both resonances enhance the output signal. Another variant of this basic idea has been proposed recently by B. J. Meers: here the second resonance is obtained by putting an additional recycling mirror at the output of the type of broadband recycling system shown in figure 12.3. The latter configuration has been called 'dual recycling'.

12.4 Resonant recycling and dual recycling

12.4.1 Resonant recycling

The original idea for resonant recycling arose while trying to find a way of detecting the very weak periodic gravitational radiation signals expected from pulsars. It grew out of the realization that if light could be stored in an arm of an interferometer for a time matching half a period of the gravity wave it might be made to exchange with light from the other arm at the same time as the phase of the gravity wave reverses, so that a signal might build up continuously (Drever, 1983). The first description of this concept, applied to both a delay line and a

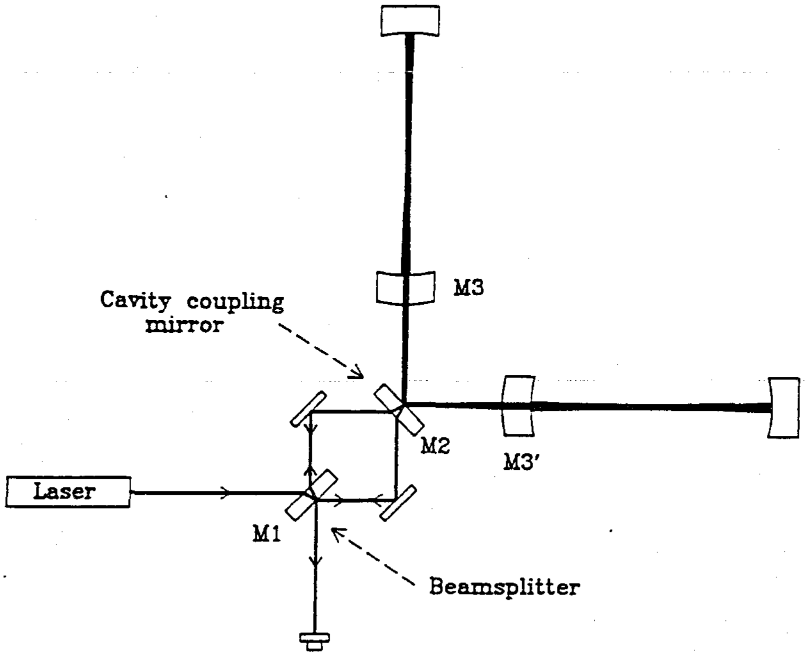


Figure 12.4. One configuration for a resonant recycling interferometer, with the main cavities coupled by mirror M2 to give one resonance at the frequency of the laser and a second resonance at a sideband frequency produced by modulation by the gravitational wave.

Fabry–Perot interferometer, was given in these terms; but in the case of the Fabry–Perot interferometer it is simpler to consider the mechanism as production of sidebands of the laser frequency by the gravity wave, and the system as providing a resonant enhancement of a sideband signal. A simple form of resonant recycling Fabry–Perot interferometer is shown in figure 12.4. Here the optical cavities in each arm are directly coupled to one another by a mirror M2. This has a high reflectivity, chosen so that its transmission approximately matches the total losses in the pair of cavities, giving maximum build-up of laser light in this part of the system.

If we consider a single axial resonance mode of the same order in each cavity, then the coupling has the effect of giving the combined system two modes of oscillation, a symmetrical one – pumped by the laser – with the light in phase in the two cavities, and an antisymmetrical one in which it is out of phase. The degree of coupling between the two cavities is determined by the transmission of their input mirrors M3 and M3′ and also by interference effects in the optical path between these mirrors. To tune the interferometer, the mirrors M3 and M3′ are adjusted so that the frequency splitting matches the frequency of the gravity wave of interest. A periodic gravity wave will then move the end mirrors and frequency modulate the cavities in such a way that sideband light which resonates with the antisymmetrical mode is parametrically pumped into that mode. There is thus both a resonance build-up of the laser light and of the signal produced by the gravity wave. For a fast pulsar signal the build-up can extend over many cycles of the gravity wave, and an enhancement of sensitivity may be obtained which is approximately equal to the number of gravity-wave periods in the light storage time. For a measurement in which the signal is integrated coherently over a total time τ, the sensitivity for amplitude h of a periodic gravity wave, at unity signal to noise ratio, is given approximately by $h = [(\lambda\hbar c(1-R)^2)/(L^2 I\eta\pi\tau)]^{1/2}$.

Here, as before, it is assumed that losses are determined only by the reflectivity R of the main mirrors. For $I = 10$ W, $\tau = 10^7$ s, and all other parameters as in the example in section 12.3, this gives a photon shot noise limit to sensitivity of the order of $h = 1 \times 10^{-28}$.

Other noise sources may, of course, prevent this sensitivity from being obtained in practice.

One possible way of exciting the interferometer and detecting the gravity-wave signal is indicated in figure 12.4. The laser sends beams in two directions to the mirror M2 via a 50% beamsplitter, and light from the antisymmetrical cavity mode returning from M2 recombines at the beamsplitter and emerges downwards towards the photodiode. Discrimination against laser intensity noise may be improved by introducing high frequency phase modulation, for example by a method like that shown in figure 12.3, in which some phase-modulated light is coherently added to the main output signal. Techniques of this type can be applied to most of these interferometer configurations, but are omitted here for simplicity.

12.4.2 Dual recycling

An alternative 'dual recycling' method (Meers, 1988) of obtaining the two resonances required for a resonant recycling interferometer is indicated in figure 12.5. Here a second recycling mirror M4 is added to the output of a standard broadband recycling interferometer. The first recycling mirror M2 acts just as in the broadband recycling system of section 12.3 to increase the laser light flux in the interferometer arms. The second recycling mirror M4 is adjusted to give a resonance with the two main cavities combined in antiphase with one another at a sideband of the laser frequency produced by the gravity wave. This enhances the gravity-wave signal just as in the resonant recycling system of section 12.4.1, and detection of the signal may be facilitated by a similar heterodyne system.

In a narrow-band operation the maximum sensitivity of this system is similar to that of the resonant recycling system of section 12.4.1, although there are differences. The presence of the two extra mirrors in the resonant part of the dual recycling system can potentially give it more flexibility in adjustment of overall response. By broadening the bandwidth of the output cavity, while keeping the

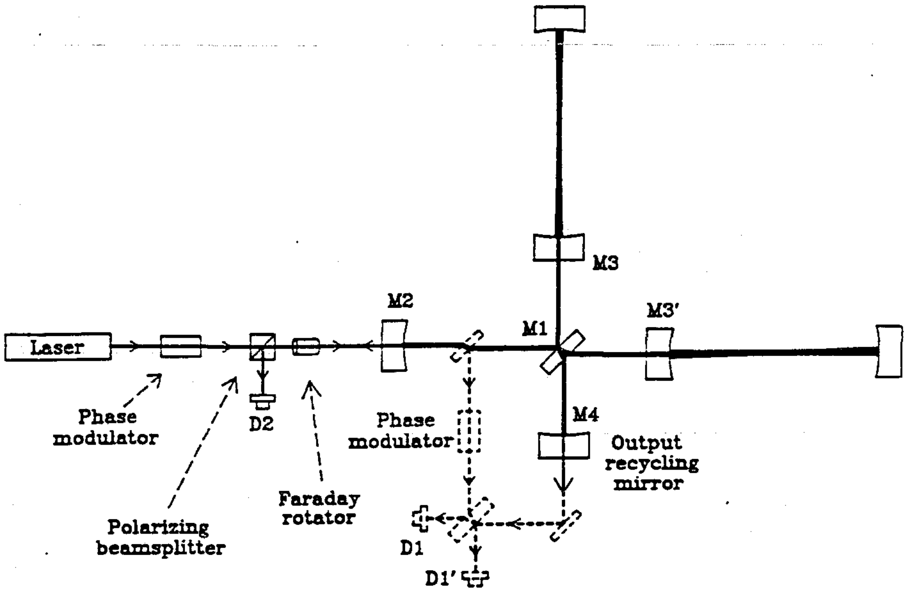


Figure 12.5. A dual recycling interferometer configuration. The output mirror M4 provides the second resonance required for resonant recycling, with the input mirror M2 giving the resonance at the laser frequency. The high-frequency side-arm modulation method outlined in figure 12.3, indicated by broken lines, may be used here also to reduce intensity noise.

input cavity narrow, a wider bandwidth can be obtained for a given peak sensitivity. However, losses in substrates may be higher. It may be noted that addition of a recycling mirror to the resonant recycling system of figure 12.4, in front of the beamsplitter M1, can give that system some similar characteristics.

12.4.3 Resonant recycling interferometers in general

The two resonant recycling systems discussed above are examples out of a wider range of configurations which can be envisaged. These systems are likely to have more general application than investigation of periodic signals alone. They can be adjusted for relatively broadband response, and then may be competitive with the broadband recycling systems of section 12.3 for pulse searches; while for particular pulse waveforms they may be more sensitive. Also in searches for a stochastic background of gravitational radiation by cross-correlation techniques they may give higher overall sensitivity. Further, it can be expected that these techniques will be relatively tolerant of optical distortions and figure errors, for the exchange of light between the interferometer arms tends to average out some imperfections. In general, resonant recycling systems of various kinds look promising for the future.

12.5 Other techniques for achieving high sensitivity

12.5.1 Use of squeezed light techniques

Theoretical and experimental work on quantum limits to optical measurements has shown that it is possible to modify the statistical properties of the light in an interferometer in such a way that a suitable measurement process may show smaller fluctuations than would be obtained from photon shot noise at the same power level in the usual system; and this may be applied to Fabry–Perot systems as well as to other types of interferometer. It has been noted by C. Caves (1981) and others that a fundamental component of the statistical fluctuations in the output from a Michelson interferometer can be regarded as arising from vacuum fluctuations at optical frequency which enter the system through the normally unused direction into the beamsplitter. It is possible to modify these fluctuations by parametric pumping with a non-linear optical medium which covers the open beamsplitter port, in such a way that the fluctuations in one quadrature phase of the output light from the interferometer are increased while those in the orthogonal quadrature phase are decreased. Effectively the fluctuations become 'squeezed' so that they become smaller in one quadrature phase than in the other. Advantage may be taken of this by arranging that detection of the output signal is made using the quieter quadrature phase alone; which may be done with another coherently pumped optical parametric amplifier. The ultimate noise in this, and other, squeezed light measurement techniques can in principle be less than in normal operation of an interferometer with the same light power.

Application of these techniques to Fabry–Perot interferometers shows promise for the future, and may lead to improvements in sensitivity without increase in light power. Some limitations should be mentioned, however. Optical losses at mirrors and beamsplitters can introduce fluctuations which are not reduced by the squeezing, and may obviate the potential gains in sensitivity. In particular, when recycling techniques are pushed to a point where losses in the interferometer arms prior to detection are the limiting factor, as in resonant recycling and dual recycling for narrow-band operation, squeezing techniques may not give significant advantage. However, with broadband recycling and wide band dual recycling where light is detected before losses have become large, squeezing may in principle give a useful improvement in sensitivity.

12.5.2 Use of auxiliary interferometers to reduce seismic noise

Up to this point we have considered photon shot noise as the main source limiting the sensitivity of an interferometric gravity-wave detector; but other noise sources may also limit practical systems. Seismic motions of the ground, which may produce stochastic forces on the test masses through the mass suspension systems, are a potential noise source which may limit performance at low frequencies. Passive isolation techniques can provide good seismic isolation at frequencies above a few hundred hertz, but their effectiveness tends to decrease as frequency falls. Various types of active feedback isolation systems have been proposed and developed with the aim of improving the low frequency performance of gravity-wave detection systems.

The relatively small diameter of the cavity mirrors and light beams in a Fabry–Perot gravity-wave detector makes it reasonable to consider putting several different interferometer beams within one set of vacuum pipes. This has made more practical a fairly simple technique for reducing low frequency seismic noise which was conceived earlier. An auxiliary interferometer is used to monitor relative motions between the suspension points of the test masses at the opposite ends of each interferometer arm. The output from this interferometer is applied to a transducer system which moves the suspension point at one end of the arm in such a way that relative motion along the direction of the arm is decreased. With the pair of suspension points locked together in this way, seismic motion components in the direction of the arm become effectively equal for the two test masses, and thus tend to cancel.

This active isolation technique is a fairly simple one for a Fabry–Perot interferometer, and can be expected to considerably reduce relative motions of the test masses. Experimental work on a system of this type has been done by Y. T. Chen. Such an arrangement is expected to usefully improve detector sensitivity at low frequencies. Further, as it reduces the dynamic range requirements for the servo systems controlling the overall lengths of the main interferometer arms it may contribute to improved performance at higher frequencies also.

12.6 Experimental strategies with Fabry–Perot systems

The Fabry–Perot interferometer system for gravity-wave detection was conceived initially mainly as a compact, and therefore relatively economical, optical system for achieving the required sensitivity. However, the compact nature of the mirrors and beams relative to those required for delay-line detectors has gradually led to new ideas about ways of using the interferometers which have influenced plans for experimental strategies for gravity-wave searches. For example, experience with searches for gravity-wave bursts using earlier broadband bar detectors showed the major advantage to be gained in discrimination against spurious signals by having at least a pair of detectors operating in coincidence at one site, along with one or more at a distant location to verify possible detections (Drever *et al.*, 1973). With laser detectors it was soon realized that if two interferometers could share a single set of beam pipes the cost of adding the second would be relatively small. Further, when planning for kilometre-scale interferometer facilities began about six years ago it became apparent that it would be wise to make any large facilities capable of accommodating delay-line interferometers as well as Fabry–Perot ones, and this implies that the beam pipes would be capable of accommodating several

Fabry–Perot interferometers. Gradually it became apparent that it could be economical to design large detection facilities to house several detectors, and this in turn has affected ideas for experimental search strategies. We summarize some of the concepts here, and their influence on the design of large detection facilities.

12.6.1 Use of interferometers of different length

The first proposed experimental strategy which became more viable with the possibility of sharing beam pipes was the concept of operating interferometers of different length side by side. This idea grew from attempts to find ways of reducing the cost of using pairs of interferometers at one site to improve the discrimination of gravity-wave signals from those due to other phenomena (Drever *et al.*, 1983c). One early arrangement proposed involved sharing of test masses and of some vacuum tanks, with separate beam pipes forming the sides of a square to give discrimination against non-gravity-wave disturbances. However, a performance nearly as good may be achieved at lower cost by sharing beam pipes, so that a single L-shaped system accommodates two interferometers which are made to respond differentially to non-gravity-wave effects by making their arm lengths significantly different. A convenient ratio of arm lengths is around 2:1, so that a vacuum tank half way along each main arm houses an end mass for the shorter interferometer.

In this case a gravitational wave will produce displacements (or forces) twice as large in the long interferometer as in the shorter one: very few other disturbances will produce displacements in this ratio. Thus a disturbance of any one of the test masses, a release of strain in any one suspension wire, or a change in optical path produced by a release of a burst of gas in any of the beam pipes, can be discriminated against. The discrimination is of course only effective with signals which are several times the mean noise – and these are likely to be the most important signals. For signals near the noise level the technique can give only marginal discrimination, but even in this case the addition of the two interferometer outputs will give slightly better sensitivity than that of a single interferometer alone.

This two-interferometer system seems capable of giving a sufficiently good signature for a gravity wave to reject many local spurious phenomena. For positive detection of a gravity-wave pulse, however, a coincidence with another detector at a distant site is still essential. The two-interferometer local system, giving signals in the required ratio, can nevertheless add greatly to the significance of a coincidence with a remote interferometer, and can reduce accidental coincidence rates by a large factor. The effective detection sensitivity for low rate burst events can thus be significantly improved.

12.6.2 Concurrent operation of interferometers for different purposes

The feasibility of operating several different interferometers within the same vacuum system can facilitate a comprehensive search for gravitational waves in other ways. In particular it can make it possible to develop new interferometer techniques without interrupting searches with existing interferometers; and it can make it viable to operate specialized interferometer systems designed to give maximum sensitivity for a specific kind of signal without stopping more general broad-range searches. We consider some aspects of these possibilities in turn.

(i) Gravity-wave searches and interferometer development

Fully realistic testing of long-baseline interferometers can only be carried out when the necessary large vacuum pipe systems become available, and it can be expected that the first interferometers to go into these systems will have far from the ultimate performance. Further development of the techniques will then be possible – and will be very important to achieve high sensitivity. Here the possibility of operating several interferometers within the same vacuum facilities can give another bonus – it can make it practicable to use one interferometer as a testbed for new developments without interrupting gravity-wave searches being made with other interferometers. A dilemma familiar to experimenters – how to divide instrument time between observation and technical development – may thus be avoided. This mode of operation – performance of continuing gravity-wave searches and observation concurrently with the development of new techniques – may be a very effective one, and is an important practical reason for attempting to accommodate additional interferometers in a large-scale facility.

(ii) Optimized searches for specific signals

It has been shown above how a suitable choice of optical system parameters can enhance the performance of an interferometer for certain types of signals; a particular example being a search for a continuing periodic wave, for which the optimum detector would have the narrowest bandwidth obtainable with available mirror losses, centred at the appropriate frequency. Even for broadband searches, the design of an interferometer for optimum performance at low frequencies differs from that for higher frequencies. In the latter case, for example, it is advantageous to make test masses relatively small so that frequencies of internal resonance modes are well above the gravity-wave frequency of interest. In this case also, the optimum cavity storage time is short; and the optimum light power on the test masses – where near the uncertainty-principle limit quantum fluctuations in radiation pressure may balance photodiode shot noise – is high. A complete search for gravity-wave signals is likely to involve a range of interferometer parameters and different types of interferometer, and here again it is clear that use of a number of interferometers in one vacuum system may make it practicable to carry out more sensitive searches in a given time than with a single one.

12.6.3 Detector and vacuum system arrangements to facilitate efficient experiments

The possibilities of running several interferometer beams within a single set of vacuum pipes can be used to the fullest extent only in vacuum systems designed for this mode of operation, and there has already been some development in concepts for doing this. In designing such a system using the typical interferometer layouts indicated earlier it can be particularly difficult to find enough space for those components which would be naturally located at the intersection of the two main arms, such as beamsplitters. To alleviate congestion in the intersection region, a concept in which separate vacuum tanks are used for test masses and beamsplitters was introduced by the writer in 1984. Tanks for test masses are located along the two main arms of the system, and beamsplitter tanks are placed along a line bisecting the angle between the two arms, with auxiliary evacuated pipes linking the tanks. One version of this arrangement is shown in figure 12.6. Auxiliary mirrors deflect the light from the beamsplitters to the appropriate main Fabry–Perot cavity. This type of system provides several virtual intersection regions, giving useful space for beamsplitters and other optical components outside the main interferometer arms. It was the basis for early plans for a system which could accommodate up to six separate Fabry–Perot interferometers. A further concept of practical importance was added more recently: the test masses on their suspensions are lowered into position on the main beam line through horizontal gate valves above the main vacuum pipe. A chamber above each gate valve can provide an evacuated housing for the test mass if it is necessary to withdraw it from the beam line, and can be used as an airlock to allow individual test masses to be removed or inserted without disturbing operation of other interferometers in the system. The beamsplitter tanks can be closed off from the main system by gate valves in the auxiliary linking pipes. This overall design thus provides a relatively economical way of accomodating several interferometers in a single pair of beam pipes, while making it possible to access the main components of each interferometer for replacement or adjustment without affecting the vacuum in the system or interrupting seriously the operation of the other interferometers.

In such a system it is important to avoid coupling of motion or vibration noise from one interferometer to another. To help ensure this, the main seismic isolation for each test mass is located in the airlock vacuum chamber above it, and is supported from the ground by a structure isolated from the main vacuum system walls by flexible metal bellows. (The latter concept is similar to one introduced in a vacuum system designed at the Rutherford Laboratory.) Additional vacuum tanks to house auxiliary optical components such as beam-conditioning cavities may be attached to the beamsplitter tanks as necessary.

There are many ways of arranging a multiple-interferometer system of this type, and the arrangement shown is just one example of a general concept which was stimulated by the compact nature of Fabry–Perot and similar optical

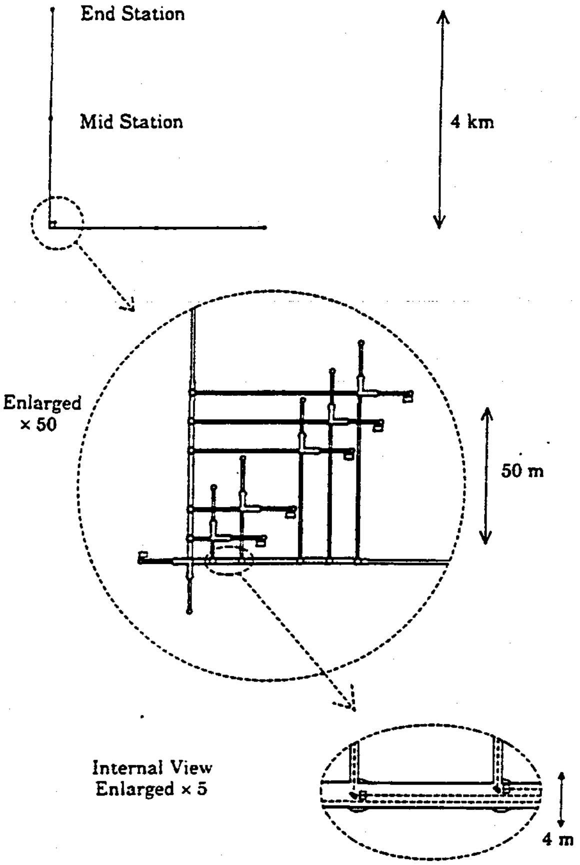


Figure 12.6. Schematic arrangement for a gravitational wave detector system which enables several Fabry–Perot interferometers to operate independently within a single pair of main vacuum pipes. Provision is shown for interferometers which monitor half the maximum arm length as well as the full arm length, to provide a gravitational wave 'signature' from the dependence of relative test mass displacement on length of baseline. The test masses are housed in vacuum chambers, with access by air-locks, along the line of each arm; while beamsplitters and other optical components are housed in separate chambers near the corner of the L, located along the bisector of the angle between the arms. The diagram is intended to indicate concepts only: layout and dimensions suggested are illustrative only of some possibilities.

configurations. The basic concept can clearly be extended to other types of optical system when necessary.

12.7 Some practical issues

The diagrams of interferometer systems shown here are greatly simplified, and many practical problems have to be dealt with to achieve high sensitivity. There is not space here to discuss these in detail, but some important aspects may be briefly mentioned.

12.7.1 Mode cleaners and fibre filters

Practical interferometers are usually designed to be insensitive to first order to fluctuations in frequency, intensity, polarization, and geometrical parameters of the input light beam. At the high sensitivity required for gravitational wave detection this is essential, and here second or higher order coupling mechanisms which are unimportant in other fields can be significant. For example, fluctuations in beam direction or position can couple energy in varying degree into modes of the main optical cavities other than the desired lowest order symmetrical mode. As these modes will in general have different resonance frequencies, they will cause varying phase shifts in the light. If there is any difference in the alignment, or in figure errors in the shape of the mirrors, in the two arms, the coupling to undesired modes may not be precisely equal, and a differential phase shift may result. If this phase shift is detected, either directly or by interference with the main mode through some other imperfection, such as photodiode non-uniformity, noise can be produced on the output. Early experiments by the Max Planck group showed that effects similar to these could seriously limit the sensitivity of a Michelson interferometer, and they demonstrated the use of an auxiliary Fabry–Perot cavity between the laser and the interferometer as a filter for geometrical fluctuations in the laser beam (Rudiger *et al.*, 1981). A filter cavity like this, used as a 'mode cleaner', is also an effective way of reducing input beam fluctuations for a Fabry–Perot interferometer.

Several other methods of reducing geometrical fluctuations in an input beam have been used or proposed. A single-mode optical fibre, used first with a Michelson interferometer by the MIT group, has been employed with several Fabry–Perot detectors. This technique has the advantage of not requiring any frequency locking to the input light; although with small-diameter fibres there is a power limitation, and a possibility of non-linear phenomena at the high power density required in the fibre. Use of one or more spatial filters for reducing beam fluctuations has been analysed by B. J. Meers, but is in general less effective for a given power loss. Most current gravity-wave interferometers employ optical fibres or mode cleaning cavities to condition the input beam.

A relatively large mode cleaner has been proposed as an intrinsic part of the optical system for a Fabry–Perot interferometer for operation with kilometre-scale arms, in a design developed two years ago by the writer. In this case the mode cleaning cavity has some additional functions and special features. The cavity is designed as the main frequency reference for the input light in the gravity-wave frequency region; it defines the input beam direction at these frequencies; and it also acts as a high-frequency amplitude, phase, and frequency filter. To make this cavity quiet and relatively free from thermal noise it is about 12 m long, and is formed between mirrors suspended by wires from seismically isolated points, in a way similar to that for the test masses of the main interferometer. The D.C. stability of the light wavelength is determined by a separate smaller rigid cavity, through a feedback system to the mode cleaning cavity which is effective only at frequencies much lower than the gravity-wave frequency. In this interferometer design a second similar cavity is also proposed as a filter between the interferometer output and the main photodetector, to help reduce effects of scattered light and other light in undesired modes. To allow passage of the modulation sidebands, the free spectral range of the cavity is arranged to match the modulation frequency for the main interferometer.

Some further points about mode cleaning cavities may be mentioned here. In a system operating with high laser power it may be necessary to limit the finesse of a mode cleaner to minimize heating effects in the cavity mirrors, and it may then be appropriate to use more than one mode cleaner in series to achieve sufficient filtering. In this context it is worth noting that useful mode cleaning action may be provided in a recycling interferometer by the effective cavity formed by the whole system.

It may be remarked that a mode cleaning cavity system may provide a more general benefit than the control of specific geometrical fluctuations. At the levels of sensitivity required for gravitational wave detection unforeseen phenomena of various kinds may disturb operation. If, however, everything involved in the sensitivity region of the system is stationary for times of order of the period of the gravity wave the chance of noise appearing near this periodicity is reduced. A mode cleaner can help ensure that all the parameters describing the light as it enters the interferometer are stationary, and thus may reduce risks of disturbance by even unpredicted optical phenomena.

12.7.2 Beam heating effects in mirrors and other components and techniques for reducing it

In laboratory-scale optical cavities typical beam diameters at the mirrors are of order a millimetre or less, and it is possible to get high flux densities in the beam spot with a fairly low power laser. Losses in the mirror can then produce large temperature gradients, leading to distortion of the mirror and changes in refractive index in the substrate. Effects of this kind have been observed in many laboratories working with optical cavities. In the interferometer arms of kilometre-scale gravitational wave detectors, the spot diameters will be larger and flux densities lower for the same input power. However in spite of the relatively small temperature gradients these can be significant over a greater distance in a thick substrate, and it was noted by W. Winkler (private communication, 1988) that first order estimates of the optical effects suggest that they may be largely independent of beam diameter for a given total beam power.

Initial observations of heating effects have been made in several laboratories, and at stored power levels of order 1 kW of green light some mirrors have been free from significant heating effects, while other mirrors, with different coating losses, have shown significant distortion of the cavity modes. It is not yet clear whether with the most suitable mirrors absorption losses are large enough to cause heating effects to set limits to power levels in practical gravity-wave interferometers.

The magnitude of the effects on an interferometer of the heat produced by a given amount of mirror absorption loss does depend on the optical configuration used, and it is possible to design optical cavity interferometers which are less vulnerable to heating effects than those illustrated earlier here. In a typical cavity mirror there are two main effects: heat produced in the coating causes a change in refractive index of the substrate, giving distortion of the input and output beams, and temperature rise in the substrate causes thermal expansion and distortion of the mirror shape. The situation may be ameliorated by arranging the cavity so that light does not have to be transmitted by the main mirror substrates. Then thermal lensing effects in the substrate are avoided, and, further, the substrate material may be chosen for high thermal conductivity and low expansion rather than for good optical properties, so the thermal effects are further reduced. One way of realizing a cavity of this type is to introduce a thin wedged plate of low-loss material, such as fused silica, into the cavity, and use this as a coupling device. One surface of the plate may be arranged to be at Brewster's angle to the beam and thus free of reflection, and the other surface oriented to give by reflection the required input and output coupling to the cavity. No coatings are required on the plate, so thermal effects in it arise only through its intrinsic absorption, which may be significantly less than in typical mirror coatings.

If cavity heating effects are sufficiently reduced by these or other means, then heating in the beamsplitter may become a limiting factor. This may be reduced by avoiding mirror coatings in the beamsplitter, and achieving the beamsplitting action by evanescent wave coupling between two prisms. If heating in the main cavities is still a limiting factor, each Fabry–Perot cavity may be folded using a delay-line mirror configuration so that the heat is distributed over several reflection spots. Such an arrangement would use up more space in the main vacuum pipes than an unfolded system. In some respects systems of this type exhibit some of the characteristics of delay line Michelson interferometers when these are used with dual recycling to reduce the number of reflection spots and economize in mirror size. With equal mirror areas the two systems will be limited at about the same power level, with a possible slight advantage to the folded Fabry–Perot due to the lower flux through the beamsplitter in that configuration. Such systems would represent some compromise between economy of beam-pipe space and power handling, and may be rendered unnecessary by improvement in mirror and substrate performance.

12.8 Conclusion

This overview of Fabry–Perot cavity gravity-wave detectors is far from complete, and has had to omit discussion of many interesting and important aspects. The aim has been to give an understanding of the basic principles and of development of some of the ideas. The technique itself is far from fully developed, and large improvements in sensitivity are to be expected. Currently, the fields of gravity-wave detection, and of the interferometers themselves, are in a state of transition. Detailed plans and designs for scaling up from laboratory instruments to kilometre-baseline detectors are being prepared, while at the same time it is becoming apparent that there is still much room for originality and variety in optical configurations. The problems of handling high light flux in the interferometer arms are likely to lead to new designs, as already indicated. Earlier systems are likely to be superseded by configurations which may combine features of both Fabry–Perot and delay line interferometers. Currently, some recycled Fabry–Perot configurations which may be folded if required seem promising and flexible, but the choices depend on properties of mirror materials, and are likely to change. In any case, the research already carried out on Fabry–Perot and other systems is likely to significantly influence the instruments which will eventually be of practical importance for investigating gravitational radiation.

Acknowledgements

Much of the work described here has been supported by the California Institute of Technology, and by the National Science Foundation in a series of grants (including in large part grant PHY-8504136, superseded by PHV-8803557). This support, and the stimulation and collaboration of many colleagues, is gratefully acknowledged.

References

Caves, C. M. (1981). *Phys. Rev. D* **23**, 1693.

Drever, R. W. P., Hough, J., Bland, R. and Lessnoff, G. W. (1973). *Nature* **246**, 340.

Drever, R. W. P. (1983). In *Gravitational Radiation*, NATO Advanced Physics Institute, Les Houches, June 1982, eds. N. Deruelle and T. Piran, p. 321, North Holland Publishing, Amsterdam.

Drever, R. W. P., Ford, G. M., Hough, J., Kerr, I. M., Munley, A. J., Pugh, J. R., Robertson, N. A. and Ward, H. (1983a). *Proceedings of the Ninth International Conference on General Relativity and Gravitation (Jena 1980)*, ed. E. Schmutzer, p. 265, VEB Deutscher Verlag der Wissenschaften, Berlin.

Drever, R. W. P., Hall, J. L., Kowalski, F. V., Hough, J., Ford, G. M., Munley, A. J. and Ward, H. (1983b). *Appl. Phys.* **B31,** 97.

Drever, R. W. P., Hough, J., Munley, A. J., Lee, S.-A., Spero, R., Whitcomb, S. E., Pugh, J., Newton, G., Meers, B. J., Brooks III, E. and Gursel, Y. (1983c). In *Quantum Optics, Experimental Gravity, and Measurement Theory*, eds. S. Meystere and M. O. Scully, p. 503, Plenum Publishing, New York.

Meers, B. J. (1988). *Phys. Rev. D* **38,** 2317.

Rudiger, A., Schilling, R., Schnupp, L., Winkler, W., Billing, H. and Maischberger, K. (1981). *Optica Acta* **28,** 641.

Weiss, R. (1972). *Quartr. Progr. Rep. Res. Lab. Electr. MIT*, **105,** 54.

PHYSICAL REVIEW D
PARTICLES AND FIELDS

THIRD SERIES, VOLUME 38, NUMBER 8 — 15 OCTOBER 1988

Recycling in laser-interferometric gravitational-wave detectors

Brian J. Meers
Department of Physics and Astronomy, University of Glasgow, Glasgow G12 8QQ, Scotland

(Received 31 March 1988)

Laser interferometers may detect gravitational waves by sensing the strain in space produced by their passage. The resultant change in intensity of an interference fringe must be observable against a background noise due to the statistical fluctuations in the number of detected photons. Optimization of the detector sensitivity thus involves devising an optical system which both maximizes the signal and minimizes the noise. This is attempted in the various arrangements known collectively as light recycling. Here, the performance of these systems is quantitatively assessed. Standard or broadband recycling functions essentially by making efficient use of the available light, but it is shown that it may also be made to further enhance the sensitivity within a narrow bandwidth, becoming tuned recycling. This works, as do all the narrow-band variants, by arranging for both the laser light and a gravitational-wave-induced sideband to be resonant in the optical system. The original narrow-band system, resonant recycling, can also be made broadband; the various sensitivity-bandwidth combinations, together with the tuning properties of such a system, are discussed. Furthermore, a new optical arrangement, dual recycling, is proposed. Its optical layout is an extension of standard recycling and its strength lies in its flexibility. It is shown that, relatively simply, it may be made into either a broadband or a narrow-band system, in each case with the same performance as the best of the other schemes. It may be tuned more efficiently and easily over a wide range of frequencies. Uniquely, optimum performance may be obtained with dual recycling without the requirement that the storage time of the optical elements in each arm of the interferometer be comparable with the period of the gravitational wave. This may allow the operation of delay line interferometers down to much lower gravitational-wave frequency and will provide great operational flexibility. Finally, it is shown that dual recycling, together with resonant recycling, is relatively insensitive to imperfections in the geometrical quality of the optical system. When implemented on interferometers with lengths greater than about a kilometer, recycling should allow the attainment of the sensitivity required in order to observe gravitational waves and open up a new window to the Universe.

INTRODUCTION

Gravitational-wave detectors based upon the use of laser interferometry to monitor the separation of widely spaced free masses are under development around the world. The kilometer scale interferometers currently being proposed should open up a new field of astronomy.[1,2] An integral feature of these detectors will be the use of variants of the optical technique known as recycling, an idea first proposed by Drever.[3] Here we quantitatively analyze these variants in a unified way, showing how they improve the sensitivity level of the detector set by photon-counting fluctuations, and propose a new variant, which may have considerable practical and operational advantages.

The basic arrangement of a laser interferometric gravitational-wave detector is shown in Fig. 1 (without mirror M_0). Principles of operation are reviewed in Refs. 2 and 3.

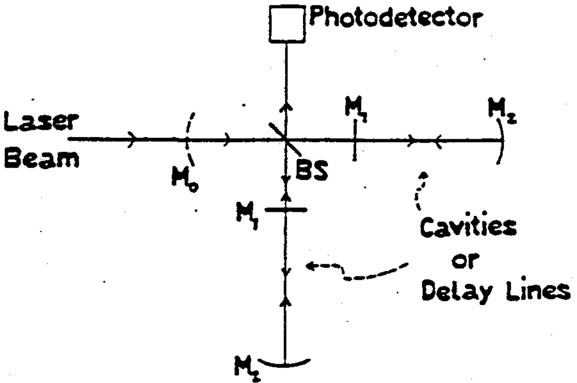


FIG. 1. A laser-interferometric gravitational-wave detector is essentially a Michelson interferometer. A gravitational wave produces opposite phase shifts on the light in the two arms of the interferometer; when interference occurs at the beamsplitter, the result is an intensity change on the light traveling to the photodetector.

©1988 The American Physical Society

A gravitational wave (amplitude h) of optimum polarization induces opposite length changes δl in each arm of the Michelson interferometer of magnitude

$$\delta l/l=\tfrac{1}{2}h\cos\omega_g t\ , \tag{1}$$

where l is the length of the interferometer arm and ω_g is the angular gravity-wave frequency. The resultant phase change $\delta\phi$ on a single beam of light spending a time τ_s in the interferometer is

$$\delta\phi=\int_{-\tau_s}^{0}\frac{4\pi}{\lambda}\delta l\,dt=\frac{h\omega}{\omega_g}\sin\frac{\omega_g\tau_s}{2}e^{-i\omega_g\tau_s/2}\ , \tag{2}$$

where λ and ω are the wavelength and angular frequency of the light, respectively. So the signal is a maximum if the storage time τ_s is equal to half the gravitational-wave period—the expected result since the gravitational wave reverses its sign with period $\frac{1}{2}\nu_g$.

In practice, multipass optical delay lines or cavities are used to match the storage time to the gravitational-wave period.[1] The choice of Fabry-Perot cavities makes the analysis more complex, for there are contributions to the signal from light with storage times of all multiples of $2l/c$. However, standard techniques for the analysis of Fabry-Perot interferometers[4] may be used as long as the signal, while arbitrarily fast, is assumed to produce a small phase change on the light and any large offset from resonance is assumed to be slow compared with the cavity storage time.[5] Thus, the field E_R reflected off a cavity with incident field E_0, input mirror of (amplitude) transmission T_1, reflectance R_1, far mirror reflectance R_2, and free spectral range $\nu_0=c/2l$ is

$$E_R/E_0=R_1-\frac{T_1^2R_2}{(1-R_1R_2)^2}\left[\frac{e^{i\delta}-R_1R_2}{1+F'\sin^2\delta/2}\left[1-\frac{ih\omega}{2\omega_g}\sin\omega_g t\right]+\frac{h\omega e^{i\omega_g t}}{4\omega_g}\frac{e^{i(\delta-\omega_g/\nu_0)}-R_1R_2}{1+F'\sin^2\left[\dfrac{\delta-\omega_g/\nu_0}{2}\right]}\right.$$
$$\left.-\frac{h\omega e^{-i\omega_g t}}{4\omega_g}\frac{e^{i(\delta+\omega_g/\nu_0)}-R_1R_2}{1+F'\sin^2\left[\dfrac{\delta+\omega_g/\nu_0}{2}\right]}\right]\ , \tag{3}$$

where $F'=4F^2/\pi^2=4R_1R_2/(1-R_1R_2)^2$; F is known as the finesse, δ is the phase offset (from resonance) for a single traverse of the interferometer. The different terms can be viewed as embodying the way in which the unmodulated laser light and the sidebands produced by the phase modulation of the gravitational wave resonate in the cavity. The static terms describe the effective reflectivity of the cavity, while the fluctuating terms give the signal. While the general form of (3) is useful (and shall be encountered again), simple interferometers operate with the cavities on resonance, at $\delta=0$. The phase change of the light is then determined by the size of the fluctuating, imaginary part of the field compared with the static, real part (i.e., relative sideband amplitude). This may be determined by straightforward algebraic manipulation of (3). If the length of the interferometer is small compared with a gravitational wavelength (so that $\omega_g/\nu_0 \ll 1$), the phase shift of the emerging light is

$$\delta\phi=\frac{\alpha_c}{2}\frac{h\omega}{\omega_g}\frac{\omega_g\tau_s}{[1+(\omega_g\tau_s)^2]^{1/2}}\ , \tag{4}$$

where the storage time is

$$\tau_s=\frac{2Fl}{\pi c}=\frac{2l(R_1R_2)^{1/2}}{c(1-R_1R_2)}\ , \tag{5}$$

where c is the speed of light and α_c, the factor by which the field emerging from the cavity is larger than the incident field, is

$$\alpha_c=\frac{T_1^2R_2}{1-R_1R_2}\sim\frac{2}{1+A^2/T_1^2}\ , \tag{6}$$

where A^2 is the loss coefficient of each mirror $(T^2+R^2+A^2=1)$. Note that the field inside the cavity is enhanced by α_c/T_1. From now on it shall be assumed that both R_1 and R_2 are close to 1.

So for low loss cavities, as envisaged, the phase change (4) on the light is the same for cavity with $\omega_g\tau_s>1$ as for a delay line $\tau_s\approx\frac{1}{2}\nu_g$. They will thus have equal potential sensitivities: if the smallest detectable intensity change (produced by the interference of two beams with phase fluctuation $\delta\phi$) is set by photon-counting statistics, then the smallest detectable gravitational-wave amplitude is

$$h_{\rm DL}=\left[\frac{\pi\hbar\lambda\Delta\nu_g}{\epsilon I_0 c}\right]^{1/2}\frac{\nu_g}{\sin(\omega_g\tau_s/2)}\ , \tag{7}$$

$$h_{\rm cav}=\left[\frac{\hbar\lambda[1+(\omega_g\tau_s)^2]\Delta\nu_g}{4\pi\epsilon I_0 c\tau_s^2}\right]^{1/2}\ , \tag{8}$$

where $\Delta\nu_g$ is the measurement bandwidth, ϵ, I_0 is the effective laser power, and $\hbar$ is the reduced Planck's constant.

BROADBAND OR STANDARD RECYCLING

It is important to realize that only a small fraction of the incident optical power is absorbed in the simple detectors considered so far.

A delay line with mirrors of amplitude reflectance R $(A^2 \ll 1)$ and N reflections has an overall intensity reflectance of

$$R_{\mathrm{DL}}^2=1-R^{2N}\sim 1-\frac{c}{l}\tau_s A^2 \tag{9}$$

while that of a cavity is [cf. (3) + algebra]

$$R_c^2=1-\frac{4FA^2/\pi}{1+F'\sin^2\delta/2} . \tag{10}$$

Thus, with $l=1$ km, $A^2=10^{-4}$, and $\tau_s=\frac{1}{2}\tau_g=0.5$ ms, the losses only total about 1.5%. With the interferometer working on a dark fringe the remaining light travels back toward the laser and, in a simple interferometer, is wasted. The simplest version of recycling (sometimes called "broadband" or "standard" recycling) consists of placing another mirror M_0 in the beam (see Fig. 1) with the correct position to coherently send light back to the interferometer.[3] The recycling mirror may also be regarded as an impedance matching device which ensures efficient transfer of power.[6] The resonant enhancement of the laser intensity in this recycling cavity reduces the significance of photon-counting errors: the power increases by the effective number of times the light is recycled, the shot-noise-limited sensitivity is enhanced by the square root of this factor. The increase in power inside the recycling system is limited by the losses in the system, principally the absorption and scattering of the cavity mirrors but with possible contributions from waveform distortions (limiting the efficiency with which the light from the two arms interferes to travel back toward M_0). Specifically, the maximum power gain P is

$$P=\frac{1}{1-R_{\mathrm{eff}}^2} , \tag{11}$$

where $1-R_{\mathrm{eff}}^2$ is the total loss on one round trip from the recycling mirror. If the losses are dominated by the cavity (or delay lines), as will be assumed from now on, then R_{eff}^2 is given by (10) or (9). The choice of storage time is a trade-off between signal and losses; for both delay line and cavity, the optimum choice is a storage time just short of giving the maximum phase shift: $\omega_g\tau_s(\mathrm{opt})=1$ for a cavity, $\tau_s=0.37\tau_g$ for a delay line. The gain in shot-noise-limited sensitivity S compared with a low loss nonrecycled system is then

$$S_c=\left[\frac{\pi\nu_g}{4\nu_0 A^2}\right]^{-1/2} = 10\left[\frac{A^2}{5\times10^{-5}}\right]^{1/2}\left[\frac{\nu_g}{1\ \mathrm{kHz}}\right]^{1/2}\left[\frac{l}{1\ \mathrm{km}}\right]^{1/2} \tag{12}$$

and

$$S_{\mathrm{DL}}=\left[\frac{1.14\nu_g}{\nu_0 A^2}\right]^{1/2} . \tag{13}$$

While this improvement in sensitivity is optimized for one gravitational-wave frequency, an enhancement is obtained at all frequencies—standard recycling produces a broadband detector. In the case of a cavity, the optimized shot-noise-limited sensitivity will be, combining (12) and (8),

$$h=\left[\frac{2\lambda\hbar A^2\nu_g\Delta\nu_g}{\epsilon I_0 l}\right]^{1/2}=7\times10^{-24}\left[\frac{\nu_g}{1\ \mathrm{kHz}}\right]^{1/2}\left[\frac{\epsilon I_0}{100\ \mathrm{W}}\right]^{-1/2}\left[\frac{A^2}{5\times10^{-5}}\right]^{1/2}\left[\frac{l}{1\ \mathrm{km}}\right]^{-1/2}\Big/\sqrt{\mathrm{Hz}} \tag{14}$$

assuming $\lambda=5\times10^{-7}$ m.

RESONANT RECYCLING

While a broadband detector is desirable when looking for unexpected sources or short bursts of gravitational radiation (GR), there are cases where a narrow-band detector is sufficient and even desirable. Examples include monochromatic sources of GR such as that from pulsars and accreting neutron stars,[1] while the observation of a stochastic background of GR and of the signals from coalescing compact binaries should benefit from an enhanced sensitivity within a restricted bandwidth. A possible way of constructing such a detector was proposed by Drever.[3] This method, known as resonant recycling, is illustrated in Fig. 2(a) (for a delay line) and Fig. 2(b) (for a cavity).

It can be seen that this is a very different optical arrangement from that of standard recycling. In the delay line case, the storage time of the light in each arm of the interferometer is arranged to be half a gravity-wave period, so that the light picks up the maximum phase shift from the gravitational wave; the light then passes directly into the other arm of the interferometer where, because the gravitational wave reverses its sign every half period it sees the *same* sign of phase shift as it did before, with the result that the signal builds up coherently. Roughly, the signal is increased by the number of times the light is cycled round the whole optical system, which is limited by the losses. If the losses are dominated by the cavity mirrors, then use of (9) and an appropriate version of (3) quickly leads to the gain in shot-noise sensitivity S_{RR} compared to a nonrecycled system:

$$S_{RR}(\mathrm{DL})=\frac{\nu_g}{\nu_0 A^2} . \tag{15}$$

Note that, because signal rather than intensity is recycled, this gain in sensitivity is approximately the square of that obtained by using standard recycling, but it is restricted to a narrow bandwidth $\Delta\nu_g$, since other frequencies become out of step with the gravitational wave:

$$\Delta\nu_g\sim\frac{\nu_g}{S_{RR}}=\nu_0A^2=8\left[\frac{A^2}{5\times10^{-5}}\right]\left[\frac{l}{1\ \mathrm{km}}\right]^{-1}\ \mathrm{Hz} . \tag{16}$$

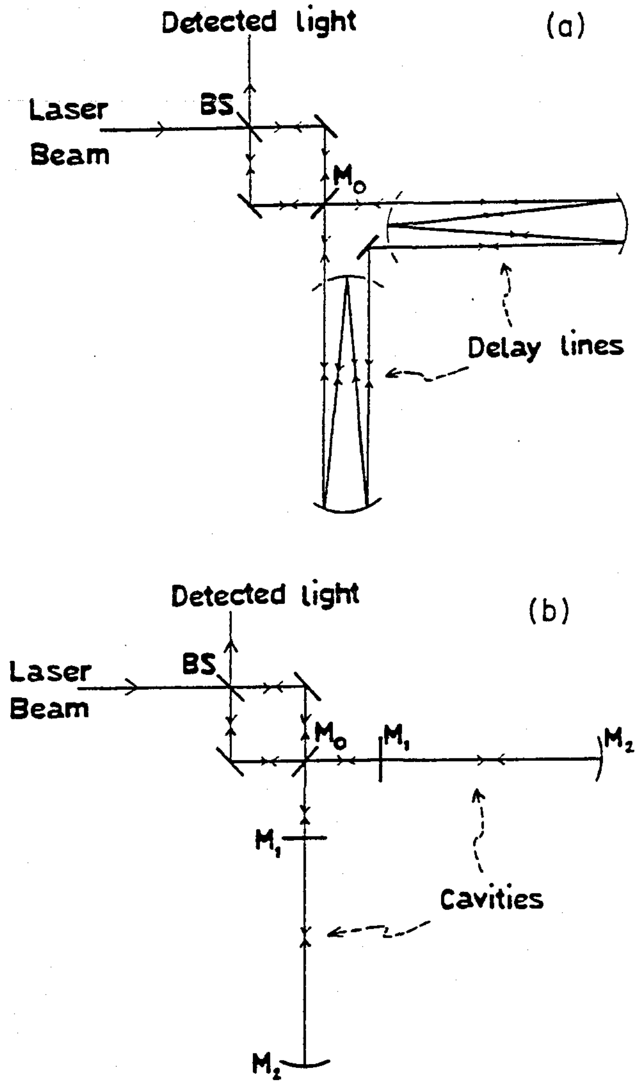


FIG. 2. A schematic diagram of the optical arrangement for resonant recycling (a) for a delay line, (b) for a cavity.

When optical cavities are used for resonant recycling, it is instructive to view the detector as a system of coupled cavities which have two normal modes;[6,7] the laser resonates with one of these, while the action of the gravitational wave is to pump energy into the other mode. In order to understand the reason for this, it is helpful to realize that light reflected off a cavity experiences a frequency-dependent phase shift θ; for the laser frequency (carrier), some manipulation of (3) gives

$$\tan\theta=\frac{(\alpha_c F/\pi)\sin\delta}{F'\sin^2\delta/2-1}\sim\frac{\alpha_c F\delta/\pi}{F^2\delta^2/\pi^2-1}\quad\text{for }\delta\ll 1\,. \tag{17}$$

Remember that δ is a measure of how far off the light is from being perfectly resonant in the cavity. For the sidebands, θ may be found by making the substitution $\delta\rightarrow\delta\pm\omega_g/\nu_0$. We see that between being on resonance ($\delta=0$) and far off ($F\delta/\pi>>1$) there is a change in the phase of $\sim 180°$ with the phase shift being 90° at $F\delta/\pi=1$. Now when a gravitational wave changes the effective length of a cavity, it imposes two sidebands on the (carrier) light emerging from the cavity, which then travels (via the recycling mirror) to the other cavity. Here it is reflected; if the length of the center cavity is adjusted so that the reflected carrier light has the same phase as that emerging from the other arm (i.e., the laser light is resonant), then, since the signal emerging from the two arms has opposite sign and the corresponding sidebands therefore have opposite phase, at least one of the sidebands must experience a phase shift on reflection which is 180° different from that of the carrier if the signal is to be increased. If the sideband and carrier also experience a relative phase shift of 180° when they are reflected off the original cavity, both carrier and sideband will be resonant, with the sideband always having the correct phase for its amplitude to be increased by the gravitational wave. This resonance condition of an 180° relative phase shift may be produced in practice in several different ways: the most symmetrical is to operate each cavity so that it gives a 90° phase shift of opposite sign to both carrier and sideband. Reference to (17) shows that this requires $F\delta/\pi=1$ or $\delta=\omega_g/2\nu_0$, $\omega_g\tau_s=2$. However, it can be seen from (17) that resonance may also be obtained for $\delta=0$ as long as $\omega_g\tau_s>>1$. It must be stressed that even if the isolated cavity is not on resonance, the coupled cavity system is.

In order to make this analysis quantitative, it is necessary to calculate the phase change produced by a gravitational wave on the light emerging from a cavity which (by itself) is not perfectly resonant. Reference to (3) plus, once more, some straightforward algebra, gives an amplitude E_s for sideband emerging from such a cavity of

$$|E_s/E_0|=\left[\frac{\alpha_c}{2}\right]\frac{h\omega}{2\omega_g}\frac{\omega_g\tau_s}{\left[1+F'\sin^2\delta/2\right]^{1/2}\left[1+F'\sin^2\dfrac{\delta-\omega_g/\nu_0}{2}\right]^{1/2}}\,. \tag{18}$$

This is the contribution, due to a single sideband, to the phase change on the reflected light from an isolated cavity. If this sideband is resonant, it will emerge from the recycling optics and be detectable with amplitude

$$|E_s/E_0|_{\text{tot}}=\frac{h\omega}{\omega_g}S_{RR}=\frac{\frac{1}{2}\alpha_c T_0^2}{(1-R_0R_c^2)(1-R_0R_{cs}^2)}\frac{h\omega}{\omega_g}\frac{\omega_g\tau_s}{(1+F'\sin\delta/2)^{1/2}\left[1+F'\sin^2\dfrac{\delta-\omega_g\nu_0}{2}\right]^{1/2}}\,, \tag{19}$$

where R_c and R_{cs} are the reflectivity of the cavity for the carrier and sideband, respectively [cf. (10)]. If we want to maximize the sensitivity at one frequency, then the best choice of recycling mirror transmission is

$$T_0^2=\frac{\frac{8F}{\pi}A^2}{(1+F'\sin^2\delta/2)^{1/2}\left[1+F'\sin^2\frac{\delta-\omega_g/\nu_0}{2}\right]^{1/2}} \qquad (20)$$

which gives for the gain S_{RR} in shot-noise-limited sensitivity over a nonrecycled system,

$$S_{RR}(\mathrm{cav})=\frac{\pi\nu_g}{4\nu_0A^2}=100\left[\frac{\nu_g}{1\ \mathrm{kHz}}\right]\left[\frac{l}{1\ \mathrm{km}}\right]\left[\frac{A^2}{5\times10^{-5}}\right]. \qquad (21)$$

Note that the gain from resonant recycling is, within its bandwidth, just the square of that obtainable with broadband recycling.

So the shot-noise-limited sensitivity is

$$h_{RR}(\mathrm{cav})=\left[\frac{2\lambda\hbar c}{\pi\epsilon I_0\tau_{\mathrm{int}}}\right]^{1/2}\frac{A^2}{l}=5\times10^{-28}\left[\frac{\epsilon I_0}{100\ \mathrm{W}}\right]^{-1/2}\left[\frac{A^2}{5\times10^{-5}}\right]\left[\frac{l}{1\ \mathrm{km}}\right]^{-1}\left[\frac{\tau_{\mathrm{int}}}{10^6\mathrm{s}}\right]^{-1/2}, \qquad (22)$$

where τ_{int} is the integration time.

Note that delay line and cavity detectors give virtually the same performance even though the latter only use one sideband: this is because cavity detectors can have a higher Q, resulting from the lower loss of a cavity "off resonance," and therefore the signal sideband has a higher coupling to the gravitational wave. One consequence is that the bandwidth of a cavity detector (at maximum sensitivity) is somewhat smaller than for a delay line detector: the different reflection phase shifts $\Delta\phi$ for different frequencies and the requirement that $S\Delta\theta<1$ gives

$$\Delta\nu_g(\mathrm{cav})=\frac{2\nu_0}{\pi}A^2\ , \qquad (23)$$

a factor of $\frac{1}{2}\pi$ smaller bandwidth than that of a corresponding delay line system.

It is possible to tune the resonant frequency of a cavity detector by altering the length of the long cavities (i.e., δ), while adjusting the length of the center cavity to keep the intensity there a maximum. With cavities of storage time τ_s one can obtain resonance at frequency ω_g with an offset δ of

$$\delta=\frac{\omega_g}{2\nu_0}\left\{1\pm\left[1-\left[\frac{2}{\omega_g\tau_s}\right]^2\right]^{1/2}\right\}. \qquad (24)$$

With a given storage time, the lowest resonant frequency is obtained at $\omega_g\tau_s=2$, $\delta=\frac{1}{2}\omega_g/\nu_0$. Different offsets then tune the optical system to a higher resonant frequency. With the same optics, this will give a sensitivity gain S for these higher frequencies which is the same as that at $\omega_g\tau_s=2$ [to see this, substitute (24) into (19)]. In other words, a resonant recycling detector which is tuned to a frequency a factor of 2 higher than that for which it was optimized will have a shot-noise-limited sensitivity a factor of 2 worse than if it had been optimized for that frequency.

The bandwidth of the detector increases as $\delta\to0$: if the lowest possible resonant frequency with a particular optical system is ν_{g0}, so that $\omega_{g0}\tau_s=2$, then the bandwidth when the detector is tuned is

$$\Delta\nu_g=\frac{2\nu_0A^2}{\pi}\frac{\nu_g}{\nu_{g0}}\ . \qquad (25)$$

In the limit of $\delta\approx0$, resonance can be obtained as long as $\omega_g\tau_s\gg1$; the sidebands will then have their phase inverted on reflection virtually independent of their frequency. Choice of $\omega_g\tau_s=(\omega_g/2\nu_0A^2)^{1/2}$ and $T_0^2=4/\omega_g\tau_s$ will then give a detector with bandwidth $\Delta\nu_g\approx\nu_g$ and sensitivity gain S just equal to that obtainable from standard recycling (13). Thus, resonant recycling can be made broadband.

TUNED RECYCLING

Just as resonant recycling can be made broadband, it is possible to make standard recycling narrow band. One scheme, known variously as tuned or detuned recycling, was suggested recently by Brillet:[7] it uses the same optical arrangement as standard recycling (Fig. 1 with mirror M_0) but the cavities are adjusted so that it is one of the gravitational-wave-induced sidebands, rather than the laser light, which is on resonance with the isolated cavities. This gives a phase shift on the light of one-half the maximum value (cf. 18 with $\delta=\omega_g/\nu_0$) while the losses for the laser light are reduced, allowing a larger build up of intensity in the center cavity and an improved shot-noise-limited sensitivity. It is also possible to view this arrangement as a coupling of the center and interferometer cavities, leading to a two-mode system. More quantitatively, the maximum power gain P in the center cavity is

$$P=\frac{1}{1-R_{\mathrm{eff}}^2}=\left[A^2\left[\eta+\frac{4F/\pi}{1+(\omega_g\tau_s)^2}\right]\right]^{-1}, \qquad (26)$$

where ηA^2 is the loss associated with one round trip of the center cavity; increasing the intensity in the center cavity enhances the importance of any losses there. If the center cavity losses are negligible, then the maximum sensitivity gain $S=P^{1/2}\,\delta\phi/(h\omega/\omega_g)$ is obtained [cf. (18)] when $T_1^2=A^2$, i.e., the highest storage time possible without losing signal. In this case, combining (26) and (18) gives [remember $(F/\pi)\omega_g/\nu_0=\omega_g\tau_s$]

$$S=\frac{\pi\nu_g}{4\nu_0 A^2} \tag{27}$$

which is exactly the same as for resonant recycling. The bandwidth is the same, also. Similarly, it is possible to choose cavity storage times so that the sensitivity gain is somewhere between the maximum value (27) and the broadband value (12), with a bandwidth such that the gain-bandwidth product is constant.

A problem with tuned recycling is that it will be hard to ensure that the losses in the center cavity are negligible; it will only be possible to obtain the maximum sensitivity gain at low frequency, where the cavity losses are more important:

$$\begin{aligned}\nu_g &\ll \frac{\nu_0}{\pi}\left[\frac{A^2}{\eta}\right]^{1/2}\\ &=100\left[\frac{l}{1\text{ km}}\right]^{-1}\left[\frac{A^2}{5\times10^{-5}}\right]^{1/2}\left[\frac{\eta}{10}\right]^{-1/2}\text{ Hz}\,.\end{aligned} \tag{28}$$

Broader band operation will be possible at higher frequencies and will probably be the most useful way of running tuned recycling.

A NEW TECHNIQUE: DUAL RECYCLING

The amplitude-phase diagram of the light emerging from a multipass delay line Michelson interferometer is shown in Fig. 3. It can be seen that when the storage time of the delay line is comparable to the gravitational-wave period, the phase changes (or sidebands) induced on the light no longer have the correct relative phase to add most efficiently. Resonant recycling is a method of making the resultant $\delta\phi$ add coherently, but there is another approach—that of attempting to add the phase changes induced on each pass so that they always have the correct relative phase to increase the signal. A proposed method of doing this is shown in Fig. 4. In its simplest form, this consists of a simple Michelson interferometer with both the standard recycling mirror M_0 and a new recycling mirror M_3 in the output port of the interferometer. When the interferometer is operating on a dark fringe, the laser frequency is directed toward M_0 while the sidebands travel to M_3, where they are recycled; the transmitted sidebands constitute the signal. It might be thought that this arrangement is equivalent to putting mirrors M_0 and M_3 into the arms of the interferometer to form optical cavities which would then enhance the signal by the effective number of bounces in the cavity (at least for $\tau_s \ll \tau_g$). However, the system of Fig. 4 contains an additional degree of freedom: namely, the position of M_3 (relative to the image in the beamsplitter of M_0). This allows the phase of the recycled sideband reflected off M_3 and reentering the interferometer to be adjusted so that it has exactly the correct phase to add coherently with the sideband being produced by the gravitational wave. In this way, the signal is increased by the effective number of bounces (set by the losses) even for a total optical storage time longer than the gravitational-wave period. Since there are two recycling mirrors, one recycling intensity and one signal, this arrangement may be termed "dual recycling".

If the arms of the interferometer have an arbitrary optical length l_{opt} then only one sideband can, in general, resonate: the mode frequency spacing $c/2l_{\text{opt}}$ only equals the sideband spacing $2\nu_g$ when $\tau_s=\frac{1}{2}\tau_g$. For a simple, single pass Michelson, the effective phase change per pass from each arm is, therefore [cf. (2)],

$$\delta\phi_1=\frac{h\omega}{4\omega_g}\omega_g\tau_s\,. \tag{29}$$

The amplitude E_s, of the single sideband emerging

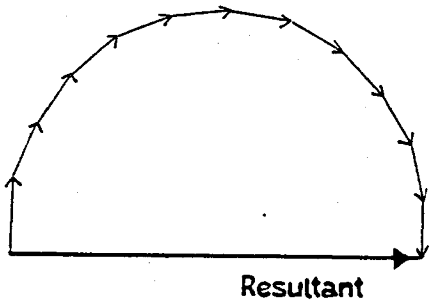


FIG. 3. An amplitude-phase diagram for the light emerging from a delay line Michelson interferometer with $\tau_s=\frac{1}{2}\tau_g$. Each vector corresponds to the phase change induced by a continuous gravitational wave on one traverse of the interferometer. The change in angle of each arrow is produced by a change in phase of the gravitational wave. For $\tau_s>\tau_g$, the curve is a spiral (or a circle if there are no losses).

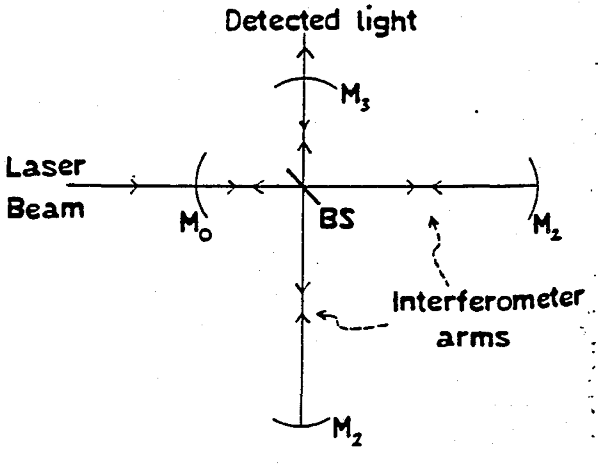


FIG. 4. The proposed new arrangement for recycling ("dual recycling"): the new element is the mirror M_3 which ensures that both signal and intensity are recycled. The optical elements in the arms of the interferometer can be either single or multipass delay lines or cavities. Viewed from M_3, the interferometer must operate on a dark fringe if differential motion of the two arms is to be enhanced.

through mirror M_3 is enhanced both by the resonance of the intensity and of the sideband, being increased by a factor

$$E_s/E_{s0}=\left[\frac{T_0R_D}{1-R_0R_D}\right]\left[\frac{T_3R_D}{1-R_3R_D}\right], \tag{30}$$

where R_D is the amplitude reflectivity of the interferometer arm, including any losses in the beamsplitter. If M_0, M_2, and M_3 have equal losses A^2 and $R_D^2=1-\eta A^2$ ($\eta>1$ allows for losses at the beamsplitter), then (30) may be rewritten as (with $\eta A^2 \ll 1$)

$$E_s/E_{s0}=\left[\frac{T_0}{\frac{1}{2}T_0^2+\frac{1}{2}A^2(\eta+1)}\right]\left[\frac{T_3}{\frac{1}{2}T_3^2+\frac{1}{2}A^2(\eta+1)}\right]. \tag{31}$$

The sideband amplitude is a maximum if the recycling mirrors are chosen so that

$$T_0^2=T_3^2=(\eta+1)A^2 \tag{32}$$

giving

$$E_s/E_{s0}=\frac{1}{(\eta+1)A^2}. \tag{33}$$

The shot-noise-limited sensitivity is determined by the sideband amplitude. The sensitivity gain $S_{\rm DR}$, normalized to the storage time limited case with no recycling (when $\delta\phi=h\omega/\omega_g$) is then

$$S_{\rm DR}=\delta\phi_1\left[\frac{\omega_g}{h\omega}\right]\frac{E_s}{E_{S0}}=\frac{1}{4(\eta+1)A^2}\omega_g\tau_s, \tag{34}$$

i.e.,

$$S_{\rm DR}=\frac{\pi\nu_g}{2(\eta+1)\nu_0A^2}. \tag{35}$$

Thus, if $\eta=1$ (i.e., the mirror losses are dominant), the sensitivity gain is just that obtainable from resonant recycling with a cavity system. This result is not surprising: dual recycling may be regarded as another way of arranging a coupled cavity system so that both the laser light and one sideband are resonant.

At frequencies slightly different from the resonant frequency, any sidebands will not resonate perfectly in the cavity formed by M_3 and the mirrors M_2, being suppressed by a factor $[1+(n_{\rm eff}\Delta\phi)^2]^{1/2}$ relative to the center frequency, where $\Delta\phi=(4\pi l/c)/\Delta\nu_g$ is the phase difference produced on a single traverse of the interferometer between the center frequency and the detuned frequency, the difference between enhanced by the effective number of bounces

$$n_{\rm eff}=\frac{R_3}{1-R_3R_D}. \tag{36}$$

Thus, the bandwidth of the resonant system is determined by $n_{\rm eff}\Delta\phi<1$ or, in the case of the maximum gain case $T_3^2=(\eta+1)A^2$,

$$\Delta\nu_g=\frac{(\eta+1)\nu_0A^2}{\pi}. \tag{37}$$

Comparison with (23) reveals that this is the same as with a cavity resonant recycling system, as long as $\eta=1$.

The bandwidth may be increased by increasing the transmission of the mirror M_3, at the cost of reducing the peak sensitivity gain. The choice of $n_{\rm eff}=\nu_0/\omega_g$ gives a system with bandwidth $\Delta\nu_g\approx\nu_g$ and a sensitivity gain just equal to that obtainable from standard recycling. Thus, greatly differing sensitivity-bandwidth combinations may be obtained by varying the transmission of a single mirror (M_3).

Another significant operational difference between this dual-recycling arrangement and resonant recycling is the possibility of tuning over a large range of gravitational-wave frequencies. In resonant recycling, signal build up requires a 180° relative phase shift between signal and carrier on reflection from a delay line or cavity. For a delay line, this requires $\tau_s=\frac{1}{2}\tau_g$ so a change in tuning requires a change in the storage time, or number of reflections. A cavity system can satisfy the resonance condition over a wide range of frequencies, but there is a minimum frequency ($\omega_g\tau_s=2$) below which the system cannot be tuned without changing the cavity mirrors. In contrast, the new dual recycling arrangement only requires that the Michelson interferometer be operating on the null of a fringe: it will work for any storage time in the arms of the interferometer and can thus be tuned to any gravitational-wave frequency. This is a qualitatively new feature: it is no longer necessary to have a storage time in the delay line (or cavity) which is comparable to the gravitational-wave period in order to obtain good sensitivity. The consequent reduction in the required number of reflections should enable delay line systems to operate well at much lower frequencies. Also, any dual recycling system should have considerably improved operational flexibility.

So far, the discussion of dual recycling has concentrated on the case of single-pass delay lines in the arms of the Michelson interferometer. It will work equally well, however, with either multipass delay lines or optical cavities. Such a choice has the advantage of reducing the significance of any losses at the beamsplitter. For a delay line with $\tau_s \ll \frac{1}{2}\tau_g$ but whose losses are greater than those at the beamsplitter, Eq. (34) holds with $\eta=1$. If $\tau_s=\frac{1}{2}\tau g$, then both sidebands are resonant but the resultant phase change is reduced by $\pi/2$ due to the curvature of the amplitude phase diagram [Fig. 3 (2)], giving a maximum gain in shot-noise-limited sensitivity of

$$S=\frac{\nu_g}{\nu_0A^2} \tag{38}$$

which is the same as that obtainable from resonant recycling [cf. (15)].

With cavities in the arms of the interferometer, the situation is similar. In the absence of recycling, two sidebands emerge from the cavities and leave the interferometer, each with an amplitude given by (18). In dual recycling, mirror M_0 is then adjusted to optimally recycle the intensity and M_3 is adjusted to optimally recycle a sideband of a particular frequency. The positioning of M_3 must take into account the frequency-dependent phase shift (17) of light on reflection off a cavity. It is this which ensures that only a narrow range of frequencies (and only one sideband) are in resonance. The choice of cavity storage time and phase offset is not at all critical: application of (18) to (34) shows that as long as the recycling mirrors are chosen such that

$$T_0^2=\frac{4FA^2/\pi}{1+F'\sin^2\delta/2} \tag{39}$$

and

$$T_3^2=\frac{\dfrac{4FA^2}{\pi}}{1+F'\sin^2\left[\dfrac{\delta-\omega_g/\nu_0}{2}\right]} \tag{40}$$

then the peak gain in shot-noise-limited sensitivity is

$$S=\frac{\pi\nu_g}{4\nu_0A^2} \tag{41}$$

which is the same as for a delay line with $\tau_s \ll \tau_g$ and for a cavity resonant recycling system.

There are many ways in which one can imagine operating a practical dual recycling system. As an example, consider a choice of cavity storage time such that $\omega_g\tau_s=1$ at the frequency of interest. With the cavities operating on resonance ($\delta=0$) and $T_0^2=4FA^2/\pi$ but without mirror M_3, this arrangement is optimized for standard recycling and broadband operation. If mirror M_3 is inserted, the peak sensitivity is increased and the bandwidth narrowed. This is illustrated in Fig. 5, which shows the effect of decreasing the transmission of M_3 while keeping the rest of the optics fixed. It can be seen that, as well as optimizing the sensitivity at one frequency or for a wide range of frequencies, it is possible to considerably enhance the sensitivity within a significant bandwidth. For example, in the case of $l=1$ km, $A^2=10^{-4}$ as shown in Fig. 5, it is possible to arrange (with $T_3^2=0.2$) for the sensitivity to be better than with standard recycling between 150 Hz and 250 Hz, with an improvement by a factor of 2 at 200 Hz. Such an arrangement would be eminently suited to searches for pulses or chirps from coalescing binaries.

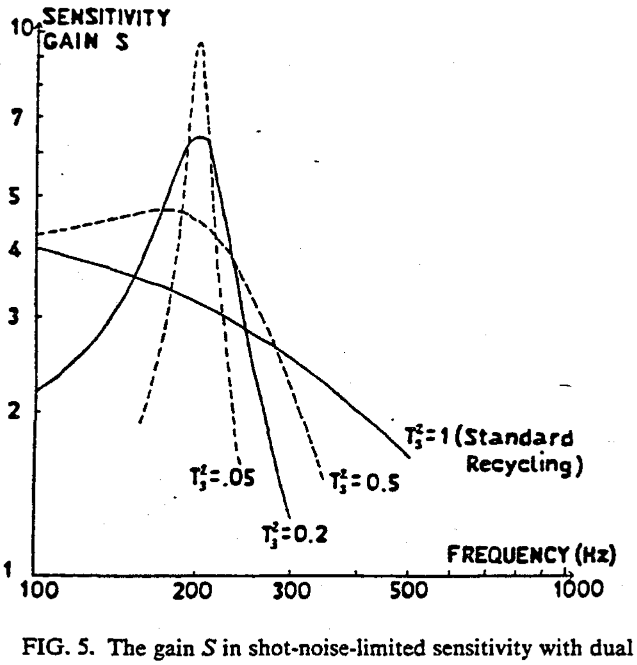


FIG. 5. The gain S in shot-noise-limited sensitivity with dual recycling, compared to a nonrecycled system, as a function of gravitational-wave frequency ν_g. The optical element in the interferometer arms is assumed to be a cavity which is optimized for standard recycling ($\omega_g\tau_s=1$) at 200 Hz, with a length $l=1$ km and mirror loss coefficient $A^2=10^{-4}$. The different curves show the effect of increasing the reflectivity of the dual recycling mirror M_3 from zero, keeping it tuned at 200 Hz: the sensitivity is enhanced within a bandwidth which decreases as the reflectivity increases

The sensitivity in these situations may be calculated by a parallel argument to that applied to the simple Michelson, essentially combining (30) and (18). This gives

$$S=\frac{S_{\max}^{1/2}T_3}{T_3^2+(2F/\pi)A^2}=\frac{S_{\max}^{1/2}T_3}{T_3^2+\frac{1}{4}S_{\max}}\ , \tag{42}$$

where $S_{\max}$ is given by (40). The bandwidth is just

$$\Delta\nu_g/\nu_g=\frac{T_3^2+\frac{1}{4}S_{\max}}{(1-T_3^2)^{1/2}}\ . \tag{43}$$

The pulse sensitivity $S\Delta\nu_g^{1/2}$ is virtually independent of bandwidth, only time resolution is lost as the bandwidth is narrowed.

The different combinations of sensitivity gain and bandwidth are obtained, in dual recycling, by changing the transmission of the mirror M_3. This may be done physically, by changing the mirror. Alternatively, it may be possible to use a variable reflectivity mirror: one way of realizing this is to make M_3 an optical cavity. Such a cavity might have mirrors of low loss but quite high transmission. If the mirrors are of equal transmission, then, when the cavity is on resonance, virtually all of the light is transmitted through it, $T_3^2\approx 1$. When the cavity is far off resonance, $R_3^2\approx 1$. So, tuning the length of this cavity alters the transmission of the "mirror" M_3 and thus the bandwidth of the detector.

It is interesting to calculate the best combination of sensitivity and bandwidth when searching for a stochastic background of gravitational radiation that may be a relic of the big bang or of oscillations of cosmic strings.[1] In such an experiment, the signal from two detectors would be cross correlated to pick out the common background; the resultant signal-to-noise ratio scales as[1,2]

$$S/N \propto S(\Delta\nu_g\tau_{\text{int}})^{1/4}\ , \tag{44}$$

where τ_{int} is the integration time. Thus, when designing a detector, $S\Delta\nu_g^{1/4}$ should be maximized. The optimum combination may be found using (42) and (43):

$$T_3^2(\text{opt})=\frac{1}{2S_{\max}}=\frac{2\nu_0A^2}{\pi\nu_g}\ , \tag{45}$$

$$\Delta\nu_g=\frac{3\nu_0A^2}{-}\ , \tag{46}$$

which gives a cross-correlation sensitivity 1.04 times better than if the detector operated at the peak narrowband sensitivity within its (narrow) bandwidth. The resulting sensitivity to a stochastic background in a bandwidth $\Delta\nu_g \approx \nu_g$ is

$$h = \left[\frac{4\hbar\lambda c \nu_g}{\pi\epsilon I_0}\right]^{1/2} \left[\frac{4}{3c\tau_{\rm int}}\right]^{1/4} \left[\frac{A^2}{l}\right]^{3/4} . \qquad (47)$$

The equivalent detectable energy density in gravitational waves[1] Ω_g, in terms of the critical density for closure of the Universe, is, therefore,

$$\Omega_g = 10^{-10} \left[\frac{\nu_g}{200\ \text{Hz}}\right] \left[\frac{\epsilon I_0}{100\ \text{W}}\right]^{-1} \left[\frac{A^2}{5\times 10^{-5}}\right]^{3/2} \times \left[\frac{l}{1\ \text{km}}\right]^{-3/2} \left[\frac{\tau_{\rm int}}{10^7\ \text{s}}\right]^{1/2} . \qquad (48)$$

Another property of interest is the possible sensitivity when the dual recycling system is tuned to another frequency by moving M_3. For a fixed cavity system, different sideband frequencies have different reflectivities off the cavities [cf. (10) with $\delta = \omega_g/\nu_0$]; if the transmission of M_3 is variable, it can be reoptimized for each frequency and the maximum sensitivity gain is just given by (41), the optimum value. If T_3 is fixed at the value for maximum sensitivity gain at, say, $\omega_{g0}\tau_s = 1$, then the sideband build up will not quite be optimum: the sensitivity gain $S(\nu_g)$ compared to the maximum value for that frequency $S_{\rm max}$ is

$$\frac{S(\nu_g)}{S_{\rm max}} = \left[\frac{8[1+(\nu_g/\nu_{g0})^2]}{[3+(\nu_g/\nu_{g0})^2]^2}\right]^{1/2} . \qquad (49)$$

Thus, at frequencies other than the optimized frequency ν_{g0}, the sensitivity gain is slightly less than optimum: this is illustrated in Fig. 6. It can be seen that the possible sensitivity is hardly reduced at all for frequencies lower than ν_{g0}; this is a result of the small change in cavity reflectivity for $\nu_g < \nu_{g0}$ and the way the sidebands resonate in the cavity. For $\nu_g > 2\nu_{g0}$ the possible sensitivity falls off more quickly, as the cavity reflectivity decreases, but the fall off is modest, much less than with resonant recycling. This results from the liberation in dual recycling from the resonance condition of resonant recycling, with the consequent change in reflectivities for both carrier and sideband as the system is tuned: in dual recycling it is only the sideband which is not recycled optimally when the tuning is altered. Thus, as can be seen from Fig. 6, a dual recycling system with fixed optics is able to tune over a factor of 20 in frequency while maintaining a sensitivity within 10% of the optimal value.

THE EFFECT OF GEOMETRICAL IMPERFECTIONS

Another important feature of any recycling system is its sensitivity to geometrical imperfections in the optics: these might be misalignments of the mirrors, deviations from flatness, or birefringence. In standard recycling, be it broadband or tuned, the effect of such imperfections is to reduce the accuracy of overlap, at the beamsplitter, of the beams from the two arms of the interferometer. This increases the rate at which light leaks out of the interferometer, tending to reduce the sensitivity of the detector both by limiting the build up of intensity in the recycling system and by degrading the contrast of the interference fringe at the output. For example, consider the case of a misalignment by angle θ of the two arms of the interferometer; this will reduce the amplitude E_0 of the fundamental mode seen by one of the beams[5] to

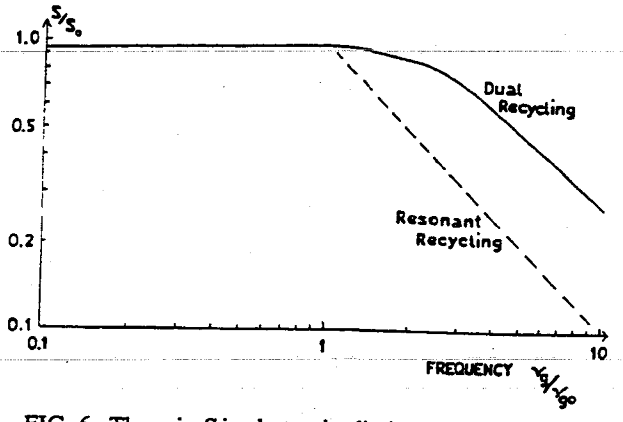


FIG. 6. The gain S in shot-noise-limited sensitivity as a function of its maximum possible value S_0 at a gravitational-wave frequency ν_g, for a fixed optical system optimized for $\nu_g = \nu_{g0}$.

$$E_0 \propto e^{-(\theta/\theta_c)^2} , \qquad (50)$$

where θ_c is the characteristic angle of the beam:

$$\theta_c = \frac{\lambda\sqrt{2}}{\pi w} \approx \left[\frac{\lambda}{l}\right]^{1/2} , \qquad (51)$$

with w the beam radius and λ, again, the wavelength. A fraction $1-\exp[-(\theta/\theta_c)^2]$ of the amplitude will therefore emerge from the output port of the beamsplitter rather than be recycled. If the sensitivity is not to be degraded, this fraction must be considerably less than the losses in the cavities (or delay lines) or equivalently

$$\theta/\theta_c \ll 1/S_{\rm opt} ,$$

where $S_{\rm opt}$ is the best possible gain is sensitivity with broadband recycling. The pointing accuracy of the masses has, therefore, to satisfy

$$\theta \ll \left[\frac{c\lambda A^2}{2\nu_g l^2}\right]^{1/2} \sim 3\times 10^{-6} \left[\frac{\nu_g}{1\ \text{kHz}}\right]^{-1/2} \left[\frac{l}{1\ \text{km}}\right]^{-1} \left[\frac{A^2}{10^{-4}}\right]^{1/2} \times \left[\frac{\lambda}{5\times 10^{-7}\ \text{m}}\right]^{1/2} \text{rad} . \qquad (52)$$

While it is probably possible to achieve such pointing accuracy, the requirements may be relaxed by using dual recycling. In a sense, this is because the rejected light is being fed back in; more precisely, any higher modes of the beam which are required to express an optical imperfection will be resonantly suppressed by the recycling cavity formed by both M_0 and M_3 of Fig. 4. This is the same phenomenon that occurs when a simple two-mirror cavity is, say, misaligned. The effect of the misalignment is to simply reduce the amplitude of the fundamental mode, as described by (50). The requirement that the sensitivity is not significantly degraded then becomes

$$\theta \ll \theta_c \sim 2\times10^{-5} \left[\frac{l}{1\ \text{km}}\right]^{-1/2} \left[\frac{\lambda}{5\times10^{-7}\ \text{m}}\right]^{1/2} \text{rad} . \tag{53}$$

This result is only strictly true if the cavity formed by the recycling mirror M_3 and the interferometer arms has a high finesse and higher modes are nonresonant. For lower finesses, the amplitude of the higher modes will be suppressed by a factor approximately equal to this finesse, and the required angular stability (53) will be relaxed by the same factor.

The mode-cleaning action of the dual-recycling system ensures that the beam emerging from the output mirror M_3 will be pure fundamental mode, giving good fringe contrast and hence good shot-noise-limited sensitivity.

Resonant recycling is also tolerant of geometrical imperfections, for two main reasons. First, the build up of intensity in the recycling system is determined by reflection at the recycling mirror (as in a simple cavity) and is not dependent on interference at a beamsplitter. Second, the two countercirculating beams travel through the same optics, so the quality of the final interference is not sensitive to geometrical imperfections.

Thus, both dual recycling and resonant recycling place significantly less stringent requirements on the geometrical quality of the optics than does standard recycling.

CONCLUSION

It has been seen that the shot-noise-limited sensitivity of a laser-interferometric gravitational-wave detector may be greatly enhanced by using recycling, but that this improvement may be obtained in several ways. In each case there is no significant difference in the possible sensitivity if delay lines or cavities are used in the arms of the interferometer. The arrangements known as standard or broadband recycling, resonant recycling, and the new technique of dual recycling all give, at least in principle, the same performance both when operating in broadband (12) and narrow-band (22) modes. Dual recycling does, however, have some desirable features. First, the optical elements in the interferometer arms, be they cavities or delay lines, are no longer required to have a storage time comparable to the gravitational-wave period in order to get the best sensitivity. Not only does this give the system great flexibility, the reduction in the required number of reflections in a delay line may make the low-frequency operation of such an arrangement considerably easier. Second, while resonant recycling can be tuned over a range of frequencies, it has been seen that dual recycling has a less restricted tuning range and gives a better performance when tuned away from its optimum frequency. Third, the sensitivity-bandwidth combination of dual recycling may be changed by altering the transmission of a single mirror. Fourth, while dual recycling retains the basic optical arrangement of standard recycling, it has a greatly reduced sensitivity to geometrical imperfections in the optics. Thus, while much work needs to be done in order to know how to implement recycling, it seems likely that the use of dual recycling will considerably enhance the operational performance of a gravitational-wave observatory.

ACKNOWLEDGMENTS

I would like to thank the rest of the Glasgow Gravitational Waves Group, together with Alain Brillet, for many interesting discussions. I am grateful to the Science and Engineering Research Council and the University of Glasgow for financial support.

[1] K. S. Thorne, in *300 Years of Gravitation*, edited by S. W. Hawking and W. Israel (Cambridge University Press, Cambridge, 1987).

[2] J. Hough, B. J. Meers, G. P. Newton, N. A. Robertson, H. Ward, B. F. Schutz, I. F. Corbett, and R. W. P. Drever, Vistas Astron. 30, 109 (1987).

[3] R. W. P. Drever, in *Gravitational Radiation*, edited by N. Deruelle and T. Piran (North-Holland, Amsterdam, 1983).

[4] M. Born and E. Wolf, *Principles of Optics* (Pergamon, London, 1959).

[5] B. J. Meers, Ph.D. thesis, University of Glasgow, 1983.

[6] R. W. P. Drever, personal communication.

[7] J.-Y. Vinet, B. J. Meers, C. N. Man, and A. Brillet, Phys. Rev. D **38**, 433 (1988).

Ref. K

THE FREQUENCY RESPONSE OF INTERFEROMETRIC GRAVITATIONAL WAVE DETECTORS

B.J. MEERS

Department of Physics and Astronomy, University of Glasgow, Glasgow G12 8QQ, Scotland, UK

Received 8 September 1989; accepted for publication 27 October 1989
Communicated by J.P. Vigier

The frequency response of interferometric gravitational wave detectors is derived. The method and result are applicable to interferometers with delay lines or cavities and which use either standard, detuned or dual recycling. This should allow detectors to be optimised for different types of gravitational wave signal.

Long baseline laser interferometers for the detection of gravitational radiation are currently being proposed in several countries (see ref. [1] for a review). The optics of these interferometers may be arranged in various ways: delay lines or cavities may be used to increase the signal, and either standard, detuned, dual or resonant recycling [2–4] may then be employed to further improve performance. Here, we take a new analytical approach to the calculation of the response of these systems to a gravitational wave. This approach is applicable to all of the optical systems, though a slight extension is necessary for the rather different optical arrangement of resonant recycling. Thus, we shall end up with a single formula describing the frequency response of a gravitational wave detector, which will be valid whether the interferometer contains delay lines or cavities, whether or not recycling is used and whether the recycling arrangement is standard, detuned or dual. This will allow the detector to be optimised for types of source with different spectral distributions – bursts, chirps from coalescing binaries, stochastic backgrounds, and continuous signals.

The basic optical arrangement of these systems is shown in fig. 1. A suitable polarised gravitational wave induces opposite length changes in the two arms of the interferometer, producing phase shifts on the sensing light which are converted into intensity changes by interference at the beamsplitter. These intensity changes constitute the gravitational wave signal.

We shall consider a single Fourier component of a gravitational wave, which we will regard as phase modulating the light to give two sidebands. The interferometer is arranged so that, when the light beams from the

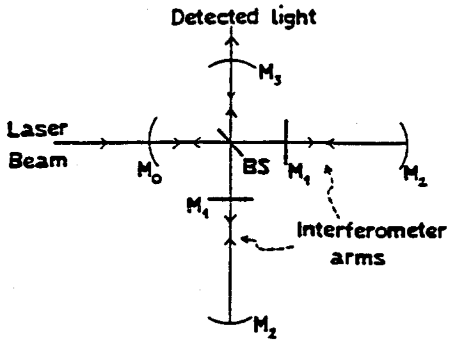


Fig. 1. The optical arrangement of an interferometric gravitational wave detector. The Michelson interferometer detects induced phase differences, which are made larger by the multiple bounce system formed between mirrors M_1 and M_2. Standard recycling places mirror M_0 to recycle the laser power while dual recycling adds M_3 to recycle signal sidebands.

0375-9601/89/$ 03.50 © Elsevier Science Publishers B.V. (North-Holland)

two arms meet at the beamsplitter, the original laser frequency heads back to mirror M_0, while any sidebands produced by differential phase modulation travel towards mirror M_3 and the output of the interferometer. The mirrors M_0 and M_3 will, in general, have different relative positions and reflectivities, which must be taken into account when calculating how both laser light and sidebands resonate. So the sideband amplitude emerging from the output mirror may be found if we regard the optical system as consisting of a split cavity, in which the laser frequency and sidebands experience different reflectivities and optical lengths. The signal in this case is just that emerging from an equivalent single cavity with (composite) input mirror M_1, and end mirror M_2, length L. The input mirror has amplitude reflection and transmission coefficients R_{1c}, T_{1c} for the carrier (laser frequency) and R_{1s}, T_{1s} for the two sidebands (we will only consider the usual case of $R_{1+}=R_{1-}$). The end mirror has reflectance R_2 for all frequencies. It is the interpretation of R_{1c}, R_{1s} etc. that determines the optical system to which the results apply.

For a simple delay line system, with no recycling, $R_{1c}=R_{1s}=0$, R_2 is the reflectivity of the whole delay line and L must be regarded as the optical length of the delay line. If standard recycling is used, the carrier reflectivity R_{1c} is just equal to the reflectivity of the recycling mirror M_0, while in dual recycling it is M_3 that determines the sideband reflectivity R_{1s}.

In a cavity system, R_2 is just the reflectivity of the end mirror and L is the length of the cavity. With no recycling, R_1 is frequency independent; it is simply the reflectivity of the cavity input mirror. In standard and detuned recycling, R_{1c} is determined by the combination of M_0 and M_1, which form a cavity. In dual recycling, R_{1s} is defined by the combination of M_1 and the output mirror M_3.

Let us consider a gravitational wave of optimum polarisation, amplitude $h\cos\omega_g t$, incident on the detector. In order to calculate the effect that this has on the light in our equivalent cavity, we will regard the light as being the sum of many beams, each of which has experienced a different number of reflections in the cavity. We will first determine the influence of the gravitational wave on one beam, then add the contributions from all the beams.

If the transit time of one beam is τ, the differential phase shift $\delta\phi$ induced on the light in one round trip is

$$\delta\phi=\int_{t-\tau}^{t}\omega h\cos\omega_g t\,\mathrm{d}t=\frac{h\omega}{\omega_g}\sin(\omega_g\tau/2)\{\exp[\mathrm{i}\omega_g(t-\tau/2)]+\exp[-\mathrm{i}\omega_g(t-\tau/2)]\}\,, \tag{1}$$

where we have used the exponential representation of $\cos\omega_g t$, and ω is the angular frequency of the light. The effect of this phase shift is to multiply the field by $e^{i\delta\phi}\approx 1+i\delta\phi$, producing sideband fields at angular frequencies offset by $\pm\omega_g$ from the laser frequency. The total sideband field is found by adding the contributions generated on each bounce: taking just the $+\omega_g$ field E_+ emerging from the cavity,

$$\begin{aligned}\frac{E_+}{E_0}&=\mathrm{i}T_{1c}T_{1s}R_2(h\omega/\omega_g)\sin(\omega_g\tau/2)\exp[\mathrm{i}\omega_g(t-\tau/2)]\\&\quad\times\sum_{N=1}^{\infty}(R_{1c}R_2)^{N-1}\exp(\mathrm{i}N\delta_c)\sum_{n=1}^{\infty}[R_{1s}R_2\exp(-\mathrm{i}\omega_g\tau)]^{n-1}\exp(\mathrm{i}n\delta_s)\,,\end{aligned} \tag{2}$$

where E_0 is the incident laser field and δ is the phase offset of the cavity from resonance: δ_c is that for the carrier, δ_s that for the sidebands. The $\exp(-\mathrm{i}n\omega_g\tau)$ term reflects the change in phase of the signal over the history of the light stored in the cavity. Summing the series gives

$$\frac{E_+}{E_0}=\frac{\mathrm{i}T_{1c}T_{1s}R_2h\omega\sin(\omega_g\tau/2)\exp[\mathrm{i}\omega_g(t+\tau/2)]\exp[\mathrm{i}(\delta_c+\delta_s-\omega_g\tau)]}{\omega_g[1-R_{1c}R_2\exp(\mathrm{i}\delta_c)]\{1-R_{1s}R_2\exp[\mathrm{i}(\delta_s-\omega_g\tau)]\}}\,, \tag{3}$$

which may be rewritten as

$$\frac{E_+}{E_0}=\frac{\mathrm{i}T_{1c}T_{1s}R_2h\omega\sin(\omega_g\tau/2)\exp[\mathrm{i}\omega_g(t+\tau/2)][\exp(\mathrm{i}\delta_c)-R_{1c}R_2]\{\exp[\mathrm{i}(\delta_s-\omega_g\tau)]-R_{1s}R_2\}}{\omega_g(1-R_{1c}R_2)^2(1-R_{1s}R_2)^2[1+F'_c\sin^2(\delta_c/2)]\{1+F'_s\sin^2[(\delta_s-\omega_g\tau)/2]\}}\,, \tag{4}$$

where

$$F_c' = \frac{4R_{1c}R_2}{(1-R_{1c}R_2)^2}, \qquad F_s' = \frac{4F_s^2}{\pi^2} = \frac{4R_{1s}R_2}{(1-R_{1s}R_2)^2}.$$

F is the finesse of the cavity. It is always optimum to arrange for $\delta_c=0$, corresponding to an isolated cavity being on resonance, or maximum power build up in a recycled system. With this choice, (4) simplifies to

$$\frac{E_+}{E_0} = \frac{\mathrm{i}T_{1c}T_{1s}R_2 h\omega \sin(\omega_g \tau/2) \exp[\mathrm{i}\omega_g(t+\tau/2)]\{\exp[\mathrm{i}(\delta_s-\omega_g\tau)]-R_{1s}R_2\}}{\omega_g(1-R_{1c}R_2)(1-R_{1s}R_2)^2\{1+F_s' \sin^2[(\delta_s-\omega_g\tau)/2]\}}. \tag{5}$$

The expression contains virtually all of the information we need (for the other sideband, $\omega_g \to -\omega_g$). For example, the enhancement of the sideband amplitude by the choice of $\delta_s=\omega_g\tau$ (as in dual and detuned recycling) can be seen.

The sidebands are detected by beating them with a local oscillator field E_L to produce an intensity change ΔI:

$$\Delta I = (E_L + E_+ + E_-)(E_L^* + E_+^* + E_-^*), \tag{6}$$

where the asterisk denotes complex conjugation. The fluctuating intensity δI is

$$\delta I = E_L E_+^* + E_L E_-^* + E_L^* E_+ + E_L^* E_-. \tag{7}$$

E_L may be an external field or may be some of the internal field which is allowed to leak out. For simplicity we will take E_L to be the original laser frequency and to be in the quadrature phase (as an internal field would be). Insertion (5) into (7) then gives, after some considerable algebra,

$$\begin{aligned}\delta I = {} & \frac{4E_0 E_L T_{1c}T_{1s}R_2 h\omega \sin(\omega_g\tau/2)}{\omega_g(1-R_{1c}R_2)(1-R_{1s}R_2)^2\{1+F_s' \sin^2[(\delta_s+\omega_g\tau)/2]\}\{1+F_s' \sin^2[(\delta_s-\omega_g\tau)/2]\}} \\ & \times\Bigg\{\sin[\omega_g(t+\tau/2)] \sin\omega_g\tau \left(\cos\delta_s + \frac{2F_s^2}{\pi^2}(\cos\delta_s - \cos\omega_g\tau)\right) \\ & +\cos[\omega_g(t+\tau/2)]\left[(\cos\delta \cos\omega_g\tau - R_{1s}R_2)\left(1+\frac{2F_s^2}{\pi^2}(1-\cos\delta_s \cos\omega_g\tau)\right) + \frac{2F_s^2}{\pi^2}\sin^2\delta_s \sin^2\omega_g\tau\right]\Bigg\}.\end{aligned} \tag{8}$$

This expression gives the intensity change at the output of the interferometer as a function of gravitational wave frequency ω_g, for any combination of detector parameters. It is a little complicated, but this is a result of its generality. Let us take a few examples.

The simplest case is that of a non-recycled delay line. Putting $R_{1c}=R_{1s}=F_s=\delta_s=0$, (8) reduces to

$$\delta I = 4E_0 E_L R_2 (h\omega/\omega_g) \sin(\omega_g\tau/2) \cos[\omega_g(t-\tau/2)], \tag{9}$$

which has the usual frequency dependence of a simple delay line. The sensitivity limit set by statistical fluctuations in the photocurrent can be found by demanding that the change in current $\eta\delta I$ (where η is the power to current efficiency of the photodiode) is greater than that due to shot noise, which has linear spectral density $(2e\eta E_L^2)^{1/2} A/\sqrt{\mathrm{Hz}}$, with e the electronic charge. Remembering that there are two noise sidebands, and converting to a quantum efficiency $\xi = 2\pi\hbar c\eta/e\lambda$ ($\hbar$ is the reduced Planck constant and λ the light wavelength), we find that the equivalent gravitational wave amplitude spectral density is

$$h_{\mathrm{rms}} = \left(\frac{\pi\hbar\lambda\nu_g^2}{\xi I_0 c \sin^2(\omega_g\tau/2)}\right)^{1/2} \mathrm{Hz}^{-1/2}, \tag{10}$$

where ν_g is the gravitational wave frequency. Reassuringly, this is the same expression that is found by rather more direct derivations [1].

If standard recycling is used with a delay line, the intensity change (9) is increased by a factor $T_{1c}/(1-R_{1c}R_2)$. With delay line mirrors of loss coefficient A^2 $(R^2+A^2=1)$ and losses dominated by the delay line, $R_2=1-c\tau A^2/2l$, where l is the length of the delay line. The optimum recycling mirror transmission is then $T_{1c}(\text{opt})=(c\tau A^2/l)^{1/2}$, and the sensitivity (10) is improved by the same factor.

If dual recycling is used in its broadband mode, where the signal storage time does not exceed a typical gravitational wave period, then the signal phase offset may be set to zero. The output response (8) then simplifies to

$$\delta I=\frac{4E_0E_LT_{1c}T_{1s}R_2h\omega\sin(\omega_g\tau/2)\,[\cos(\omega_g\tau/2)\cos\omega_g t+(2F_s/\pi)\sin(\omega_g\tau/2)\sin\omega_g t]}{\omega_g(1-R_{1c}R_2)(1-R_{1s}R_2)[1+F'_s\sin^2(\omega_g\tau/2)]}\,, \tag{11}$$

or

$$|\delta I|=\frac{4E_0E_LT_{1c}T_{1s}R_2h\omega\sin(\omega_g\tau/2)}{\omega_g(1-R_{1c}R_2)(1-R_{1s}R_2)[1+F'_s\sin^2(\omega_g\tau/2)]^{1/2}}\,. \tag{12}$$

This relation applies to a delay line with dual recycling, an isolated cavity on resonance, or a cavity system with either standard or dual recycling. A delay line looks just like a longer cavity (remember that τ is the storage time for a delay line, the round trip time $2l/c$ for a cavity). Indeed, a delay line interferometer with dual recycling may be regarded as a folded, split Fabry–Perot cavity. It is interesting to see that in the low frequency limit $\omega_g\tau\ll 1$ all of these optical systems have the same frequency response.

The broadband sensitivity will be optimised at frequency ν_{g0} if the signal finesse is chosen so that $F'_s\sin^2(\pi\nu_{g0}\tau)\approx 1$. If we take the case of $\omega_{g0}\tau\ll 1$ and the losses being limited by the delay line/cavity mirrors, the shot noise limited sensitivity at ν_{g0} is

$$h_{\text{rms}}=\left(\frac{\hbar\lambda A^2\nu_{g0}\Delta\nu_g}{\xi I_0 l}\right)^{1/2}. \tag{13}$$

This is the same as that obtained with an optimised cavity system with standard recycling.

The frequency response of a broadband dual recycling system is described by (12) and illustrated in fig. 2.

If a dual recycling system is tuned to a frequency ν_{g0}, so that $\delta_s=\pm\omega_g\tau$, then the full expression (8) must usually be used to describe the frequency response. Some different combinations of sensitivity and bandwidth (set by the signal storage time $F_s\tau$) are shown in fig. 3. The best combination will depend on the type of source being searched for.

In the case of a narrow bandwidth, $F_s\nu_g\tau\gg 1$, but with $\omega_g\tau\ll 1$, the frequency response (8) is approximately

$$\delta I=\frac{2E_0E_LT_{1c}T_{1s}R_2h\omega\tau}{(1-R_{1c}R_2)(1-R_{1s}R_2)\{1+[2F_s\tau(\nu_g-\nu_{g0})]^2\}^{1/2}} \tag{14}$$

for frequencies ν_g close to the centre frequency ν_{g0}. So the FWHM bandwidth is $1/F_s\tau$. If the maximum peak sensitivity is desired, the recycling mirrors should be chosen so that $T^2_{1c}=T^2_{1s}=c\tau A^2/l$. The resultant shot noise imited sensitivity is

$$h_{\text{rms}}=\frac{A^2}{l}\left(\frac{2\hbar\lambda c\Delta\nu_g}{\pi\xi I_0}\right)^{1/2}\left[1+\left(\frac{2\pi l}{cA^2}(\nu_g-\nu_{g0})\right)^2\right]^{1/2}, \tag{15}$$

which has an FWHM bandwidth of $cA^2/\pi l$. If the signal is integrated for a time τ_{int}, such as in a search for a continuous gravitational wave, the measurement bandwidth is $\Delta\nu_g=1/\tau_{\text{int}}$.

Note that, for a detector with $\omega_g\tau\ll 1$, this optimum sensitivity can be obtained at any frequency, with the same set of optics. Only the position of M_3 needs to be changed to alter the tuning.

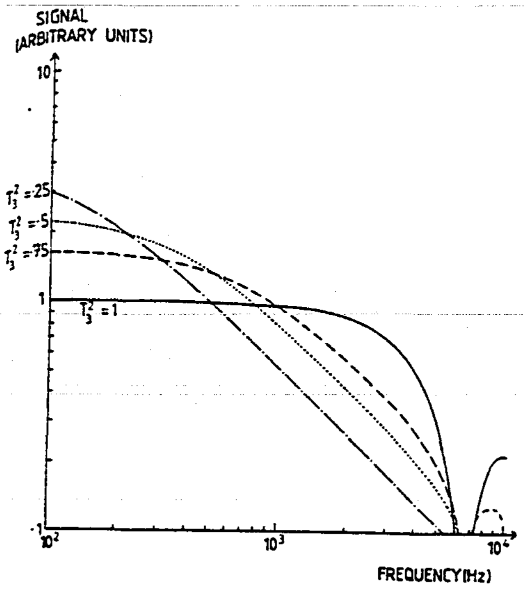


Fig. 2. The frequency response of a 3 km, 16 reflection delay line with different degrees of broadband ($\delta_s=0$) dual recycling. The curves differ only in the value of the transmission coefficient $T_3^2=T_{1s}^2$ of the signal recycling mirror. So dual recycling enables good low frequency performance to be obtained even with a low storage time delay line.

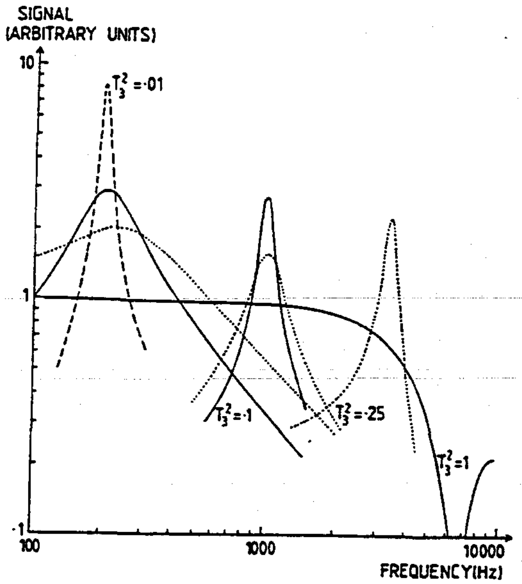


Fig. 3. The frequency response of a 3 km, 16 reflection delay line with different degrees of tuned ($\delta_s=\omega_{g0}\tau$) dual recycling. Three different tunings are shown, with different sensitivity/bandwidth combinations determined by the reflectivity of the signal recycling mirror M_3.

As can be seen from (8) and fig. 3, a narrowband delay line system with $\sin(\omega_g\tau/2)=1$ gives performance better than a system with $\omega_g\tau\ll 1$ by a factor $4/\pi$. This may be understood if it is realised that, although a factor of $\pi/2$ is lost as the sidebands add in the longer delay line [4], *both* sidebands resonate in the dual recycling system.

If cavities are used in the arms of the interferometer, then the equivalent input "mirrors" are also cavities. We need to know how to choose the reflectivity of M_0 and M_3. The transmission coefficient T_{1c} is

$$T_{1c}=\frac{T_0T_1}{(1-R_0R_1)[1+F_0'\sin^2(\delta_0/2)]}, \tag{16}$$

where F_0 is the finesse and δ_0 the phase offset of the input "mirror". If the interferometer cavity is run on resonance (as an isolated cavity), then this (isolated) recycling cavity should be off resonance, in order to act as a perfect mirror. In this case, we have

$$T_{1c}=\tfrac{1}{2}T_0T_1, \tag{17}$$

which allows the choice of T_0. A similar relation applies to the signal recycling mirror. Note that a signal recycling cavity which is off resonance gives the same reflection coefficient and phase offset for both sidebands, justifying our earlier assumption.

In detuned recycling, the interferometer cavities are run off resonance, which means the power recycling cavity must be operated nearly on resonance if the intensity in the interferometer is to be kept high (i.e. $\delta_c=0$). However, this means that the power circulating in the recycling cavity is greatly increased, enhancing the significance of any losses there. Our assumption that the losses are dominated by the cavity mirrors is probably unjustified in this case [3,4]. Since our analysis has shown that dual recycling has exactly the same frequency response as ideal detuned recycling, it would seem that dual recycling gives both the best performance and the most flexibility.

We have not specifically discussed resonant recycling here, but the same approach could be taken: in this case the frequency dependent reflection does not follow a physical separation of the different frequencies but arises from reflection of all of the light from a cavity or delay line. This imposes constraints on tuning [4] but a similar frequency response to that of dual recycling will be obtained.

Conclusion. We have seen that the various forms of interferometric gravitational wave detector can be regarded as equivalent to a cavity whose length is modulated by a gravitational wave and the input mirror of which has a frequency dependent complex reflectivity. This viewpoint allows a common analysis of the different optical arrangements. We have seen that a low storage time delay line with dual recycling, an optimized cavity system with standard recycling and a low storage time cavity with dual recycling all give the same sensitivity and frequency response, as long as the signal storage time $F_s\tau$ is the same in all cases. The general expression which we have derived for the frequency response of recycled interferometers should enable the optimum combination of sensitivity, bandwidth and tuning to be found for each type of gravitational wave.

I would like to thank SERC and the University of Glasgow for financial support.

References

[1] K.S. Thorne, in: 300 years of gravitation, eds. S.W. Hawking and W. Israel (Cambridge Univ. Press, Cambridge, 1987).

[2] R.W.P. Drever, in: Gravitational radiation, eds. N. Deruelle and T. Piran (North-Holland, Amsterdam, 1983).

[3] J.-V. Vinet, B.J. Meers, C.N. Man and A. Brillet, Phys. Rev. D 38 (1988) 433.

[4] B.J. Meers, Phys. Rev. D 38 (1988) 2317.

Doubly-resonant signal recycling for interferometric gravitational-wave detectors.

Brian J. Meers
Department of Physics and Astronomy, University of Glasgow,
Glasgow G12 8QQ, Scotland

Ronald W. P. Drever
California Institute of Technology, Pasadena,
California 91125.

Abstract

We describe a new optical system for laser-interferometric gravitational-wave detectors. An extension of dual recycling, the arrangement that we propose uses a long storage-time cavity to split the resonance of the signal-recycling cavity, allowing two signal sidebands to be simultaneously resonant. We show that this couples out the gravitational-wave signal more efficiently than a conventional singly-resonant interferometer. The result is both better potential gravitational-wave sensitivity and the possibility of making narrowband observations simultaneously at two different frequencies.

1 Introduction

The detection of gravitational radiation remains one of the great experimental challenges. With potential sources including supernovae, coalescing compact-object binaries, pulsars, cosmic strings or even the early stages of the big bang (see Thorne [1] for a review), the rewards of detection will be high for both astronomy and physics. Such is the weakness of the interaction of gravitational waves with matter that every effort must be made to optimise this interaction in the detectors. Better detectors will be able to see less efficient or more distant sources, or extract more information from the signal.

A particularly promising type of gravitational-wave detector is the long-baseline laser interferometer [1], the basic arrangement of which is shown in fig. 1. The fluctuations in the curvature of spacetime associated with the gravitational wave induce different phase shifts on the light in the two arms of the interferometer, resulting in changes in the power emerging from the output port of the beamsplitter. These changes in output power should be as large as possible if even weak signals are to be seen above the statistical fluctuations in photon number. It is therefore sensible to employ high power lasers and to use the light efficiently: recycling [2] of the light (as shown in fig. 1) maximises the circulating power and increases the signal. Furthermore, the phase change induced on the light is enhanced if the light is allowed to interact with the gravitational wave for a long time, by using multiple reflections in the arms of the interferometer for example. Either optical cavities or delay lines may be used to achieve this. Recycling with a delay line is made easier by using a mirror to reflect the light back along itself, so that it enters and leaves on overlapping paths. However, the gain from the use of multiple reflections is limited, for the sign reverses every half-period of the gravitational wave, tending to cancel out the signal if longer light storage times are used. This is why, in the simple arrangement of fig. 1, the signal should be extracted promptly even though the light power (carrier) can be recycled so that it has the maximum possible interaction time, set by the losses of the system. Nevertheless, a larger signal within a reduced bandwidth can be obtained with somewhat different optical arrangements. This improvement of the signal to noise ratio within a narrow band will be especially useful when looking for signals that are monochromatic (such as those from pulsars), quasi-monochromatic (from coalescing compact binaries [3, 4]) or stochastic (from cosmic strings, for example [1, 5]). The first narrowband arrangement suggested was resonant recycling [2]. The later variants detuned recycling [6] and dual recycling [5] more closely resemble the Michelson interferometer of fig. 1. While these optical systems share the principle of simultaneously resonating the light power and at least one signal sideband induced by the gravitational wave, the most flexible and practical seems to be dual recycling [5]. The optical arrangement of dual recycling is indicated in fig. 2. It can be seen that the interferometer layout is the same as that shown in fig. 1, but with the addition of the partially transmitting signal recycling mirror M_3 at the output. The system works because the interference at the beamsplitter directs light of the original laser frequency back to the power recycling mirror M_0, while sidebands resulting from the differential phase modulation produced by the gravitational wave travel to M_3. The position of M_3 can then be chosen so that at least one signal sideband is resonant within the signal recycling cavity that is formed between M_3 and the arms of the interferometer. The flexibility of dual recycling is a consequence of the independent resonance of light power and signal sidebands. With the position of the signal recycling mirror M_3 defining the tuning, the transmission determines the signal storage time, hence the signal bandwidth. Transfer functions and different sensitivity-bandwidth combinations are discussed in ref. [7]. The operation of a small dual recycling system, including a seven-fold enhancement of the signal to noise ratio, has been experimentally demonstrated recently [8].

While the power recycling component of dual recycling may be viewed as a method of impedance matching in of the laser beam, the signal recycling part may be considered to be impedance matching out of the signal.

While narrowbanding can improve the gravitational-wave sensitivity by an order of magnitude, none of the previously proposed systems gives what might be considered to be ideal performance. The reason for this is that only *one* of the signal sidebands can, in general, be resonant within the optical system. The other sideband usually lies at some arbitrary fraction of the free spectral range of the signal recycling cavity away from the resonant frequency, so is of small amplitude. This lack of contribution to the signal from one sideband is evidently not efficient. There is one situation in which both sidebands are resonant, that of a delay line being used in the arms of the interferometer to increase the storage time to half of the gravitational-wave period. The free spectral range of the signal recycling cavity is then equal to the sideband separation, allowing both sidebands to be resonant. Even in this case, the somewhat inefficient way in which the sidebands add within the arms of the interferometer reduces the signal by a factor $\pi/2$ from what may be regarded as its maximum value [5, 7]. Possibly more serious, such a system is very inflexible: changing frequency is difficult, and it may be impossible to obtain a sufficiently long storage time in the arms for low frequency operation. What we want is a system that allows both signal sidebands to be resonant for essentially any tuning frequency. We need a round-trip time within the arms of the interferometer that is small compared with the gravitational-wave period, so that the sidebands generated during one round trip add in phase, but we also need a way of recycling the two sidebands so that they are *both* in resonance. Phrasing the problem in this way points to the solution—the use of a frequency-dependent signal recycling mirror. If the two sidebands see sufficiently different phase shifts on reflection off M_3, it can be arranged that both the signal sidebands add coherently when recycled. The sidebands will then build up resonantly, giving an increased gravitational-wave signal.

So, how do we make a frequency-dependent mirror with the desired properties? It is clear that we only need the phase shift on reflection to be different for the two sidebands; the magnitude of the reflectivity should be the same. We can imagine using some sort of delay line system to change the relative phases of the sidebands, but this suffers from the same inflexibility mentioned earlier. The alternative is to use a high storage time cavity as the signal recycling mirror, as shown in fig. 3. The variation of phase-shift-on-reflection across the cavity resonance then ensures resonance for both sidebands. This scheme may also be regarded as a coupling of the output cavity (M_3) and the signal recycling cavity, producing a double resonance in which both sidebands may resonate. The frequency splitting will be determined by the degree of coupling of the cavities (the transmission T_{3m} of the middle mirror in the cavity that makes up M_3), while the sensitivity-bandwidth combination will be set by the rate of energy loss from the system (the transmission T_{3e} of the end mirror).

In the next section we will discuss quantitatively the properties of this method of doubly-resonant signal recycling. We shall consider the enhancement of photon-noise-limited sensitivity and the dependence of the transfer function on the properties of the output cavity. The first case to be discussed will be that of the two sidebands corresponding to a single gravitational-wave frequency being resonant. We will then consider the option of resonating one sideband from two different frequencies.

2 Response of doubly-resonant signal-recycling

2.1 Analytic results

Ref. [7] describes a method of analysing general interferometer configurations. Any detector with the same basic layout as that shown in fig. 2 may be represented by a single equivalent cavity with a frequency-dependent input mirror. Thus, the light of the original frequency (the carrier) and the signal sidebands can have different storage times and resonance conditions. In most situations, the frequency dependence of the equivalent input mirror is a consequence of the physical separation of the carrier and sidebands, followed by their independent recycling. This formalism is also ideally suited to situations in which the two sidebands may have different reflectivities. With the effective input mirror having amplitude transmission and reflectivity seen by the carrier of T_{1c} and R_{1c}, while T_{1+} and R_{1+} are the corresponding quantities seen by the upshifted (plus) sideband, the emerging amplitude E_+ of one of the sidebands induced by a gravitational wave $h = h_0 \cos \omega_g t$ is just given by equation (5) of ref. [7]:

$$\frac{E_+}{E_0} = \frac{iT_{1c}T_{1+}R_2h_0\sin(\omega_g\tau/2)e^{i\omega_g(t+\tau/2)}\left[e^{i(\delta_+-\omega_g\tau)} - R_{1+}R_2\right]}{\omega_g(1-R_{1c}R_2)(1-R_{1+}R_2)^2\left[1+F'_+\sin^2[(\delta_+-\omega_g\tau)/2]\right]^{\frac{1}{2}}}, \qquad (1)$$

where τ is the round trip time for light in the arms of the interferometer, R_2 is the amplitude reflectivity of the arms and δ_+ is the phase offset from resonance seen by the light on a round trip of the signal recycling cavity. The finesse F_+ seen by the sideband is given by

$$\frac{4F_+^2}{\pi^2} = F_+' = \frac{4R_{1+}R_2}{(1-R_{1+}R_2)^2} \, . \tag{2}$$

For the other (minus) sideband, $\omega_g \rightarrow -\omega_g$, $R_{1+} \rightarrow R_{1-}$, *etc.* Equation (1) just describes how well a signal sideband resonates within the optical system. One thing that we can see immediately is that the condition for both sidebands to be resonant is that

$$\sin[(\delta_+ - \omega_g\tau)/2] = \sin[(\delta_- + \omega_g\tau)/2] = 0 \, . \tag{3}$$

This may be achieved if the sidebands experience phase shifts θ_+, θ_- on reflection off the output cavity-mirror such that

$$\theta_+ - \theta_- = 2\pi - 2\omega_g\tau \, . \tag{4}$$

As long as this condition is satisfied, the overall position of M_3 may be adjusted to ensure resonance. Now the phase shift on reflection off a cavity [5] is given by

$$\tan\theta_+ = \frac{(\alpha_3 F_3/\pi)\sin(\delta_3 + \omega_g\tau_3)}{1 - \alpha_3 + F_3' \sin^2[(\delta_3 - \omega_g\tau_3)/2]} \, , \tag{5}$$

with the subscript of δ_3, τ_3 referring to the cavity M_3, and

$$\alpha_3 = \frac{T_{3m}^2 R_{3e}}{1 - R_{3m}R_{3e}} \, . \tag{6}$$

R_{3m}, R_{3e} are the reflectivities of the middle and end mirrors, respectively, of the cavity M_3. For typical cavities of interest here, the ratio of amplitudes factor α_3 will be a little less than 2. (Strictly speaking, the factor F_3' in equation (5) should be $\frac{2F_3}{\pi}\left(\frac{2F_3}{\pi} + \alpha_3\right)$.) The variation of the phase shift on reflection across the cavity resonance curve is shown in fig. 4(a). It can be seen that θ changes by almost 2π on going from one side of resonance to the other. So, if we arrange for the two gravitational-wave sidebands to lie on opposite sides of the output cavity resonance the resonance condition (4) may be satisfied. It is evident that the cavity must have a sufficiently long storage time for its linewidth to be less than the gravitational-wave frequency. If the losses associated with this cavity are to be small, it must also be physically long.

This viewpoint and analysis applies equally whether cavities or delay lines are used in the arms of the interferometer. However, the interpretation of what is meant by the mirror M_{3m} does vary. With delay lines, this is simply the middle mirror of the output cavity. But with cavities in the arms, the 'mirror' M_{3m} must be considered to be a composite mirror, consisting of the cavity formed by the input mirror to the cavity in the arms and the middle mirror of the output cavity. T_{3m} may be easily calculated using standard relations, similar to equation (10) below. Since the length of this cavity-mirror M_{3m} will be much shorter than the armlength, this extra complication will have a very small effect upon the frequency response. The use of multiple reflections on either side of the beamsplitter is only necessary in order to make the losses associated with the beamsplitter negligible.

Let us try and obtain some feeling for the required properties of the output cavity. If we restrict our attention initially to the case of $\delta_3 = 0$, so that the sidebands lie symmetrically on either side of the resonance curve of M_3, and low output transmission, with the sidebands lying reasonably far down the resonance curve (see fig. 4(b)), $|1 - \alpha_3| \ll F_3' \sin^2[(\delta_3 - \omega_g \tau_3)/2]$, then a little manipulation of (4) and (5) gives the requirement for double resonance:

$$T_{3m}^2 = 2 \tan(\omega_g' \tau_3/2) \tan \omega_g' \tau \, . \tag{7}$$

It is interesting to rewrite this to give the frequency ω_g' for resonance in the limit of $\omega_g' \tau \ll 1$:

$$\omega_g' \approx \frac{T_{3m}}{\sqrt{\tau \tau_3}} \, . \tag{8}$$

Thus, the resonant frequency does only depend on the transmission T_{3m} of the middle mirror, not on T_{3e}. This confirms and quantifies the intuition that the splitting of the resonance is determined by the coupling of the two cavities.

The signal buildup is determined by the transmission of the signal recycling cavity-mirror: peak signal $\propto |T_{1+}|^{-1}$ if the losses are negligible. This transmission is

$$T_{1+} = \frac{T_{3m} T_{3e} \left[e^{i(\delta_3 - \omega_g \tau_3)} - R_{3m} R_{3e} \right]}{(1 - R_{3m} R_{3e})^2 \left[1 + F_3' \sin^2[(\delta_3 - \omega_g \tau_3)/2] \right]} \, , \tag{9}$$

or

$$|T_{1+}| = \frac{T_{3m} T_{3e}}{(1 - R_{3m} R_{3e}) \left[1 + F_3' \sin^2[(\delta_3 - \omega_g \tau_3)/2] \right]^{\frac{1}{2}}} \, . \tag{10}$$

With symmetrical operation and fairly low transmission, putting in the resonance condition (8) gives the signal transmission purely in terms of the transmission of the end mirror in the output cavity:

$$|T_{1+}| \approx T_{3e} \sqrt{\tau / \tau_3} \, . \tag{11}$$

So the peak signal size is indeed entirely determined by the leakage rate through the end mirror.

It is both interesting and straightforward to compare the bandwidth of a doubly-resonant signal recycling interferometer with that of a simple system with the same signal transmission and storage time. The bandwidth of the doubly-resonant system will be narrower because of the change in phase shift on reflection θ away from the resonant frequency. Differentiating (5) and again using the condition (8) for resonance gives the additional phase change $\Delta\theta$ on moving away $\Delta\omega_g$ from resonance:

$$\Delta\theta \approx \Delta\omega_g \tau \, . \tag{12}$$

But, as can be seen from inspecting (1), this is just enough to *double* the phase shift from resonance. So the bandwidth for a given output transmission is simply halved.

As long as the signal finesse is sufficiently low for the losses to be negligible, the system considered above will give twice the signal (two sidebands) with half the bandwith of a singly-resonant interferometer. With the same bandwidth, the peak signal would be increased by a factor of $\sqrt{2}$. Alternatively, a doubly-resonant interferometer with the same peak sensitivity will have twice the bandwidth of a singly-resonant system.

If the maximum peak signal is desired then the losses, both in the arms of the interferometer and in the output cavity, must be taken into account. The reflectivity of the output cavity may be found from (10), but it is sometimes more convenient [5] to write it as

$$R_{1+}^2 \approx 1 - \frac{(2F_3/\pi)(2A^2 + T_{3e}^2)}{1 + F_3' \sin^2[(\delta_3 - \omega_g \tau_3)/2]} \, . \tag{13}$$

A^2 is the loss coefficient of the mirrors ($T^2 + R^2 + A^2 = 1$), assumed to be the same for all of the interferometer mirrors. A typical value for A^2 might be 5×10^{-5}. With the same simplifying assumptions that we have used above, it is straightforward to show that, if the output cavity produces double resonance and is a factor η times the armlength of the interferometer, the dissipation associated with it is greater than that in the arms by a factor $2/\alpha_3\eta$. Since $2/\alpha_3 \approx 1$ (typically to within a few percent) for a narrowband system and the maximum signal buildup is proportional to the square root of the total loss [5], the enhancement factor G_{max} for the maximum signal in doubly-resonant signal recycling, compared to that in a simple system, is

$$G_{max} \approx \frac{2}{(1 + 1/\eta)^{\frac{1}{2}}} \, . \tag{14}$$

So with an output cavity-mirror much longer than the arms of the interferometer ($\eta \gg 1$) it is possible to gain a factor of 2 in peak signal compared with a singly-resonant system. If the output cavity is the same length (perhaps using the same vacuum system), $\eta = 1$, a factor of $\sqrt{2}$ is gained. A long output cavity is required if the signal to noise ratio is to be significantly improved.

Note that placing the output cavity along one of the arms of the interferometer does not change significantly the detector's output: the cavity contains a negligible amount of power out of which energy may be pumped.

We have described two complementary ways of looking at doubly-resonant signal recycling: that of a coupled cavity system with the frequency splitting determined by the coupling (T_{3m}), the bandwidth determined by the energy loss rate (T_{3e}); and that of a frequency-dependent signal recycling mirror. An understanding of both viewpoints is useful. The coupled cavity view gives a beautiful insight into the dependence of the tuning and bandwidth on the mirror properties. Similarly, the frequency-dependent mirror view clarifies the origin of the property of doubly-resonant signal recycling that the bandwidth is smaller than that of a singly-resonant system of the same signal transmission. If we seem to concentrate on the frequency-dependent mirror model, it is because it is a very powerful and general aid to both computation and insight.

2.2 Numerical results

By restricting our attention to symmetric operation of the output cavity and fairly narrow bandwidth we have been able to see easily the essential features of doubly-resonant signal recycling. The algebra is more complicated in broadband or asymmetric situations. It is possible to minimise the inconvenience this causes by using a computer to perform the algebraic manipulations. We have done this using *Mathematica.*

The output signal is a change in light power δI produced by beating the sidebands with a local oscillator field [7] E_L:

$$\delta I = E_L E_+^\star + E_L E_-^\star + E_L^\star E_+ + E_L^\star E_- \ , \tag{15}$$

where $\star$ indicates complex conjugation. We know E_+ and E_- from equation (1) and following relations. The optimum phase Φ for the local oscillator field is the one for which all three phasors (E_L, E_+ and E_-) are parallel at one point of the cycle, at the resonant frequency $\omega_g^{'}$:

$$\Phi = \frac{1}{2}\left\{\arg\left[E_+(\omega_g^{'})e^{-i\omega_g^{'}t}\right] + \arg\left[E_-(\omega_g^{'})e^{i\omega_g^{'}t}\right]\right\} \ . \tag{16}$$

This is straightforward to calculate with a system such as *Mathematica.*

Our analysis so far has been general in that both cavities or delay lines are allowed to be within the arms of the interferometer. Doubly resonant signal recycling works equally well in each case. Nevertheless, it should be remembered that with cavities in the arms, the middle mirror M_{3m} is itself another cavity, that formed by the mirrors in the interferometer and output cavities closest to each other. This may well have advantages, such as having a variable transmission. But it is also somewhat more complex. For simplicity, the examples that we will specifically consider will be with delay lines in the interferometer arms. A further convenience is that a choice of a 3 km, 16 reflection delay line, with all mirrors having loss coefficient $A^2 = 5 \times 10^{-5}$, allows a direct comparison with the transfer functions plotted in ref. [7]; we also will plot signal size compared with the low frequency limit of the same optical system with no signal recycling.

2.2.1 A single resonant frequency

Fig. 5 shows the response of such an interferometer with doubly-resonant signal recycling, using an output cavity-mirror the same length as the arms of the interferometer. This example is of a symmetrically run system tuned to a gravitational-wave frequency of 200 Hz. The transmission T_{3m} of the middle mirror of the output cavity is just given by equation (7). The evolution of the detector sensitivity-bandwidth combination as the output transmission T_{3e} is varied, embodied in equation (11), is evident. It can be seen that doubly-resonant signal recycling works well even when the bandwidth is comparable to the observing frequency. The improvement can be seen clearly in fig. 6, in which the transfer function of a singly-resonant interferometer is compared with those of two doubly-resonant systems, one with the same peak response, one with the same FWHM bandwidth. As predicted, the doubly-resonant system with the same bandwidth has $\sqrt{2}$ greater peak response; that with the same peak signal has twice the bandwidth.

In the absence of a conveniently variable middle-mirror transmission T_{3m}, a way of tuning the system to a different frequency is to run the output cavity asymmetrically, slightly off resonance for the original laser frequency. As can be seen from fig. 4 or equation (5), a greater sideband separation will then be required in order to give enough relative phase shift to satisfy the double resonance condition (4). So the doubly-resonant frequency increases with tuning offset δ_3. This is illustrated in fig. 7, which shows the change in the transfer function of a fixed optical system as the tuning of the output cavity is varied. Note the distortion of the shape of the frequency response as the reflectivity for one of the sidebands is increased while that of the other is decreased by changing the tuning of the output cavity.

2.2.2 Multiple resonances

So far we have restricted our attention to situations in which it is arranged that the double resonance of the coupled-cavity system coincides with the two signal sidebands of a single gravitational-wave frequency. There is no reason, however, why one sideband from each of two different frequencies could not be resonated. We could then make narrowband observations simultaneously at two frequencies, for example. This would be useful when looking for different harmonics of a periodic signal, such as that from a pulsar; simultaneous observation of two pulsars; and when making a search for a stochastic gravitational-wave background, yet retaining good sensitivity and timing accuracy for signals from coalescing compact binaries. In addition, with sufficiently sophisticated control systems, one of the resonances could be made to try and track dynamically the evolving chirp from a coalescing binary [4], with a subsequent increase in signal to noise ratio, while the other resonance would remain at the original frequency: this might allow other components (such as those produced at post-Newtonian order [3]) of the chirp waveform to be observed.

Some feeling for the characteristics of such a system with a delay line interferometer may be obtained from fig. 8. This shows the splitting of the double resonance as the overall position of the output cavity-mirror is adjusted, changing the frequency at which the plus sideband is resonant from 200 Hz to $(200 + f)$ Hz. A relatively broad bandwidth has been chosen for clarity. In this example the output cavity is maintained at the same length, resonant for the original laser frequency. Increasing the upper sideband's resonant frequency effectively pushes it to the right on the output cavity resonance curve (fig. 4), pulling the other resonance also to the right. This reduces the frequency of the other resonance. If the phase shift on reflection θ varied linearly with frequency, the lower resonance would decrease the same amount as the upper increased. Fig. 8 shows that this is a good approximation for small splittings. It can also be seen that the bandwidth of the upper resonance decreases while that of the lower increases as the resonance is split further. This is a consequence of the transmission through the output cavity-mirror being reduced for the upper sideband, and increased for the lower sideband, as the frequencies are shifted.

Another way of splitting the resonance is to alter the tuning of the output cavity ($\delta_3 \neq 0$). This changes the difference in phase shift upon reflection for the two sidebands. A combination of adjusting the lengths of both the output and signal recycling cavities will allow the frequencies of both resonances to be controlled.

It is interesting to think about an extension of a system with two resonances: would it be possible to add on many coupled cavities, producing many resonances which together would give broadband frequency response? Could such a system have much better sensitivity than a conventional broadband detector? It is certainly possible in principle to have n cavities producing n resonances. But these cavities also have losses associated with them: with equal lengths and optical quality, the dissipation is increased by n. Just as the presence of two cavities reduces the maximum signal build up by $\sqrt{2}$ from the ideal value, the presence of n reduces it by $\sqrt{n}$. To cover a frequency span f with a system of resonances of width B requires $\sqrt{n} \approx \sqrt{f/B}$. But this is just the same factor that is gained by reducing the bandwidth. With $\sqrt{f/B}/\sqrt{n} \approx 1$ the broadband sensitivity cannot change substantially and there could be little advantage in having a multiply-coupled system.

3 Discussion

We have seen that the use of a long cavity as the signal-recycling mirror can ensure resonance for both signal sidebands, thereby increasing the signal to noise ratio of the gravitational-wave detector. With a cavity the same length as the arms of the interferometer, such a system may be a factor of $\sqrt{2}$ better than a conventional singly-resonant interferometer. Since the dissipation rate determines the photon-noise limited sensitivity of an interferometer, the improvement factor may be increased to a factor of 2 by lengthening the output cavity, as long as we do not care about the bandwidth. The sensitivity-bandwidth combination in broader band situations, the pulse sensitivity, would still only be improved by $\sqrt{2}$, however. This is a consequence of the nature of the frequency-dependence used to arrange double resonance. It is not clear whether or not the reduction of the signal bandwidth is a fundamental part of the process of resonating both sidebands. It may be that there is some better way of arranging doubly-resonant signal recycling, waiting to be thought of.

A gain in signal to noise ratio of a factor of $\sqrt{2}$ may seem modest. Nevertheless, it allows a given sensitivity level to be reached with a factor of 2 less laser power. Alternatively, the volume of space that can be observed is increased by a factor of $2^{3/2} \approx 2.8$.

Not only does the use of a cavity as the signal recycling mirror allow a significant improvement in photon-noise limited sensitivity, it also has some other advantages. In particular, a non-confocal cavity suppresses the transmission of higher order spatial modes out of the system, conferring much greater tolerance of imperfections in the optics. This itself can allow substantially improved sensitivity [9]. A signal recycling cavity-mirror that is both geometry-dependent and frequency-dependent would be a powerful combination.

An output cavity may also reduce the detector noise resulting from light scattered off the walls of the vacuum pipe [10]. This conclusion will have to be re-assessed, however, if the output cavity is arranged alongside one of the arms of the interferometer.

We have seen that the tuning of even the simplest doubly-resonant system may be adjusted without changing the transmission of the mirrors. Nevertheless, going to a long storage-time cavity and double resonance does lose some of the tuning flexibility of a singly-resonant system, especially one with a short output cavity [5]. This flexibility may be recovered if the mirrors of the output cavity are themselves (short) cavities and are therefore of variable transmission. As we have seen, changing the output mirror transmission T_{3e} will vary the detector sensitivity-bandwidth combination, while changing T_{3m} will adjust the tuning frequency.

The system that we have proposed is not simple. The control of the various optical elements will be complex. However, if we can learn how to do this, the rewards are great. An interferometer using doubly-resonant signal recycling has the potential for truly excellent gravitational-wave sensitivity, giving an even cleaner new window on the universe.

Acknowledgements

BJM would like to gratefully acknowledge support from the Royal Society, the University of Glasgow and the Science and Engineering Research Council, as well as hospitality from Caltech during a visit in June 1989. The work at Caltech was supported in part by NSF grant PHY-8803557.

References

[1] K. S. Thorne, in *300 Years of Gravitation*, edited by S. W. Hawking and W. Israel (Cambridge University Press, Cambridge, 1987).

[2] R. W. P. Drever, in *Gravitational Radiation*, edited by N. Deruelle and T. Piran (North-Holland, Amsterdam, 1983)

[3] A. Krolak, A. Lobo and B. J. Meers, Phys. Rev. D **43**, 2470 (1991)

[4] B. J. Meers and A. Krolak, to be published

[5] B. J. Meers, Phys. Rev. D **38**, 2317 (1988)

[6] J.-Y. Vinet, B. J. Meers, C. N. Man and A. Brillet, Phys. Rev. D **38**, 433 (1988)

[7] B. J. Meers, Phys. Lett. A **142**, 465 (1989)

[8] K. A. Strain and B. J. Meers, Phys. Rev. Lett. **66**, 1391 (1991)

[9] B. J. Meers and K. A. Strain, Phys. Rev. D **43**, 3117 (1991)

[10] K. S. Thorne, Caltech Goldenrod Preprint GRP-200 (1989)

Figure Captions

Figure 1: The essential optics of a laser-interferometric gravitational-wave detector. The arms of the Michelson interferometer would, in practice, contain a multiple-reflection optical delay line or cavity to increase the induced phase change, which is then converted to an observable change in light power by interference at the beamsplitter. The interferometer is arranged to operate on a dark fringe at the output, so most of the power is directed back towards the laser. The partially transmitting mirror M_0 recycles this power by coherently adding it to the incoming power.

Figure 2: An interferometer with dual recycling. With interference at the beamsplitter physically separating the original laser frequency and sidebands induced by the gravitational wave, the light power resonates in the cavity formed by M_0 and the arms of the interferometer, while the sidebands independently resonate in that formed by M_3 and the arms.

Figure 3: An interferometer using a cavity as the signal-recycling mirror. The middle and end mirrors, M_{3m} and M_{3e}, of the output cavity together form a compound mirror M_3. The frequency-dependent reflectivity of this compound mirror allows two signal sidebands to be simultaneously resonant.

Figure 4: (a) the phase shift on reflection θ and (b) the transmission of a typical cavity ($T_{3e}^2 = .002, T_{3m}^2 = .0049$). The frequency offset from resonance is normalised to the cavity corner frequency. Note that θ changes by 2π going from one side of the resonance to the other.

Figure 5: The frequency response with doubly-resonant signal recycling, tuned to 200 Hz, for different transmissions T_{3e} of the end mirror of the output cavity. The signal size is compared to that of the same optical system (a 16-reflection, 3 km delay line) with no signal recycling. Similar behaviour would be obtained with cavities in the arms, with T_{3e} smaller by a factor of $\sqrt{16}$.

Figure 6: The frequency response of a 16-reflection, 3 km delay line with: (a) singly-resonant signal recycling ($T_3^2 = .018$); (b) doubly-resonant signal recycling, $T_{3e}^2 = .01$; (c) doubly-resonant signal recycling, $T_{3e}^2 = .005$. The doubly-resonant system has greater peak response for the same bandwidth, greater bandwidth for the same peak response.

Figure 7: The frequency response of a doubly-resonant interferometer with $T_{3e}^2 = .01$. The system is initially tuned (a) to 200 Hz ($\delta_3 = 0$). Adjusting the length of the output cavity allows resonance at (b) 230 Hz ($\delta_3 = -.016$) or (c) 280 Hz ($\delta_3 = -.032$).

Figure 8: The splitting of the resonance by alteration of the length of the signal recycling cavity (moving M_3) so that the upper sideband is resonant at $(200 + f)$ Hz. $T_{3e}^2 = .01$.

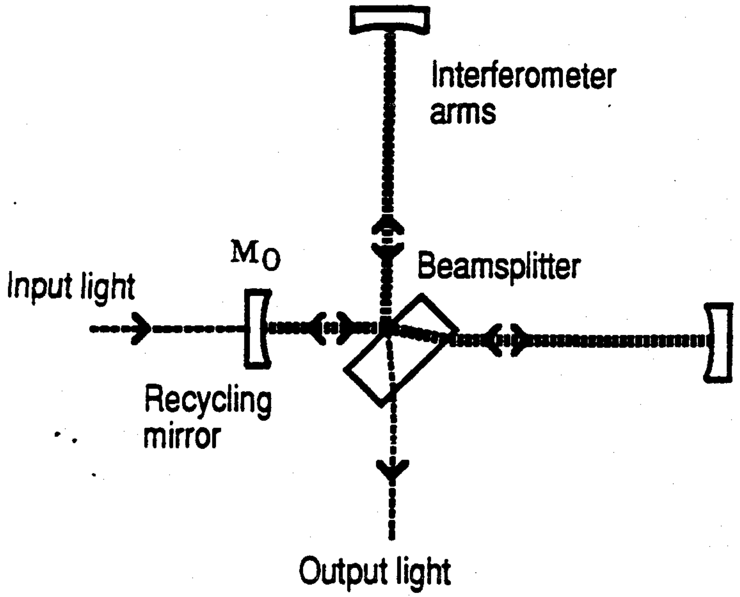

Interferometer arms
M$_O$
Beamsplitter
Input light
Recycling mirror
Output light

Figure 1

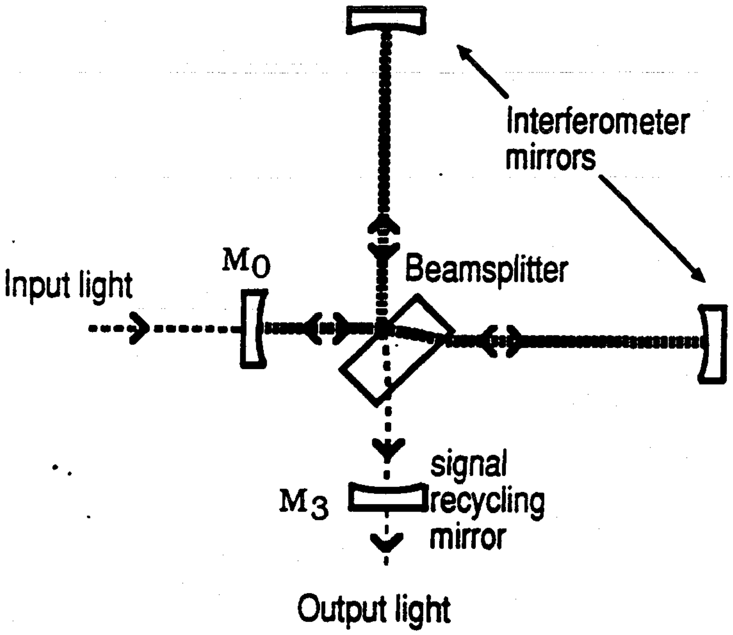

Interferometer mirrors
Input light
M0
Beamsplitter
signal recycling mirror
M3
Output light

Figure 2

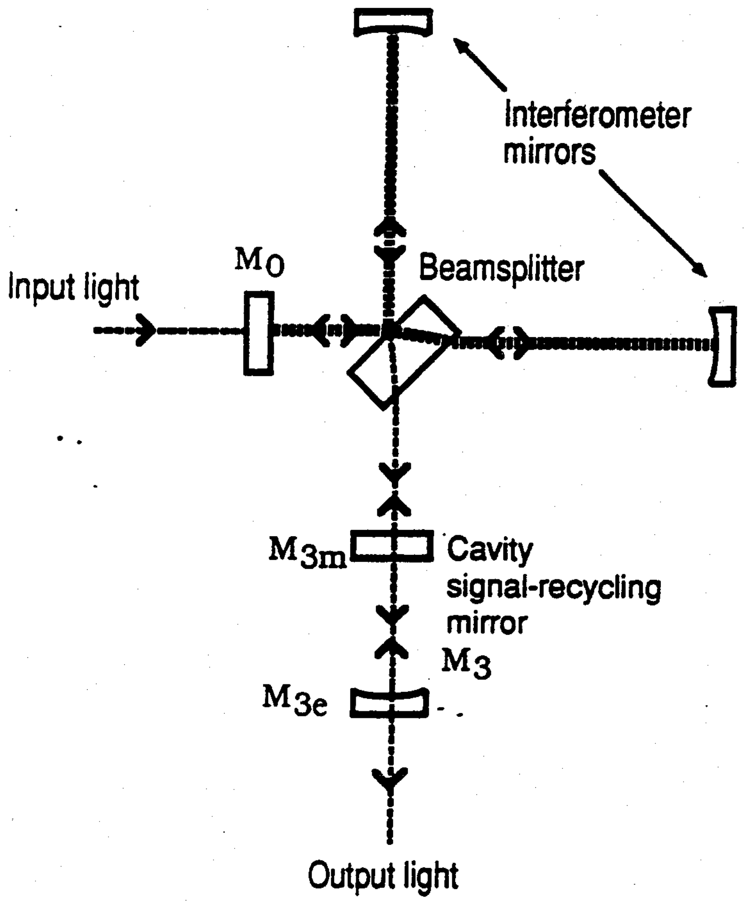

Interferometer mirrors
M0
Beamsplitter
Input light
M3m
Cavity signal-recycling mirror
M3
M3e
Output light

Figure 3

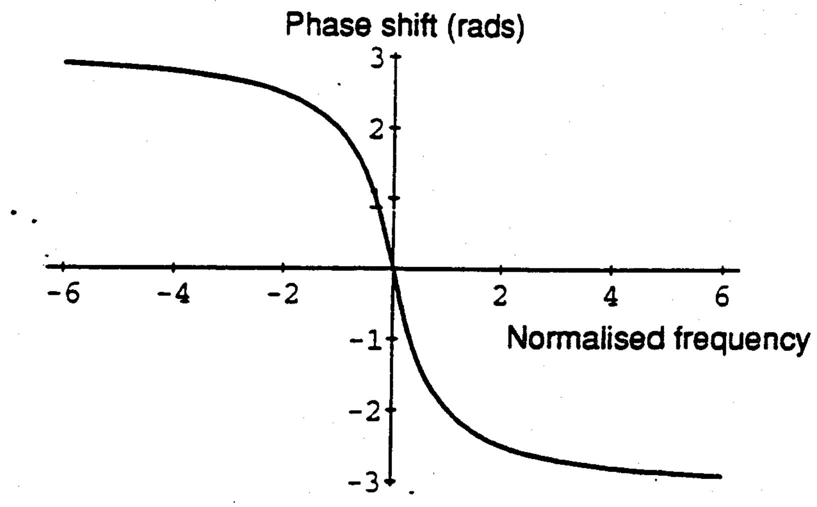


Figure 4(a)

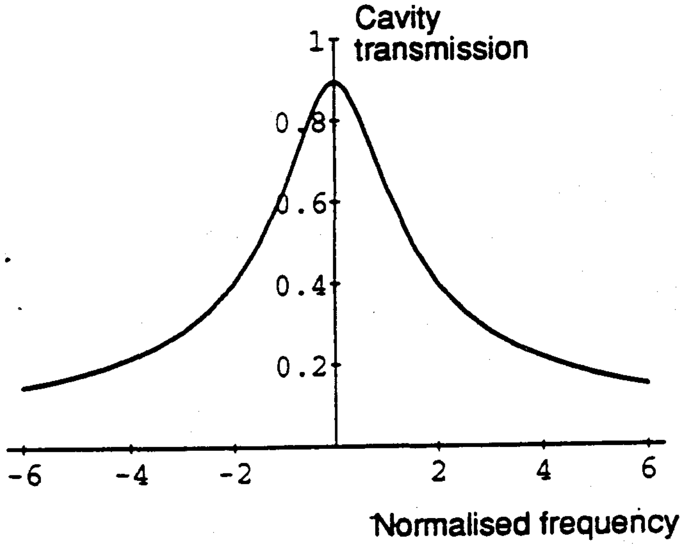

Cavity transmission
1
0.8
0.6
0.4
0.2
-6
-4
-2
2
4
6
Normalised frequency

Figure 4(b)

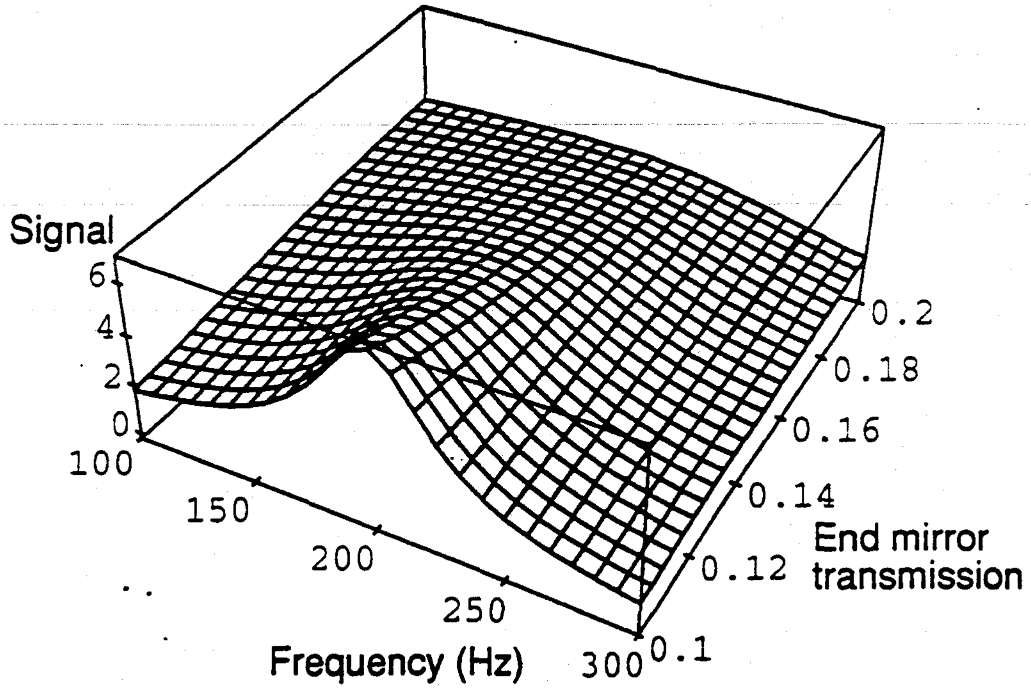

Signal
6
4
2
0
100
150
200
250
300
Frequency (Hz)
0.1
0.12
0.14
0.16
0.18
0.2
End mirror transmission

Figure 5

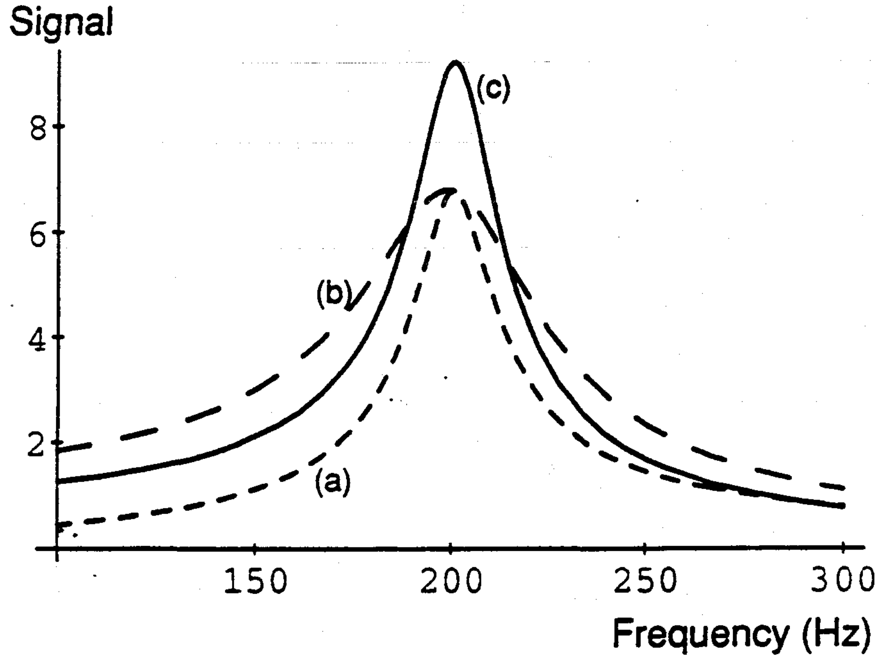


Figure 6

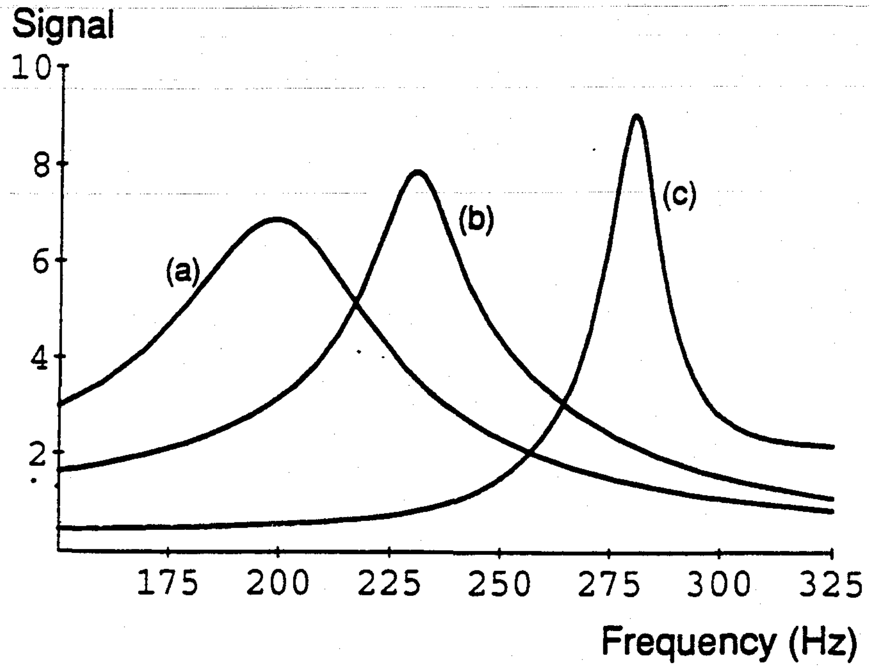


Figure 7

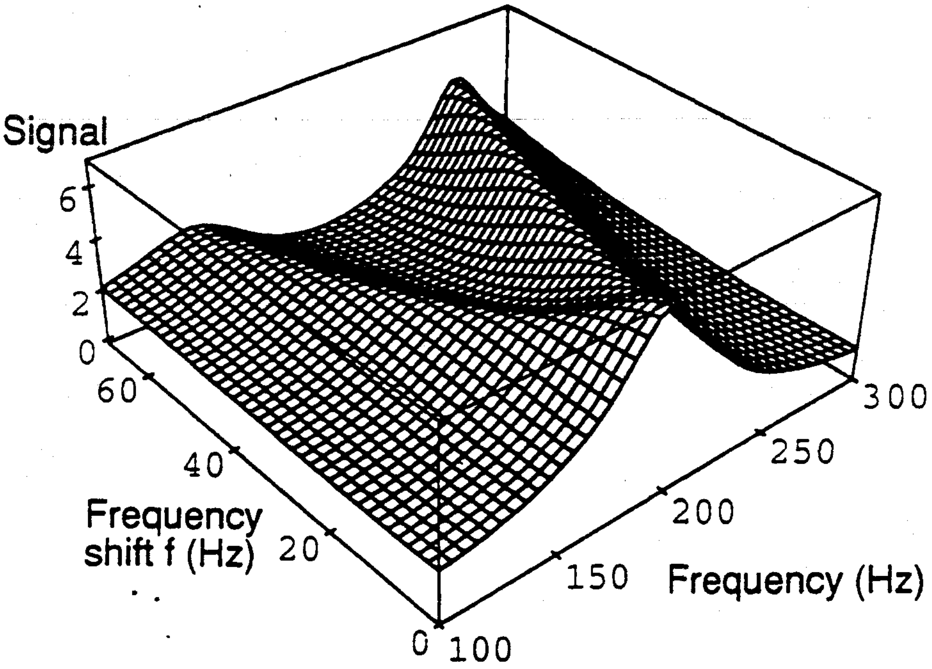

Signal
6
4
2
0
60
40
20
0
Frequency
shift f (Hz)
100
150
200
250
300
Frequency (Hz)

Figure 8

Physics Letters A 175 (1993) 273–276
North-Holland

PHYSICS LETTERS A

Resonant sideband extraction: a new configuration for interferometric gravitational wave detectors

J. Mizuno, K.A. Strain, P.G. Nelson, J.M. Chen, R. Schilling, A. Rüdiger, W. Winkler and K. Danzmann

Max-Planck-Institut für Quantenoptik, Ludwig-Prandtl-Strasse 10, W-8046 Garching near Munich, Germany

Received 10 December 1992; revised manuscript received 27 January 1993; accepted for publication 15 February 1993
Communicated by J.P. Vigier

We introduce a new Fabry–Perot based interferometric gravitational wave detector that, compared with previous designs, greatly decreases the amount of power that must be transmitted through optical substrates to obtain a given light power in its arms. This significantly reduces the effects of wavefront distortions caused by heating due to absorption in the optics, and allows an improved broadband sensitivity to be achieved.

To obtain a good sensitivity in long-baseline interferometric gravitational wave detectors, one requires high light power in the arms of the interferometer to increase the photon-statistic limited signal-to-noise ratio. In the *standard Fabry–Perot configuration*, which consists of a Michelson interferometer having a Fabry–Perot cavity in each arm [1], this can be done by increasing the finesse of the cavities (made possible by the availability of very low-loss mirror coatings). The storage-time for the signal sidebands must, however, be kept short enough to give the desired detection bandwidth (since cavities act like low-pass filters). This determines what is referred to as the *storage-time limit*. A high laser power, or a "power recycling mirror" [1], must then be used to compensate for the limitation that the storage-time puts on the power enhancement in the arm cavities. To obtain a sensitivity and a bandwidth which are desirable in future advanced detectors, one is then required to have extremely high light power incident on the beamsplitter, potentially in excess of 10 kW. Thermally induced lensing in the beamsplitter and mirror substrates will make it exceedingly difficult to reach this power [2,3].

By using the principles of coupled cavities, it is possible to increase the finesse of the arm cavities beyond

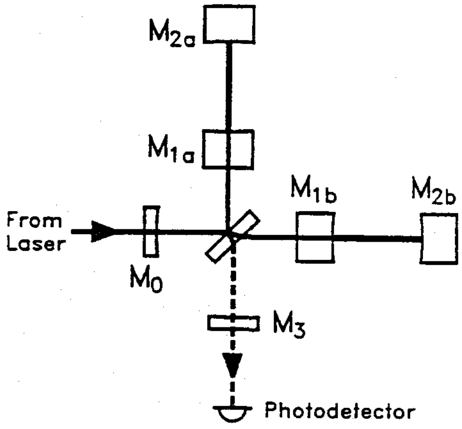


Fig. 1. Schematic diagram of the optical configuration of *resonant sideband extraction*. The storage-time for the carrier light in the arms is longer than the *storage-time limit*, whereas the storage-time for differentially generated sidebands is reduced by the existence of the signal extraction mirror (M_3).

0375-9601/93/$ 06.00 © 1993 Elsevier Science Publishers B.V. All rights reserved.

that given by the storage-time limit, while keeping the storage-time for the signal sidebands consistent with the desired detection bandwidth. This is accomplished by adding a *signal extraction mirror* (M_3 in fig. 1) to the standard configuration, the purpose of which is to *decrease* the storage-time for the signal sidebands. This is not to be confused with a "signal recycling mirror" [4], which was proposed as a means of *increasing* the signal sideband storage-time. The result is that the new system can obtain very high power in its arms with only very modest power being present at the beamsplitter, thus greatly reducing the heating problem. In principle, the power enhancement in the arm cavities could be so great that no power recycling would be required.

A suitably polarized gravitational wave will phase modulate the light in the arms, and will thus produce upper and lower sidebands on the laser light with a frequency spacing from the carrier equal to the gravitational wave frequency. These sidebands are generated with opposite phase in the two arms. In a perfect interferometer operated on a "dark fringe", the beamsplitter separates these sidebands from any carrier light leaving the arms. The carrier light is directed towards the laser (where it may be recycled by a power recycling mirror M_0), and the sideband light towards the signal extraction mirror M_3. The mirror M_3 thus forms a coupled cavity with the cavities in the arms of the interferometer. It can be used to make the apparent reflectivity at the inboard mirrors (M_{1a} and M_{1b}) lower for the signal sidebands than it is for the carrier light. We will refer to the cavity formed by the inboard mirrors and the signal extraction mirror as the signal extraction cavity (SEC).

The frequency response of the interferometer is determined by the properties of the SEC (assuming the length of the arms are fixed by practical considerations). The reflectivity of the SEC on resonance is determined by the reflectivities of the signal extraction mirror and the inboard mirrors, and sets the approximate detection bandwidth and the sensitivity at zero hertz. Both the length (order) of the SEC and its tuning relative to the carrier light affect the shape of response. We classify the responses into *symmetric* and *asymmetric* cases. In the symmetric case, the SEC is chosen to be resonant at the carrier frequency, and the upper and lower signal sidebands generated in the arms see the same reflectivity of the SEC. In the asymmetric case, the SEC is slightly detuned from the carrier frequency, and the sidebands see different reflectivities.

First consider the symmetric case. If the length of the SEC is infinitesimal, its bandwidth is broad, and all signal sidebands see the same reflectivity and experience the same phase shift upon reflection from it. Thus, the frequency response of the interferometer has the same shape as in the standard configuration. The reflectivity of the SEC determines the storage-time for the sidebands in the system, and thus the detection bandwidth. The peak sensitivity is also affected, and increases as the square root of this storage-time. As one increases the length of the SEC (by half wavelength steps, to keep it tuned to the carrier frequency), its bandwidth narrows,

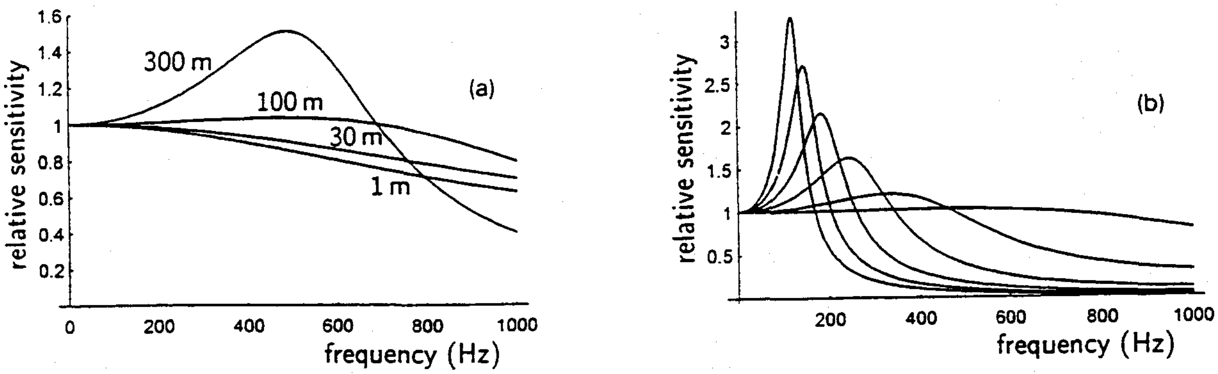


Fig. 2. Typical frequency response curves for the proposed system. (a) Dependence of the frequency response on the length of the SEC in the *symmetric case*. The curves correspond to lengths of 1, 30, 100, and 300 m, respectively. (b) Dependence on the tuning of the SEC to carrier. The broadest response corresponds to the *symmetric case*, and the others are detuned from it (the *asymmetric case*). The length of the SEC is 100 m, and the detuning is by steps of $2\pi/1000$ in δ. In both figures, the fractional power loss in the SEC is assumed to be 5×10^{-3} (dominated by imperfect interference of the beams from the two arms), $t_1^2=5\times10^{-4}$, and $t_3^2=5\times10^{-3}$. The vertical units are arbitrary.

and the phase of the sideband light reflected from it changes. If we choose the SEC to be undercoupled [#1], the frequency dependence of this phase tends to cancel that associated with the transit time in the arms (over a limited frequency range). This has the effect of flattening the response curve near its peak (at zero hertz), at the expense of making the high-frequency cutoff sharper than in the standard configuration (fig. 2a). The bandwidth of the whole system (measured at -3 dB of the peak) increases as the SEC is elongated, but typically by a factor less than two. If the length of the SEC is increased further, then a peaked response is obtained. In most cases, however, this is not practical, and narrow-banding is more readily accomplished in the asymmetric case described below. By an appropriate choice of the length and reflectivity of the SEC, a suitable compromise between the sensitivity, bandwidth and flatness of response can be found. Making the bandwidth of the SEC comparable with the desired detection bandwidth provides the flattest response.

In the asymmetric case, one starts with a suitable symmetric response, and tunes the resonance of the SEC slightly away from the carrier frequency (by moving the signal extraction mirror a small fraction of a wavelength). This produces a peak in the response, as illustrated in fig. 2b. As the detuning continues, the peak moves to lower frequencies, narrows in bandwidth, and increases in height. The bandwidth is roughly proportional to the peak frequency, and the height approximately inversely proportional (for all but the narrowest bandwidths). This dependence on the position of the signal extraction mirror differs from the case of signal recycling, where the peak sensitivity and bandwidth are independent of the mirror's position, and only the frequency of the peak response changes. If both configurations are optimized for the same peak frequency and bandwidth, however, similar frequency responses can be obtained.

To evaluate the frequency response of the system, it is helpful to apply the formalism of Meers [5]. He derives the frequency response for interferometers by considering the cavity in which the signal sidebands are stored. In an interferometer such as ours, this is a three-mirror coupled cavity system, consisting of an end mirror, and a compound output mirror. In this case the end mirror corresponds to M_{2a} or M_{2b}, and the compound mirror to the SEC. As a major difference from those considered in ref. [5], our configuration exploits the behavior of the system when the compound output mirror (one SEC) is nearly resonant with the sideband light (ref. [5] treats only the non-resonant case). The following equation gives the ratio of the signal sideband amplitude appearing at the output of the interferometer to the input carrier amplitude incident on the beamsplitter,

$$|G(\omega_g)| = h\frac{\omega\sin(\frac{1}{2}\omega_g\tau_A)}{\omega_g}\frac{t_1 r_2}{1-r_1 r_2}\left|\frac{t_{SEC}(\omega_g)}{1-r_{SEC}(\omega_g) r_2 \exp(i\tau_A\omega_g)}\right|$$

$$= h\frac{\omega\sin(\frac{1}{2}\omega_g\tau_A)}{\omega_g}\frac{t_1 r_2}{1-r_1 r_2}\frac{t_1 t_3}{|1-r_1 r_2\exp(i\tau_A\omega_g) - r_1 r_3\exp[i(\delta+r_S\omega_g)] + r_2 r_3\exp\{i[\delta+(\tau_A+r_S)\omega_g]\}|}\,.$$

Here, h is the strain amplitude induced by the gravitational wave. The complex amplitude transmittance and reflectivity of the SEC are $t_{SEC}(\omega_g)$ and $r_{SEC}(\omega_g)$. The amplitude transmission and reflection coefficients for the nth mirror (where "a" and "b" mirrors are assumed identical) are t_n and r_n. The sum of their squares can be less than one, depending on the model for losses in the system. The angular frequency of the light and the gravitational wave are given by ω and ω_g, respectively. The round-trip transit time for light in the arms is τ_A and for the SEC is τ_S. The phase offset of the SEC is given by δ (equal to zero in the symmetric case). Figure 2 shows the frequency responses obtained by summing over the upper and lower sidebands, taking their relative phase into account. Responses other than the ones described here are possible. Direct comparison of these curves with those for the standard configuration is highly model dependent, so we do not do so here [6]. Optimization will require a detailed analysis which includes the non-ideal aspects of the interferometer.

To choose the mirror reflectivities and separations, one must consider the losses in the system. The important

[#1] In an undercoupled cavity, the amplitude of the light directly reflected from the input mirror is larger than that leaking out from inside the cavity.

losses are those in the arms, and those in the SEC. The loss in the arms is due to scattering and absorption in the mirror coatings, and it dominates the power loss in the system. It does not directly affect the frequency response. The loss in the SEC is mainly due to imperfect interference of the beams from the two arms. Rayleigh scattering in the optics, and losses at anti-reflection coatings. It limits the efficiency of signal extraction that can be achieved and alters the shape of the frequency response. To optimize the frequency response and sensitivity, the losses in the SEC should be small compared with the transmission of the signal extraction mirror. To achieve this, one may need to increase the transmission of the inboard mirrors M_{1a} and M_{1b}. It should be noted that the resulting reduction in power buildup in the arms can be compensated by a relatively modest power recycling factor (perhaps on the order of ten).

In *any* proposed configuration optimized for a chosen bandwidth, there is a minimum energy (number of photons) that must be stored in the arms to achieve a given sensitivity. Future detectors aiming for a strain sensitivity of 10^{-22} in a 1 kHz bandwidth will need at least 20 J of 1 μm wavelength photons to be stored. With arm lengths of 3 km, this requires an optical power in each arm of 500 kW [#2]. For the standard configuration, the storage-time limit requires the effective number of beams in each arm to be less than 50, and thus the power incident on the beamsplitter must be at least 40 kW. At such power levels, severe thermal distortions can be expected given the currently available substrate materials [2,3]. In the proposed system, numerical models [6] show that finesse can be increased to approximately 10 000, which would require a power incident on the beamsplitter of only 150 W. Assuming 50 ppm loss per round-trip, the reflectivity of the arm cavities should be approximately 0.65, and these powers should be obtainable with only a modest power recycling factor. Of course, the power in the arms remains high, and careful consideration must be paid to the thermal effects due to absorption at the mirror coatings. This, however, is common to all proposed configurations that use Fabry–Perot cavities in the arms.

In conclusion, using this technique of resonant sideband extraction allows Fabry–Perot based interferometers to have long storage-time, very high finesse arm cavities without sacrificing the detector's bandwidth. This makes it possible to have high light power in the arms without requiring high power to be transmitted through any optics. This effectively eliminates one of the dominant distortion problems in the standard configuration, and thus the broadband sensitivity of this configuration should exceed those of previously proposed designs.

The authors would like to thank the gravitational wave group at the University of Glasgow, especially Jim Hough, Euan Morrison, David Robertson, and Harry Ward who all provided helpful insights into the understanding of this system. This work has been supported by the Bundesministerium für Forschung und Technologie (BMFT). One of the authors (JM) would like to thank the JSPS for its support.

[#2] The power actually incident on the mirrors may be less than this if folded Fabry–Perot cavities or delay lines are used.

References

[1] R.W.P. Drever et al., in: Quantum optics, experimental gravitation, and measurement theory, eds. P. Meystre and M.O. Scully (Plenum, New York, 1983) pp. 503–514.
[2] W. Winkler et al., Phys. Rev. A 44 (1991) 7022.
[3] K.A. Strain et al., in preparation.
[4] B.J. Meers, Phys. Rev. D 38 (1988) 2317.
[5] B.J. Meers, Phys. Lett. A 142 (1989) 465.
[6] J. Mizuno et al., to be published.

Interferometric detectors for gravitational radiation.
Drever, R.W.P.. "Gravitational Radiation," NATO Advanced
Physics Institute, Les Houches, (ed. N. Deruelle and
T. Piran), 321-337, (North Holland Publishing, 1983).

INTERFEROMETRIC DETECTORS FOR GRAVITATIONAL RADIATION

R.W.P. Drever

California Institute of Technology, Pasadena
and
University of Glasgow, Glasgow, Scotland

1. INTRODUCTION

Most of the experiments aimed at the detection of gravitational radiation carried out to date have employed resonant bar gravity wave detectors, in which changes in longitudinal vibration of a suspended metal bar, due to the apparent differential action of the gravity wave on the material towards the ends of the bar, are sought. An alternative approach, in which changes in the relative motions of two or more widely separated and nearly free test masses are monitored using laser interferometry, is now being developed in several laboratories. We shall outline here some of the principles and ideas behind these laser interferometer gravitational wave detectors, and also some possibilities for further development of these techniques which seem interesting for future experiments.

An obvious way one might consider detecting gravity waves is through the changes in separation of free test particles, and the idea of using optical interferometers for observing this has certainly occurred to many physicists: indeed one might wonder why so few searches for gravity waves have been made this way. The smallness of the expected effects provides the main answer: for even the relatively strong signal from a supernova in our galaxy would give a relative motion in a pair of test masses 10 meters apart of only 10^{-16} m, and the idea of measuring this in a time of order of a millisecond using light of wavelength nearly 10^{10} times larger than this is not initially attractive. With the resonant bar technique, however, the test masses are linked together by the elasticity of a metal bar chosen to resonate near the frequency of interest, extending the time available for the measurement up to the damping time of the bar, and enabling piezo-electric or capacitative transducers to be used which approach this order of sensitivity. These resonant bar detectors, *pioneered by Joseph Weber, have* been, and are being, successfully developed in many laboratories.

If one looks for much higher sensitivity, however, as may be required for detection of supernova signals from the distance of the Virgo cluster, the very small energy changes to be sensed in a resonant bar detector do impose serious practical difficulties. Thermal noise in the bar, and transducer sensitivity,are both severe problems, even in a low temperature system. It used to be thought that quantum limits would set a fundamental barrier, but there now appear to be ways of avoiding these in principle and the thermal noise seems to be the dominant problem. In this region of gravity wave sensitivity, there are of course severe problems with free mass detectors too; but with these, there is the important advantage that the displacements to be observed may be increased considerably by increasing the distance between the masses - in principle till the distance becomes comparable to half the wavelength of the gravity wave, typically many tens of kilometers. Thermal noise is made even less important by the fact that there is no need for any material connection between the test masses to resonate near the frequencies of interest. A free mass detector also looks likely to operate over a wider range of frequencies than resonant bar instruments. These considerations have made it seem worth while to investigate and develop possibilities for free-mass gravity wave detectors with optical displacement sensing.

2. SENSITIVITY DESIRABLE

The wide spectrum of gravitational radiation signals expected from astronomical phenomena of various types presents a considerable range of possible targets for experimental searches, if adequate sensitivity can be achieved. Much effort has gone into estimating radiation fluxes, and recent work in this area is reviewed elsewhere in this volume. Several summaries of spectra of gravitational radiation anticipated at the earth have also been published.[1] A part of the spectrum relatively accessible to earth-bound experiments is that in the region between a few tens of hertz and a few kilohertz and here signals from stellar collapse in our galaxy may give amplitudes of order 10^{-17}, but probably occurring at a rate less than one per year. For a pulse rate of order one per month a sensitivity of order 10^{-21} or better in gravity wave amplitude may be required. Indeed at sensitivities around this several types of sources may become detectable and this might be an interesting target to consider for future pulse experiments-at least in the long term. Other types of signals are also possible targets, such as the continuous gravitational radiation from pulsars, or a stochastic background from the big bang or from stellar collapse processes occurring at an early epoch. In these cases the gravity wave amplitudes expected are even smaller than from impulsive events, but the continuous nature of the signal may in principle enable higher sensitivity to be achieved with an appropriate detector and suitable mode of operation, as we will consider later.

3. BASIC ARRANGEMENT

To attempt to measure by optical means a change in distance between a pair of masses of order one part in 10^{21} would require an exceptionally stable wavelength standard; and there are obvious advantages in making a differential measurement of two almost equal baselines perpendicular to one another, which may be oppositely affected by a gravitational wave of suitable polarisation and direction of propagation. In principle this might be done using a Michelson interferometer to compare distances between mirrors attached to three test masses suspended like pendulums, as schematically indicated in Figure 1. Monochromatic light is not essential for such a measurement, but a laser is a convenient source because of the intensity and directional properties of its output. Let us assume that the gravitational waves of interest have a period short compared with the period of the pendulum suspensions, so that the test masses may be regarded as effectively free for small horizontal motions, and consider some of the more basic noise sources expected in a measurement of a short gravitational wave pulse.

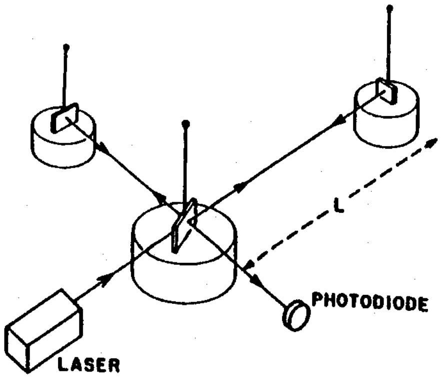


Figure 1

3.1 Quantum Limit.

One limit to the sensitivity of a pulse measurement would be set by the quantum uncertainty in the position of each of the free masses. For detection of change in position over a time τ, the duration of the gravity wave pulse, the smallest displacement detectable is approximately $(2 \hbar \tau/ m)^{1/2}$, where m is the mass and $2\pi \hbar$ is Planck's Constant. If the baseline L between the masses is 40 m, the pulse duration τ is one millisecond, and test mass $m = 100$ kg, this quantum uncertainty would set a limit to gravity wave amplitude, h, detectable at unity signal to noise ratio, of order $h = 10^{-21}$. Thus for a 40 m detector, the quantum limit is of the same order as our target sensitivity for short pulses, and it could be reduced further by increasing the baseline. In fact the quantum limit is not likely to be a serious difficulty in searches for short pulses with this type of detector, although it may become important in measurements of long pulses or continuous signals. In practice, photon counting error is more likely to be a problem.

3.2 Photon Counting Error.

As the motions expected are small compared with the wavelength of the light, only a small change in output light intensity can be anticipated, and in a simple system this must be detected in the presence of intensity fluctuations due to photon counting statistics, at the least.[2] In the arrangement in Figure 1, a single photodetector is indicated receiving light from one side of the beamsplitter. In this situation, it may be shown that the displacement sensitivity set by photon counting error depends on the initial phase difference between the two components making up the output light, and optimum photon-shot-noise-limited sensitivity is obtained when the phase difference tends to π, that is near a dark fringe in the output light. If it is assumed that the photodiode has unity quantum efficiency, then the corresponding limit to the amplitude of gravitational wave detectable is $h = (\lambda \hbar c/ 8\pi L^2 I \tau)^{1/2}$, where I is the laser power, λ is the wavelength of the light, and c is the velocity of light. An alternative mode of operation would be to use two photodiodes, one detecting light from each side of the beamsplitter; and in this case the same overall sensitivity may be obtained anywhere in the fringe pattern. In either case, if we take a laser power of 1 watt at a wavelength of 500 nm, a baseline L = 40 m and measuring time τ of one millisecond as before, the gravity wave amplitude giving unity signal to noise ratio against the photon counting error is $h = 2 \times 10^{-17}$. This is far from our target sensitivity; and if we were to attempt to approach the quantum limit by merely increasing the laser power in this simple configuration we would require many megawatts of light.

It may be noted that some pioneering experiments with this type of gravity wave detector have been carried out using a configuration essentially similar to this by Moss and Forward[3], who showed that performance near the photon noise limit for a low power laser could be achieved.

An ingenious proposal for reducing photon counting error was recently made by C. M. Caves[4], who suggested use of squeezed photon states. By altering the distribution of vacuum fluctuations between two orthogonal phases, the photon counting fluctuations may be decreased at the expense of increased but less significant fluctuations in differential radiation pressure on the test masses. Practical application looks difficult at present, due partly to losses in the non-linear optical elements required; but the idea is in principle an extremely interesting one.

An important practical method for improving photon-noise limited sensitivity was suggested by R. Weiss[5]: the use of optical delay lines to reflect the light beam many times between the test masses, and thus increase the optical phase shift resulting from relative motion. Experimental work on this type of multiple-reflection Michelson interferometer has been, or is being, carried out at several laboratories including M.I.T., the Max Planck Institute at Munich[6], and the University of Glasgow.[7]

4. MULTIREFLECTION MICHELSON INTERFEROMETERS

A simplified schematic arrangement for a multireflection Michelson interferometer gravity wave detector is shown in Figure 2. Here the light is made to traverse the distance between the test masses many times by a suitable optical system, such as a Herriott delay line, before it recombines at the beamsplitter. This increases the total phase shift experienced by the light for a given movement of the masses by the number of round trips in each arm: and by using a suitable concave mirror configuration to minimize diffraction losses this number can be made large. In practice the useful number of reflections may be limited by one of two factors: the reflection losses at the mirror, or the total light travel time within each arm of the interferometer. In the first case, with mirror losses important, it may be shown that optimum photon-noise-limited sensitivity is obtained with a number of reflections in each arm equal to $2/(1-R)$, where R is the mirror reflectivity. The sensitivity achieved is then better than that of a simple Michelson interferometer by a factor of $e(1-R)$, where $e= 2.72...$In a system with baseline $L = 40m$, mirror reflectivity $R = 0.997$, and other parameters as before, this would give a gravity wave sensitivity of order $h = 10^{-19}$.

If, however, the baseline were large enough, and the mirror reflectivity high enough to cause light making $2/(1-R)$ reflections to spend a time within the system longer than the time scale in which the gravity wave reverses its sign, then some cancellation of signal could occur. In this case it would be nearer optimum to choose the number of reflections to make the light spend a time in each arm equal to the time scale of the gravity wave. The sensitivity then becomes independent of arm length; and for a storage time of one millisecond and a laser power of 1 W, this could correspond to $h= 2 \times 10^{-21}$ for either a 40 m armlength system with 8000 reflections or a 4 km armlength system with 80 reflections.

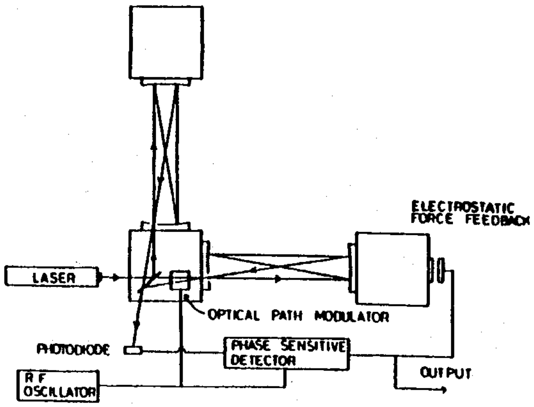


Figure 2

These examples are idealised, of course, but they do suggest that interesting sensitivities might be achieved with this type of gravity wave detector if the many practical problems could be overcome. Also, there are some new ideas for improving the photon-noise-limited sensitivity even further, as we shall discuss later.

One practical difficulty in the optical sensing system just described became apparent in early experiments at Munich and at Glasgow - the potentially serious effect of incoherent scattering of light at the multireflection mirrors or elsewhere in the system. If scattered light reaches the photodetector having traversed a path different from that of the main beam, it will differ in phase from it by an amount dependent both on the path difference and the instantaneous wavelength of the light. The path difference involved can be very long - comparable to the total travel distance through the system - so small fluctuations in wavelength may give relatively large phase fluctuations in the output, particularly as the phase fluctuation in the resultant beam is proportional to the relative amplitude of the scattered light and not its relative intensity. The effect may be reduced by precise stabilisation of the wavelength of the laser, and also by arranging that the spots on the multireflection mirrors where successively reflections take place do not overlap, but it may still be important in a large system. The Munich group suggested[8] that the effect might be reduced further by modulating the wavelength of the laser light through an amount chosen to make the phase difference of the major components of the scattered light average to zero over the integration time of the measurement. Another approach would be to make the path traveled by scattered light equal to that of the main beam, and this may in fact be achieved if another type of optical system, a Fabry-Perot cavity, is used instead of a Michelson interferometer with many discrete reflections in each arm.

5. OPTICAL CAVITY INTERFEROMETERS

5.1 Principle.

The idea of using changes in the resonant frequency of an optical (or microwave) cavity to detect small motions is an old one, but one practical method of using optical cavities in gravity wave detectors was outlined only relatively recently.[9] The principle is indicated in Figure 3. Light from a laser passes through a beamsplitter to a pair of Fabry-Perot cavities formed between mirrors attached to the three test masses. If the lengths of the two cavities are adjusted to give resonance with the light from the laser, then differential changes in length may be sensed by changes in the resonance conditions and small changes near resonance may be detected by measuring phase changes between light within each cavity and the input beam, or directly between one cavity and another. The phase difference might be detected by interference between light emerging from each of the cavities back through its input mirror, possibly by using the high frequency phase modulation technique shown in Fig. 2 in connection with the multireflection Michelson interferometer. One way of carrying this out is indicated schematically in Figure 4. Here light from the cavities is phase modulated by two Pockels cells, P1 and P2, driven in antiphase at a suitable radiofrequency, and is detected by photodiode D1. The output from this photodiode, when synchronously demodulated, can give a measure of the phase difference between the light in the cavities. An optical isolator I is used to prevent reflected light affecting the operation of the laser. The additional photodetectors D2 and D3

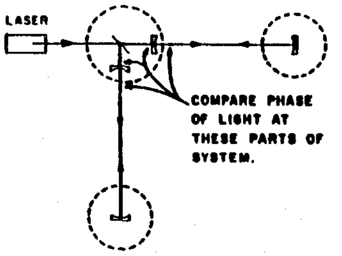


Figure 3

are for auxiliary functions which will be described later.

The maximum sensitivity in this arrangement is obtained when the second mirror in each cavity has the highest reflectivity available, R, say, and the reflectivity of the input mirror for each cavity is chosen either to give optimum photon-noise-limited displacement measurement or to give a light storage time equal to the time scale of the gravity wave, whichever is appropriate. In the former case it may be shown that the gravity wave sensitivity is given approximately by $h = (1-R)\ (\lambda\ \hbar\ c/\ 8\pi\ L^2\ I\ \tau)^{\frac{1}{2}}$, where it is assumed that absorption and scattering losses in the transmission of light by the input mirrors of the cavities are negligible (that is, transmission coefficient = 1-R). This expected

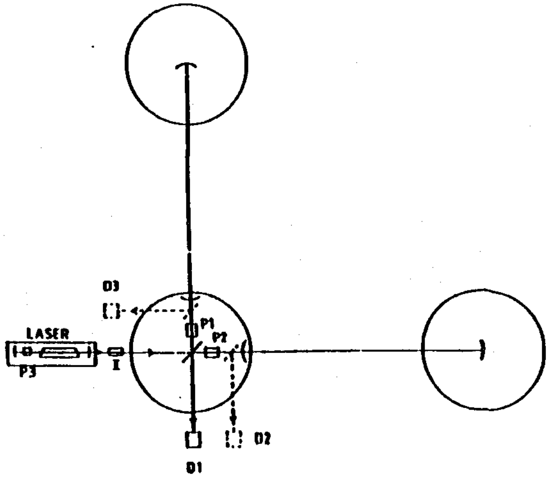


Figure 4

sensitivity is better by a factor of about 2 than the sensitivity obtainable by a multireflection Michelson interferometer using mirrors of the same reflectivity, but transmission losses may degrade this so that in practice the photon noise limit to sensitivity of the two types of optical system is likely to be roughly equal.

There are several advantages of this type of cavity interferometer over the delay-line Michelson system, apart from the possibility of reduced phase noise from scattered light. The diameter of the cavity mirrors can be considerably smaller than that of delay-line mirrors; and for example, even with cavities 10 km long, a mirror diameter of 18 cm is sufficient to make diffraction losses less than 1 part in 10^5 for light of wavelength 500 nm. This reduces the diameter of vacuum pipe required, and also may make it easier to keep mechanical resonances in the mirrors and their mountings high compared with the frequency of the gravity waves, thus minimizing thermal noise. The Fabry-Perot system has, however, some obvious disadvantages too- particularly the requirement for very precise control of the wavelength of the laser and of the lengths of the cavities. Indeed with long cavities of the high finesse desirable here exceptional short-term wavelength stability is required from the laser. A special laser stabilisation technique has been developed to provide this.

5.2 Laser wavelength stabilisation.

The principle of the laser wavelength control system being used is shown in Figure 5. Plane polarised light from the laser is phase modulated by passage through a Pockels cell, at a frequency in the range of 10 to 40 MHz, and then enters one of the Fabry-Perot cavities through a polarising beamsplitter and a quarter-wave plate. The axes of the quarter-wave plate are oriented at $\pm 45^{\circ}$ to the polarisation of the input light, so that circularly polarised light enters the cavity. Light coming back from the input mirror of the cavity is circularly polarised in the opposite sense, is transformed into plane polarised light with polarisation orthogonal to that of the input beam, and is reflected by the polarising beamsplitter to the photodetector. The light arriving at the photodiode can be considered to have two components: the

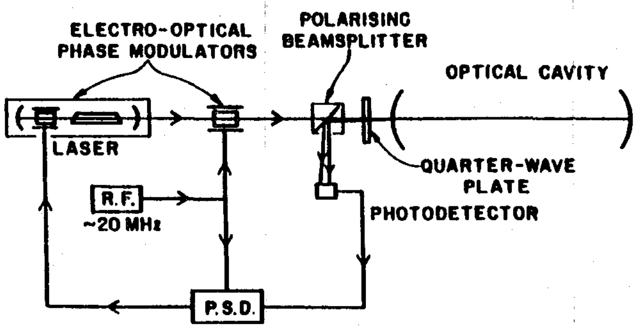


Figure 5

phase-modulated laser light directly reflected by the input cavity mirror, and light emerging from within the cavity - which has built up over the cavity storage time and thus has had its modulation sidebands removed. In the figure, these components are drawn as diverging rays to make the operation clearer, although they are of course coincident in reality. If the laser light is precisely in resonance with the cavity these two components have opposite average phase, and the photodiode output has no component at the modulation frequency. If the laser is slightly off resonance, the photodiode gives a signal at the modulation frequency whose amplitude and phase indicates the magnitude and sign of the error. Demodulation of the photodiode signal by a phase sensitive detector (P.S.D.) gives a voltage signal which may be applied to a second Pockels cell within the laser cavity itself, so that the wavelength of the light from the laser is driven closer to the cavity resonance, and the laser becomes locked in wavelength to the cavity. To achieve a high degree of stabilisation at the gravity-wave frequency it is important that the control system has a wide bandwidth, and a useful feature of the arrangement is that the rise time of the phase error signal is not affected by the fact that the cavity may have a very long storage time.

Early experimental work on this laser-cavity stabilisation technique has been done at the Joint Institute for Laboratory Astrophysics, Boulder, using dye and helium-neon lasers* and at Glasgow, and subsequently Caltech, with argon ion lasers[10,11], and has shown that adequate stabilisation for at least the current stage of development of the Fabry-Perot interferometers can be achieved.

A considerable amount of experimental work relating to gravity wave detectors using Fabry-Perot interferometers has been carried out at Glasgow and at Caltech, much of the earlier work being done with a slightly different arrangement, shown in simplified form in Figure 6. Here triangular cavities are used instead of 2-mirror cavities so that optical feedback to the laser is avoided without use of isolators, and separate detection of signals from the two cavities is used to simplify the optical system. The laser is shown arranged to be locked in wavelength to the right-hand cavity (arm 1) in the figure by the method described above; and a similar phase measuring system is then used to give a fine adjustment to the length of the lower cavity (arm 2) by a piezoelectric transducer (PZT) on which one of the cavity mirrors is mounted, so that this cavity becomes locked into resonance with the wavelength of the light. With this arrangement, a signal corresponding to a gravity wave disturbance may be obtained, approximately, from the feedback voltage applied to the piezoelectric transducer, or, more precisely, from a suitable combination of this signal with the phase error signals ϕ_1 and ϕ_2 from the two feedback loops.

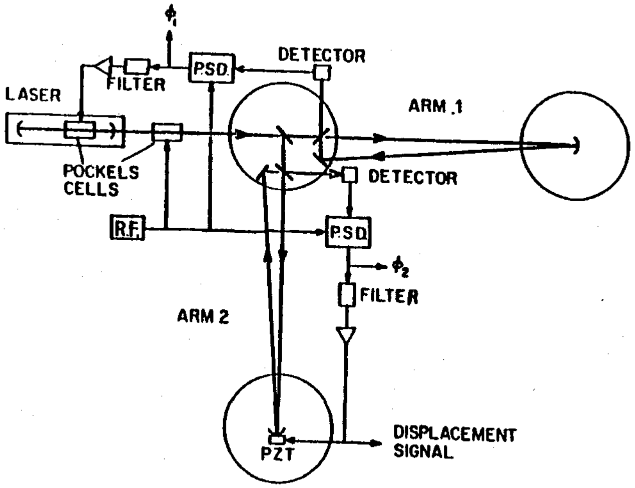


Figure 6

It may be noted that separate detection of resonance in the two cavities, as indicated here, does in principle degrade the photon-noise-limited sensitivity by a factor of 2, but it is convenient for initial experiments since it avoids need for Pockels cells on any of the test masses, or matching of cavity finesse. Optical systems more like that shown in Figure 4 are now being developed. Here the auxiliary photodiodes D2 and D3 shown dotted are intended for laser stabilisation and to facilitate monitoring of the two cavities, and would be arranged to use only a small sample of the total available light.

The experimental development of laser interferometer gravity wave detectors based on either the multireflection Michelson or the optical cavity system outlined has led to overall arrangements which are in practice considerably more complex than suggested by the diagrams here, for additional feedback systems have to be incorporated to control orientation and position of the test masses in the presence of seismic disturbances, and fluctuations in direction and position of the laser beam have to be reduced by passive or active optical systems. Although many problems remain to be overcome, the experimental work has gone far enough to make it seem likely that optical sensing performance close to that indicated by the simple theoretical estimates given above may indeed be achievable by either of the techniques we have outlined. It may be useful to consider some other basic noise sources at this point, and in particular thermal noise - for this is a limiting factor in current resonant bar gravity wave detectors, and is not negligible here.

6. THERMAL NOISE IN LASER INTERFEROMETERS

The real or apparent fluctuations in motion of test masses have to be carefully considered in any gravity wave experiment, for the displacements to be observed are usually small compared with the mean amplitudes of thermal motion. In the type of nearly free mass detector discussed here the mechanical thermal noise may be conveniently divided into two parts - that associated with the low frequency pendulum mode of oscillation of the test masses, and that associated with internal degrees of freedom of the test mass and mirror structure itself. Relevant analyses of thermal noise fluctuations have been given by Weiss[5], Braginsky and Manukin[12], and others; we will just summarise some results here.

6.1 Thermal noise - pendulum mode.

With a simple pendulum suspension of convenient length, the resonant frequency for the pendulum mode of a test mass is of order 1 Hz or lower, well below the frequency of interest for initial gravity wave searches. The power spectral density of displacement is given approximately by $(\delta x)^2/\delta f = 4\ kT\ \omega_0/m\ Q\ \omega^4$, where ω_0 is the angular frequency of the pendulum resonance, Q the quality factor of the resonance, m the test mass, ω the angular frequency of interest, k = Boltzman's Constant, and T the temperature. Some early tests at Glasgow suggest that construction of a simple pendulum with Q near 10^6 is quite practicable;and if we take this value for Q, a mass m=10 kg, ω_0=2π, ω=2π.1000,and T=300 we find that in a system with a 40 m baseline, thermal noise would set a limit to sensitivity of the order of $h = 3 \times 10^{-22}$ in a bandwidth of 1 kHz. This component of thermal noise is therefore not expected to be very serious at these frequencies, if an effective high Q can be maintained in a practical suspension system;and a longer baseline will reduce the noise further. However thermal noise may well become important at lower frequencies.

6.2 Thermal noise - internal modes.

Internal vibrations of the test mass structure can be very complex and there may be many modes near the frequency of interest when the test mass incorporates several mirrors and other components. To minimize the thermal noise in the frequency region of interest, it is desirable to keep the resonant frequency of all modes as high as possible, and certainly high compared with the gravity wave frequency - which may not be easy. In this case, for a single mode of angular resonant frequency ω_o, the power spectrum of displacement is given approximately by $(\delta x)^2/\delta f = 4kT/m\,Q\omega_o^3$. If we take as example $m = 10$ kg, $T = 300$, $\omega_o = 2\pi.5000$ and $Q \sim 10^6$, we find that this sets a limit of about $h = 6 \times 10^{-21}$ in a bandwidth of 1 kHz with a system of baseline 40 m. These values for Q and ω_o are however not easy to achieve in a complex structure. Increase of baseline makes this component of thermal noise less significant, but it is evident that careful design of the test mass structure is required.

It may be noted that the Fabry-Perot cavity type of interferometer may have a disadvantage here in that it is likely to require more precise mirror adjustment than a Michelson system, and thus lead to a more complex structure for at least one of the test masses. One arrangement which we suggest may ameliorate this problem involves use of two separate and very simple test masses at the junction of the two baselines, each containing merely a single cavity mirror,with a separate and more complex suspended structure incorporating the rest of the optical components. In this way, the thermal noise may be minimized in the parts where it is most significant, although the system as a whole does become more complex. At the present stage these problems have not been fully investigated, although considerable advances in reducing thermal noise in a Michelson interferometer system have been made by the Munich group. At present, it appears that to keep internal thermal noise sufficiently small does require careful design of the test masses, but we feel that the problems involved are by no means insoluble ones.

Some notes on the question of seismic isolation of an interferometer gravity wave detector may be appropriate at this point.

7. SEISMIC ISOLATION

Isolation from seismic disturbance is an important practical problem for any type of gravitational wave detector. In the region of the spectrum around 1 kHz, however, it has been tackled very successfully in work with resonant bar gravity wave detectors. At these frequencies good vibration attenuation can be obtained by simple stacks of lead or steel masses alternating with layers of rubber, of the general type developed and widely used since the initial experiments of Joseph Weber. These same methods are applicable for interferometer detectors, and indeed in some ways the problems are simpler than for resonant bar detectors of the same sensitivity, for the displacements to be observed are larger with the laser detectors due to the much longer baselines involved. The seismic motions at the ends of a long baseline are of course less correlated than the motions at the two ends of a resonant bar, but the isolation of the simple pendulum suspension of a single test mass is sufficient on its own to give good attentuation at 1 kHz. Overall, it seems that seismic isolation is unlikely to be a very serious problem for gravity wave frequencies near 1 kHz, although it becomes rapidly more difficult at lower frequencies. It may be noted that low frequency motions of the suspended test masses can give dynamic range problems in optical interferometer systems, and active feedback systems are necessary to damp and restrain the low frequency movements of the masses. The masses may be controlled by applying magnetic or electrostatic forces, or by mechanical motion of the points from which the suspension wires are supported,and all of these methods have proved satisfactory to some degree. The problems involved are technically quite challenging ones, and the solutions are interesting, but it is not appropriate to discuss these in detail here in this article which relates more to basic limitations to the interferometer techniques.

We have now discussed many aspects of laser interferometer gravity wave detectors and have indicated how there may be real possibilities for achieving gravity wave amplitude sensitivities of the order of 10^{-21} for 1 millisecond pulses, with large scale instruments of this type. The most serious limitation to sensitivity in this part of the spectrum looks likely to come from photon counting noise, and although this may possibly be reduced by increases in laser power, or use of multiple lasers, there would seem to be practical limits to these solutions. It may be useful to briefly discuss here some relatively new ideas which suggest alternative ways of improving sensitivity, although it should be emphasized that these suggestions relate more to future possibilities than to the current stage in the experimental development of the techniques.

8. POSSIBILITIES FOR FUTURE ENHANCEMENT IN SENSITIVITY

8.1 Possibility for more efficient use of the light.

It has been mentioned in Section 3.2 that in a Michelson interferometer using a single photodetector maximum sensitivity is obtained when the detector is near a dark fringe; and if the system is efficient and adjusted so that one fringe extends over the whole width of the output beam this implies that most of the light leaves the interferometer through the other side of the beamsplitter. It has occurred to us that this light may be fed back into the interferometer by making it add coherently to the initial laser beam by means of an additional mirror of suitably chosen reflectivity, as indicated in Figure 7. Accurate adjustment of path lengths or of laser wavelength would, of course, be necessary to insure that maximum enhancement of light is achieved, and one way of doing this might be with the phase modulation laser-cavity locking system described in Section 5, using a phase modulating Pockels cell P3 and additional photodetector D2, with the system arranged to minimise the light intensity at D2. The whole optical system then

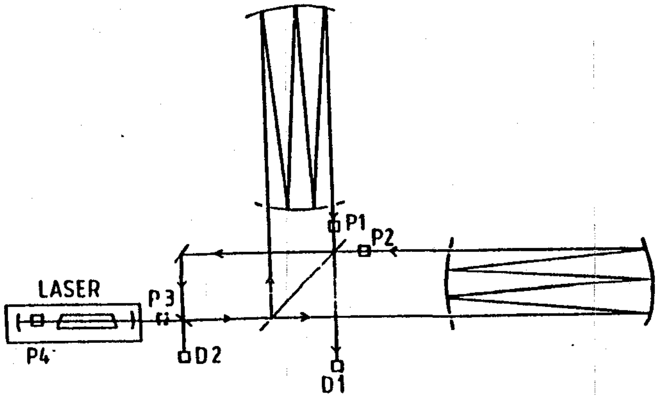


Figure 7

functions like a large Fabry-Perot cavity, and if losses are small and the input mirror reflectivity is suitably chosen there can be considerable enhancement of internal light flux. This arrangement is only useful if the combination of arm length and reflectivity of the delay line mirrors is such that the maximum achievable storage time of the light within each delay line is longer than the time-scale of the gravity waves of interest. The number of reflections in each arm would then be chosen to give a storage time which matches the gravity wave time-scale, and the light intensity within the whole system can then build up over a time approaching the maximum storage time permitted by the losses in the mirrors and other components. If the system is such that the dominant losses are those associated with delay line mirrors of reflectivity R, and the reflectivity of the feedback mirror is chosen for maximum light buildup, then the sensitivity is given approximately by $h = (\lambda \hbar (1-R)/2 \pi L I \tau^2)^{\frac{1}{2}}$, where I= output power of laser. If one considers a large system, with baseline L = 10km, and $(1-R) = 10^{-4}$, then the sensitivity would be of order 10^{-22} for 1 millisecond gravity wave pulses, with a laser power of 10 watts. These parameters are not impossible ones for future experiments.

The same method may be applied to optical cavity interferometers also, as shown, for example in Figure 8. Again, the system is only useful if achievable storage times exceed the time-scale of the gravity waves of interest. The reflectivity of the input mirror of each cavity is chosen to give a storage time within the cavity which matches the time-scale of the gravity wave, which under these conditions would lead to reflection of a large fraction of the light incident on each cavity input mirror back towards the laser. An additional mirror is added in front of the laser to return most of this light back to the interferometer, with phase adjusted to enhance the input light from the laser. If the reflectivity of the input mirror is suitably chosen for maximum light buildup, then the photon-noise-limited sensitivity of this system becomes essentially equal to that of the Michelson interferometer system just described. Precise adjustment of optical paths as well as laser wavelength is required to achieve correct phasing within this system, and auxiliary photodiodes D2, D3 and D4 along with phase modulators P1, P2 and P3 are indicated as means of achieving this. As the internal phase adjustment requires only a narrow bandwidth the photodiodes D2 and D3 need only remove a very small fraction of the light circulating within the system.

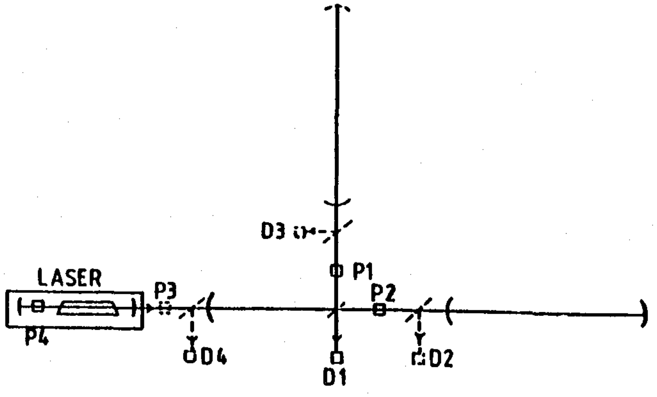


Figure 8

With these proposed techniques for re-use of light within an interferometer,[13] the optical system as a whole may be regarded as a large cavity which stores up light to an extent limited in principle only by the losses in the components. When a gravity wave pulse arrives, the resultant phase changes allow a part of this stored energy to pass out quickly to the output photodiode. The system may thus be quite energy-efficient, and it looks a promising one for future experiments.

8.2 A possibility for enhancing sensitivity for periodic signals.

Our discussion so far has concerned principally the detection of short gravity wave pulses, but it is evident that the same kind of apparatus could be used for searching for periodic gravity wave signals, such as those expected from pulsars, from rapidly rotating neutron star binaries, or possibly from vibrations of neutron stars or other objects. By use of appropriate data processing and integration over many periods of the gravity wave it is clear that better amplitude sensitivity may be obtained with periodic signals than with single pulses. Consideration of expected signal strengths from known sources, such as the Crab or Vela pulsar, does however suggest that it would be useful to have a sensitivity higher than obtainable in this way. We propose now a possible method for further enhancing the sensitivity of a laser interferometer detector for periodic signals. This technique, like the ones described in the previous section, depends on use of an optical system capable of giving very long light storage times - the condition in this case being that the combination of baseline length and mirror losses should enable light to be kept in the system for times long compared with the period of the expected signal.

The idea is most easily explained for a multireflection Michelson interferometer which in this application might have its optical system re-arranged as shown in Figure 9. Light from the laser enters the system through a beamsplitter, M1,

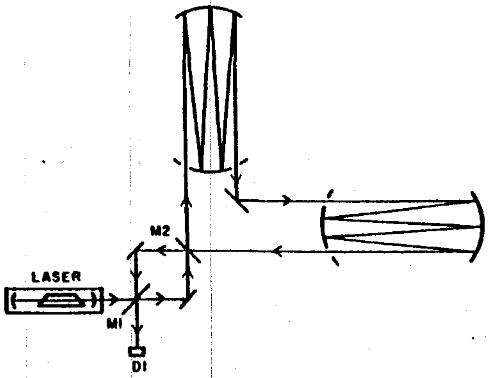


Figure 9

which divides it into two equal parts which pass through a mirror, M2, of suitably chosen high reflectivity and then traverse the delay lines in opposite directions. It is arranged that each delay line introduces a delay equal to half of the gravity wave period. Light travelling in the direction of the arrows which enters the upper delay line at a time when the gravity-wave-induced displacement of the test masses is changing its sign will have its phase shifted in one direction while it is within this delay line. It then leaves this delay line and enters the right-hand one just as the gravity-wave displacement is reversing, so that this light experiences a further shift of phase in the same direction in the second delay line. Most of the light then retraverses the first delay line where further phase shift takes place. Light passing through the delay lines in the opposite direction experiences a buildup of the opposite phase shift, and the phase differences generated over the total storage time of the system may eventually be detected at photodiode D1, possibly using a radiofrequency phase modulation system (omitted from the diagram for simplicity).

An optical cavity gravity wave detector can also be arranged in a similar way to have enhanced sensitivity for periodic signals, as indicated in Figure 10. Here the storage time of each cavity is made to equal half of the period of the expected signal by suitably choosing the reflectivity of mirrors M3 and M3';and by use of polarising beamsplitters (labelled POL. in the figure) and quarter-wave plates (labelled λ/4)the light is made to circulate from one cavity to the other, building up phase shift from a gravity wave signal over the total storage time of the system.

The buildup of phase differences over a long storage time can give a considerable improvement in the photon counting limit to the sensitivity of both these types of interferometers. In essence the sensitivity is improved over that obtainable in one period of the signal with a conventional multireflection system by a factor approximately equal to the ratio of the storage time to the period of the signal, with a further improvement by the square root of the number of periods integrated over. The photon counting limit to sensitivity becomes approximately $h = (\lambda \hbar c (1-R)^2 / 2\pi I \tau' L^2)^{1/2}$, where R is the maximum mirror reflectivity available, and τ' is the total duration of the measurement. This arrangement can in principle give such a good photon counting limit to sensitivity that substitution of parameters for a large low-loss interferometer might make it seem straightforward to detect the expected gravity wave flux from the Crab or Vela pulsars. This is misleading, however, since other noise sources have to be considered also, and in this case thermal noise from the suspension and even the quantum limit for the test masses are likely to be serious problems. This type of interferometer will probably be more useful at slightly higher frequencies, perhaps for the more intense periodic signals which may follow some collapse processes.

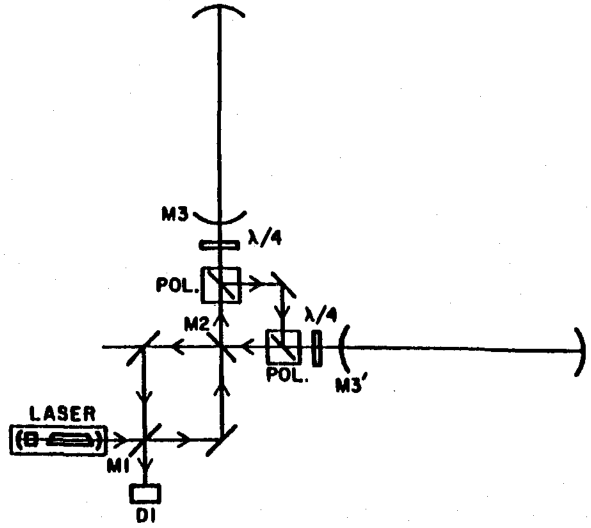


Figure 10

Having now discussed techniques for detection of pulses and of periodic signals using laser interferometer gravity wave detectors, it might be worth mentioning briefly how these same instruments might be used to detect a stochastic background of gravitational radiation.

9. DETECTION OF A STOCHASTIC BACKGROUND

Laser interferometer detectors seem quite promising instruments for searches for a stochastic background of gravitational radiation, as might arise for example, from collapse of black holes at an early epoch. In a search of this type, the signal has the form of noise itself, and a considerable improvement in effective sensitivity can be obtained by use of a pair of detectors in a correlation mode to provide discrimination against internal noise from either instrument. Early experiments of this type have been carried out using resonant bar gravity wave detectors at Glasgow[14] and similar experiments have also been performed at Tokyo[15]. There is, however, a real possibility of achieving a much more interesting sensitivity with laser interferometers due to their higher expected intrinsic sensitivity together with wide bandwidth.

The principle of such an experiment is indicated in Figure 11 where the outputs

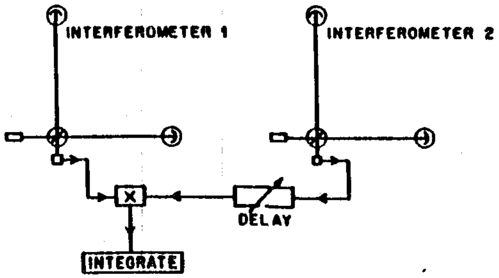


Figure 11

from two laser interferometric gravity wave detectors are multiplied together and integrated over a suitable observing time τ'. If the distance between the detectors is small compared with the wavelength of the gravitational radiation of interest, any common signal will give a correlated output and the power sensitivity obtained in the measurement becomes better than that achievable with a single detector alone by a factor of $(\pi B\tau')^{-\frac{1}{2}}$, where B is the bandwidth of each instrument by itself. If we take as example, a pair of optical cavity detectors with armlengths of 40 m, mirrors having reflectivity corresponding to $(1-R)=10^{-3}$ operating at a frequency centered on 100 Hz with a bandwidth of 100 Hz, and with other parameters as before, we might expect a gravity wave amplitude sensitivity for the individual detectors of the order of $2 \times 10^{-21}/\sqrt{Hz}$. An experiment involving correlation over 10^5 seconds could then set a limit of about $3 \times 10^{-23}/\sqrt{Hz}$ for radiation with the most favorable polarisation and flux direction.If it were assumed that gravitational radiation flux is concentrated in a frequency region near the frequency investigated, then an experiment of this type might set a limit to this gravity wave energy corresponding to a few percent of the closure density for the universe. With larger and more sensitive detectors such as those discussed earlier, correlation searches become feasible at a quite interesting level, and such experiments could form a very useful part of a general search for gravitational radiation of all types. They might, for example, provide the best sensitivity achievable for a wide range of unpredicted kinds of signals, such as large numbers of very small pulses, or nearly periodic signals of various types.

10. GENERAL REMARKS

It is hoped that the account given here of some current developments and ideas relating to laser interferometer gravity wave detectors gives a fair picture of the present state of this type of research. These instruments are complex and difficult ones, and their development presents a real challenge to the experimental physicist. It is too early yet to know which experimental approaches will prove most successful for the eventual unambiguous detection and investigation of gravitational wave signals, but the detectors discussed here seem at least as promising as other instruments of comparable cost and difficulty and the possibility of tailoring a single detector for several different types of experiment, suggested by some of the ideas outlined here, seems an interesting one. A considerable amount of difficult experimental work will still be necessary before experiments near the limits of sensitivity discussed here are likely to be made, but the prospects look good and the possibilities for real development of gravitational wave astronomy look interesting and exciting.

ACKNOWLEDGMENTS

The author would like to acknowledge the essential contributions of his colleagues at the California Institute of Technology and the University of Glasgow both in relation to discussions of principles and ideas and to underlying experimental work. The research in the USA is supported by the National Science Foundation (Grant NSF PHY82 04056) and in the UK by the Science and Engineering Research Council.

REFERENCES

(1) Thorne, K.S., Rev. Mod. Phys. 52 (1980) 285.
Smarr, L. Sources of Gravitational Radiation (Cambridge University Press, 1979).
Epstein, R. and Clark, J.P.A. in: Smarr, L. (ed.) Sources of Gravitational Radiation (Cambridge University Press, 1979).
Thorne, K.S. in: Lebovitz, N.R., Reid, W.H. and Vandervoort, P.O. (eds.) Theoretical Principles in Astrophysics and Relativity (University of Chicago Press, 1978) 149.
Douglass, D.H. and Braginsky,V.B., in:Hawking,S.W. and Israel, W.(eds.), General Relativity: An Einstein Centenary Survey (Cambridge University Press, 1979).
Tyson, J.A. and Giffard, R.P., Ann Rev. of Astro. Astrophys. 16 (1978) 521.
Press, W.H. and Thorne, K.S., Ann. Rev. of Astro. Astrophys. 10 (1972) 335.

(2) Edelstein, W.A., Hough, J., Pugh, J.R., and Martin, W, J. Phys. E. Sci. Instrum., 11 (1978) 710.
Caves, C.M., Phys. Rev. Letters 45 (1980) 75.

(3) Moss, G.E., Miller, L.R. and Forward, R.L., Appl.Opt. 10 (1971) 2495.
Forward, R.L., Phys. Rev. D17 (1978) 379.

(4) Caves, C.M., Phys. Rev. D23 (1981) 1693.

(5) Weiss,R., Progress Report 105, Res. Lab Electronics, MIT (1972) 54.

(6) Billing, H., Maischberger, K., Rudiger, A., Schilling, S., Schnupp, L. and Winkler, W., J. Phys. E12 (1979) 1043.

(7) Drever, R.W.P., Hough, J., Edelstein, W.A., Pugh, J.R., Martin, W., Proc. of the Intern. Sympos. on Experimental Gravitation, Pavia 1976, B. Bertotti (ed.) (Accad. Nazionale dei Lincei, 1977).

(8) Schilling, R., Schnupp, L., Winkler, W., Billing, H., Maischberger, K. and Rudiger, A., J. Phys. E. Sci. Instrum., 14 (1981) 65.

(9) Drever, R.W.P., Ford, G.M., Hough, J., Kerr, I., Munley, A.J., Pugh, J.R., Robertson, N.A. and Ward, H., 9th International Conference on General Relativity and Gravitation, GR9, Jena (1980), (in press).

(10) Drever, R.W.P., Hough, J., Munley, A.J., Lee, S-A., Spero, R., Whitcomb, S.E., Ward, H., Ford, G.M., Hereld, M., Robertson, N.A., Kerr , I., Pugh, J.R., Newton, G.P., Meers, B., Brooks III, E.D. and Gursel, Y., Proc. of the 5th International Conference on Laser Spectroscopy (Springer-Verlag, 1981) 33.

(11) Hough, J., Drever, R.W.P., Munley, A.J., Lee, S-A., Spero, R., Whitcomb, S.E., Ward, H., Ford, G.M., Hereld, M., Robertson, N.A., Kerr, I., Pugh, J.R., Newton, G.P., Meers, B., Brooks III, E.D. and Gursel, Y. Proc. of the NATO Advanced Study Institute, Bad Windsheim, West Germany 1981 (in press).

(12) Braginsky, V.B. and Manukin, A.B., Measurement of Small Forces in Physical Experiments (Nauka, Moscow, 1974; University of Chicago Press, 1977).

(13) Drever, R.W.P., Hough, J., Munley, A.J., Lee, S-A., Spero, R., Whitcomb, S.E., Ward, H., Ford, G.M., Hereld, M., Robertson, N.A., Kerr, I., Pugh, J.R., Newton, G.P., Meers, B., Brooks III, E.D. and Gursel, Y. Proc. of the NATO Advanced Study Institute, Bad Windsheim, West Germany 1981 (in press).

(14) Hough, J., Pugh, J.R., Bland, R. and Drever, R.W.P. Nature 254 (1975) 498.

(15) Hirakawa, H. and Narihara, K., Phys. Rev. Lett. 35 (1975) 330.

FOOTNOTE

*Some of the initial development and testing of this stabilisation technique was done in collaborative work by J.L. Hall and F.W. Kowalski of the Joint Institute for Laboratory Astrophysics, University of Colorado, and the University of Glasgow gravity wave group.[9]

PHYSICAL REVIEW D

PARTICLES AND FIELDS

THIRD SERIES, VOLUME 38, NUMBER 2 15 JULY 1988

Noise behavior of the Garching 30-meter prototype gravitational-wave detector

D. Shoemaker,* R. Schilling, L. Schnupp, W. Winkler, K. Maischberger, and A. Rüdiger
Max-Planck-Institut für Quantenoptik, D-8046 Garching bei München, Federal Republic of Germany

(Received 22 February 1988)

The prototype gravitational-wave detector at Garching is described: in a laser-illuminated Michelson interferometer having arms 30 m in length, a folded optical path of 3 km is realized. The origin, action, and magnitude of possible noise sources are given. The agreement between the expected and measured noise is good. For a band of astrophysical interest, extending from 1 to 6 kHz, the quantum shot noise corresponding to a light power of $P = 0.23$ W is dominant. In terms of the dimensionless strain h the best sensitivity in a 1-kHz bandwidth is $h = 3\times 10^{-18}$, comparable to the most sensitive Weber-bar-type antennas.

I. INTRODUCTION

Various methods have been proposed for the detection of gravitational radiation; two methods have been developed to a point where the chances of a successful search for gravitational-wave events can be realistically assessed. The *resonant bar technique* was pioneered by Weber[1] and followed by further efforts, first with room-temperature bars (Billing *et al.*[2]) and more recently with cooled bars using very-low-noise superconducting transducers (see Ref. 3 and references therein; also, for the current sensitivities, see Ref. 4). These experiments have put important upper limits on the level of gravitational radiation. The best sensitivities for the gravitational strain h that have been obtained so far are of the order of $h \approx 10^{-18}$.

A different approach is to use *interferometric techniques* to sense changes in the optical path length between widely separated test masses, as first discussed by Gertsenshtein and Pustovoit.[5] Early workers in this field include Weiss[6] and Forward.[7] Since that time a number of groups have pursued this method, and there are now prototype interferometers at MIT,[8] Glasgow,[9] Caltech,[10] Orsay,[11] and at the Max-Planck-Institut für Quantenoptik, Garching. From the best sensitivity values of these prototypes, as obtained at Garching,[12] and recently also at Glasgow,[13] one can extrapolate to the sensitivity of large interferometric antennas; an ultimate goal of better than $h \approx 10^{-21}$ seems attainable.

Work on laser interferometers at the Max-Planck-Institute started as early as 1974 (initially at the Institut für Astrophysik, now at the MPI für Quantenopik) and first concentrated on a 3-m arm-length prototype (Ref. 14 and references therein). After encouraging results it was decided that one could profit from a longer baseline interferometer (30 m), the construction of which was completed in mid-1983. The goal of the research is to investigate the noise sources in prototype gravitational-wave detectors as an aid in planning full-scale detectors.[15] This paper will describe the 30-m instrument, and particularly its 1986 improvements, in some detail, and it will present the status of the understanding of noise sources observed therein.

The basic design of the 30-m prototype is quite similar to the earlier 3-m prototype;[16] a schematic diagram is shown in Fig. 1. A Michelson interferometer formed of effectively "free" masses is illuminated by an argon-ion laser, with the light path folded in the optical delay-line configuration[17] to increase the sensitivity of the interferometer to gravitational waves. The interference pattern (on diode $D1$) is held to a minimum of intensity by a servosystem, and the control signal of this servo would contain the gravitational-wave signal. This control signal can be interpreted in terms of equivalent mirror displacement x; this is convenient for comparison with noise sources.

Because the response of the interferometer to gravitationally induced strains is broadband in nature, as are most of the noise sources encountered, it is helpful to work with noise spectral densities (that is, the noise contributed per unit bandwidth). Rigorously, such spectral densities are defined as the *squared* deviations per unit bandwidth, say, in m^2/Hz. It has become customary, however, to express them in *linear* measure, so that the units come out as m/$\sqrt{\text{Hz}}$. In this paper, such linear spectral densities will be characterized by a tilde above the symbol. In Fig. 2 the spectra of anticipated noise sources are illustrated, expressed as equivalent mirror motion $\tilde{x}$, calibrated in units of m/$\sqrt{\text{Hz}}$. To calculate the corresponding sensitivity in gravitational strain.

©1988 The American Physical Society

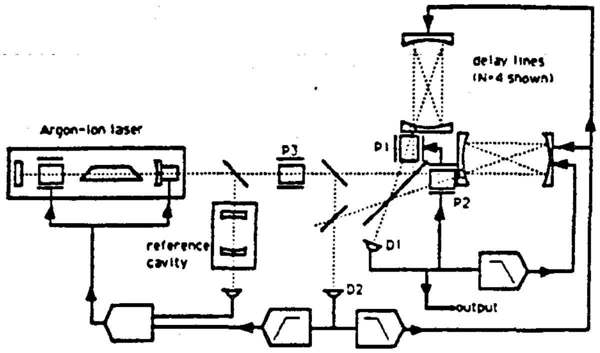


FIG. 1. A schematic view of the interferometer.

$\tilde{h} \equiv \tilde{\delta l}/l$, the mirror motion $\tilde{x}$ is divided by the interferometer arm length l (30 m for the Garching interferometer).

II. QUANTUM NOISE

The two arms of the interferometer are at right angles to each other (Fig. 1). The optical path in each arm is folded in a delay-line configuration and the beams returning from the two delay lines are brought to interference. The separation between the mirrors (which have a radius of curvature of 31.6 m) of the delay line can be varied between 29 and 32 m to obtain the desired number of beams in the delay line. For the data presented here, $N=90$ beams are used, giving a light storage time of $\tau=9$ μs. The separation of the delay-line mirrors can be adjusted with motion-driven translation stages to find the "reentrant condition" for the delay line. This leads to a first-order independence of the path length on tilts, rotations, and lateral translations of the far delay-line mirrors.[18,19]

The beam splitter and each of the delay-line mirrors are balanced in a simple wire sling pendulum which forms part of the mechanical isolation system (see Secs. IV and V, and Fig. 3), and the other optical components are treated similarly. The coarse alignment of the interferometer is achieved in the vertical axis with rotation of the pendulum suspension points (motor driven) and in the horizontal axis by screw adjustment of the point at which the pendulum wires leave the optical component. Fine adjustment is achieved with offset currents in the coils of the active pendulum damping system (see below); this allows optimization of the contrast in the interferometer. The best contrast $K=(I_{\max}-I_{\min})/(I_{\max}+I_{\min})$ observed with the 90 beam delay line is $K=0.992$, $I_{\max}$ and $I_{\min}$ being the photocurrents at the maximum and the minimum of the interference pattern. The contrast is limited primarily by imperfections in the delay-line mirrors themselves. The optical system holds a contrast of $K \geq 0.96$ for several days without readjustment.

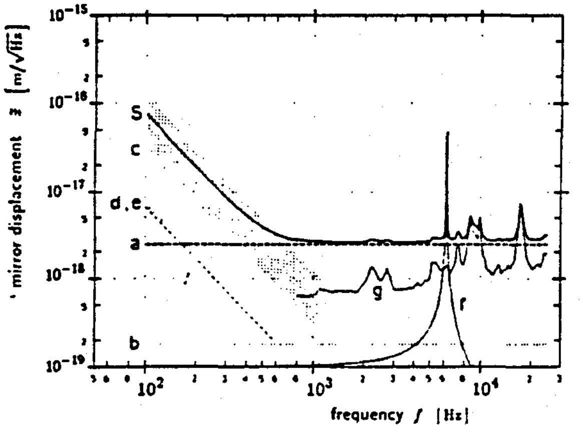


FIG. 2. Spectral densities of various noise sources, expressed as equivalent mirror motion $\tilde{x}$, in units of m/$\sqrt{\text{Hz}}$. a, photon shot noise; b, residual gas fluctuations; c, filtered ground motion; d, electronic damping system; e, pendulum thermal motion; f, mirror thermal motion; g, laser frequency fluctuations; S, quad-

If the only limit to determining the position of the masses were the shot noise of the photocurrent in the photodetector, and if the contrast of the interferometer were perfect, the noise-equivalent position fluctuation would be

$$\tilde{x}=\frac{1}{N}\frac{\lambda}{2\pi}\left[\frac{2e}{I_{\max}}\right]^{1/2}\left[\frac{\pi f\tau}{\sin\pi f\tau}\right], \qquad (1)$$

where λ is the wavelength of the light used to illuminate the interferometer, e the elementary charge, and f the measurement frequency. For the storage time τ given, and the frequencies f considered, the factor in the second set of parentheses is close to unity. The finite contrast compromises this sensitivity, as do technical noise sources (Johnson noise in the photodetector, amplifier noise); an expression which takes these factors into account, and which is relevant for the modulation scheme used, is derived in Appendix A. In the ideal case ($K=1.00$, no technical noise), it reduces to the simple form above. For the data presented here (with the experimental conditions $N=90$, $\lambda=514.5$ nm, $K=0.96$, $I_{\max}=70$ mA, corresponding to a maximum of about 0.23 W on the photodetector), the calculation results in a shot-noise level which is equivalent to a displacement of 2.5×10^{-18} m/$\sqrt{\text{Hz}}$, shown as curve a in Fig. 2. This is a factor 1.25 (2 dB) greater than the ideal case of perfect contrast and no additional noise sources. The influence of radiation pressure fluctuations[20,6] is completely negligible at the power levels encountered here.

III. FLUCTUATIONS OF RESIDUAL GAS PRESSURE

The entire interferometer is contained in a vacuum system to reduce the effect of refractive index fluctuations and ambient acoustic noise on the apparent path length. The vacuum system has three vertical tanks, 1.0 m in diameter, 1.1 m in height. The "central tank," which houses the beam splitter, the near delay-line mirrors, and the input optics, is connected to the two "end tanks" (which contain the far delay-line mirrors) by horizontal tubes 0.4 m in diameter. The end tanks are on a system of rails, and it is possible to add extension tubes (max-tanks. Only the central tank is in the laboratory; the end tanks are in separate end houses. The horizontal tubes are supported by steel guides on a concrete bed, which in turn is covered by a semicircular concrete cover and about one-half meter of earth. Rotary and turbomolecular pumps allow the system to be pumped from atmospheric pressure (10^5 Pa) to 10^{-2} Pa in 6 h; with the pumps turned off, the system pressure rises to 1 Pa in 24 h. The measurements presented here were performed with pressures between 10^{-1} and 1 Pa.

A noise source to be considered, although not a significant one at present, stems from the residual gas in the vacuum system. The number of molecules in the light path fluctuates, leading to small changes in the apparent optical index, and hence in path length. An estimate[21] of the magnitude of this effect, valid for the experimental conditions in the 30-m prototype, gives an equivalent mirror motion with a linear spectral density

$$\tilde{x} \approx \left[\frac{2\sqrt{3}\pi(n_0-1)^2}{(A_0/V_0)c_0\sqrt{\lambda}} \frac{\sqrt{l}}{N} \left[\frac{p}{p_0}\right] \left[\frac{T_0}{T}\right]^{3/2} \right]^{1/2} , \qquad (2)$$

where A_0 is Avogadro's number (6.02×10^{23} molecules/mole), V_0 is the volume of one mole of gas at standard temperature and pressure (22.4×10^{-3} m^3/mole), n_0 is the index of refraction of the gas, c_0 is the most probable thermal molecular speed in the gas (for nitrogen at room temperature $T=300$ K, $n_0\approx1+2.7\times10^{-4}$ and $c_0\approx400$ m/s), and p_0 and T_0 are the standard pressure and temperature. For these values and typical measurement pressures p one finds $\tilde{x}\approx2\times10^{-19}$ m/$\sqrt{\text{Hz}}$ (curve b in Fig. 2).

IV. MOTIONS OF THE OPTICAL COMPONENTS

As mentioned, to isolate the optical components from movement of the suspension point, they are hung as pendulums of length l, and thus of resonant angular frequency $\omega_0=\sqrt{g/l}$. The $(\omega_0/\omega)^2$ isolation that one would expect from an ideal pendulum is compromised by the finite quality factor Q of the pendulum, and by the suspension wire resonances ("violin string" modes); a model which predicts well the measured transfer function $H(\omega)$ (ratio of mirror motion to suspension point motion) of the suspension system is derived in Appendix B, and the measured and predicted transfer function are plotted in Fig. 6 below. For the pendulums in use at Garching, one finds a transfer function which can be roughly characterized as $(\omega_0/\omega)^2$ to the first "violin string" wire resonance of 212 Hz, then exhibiting a complex resonant structure with an isolation typically better than 10^{-4}.

This wire sling pendulum is suspended in turn from an upper pendulum, consisting of a massive plate suspended by coil springs 0.1 m in length (Fig. 3). In addition to the improvement of the isolation in the longitudinal direction due to the two pendulums in series (see Appendix B), the vertical compliance of the coil spring provides isolation from vertical and rotational motions of the overall support structure which could be cross-coupled into horizontal motion of the optical component. Up to 60 Hz, the measured transfer function (see Fig. 6) of the compound (two-stage) pendulum falls as $(\omega_1\omega_0)^2/\omega^4$, where ω_1 and ω_0 are the angular resonant frequencies of the top and bottom pendulums; above a transition section extending from 60 to 200 Hz, the attenuation is typically 10^{-6}. The ground-noise spectrum of the laboratory in Garching, while neither stationary nor smooth, can be roughly characterized by 3×10^{-7} (1 Hz/f)^2 m/$\sqrt{\text{Hz}}$ between 1 Hz and 1 kHz. Hence, the residual relative motion of the optical components can be estimated, and it is shown as the hatched area labeled as c in Fig. 2.

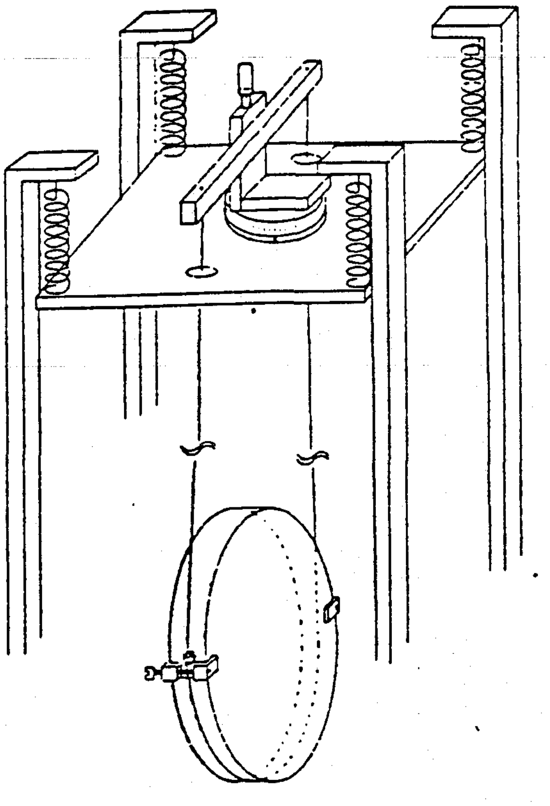


FIG. 3. The seismic isolation system. The upper stage, supported by four coil springs, carries translation and rotation stages for the mirror suspension wire. The upper suspension is 0.10 m in length, the lower 0.72 m. One of the screws for coarse adjustment of mirror tilt is indicated.

The pendulums are electronically damped at low frequencies[22] to prevent large motions due to the ground noise at the resonance frequency. To detect the motion of the pendulum with respect to the suspension point, for each degree of freedom to be damped a small vane is mounted on the optical component, and an infrared light-emitting diode (LED) and opposing silicon photodiode are mounted on the suspension point base. The vane partially interrupts the light, developing a signal proportional to the displacement of the vane. The electronic noise of this transducer, for the geometry used, has a displacement equivalent of 5×10^{-9} m/$\sqrt{\text{Hz}}$ at 0.75 Hz, and thus is well below the values of the order 10^{-7} m/$\sqrt{\text{Hz}}$ actually being measured. This electronic noise, from a total of 16 transducers, causes motions of the mirrors via the damping coils (see below). However, the influence of this noise is reduced both electronically b filters in the damping servoamplifiers and mechanically by the mass of the optical component, leading to expected motions of $10^{-10}(1\ \text{Hz}/f)^4\ \text{m}/\sqrt{\text{Hz}}$, much less than the present sensitivity.

The damping forces are applied by a small permanent magnet mounted integrally with the vane, and an aircore electromagnet mounted concentrically with the LED-photodiode assembly. The coil is positioned so that the magnetic field gradient is maximized at the magnet; this results in the best decoupling of the forces exerted on the optical component from (ground-noise-induced) motion of the coil. Assuming an error in the coil position of as much as 1 mm, and the largest control current possible, the natural ground motion could lead to optical component motions of the order of $4\times10^{-14}(1\ \text{Hz}/f)^4\ \text{m}/\sqrt{\text{Hz}}$, a negligible level at the present sensitivity. Noise in the coil current due to the final amplifier, which is left wideband to allow the use of the coil-magnet system in the interferometer locking servosystems, must be taken into account; for the present system, this noise results in a motion of the optical components of $6.8\times10^{-14}(1\ \text{Hz}/f)^2\ \text{m}/\sqrt{\text{Hz}}$, illustrated as curve *d* in Fig. 2.

The thermally driven noise of the pendulum can be estimated by considering it as a damped harmonic oscillator; here the Q is that due to the pendulum without electronic damping, because the damping servosystem gain is rolled off at high frequencies. For the thermal motion of an oscillator with an internal energy of $\frac{1}{2}k_BT$ one expects a spectral density

$$\tilde{x}=\left[\frac{4k_BT}{mQ\omega_0^3}\right]^{1/2}\left[\frac{1}{\left[1-\left[\frac{\omega}{\omega_0}\right]^2\right]^2+\frac{1}{Q^2}\left[\frac{\omega}{\omega_0}\right]^2}\right]^{1/2} \tag{3}$$

which for the pendulum, in the limit of frequencies high compared with the pendulum resonant frequency, gives $7.4\times10^{-14}(1\ \text{Hz}/f)^2\ \text{m}/\sqrt{\text{Hz}}$. This is shown as curve *e* in Fig. 2. Lossy isolation systems (for instance, lead and rubber stacks) have been avoided, as measurements show that only relatively little isolation with a complicated resonance spectrum is achieved. The tendency of such systems to "creep," and their unknown thermally driven motion, make them unattractive for future designs.

The thermally driven motion of the internal modes of the mirrors themselves must also be considered. The delay-line mirrors and the beam splitter are made of circular substrates 150 mm in diameter and 25 mm thick. The previously mentioned wire sling suspension system keeps the mechanical resonances simple and of high Q. The observed frequency of 6.3 kHz for the lowest mode of the mirrors agrees well with calculation[23] for a free cylinder; the Q of this resonance is 500 (limited by the plastic clips which guide the suspension wires, and the material of construction, Zerodur). The motion measured at the peak, $5\times10^{-17}\ \text{m}/\sqrt{\text{Hz}}$, is about a factor of 7 less than that which one calculates. This can be explained by noting that the lowest mode has radial nodal lines, and the pattern is sampled by the beam spots roughly equally often on the approaching and receding sections of the mirror, leading to a noticeable cancellation of the effect of the path-length change. The calculated contribution to the displacement noise is shown in curve *f* of Fig. 2. For frequencies much lower than the resonance the noise is at a level of $1\times10^{-19}\ \text{m}/\sqrt{\text{Hz}}$.

V. FLUCTUATIONS OF LASER LIGHT POWER

The interferometer output falling on diode *D*1 (see Fig. 1) is held to a dark fringe with a modulation method:[6] In both arms of the interferometer, Pockels cells (*P*1 and *P*2 in Fig. 1) in the light path between the beam splitter and the near-delay-line mirror are used to impress a high-frequency (10-MHz) phase modulation on the light. The light falling on the measurement photodiode *D*1 is demodulated, and the resulting error signal is amplified, filtered, and applied to the Pockels cells to hold the intensity on the photodiode to a minimum. The voltage applied to the Pockels cells, which is a linear function of the change of the light path in the delay lines, would carry the gravitational-wave signal. The phase modulation frequency is chosen to be in the frequency range where the amplitude noise of the argon laser is limited (at the power levels used) by the photon shot noise, typically above 5 MHz for the Coherent Innova 90-5 employed. To reduce the light reflected from the measurement photodiode *D*1 (EGG type DT110), it is held at the Brewster angle (for silicon ≈75°), thus achieving a quantum efficiency of about 80%.

Keeping the interference pattern at a minimum of intensity reduces the sensitivity of the measurement to amplitude noise in the illuminating laser beam, keeps the intensity on the measurement photodiode *D*1 at a manageable level, and allows the other output beam of the interferometer to be used for other purposes. The unity gain frequency in this servoloop must be high for two reasons: first, to give a very large gain at dc, ensuring that the interferometer is held accurately to the dark fringe thus eliminating the influence of low-frequency laser intensity fluctuations on the signal; and second, to give a reliable measurement signal at the highest signal frequency of interest (about 10 kHz). Because, in the present optical arrangement, the light passes through the Pockels cells both before and after the delay line, the unity gain frequency is limited by the time delay in the delay line to about 60 kHz (for 90 beams in the delay line). To reduce the dynamic range of the signal applied to the Pockels cells at very low frequencies (where the signals are largest) this control signal is also sent, suitably filtered, to the damping coils which exert forces on the far mirrors. The unity-gain point in this slow servoloop is set to about 30 Hz, a compromise between dynamic range reduction and the desire not to be acting on the mirrors mechanically in the frequency range of interest (above 100 Hz). This servosystem reduces the influence of amplitude noise to a negligible level.

VI. FLUCTUATIONS OF LASER LIGHT FREQUENCY AND POSITION

The light which illuminates the interferometer must be stabilized in frequency because of imperfections in the optical system. The delay-line mirrors are not all of exactly the same radius of curvature, which means that meeting the reentrant condition for both delay lines does not result in total path lengths that are exactly equal, but which differ by a static offset ΔL. This directly translates frequency fluctuations $\tilde{\delta\nu}$ into apparent mirror motion $\tilde{x}$ (Ref. 16):

$$\tilde{x}=\frac{\Delta L}{N}\frac{\tilde{\delta\nu}}{\nu}\ . \qquad (4)$$

A total path-length difference of about 2 m is observed when both 90 beam delay lines are reentrant, implying a net mismatch in curvature of about 0.02 m for the present mirrors.

In addition, stray light in the delay lines (due, for instance, to scattering from the mirror surfaces) allows interference between light having traveled different path lengths;[16,18] for $N=90$ beams in the delay line, the characteristic difference is about 3 km. The high reflectivity of the mirrors allows significant contributions even from "trapped" stray light which has made a large number of round trips of the delay line (corresponding to path differences up to the order of 100 km). The current mirrors employed show a scattering coefficient σ (relative amplitude of the scattered light which then interferes with the main beam) of about 10^{-4}; given the mirror reflectivity of 0.997, one finds that the fluctuation in apparent path length due to this effect is of the same order of magnitude as that due to the static path-length difference. Thus, the coefficient that relates frequency noise $\tilde{\delta\nu}$ to apparent mirror motion $\tilde{x}$ is about 7×10^{-17} m/Hz.

The frequency is stabilized in two steps (see Fig. 1). In the first step,[14] some of the light leaving the laser is directed through a Fabry-Perot reference cavity; the transmitted light intensity is compared with a reference beam intensity, and an error signal is formed. Suitably amplified and filtered, it is applied at low frequencies to a piezoelectric transducer holding one of the laser mirrors, and at high frequencies to an intracavity Pockels cell. The unity-gain frequency of this servoloop is about 400 kHz, limited by the bandwidth of the electronics.

In the second step,[12] a fraction of the light leaving the interferometer on the input side of the beam splitter is brought to interference with a fraction of the input light. The phase difference between these two beams is a function of the average path length in the two arms and the remaining frequency noise on the laser light. The error signal is developed with a modulation-demodulation scheme similar to that described for the main signal path, utilizing Pockels cell $P3$ and photodiode $D2$. At frequencies in the range of a few hertz, the Fabry-Perot cavity is the more stable reference, and the error signal is used to prevent common-mode motion of the delay-line mirrors (i.e., the interference pattern on $D2$ is held to a minimum of intensity). At frequencies higher than a few tens of hertz, the length of the delay line is a quiet and more sensitive length reference; the error signal in this frequency range is fed back to the laser. The gain possible in this servoloop is primarily limited by the time delay in the delay line; for a storage time of 9 μs (corresponding to $N=90$ beams in the delay line) it is about 60 kHz. An estimate of the influence on the interferometer of the remaining frequency noise is shown as curve g in Fig. 2. The resonance structures seen around 10 kHz are due to thermally driven motions in the Fabry-Perot cavity.

Fluctuations in the beam geometry can be translated into apparent mirror motions.[22] In a simple Michelson interferometer, a misalignment of the beam splitter leads to a sensitivity to changes in the beam position; a difference in the length of the two arms leads to a sensitivity to beam direction. In addition, there remains a sensitivity due to irregularities in the optical components (mirrors and Pockels cells in particular) even when the interferometer is well aligned and symmetrized.

A single-mode optical fiber is used to suppress such fluctuations in the beam geometry[24] and to carry the light from the laser table into the vacuum. The fiber employed here (the fiber used is an experimental type, unfortunately not in production, kindly donated by AEG) has a mode radius $w=2.7$ μm, and microscope objectives (10×) are used to couple into and out of the fiber. It is not a polarization-holding fiber, and to correct for slow thermal drifts in the polarization a $\lambda/2$ plate in front of the input to the fiber is used. The fiber output assembly is isolated from ground noise by suspending it on an electronically damped pendulum mass; this is necessary to obtain a jitter-free beam, and allows convenient adjustment of the input beam angle and position as well.

A coefficient for the sensitivity to beam motion can be obtained by giving a calibrated motion to the beam and viewing the effect in the apparent mirror motion spectrum, which in the system with $N=90$ bounces gave 2×10^{-6} m/rad. Attempts to measure the residual beam jitter after the fiber are limited by measurement noise (shot noise in the quadrant photodiode current) at 3×10^{-12} rad/$\sqrt{\text{Hz}}$ for all frequencies higher than 50 Hz. Thus the *upper limit* for the influence of beam jitter on the interferometer spectrum lies at 6×10^{-18} m/$\sqrt{\text{Hz}}$, a factor of 2 higher than the observed interferometer noise level in the kilohertz range; clearly, a more sensitive independent measurement of the beam motion after the fiber is needed before it can be eliminated as a possible noise source.

VII. PERFORMANCE OF THE INTERFEROMETER

Estimates of the contribution to the noise "budget" by each of the noise sources mentioned above have been made in Fig. 2, as well as a quadratic sum of all of these sources (curve S). The figure for beam jitter is not included in this sum because it is only an upper limit. The output signal of the interferometer—the control signal for the Pockels cells in the arms of the interferometer—is usually analyzed by a Fourier-transform spectrum analyzer (hp3582A). Continuous measurements of 30 min or an hour are possible, and when the interferometer loses "lock" (usually due to a longitudinal mode hop in the laser) the servosystems are automatically sequentially "relocked." Figure 4, curve A, is a composition of several 2-min averages at different sampling rates (to cover the broad frequency range presented), resulting in the relatively small uncertainty in the noise level. Also

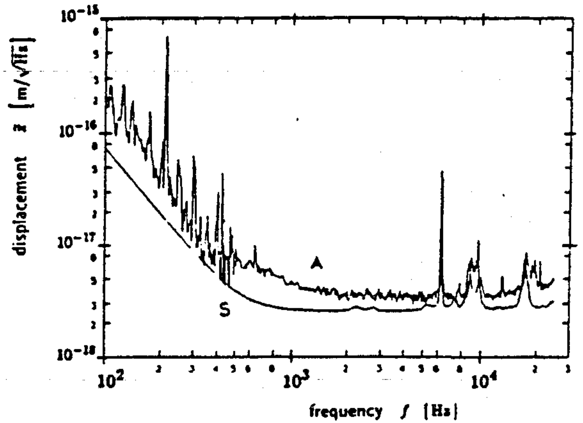


FIG. 4. The interferometer noise spectrum. *A*, measured; *S*, predicted.

shown is the quadratic sum of all noise sources (curve *S*); it is seen that the noise is in excess of the estimate.

Up to frequencies of several hundred hertz, multiples of the 50-Hz main supply are quite evident. Also seen are several suspension wire resonances (at 212 and 424 Hz). The overall level lies above that predicted; possible explanations are that remaining cross coupling in the pendulum isolation system allows ground motion to drive the mirrors, or that there is insufficient decoupling between the laser and the interferometer.

At higher frequencies the mirror resonances at 6.3 kHz are visible, as is the remaining influence of resonances in the Fabry-Perot cavity used as the frequency stabilization reference. There are still contributions from the mains harmonics, although they are not resolved in the spectrum. In the quietest frequency band, between 1 and 6 kHz, there is a discrepancy of a factor of 1.4 (or in logarithmic measure 3 dB) between the predicted shot-noise level and that observed. Experiments with the 0.3-m interferometer (see Appendix C), and with different power levels (see Appendix A), suggest that this excess consists of two parts: a discrepancy between the calculated and observed shot noise, which causes a scaling error (i.e., a constant error in logarithmic measure) of 1.06; and a constant noise (measured in mirror displacement) at a level of 2.5×10^{-18} m/$\sqrt{\text{Hz}}$. The latter noise source could be one (or several) of the previously mentioned white-noise sources. To gain more knowledge in this frequency regime the shot-noise limit must be significantly reduced, requiring considerably more light power.

VIII. CONCLUSION

The Garching prototype interferometer, with an optical path of 90×30 m, is very close to the shot-noise limit calculated for the relatively high power available, and over a broad frequency range of astrophysical interest. Even though this *prototype* has a delay-line storage time of only 9 μs, it already has a sensitivity (expressed as an equivalent dimensionless strain *h* of 3×10^{-18} in a 1-kHz bandwidth) which for many predicted sources is comparable to the most sensitive bar-type antennas.[25] Optimizing the storage time for the frequency range envisaged would allow an increase in sensitivity by almost 2 orders of magnitude. Much of the technology developed here can be extended to full-scale gravitational-wave interferometers,[15] leading to optimism for the feasibility of gravitational-wave detection in the near future.

APPENDIX A: CALCULATION OF THE SHOT NOISE

The current $I_{\text{ph}}(t)$ in the measurement photodiode $D1$ is

$$I_{\text{ph}}(t)=I_{\min}+\frac{I_{\max}-I_{\min}}{2}[1-\cos\phi(t)] \tag{A1}$$

with

$$\phi(t)=kx+\phi_m\sin\omega_m t \tag{A2}$$

and

$$x=x_0+\delta x(t)\ , \tag{A3}$$

where $k=(2\pi)/\lambda$, x_0 the static difference in optical path between the arms of the interferometer (the "operating point"), $\delta x(t)$ the signal (small in comparison with λ, and slowly changing in comparison with the modulation frequency), ϕ_m the amplitude of the high-frequency phase modulation applied by the Pockels cells, and ω_m the modulation frequency. This photodiode current can be expanded in a series of Bessel functions $J_n(\phi_m)$; keeping only the lowest-order terms (this corresponds to bandpass filtering the photodiode signal around the modulation frequency and at dc) one finds

$$I_{\text{dc}}=I_{\min}+\frac{I_{\text{eff}}}{2}[1-J_0(\phi_m)\cos kx] \tag{A4}$$

and

$$I_{\omega_m}=\frac{I_{\text{eff}}}{2}2J_1(\phi_m)\sin kx\ \sin\omega_m t\ , \tag{A5}$$

with an effective current swing of $I_{\text{eff}}=(I_{\max}-I_{\min})$.

The servosystem is arranged to hold the output of the interferometer on diode $D1$ at a minimum of intensity; this corresponds to $kx\ll1$ (modulo 2π), or

$$I_{\text{dc}}\approx I_{\min}+\frac{I_{\text{eff}}}{2}[1-J_0(\phi_m)] \tag{A6}$$

and

$$I_{\omega_m}\approx I_{\text{eff}}J_1(\phi_m)k\,\delta x(t)\sin\omega_m t\ . \tag{A7}$$

The signal is demodulated by multiplying by a square wave at ω_m; taking a net mixer gain of R, and keeping only terms around dc, one finds

$$V_{\text{sig}}=RI_{\text{eff}}J_1(\phi_m)k\,\delta x(t)\ . \tag{A8}$$

The noise that competes with this signal is the shot noise due to the flow of current in the photodiode I_{dc}, and the technical noise sources (Johnson noise of the photodiode internal resistance, amplifier electronic noise). The latter can be characterized as an additional fixed virtual current I_{det} in the photodiode. The noise due to these two currents is white and sity of $\sqrt{2e(I_{dc}+I_{det})}$ at the photodiode; it is filtered by the photodiode amplifier so that the amplitude is negligible for frequencies outside of the band $\omega_m \pm \omega_{sig}$, with ω_{sig} the (angular) signal frequency. This noise is then multiplied by the square wave in the mixer, resulting in noise components which are mixed down to the signal frequency:

$$\tilde{V}_{noise}=R\sqrt{2}\sqrt{2e(I_{dc}+I_{det})}\ . \tag{A9}$$

The $\sqrt{2}$ is due to the fact that the noise above and below the carrier are mixed down to the same low (positive) frequency, and add incoherently.

Equating the signal and the noise, one finds that the equivalent displacement noise due to the shot noise can be expressed by the linear spectral density

$$\tilde{x}_{shot}=\frac{1}{k}\sqrt{2}\frac{\sqrt{2e(I_{dc}+I_{det})}}{J_1(\phi_m)I_{eff}}\ . \tag{A10}$$

The expression for $\tilde{x}_{shot}$ as a function of ϕ_m has a gentle minimum, and the ϕ_m corresponding to this minimum should be applied to observe the highest sensitivity. The signal which is used for the output of the interferometer is the control signal to the Pockels cells; in the limit of large loop gain, this voltage is related to the position noise via the Pockels cell voltage V_λ that causes a change in optical path by one wavelength λ. Then the noise voltage expected for the shot-noise limit is

$$\tilde{V}_{shot}=\frac{V_\lambda \tilde{x}_{shot}}{\lambda}\ . \tag{A11}$$

An approximate expression, valid for typical contrasts ($K>0.9$) and modulation depths ($I_{dc}<0.2I_{max}$), simplifies the calculation of the voltage shot noise $\tilde{V}_{shot}$:

$$\tilde{V}_{shot}\approx\frac{V_\lambda}{2\pi}\left[\frac{2e(I_{dc}+I_{det})}{(I_{dc}-I_{min})I_{eff}-\frac{3}{2}(I_{dc}-I_{min})^2}\right]^{1/2}\ . \tag{A12}$$

Under these circumstances, an approximation giving the optimum modulation depth can be found. The observable quantity $I_{dc,opt}$ corresponding to this optimum is

$$I_{dc,opt}\approx[\tfrac{2}{3}(I_{min}+I_{det})I_{max}]^{1/2}\ . \tag{A13}$$

Measurements were performed on the 30-m and simplified 0.3-m interferometers (Appendix C) to verify the accuracy of the exact expressions. The photodiode currents I_{max}, I_{min}, and I_{dc} were monitored with a precision resistor; the technical noise was measured by comparison with the noise due to an incandescent lamp at frequencies near the modulation frequency of 10 MHz where it is assumed that the lamp has no noise above shot noise. The V_λ for the Pockels cells was measured *in situ* by finding the voltage step which resulted in a change of exactly one λ of optical path length in the cell, and tested separately for nonlinearity (the voltage for a jump of 2λ is within high precision twice that needed for 1λ).

With the $N=90$ beam interferometer operating under normal conditions, the voltage noise on the Pockels cells in the quietest frequency range (4–5 kHz) was observed for a variety of conditions (changes in illuminating intensity, depth of modulation, contrast, and excess white light on the photodetector). A fit was made to the experimental data with three free parameters: scale factors β and κ which describe errors in V_λ and I_{det}, respectively, and an additive constant ϵ which describes a constant excess noise motion of the mirrors:

$$\tilde{V}^2_{shot,fit}=\left[\frac{\beta V_\lambda}{2\pi}\right]^2\frac{4e(\kappa I_{det}+I_{dc})}{J_1^2(\phi_m)I_{eff}^2}+\epsilon^2\ . \tag{A14}$$

The argument ϕ_m in the Bessel function $J_1(\phi_m)$ can be derived from the measured values I_{dc}, I_{min}, I_{eff}, as shown in the beginning of this appendix.

The result for the 30-m interferometer is shown in Fig. 5, where the *measured* noise in V^2/Hz is the independent variable and the *calculated* noise in V^2/Hz is the dependent variable, and the measured points are compared with the ideal case, a one-to-one relationship. A logarithmic graph was chosen to allow the large range of values to be clearly presented; however, a graph *linear* in both axes is a more useful diagnostic tool, with the advantage of allowing errors in the fit to appear in a simple graphical form. For instance, an error in the Pockels cell coefficient V_λ appears as a slope error, a constant additive noise ϵ as a displacement, and an error in the effective technical noise current I_{det} as a deviation from that which should be a straight line. For the data presented, the best β is 1.06, the best κ is 0.88, and the best ϵ is 2.37×10^{-7} V/$\sqrt{\text{Hz}}$ (the latter corresponding to a noise in mirror displacement of 2.5×10^{-18} m/$\sqrt{\text{Hz}}$). Because the origin of these differences from the expected values is not yet explained, the equation without modification is used to calculate the expected shot-noise limit for the 90 beam delay-line interferometer, as given

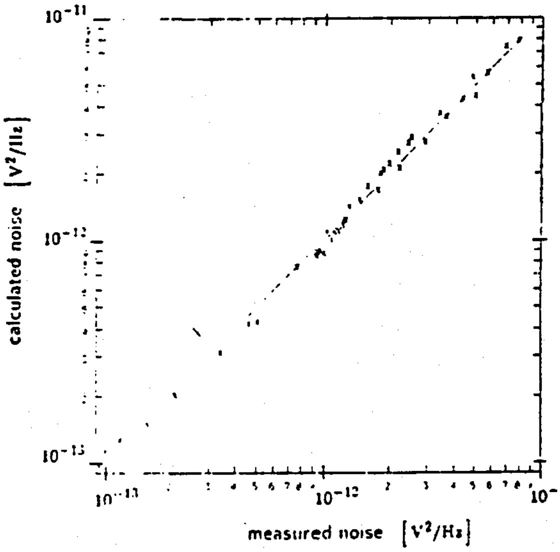


FIG. 5. Comparison of measured and calculated shot noise. The solid line indicates the locus of perfect agreement between measured and calculated noise levels.

in Fig. 2. However, it is reassuring that the errors are small, and that the functional dependence of the noise is correct.

APPENDIX B: TRANSFER FUNCTION OF THE PENDULUM ISOLATION SYSTEM

The pendulum can be treated in close analogy to an electrical (loss-free) transmission line, terminated with an inductance (to represent the inertial termination by the impedance $Z_P = i\omega m$ of the pendulum mass m). The characteristic impedance $Z = \sqrt{mg\gamma}$ of the mechanical transmission line is given by the tensile force mg on the wire and the linear mass density γ. The propagation constant $k = \omega / v_{tr}$ is determined by the velocity $v_{tr} = \sqrt{mg/\gamma}$ with which a transverse motion propagates along the wire.

As in an electrical transmission line, the displacement x_P at the termination (pendulum mass) is transformed to the front end (suspension point) via a transformation

$$x_0 = x_P \left[\frac{Z_P}{Z} i \sin kl + \cos kl \right] , \qquad \text{(B1)}$$

and one arrives at the transfer function magnitude

$$H(\omega) \equiv \frac{x_P}{x_0} = \frac{1}{\cos kl - \frac{\omega m}{Z} \sin kl} . \qquad \text{(B2)}$$

The lowest resonance $\omega_P = \sqrt{g/l}$ (the pendulation mode) and the well-known low-frequency transfer function $H(\omega) = [1 - (\omega/\omega_P)^2]^{-1}$ are easily derived by expanding for $kl \ll 1$; for $l = 0.72$ m we have $f_P = \omega_P / 2\pi \approx 0.59$ Hz.

All further resonances (the "violin string" resonances at ω_n) can be found from the approximation $kl \approx n\pi$, leading to

$$\omega_n \approx n\pi\omega_P \left[\frac{m}{\mu} \right]^{1/2} , \qquad \text{(B3)}$$

with $\mu = \gamma l$ the mass of the wire sling (two wires). In between these resonances, the transfer function $H(\omega)$ provides an isolation that is at best

$$H(\omega) \approx \frac{Z}{\omega m} = \frac{\omega_P}{\omega} \left[\frac{\mu}{m} \right]^{1/2} . \qquad \text{(B4)}$$

For the values used ($m = 1.1$ kg, steel wire 0.1 mm in diameter), the mass ratio m/μ is about 12 500, and the wire resonances are in very good agreement with the measured peaks at multiples of $f_1 \approx 212$ Hz. At these frequencies, the pendulum suspension not only loses its isolation feature, it may even enhance the pendulum mass motion. Figure 6 shows the measured (curve *a*) and calculated (curve *b*) transfer functions. The additional peaks observed in the measured transfer function at 300 and 550 Hz are due to pickup of harmonics of the 50-Hz line frequency.

The influence of the damping due to internal losses in the suspension system can be thought of as entering in two ways: first, as a modification in the high-frequency transfer function of the simple pendulum model, and second, as a change in the form of the "violin string" resonances. The finite Q of the pendulum motion, in particular if it is due to internal friction in the suspending wires, causes a transition from $H \approx (\omega_0/\omega)^2$ to $H \approx \omega_0/(\omega Q)$ at a transition frequency of $(\omega_0/2\pi)Q$. However, under normal conditions, the pendulum Q is primarily limited by damping from the residual gas: the measured pendulum Q shows a monotonic dependence on the residual gas pressure. At the lowest pressure attainable (10^{-3} Pa) it reaches $Q_{max} = 10^5$; at the normal operating pressure of 1 Pa, it is $Q = 3 \times 10^4$. If we take this Q_{max} as an upper limit for the internal losses, one calculates a transition frequency f_Q on the order of 10^5 Hz where the transfer function is dominated by the string resonances. These string resonances have measured Q's of the order of 2×10^4, which does not lead to a significant compromise of the transfer function; the primary effect is to keep the wire resonant peaks in the transfer function finite, in our case at values below unity. Thus, damping mechanisms do not significantly influence the performance of the isolation system.

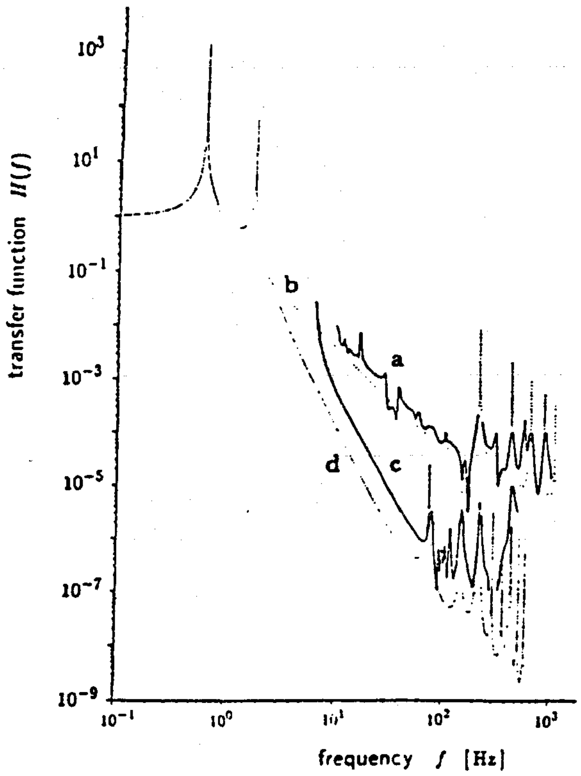


FIG. 6. Pendulum transfer functions. *a*, measured, single stage; *b*, calculated, single stage; *c*, measured, double stage; *d*, calculated, double stage.

The two-stage pendulum currently in use at Garching can be similarly analyzed. By twofold application of the transformation from the pendulum mass to the driving point, a transfer function for the compound pendulum is found. In contrast with the single pendulum, where there are no free parameters, it is necessary to characterize the coil springs (which form the upper pendulums) in terms of a linear mass density based on the measured transverse resonant frequencies and the length of the springs. With this parameter adjusted for the best fit, the measured and calculated transfer functions (in Fig. 6, curves c and d, respectively) again are in reasonably good agreement.

The measurement of the pendulum transfer function is made difficult by the large (120 dB) difference in amplitudes of mechanical motion between the pendulum suspension point and the mirror, and all acoustical or mechanical "short circuits" must be carefully avoided. For this reason, the measurement is performed entirely in the vacuum system of the interferometer. The pendulum suspension point of one of the far delay-line mirrors is driven parallel to the interferometer arm axis by an electromagnetic "shaker." The motion at this point is monitored with a piezoelectric accelerometer (Endevco model 7705-1000), and the motion of the mirror is measured with the aid of the $N=90$ beam interferometer.. The driving function for the "shaker" is a swept sine wave, the frequencies close to the "violin string" resonances having been avoided. A two-channel Fourier transform spectrum analyzer (hp3582A) calculates the raw transfer function, which is then corrected for the accelerometer response and the interferometer low-frequency servoloop. The resulting measured transfer functions agree well with the predictions (see Fig. 6).

APPENDIX C: THE 0.3-m INTERFEROMETER

An alternative configuration of the interferometer is formed by turning the near delay-line mirrors around so that the beam is immediately returned to the beam splitter, for a total path of 0.6 m. The sensitivity of the interferometer to mirror motion is much reduced, and the optical path is simplified. The reduction of scattered light and the ability to bring the interferometer to near perfect symmetry strongly reduces the constraints on the laser frequency stabilization. The frequency stabilization servoloop bandwidth achievable with the piezo-controlled laser mirror suffices, and this allows the operation of the laser without its internal Pockels cell, leading to higher output powers and thus more light in the interferometer.

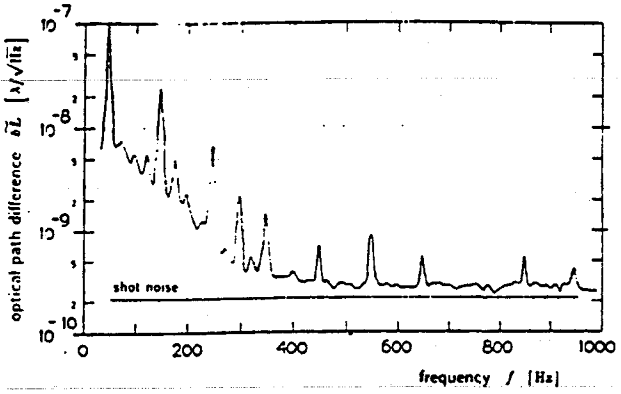


FIG. 7. Noise spectrum of the 30-cm test interferometer. The horizontal line represents the calculated shot noise for the experimental conditions. The regularly spaced peaks are caused by electrical disturbances from harmonics of the 50-Hz line frequency.

However, the interferometer fringe detection electronics are the same as for the 30-m interferometer, and the Pockels cells are still used in the same manner; this allows tests of much of the system without the added complications associated with the delay lines. In particular it has proven invaluable for helping to locate mechanical resonances. Except for the expected contributions of ground noise (below several hundred hertz) and thermally driven resonances of the optical components (above 6 kHz), the noise spectrum is limited by the shot noise in the measurement. A sample spectrum is shown in Fig. 7, for which $I_{\max}=233$ mA, and $I_{\min}=0.7$ mA; this corresponds to a contrast $K=0.994$. In addition to indicating that the measurement system works properly at this level, $\tilde{\delta L}=2.4\times10^{-10}\lambda/\sqrt{\mathrm{Hz}}$, it is reassuring to see that the optical fiber and the Pockels cells do not display anomalous effects with light powers on the order of 1 W.

*Present address: CNRS, Bâtiment 104, 91405 Orsay, France.

[1] J. Weber, Phys. Rev. Lett. 22, 1320 (1969).

[2] H. Billing, P. Kafka, K. Maischberger, F. Meyer, and W. Winkler, Lett. Nuovo Cimento 12, 111 (1975).

[3] V. B. Braginsky and K. S. Thorne, in *Present Status of Gravitational-Wave Experiments*, proceedings of the Ninth International Conference on General Relativity, Jena, Germany, 1980, edited by E. Schmutzer (Cambridge University Press, Cambridge, England, 1983), pp. 239–253.

[4] *Proceedings of the Fourth Marcel Grossmann Meeting on General Relativity*, Rome, 1985, edited by R. Ruffini (Elsevier, New York, 1986).

[5] M. E. Gertsenshtein and V. I. Pustovoit, Zh. Eksp. Teor. Fiz. 43, 605 (1962) [Sov. Phys. JETP 16, 433 (1963)].

[6] R. Weiss, MIT Research Laboratory of Electronics Report No. 105, 1972 (unpublished).

[7] R. L. Forward, Phys. Rev. D 17, 379 (1978).

[8] J. Livas, R. Benford, D. Dewey, A. Jeffries, P. Saulson, D. Shoemaker, and R. Weiss, in *Proceedings of the Fourth Marcel Grossmann Meeting on General Relativity* (Ref. 4), pp. 591–597.

[9] G. P. Newton, J. Hough, G. A. Kerr, B. J. Meers, N. A. Robertson, H. Ward, J. B. Mangan, and S. Hoggan, in *Proceedings of the Fourth Marcel Grossmann Meeting on General Relativity* (Ref. 4), pp. 599–604.

[10] R. Spero, in *Proceedings of the Fourth Marcel Grossmann Meeting on General Relativity* (Ref. 4), pp. 615–620.

[11] A. Brillet (private communication).

[12] D. H. Shoemaker, W. Winkler, K. Maischberger, A. Rüdiger, R. Schilling, and L. Schnupp, in *Proceedings of the Fourth Marcel Grossmann Meeting on General Relativity* (Ref. 4), pp. 605–614.

[13] H. Ward *et al.*, in *The Glasgow Gravitational Wave Detector—Present Progress and Future Plans*, proceedings of the International Symposium on Experimental Gravitational Physics, Guangzhou, People's Republic of China, 1987 (World Scientific, Singapore, in press).

berger, and L. Schnupp, in *Quantum Optics, Experimental Gravitation, and Measurement Theory,* edited by P. Meystre and M. O. Scully (Plenum, New York, 1983), pp. 525–566.

[15] W. Winkler, K. Maischberger, A. Rüdiger, R. Schilling, L. Schnupp, and D. H. Shoemaker, in *Proceedings of the Fourth Marcel Grossmann Meeting on General Relativity* (Ref. 4), pp. 621–630.

[16] H. Billing, K. Maischberger, A. Rüdiger, R. Schilling, L. Schnupp, and W. Winkler, J. Phys. E 12, 1043 (1979).

[17] D. Herriot, H. Kogelnik, and R. Kompfner, Appl. Opt. 3, 523 (1964).

[18] W. Winkler, Ph.D. thesis, München Internal Report No. MPQ 74, 1983 (unpublished).

[19] P. Linsay, P. Saulson, and R. Weiss, *A Study of a Long Baseline Gravitational Wave Antenna System* (MIT, Cambridge, MA, 1983).

[20] W. Winkler, in *A Laser Interferometer to Search for Gravitational Radiation,* proceedings of the International Meeting on Experimental Gravitation, Pavia, 1976 (Accademia Nazionale dei Lincei, Rome, 1977), pp. 351–363.

[21] A. Rüdiger (unpublished).

[22] K. Maischberger, A. Rüdiger, R. Schilling, L. Schnupp, W. Winkler, and H. Billing, in *Proceedings of the Second Marcel Grossmann Meeting on General Relativity,* Trieste, 1979, edited by R. Ruffini (North-Holland, Amsterdam, 1982), pp. 1083–1100.

[23] J. R. Hutchinson, ASME J. Appl. Mech. 47, 901 (1980).

[24] R. Weiss (private communication).

[25] D. Dewey, Phys. Rev. D 36, 1577 (1987).

REF. P

Automatic Control Systems

fourth edition

Benjamin C. Kuo

Professor of Electrical Engineering
University of Illinois at Urbana-Champaign

PRENTICE-HALL, INC., Englewood Cliffs, NJ 07632

chapter one

Introduction

1.1 CONTROL SYSTEMS

In this introductory chapter we attempt to familiarize the reader with the following subjects:

1. What a control system is.
2. Why control systems are important.
3. What the basic components of a control system are.
4. Why feedback is incorporated into most control systems.
5. Types of control systems.

With regard to the first two items, we cite the example of the human being as perhaps the most sophisticated and the most complex control system in existence. An average human being is capable of performing a wide range of tasks, including decision making. Some of these tasks, such as picking up objects, or walking from one point to another, are normally carried out in a routine fashion. Under certain conditions, some of these tasks are to be performed in the best possible way. For instance, an athlete running a 100-yard dash has the objective of running that distance in the shortest possible time. A marathon runner, on the other hand, not only must run the distance as quickly as possible, but in doing so, he or she must control the consumption of energy so that the best result can be achieved. Therefore, we can state in general that in life there are numerous "objectives" that need to be accomplished, and the means of achieving the objectives usually involve the need for control systems.

In recent years control systems have assumed an increasingly important role in the development and advancement of modern civilization and technology. Practically every aspect of our day-to-day activities is affected by some type of control system. For example, in the domestic domain, automatic controls in heating and air-conditioning systems regulate the temperature and humidity of homes and buildings for comfortable living. To achieve maximum efficiency in energy consumption, many modern heating and air-conditioning systems in large office and factory buildings are computer controlled.

Control systems are found in abundance in all sectors of industry, such as quality control of manufactured products, automatic assembly line, machine-tool control, space technology and weapon systems, computer control, transportation systems, power systems, robotics, and many others. Even such problems as inventory control, and social and economic systems control, may be approached from the theory of automatic controls.

Regardless of what type of control system we have, the basic ingredients of the system can be described by

1. Objectives of control
2. Control system components
3. Results

In block diagram form, the basic relationship between these three basic ingredients is illustrated in Fig. 1-1(a).

In more scientific terms, these three basic ingredients can be identified with inputs, system components, and outputs, respectively, as shown in Fig. 1-1(b).

In general, the objective of the control system is to control the outputs c in some prescribed manner by the inputs u through the elements of the control system. The inputs of the system are also called the *actuating signals*, and the outputs are known as the *controlled variables*.

As a simple example of the control system fashioned in Fig. 1-1, consider the steering control of an automobile. The direction of the two front wheels may be

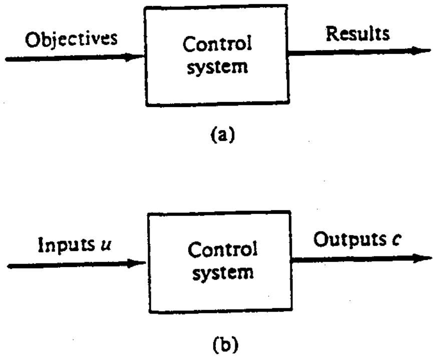


Figure 1-1 Basic components of a control system.

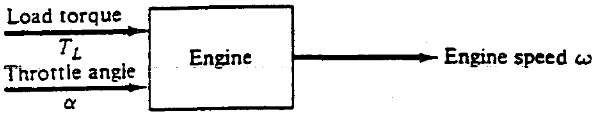


Figure 1-2 Idle-speed control system.

regarded as the controlled variable c, or the output; the direction of the steering wheel is the actuating signal u, or the input. The control system or process in this case is composed of the steering mechanisms and the dynamics of the entire automobile. However, if the objective is to control the speed of the automobile, then the amount of pressure exerted on the accelerator is the actuating signal, and the vehicle speed is the controlled variable. As a whole, we may regard the automobile control system as one with two inputs (steering and accelerator) and two outputs (heading and speed). In this case, the two controls and outputs are independent of each other; but in general, there are systems for which the controls are coupled. Systems with more than one input and one output are called *multivariable systems*.

As another example of a control system, we consider the idle-speed control of an automobile engine. The objective of such a control system is to maintain the engine idle speed at a relatively low value (for fuel economy) regardless of the applied engine loads (e.g., transmission, power steering, air conditioning, etc.). Without the idle-speed control, any sudden engine load application would cause a drop in engine speed which might cause the engine to stall. Thus, the main objectives of the idle-speed control system are (1) to eliminate or minimize the speed droop when engine loading is applied, and (2) to maintain the engine idle speed at a desired value. Figure 1-2 shows the block diagram of the idle-speed control system from the standpoint of inputs–system–outputs. In this case, the throttle angle α and the load torque T_L (due to the application of air conditioning, power steering, transmission, or brakes, etc.) are the inputs, and the engine speed ω is the output. The engine is the controlled process or system.

Open-Loop Control Systems (Nonfeedback Systems)

The idle-speed control system illustrated in Fig. 1-2 is rather unsophisticated and is called an *open-loop control system*. It is not difficult to see that the system as it is shown would not satisfactorily fulfill the desired performance requirements. For instance, if the throttle angle α is set at a certain initial value, which corresponds to a certain engine speed, when a load torque T_L is then applied, there is no way to prevent a drop in the engine speed. The only way to make the system work is to have means of adjusting α in response to a change in the load torque, in order to maintain ω at the desired level.

Because of the simplicity and economy of open-loop control systems, we may find this type of system in practical use in numerous situations. In fact, practically all automobiles manufactured prior to 1981 did not have an idle-speed control system.

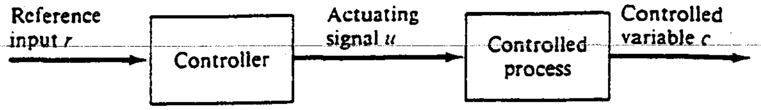


Figure 1-3 Elements of an open-loop control system.

An electric washing machine is another example of an open-loop system because, typically, the amount of machine wash time is entirely determined by the judgment and estimation of the human operator. A true automatic electric washing machine should have the means of checking the cleanliness of the clothes being washed continuously and turn itself off when the desired degree of cleanliness is reached.

The elements of an open-loop control system can usually be divided into two parts: the controller and the controlled process, as shown by the block diagram in Fig. 1-3. An input signal or command r is applied to the controller, whose output acts as the actuating signal u; the actuating signal then controls the controlled process so that the controlled variable c will perform according to some prescribed standards.

In simple cases, the controller can be an amplifier, mechanical linkages, or other control means, depending on the nature of the system. In more sophisticated electronics control, the controller can be an electronic computer, such as a microprocessor.

Closed-Loop Control Systems (Feedback Control Systems)

What is missing in the open-loop control system for more accurate and more adaptive control is a link or feedback from the output to the input of the system. To obtain more accurate control, the controlled signal $c(t)$ should be fed back and compared with the reference input, and an actuating signal proportional to the difference of the input and the output must be sent through the system to correct the error. A system with one or more feedback paths such as that just described is called a *closed-loop system*.

The block diagram of a closed-loop idle-speed control system is shown in Fig. 1-4. The reference input ω_r sets the desired idling speed. Ordinarily, when the load torque is zero, the engine speed at idle should agree with the reference value ω_r, and any difference between the actual speed and the desired speed caused by any disturbance such as the load torque T_L is sensed by the speed transducer and the error detector, and the controller will operate on the difference and provide a signal to adjust the throttle angle α to correct the error.

Figure 1-5 illustrates a comparison of the typical performances of the open-loop and closed-loop idle-speed control systems. In Fig. 1-5(a), the idle speed of the open-loop system will drop and settle at a lower value after a load torque is applied. In Fig. 1-5(b) the idle speed of the closed-loop system is shown to recover quickly to the preset value after the application of T_L.

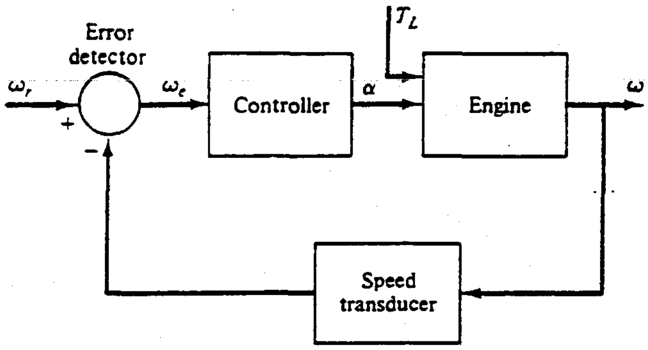


Figure 1-4 Closed-loop idle-speed control system.

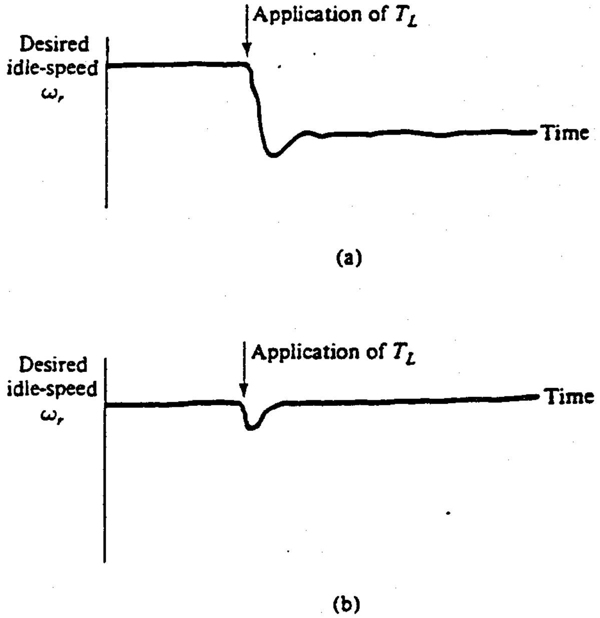


Figure 1-5 (a) Typical idle-speed response of an open-loop system. (b) Typical idle-speed response of a closed-loop system.

The idle-speed control system illustrated above is also known as a *regulator system* whose objective is to maintain the system output at some prescribed level.

As another illustrative example of a closed-loop control system, Fig. 1-6 shows the block diagram of the printwheel control system of a word processor or electronic typewriter. The printwheel, which typically has 96 or 100 characters, is to be rotated to position the desired character in front of the hammer for printing. The character selection is done in the usual manner from a keyboard. Once a certain key on the

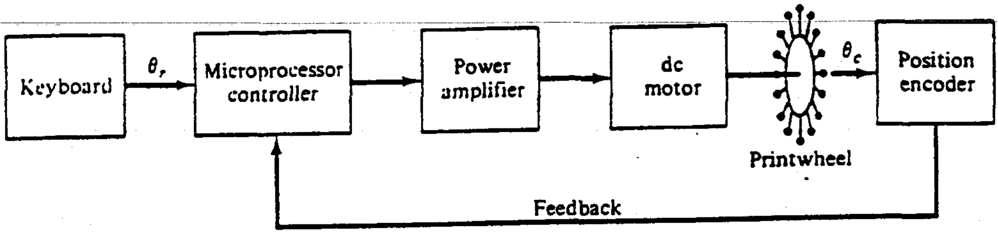


Figure 1-6 Printwheel control system.

keyboard is depressed, a command for the printwheel to rotate from the present position to the next position is initiated. The microprocessor computes the direction and the distance to be traveled, and sends out a control logic signal to the power amplifier, which in turn controls the motor that drives the printwheel. The position of the printwheel is detected by a position sensor whose output is compared with the desired position in the microprocessor. The motor is thus controlled in such a way as to drive the printwheel to the desired position. In practice, the control signals generated by the microprocessor controller should be able to drive the printwheel from one position to another sufficiently fast so that the printing can be done accurately within the specified time frame.

Figure 1-7 shows a typical set of input and output of the system. When a reference command input is given, the signal is represented as a step function. Since the electric circuit of the motor has inductance and the mechanical load has inertia, the printwheel cannot move to the desired position instantaneously. Typically, it will follow the response as shown, and settle at the new position after some time t_1. Printing cannot begin until the printwheel has come to a stop; otherwise, the character will be smeared. Figure 1-7 shows that after the printwheel has settled, the period from t_1 to t_2 is reserved for printing, so that after $t=t_2$, the system is ready to receive a new command.

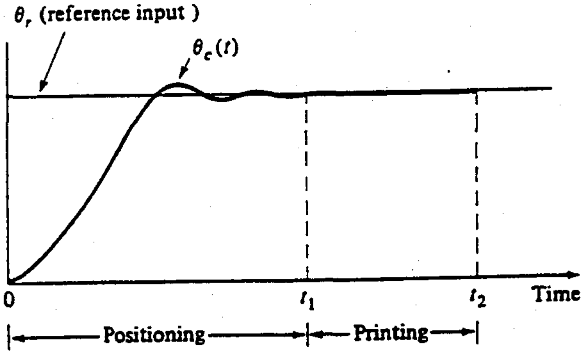


Figure 1-7 Typical input and output of the printwheel control system.

1.2 WHAT IS FEEDBACK AND WHAT ARE ITS EFFECTS?

The motivation for using feedback illustrated by the examples in Section 1.1 is somewhat oversimplified. In these examples the use of feedback is shown to be for the purpose of reducing the error between the reference input and the system output. However, from a theoretical standpoint the significance of the effects of feedback in control systems is much more profound than is demonstrated by these examples. The reduction of system error is merely one of the many important effects that feedback may have upon a system. We show in the following sections that feedback also has effects on such system performance characteristics as stability, bandwidth, overall gain, impedance, and sensitivity.

To understand the effects of feedback on a control system, it is essential that we examine this phenomenon with a broad mind. When feedback is deliberately introduced for the purpose of control, its existence is easily identified. However, there are numerous situations wherein a physical system that we normally recognize as an inherently nonfeedback system may turn out to have feedback when it is observed in a certain manner. In general we can state that whenever a closed sequence of *cause-and-effect relationships* exists among the variables of a system, feedback is said to exist. This viewpoint will inevitably admit feedback in a large number of systems that ordinarily would be identified as nonfeedback systems. However, with the availability of the feedback and control system theory, this general definition of feedback enables numerous systems, with or without physical feedback, to be studied in a systematic way once the existence of feedback in the above-mentioned sense is established.

We shall now investigate the effects of feedback on the various aspects of system performance. Without the necessary background and mathematical foundation of linear system theory, at this point we can only rely on simple static system notation for our discussion. Let us consider the simple feedback system configuration shown in Fig. 1-8, where r is the input signal, c the output signal, e the error, and b the feedback signal. The parameters G and H may be considered as constant gains. By simple algebraic manipulations it is simple to show that the input–output

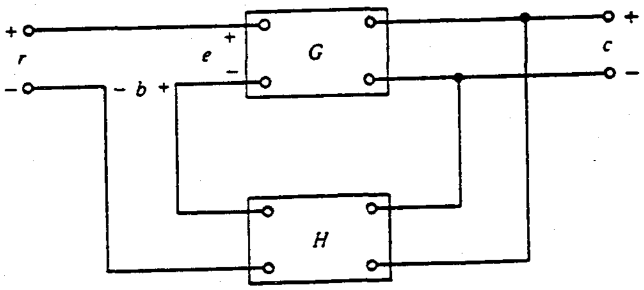


Figure 1-8 Feedback system.

relation of the system is

$$M = \frac{c}{r} = \frac{G}{1+GH} \tag{1-1}$$

Using this basic relationship of the feedback system structure, we can uncover some of the significant effects of feedback.

Effect of Feedback on Overall Gain

As seen from Eq. (1-1), feedback affects the gain G of a nonfeedback system by a factor of $1+GH$. The reference of the feedback in the system of Fig. 1-8 is negative, since a minus sign is assigned to the feedback signal. The quantity GH may itself include a minus sign, so the general effect of feedback is that it may increase or decrease the gain. In a practical control system, G and H are functions of frequency, so the magnitude of $1+GH$ may be greater than 1 in one frequency range but less than 1 in another. Therefore, feedback could increase the gain of the system in one frequency range but decrease it in another.

Effect of Feedback on Stability

Stability is a notion that describes whether the system will be able to follow the input command. In a nonrigorous manner, a system is said to be unstable if its output is out of control or increases without bound.

To investigate the effect of feedback on stability, we can again refer to the expression in Eq. (1-1). If $GH=-1$, the output of the system is infinite for any finite input. Therefore, we may state that feedback can cause a system that is originally stable to become unstable. Certainly, feedback is a two-edged sword; when it is improperly used, it can be harmful. It should be pointed out, however, that we are only dealing with the static case here, and, in general $GH=-1$ is not the only condition for instability.

It can be demonstrated that one of the advantages of incorporating feedback is that it can stabilize an unstable system. Let us assume that the feedback system in Fig. 1-8 is unstable because $GH=-1$. If we introduce another feedback loop through a negative feedback of F, as shown in Fig. 1-9, the input–output relation of the overall system is

$$\frac{c}{r} = \frac{G}{1+GH+GF} \tag{1-2}$$

It is apparent that although the properties of G and H are such that the inner-loop feedback system is unstable, because $GH=-1$, the overall system can be stable by properly selecting the outer-loop feedback gain F.

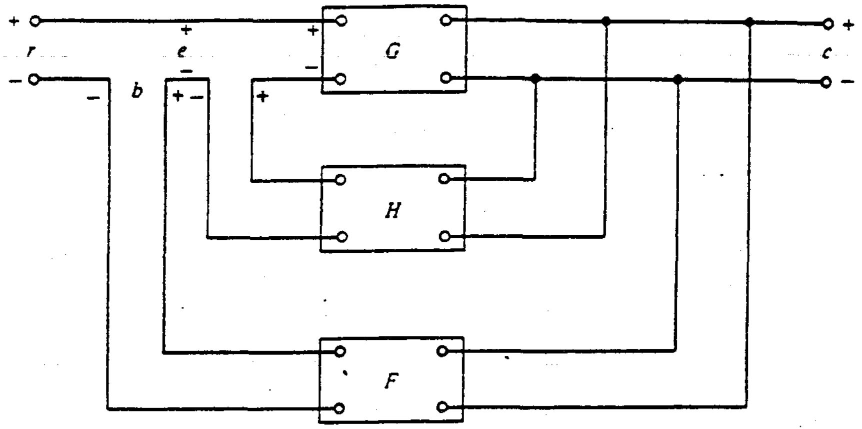


Figure 1-9 Feedback system with two feedback loops.

Effect of Feedback on Sensitivity

Sensitivity considerations often play an important role in the design of control systems. Since all physical elements have properties that change with environment and age, we cannot always consider the parameters of a control system to be completely stationary over the entire operating life of the system. For instance, the winding resistance of an electric motor changes as the temperature of the motor rises during operation. In general, a good control system should be very insensitive to these parameter variations while still able to follow the command responsively. We shall investigate what effect feedback has on the sensitivity to parameter variations.

Referring to the system in Fig. 1-8, we consider G as a parameter that may vary. The sensitivity of the gain of the overall system M to the variation in G is defined as

$$S_G^M = \frac{\partial M/M}{\partial G/G} \tag{1-3}$$

where ∂M denotes the incremental change in M due to the incremental change in G; $\partial M/M$ and $\partial G/G$ denote the percentage change in M and G, respectively. The expression of the sensitivity function S_G^M can be derived by using Eq. (1-1). We have

$$S_G^M = \frac{\partial M}{\partial G}\frac{G}{M} = \frac{1}{1+GH} \tag{1-4}$$

This relation shows that the sensitivity function can be made arbitrarily small by increasing GH, provided that the system remains stable. It is apparent that in an open-loop system the gain of the system will respond in a one-to-one fashion to the variation in G.

In general, the sensitivity of the system gain of a feedback system to parameter variations depends on where the parameter is located. The reader may derive the sensitivity of the system in Fig. 1-8 due to the variation of H.

Effect of Feedback on External Disturbance or Noise

All physical control systems are subject to some types of extraneous signals or noise during operation. Examples of these signals are thermal noise voltage in electronic amplifiers and brush or commutator noise in electric motors.

The effect of feedback on noise depends greatly on where the noise is introduced into the system; no general conclusions can be made. However, in many situations, feedback can reduce the effect of noise on system performance.

Let us refer to the system shown in Fig. 1-10, in which r denotes the command signal and n is the noise signal. In the absence of feedback, $H=0$, the output c is

$$c=G_1G_2e+G_2n \tag{1-5}$$

where $e=r$. The signal-to-noise ratio of the output is defined as

$$\frac{\text{output due to signal}}{\text{output due to noise}}=\frac{G_1G_2e}{G_2n}=G_1\frac{e}{n} \tag{1-6}$$

To increase the signal-to-noise ratio, evidently we should either increase the magnitude of G_1 or e relative to n. Varying the magnitude of G_2 would have no effect whatsoever on the ratio.

With the presence of feedback, the system output due to r and n acting simultaneously is

$$c=\frac{G_1G_2}{1+G_1G_2H}r+\frac{G_2}{1+G_1G_2H}n \tag{1-7}$$

Simply comparing Eq. (1-7) with Eq. (1-5) shows that the noise component in the output of Eq. (1-7) is reduced by the factor $1+G_1G_2H$, but the signal component is also reduced by the same amount. The signal-to-noise ratio is

$$\frac{\text{output due to signal}}{\text{output due to noise}}=\frac{G_1G_2r/(1+G_1G_2H)}{G_2n/(1+G_1G_2H)}=G_1\frac{r}{n} \tag{1-8}$$

and is the same as that without feedback. In this case feedback is shown to have no direct effect on the output signal-to-noise ratio of the system in Fig. 1-10. However, the application of feedback suggests a possibility of improving the signal-to-noise ratio under certain conditions. Let us assume that in the system of Fig. 1-10, if the magnitude of G_1 is increased to G_1' and that of the input r to r', with all other parameters unchanged, the output due to the input signal acting alone is at the same

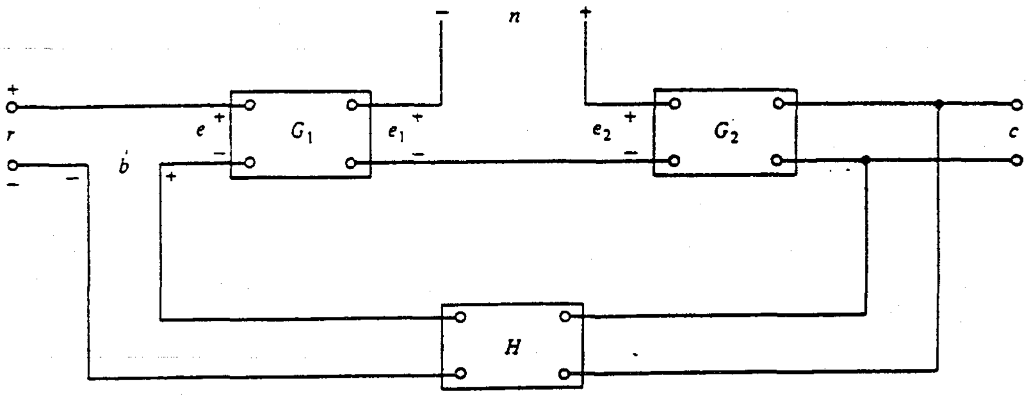


Figure 1-10 Feedback system with a noise signal.

level as that when feedback is absent. In other words, we let

$$c|_{n=0} = \frac{G_1'G_2r'}{1+G_1'G_2H} = G_1G_2r \tag{1-9}$$

With the increased G_1, G_1', the output due to noise acting alone becomes

$$c|_{r=0} = \frac{G_2n}{1+G_1'G_2H} \tag{1-10}$$

which is smaller than the output due to n when G_1 is not increased. The signal-to-noise ratio is now

$$\frac{G_1G_2r}{G_2n/(1+G_1'G_2H)} = \frac{G_1r}{n}(1+G_1'G_2H) \tag{1-11}$$

which is greater than that of the system without feedback by a factor of $(1+G_1'G_2H)$.

In general, feedback also has effects on such performance characteristics as bandwidth, impedance, transient response, and frequency response. These effects will become known as one progresses into the ensuing material of this text.

1.3 TYPES OF FEEDBACK CONTROL SYSTEMS

Feedback control systems may be classified in a number of ways, depending upon the purpose of the classification. For instance, according to the method of analysis and design, feedback control systems are classified as linear and nonlinear, time varying or time invariant. According to the types of signal found in the system, reference is often made to continuous-data and discrete-data systems, or modulated and unmodulated systems. Also, with reference to the type of system components, we often come across descriptions such as electromechanical control systems, hydraulic control systems, pneumatic systems, and biological control systems. Control systems are often classified according to the main purpose of the system. A positional control system and a velocity control system control the output variables according to the way the names imply. In general, there are many other ways of identifying control systems according to some special features of the system. It is important that some of these more common ways of classifying control systems are known so that proper perspective is gained before embarking on the analysis and design of these systems.

Linear versus Nonlinear Control Systems

This classification is made according to the methods of analysis and design. Strictly speaking, linear systems do not exist in practice, since all physical systems are nonlinear to some extent. Linear feedback control systems are idealized models that are fabricated by the analyst purely for the simplicity of analysis and design. When the magnitudes of the signals in a control system are limited to a range in which system components exhibit linear characteristics (i.e., the principle of superposition applies), the system is essentially linear. But when the magnitudes of the signals are extended outside the range of the linear operation, depending upon the severity of the nonlinearity, the system should no longer be considered linear. For instance, amplifiers used in control systems often exhibit saturation effect when their input signals become large; the magnetic field of a motor usually has saturation properties. Other common nonlinear effects found in control systems are the backlash or dead play between coupled gear members, nonlinear characteristics in springs, nonlinear frictional force or torque between moving members, and so on. Quite often, nonlinear characteristics are intentionally introduced in a control system to improve its performance or provide more effective control. For instance, to achieve minimum-time control, an on–off (bang-bang or relay) type of controller is used. This type of control is found in many missile or spacecraft control systems. For instance, in the attitude control of missiles and spacecraft, jets are mounted on the sides of the vehicle to provide reaction torque for attitude control. These jets are often controlled in a full-on or full-off fashion, so a fixed amount of air is applied from a given jet for a certain time duration to control the attitude of the space vehicle.

For linear systems there exists a wealth of analytical and graphical techniques for design and analysis purposes. However, nonlinear systems are very difficult to treat mathematically, and there are no general methods that may be used to solve a wide class of nonlinear systems.

Time-Invariant versus Time-Varying Systems

When the parameters of a control system are stationary with respect to time during the operation of the system, the system is called a time time-invariant system. In practice, most physical systems contain elements that drift or vary with time. For example, the winding resistance of an electric motor will vary when the motor is being first excited and its temperature is rising. Another example of a time-varying system is a guided-missile control system in which the mass of the missile decreases as the fuel on board is being consumed during flight. Although a time-varying system without nonlinearity is still a linear system, the analysis and design of this class of systems are usually much more complex than that of the linear time-invariant systems.

Continuous-Data Control Systems

A continuous-data system is one in which the signals at various parts of the system are all functions of the continuous time variable t. Among all continuous-data control systems, the signals may be further classified as ac or dc. Unlike the general definitions of ac and dc signals used in electrical engineering, ac and dc control systems carry special significances. When one refers to an ac control system it usually means that the signals in the system are modulated by some kind of modulation scheme. On the other hand, when a dc control system is referred to, it does not mean that all the signals in the system are of the direct-current type; then there would be no control movement. A dc control system simply implies that the signals are unmodulated, but they are still ac signals according to the conventional definition. The schematic diagıam of a closed-loop dc control system is shown in Fig. 1-11. Typical waveforms of the system in response to a step function input are shown in the figure. Typical components of a dc control system are potentiometers, dc amplifiers, dc motors, and dc tachometers.

The schematic diagram of a typical ac control system is shown in Fig. 1-12. In this case the signals in the system are modulated; that is, the information is transmitted by an ac carrier signal. Notice that the output controlled variable still behaves similar to that of the dc system if the two systems have the same control objective. In this case the modulated signals are demodulated by the low-pass

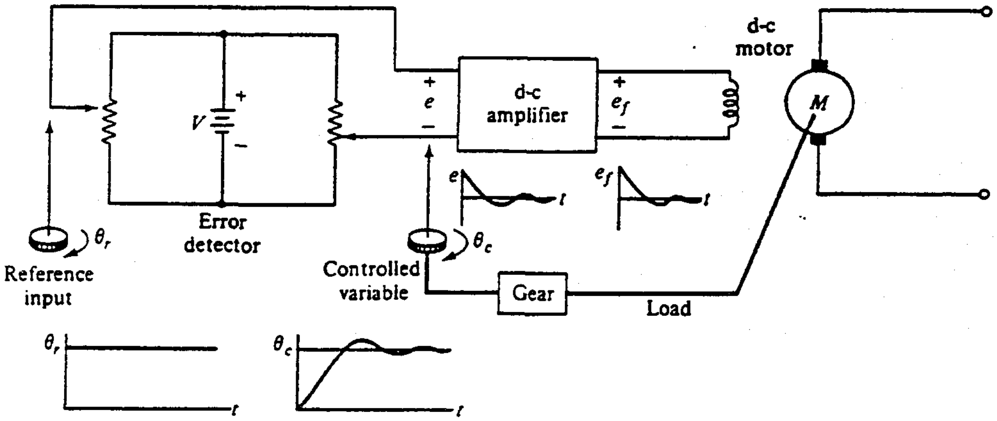


Figure 1-11 Schematic diagram of a typical dc closed-loop control system.

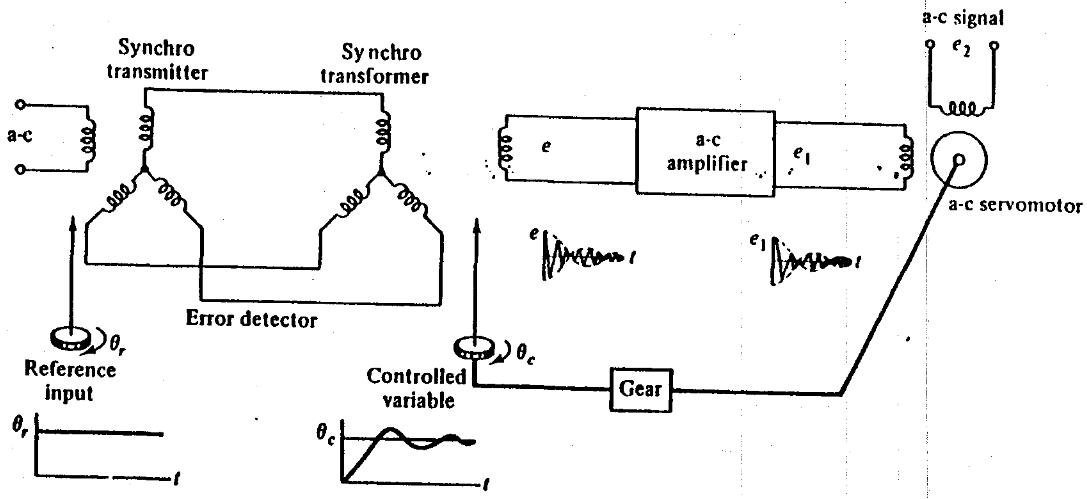


Figure 1-12 Schematic diagram of a typical ac closed-loop control system.

characteristics of the control motor. Typical components of an ac control system are synchros, ac amplifiers, ac motors, gyroscopes, and accelerometers.

In practice, not all control systems are strictly the ac or the dc type. A system may incorporate a mixture of ac and dc components, using modulators and demodulators to match the signals at various points of the system.

Sampled-Data and Digital Control Systems

Sampled-data and digital control systems differ from the continuous-data systems in that the signals at one or more points of the system are in the form of either a pulse train or a digital code. Usually, sampled-data systems refer to a more general class of systems whose signals are in the form of pulse data, where a digital control system refers to the use of a digital computer or controller in the system. In this text the term "discrete-data control system" is used to describe both types of systems. For example, the printwheel control system shown in Fig. 1-6 is a typical discrete-data or digital control system, since the microprocessor receives and outputs digital data.

In general, a sampled-data system receives data or information only intermittently at specific instants of time. For instance, the error signal in a control system may be supplied only intermittently in the form of pulses, in which case the control system receives no information about the error signal during the periods between two consecutive pulses. Figure 1-13 illustrates how a typical sampled-data system operates. A continuous input signal $r(t)$ is applied to the system. The error signal

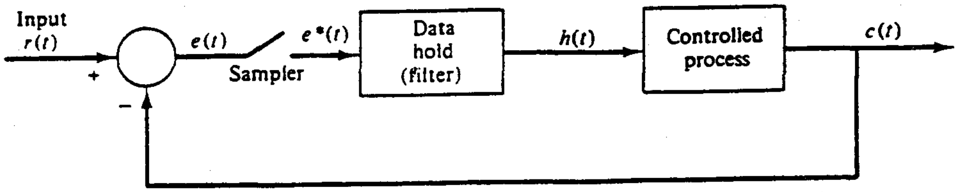


Figure 1-13 Block diagram of a sampled-data control system.

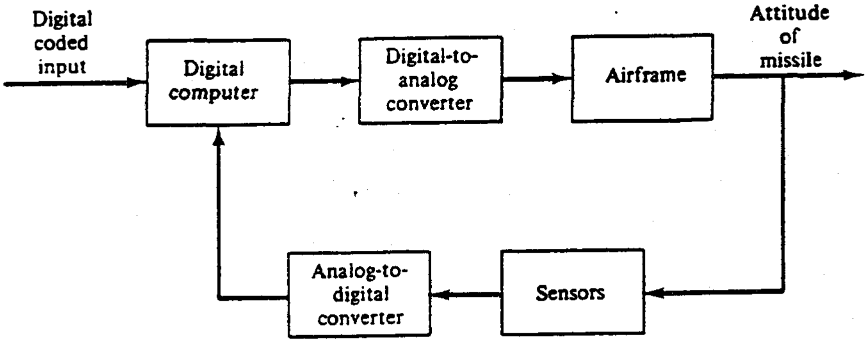


Figure 1-14 Digital autopilot system for a guided missile.

$e(t)$ is sampled by a sampling device, the sampler, and the output of the sampler is a sequence of pulses. The sampling rate of the sampler may or may not be uniform. There are many advantages of incorporating sampling into a control system. One advantage easily understood provides time sharing of expensive equipment among several control channels.

Because digital computers provide many advantages in size and flexibility, computer control has become increasingly popular in recent years. Many airborne systems contain digital controllers that can pack several thousand discrete elements in a space no larger than the size of this book. Figure 1-14 shows the basic elements of a digital autopilot for a guided missile.

BATCH START

Lecture 7: Lasers and Input Optics - 1

STAPLE OR DIVIDER

Ph 103c: The Physics of LIGO 20 April 1994

LECTURE 7
LASERS AND INPUT OPTICS—I

Lecture by Robert Spero

Assigned Reading:

Q. A. Rüdiger, R. Schilling, L. Schnupp, W. Winkler, H. Billing and K. Maischberger, "A mode selector to suppress fluctuations in laser beam geometry," *Optica Acta,* **28**, 641–658 (1981).

R. "Noise in Optical Detection and Generation," Chapter 10 of A. Yariv, *Optical Electronics* (Saunders College Publishing, 1991).

Suggested Supplementary Reading:

The following two articles explain how frequency noise originating in vacuum fluctuations is fundamentally independent of the properties of atoms and depends only on the mirror properties and other cavity losses, and that the sensitivity of interferometric gravitational wave detectors is the same whether the arms are empty ("passive") cavities, as in LIGO, or idealized ("active" cavity) lasers.

S. "Comparison Between Active-cavity and Passive-cavity Interferometers," Abramovici A., Vager Z, *Phys. Rev.* **A33** (5), 3181-3184 (1986).

T. "Passive Versus Active Interferometers–Why Cavity Losses Make them Equivalent", J. Geabanacloche, *Phys. Rev. A* **35**(6), 2518–2522 (1987).

U. T.M. Niebauer, R. Schilling, K. Danzmann, A. Rudiger, and W. Winkler, "Nonstationary Shot Noise and its Effect on the Sensitivity of Interferometers" *Phys. Rev A* **43**(9), 5022–5029(1991). This paper resolves a long-standing 15% discrepancy between the calculated and observed shot noise in the German 30 m interferometer, having to do with the shape of the waveforms used for modulation and demodulation.

V. P.H. Roll, R. Krotkov, and R.H. Dicke, *Ann. Phys* **26**, 442 (1964). This paper, though long, is fun to read. It describes a classic experiment to measure the equivalence of inertial and gravitational mass, and is an excellent example of how experiments are designed, operated, and analyzed. The pages excerpted for the handout show how an optical lever reads the torsion balance deflection. Dicke's clever design, using a vibrating wire that casts a shadow on a photdetector, is a prototype for the use of modulation to reduce noise.

A Few Suggested Problems:

The 40 m interferometer operates in an "unrecombined" configuration: the reflected beams from the two arms' input mirrors do not interfere, and are separately detected. The shot noise levels from the two photodetectors add in quadrature; in the case of identical arms the total shot noise equivalent displacement is

$$\Delta\tilde{L}(f) \equiv \tilde{x}(f) = \frac{l}{4\pi\tau_E}\sqrt{3\mathcal{F}}\left[\frac{\lambda \mathrm{h}}{cP}\left(1+[f/f_k]^2\right)\right]^{1/2}$$

where $\mathcal{F}$ contains terms that depend on the depth of modulation Γ. ($\Gamma = 1$ corresponds to phase modulation of amplitude 1 radian.)

$$\mathcal{F} = \frac{1}{3}\left[\frac{M^{-1} + A^2J_0^2 - 2AJ_0^2 + 2AJ_0J_2}{MA^2J_0^2J_1^2}\right]$$

M is the (energy) mode matching fraction; $0 < M < 1$, $M = 1$ being the case of perfect alignment of the mirrors and proper matching of the laser gaussian beam parameters to the cavity mirror curvatures and separation. The mismatched fraction of the laser beam $(M - 1)$ does not participate in the interfence, but does add to the shot noise. The 40 m interferometer operates with $M \approx 0.9$. J_0, J_1, J_2 are Bessel functions evaluated at Γ. Each cavity has input mirror transmission T and the sum of other losses L. τ_E, the cavity energy storage time, is the time it takes the intensity of the light "leaking" out of one of the arm cavities to drop from its starting level by a factor of e, after the input light is turned off; $\tau_E = \tau_t/(L+T)$, with $\tau_t = 2l/c$ the round-trip transit time and l the length of each arm. The cavity knee frequency is $f_k = 1/(4\pi\tau_E)$. $A = 2T/(L+T)$ is the amplitude of the cavity field leaking back out through the input mirror on resonance, in the absence of modulation. It is normalized to the input amplitude, and is constrained by $0 < A < 2$. λ is the optical wavelength, P is the total power (corrected for inefficiency in the photodiodes and other losses outside the arm cavities) incident on the beamsplitter, and f is the signal frequency.

1. Suppose the beamsplitter is not symmetric: that is, if P_1 and P_2 represent the power incident on the two arms, $P_1 = P\alpha$, $P_2 = P(1-\alpha)$, $\alpha \neq 0.5$. How does the sensitivity change from the symmetric case? How much asymmetry is required to degrade the sensitivity by 10%?
2. Verify that the modulation function $\mathcal{F}$ has a minimum value of 1. What parameters are required to approach this value? Optimization of interferometer sensitivity requires minimization of $\mathcal{F}$. Explain how the optimum value of Γ depends on the mode matching M and the mirror transmission and loss, T and L.
3. Even with l as short as 40 m, it is possible–using readily available very low-loss mirrors–to make f_k lower than the lowest expected detectable signal frequency f of approximately 100 Hz. Verify that for $f > f_k$, the shot-noise limited strain sensitivity $\tilde{h}(f) = \tilde{x}(f)l$ is independent of l, and make a plot sketching $\tilde{h}(f)$ for various values of l, all other parameters held fixed. The currently achieved shot-noise limited displacement sensitivity of the 40 m interferometer is approximately the same as the requirement for initial LIGO ($l = 4$ km) detectors. What are the implications for the design of LIGO detectors? For R&D on the 40 m interferometer?
4. Compare the shot noise sensitivity above to the "recombined" but not recycled calculation of Lecture 4. Explain why the sensitivity is worse for the unrecombined configuration.

LECTURE 8.
LASERS AND INPUT OPTICS — II
Lecture by Alex Abramovici

Assigned Reading:

W. A. Rüdiger, R. Schilling, L. Schnupp, W. Winkler, H. Billing and K. Maischberger, "A mode selector to suppress fluctuations in laser beam geometry," *Optica Acta,* **28**, 641–658 (1981). [This paper describes, first in simple terms and then in terms of a mode decomposition, the use of a *mode cleaning Fabry-Perot cavity* to precondition the light that is injected into an interferometer. The preconditioning includes suppression of beam wiggle, suppression of beam-diameter pulsations, and suppression of other unwanted spatial modes of the laser light. Also described is the use of lenses to adjust the radius of curvature of the beam's phase fronts so as to match the desired eigenmodes of the mode cleaner and of each arm of the interferometer. Note that, at the time this paper was written, "supermirrors" with losses far far less than 0.01 were not yet available, and the degree to which one can control mirror heating by keeping the mirrors extremely clean was not yet understood.]

X. Those students who are not familiar with the physics of lasers should also read the introductory chapter of a good text on laser physics; for example, Chapter 1, "Introduction" of W. Koechner, *Solid-State Laser Engineering* (Springer Verlag, Berlin, 1988), which is being passed out.

Suggested Supplementary Reading:

H. Read more deeply into your favorite laser physics text. Most especially, read the material dealing with Gaussian beams and their manipulation, e.g. the material already suggested in Lecture 4: Reference H — chapter 17, "Physical Properties of Gaussian Beams," of A. E. Siegman, *Lasers* (University Science Books, Mill Valley CA, 1986).

A Few Suggested Problems

1. *Laser Stabilization by Locking to a Cavity.* The frequency of a laser is stabilized by locking it to an eigenmode of an optical cavity using a feedback loop. Suppose that, in the absence of the feedback loop, the laser's frequency differs from the cavity's eigenfrequency by an amount $\Delta\nu \equiv \nu_o$ (the "initial detuning"). When the feedback system is turned on, the residual detuning is $\Delta\nu \equiv \nu_1 = \nu_o/(1+G)$, where $G \gg 1$ is the gain of the feedback system, which is proportional to the power I of the light beam. What is the rms frequency fluctuation σ_ν induced by an rms fluctuation σ_I of the optical power?
2. *Mode Cleaning Cavity—I.* The Fabry-Perot cavity that will make up each arm of LIGO's standard, broad-band interferometric gravitational-wave detector will have a corner mirror with modest power transmisivity $\mathcal{T}_c \equiv 1-\mathcal{R}_c \sim 10^{-2}$ and an end mirror with tiny transmissivity $\mathcal{T}_e \sim 10^{-5}$. With this huge difference of transmissivities, almost all the light injected into the cavity through the corner mirror ultimately leaves back through the corner mirror; hardly any leaks out the end mirror. In a mode cleaning Fabry-Perot cavity, by contrast, the two mirrors are chosen to have identical transmisivities $\mathcal{T}$. In this case show that, if the cavity (which is idealized as having no absorption or scattering) is driven on resonance through the left mirror, all the light leaves the cavity through the right mirror. If the cavity is driven off resonance, what fraction of the light goes out each end?
3. *Mode Cleaning Cavity—II.* Consider a mode-cleaning cavity consisting of two identical concave mirrors with radii of curvature $R = 1$m and power transmissivities $\mathcal{T} = 0.005$, separated by 1.5m along the optic axis. The cavity is driven by laser light that is primarily in the TEM$_{00}$ mode, with a small admixture of TEM$_{01}$; and it is driven on a TEM$_{00}$ resonance so all the light in that mode passes through the cavity from one side to the other. What fraction of the light in the TEM$_{01}$ mode passes through?
4. *Changes in wavefront curvature on reflection from a curved mirror.* The phase variation in the transverse plane (i.e. at constant z) for a diverging Gaussian beam propagating in the z direction (so $\psi \propto e^{+i(kz-\omega t)}$), with a phase-front radius of curvature R, is $\psi \propto e^{ikr^2/2R}$ [cf. Eq. (3.5) in Reference W above, or Eq. (7.35) of Reference G: chapter 7, "Diffraction," of Blandford and Thorne, *Applications of Classical Physics.*] What will be the transverse variation of this same wave (a) after reflecting off a planar mirror set up normal to the z axis? (b) after reflecting off a concave spherical mirror with radius of curvature $R_m = R$? (c) after passing through a converging lens with focal length f?

LECTURE 9.
OPTICAL ELEMENTS

Lecture by Rick Savage

Assigned Reading:

Y. W. Winkler, K. Danzmann, A. Rüdiger and R. Schilling, "Optical Problems in Interfereometric Gravitational Wave Antennas," in *The Sixth Marcel Grossmann Meeting*, eds. H. Sato and T. Nakamura (World Scientific, Singapore, 1991), pp. 176–191.

H. A. E. Siegman, *Lasers* (University Science Books, Mill Valley CA, 1986), chapter 17 "Physical Properties of Gaussian Beams": Section 17.1 "Gaussian Beam Propagation," (pages 663-674); section 17.4 "Axial Phase Shifts: The Guoy Effect," (pages 682-685), and section 17.5 "Higher-Order Gaussian Modes," (pages 685-691). [This material was suggested reading in Lecture 4.] If you did not read it then, you should read it now.]

Suggested Supplementary Reading:

Z. D. Malacara, *Optical Shop Testing* (John Wiley and Sons, New York, 1978), section 1.2, "Fizeau Interferometer," pp. 19–37.

AA. H. A. Macleod, *Thin-Film Optical Filters,* 2nd edition (Adam Hilger Ltd., Bristol, 1986), "Introduction," pp. 1–10.

SS. J. M. Elson, H. E. Bennett, and J. M. Bennett, "Scattering from Optical Surfaces," in *Applied Optical Engineering*, Vol. VII (Academic Press 1979), Chapter 7, page 191. [This is reproduced later, in connection with Lecture 15.]

A Few Suggested Problems

1. *Gaussian Beam Propagation.* The present conceptual design for the 4 km long LIGO arm cavities specifies that the input mirror be flat and the end mirror curved, with a radius of curvature of 6 km. The beam waist is therefore located on the flat input mirror and the spot size is significantly larger on the curved mirror than on the flat.
   a. Consider employing a symmetrical curved-curved mirror configuration instead of the flat-curved geometry. What is the required radius of curvature of the mirrors ($R_1 = R_2$) to maintain the cavity g factor product at $g_1 g_2 = 1/3$?
   b. Calculate the spot size at the beam waist and on the mirrors. What is the Rayleigh range for this configuration?
   c. What factors might influence the decision to adopt either the flat-curved or the symmetrical, curved-curved geometry?
2. *Scattering from mirror surface irregularities.* Consider the following simple model for scattering from mirror surface irregularities. Represent the mirror surface height $z = \mu(x, y)$ as a superposition of a number of spatially monochromatic terms. Assume, for the moment, that only one term is non-zero, and let that term have peak height a and wavelength Λ; i.e. set

$$\mu(x, y) = a \cos(2\pi x/\Lambda).$$

   Idealize the mirror to be of infinite extent and irradiated by a plane wave at normal incidence.
   a. Show that the wave reflected from the surface has spatial sidebands that propagate at some angle θ relative to the specularly reflected beam. What is θ? [Hint: This can be regarded as an exercise in Fraunhofer diffraction (Why?); see, e.g., Section 7.3 of chapter 7, "Diffraction", of Blandford and Thorne, *Applications of Classical Physics* (which was passed out in Lecture 4).]
   b. Find the power scattered into these sidebands as a function of the amplitude of the surface variation, a.
   c. Generalize to find the total light scattered into all angles from a surface with a total rms irregularity σ.

REF. W8(

OPTICA ACTA, 1981, VOL. 28, NO. 5, 641–658

A mode selector to suppress fluctuations in laser beam geometry

A. RÜDIGER, R. SCHILLING, L. SCHNUPP,
W. WINKLER, H. BILLING and K. MAISCHBERGER
Max-Planck-Institut für Physik und Astrophysik, Institut für Astrophysik, D-8046 Garching bei München, F.R. Germany

(*Received 3 July 1980; revision received 11 July 1980*)

Abstract. Our development of a gravitational wave detector requires a Michelson interferometer of extreme sensitivity capable of measuring 10^{-16} m (i.e. some 10^{-10} of a wavelength λ of the illuminating laser light). Even after painstaking alignment of the interferometer components, and after considerable improvement of the laser stability, noise contributions much in excess of this goal were observed, due partly to fluctuations of the laser beam geometry. The two most obvious types of geometric beam fluctuations are a lateral beam jitter and a pulsation in beam width; these lead to spurious interferometer signals if the interfering wavefronts are misaligned in their tilts or in their curvatures respectively.

The geometry of the laser beam can be considerably stabilized by passing it through an optical resonator. The geometric beam fluctuations, as viewed from this resonator, can be described by a well-centred ground mode TEM_{00}, contaminated by transverse modes TEM_{mn}, with amplitudes decreasing rapidly with the mode order $m+n$. In the simplest case, the resonator consists of two identical concave mirrors of high reflectance ρ^2. The mirror separation can be chosen such that, while the resonator is tuned for maximum transmittance for the TEM_{00} mode, the low order transverse modes TEM_{mn} are almost totally suppressed, in amplitude by factors of the order of $1-\rho^2$. Considerations leading to a practical implementation are discussed, and experimental results are given.

1. Introduction

The gravitational wave detector being developed [1, 2] at the Max-Planck-Institut für Physik und Astrophysik is intended to measure the small strains $\delta L/L$ induced by gravitational radiation. The most promising sources are catastrophic events lasting only a few milliseconds, which transmit their largest spectral contributions in the range from a few hundred to a few thousand hertz. The goal is to reach a sensitivity $\delta L/L$ of about 10^{-21} so as to be able to detect events as far away as the Virgo cluster.

1.1. *A Michelson interferometer*

The detector is a laser-illuminated Michelson interferometer. Its total optical path length L has an optimum at half the wavelength of the gravitational radiation, say 100 km. One can realize such long paths by reflecting the interferometer beam N times (e.g. several hundred times) between two concave mirrors at distance l, as indicated in figure 1 for $N=4$ passes. But even then, variations δL in path difference of as little as 10^{-16} m have to be measured, i.e. some 10^{-10} of a wavelength λ of the

0030-3909/81/2805 0641 $02·00 © 1981 Taylor & Francis Ltd

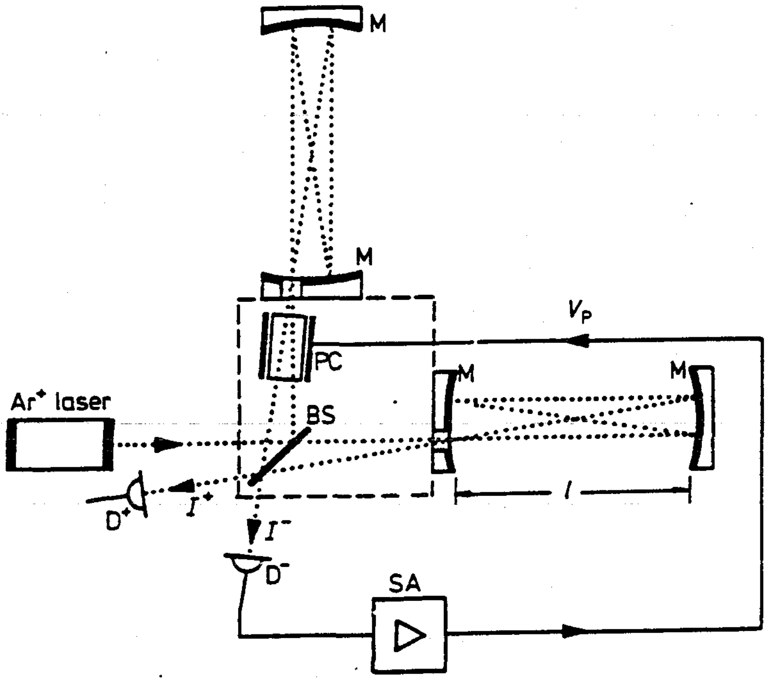


Figure 1. Michelson interferometer with folded light paths $L=Nl$, $N=4$. BS beam splitter; M mirror; PC Pockels cell; D^+, D^- photodiodes; SA servo amplifier.

illuminating laser light (Ar$^+$ laser, $\lambda=514\,\mathrm{nm}$). Throughout this paper, path differences δL will be expressed by the resulting phase differences

$$\phi=k\delta L=2\pi\delta L/\lambda, \tag{1.1}$$

with $k=2\pi/\lambda$ the propagation constant. In terms of these phase differences, our sensitivity goal ϕ_g is of the order of 10^{-9} rad. The interference of the recombined beams is monitored at one or both of the output ports by photodiodes D^+ and D^-, where, for parallel wavefronts, the light powers $P^{\pm}$ would be

$$P^{\pm}=\frac{P_0}{2}(1\pm\cos\Phi). \tag{1.2}$$

A chosen operating point Φ_0 is maintained via a servo loop, by applying a voltage $V_P(t)$ to a Pockels cell such that the change in optical phase $\chi(t)$ in the Pockels material compensates the changes in geometric phase $\phi_0(t)$:

$$\Phi_0=\phi_0(t)-\chi(t)=\text{const.} \tag{1.3}$$

One choice [3] of the operating point is to make the two output powers equal; the other [4], which we have adopted, uses a modulation technique, allowing operation at a power minimum of P^-, i.e. at

$$\Phi_0=0 \bmod 2\pi. \tag{1.4}$$

Both choices represent nulling methods, making the measured signal $\chi(t)$ very insensitive to fluctuations in the laser power.

1.2. *Local phase variations and field fluctuations*

This paper is concerned with the noise effects that can arise when—in contrast to the assumptions of equation (1.2)—the interfering wavefronts are not parallel, but rather arrive at the photodiode D$^-$ (x,y-plane) with a phase difference

$$\Phi(x,y;t)=\phi(x,y;t)-\chi(t) \tag{1.5}$$

that is a function of x and y, owing to imperfections in the interferometer. In the presence of such local phase variations $\phi(x,y;t)$, the Pockels signal $\chi(t)$ becomes a function of the (possibly time-dependent) field distributions $E_1(x,y;t)$, $E_2(x,y;t)$ of the two interfering beams.

Quite generally, the light power $P(t)$ striking the photodiode D$^-$ can be written as

$$P(t)=\frac{1}{2Z}\int\int|E_1-E_2\exp(i\Phi)|^2\,dxdy, \tag{1.6}$$

or, showing more clearly the phase dependence,

$$P(t)=\frac{1}{2Z}\int\int\{E_1^2+E_2^2-2E_1E_2\cos\Phi\}\,dxdy \tag{1·7}$$

(with $Z=\sqrt{(\varepsilon_0/\mu_0)}$ the vacuum impedance). For sufficiently narrow beams, the integration over the photodiode surface can be replaced by an integration over the infinite plane x,y.

The interferometer signal is given by the Pockels cell phase $\chi(t)$, which is servo-controlled in such a way as to make the power $P(t)$ on the photodiode a minimum. A necessary condition for a minimum is $\partial P/\partial\chi=0$, which leads to

$$\tan\chi(t)=\frac{\int\int E_1E_2\sin\phi\,dxdy}{\int\int E_1E_2\cos\phi\,dxdy}, \tag{1.8}$$

where E_1, E_2 and ϕ are functions of x, y and t. If $\chi_{\min}(t)$ is a solution leading to a power minimum, so are the solutions given by $\chi_{\text{even}}(t)=\chi_{\min}(t)\bmod 2\pi$, whereas the odd solutions $\chi_{\text{odd}}(t)=\chi_{\text{even}}(t)+\pi$ would lead to power maxima on the photodiode.

1.3. *The error signal*

We will narrow down the discussion to cases in which the geometric phase can be written

$$\phi(x,y;t)=\phi_0(t)+\tilde{\phi}(x,y), \tag{1.9}$$

i.e. separated into a main signal $\phi_0(t)$ and a stationary phase 'ripple' $\tilde{\phi}(x,y)$ that does not depend on the time t.

Such time-independent variations $\tilde{\phi}(x,y)$ in the geometric phase can result from misalignments of the beam splitter, or from mismatches between the optical components in the two interferometer arms. They can nevertheless lead to time-dependent spurious fluctuations $\delta\chi(t)$ in the compensating signal $\chi(t)$ if the interfering field strengths E_1 and E_2 are time-dependent functions $E_1(x,y;t)$, $E_2(x,y;t)$ of x and y.

By minimizing $P(t)$ of equation (1.7) with respect to the 'error signal'

$$\Delta\chi(t)=\chi(t)-\phi_0(t), \tag{1.10}$$

i.e. to the deviation of $\chi(t)$ from the main signal $\phi_0(t)$, one can derive, as an alternative form for (1.8),

$$\tan \Delta\chi(t) = \frac{\iint E_1 E_2 \sin \tilde{\phi}\, dx dy}{\iint E_1 E_2 \cos \tilde{\phi}\, dx dy}. \qquad (1.11)$$

If we make the assumption that the two interfering beams have identical intensity profiles $I(x, y; t)$, i.e. the fields E_1 and E_2 differ only in their phase $\phi(x, y; t)$, then one can substitute $I(x, y; t)$ for the product $E_1 E_2$ in equations (1.8) and (1.11).

With the further assumption that the phase ripple $\tilde{\phi}(x, y)$ does not change drastically inside the spot illuminated by $I(x, y; t)$, we arrive at a linearized solution

$$\Delta\chi(t) = \frac{\iint I(x, y; t)\tilde{\phi}(x, y)\, dx dy}{\iint I(x, y; t)\, dx dy}, \qquad (1.12)$$

the error signal $\Delta\chi(t)$ being represented by the original phase ripple $\tilde{\phi}(x, y)$ averaged over the photodiode surface with the locally fluctuating intensity $I(x, y; t)$ as weighting factor.

Finally, assuming that the intensity distribution $I(x, y; t)$ fluctuates by only a small amount $\delta I(x, y; t)$ around a constant time average $I_0(x, y)$, the fluctuation $\delta\chi(t)$ in the signal $\chi(t)$ becomes

$$\delta\chi(t) = \frac{\iint \delta I(x, y; t)\tilde{\phi}(x, y)\, dx dy}{\iint I_0(x, y)\, dx dy}. \qquad (1.13)$$

2. Fluctuations of laser beam geometry

In this section, two types of fluctuations in the beam intensity distribution are considered, both of which lend themselves to simple geometrical interpretations. It is only after the discussion of these two special cases that a more general view is given, as required for the understanding of the proposed mode selector scheme for reducing beam geometry noise.

2.1. *Lateral beam jitter*

Let us assume an incoming laser beam which, preserving its shape, moves laterally in the x-direction by a time-dependent distance $a(t)$: we have to substitute $I[x - a(t)]$ for $I(x)$. Furthermore, let us assume a tilt α_x between the two interfering wavefronts, then

$$\tilde{\phi}(x) = \alpha_x k x. \qquad (2.1)$$

Such a tilt could, for instance, be the result of a misalignment of the beam splitter by an angle $\alpha_x/2$. Within the approximation of equation (1.13), the spurious signal due to the excursions $a(t)$ becomes

$$\delta\chi = \frac{\int \{I(x-a) - I(x)\}\alpha_x k x\, dx}{\int I(x)\, dx}, \qquad (2.2)$$

which reduces, when the integration is taken from $-\infty$ to $+\infty$, to

$$\delta\chi(t) = \alpha_x k a(t). \qquad (2.3)$$

This seemingly trivial result would, of course, also have followed from a model considering only a single ray, displaced sideways by $a(t)$, but here it has been derived for an arbitrary intensity profile $I(x)$.

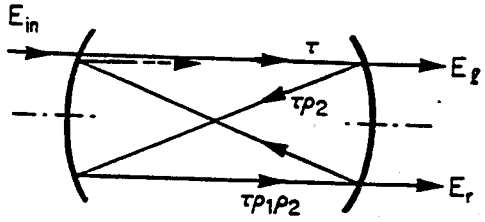


Figure 2. Confocal resonator used as beam symmetrizer.

The beam splitter can, in a rather straightforward procedure, be aligned such that the interfering wavefronts differ in tilt by less than $\alpha_x \simeq 10^{-5}$ rad. Adjusting and maintaining a tilt of 10^{-6} is about the best that can be done in our present set-up.

In the frequency range of interest (from a few hundred to a few thousand hertz), the laser beam exhibits lateral excursions $a(t)$ of the order of 10^{-9} m [2]. Although this is extremely small when compared with the width $2w$ of the laser beam ($2w \simeq 1{\cdot}5$ mm), the resulting noise $\delta\chi$ is thus one or two orders of magnitude above the eventual sensitivity goal of $\phi_g = 10^{-9}$ rad.

2.2. *The 'beam symmetrizer'*

The estimate of the noise contribution $\delta\chi(t)$ demonstrates clearly the need for reducing the lateral jitter $a(t)$ of the beam, before it enters the interferometer. A first step in that direction was the introduction of a 'beam symmetrizer' [5].

An optical resonator is placed into the light path, consisting of two confocal mirrors (figure 2), i.e. mirrors with a spacing d equalling their radius of curvature R. When tuned to the laser frequency, this resonator has a high transmittance for light components which are symmetrical with respect to its optical axis. A small lateral jitter $a(t)$ can be approximated by adding a small off-axis component $a(t)I'(x)$. In figure 2, the admixture of such an asymmetrical component is indicated schematically by a single incoming ray E_{in}. It splits up into two outgoing rays, symmetrical with respect to the optical axis, and with a ratio E_ℓ/E_r of their field-strengths which is given by the product $\rho^2 = \rho_1\rho_2$ of the field reflectivities ρ_1 and ρ_2 of the two mirrors.

The relative field strength difference $(E_\ell - E_r)/(E_\ell + E_r)$, and thus the beam's lateral excursion, appears reduced by a factor

$$\frac{1}{S_1} = \frac{1-\rho^2}{1+\rho^2} \simeq \frac{1-\rho^2}{2\rho}. \qquad (2.4)$$

With reflectances of $\rho^2 = 95$ per cent, the resulting suppression by a factor $S_1 \simeq 40$ was sufficient for our present requirements.

2.3. *Pulsation in beam width*

Another type of fluctuation in the intensity profile can be interpreted as a pulsation $\delta w/w$ in the beam width $2w$. For simplicity, let us consider a case where $I(x)$ is replaced by

$$(1+\kappa)I[(1+\kappa)x], \qquad (2.5)$$

i.e. a case where only the width in the x-direction is affected, and, moreover, the total power is conserved.

In the frequency window from 500 Hz to a few kilohertz, our Ar$^+$ laser exhibited relative width variations $\delta w/w = \kappa(t)$ of the order of 10^{-6} [5].

The symmetrical beam pulsation (2.5) is an even function of x, and it leads to spurious signals $\delta\chi(t)$ only if the phase difference $\tilde{\phi}(x)$ between the wavefronts also includes even functions of x, in the simplest case

$$\tilde{\phi}(x) = cx^2. \tag{2.6}$$

An appropriate measure for the curvature of the phase ripple $\tilde{\phi}(x)$ is cw^2, for which an upper limit can be gained from a measurement of the light power at interference minimum. It was found to be of the order of 10^{-1} rad.

The spurious signal

$$\delta\chi = \frac{\int \{(1+\kappa)I[(1+\kappa)x] - I(x)\} cx^2\, dx}{\int I(x) dx} \tag{2.7}$$

becomes (neglecting a term in κ^2)

$$\delta\chi = -2\kappa c w^2 v^2, \tag{2.8}$$

where the numerical factor

$$v^2 = \frac{\int I(x) x^2\, dx}{w^2 \int I(x)\, dx} \tag{2.9}$$

depends on the model assumed for $I(x)$, but not very strongly. For an $I(x)$ constant from $-w$ to $+w$, one obtains $v^2 = 1/3$; for a gaussian, down to e^{-2} at $x = w$, v^2 would be 1/2; and for the slowly decreasing profile $I(x) = [1 + (x/w)^2]^{-2}$, we have $v^2 = 1$.

With the empirical data for $\kappa = \delta w/w$ and cw^2 as quoted above, equation (2.8) gives a noise contribution of about one power of 10 above the sensitivity goal of $\phi_g = 10^{-9}$. So the beam pulsation also calls for a suppression with similar effectiveness as we had been able to provide in the case of the lateral beam jitter.

The confocal optical resonator used there does not suppress the pulsations in beam width. We are particularly indebted to Dr. A. Brillet for pointing out that, by a different choice of mirror separation, the extreme degeneracy of the confocal case is lifted, and a strong suppression both of the lateral jitter and of the width pulsation is made possible.

The following sections will treat, in a more general way, the characteristics of spherical optical resonators, and how they can be used to suppress deviations from the ideal shape of a laser beam.

3. Modes in optical resonators

The view taken in this section can be summarized in the following manner. The laser beam, on its way to the interferometer, passes through a high-Q optical resonator. As viewed from this resonator, an incident field distribution $E(x, y; t)$ can be expanded into a series of eigenmodes $e_{mn}(x, y)$ with (possibly time-dependent) amplitudes $a_{mn}(t)$.

The desired fundamental mode $e_{00}(x, y)$ is a gaussian beam centred on the resonator's optical axis. By adjusting the mirror distance, the resonator will be tuned so that this fundamental mode has maximum transmission. The higher modes e_{mn}, describing the unwanted perturbations of the gaussian beam, will in general not be in resonance, and will thus be almost totally suppressed.

3.1. *Theory of optical resonators*

The theory of confocal optical resonators was first expounded simultaneously by Fox and Li [6] and by Boyd and Gordon [7], and later generalized by Boyd and Kogelnik [8].

Some formulae used below are taken from reviews and applications [9–11], but in general they are easily derived from the original representations.

3.2. *Notation*

The notation used by the authors quoted is not consistent, so we will start off by defining some of the symbols to be used below.

The optical resonator is composed of two highly reflective mirrors, separated by a distance d, and usually facing each other with their concave surfaces (figure 3 (*a*)). Their radii ofcurvature are denoted by R_1 and R_2 (or R if they are identical). As an important special case we have the confocal resonator, said to have the characteristic length b, if distance d and curvature radius R have a common value b (see figure 3 (*b*)).

The field propagates in the z-direction, i.e. in the optical axis interconnecting the centres of curvature of the two mirrors. The transverse directions are denoted by x and y, and the distance from the optical axis by $r=\sqrt{(x^2+y^2)}$. The lateral field distribution $E(x,y)$ of the beam will be expressed by normal modes which all have a common gaussian factor $\exp(-r^2/w^2)$. This factor is down to $1/e$ at the characteristic half-width w which is a function of z. Sometimes it is convenient to use the normalized transverse coordinates.

$$\xi=\frac{x}{w}, \quad \eta=\frac{y}{w}. \tag{3.1}$$

The mirror diameters are considered to be large compared with the beam widths.

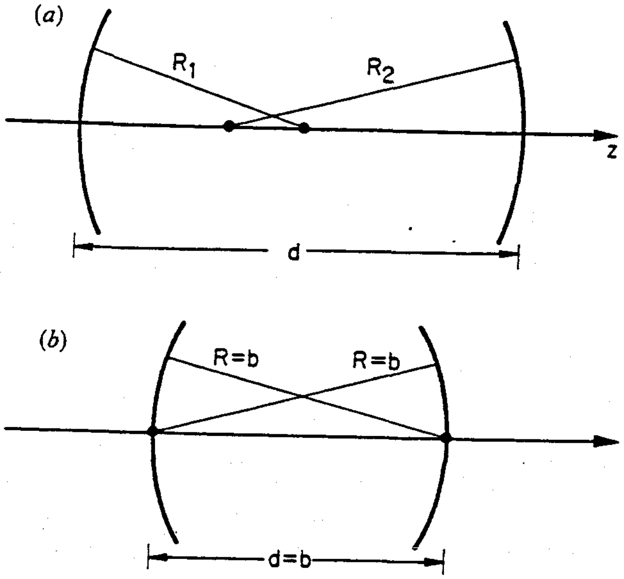


Figure 3. Optical resonators: (*a*) general configuration; and (*b*) confocal case.

3.3. *Modes in the confocal resonator*

Starting out from Huygens's principle, Boyd and Gordon [7] have determined self-consistent field patterns $E(x, y)$ such that a given field distribution on one mirror reproduces itself on the opposite mirror, allowing only a constant (possibly complex) factor. The problem can be formulated as an integral equation which, in the limit of infinite mirror diameters, is solved by normal modes $e_{mn}(x, y)$ that can be expressed as a product

$$e_{mn} \propto h_m(\xi) h_n(\eta) \tag{3.2}$$

of two gaussian–hermite functions of the type

$$h_m(\xi) = \frac{\Gamma(\frac{1}{2}m+1)}{\Gamma(m+1)} H_m(\sqrt{2}\xi) \exp(-\xi^2), \tag{3.3}$$

where the arbitrary factor $\Gamma(\frac{1}{2}m+1)/\Gamma(m+1)$ was chosen such that for even orders $m=2j$ we have $|h_{2j}(0)|=1$. The hermite polynomials H_m of the argument $\sqrt{2}\xi$ are defined by

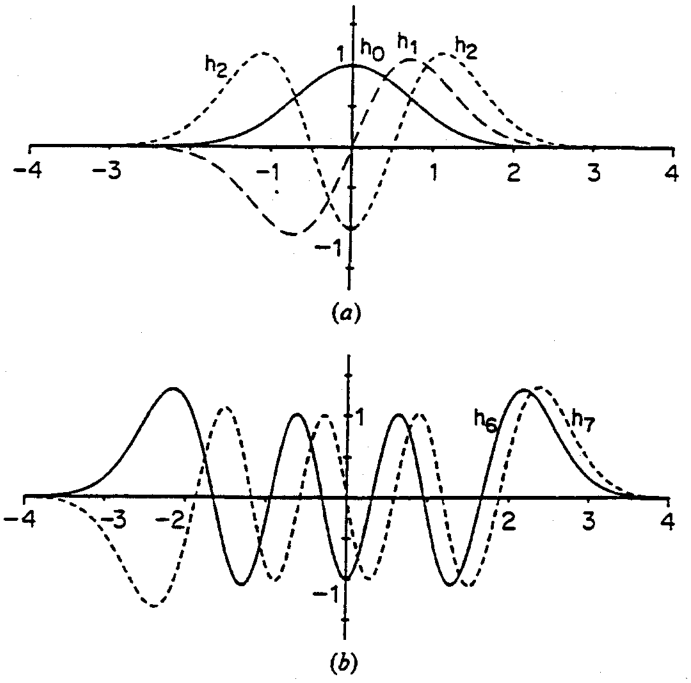
$$H_m(\sqrt{2}\xi) = 2^{-m/2} \exp(2\xi^2) \left(\frac{-d}{d\xi}\right)^m \exp(-2\xi^2). \tag{3.4}$$

Figure 4 (*a*) shows the first three of the gaussian–hermite functions,

$$h_0(\xi) = \exp(-\xi^2); \quad h_1(\xi) = \sqrt{(2\pi)}\xi \exp(-\xi^2); \quad h_2(\xi) = (4\xi^2 - 1) \exp(-\xi^2),$$

and figure 4 (*b*), as examples of higher orders, $h_6(\xi)$ and $h_7(\xi)$.

When we add to the purely gaussian ground mode $h_0(\xi)$ a small contribution of $h_1(\xi)$ (of relative amplitude a_1) this is equivalent to a lateral shift of the beam axis by



(*a*)

(*b*)

Figure 4. Gaussian–hermite functions: (*a*) $h_0(\xi)$, $h_1(\xi)$, $h_2(\xi)$; and (*b*) $h_6(\xi)$, $h_7(\xi)$.

$\delta\xi=\delta x/w\sqrt{(\pi/2)}a_1$. A small contribution of $h_2(\xi)$ (amplitude a_2), on the other hand, leads to a change in the beam width by $\delta w/w=2a_2$. The phenomena of lateral beam jitter and pulsation in the beam width, discussed in the previous sections (§§ 2.1 and 2.3), are thus expressed as contaminations by low-order transverse modes.

3.4. *Non-confocal resonators*

From the field distributions (3.2) on the confocal mirrors, the field distribution at any value z can be calculated, again using Huygens' principle. With the use of the gaussian–hermite functions introduced above, Boyd and Gordon's equation (20) becomes very simple:

$$v_{mn}(x,y;z)=\frac{w_0}{w(z)}h_m\left(\frac{x}{w(z)}\right)h_n\left(\frac{y}{w(z)}\right)$$
$$\times\exp\left(-i\left\{kz+(1+m+n)\psi(z)+\frac{kr^2}{2R(z)}\right\}\right),\qquad(3.5)$$

with $w(z)$, $R(z)$ and $\psi(z)$ to be explained below.

If for the moment we ignore the phase factor, we see that the modes maintain their general appearance, only that all lateral dimensions of the field distribution are scaled by a common factor $w(z)$. At the waist of the beam, at $z=0$, the gaussian

$$\exp(-\xi^2-\eta^2)=\exp(-r^2/w^2),$$

included in $h_m(\xi)h_n(\eta)$, has a $1/e$ radius w given by

$$w_0=\sqrt{\frac{b}{k}}=\sqrt{\left(\frac{b\lambda}{2\pi}\right)}.\qquad(3.6)$$

The minimum width w_0 is uniquely determined by the wavelength λ and the characteristic length b of the confocal resonator. As indicated in figure 5, the beam radius w increases with the distance z from the central plane according to

$$w(z)=w_0\sqrt{(1+\zeta^2)}.\qquad(3.7)$$

Here, we have introduced the normalized longitudinal coordinate

$$\zeta=\frac{2z}{b}\qquad(3.8)$$

to describe the propagation along the optical axis (z-axis), counting from the central plane.

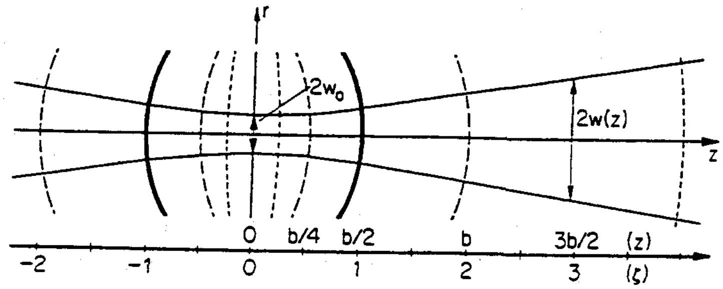


Figure 5. Gaussian beam and surfaces of constant phase.

The surfaces of constant phase have radii of curvature

$$R(z)=\frac{1+\zeta^2}{2\zeta}b=\left(\zeta+\frac{1}{\zeta}\right)\frac{b}{2}, \tag{3.9}$$

having a minimum ($R=b$) at the confocal distance $z=\pm b/2$ ($\zeta=\pm 1$), and approaching infinity (plane wave) at the beam waist ($z=0$), as well as at infinite distance ($z=\pm\infty$).

A gaussian beam with given propagation constant $k=2\pi/\lambda$ is fully determined for all values of z if at some distance z_0 from the beam's waist any two out of the (five) following variables are known:

$$b;\quad z_0;\quad \zeta_0=\frac{2z_0}{b};\quad R(z_0)=\left(\zeta_0+\frac{1}{\zeta_0}\right)\frac{b}{2};\quad w(z_0)=\sqrt{\left[\frac{b}{k}(1+\zeta_0^2)\right]}. \tag{3.10}$$

The mode pattern remains unchanged if one places two appropriately curved mirrors at any two of the surfaces of constant phase. The field distribution in the non-confocal resonator thus formed can be derived from an equivalent confocal resonator of characteristic length b, where b can be determined from R_1, R_2 and d with the relation

$$b^2=\frac{4d(R_1-d)(R_2-d)(R_1+R_2-d)}{(R_1+R_2-2d)^2}. \tag{3.11}$$

The condition that b^2 be positive describes the region of stable configurations, i.e. configurations in which the beam is being refocused on successive round trips, rather than becoming progressively more divergent.

3.5. *Mode selective Fabry–Pérot resonator*

The resonator can be treated as a Fabry–Pérot, in which the field, on each return trip $z_1\to z_2\to z_1$, is weakened by a factor $\rho_1\rho_2$, when ρ_1^2 and ρ_2^2 are the power reflectances of the two mirrors. We will from now on assume mirrors of equal reflectances, ρ^2, and of common transmittances τ^2.

In contrast to plane-wave optics, the light phase ϕ varies with z (or with ζ) not in a linear fashion $\phi(z)=kz$, but rather in the form

$$\phi(z)=kz+(1+m+n)\psi(z), \tag{3.12}$$

with

$$\psi(z)=\arctan\zeta. \tag{3.13}$$

The additional phase contribution $\Delta\phi=(1+m+n)\psi(z)$ has a high rate of change in a region around $z=0$, extending, say, out to $z=\pm b$.

What is important in our application is that $\Delta\phi$ depends on the mode order $N=m+n$, i.e. on the sum of the hermite indices m and n. This will allow us to discriminate between modes that have different orders N.

The resonator can be tuned, by fine adjustment of the mirror distance $d=z_1-z_2$, such that the ground mode ($N=0$) reproduces itself with zero phase difference

$$\Delta\phi_0=2kd+2(\psi_1-\psi_2)=0 \bmod 2\pi \tag{3.14}$$

on one return trip $z_1\to z_2\to z_1$.

For the higher modes $m+n>0$, the successive field contributions being superimposed are then out of phase by an angle

$$\Delta\phi_N=(m+n)2(\psi_1-\psi_2)=N2\Psi. \tag{3.15}$$

Here, Ψ stands for the single-trip phase difference

$$\Psi=\psi_1-\psi_2=\arctan\zeta_1-\arctan\zeta_2. \tag{3.16}$$

By adding up the field contributions one derives the fraction T_N of the incident power transmitted by this Fabry–Pérot resonator as

$$T_N=\frac{\tau^4}{(1-\rho^2)^2}\,\frac{1}{1+\left(\dfrac{2\rho}{1-\rho^2}\sin N\Psi\right)^2}, \tag{3.17}$$

which we will call the (power) 'throughput' to distinguish it from the single-mirror transmittance τ^2.

4. Choice of resonator parameters

In this section we will discuss some of the considerations in choosing the parameters of an optical resonator so that it best utilizes the mode selectivity expressed in equation (3.17).

4.1. *Mirror properties*

In equation (3.17), the first term $T_0=\tau^4/(1-\rho^2)^2$ is the power throughput at resonance. It would reach unity only for ideal mirrors in which power reflectance ρ^2 and power transmittance τ^2 add up to unity: $\rho^2+\tau^2=1$. In the presence of losses σ^2 (absorption or scattering), with $\rho^2+\sigma^2+\tau^2=1$, one can express T_0 by

$$T_0=\frac{1}{(1+\sigma^2/\tau^2)^2}. \tag{4.1}$$

For a high throughput T_0 of the ground mode, the losses σ^2 have to be small compared with the mirror transmittance τ^2, which in itself is a small quantity in our high-Q resonator. With a commercial resonator (Tropel, $\rho^2=95$ per cent), we were able to obtain a throughput T_0 of about 75 per cent, but better values are technically realizable.

The relative suppression of the non-fundamental modes ($N>0$) is determined by the second term in equation (3.17). We shall define it by the reverse square root of this second term:

$$S_N=\sqrt{\left[1+\left(\frac{2\rho}{1-\rho^2}\sin N\Psi\right)^2\right]}. \tag{4.2}$$

Unless the phase angle $N\Psi$ is extremely close to a resonance $N\Psi_0=0 \bmod \pi$, the suppression (in field amplitude) is

$$S_N\simeq\frac{2\rho}{1-\rho^2}\sin N\Psi=S_{\max}\,s_N.$$

With mirrors having a reflectance ρ^2 of 95 per cent, a maximum suppression of $S_{\max}=2\rho/(1-\rho^2)\simeq 40$ can be obtained, which is considered sufficient for our current needs (§ 2.2).

Should higher suppression ratios $S_{\max}$ be required, one can either increase the reflectance ρ^2 of the mirrors, or use two resonators in series. In both cases one will have to expect a deterioration in the throughput T_0 of the fundamental mode.

4.2. *The phase angle* Ψ

Perturbing modes of the orders $N=1$ and $N=2$ have a simple geometric interpretation, which allowed a direct measurement (cf. §§ 2.1 and 2.3). Even though their field strengths are about six powers of 10 below the fundamental mode, they nevertheless give rise to intolerable error signals. Thus, in the choice of an appropriate phase angle Ψ, one has to make sure that the modes with $N=1$ and $N=2$ are suppressed sufficiently, i.e. one wants $\sin\Psi$ and $\sin 2\Psi$ to be sufficiently far away from zero.

Modes of order higher than $N=2$ represent rather complicated patterns. It can be assumed that they are contained in the incoming beam with amplitudes that decrease rapidly with increasing N. Furthermore, in the interferometer they will become effective only in combination with a corresponding phase ripple of the interfering wavefronts.

The need to suppress modes of medium order ($N=3, 4$) has not yet been clearly established. But a suppression would be welcome, and it can be had at little extra expense by an appropriate choice of Ψ.

Using (3.16) and (3.8) and well-known addition theorems, one can express Ψ by

$$\cos\Psi=\sqrt{\left[\left(1-\frac{d}{R_1}\right)\left(1-\frac{d}{R_2}\right)\right]}. \qquad (4.3)$$

The condition that the radicand be positive, and not exceed 1, leads to the same stability criteria as those derived from equation (3.11).

Figure 6 shows plots of $s_N=\sin N\Psi$ for $N=1,\ldots,4$ (for $N=5,6$, only the zero crossings are indicated). From these curves (special cases of Lissajous' traces) one can pick values $\cos\Psi$ for which s_1 and s_2 have high values (above 0·5, say), and for which none of the further traces is very close to zero.

4.3. *Symmetrical resonator configurations*

In practical applications one prefers the symmetric configuration, i.e. resonators with mirrors of equal radii of curvature. For brevity, we will restrict the following discussion to this special case, in which (4.3) becomes

$$\cos\Psi=1-\frac{d}{R}=1-\delta. \qquad (4.4)$$

Only $\delta=d/R$, the ratio between the mirror separation d and the radius of curvature, R, enters into Ψ.

In figure 6, the lower axis is denoted by this relative mirror separation $\delta=d/R=1-\cos\Psi$. Beneath it some representative resonator configurations are indicated.

The traces are symmetrical with respect to $\cos\Psi=0$, i.e. to the confocal case $d=R$. Here, the odd modes undergo maximum suppression $S_{\max}=2\rho/(1-\rho^2)$, as all $|s_{2j+1}|=1$. The even modes, however, pass the Fabry–Pérot with no extra suppression, as all even-order traces s_{2j} vanish at $\cos\Psi=0$.

One has to go rather far in either direction, before the most important of the even modes, $N=2$, shows an appreciable suppression, e.g. to $|\cos\Psi|>0{\cdot}126$ for $|s_2|>0{\cdot}25$.

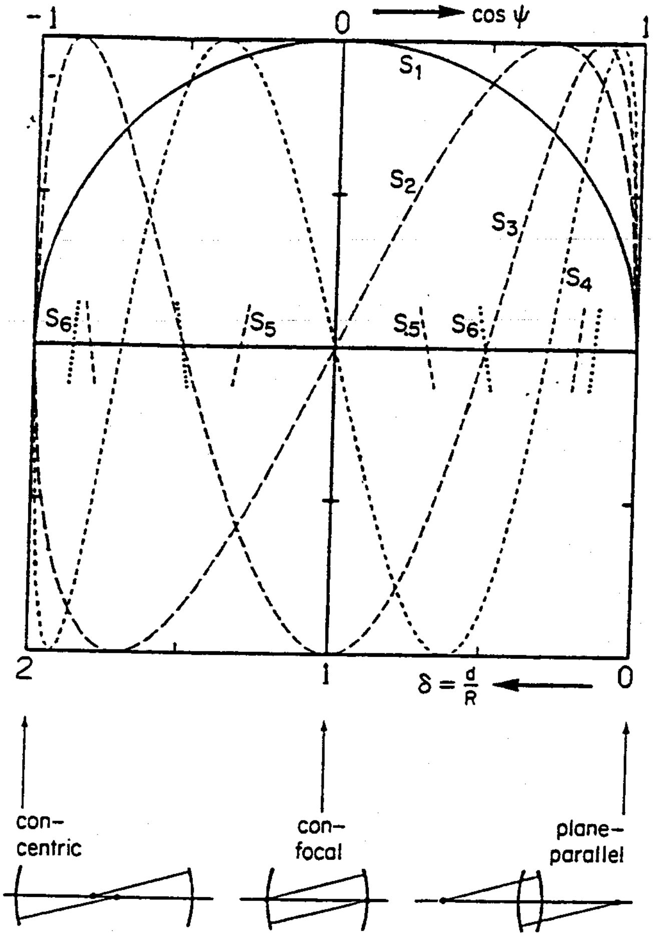


Figure 6. Coefficients $s_N = \sin N\Psi$, plotted versus $\cos\Psi$ (upper axis), or versus $\delta = 1 - \cos\Psi$ (lower axis). Three representative mirror configurations are indicated.

‾he plane-parallel limit ($\cos\Psi = 1$, $\delta = 0$) and the concentric limit ($\cos\Psi = -1$, $\delta = 2$) are degenerate with respect to all orders, so here we would have no mode selection at all. The suppression of the most important perturbing mode ($N=1$) increases most slowly with departure from these limiting cases, but rapidly enough to assume values $s_1 > 0{\cdot}5$ in the range from $\cos\Psi = -0{\cdot}866$ to $\cos\Psi = +0{\cdot}866$ ($\delta = 0{\cdot}134$ to $\delta = 1{\cdot}866$).

Configurations that come close to the plane-parallel or concentric limits should be avoided anyway, for reasons of mechanical stability. Small tilts of the resonator mirrors would cause the optical axis to deviate strongly in its position (plane-parallel case) or in its direction (concentric).

Vithin the limits dictated by s_1 and s_2, one can easily pick configurations which pro.ide reasonable suppression also for the medium-order modes.

4.4. *Mode matching*

In order to achieve a high throughput T_0, the beam emerging from the laser must be matched to the fundamental mode of the resonator. This matching can, for instance, be made with the help of a single adaptation lens as shown in figure 7.

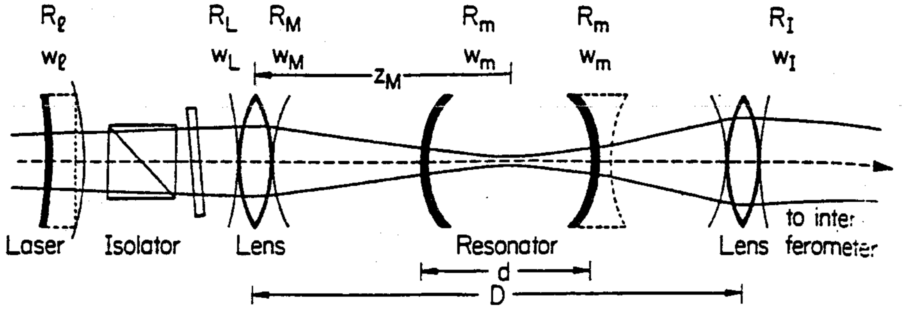


Figure 7. Matching of resonator modes (R_m, w_m) to given laser beam (R_ℓ, w_ℓ) and requirements of the interferometer (R_I, w_I).

In calculations for this mode matching, repeated use must be made of the interrelations (3.10) between the characteristic variables of a gaussian beam.

From the configuration of the laser mirrors, one can derive the spot diameter $2w_\ell$ on the output mirror (with curvature radius R_ℓ). After passing through the mirror substrate and an optical isolator, the beam arrives with w_L and R_L at the adaptation lens.

In a similar fashion, one can trace the gaussian mode of the resonator (characterized by R_1, R_2, d) back to the position of the lens, and one must choose the distance z_M from waist to lens such that the mode radius w_M at the lens matches the actual beam radius w_L.

This choice of z_M also determines the radius of curvature, R_M. The matching in the curvature of the beams is provided by a lens with a focal length f given by

$$\frac{1}{f}=\frac{1}{R_L}+\frac{1}{R_M}. \tag{4.5}$$

In case of a mismatch, either in the curvature radii R of the phase fronts, or in the beam widths $2w$, the incoming field distribution $E(x, y)$, expanded with respect to the resonator modes, will have appreciable contributions in the non-fundamental modes ($N>0$). These extra modes are almost totally reflected, the loss in power throughput being proportional to the square of the relative mismatch $\delta R/R$ or $\delta w/w$ respectively. This loss becomes negligible when the relative mismatch is below, say, 10 per cent.

4.5. *Power density on optical surfaces*

In our application, we want to transmit high laser powers P, at present about 1 W, and perhaps up to 100 W in the eventual experiment. Intensities (power densities) $I=P/\pi w^2$ above an $I_{\max}$ in the order of 10 W/mm^2 can inflict thermal damage on the multi-layer coatings. So we have to keep the spot area πw^2 on the mirrors above

$$\pi w_m^2=\frac{P}{I_{\max}}. \tag{4.6}$$

For mirrors of a given curvature radius R, the equations (3.7) and (3.6) define a characteristic spot radius

$$W_R=\sqrt{\left(\frac{2R}{k}\right)}=\sqrt{\left(\frac{R\lambda}{\pi}\right)}, \tag{4.7}$$

describing the spot size on the mirrors in a confocal arrangement ($d=R, \zeta=1$).

The intensity in such a confocal arrangement would be

$$I_R = \frac{P}{R\lambda}. \tag{4.8}$$

For the same mirrors in a non-confocal configuration ($d \neq R$), one can derive from (3.7) and (3.9) that the ratio of actual spot area πw^2 to the 'confocal' spot area πW_R^2 is

$$\frac{w^2}{W_R^2} = \zeta = \frac{d}{b}. \tag{4.9}$$

Thus, only configurations with sufficiently high

$$\zeta > \zeta_{\text{min}} = \frac{P}{R\lambda I_{\text{max}}} \tag{4.10}$$

are allowed, i.e. configurations with relative mirror separations

$$\delta_{\text{min}} = \frac{d_{\text{min}}}{R} > \frac{2\zeta_{\text{min}}^2}{1+\zeta_{\text{min}}^2}. \tag{4.11}$$

In an example with $P=1$ W, $I_{\text{max}}=5$ W/mm^2, $R=0{\cdot}1$ m, $\lambda=0{\cdot}514\times10^{-6}$ m, we have $\zeta_{\text{min}}=3{\cdot}9$. The resulting configuration has a relative mirror separation of $d_{\text{min}}/R=1{\cdot}88$, which is about as close to the concentric limiting case as one would dare to go, from the considerations of mode suppression and mechanical stability discussed in § 4.3. With mirrors of curvature radius less than $R_{\text{min}}=0{\cdot}1$ m, one would no longer be able to reconcile these requirements.

4.6. *Structural length*

A further consideration in the implementation of a mode selector is the length d of the resonator and the overall length D required for mode matching.

If, from considerations of mode suppression and power density, minimum values for $s_1=\sin\Psi$ and spot diameter $2w_{\text{m}}$ are given, the mirror spacing d_{min} of a (symmetrical) resonator becomes

$$d_{\text{min}} = \frac{kw_{\text{m}}^2}{2} s_1. \tag{4.12}$$

With $w_{\text{m}}=0{\cdot}25$ mm and $s_1=0{\cdot}5$, we arrive at a resonator of length $d=0{\cdot}19$ m, to be realized with mirrors of a curvature radius of either $R_{\text{c}}=0{\cdot}10$ m (near concentric case), or $R_{\text{p}}=1{\cdot}43$ m (near plane-parallel).

For an estimate of the overall length D, including the mode-matching optics, we have to make a few further assumptions. In our application, the beam coming from the laser has similar characteristics ($R_{\text{L}}, w_{\text{L}}$) as those required for the Michelson interferometer ($R_{\text{I}}, w_{\text{I}}$). In a completely symmetric configuration, the overall length D equals $2z_{\text{M}}$ (see figure 7).

If only one mode matching lens each (at $z=\pm z_{\text{M}}$) is used, the overall length can be approximated by

$$D \simeq \frac{kw_{\text{m}}w_{\text{L}}}{2} s_1. \tag{4.13}$$

With $w_{\text{m}}=0{\cdot}25$ mm, $w_{\text{L}}=0{\cdot}75$ mm and $s_1=0{\cdot}5$, the total length D would be $0{\cdot}57$ m. This length can be reduced by the use of a diverging lens near the resonator mirrors.

By using plane-concave or symmetric bi-concave substrates for the resonator mirrors (the latter case being indicated in figure 7 for the right-hand resonator mirror) one can reduce D to 0·48 and 0·40 m respectively.

In our current gravitational wave experiment, bi-concave mirror substrates are going to be used, both radii of curvature being 0·10 m. The adaptation lenses will then need focal lengths of 0·15 m.

4.7. *Performance*

The suppression of the low-order modes was tested with a commercial optical resonator (Tropel, $R=50$ mm), at reduced laser powers to avoid damage to the optical coatings. The transmittance τ^2 of the mirrors is approximately 6 per cent, and to explain the resonance throughput of only 50 per cent we have to assume losses σ^2 of 2·5 per cent. This surprisingly high value may be due to an inadvertent thermal damage. With the resulting reflectance of $\rho^2=91·5$ per cent we find a maximum suppression $S_{\max}=2\rho/(1-\rho^2)$ of 22·5.

The resonator was operated at its maximum mirror spacing of $d=61$ mm, a configuration rather close to the confocal case, with a first-order coefficient $s_1=0·975$, and reasonable coefficients s_N for the next few orders N.

As pointed out in § 3.3, an admixture of a first-order mode (of relative field amplitude a_1) results in a lateral beam jitter (by $\delta x/w=1·25a_1$). This jitter can be measured with a position-sensitive photodiode. Two plots of the spectral distribution of this jitter are shown in figure 8, with and without the mode selector. The spikes near 1·8 kHz represent an artificially introduced lateral jitter which clearly confirms the expected suppression by $s_1S_{\max}=22$. The continuous spectrum with mode selector (lower trace) is dominated by additional perturbations such as shot noise and harmonics of the 50 Hz mains, making a judgement of the suppression impossible. For the modes of order $N=2$ (the width pulsation), and even more for the higher modes, measurement as well as artificial generation become much more difficult. The suppression of such higher modes has been checked only qualitatively, by observing that despite gross variations in the incoming beam shape (shadowing by masks) the outgoing beam maintained a sufficiently gaussian appearance of practically constant width.

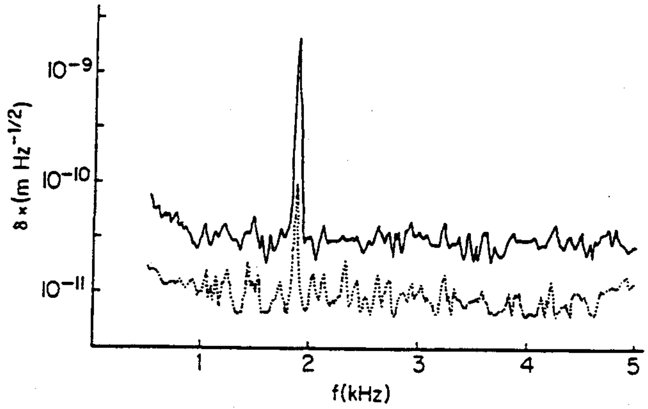


Figure 8. Spectrum of lateral beam jitter, without (upper trace) and with (lower trace) mode selection.

Notre développement d'un détecteur d'onde gravitationnelle nécessite un interféromètre
de Michelson d'extrême sensibilité capable de mesurer 10^{-6} m (c'est-à-dire quelques 10^{-10}
de la longueur d'onde de la lumière d'éclairage). Même après des réglages délicats des pièces
composant l'interféromètre, et après de considérables améliorations de la stabilité du laser, des
bruits très supérieurs au but à atteindre ont été observés, qui provenaient des fluctuations de la
géométrie du faisceau laser. Les deux principaux types de fluctuations géométriques du
faisceau sont une oscillation latérale et une pulsation en largeur du faisceau; ceci conduit à des
signaux d'interféromètre perturbés si les fronts d'onde qui interfèrent sont mal réglés en
direction et en courbure respectivement.

La géométrie du faisceau peut être considérablement stabilisée en le faisant passer au
travers d'un résonateur optique. Les fluctuations géométriques du faisceau, vues de ce
résonateur, peuvent être considérées comme un mode fondamental bien centré TEM$_{00}$, altéré
par des modes transverses TEM$_{mn}$, dont les amplitudes décroissent rapidement avec les ordres
m et n. Le résonateur, dans le cas le plus simple, consiste en deux miroirs concaves
identiques de facteur de réflexion élevé ρ^2. La distance des miroirs peut être choisie de manière
que, lorsque le résonateur est accordé pour que la transmission soit maximale pour le mode
TEM$_{00}$, les modes transverses TEM$_{mn}$ d'ordres inférieur soient presque totalement
supprimés leurs amplitudes étant réduites par un facteur de l'ordre de $1-\rho^2$. Des
considérations conduisant à une mise en oeuvre pratique sont discutées et des résultats
expérimentaux sont donnés.

Unsere Entwicklung eines Gravitationswellen-Detektors erfordert ein Michelson-
Interferometer mit einer Empfindlichkeit bis herab zu 10^{-16} m, also etwa 10^{-10} einer
Wellenlänge λ des beleuchtenden Laserlichts. Selbst nach sorgfältiger Ausrichtung der
Interferometer-Komponenten, und nach erheblichen Verbesserungen in der Laserstabilität,
wurden noch Störsignale beobachtet, die weit über dieser angestrebten Empfindlichkeit
lagen. Diese Störsignale rühren u.a. von Schwankungen der Laserstrahl-Geometrie. Zwei
typische Beispiele für solche Schwankungen sind (a) ein seitliches Wackeln der Strahl-
Position, und (b) ein Pulsieren des Strahl-Durchmessers; sie führen zu Störsignalen im
Interferometer, wenn die interferierenden Wellenfronten fehlangepaßt sind bezüglich ihrer
Richtung (a), bzw. bezüglich ihrer Krümmung (b).

Die geometrische Stabilität kann wesentlich verbessert werden, wenn man den
Laserstrahl einen optischen Resonator durchlaufen läßt. Aus der Sicht dieses Resonators
können die Geometrie-Schwankungen beschrieben werden durch einen wohlzentrierten
Grund-Modus TEM$_{00}$, verunreinigt durch höhere Moden TEM$_{mn}$, mit Amplituden, die
rasch mit zunehmender Ordnung $m+n$ abfallen. Der Resonator besteht im einfachsten Fall
aus zwei identischen Hohlspiegeln hoher Reflektivität ρ^2. Der Spiegelabstand läßt sich so
wählen, daß—bei Abstimmung auf maximale Transmission für den Grund-Modus TEM$_{00}$—
die Transversal-Moden TEM$_{mn}$ niedriger Ordnung $m+n$ fast vollständig unterdrückt
werden, nämlich um Amplituden-Faktoren der Größenordnung $1-\rho^2$. Gesichtspunkte, die
bei einer praktischen Ausführung zu beachten sind, werden diskutiert, und experimentelle
Ergebnisse werden wiedergegeben.

References

[1] Winkler, W., 1977, *Proceedings of the International Symposium on Experimental Gravitation*, Pavia (Rome: Accademia Nazionale dei Lincei), pp. 351–363.

[2] Billing, H., Maischberger, K., Rüdiger, A., Schilling, R., Schnupp, L., and Winkler, W., 1979, *J. Phys.* E, **12**, 1043.

[3] Moss, G. E., Miller, L. R., and Forward, R. L., 1971, *Appl. Optics*, **10**, 2495.

[4] Weiss, R., 1972, *Q. Prog. Rep. Res. Lab. Electron., M.I.T.*, **105**, 54.

[5] Maischberger, K., Rüdiger, A., Schilling, R., Schnupp, L., Winkler, W., and Billing, H., 1980, *Proceedings of the Second Marcel Grossmann Meeting*, Trieste, edited by R. Ruffini (in the press); preprint MPI-PAE/Astro 209.

[6] Fox, A. G., and Li, T., 1961, *Bell Syst. tech. J.*, **40**, 453.

[7] Boyd, G. D., and Gordon, J. P., 1961, *Bell Syst. tech. J.*, **40**, 489.

[8] Boyd, G. D., and Kogelnik, H., 1962, *Bell Syst. tech. J.*, **41**, 1347.

[9] Röss, D., 1969, *Lasers, Light Amplifiers and Oscillators* (London, New York: Academ Press).

[10] Abderrazik, J.-E., 1967, *Annls Télécommun.*, **22**, 41.

[11] Fork, R. L., Herriot, D. R., and Kogelnik, H., 1964, *Appl. Optics*, **3**, 1471.

REF. :

Optical Electronics

Fourth Edition

Amnon Yariv
California Institute of Technology

HRW
SAUNDERS

Saunders College Publishing
A Division of Holt, Rinehart and Winston, Inc.
Philadelphia Chicago Fort Worth San Francisco
Montreal Toronto London Sydney Tokyo

Copyright © 1991, 1985, 1976, 1971 by Saunders College Publishing, a division of Holt, Rinehart and Winston, Inc.

All rights reserved. No part of this publication may be reproduced or transmitted in a form or by any means, electronic or mechanical, including photocopy, recording, or a information storage and retrieval system, without permission in writing from the publi

Requests for permission to make copies of any part of the work should be mailed to: Permissions Department, Holt, Rinehart and Winston, Inc., 8th floor, Orlando, Florida 32887.

Text Typeface: Times Roman
Compositor: General Graphic Services
Senior Acquisitions Editor: Barbara Gingery
Managing Editor: Carol Field
Project Editor: Martha Brown
Manager of Art and Design: Carol Bleistine
Art Director: Christine Schueler
Art and Design Coordinator: Doris Bruey
Text Designer: Rita Naughton
Cover Designer: Lawrence R. Didona
New Text Artwork: Vantage Art
Director of EDP: Tim Frelick
Production Manager: Bob Butler
Marketing Manager: Denise Watrobsky

Printed in the United States of America

OPTICAL ELECTRONICS, 4e

0-03-047444-2

Library of Congress Catalog Card Number: 90-050825

1234 039 987654321

THIS BOOK IS PRINTED ON ACID-FREE, RECYCLED PAPER

Noise in Optical Detection and Generation

10.0 INTRODUCTION

In this chapter we study the effect of noise in a number of important physical processes. We will take the term noise to represent random electromagnetic fields occupying the same spectral region as that occupied by some "signal." The effect of noise will be considered in the following cases.

1. *Measurement of optical power.* In this case the noise causes fluctuations in the measurement, thus placing a lower limit on the smallest amount of power that can be measured.
2. *Linewidth of laser oscillators.* The presence of incoherent spontaneous emission power will be found to be the cause for a finite amount of spectral line broadening in the output of single-mode laser oscillators. This broadening manifests itself as a limited coherence time.
3. *Optical communication system.* We will consider the case of an optical communication system using a binary pulse code modulation in which the information is carried by means of a string of 1 and 0 pulses. The presence of noise will be shown to lead to a certain probability that any given pulse in the reconstructed train pulse is in error.

In this chapter we consider optical detectors utilizing light-generated charge carriers. These include the photomultiplier, the photoconductive detector, the *p-n* junction photodiode, and the avalanche photodiode. These detectors are the main ones used in the field of quantum electronics, because they combine high sensitivity with very short response times. Other types of detectors, such as bolometers, Golay cells, and thermocouples, whose operation depends on temperature changes induced by the absorbed radiation, will not be discussed.[1]

Two types of noise will be discussed in detail. The first type is thermal (Johnson) noise, which represents noise power generated by thermally agitated charge carriers. The expression for this noise will be derived by using the conventional thermodynamic treatment as well as by a statistical analysis of a particular model in which the physical origin of the noise is more apparent. The second type, shot noise (or generation-recombination noise in photoconductive detectors), is attributable to the random way in which electrons are emitted or generated in the process of interacting with a radiation field. This noise exists even at zero temperature, where thermal agitation or generation of carriers can be neglected. In this case it results from the randomness with which carriers are generated by the *very signal that is measured.* Detection in the limit of signal-generated shot noise is called quantum-limited detection, since the corresponding sensitivity is that allowed by the uncertainty principle in quantum mechanics. This point will be brought out in the next chapter.

10.1 LIMITATIONS DUE TO NOISE POWER

Measurement of Optical Power

Consider the problem of measuring an optical signal field

$$v_S(t) = V_S \cos \omega t \qquad (10.1\text{-}1)$$

in the presence of a noise field. The instantaneous noise field that adds to that of the signal can be taken as the sum of an in-phase component and a quadrature component according to

$$v_N(t) = V_{NC}(t) \cos \omega t + V_{NS}(t) \sin \omega t \qquad (10.1\text{-}2)$$

where $V_{NC}(t)$ and $V_{NS}(t)$ are slowly [compared to exp $(i\omega t)$] varying random uncorrelated quantities with a zero mean. The total field at the detector $v(t) = v_S(t) + v_N(t)$ can be written as

$$v(t) = \text{Re}\{[V_S + V_{NC}(t) - iV_{NS}(t)]e^{i\omega t}\} \qquad (10.1\text{-}3)$$

$$\equiv \text{Re}[V(t)e^{i\omega t}] \qquad (10.1\text{-}4)$$

The total (signal plus noise) field phasor $V(t)$ is shown in Figure 10-1.

In most situations of interest to optical detection the sources of noise are due to the concerted action of a large number of independent agents. In this case the central limit theorem of statistics [1] tells us that the probability function for finding $V_{NC}(t)$ at time t between V_{NC} and $V_{NC} + dV_{NC}$ is de-

[1] The interested reader will find a good description of these devices in Reference [6].

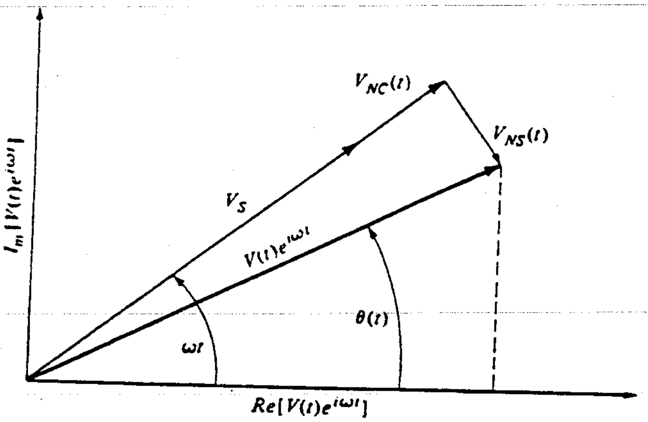


Figure 10-1 A phasor diagram showing the total (signal plus noise) field phasor $V(t)$ at time t. The instantaneous field is given by the horizontal projection of $V(t)$ exp $(i\omega t)$.

scribed by a Gaussian

$$p(V_{NC})\,dV_{NC} = \frac{1}{\sqrt{2\pi}\sigma} e^{-V_{NC}^2/2\sigma^2}\,dV_{NC} \qquad (10.1\text{-}5)$$

and by a similar expression in which V_{NS} replaces V_{NC} for $p(V_{NS})$. Since $V_{NC}(t)$ has a unity probability of having some value between $-\infty$ and ∞, it follows that

$$\int_{-\infty}^{\infty} p(V_{NC})\,dV_{NC} = 1 \qquad (10.1\text{-}6)$$

It follows from (10.1-5) that $\overline{V}_{NC}$, the ensemble average[2] (denoted by a horizontal bar) of V_{NC}, is zero,[3] whereas the mean square value is

[2] The ensemble average $\overline{A(t)}$ of a quantity $A(t)$ is obtained by measuring A simultaneously at time t in a very large number of systems that, *to the best of our knowledge*, are identical. Mathematically,

$$\overline{A(t)} = \lim_{N\to\infty}\left[\frac{1}{N}\sum_{n=1}^{N} A_n(t)\right]$$

where $A_n(t)$ denotes the observation in the nth system. In a truly random phenomenon, the time averaging and ensemble averaging lead to the same result, so the ensemble average is independent of the time t in which it is performed and can also be obtained from

$$\overline{A} = \int_{-\infty}^{\infty} Ap(A)\,dA$$

where $p(A)$ is the probability function, in the sense of (10.1-5), of the variable A.

[3] The reason for $V_{NC}(t) = 0$ can be appreciated from Figure 10-1. $V_{NC}(t)$ has an equal probability of being in phase with V_S as of being out phase, thus averaging out to zero.

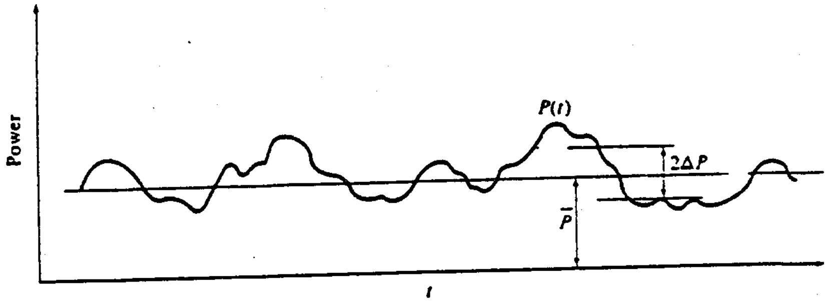


Figure 10-2 The intermingling of noise power with that of a signal causes the total power to fluctuate. The rms fluctuation ΔP limits the accuracy of power measurements.

$$\overline{V_{NC}^2} = \overline{V_{NS}^2} = \int_{-\infty}^{\infty} V_{NC}^2 p(V_{NC})\, dV_{NC} = \sigma^2 \qquad \text{(10.1-7)}$$

The "power" in $v(t)$ is obtained using (1.1-12) as

$$\begin{aligned} P(t) &\equiv [V(t)e^{i\omega t}][V^*(t)e^{-i\omega t}] \\ &= V_S^2 + 2V_S V_{NC} + V_{NC}^2 + V_{NS}^2 \end{aligned} \qquad \text{(10.1-8)}$$

The ensemble average (or *long* time average) of $P(t)$ is

$$\overline{P} \equiv \overline{P(t)} = \overline{V_S^2} + \overline{V_{NC}^2} + \overline{V_{NS}^2} = \overline{V_S^2} + 2\sigma^2 \qquad \text{(10.1-8a)}$$

where use has been made of the fact that $\overline{V}_{NC} = 0$ and of (10.1-7).

The physical significance of the time-varying power $P(t)$ and its long-time (or ensemble) average $\overline{P}$ is illustrated by Figure 10-2.

It is clear from the fluctuating nature of $P(t)$ that any measurement of this power is subject to an uncertainty due to the random nature of V_{NC} and V_{NS} in (10.1-8). As a measure of the uncertainty in power measurement, we may reasonably take the root mean square (rms) power deviation

$$\Delta P \equiv [\overline{(P(t) - \overline{P})^2}]^{1/2}$$

Using (10.1-8) and (10.1-8a), we obtain after some algebra

$$\Delta P = (4\overline{V_S^2 V_{NC}^2} + 2\overline{V_{NC}^4} - 2\overline{V_{NS}^2}\;\overline{V_{NC}^2})^{1/2} \qquad \text{(10.1-9)}$$

Using (10.1-5) we obtain

$$\overline{V_{NC}^4} = \int_{-\infty}^{\infty} V_{NC}^4 p(V_{NC})\, dV_{NC} = 3\sigma^4 \qquad \text{(10.1-10)}$$

so that using $\overline{V_{NC}^2} = \overline{V_{NS}^2} = \sigma^2$ in (10.1-9) results in

$$\Delta P = 2\sigma(\overline{V_S^2} + \sigma^2)^{1/2} = 2\sigma(P_S + \sigma^2)^{1/2} \qquad \text{(10.1-11)}$$

where according to (10.1-8) we may associate $P_S = \overline{V_S^2}$ with the signal power that is, the power that would be measured if V_{NC} and V_{NS} were, hypothetically, rendered zero.

A question of practical importance involves the minimum signal power that can be measured in the presence of noise. We may, somewhat arbitrarily, take this power P_{limit} to be that at which the uncertainty ΔP becomes equal to the signal power P_S. At this point we have from (10.1-11)

$$P_{\text{limit}} = 2\sigma(P_{\text{limit}} + \sigma^2)^{1/2}$$

or, after solving for P_{limit},

$$P_{\text{limit}} = 2\sigma^2(1 + \sqrt{2}) = P_N(1 + \sqrt{2}) \qquad (10.1\text{-}12)$$

where $P_N = 2\sigma^2 = \overline{V}^2_{NC} + \overline{V}^2_{NS}$ is the noise power. Widespread convention chooses to define the minimum detectable signal power as equal to P_N instead of $2.414P_N$, as obtained above. This simplification is understandable, since our choice of the limit of detectability $\Delta P = P_S$ was somewhat arbitrary. In any case the main conclusion to remember is that near the limit of detectivity, the rms power fluctuation is comparable to the signal power. The next task, which will be taken up in this chapter and in Chapter 11, is to find out the main sources of noise power and consequently ways to minimize them. Before tackling this task, however, we need to develop some mathematical tools for dealing with random processes.

10.2 NOISE—BASIC DEFINITIONS AND THEOREMS

A real function $v(t)$ and its Fourier transform $V(\omega)$ are related by

$$V(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} v(t)e^{-i\omega t}\,dt \qquad (10.2\text{-}1)$$

and

$$v(t) = \int_{-\infty}^{\infty} V(\omega)e^{i\omega t}\,d\omega \qquad (10.2\text{-}2)$$

In the process of measuring a signal $v(t)$, we are not in a position to use the infinite time interval needed, according to (10.2-1), to evaluate $V(\omega)$. If the time duration of the measurement is T, we may consider the function $v(t)$ to be zero when $t \le -T/2$ and $t \ge T/2$ and, instead of (10.2-1), get

$$V_T(\omega) = \frac{1}{2\pi}\int_{-T/2}^{T/2} v(t)e^{-i\omega t}\,dt \qquad (10.2\text{-}3)$$

Since $v(t)$ is real, it follows that

$$V_T(\omega) = V_T^*(-\omega) \qquad (10.2\text{-}4)$$

T is usually called the resolution or integration time of the system.

Let us evaluate the average power P associated with $v(t)$. Taking the instantaneous power as $v^2(t)$, we obtain[4]

$$P = \frac{1}{T}\int_{-T/2}^{T/2} v^2(t)\,dt = \frac{1}{T}\int_{-T/2}^{T/2}\left\{v(t)\left[\int_{-\infty}^{\infty} V_T(\omega)e^{i\omega t}\,d\omega\right]\right\}dt \tag{10.2-5}$$

Using (10.2-3) and (10.2-4) in the last equation and interchanging the order of integration leads to

$$P = \frac{2\pi}{T}\int_{-\infty}^{\infty} |V_T(\omega)|^2\,d\omega \tag{10.2-6}$$

or

$$P = \frac{4\pi}{T}\int_{0}^{\infty} |V_T(\omega)|^2\,d\omega \tag{10.2-7}$$

where we used

$$\lim_{T\to\infty}(2\pi)^{-1}\int_{-T/2}^{T/2} dt\,\exp\,[i(\omega+\omega')t] = \delta(\omega+\omega')$$

If we define the *spectral density function* $S_v(\omega)$ of $v(t)$ by

$$S_v(\omega) = \lim_{T\to\infty}\frac{4\pi|V_T(\omega)|^2}{T} \tag{10.2-8}$$

then, according to (10.2-7), $S_v(\omega)d\omega$ is the portion of the average power of $v(t)$ that is due to frequency components between ω and $\omega + d\omega$. According to this physical interpretation, we may measure $S_v(\omega)$ by separating the spectrum of $v(t)$ into its various frequency classes as shown in Figure 10-3 and then measuring the power output $S_v(\omega_i)\Delta\omega_i$ of each of the filters [2].

Wiener–Khintchine Theorem

We will next derive another formal result involving the spectral density function.

Consider the time average of the product of some field quantity $v(t)$ with its delayed version $v(t+\tau)$

$$C_v(t) = \overline{v(t)v(t+\tau)} \tag{10.2-9}$$

The function $C_v(\tau)$ is termed the autocorrelation function of $v(t)$. We use (10.2-2) to carry out the integration indicated in (10.2-9)

$$\begin{aligned} C_v(\tau) &= \frac{1}{T}\int_{-T/2}^{T/2} v(t)v(t+\tau)\,dt \\ &= \frac{1}{T}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-T/2}^{T/2} d\omega\,d\omega'\,dt\,V_T(\omega)V_T(\omega')e^{i(\omega+\omega')t}e^{i\omega\tau} \end{aligned} \tag{10.2-10}$$

[4] It may be convenient for this purpose to think of $v(t)$ as the voltage across a one-ohm resistance.

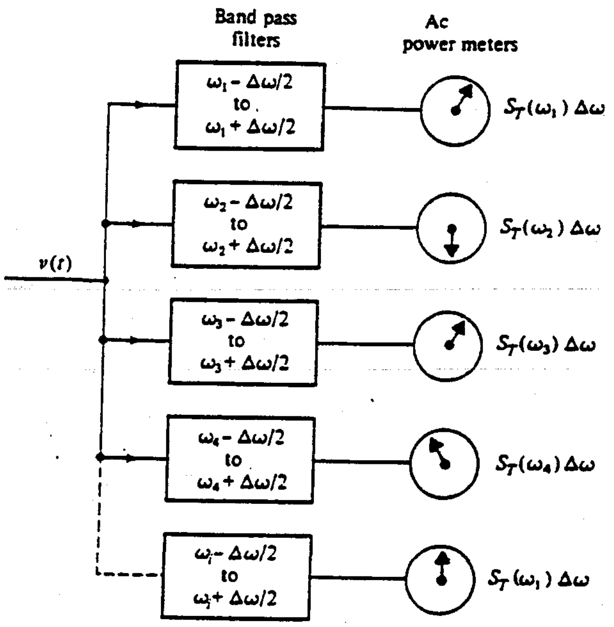

Band pass filters
Ac power meters
$v(t)$
$\omega_1 - \Delta\omega/2$ to $\omega_1 + \Delta\omega/2$
$S_T(\omega_1)\,\Delta\omega$
$\omega_2 - \Delta\omega/2$ to $\omega_2 + \Delta\omega/2$
$S_T(\omega_2)\,\Delta\omega$
$\omega_3 - \Delta\omega/2$ to $\omega_3 + \Delta\omega/2$
$S_T(\omega_3)\,\Delta\omega$
$\omega_4 - \Delta\omega/2$ to $\omega_4 + \Delta\omega/2$
$S_T(\omega_4)\,\Delta\omega$
$\omega_i - \Delta\omega/2$ to $\omega_i + \Delta\omega/2$
$S_T(\omega_1)\,\Delta\omega$

Figure 10-3 Diagram illustrating how the spectral density function $S_T(\omega)$ of a signal $v(t)$ can be obtained by measuring the power due to different frequency intervals.

In the limit $T \to \infty$,

$$\lim_{T\to\infty} \int_{-T/2}^{T/2} dt\, e^{i(\omega+\omega')t} = 2\pi\delta(\omega + \omega') \tag{10.2-11}$$

so that

$$C_v(\tau) = \lim_{T\to\infty} \frac{2\pi}{T} \iint_{-\infty}^{\infty} V_T(\omega')V_T(\omega)\delta(\omega + \omega')e^{i\omega\tau}\, d\omega\, d\omega'$$

$$= \lim_{T\to\infty} \frac{1}{2} \int_{-\infty}^{\infty} \frac{4\pi|V_T(\omega)|^2}{T} e^{i\omega\tau}\, d\omega \tag{10.2-12}$$

The quantity $4\pi|V_T(\omega)|^2/T$ is, according to (10.2-8), the spectral density function of $S_v(\omega)$ of $v(t)$, so that

$$C_v(\tau) = \frac{1}{2} \int_{-\infty}^{\infty} S_v(\omega)e^{i\omega\tau}\, d\omega \tag{10.2-13}$$

so that using (10.2-1)

$$S_v(\omega) = \frac{1}{\pi} \int_{-\infty}^{\infty} C_v(\tau)e^{-i\omega\tau}\, d\tau \tag{10.2-14}$$

The last two equations state that the spectral density function $S_v(\omega)$ and the autocorrelation function $C_v(\tau)$ form a Fourier transform pair. This result is one of the more important theoretical and practical tools of information theory and of the mathematics of random processes, and it is known, after the American and Russian mathematicians who, independently, formulated it, as the Wiener–Khintchine theorem. Its main importance for our purposes lies in the fact that it is often easier to obtain, experimentally or theoretically, $C_v(\tau)$ rather than $S_v(\omega)$, so that $S_v(\omega)$ is derived by a Fourier transformation of $C_v(\tau)$.

10.3 THE SPECTRAL DENSITY FUNCTION OF A TRAIN OF RANDOMLY OCCURRING EVENTS

Consider a time-dependent random variable $i(t)$ made up of a very large number of individual events $f(t - t_i)$ that occur at random times t_i.[5] An observation of $i(t)$ during a period T will yield

$$i_T(t) = \sum_{i=1}^{N_T} f(t - t_i) \qquad 0 \leq t \leq T \tag{10.3-1}$$

where N_T is the total number of events occurring in T. Typical examples of a random function $i(t)$ are provided by the thermionic emission current from a hot cathode (under temperature-limited conditions), or the electron current caused by photoemission from a surface. In these cases $f(t - t_i)$ represents the current resulting from a single electron emission occurring at t_i.

The Fourier transform of $i_T(t)$ is given according to (10.2-3) by

$$I_T(\omega) = \sum_{i=1}^{N_T} F_i(\omega) \tag{10.3-2}$$

where $F_i(\omega)$ is the Fourier transform[6] of $f(t - t_i)$

$$\begin{aligned} F_i(\omega) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} f(t - t_i)e^{-i\omega t}\,dt = \frac{e^{-i\omega t_i}}{2\pi}\int_{-\infty}^{\infty} f(t)\,e^{-i\omega t}\,dt \\ &= e^{-i\omega t_i}F(\omega) \end{aligned} \tag{10.3-3}$$

[5] This means that the *a priori* probability that a given event will occur in any time interval is distributed uniformly over the interval, or equivalently, that the probability $p(n)$ for n events to occur in an observation period T is given by the Poisson distribution function [2]

$$p(n) = \frac{(\bar{n})^n e^{-n}}{n!}$$

where $\bar{n}$ is the average number of events occurring in T.

[6] We assume that the individual event $f(t - t_i)$ is over in a short time compared to the observation period T, so the integration limits can be taken as $-\infty$ to ∞ instead of 0 to T.

From (10.3-2) and (10.3-3) we obtain

$$|I_T(\omega)|^2 = |F(\omega)|^2 \sum_{i=1}^{N_T} \sum_{j=1}^{N_T} e^{-i\omega(t_i - t_j)}$$

$$= |F(\omega)|^2 \left(N_T + \sum_{i \neq j}^{N_T} \sum_{j}^{N_T} e^{i\omega(t_j - t_i)} \right) \tag{10.3-4}$$

If we take the average of (10.3-4) over an ensemble of a very large number of physically identical systems, the second term on the right side of (10.3-4) can be neglected in comparison to N_T, since the times t_i are random. This results in

$$\overline{|I_T(\omega)|^2} = \overline{N}_T |F(\omega)|^2 \equiv \overline{N}T\, |F(\omega)|^2 \tag{10.3-5}$$

where the horizontal bar denotes ensemble averaging and where $\overline{N}$ is the average rate at which the events occur so that $\overline{N}_T = \overline{N}T$. The spectral density function $S_T(\omega)$ of the function $i_T(t)$ is given according to (10.2-8) and (10.3-5) as

$$S(\omega) = 4\pi \overline{N} |F(\omega)|^2 \tag{10.3-6}$$

In practice, one uses more often the spectral density function $S(\nu)$ defined so that the average power due to frequencies between ν and $\nu + d\nu$ is equal to $S(\nu)\, d\nu$. It follows then, that $S(\nu)\, d\nu = S(\omega)\, d\omega$; thus, since $\omega = 2\pi\nu$,

$$S(\nu) = 8\pi^2 \overline{N} |F(2\pi\nu)|^2 \tag{10.3-7}$$

The last result is known as Carson's theorem and its usefulness will be demonstrated in the following sections where we employ it in deriving the spectral density function associated with a number of different physical processes related to optical detection.

Equation (10.3-7) was derived for the case in which the individual events $f(t - t_i)$ were displaced in time but were otherwise identical. There are physical situations in which the individual events may depend on one or more additional parameters. Denoting the parameter (or group of parameters) as α, we can clearly single out the subclass of events $f_\alpha(t - t_i)$ whose α is nearly the same and use (10.3-7) to obtain directly

$$S_\alpha(\nu) = 8\pi^2 \overline{N}(\alpha) |F_\alpha(2\pi\nu)|^2\, \Delta\alpha \tag{10.3-8}$$

for the contribution of this subclass of events to $S(\nu)$. $F_\alpha(\omega)$ is the Fourier transform of $f_\alpha(t)$, and thus $\overline{N}(\alpha)\Delta\alpha$ is the average number of events per second whose α parameter falls between α and $\alpha + \Delta\alpha$.

$$\int_{-\infty}^{\infty} \overline{N}(\alpha)\, d\alpha = \overline{N}$$

The probability distribution function for α is $p(\alpha) = \bar{N}(\alpha)/\bar{N}$; therefore,

$$\int_{-\infty}^{\infty} p(\alpha)\, d\alpha = \frac{1}{\bar{N}} \int_{-\infty}^{\infty} \bar{N}(\alpha)\, d\alpha = 1 \tag{10.3-9}$$

Summing (10.3-8) over all classes α and weighting each class by the probability $p(\alpha)\,\Delta\alpha$ of its occurrence, we obtain

$$\begin{aligned} S(\nu) &= \sum_{\alpha} S_\alpha(\nu) = 8\pi^2 \sum_{\alpha} \bar{N}(\alpha)|F_\alpha(2\pi\nu)|^2\, \Delta\alpha \\ &= 8\pi^2\bar{N} \sum_{\alpha} |F_\alpha(2\pi\nu)|^2 p(\alpha)\, \Delta\alpha \\ &= 8\pi^2\bar{N} \int_{-\infty}^{\infty} |F_\alpha(2\pi\nu)|^2 p(\alpha)\, d\alpha = 8\pi^2\bar{N}\overline{|F(2\pi\nu)|^2} \end{aligned} \tag{10.3-10}$$

where the bar denotes averaging over α. Equation (10.3-10) is thus the extension of (10.3-7) to the case of events whose characterization involves, in addition to their time t_i, some added parameters. We will use it further in this chapter to derive the noise spectrum of photoconductive detectors in which case α is the lifetime of the excited photocarriers.

10.4 SHOT NOISE [3]

Let us consider the spectral density function of current arising from random generation and flow of mobile charge carriers. This current is identified with "shot noise." To be specific, we consider the case illustrated in Figure 10-4, in which electrons are released at random into the vacuum from electrode

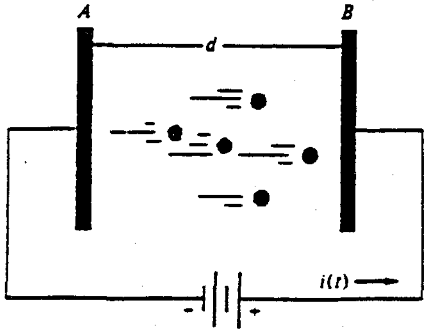


Figure 10-4 Random electron flow between two electrodes. This basic configuration is used in the derivation of shot noise.

A to be collected at electrode B, which is maintained at a slight positive potential relative to A.

The average rate $\bar{N}$ of electron emission from A is $\bar{N} = \bar{I}/e$, where $\bar{I}$ is the average current and the electronic charge is taken as $-e$. The current pulse due to a single electron as observed in the external circuit is

$$i_e(t) = \frac{ev(t)}{d} \tag{10.4-1}$$

where $v(t)$ is the instantaneous velocity and d is the separation between A and B. To prove (10.4-1), consider the case in which the moving electron is replaced by a thin sheet of a very large area and of total charge $-e$ moving between the plates, as illustrated in Figure 10-5.

It is a simple matter to show (see Problem 10.1), using the relation $\nabla \cdot \mathbf{E} = \rho/\epsilon$, that the charge induced by the moving sheet on the left electrode is

$$Q_1 = \frac{e(d-x)}{d} \tag{10.4-2}$$

and that on the right electrode is

$$Q_2 = \frac{ex}{d} \tag{10.4-3}$$

where x is the position of the charged sheet measured from the left electrode. The current in the external circuit due to a single electron is thus

$$i_e(t) = \frac{dQ_2}{dt} = \frac{e}{d}\frac{dx}{dt} = \frac{e}{d}v(t) \tag{10.4-4}$$

in agreement with (10.4-1).

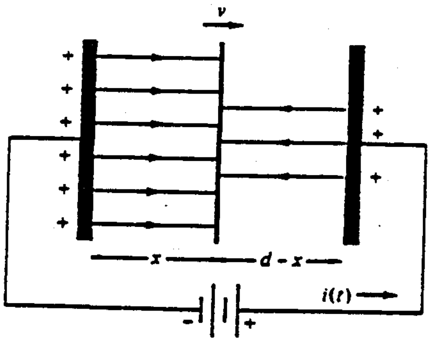


Figure 10-5 Induced charges and field lines due to a thin charge layer between the electrodes.

The Fourier transform of a single current pulse is

$$F(\omega) = \frac{e}{2\pi d}\int_0^{t_a} v(t)e^{-i\omega t}\,dt \qquad (10.4\text{-}5)$$

where t_a is the arrival time of an electron emitted at $t = 0$. If the transit time of an electron is sufficiently small that, at the frequency of interest ω,

$$\omega t_a \ll 1 \qquad (10.4\text{-}6)$$

i.e., $i_e(t) \propto \delta(t)$, we can replace exp $(-i\omega t)$ in (10.4-5) by unity and obtain

$$F(\omega) = \frac{e}{2\pi d}\int_0^{t_a} \frac{dx}{dt}\,dt = \frac{e}{2\pi} \qquad (10.4\text{-}7)$$

since $x(t_a)$ is, by definition, equal to d. Using (10.4-7) in (10.3-7) and recalling that $\bar{I} = e\bar{N}$ gives

$$S(\nu) = 8\pi^2\bar{N}\left(\frac{e}{2\pi}\right)^2 = 2e\bar{I} \qquad (10.4\text{-}8)$$

The power (in the sense of 10.2-5) in the frequency interval ν to $\nu + \Delta\nu$ associated with the current is, according to the discussion following (10.2-8), given by $S(\nu)\,\Delta\nu$. It is convenient to represent this power by an *equivalent noise generator* at ν with a mean-square current amplitude

$$\overline{i_N^2}(\nu) \equiv S(\nu)\,\Delta\nu = 2e\bar{I}\,\Delta\nu \qquad (10.4\text{-}9)$$

The noise mechanism described above is referred to as *shot noise*.

It is interesting to note that e in (10.4-9) is the charge of the particle responsible for the current flow. If, hypothetically, these carriers had a charge of $2e$, then at the *same average current* $\bar{I}$ the shot-noise power would double. Conversely, shot noise would disappear if the magnitude of an individual charge tended to zero. This is a reflection of the fact that shot noise is caused by fluctuations in the current that are due to the discreteness of the charge carriers and to the random electronic emission (for which the number of electrons emitted per unit time obey Poisson statistics [2]). The ratio of the fluctuations to the average current decreases with increasing number of events.[7]

Another point to remember is that, in spite of the appearance of $\bar{I}$ on the right side of (10.4-9), $\overline{i_N^2}(\nu)$ represents an alternating current with frequencies near ν.

[7] More precisely, for events obeying Poisson statistics we have (Reference [1] or derivable directly from footnote 5)

$$\frac{[\overline{(\Delta N)^2}]^{1/2}}{\bar{N}} = \frac{1}{(\bar{N})^{1/2}}$$

where N is the number of events in an observation time, $\bar{N}$ is the average value of N, and $\overline{(\Delta N)^2} \equiv \overline{(N - \bar{N})^2}$.

10.5 JOHNSON NOISE

Johnson, or *Nyquist noise* describes the fluctuations in the voltage across a dissipative circuit element; see References [4, 5]. These fluctuations are most often caused by the thermal motion of the charge carriers.[8] The charge neutrality of an electrical resistance is satisfied when we consider the whole volume, but locally the random thermal motion of the carriers sets up fluctuating charge gradients and, correspondingly, a fluctuating (ac) voltage. If we now connect a second resistance across the first one, the thermally induced voltage described above will give rise to a current and hence to a power transfer to the second resistor.[9] This is the so-called *Johnson noise*, whose derivation follows.

Consider the case illustrated in Figure 10-6 of a transmission line connected between two similar resistances R, which are maintained at the same temperature T. We choose the resistance R to be equal to the characteristic impedance Z_0 of the lines, so that no reflection can take place at the ends. The transmission line can support traveling voltage waves of the form

$$v(t) = A \cos (\omega t \pm kz) \tag{10.5-1}$$

where $k = 2\pi/\lambda$ and the phase velocity is $c = \omega/k$.

For simplicity we require that the allowed solutions be periodic in the distance L,[10] so if we extend the solution outside the limits $0 \le z \le L$ we

[8] We use the word "carriers" rather than "electrons" to include cases of ionic conduction or conduction by holes.

[9] The same argument applies to the second resistor, so at thermal equilibrium the net power leaving each resistor is zero.

[10] This seemingly arbitrary type of boundary condition is used extensively in similar situations in thermodynamics to derive the blackbody radiation density, or in solid-state physics to derive the density of electronic states in crystals.

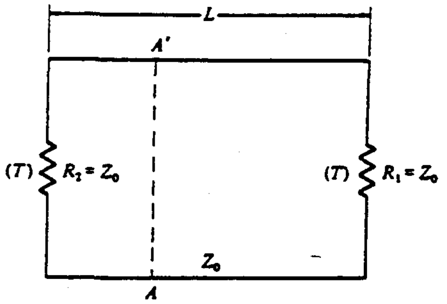


Figure 10-6 Lossless transmission line of characteristic impedance Z_0 connected between two matched loads ($R = Z_0$) at temperature T.

obtain

$$v(t) = A\cos[\omega t \pm k(z+L)] = A\cos(\omega t \pm kz)$$

This condition is fulfilled when

$$kL = 2m\pi \qquad m = 1, 2, 3, \ldots \tag{10.5-2}$$

Therefore, two adjacent modes differ in their value of k by

$$\Delta k = \frac{2\pi}{L} \tag{10.5-3}$$

and the number of modes having their k values somewhere between zero and $+k$ is[11]

$$N_k = \frac{kL}{2\pi} \tag{10.5-4}$$

or, using $k = 2\pi\nu/c$, we obtain

$$N(\nu) = \frac{\nu L}{c}$$

for the number of positively traveling modes with frequencies between zero and ν.

The number of modes per unit frequency interval is

$$p(\nu) = \frac{dN(\nu)}{d\nu} = \frac{L}{c} \tag{10.5-5}$$

Consider the power flowing in the $+z$ direction across some arbitrary plane, $A - A'$ say. It is clear that due to the lack of reflection this power must originate in R_2. Since the power is carried by the electromagnetic modes of the system, we have

$$\text{Power} = \frac{\text{energy}}{\text{distance}}\,(\text{velocity of energy})$$

We find, taking the velocity of light as c, that the power P due to frequencies between ν and $\nu + \Delta\nu$ is given by

$$\begin{aligned} P &= \left(\frac{1}{L}\right)\left(\begin{matrix}\text{number of modes between}\\ \nu \text{ and } \nu + \Delta\nu\end{matrix}\right)(\text{energy per mode})(c) \\ &= \left(\frac{1}{L}\right)\left(\frac{L}{c}\Delta\nu\right)\left(\frac{h\nu}{e^{h\nu/kT} - 1}\right)(c) \end{aligned}$$

or

$$P = \frac{h\nu\Delta\nu}{e^{h\nu/kT} - 1} \approx kT\Delta\nu \qquad (kT \gg h\nu) \tag{10.5-6}$$

[11] Negative k values correspond, according to (10.5-1), to waves traveling in the $-z$ direction. Our bookkeeping is thus limited to modes carrying power in the $+z$ direction.

where we used the fact that in thermal equilibrium the energy of a mode is given by [7]

$$\mathscr{E} = \frac{h\nu}{e^{h\nu/kT} - 1} \tag{10.5-7}$$

This result is also obtained in Appendix D from a different point of view. An equal amount of noise power is, of course, generated in the right resistor and is dissipated in the left one, so in thermal equilibrium the net power crossing any plane is zero.

The power given by (10.5-6) represents the maximum noise power available from the resistance, since it is delivered to a matched load. If the load connected across R has a resistance different from R, the noise power delivered is less than that given by (10.5-6). The noise-power bookkeeping is done correctly if the resistance R appearing in a circuit is replaced by either one of the following two equivalent circuits: a noise generator in series with R with mean-square voltage amplitude

$$\overline{v_N^2}(\nu) = \frac{4h\nu R\Delta\nu}{e^{h\nu/kT} - 1} \underset{kT \gg h\nu}{\simeq} 4kTR\Delta\nu \tag{10.5-8}$$

or a noise current generator of mean square value

$$\overline{i_N^2}(\nu) = \frac{4h\nu\Delta\nu}{R(e^{h\nu/kT} - 1)} \underset{kT \gg h\nu}{\simeq} \frac{4kT\Delta\nu}{R} \tag{10.5-9}$$

in parallel with R. The noise representations of the resistor are shown in Figure 10-7. There are numerous other derivations of the formula for Johnson noise. For derivations using lumped-circuit concepts and an antenna example, the reader is referred to References [6, 7], respectively.

Statistical Derivation of Johnson Noise

The derivation of Johnson noise leading to (10.5-6) leans heavily on thermodynamic and statistical mechanics considerations. It may be instructive to obtain this result using a physical model for a resistance and applying the

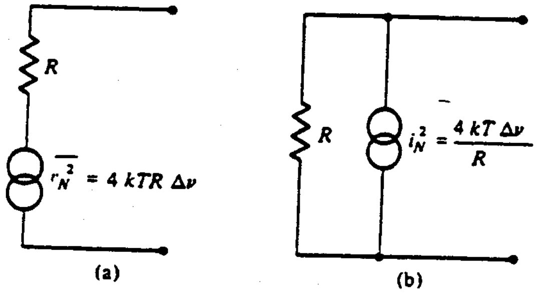

R
$\overline{r_N^2} = 4\,kTR\,\Delta\nu$
(a)
R
$\overline{i_N^2} = \frac{4\,kT\,\Delta\nu}{R}$
(b)

Figure 10-7 (a) Voltage and (b) current noise equivalent circuits of a resistance.

mathematical tools developed in this chapter. The model used is shown in Figure 10-8.

The resistor consists of a medium of volume $V = Ad$, which contains N free electrons per unit volume. In addition, there are N positively charged ions, which preserve the (average) charge neutrality. The electrons move about randomly with an average kinetic energy per electron of

$$\bar{E} = \tfrac{3}{2}kT = \tfrac{1}{2}m(\overline{v_x^2} + \overline{v_y^2} + \overline{v_z^2}) \tag{10.5-10}$$

where $\overline{v_x^2} = \overline{v_y^2} = \overline{v_z^2}$ refer to thermal averages. A variety of scattering mechanisms including electron–electron, electron–ion, and electron–phonon collisions act to interrupt the electron motion at an average rate of τ_0^{-1} times per second. τ_0 is thus the mean scattering time. These scattering mechanisms are responsible for the electrical resistance and give rise to a dc conductivity[12]

$$\sigma = \frac{N e^2 \tau_0}{m} \tag{10.5-11}$$

where m is the mass of the electron.[13] The sample dc resistance is thus

$$R = \frac{d}{\sigma A} = \frac{md}{Ne^2\tau_0 A} \tag{10.5-12}$$

while its ac resistance $R(\omega)$ is $md(1 + \omega^2\tau^2)/Ne^2\tau_0 A$.

We apply next the results of Section 10.3 to the problem and choose as our basic single event the current pulse $i_e(t)$ in the external circuit due to the motion of *one* electron between two successive scattering events. Using (10.4-1), we write

$$i_e(t) = \begin{cases} \dfrac{ev_x}{d} & 0 \le t \le \tau \\ 0 & \text{otherwise} \end{cases} \tag{10.5-13}$$

[12] The derivation of (10.5-11) can be found in any introductory book on solid-state physics.

[13] In a semiconductor we use the effective mass of the charge carrier.

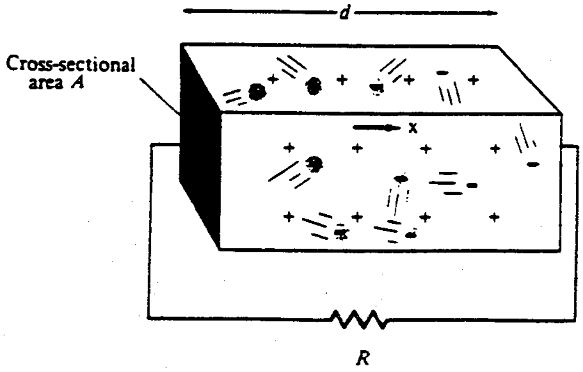


Figure 10-8 Model of a resistance used in deriving the Johnson-noise formula.

where v_x is the x component of the velocity (assumed constant) and where τ is the scattering time of the electron under observation. Taking the Fourier transform of $i_e(t)$, we have

$$I_e(\omega, \tau, v_x) = \frac{1}{2\pi}\int_0^\tau i_e(t)e^{-i\omega t}\,dt = \frac{(1/2\pi)ev_x}{-i\omega d}\,[e^{-i\omega\tau} - 1] \qquad (10.5\text{-}14)$$

from which

$$|I_e(\omega, \tau, v_x)|^2 = \frac{e^2v_x^2}{4\pi^2\omega^2d^2}\,[2 - e^{i\omega\tau} - e^{-i\omega\tau}] \qquad (10.5\text{-}15)$$

According to (10.3-10) we need to average $|I_e(\omega, \tau, v_x)|^2$ over the parameters τ and v_x. We assume that τ and v_x are independent variables—that is, that the probability function

$$p(\alpha) = p(\tau, v_x) = g(\tau)f(v_x)$$

is the product of the individual probabilities [1]—and take $g(\tau)$ as[14]

$$g(\tau) = \frac{1}{\tau_0}\,e^{-\tau/\tau_0} \qquad (10.5\text{-}16)$$

and, performing the averaging over τ, obtain

$$\overline{|I_e(\omega, v_x)|^2} = \int_0^\infty g(\tau)|I_e(\omega, v_x, \tau)|^2\,d\tau = \frac{2e^2v_x^2\tau_0^2}{4\pi^2d^2(1 + \omega^2\tau_0^2)} \qquad (10.5\text{-}17)$$

The second averaging over v_x^2 is particularly simple, since it results in the replacement of v_x^2 in (10.5-17) by its average $\overline{v_x^2}$, which, for a sample at thermal equilibrium, is given according to (10.5-10) by $\overline{v_x^2} = kT/m$. The final result is then

$$\overline{|I_e(\omega)|^2} = \frac{2e^2\tau_0^2kT}{4\pi^2md^2(1 + \omega^2\tau_0^2)} \qquad (10.5\text{-}18)$$

[14] If the collision probability per carrier per unit times is $1/\tau_0$ and $q(t)$ is the probability that an electron *has not* collided by time t, we have:

$$q'(t) = -q(t)\frac{1}{\tau_0} \Rightarrow q(t) = e^{-t/\tau_0}$$

Taking $g(\tau)\,d\tau$ as the probability that a collision will occur between τ and $\tau + d\tau$, it follows that

$$q(t) = 1 - \int_0^t g(t')\,dt'$$

and thus

$$g(t) = -\frac{dq}{dt} = \frac{1}{\tau_0}\,e^{-t/\tau_0}$$

as in (10.5-16).

The average number of scattering events per second $\bar{N}$ is equal to the total number of electrons NV divided by the mean scattering time τ_0

$$\bar{N} = \frac{NV}{\tau_0} \tag{10.5-19}$$

thus, from (10.3-10), we obtain

$$S(\nu) = 8\pi^2\overline{N|I_e(\omega)|^2} = \frac{4NVe^2\tau_0 kT}{md^2(1 + \omega^2\tau_0^2)}$$

and, after using (10.5-12) and limiting ourselves as in (10.4-6) to frequencies where $\omega\tau_0 \ll 1$, we get

$$\overline{i_N^2}(\nu) \equiv S(\nu)\Delta\nu = \frac{4kT\Delta\nu}{R(\nu)} \tag{10.5-20}$$

in agreement with (10.5-9).

10.6 SPONTANEOUS EMISSION NOISE IN LASER OSCILLATORS

Another type of noise that plays an important role in quantum electronics is that of spontaneous emission in laser oscillators and amplifiers. As shown in Chapter 5, a necessary condition for laser amplification is that the atomic population of a pair of levels 1 and 2 be inverted. If $E_2 > E_1$, gain occurs when $N_2 > N_1$. Assume that an optical wave with frequency $\nu = (E_2 - E_1)/h$ is propagating through an inverted population medium. This wave will grow coherently due to the effect of stimulated emission. In addition, its radiation will be contaminated by noise radiation caused by spontaneous emission from level 2 to level 1. Some of the radiation emitted by the spontaneous emission will propagate very nearly along the same direction as that of the stimulated emission and cannot be separated from it. This has two main consequences. First, the laser output has a finite spectral width. This effect is described in this section. Second, the signal-to-noise ratio achievable at the output of laser amplifiers [7] is limited because of the intermingling of spontaneous emission noise power with that of the amplified signal. (See Figure 10-9 and Appendix D.)

Returning to the case of a laser oscillator, we represent it by an RLC circuit, as shown in Figure 10-10. The presence of the laser medium with

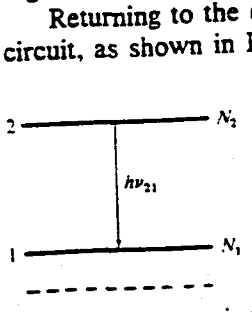


Figure 10-9 An atomic transition with $N_2 > N_1$ providing gain for laser oscillation.

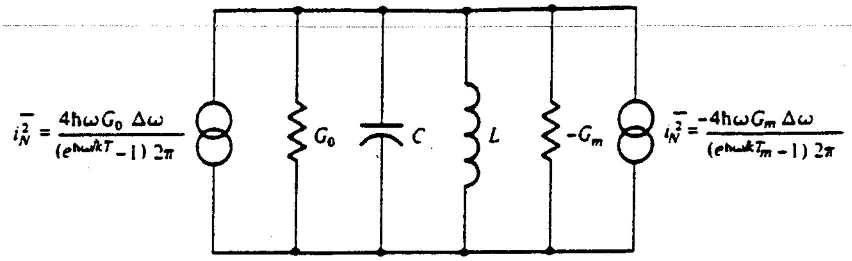

$\overline{i_N^2} = \frac{4\hbar\omega G_0\,\Delta\omega}{(e^{\hbar\omega/kT}-1)\,2\pi}$
G_0
C
L
$-G_m$
$\overline{i_N^2} = \frac{-4\hbar\omega G_m\,\Delta\omega}{(e^{\hbar\omega/kT_m}-1)\,2\pi}$

Figure 10-10 Equivalent circuit of a laser oscillator.

negative loss (that is, gain) is accounted for by including a negative conductance $-G_m$ while the ordinary loss mechanisms described in Chapter 6 are represented by the positive conductance G_0. The noise generator associated with the losses G_0 is given according to (10.5-9) as

$$\overline{i_N^2} = \frac{4\hbar\omega G_0(\Delta\omega/2\pi)}{e^{\hbar\omega/kT} - 1}$$

where T is the actual temperature of the losses. Spontaneous emission is represented by a similar expression[15]

$$(\overline{i_N^2})_{\substack{\text{spont}\\\text{emission}}} = \frac{4\hbar\omega(-G_m)(\Delta\omega/2\pi)}{e^{\hbar\omega/kT_m} - 1} \tag{10.6-1}$$

where the term $(-G_m)$ represents negative losses and T_m is a temperature determined by the population ratio according to

$$\frac{N_2}{N_1} = e^{-\hbar\omega/kT_m} \tag{10.6-2}$$

Since $N_2 > N_1$, then $T_m < 0$, $(\overline{i_N^2})$ in (10.6-1) is positive definite.

Although a detailed justification of (10.6-1) is outside the scope of the present treatment, a strong case for its plausibility can be made by noting that since $G_m \propto N_2 - N_1$, $(\overline{i_N^2})$ in (10.6-1) can be written, using (10.6-2), as[16]

$$(\overline{i_N^2})_{\substack{\text{spont}\\\text{emission}}} \propto \frac{-4\hbar\omega\Delta\omega(N_2 - N_1)}{(N_1/N_2) - 1} = 4\hbar\omega\Delta\omega N_2 \tag{10.6-3}$$

and is thus proportional to N_2. This makes sense, since spontaneous emission power is due to 2 → 1 transitions and should consequently be proportional to N_2.

[15] The 2π factor appearing in the denominators of $\overline{i_N^2}$ is due to the fact that here we use $\overline{i_N^2}(\omega)$ instead of $\overline{i_N^2}(\nu)$ with

$$\overline{i_N^2}(\omega)\Delta\omega = \overline{i_N^2}(\nu)\,\Delta\nu \qquad \Delta\omega = 2\pi\Delta\nu$$

[16] The proportionality of G_m to $N_2 - N_1$ can be justified by noting that in the equivalent circuit (Figure 10-10) the stimulated emission power is given by v^2G_m where v is the voltage. Using the field approach, this power is proportional to $E^2(N_2 - N_1)$ where E is the field amplitude. Since v is proportional to E, G_m is proportional to $N_2 - N_1$.

Returning to the equivalent circuit, its quality factor Q is given by

$$Q^{-1} = \frac{G_0 - G_m}{\omega_0 C} = \frac{1}{Q_0} - \frac{1}{Q_m} \tag{10.6-4}$$

where $\omega_0^2 = (LC)^{-1}$. The circuit impedance is

$$\begin{aligned} Z(\omega) &= \frac{1}{(G_0 - G_m) + (1/i\omega L) + i\omega C} \\ &= \frac{i\omega}{C} \frac{1}{(i\omega\omega_0/Q) + (\omega_0^2 - \omega^2)} \end{aligned} \tag{10.6-5}$$

so the voltage across this impedance due to a current source with a complex amplitude $I(\omega)$ is

$$V(\omega) = \frac{i}{C} \frac{I(\omega)}{[(\omega_0^2 - \omega^2)/\omega] + (i\omega_0/Q)} \tag{10.6-6}$$

which, near $\omega = \omega_0$, becomes

$$\overline{|V(\omega)|^2} = \frac{1}{4C^2} \frac{\overline{|I(\omega)|^2}}{(\omega_0 - \omega)^2 + (\omega_0^2/4Q^2)} \tag{10.6-7}$$

The current sources driving the resonant circuit are those shown in Figure 10-10; since they are not correlated, we may take $\overline{|I(\omega)|^2}$ as the sum of their mean-square values

$$\overline{|I(\omega)|^2} = 4\hbar\omega \left[\frac{G_m N_2}{N_2 - N_1} + \frac{G_0}{e^{\hbar\omega/kT} - 1}\right] \frac{d\omega}{2\pi} \tag{10.6-8}$$

where in the first term inside the square brackets we used (10.6-2). In the optical region, $\lambda = 1\ \mu$m say, and for $T = 300°$K we have $\hbar\omega/kT \simeq 50$; thus, since near oscillation $G_m \simeq G_0$, we may neglect the thermal (Johnson) noise term in (10.6-8), thereby obtaining

$$\overline{|V(\omega)|^2}_{\omega \to \omega_0} = \frac{\hbar G_m}{2\pi C^2} \left(\frac{N_2}{N_2 - N_1}\right) \frac{\omega\, d\omega}{(\omega_0 - \omega)^2 + (\omega_0^2/4Q^2)} \tag{10.6-9}$$

Equation (10.6-9) represents the spectral distribution of the laser output. If we subject the output to high-resolution spectral analysis, we should, according to (10.6-9), measure a linewidth

$$\Delta\omega = \frac{\omega_0}{Q} \tag{10.6-10}$$

between the half-intensity points. The trouble is that, though correct, (10.6-10) is not of much use in practice. The reason is that according to (10.6-4), Q^{-1} is equal to the difference of two nearly equal quantities neither of which is known with high enough accuracy. We can avoid this difficulty by showing that Q is related to the laser power output, and thus $\Delta\omega$ may be expressed in terms of the power.

The total optical oscillation power extracted from the atoms comprising the laser is

$$P = G_0 \int_0^\infty \frac{|V(\omega)|^2}{d\omega}\, d\omega$$
$$= \frac{\hbar G_m G_0}{2\pi C^2}\left(\frac{N_2}{N_2 - N_1}\right)\int_0^\infty \frac{\omega\, d\omega}{(\omega_0 - \omega)^2 + (\omega_0/2Q)^2} \tag{10.6-11}$$

Since the integrand peaks sharply near $\omega = \omega_0$, we may replace ω in the numerator of (10.6-11) by ω_0 and after integration obtain

$$P = \frac{\hbar G_m G_0 Q}{C^2}\left(\frac{N_2}{N_2 - N_1}\right) \tag{10.6-12}$$

which is the desired result linking P to Q. In a laser oscillator the gain very nearly equals the loss, or in our notation, $G_m = G_0$. Using this result in (10.6-12), we obtain

$$Q = \frac{C^2}{\hbar G_0^2}\left(\frac{N_2 - N_1}{N_2}\right) P$$

which, when substituted in (10.6-10), yields

$$\Delta\nu = \frac{2\pi h\nu_0(\Delta\nu_{1/2})^2}{P}\left(\frac{N_2}{N_2 - N_1}\right) \tag{10.6-13}$$

where $\Delta\nu_{1/2}$ is the full width of the passive cavity resonance given in (4.7-6) as $\Delta\nu_{1/2} = \nu_0/Q_0 = (1/2\pi)(G_0/C)$. It is worthwhile to recall here that $\Delta\nu$ represents, in the quantum limit, the laser field spectral width. The expression (10.6-13) is known as the Schawlow–Townes linewidth after the two American co-inventors of the laser [18] who first derived it.

Equation (10.6-13) does not predict an inverse dependence of $\Delta\nu$ on P, as may be deduced at a first glance, because of the dependence of N_2 on P. For very large powers, $P \rightarrow \infty$, N_2 is proportional to P, while $N_2 - N_1$ remains clamped at its threshold value. This leads to a residual power independent value of $\Delta\nu$. To appreciate this argument qualitatively, we note that unless the lifetime t_1 of the lower laser level is zero, as P increases, N_1 must increase since the increased (net)-induced transition rate into level 1 must equal in steady state N_1/t_1, the rate of emptying of level 1. This causes the population N_2 to increase in order to keep $N_2 - N_1$ and thus the gain, a constant. At sufficiently high values of P, N_2 becomes and stays proportional to P and the ratio N_2/P in (10.6-13) approaches a constant value, thus leading to a residual power independent linewidth.

To obtain the power dependence of the factor

$$\mu \equiv \frac{N_2}{(N_2 - N_1)_{\text{th}}}$$

we solve the rate equations for the atomic populations plus the equation for the photon number $p(p$ = number of photons in the optical resonator)

$$\frac{dN_2}{dt} = R - \frac{N_2}{t_2} - (N_2 - N_1)W_i$$

$$\frac{dN_1}{dt} = \frac{N_1}{t_1} + (N_2 - N_1)W_i + \frac{N_2}{t_2}$$

$$\frac{dp}{dt} = (N_2 - N_1)W_i - \frac{p}{t_c} \tag{10.6-14}$$

The first two equations are similar to (5.6-3) and (5.6-4) with $R_1 = 0$, $t_2 \rightarrow t_{\text{spont}}$, $R_2 \rightarrow R$, W_i is the induced transition rate and N_2, N_1, representing the total atomic populations of the laser transition levels 2 and 1, respectively. The third equation is a conservation equation for the total number of photons. W_i is the induced transition rate. The photon lifetime t_c is related to the cavity linewidth $\Delta\nu_{1/2}$ by $\Delta\nu_{1/2} = (2\pi t_c)^{-1}$. At equilibrium, $d/dt = 0$, we can solve (10.6-14) to obtain

$$N_2 - N_1 = \frac{R(t_2 - t_1)}{1 + W_i t_2}$$

$$N_2 = \frac{Rt_2(1 + W_i t_1)}{1 + W_i t_2} \tag{10.6-15}$$

so that

$$\frac{N_2}{(N_2 - N_1)_{\text{th}}} = \frac{t_2}{t_2 - t_1}(1 + W_i t_1) \tag{10.6-16}$$

where the subscript "th" indicates the value at threshold. The power output, including "wall losses" of the laser, is

$$P = (N_2 - N_1)_{\text{th}} W_i h\nu \tag{10.6-17}$$

which, when used together with (10.6-19) in (10.6-13) gives

$$\Delta\nu_{\text{laser}} = \frac{2\pi h\nu(\Delta\nu_{1/2})^2}{P}\,\frac{t_2}{t_2 - t_1} + \frac{c\Delta\nu_{1/2}\lambda_0^2}{8\pi n^3 \Delta\nu_{\text{gain}} V}\,\frac{t_1}{t_2 - t_1} \tag{10.6-18}$$

where $\Delta\nu_{\text{gain}}$ is the linewidth of atomic transition responsible for the laser gain. V is the mode volume. In obtaining (10.6-18), we use

$$(N_2 - N_1)_{\text{th}} = \frac{8\pi\nu^2 n^3 \Delta\nu_{\text{gain}} V t_2}{c^3 t_c} \quad (t_2 = t_{\text{spont}}) \tag{10.6-19}$$

which is obtained from (6.1-11) if we put $\Delta\nu_{\text{gain}} = 1/g(\nu)$. The first term on the right-hand side of (10.6-18) is the conventional Schawlow–Townes expression containing the inverse P dependence. The second term is power independent and corresponds to a residual linewidth as $P \rightarrow \infty$.

To get an idea of the magnitudes involved, we consider the case of a

0.6328 μm He–Ne laser with mirror reflectivities of $R = 0.99$, a resonator length of $1 = 30$ cm, and take $t_1/t_2 = 0.1$. We obtain

$$\Delta\nu_{1/2}(\text{Hz}) = \frac{(1-R)c}{2\pi nl} = 1.6 \times 10^6$$

and

$$\Delta\nu_{\text{laser}}(\text{Hz}) = \frac{10^{-3}}{P(\text{mW})} + 3.8 \times 10^{-4}$$

The residual linewidth thus dominates at power levels exceeding a few milliwatts.

10.7 PHASOR DERIVATION OF THE LASER LINEWIDTH

The derivation of the laser linewidth in Section 10.6 takes advantage of the highly sophisticated and efficient concepts and phenomena represented by the seemingly simple circuit model of a laser oscillator. The price we pay when taking this approach is a certain loss of physical insight into the mechanisms whereby spontaneous emission affects the laser linewidth.

In this section we will derive the expression (10.6-13) for the laser linewidth using a different approach. This is done not only for pedagogic purposes, but because some of the interim results involving phase fluctuations are useful in their own right.

The Phase Noise

An ideal monochromatic radiation field can be written as

$$\mathscr{E}(t) = \text{Re}[E_0 e^{i(\omega_0 t+\theta)}] \tag{10.7-1}$$

where ω_0 the radian frequency, E_0 the field amplitude, and θ are constant. A real field including that of lasers undergoes random phase and amplitude fluctuations that can be represented by writing

$$\mathscr{E}(t) = \text{Re}[E(t)e^{i[\omega_0 t+\theta(t)]}] \tag{10.7-2}$$

where $E(t)$ and $\theta(t)$ vary only "slightly" during one optical period.

There are many reasons in a practical laser for the random fluctuation in amplitude and phase. Most of these can be reduced, in theory, to inconsequence by various improvements such as ultrastabilization of the laser cavity length and the near elimination of microphonic and temperature variations. There remains, however, a basic source of noise that is quantum mechanical in origin. This is due to spontaneous emission that continually causes new power to be added to the laser oscillation field. The electromagnetic field represented by this new power, not being coherent with the old field, causes phase, as well as amplitude, fluctuations. These are responsible ultimately for the deviation of the evolution of the laser field from that of an ideal monochromatic field, i.e., for the quantum mechanical noise.

Let us consider the effect of one spontaneous emission event on the electromagnetic field of a single oscillating laser mode. A field such as (10.7-1) can be represented by a phasor of length E_0 rotating with an angular (radian) rate ω_0. In a frame rotating at ω_0 we would see a constant vector E_0. Since $E_0^2 \propto \bar{n}$, the average number of quanta in the mode, we shall represent the laser field phasor before a spontaneous emission event by a phasor of length $\sqrt{\bar{n}}$ as in Figure 10-11. The spontaneous emission adds *one* photon to the field, and this is represented, according to our conversion, by an incremental vector of unity length. Since this field increment is not correlated in phase with the original field, the angle ϕ is a random variable (i.e., it is distributed uniformly between zero and 2π). The resulting change $\Delta\theta$ of the field phase can be approximated for $\bar{n} \gg 1$ by

$$\Delta\theta_{\text{one emission}} = \frac{1}{\sqrt{\bar{n}}}\cos\phi \tag{10.7-3}$$

Next consider the effect of N spontaneous emissions on the phase of the laser field. The problem is one of random walk, since ϕ may assume with equal probability any value between 0 and 2π. We can then write

$$\langle[\Delta\theta(N)]^2\rangle = \langle(\Delta\theta_{\text{one emission}})^2\rangle N \tag{10.7-4}$$

and from (10.7-3)

$$\langle[\Delta\theta(N)]^2\rangle = \frac{1}{\bar{n}}\langle\cos^2\phi\rangle N$$

where $\langle\ \rangle$ denotes an ensemble average taken over a very large number of individual emission events.

Equation (10.7-4) is a statement of the fact that in a random walk problem the mean squared distance traversed after N steps is the square of the size

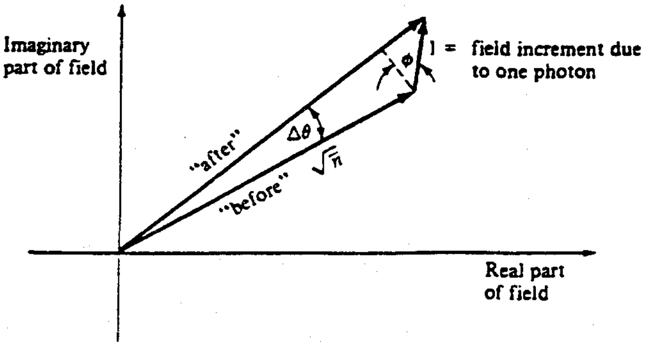


Figure 10-11 The phasor model for the effect of a single spontaneous emission event on the laser field phase.

of one step times N. The mean deviation $\langle \Delta\theta(N)\rangle$ after N spontaneous emissions is, of course, zero. Any one experiment, however, will yield a nonzero result. The mean squared deviation is thus nonzero and is a measure of the phase fluctuation. To obtain the root-mean-square (rms) phase deviation in a time t, we need to calculate the average number of spontaneous emission events $N(t)$ into a single laser mode in a time t.

The total number of spontaneous transitions per second into all modes is N_2/t_{spont}, where N_2 is the total number of atoms in the upper laser level 2 and t_{spont} is the spontaneous lifetime of an atom in 2. The total number of transitions per second into one mode is thus

$$\frac{N_{\text{spont}}}{\text{second-mode}} = \frac{N_2}{t_{\text{spont}}p} \tag{10.7-5}$$

where

$$p = \frac{8\pi\nu_0^2\,\Delta\nu V n^3}{c^3} \tag{10.7-6}$$

is the number of modes interacting with the laser transition, i.e., partaking in the spontaneous emission. V is the mode volume, and $\Delta\nu$ is the linewidth of the atomic transition responsible for the laser gain. We can rewrite (10.7-5) as

$$\frac{N_{\text{spont}}}{\text{second-mode}} = \left(\frac{N_2}{\Delta N_t}\right)\frac{(\Delta N_t)}{t_{\text{spont}}p} \tag{10.7-7}$$

where ΔN_t is the population inversion $(N_2 - N_1)$ at threshold. Next we use the result (6.1-11)

$$\Delta N_t = \frac{p t_{\text{spont}}}{t_c}$$

where t_c is the photon lifetime in the resonator, and obtain

$$\frac{N_{\text{spont}}}{\text{second-mode}} = \frac{\mu}{t_c}, \qquad \mu \equiv \frac{(N_2)_t}{\Delta N_t} = \frac{(N_2)_t}{(N_2 - N_1)_t} \tag{10.7-8}$$

The number of spontaneous transitions into a single mode in a time τ is thus

$$N(\tau) = \frac{\mu\tau}{t_c} \tag{10.7-9}$$

We recall here that in an ideal four-level laser $N_1 = 0$ and $\Delta N_t = N_2$, i.e., $\mu = 1$. In a three-level laser, on the other hand, μ can be appreciably larger than unity. In a ruby laser at room temperature, for example (see Section 7.3), $\mu = 50$. This reflects the fact that for a given gain the total excited population N_2 of a three-level laser must exceed that of a four-level laser by the factor μ, since gain is proportional to $N_2 - N_1$. Equation (10.7-8) is also equivalent to stating that above threshold there are μ spontaneously emitted photons present in a laser mode.

Using (10.7-9) in (10.7-4), we obtain for the root-mean-square phase deviation after τ seconds

$$\Delta\theta(t) \equiv \langle[\Delta\theta(t)]^2\rangle^{1/2} = \sqrt{\frac{1}{2\bar{n}}\frac{\mu t}{t_c}}$$

The maximum time t available for such an experiment is the integration time T of the measuring apparatus so that

$$\Delta\theta(T) = \sqrt{\frac{1}{2\bar{n}}\frac{\mu T}{t_c}} \tag{10.7-10}$$

The rms frequency excursion caused by $\Delta\theta$ is

$$(\Delta\omega)_{\text{RMS}} = \frac{\Delta\theta(T)}{T} = \sqrt{\frac{\mu}{2\bar{n}t_c T}} \tag{10.7-11}$$

We can cast the last result in a more familiar form by using the relations

$$P_e = \frac{\bar{n}\hbar\omega_0}{t_c} \qquad B = \frac{1}{2T} \tag{10.7-11a}$$

Here P_e is the power emitted by the atoms (i.e., the sum of the useful power output plus any power lost by scattering and absorption), and B is the bandwidth in hertz of the phase-measuring apparatus. The result is

$$(\Delta\omega)_{\text{RMS}} = \sqrt{\frac{\mu\hbar\omega_0}{P_e t_c^2}B} \tag{10.7-12}$$

From the experimental point of view $(\Delta\omega)_{\text{RMS}}$ is the root-mean-square deviation of the reading of an instrument whose output is the frequency $\omega(t) \equiv d\theta/dt$. We will leave it as an exercise (Problem 10.11) for the student to design an experiment that measures $(\Delta\omega)_{\text{RMS}}$.

Ring laser gyroscopes sense rotation by comparing the oscillation frequencies of two counter-propagating modes in a rotating ring resonator. Their sensitivity, i.e., the smallest rotation rate that they can sense, is thus limited by any uncertainty $\Delta\omega$ in the laser frequency. Experiments have indeed demonstrated a rotation measuring sensitivity approaching the quantum limit as given by (10.7-12).

The Laser Field Spectrum

Next we address the case where one measures directly the spectrum of the optical field

$$\mathscr{E}(t) = \text{Re}[E(t)e^{i[\omega_0 t+\theta(t)]}] \tag{10.7-13}$$

using, say, a scanning Fabry–Perot etalon. If the etalon has a sufficiently high spectral resolution, the measurement should yield the spectral density function $S_\epsilon(\omega)$ of the laser field. We will, consequently, proceed to obtain an expression for this quantity. We make use of the Wiener–Khintchine theorem (10.2-14) according to which $S_\epsilon(\omega)$ is the Fourier integral transform of the field autocorrelation function $C_\epsilon(\tau)$

$$S_{\mathcal{E}}(\omega) = \frac{1}{\pi}\int_{-\infty}^{\infty} C_{\mathcal{E}}(\tau)e^{-i\omega\tau}\,d\tau \tag{10.7-14}$$

$$S_{\mathcal{E}}(\omega) = \frac{4\pi}{T}|\mathcal{E}_T(\omega)|^2 \qquad \mathcal{E}_T(\omega) = \frac{1}{2\pi}\int_{-T/2}^{T/2} \mathcal{E}(t)e^{-i\omega t}\,dt$$

$$C_{\mathcal{E}}(\tau) \equiv \langle \mathcal{E}(t)\mathcal{E}(t+\tau)\rangle \tag{10.7-15}$$

where the symbol $\langle\ \rangle$ represents an ensemble, or time, average.

Using (10.7-13) we obtain

$$\begin{aligned} C_{\mathcal{E}}(\tau) = \frac{1}{4}\langle [&E(t)e^{i[\omega_0 t+\theta(t)]} + E^*(t)e^{-i[\omega_0 t+\theta(t)]}] \\ &\times [E(t+\tau)e^{i[\omega_0(t+\tau)+\theta(t+\tau)]} + E^*(t+\tau)e^{-i[\omega_0(t+\tau)+\theta(t+\tau)]}]\rangle \end{aligned} \tag{10.7-16}$$

Now, for example,

$$\langle E(t)E(t+\tau)e^{i[2\omega_0 t+\theta(t)+\theta(t+\tau)]}\rangle = 0$$

since it corresponds to averaging a signal oscillating at twice the optical frequency over many periods. So if we keep only the slowly varying terms in $C_{\mathcal{E}}(\tau)$, we obtain

$$\begin{aligned} C_{\mathcal{E}}(\tau) &= \frac{1}{4}\langle E(t)E^*(t+\tau)e^{i[-\omega_0\tau+\theta(t)-\theta(t+\tau)]} + E^*(t)E(t+\tau)e^{i[\omega_0\tau-\theta(t)+\theta(t+\tau)]}\rangle \\ &= [I(\tau) + I^*(\tau)] \end{aligned} \tag{10.7-17}$$

$$I(\tau) = \langle E^*(t)E(t+\tau)e^{i[\Delta\theta(t,\tau)+\omega_0\tau]}\rangle \tag{10.7-18}$$

$$\Delta\theta(t,\tau) \equiv \theta(t+\tau) - \theta(t) \tag{10.7-19}$$

The main contributions to the laser noise are due to fluctuations of the phase $\theta(t)$ and not the amplitude $E(t)$, since the amplitude fluctuations are kept negligibly small by gain saturation. Taking advantage of this fact, we write $\langle E^*(t)E(t+\tau)\rangle = \langle E^2\rangle \approx$ constant so that

$$I(\tau) = \langle E^2\rangle e^{i\omega_0\tau}\langle e^{i\Delta\theta(t,\tau)}\rangle \tag{10.7-20}$$

Given a (normalized) probability distribution function for $\Delta\theta$, $g(\Delta\theta)$, the expectation value of $\exp\{i\Delta\theta(t,\tau)\}$ is obtained from

$$\langle e^{i\Delta\theta(t,\tau)}\rangle = \int_{-\infty}^{\infty} e^{i\Delta\theta(t,\tau)}g(\Delta\theta)\,d(\Delta\theta) \tag{10.7-21}$$

Since the total phase excursion $\Delta\theta$ is the net result of many small and statistically independent (spontaneous transitions) excursions, the central limit theorem of statistics applies, and $g(\Delta\theta)$ is a Gaussian, which we write as

$$g(\Delta\theta) = \frac{1}{\sqrt{2\pi\langle(\Delta\theta)^2\rangle}}\, e^{-(\Delta\theta)^2/2\langle(\Delta\theta)^2\rangle} \qquad (10.7\text{-}22)$$

where

$$\langle(\Delta\theta)^2\rangle = \int_{-\infty}^{\infty} (\Delta\theta)^2 g(\Delta\theta)d(\Delta\theta) \qquad (10.7\text{-}23)$$

Using (10.7-22) in (10.7-21), we obtain

$$\langle e^{i\Delta\theta(t,\tau)}\rangle = e^{-\langle(\Delta\theta)^2\rangle/2} = e^{-\mu\tau/(4\bar{n}t_c)} \qquad (10.7\text{-}24)$$

where in order to obtain the last result, we used (10.7-10) with $T = \tau$. Using (10.7-24) in (10.7-20),

$$C_{\mathscr{E}}(\tau) = \frac{1}{4}\langle E^2\rangle e^{-\mu\tau/4\bar{n}t_c}(e^{i\omega_0\tau} + e^{-i\omega_0\tau}) \qquad (10.7\text{-}25)$$

The spectral density function of the laser field $S_{\mathscr{E}}(\omega)$, the quantity observed by a spectral analysis of the field, is given according to (10.7-14) and (10.7-25) by

$$S_{\mathscr{E}}(\omega) = \frac{\langle E^2\rangle}{4\pi}\int_{-\infty}^{\infty} e^{(-\mu\tau/4\bar{n}t_c)-i\omega\tau}\,(e^{i\omega_0\tau} + e^{-i\omega_0\tau}\, dt)\, d\tau \qquad (10.7\text{-}26)$$

$$= \frac{\langle E^2\rangle}{4\pi}\left(\frac{\mu/4\bar{n}t_c}{(\mu/4\bar{n}t_c)^2 + (\omega-\omega_0)^2} + \frac{\mu/4\bar{n}t_c}{(\mu/4\bar{n}t_c)^2 + (\omega+\omega_0)^2}\right) \qquad (10.7\text{-}27)$$

We have defined in (10.2-7) the spectral density function in such a way that only positive frequencies need to be considered. For $\omega > 0$ the second term on the right side of (10.7-27) contributes negligibly so that

$$S_{\mathscr{E}}(\omega) = \frac{\langle E^2\rangle}{4\pi}\,\frac{\mu/4\bar{n}t_c}{(\mu/4\bar{n}t_c)^2 + (\omega-\omega_0)^2} \qquad (10.7\text{-}28)$$

which corresponds to a Lorentzian-shaped function centered on the nominal laser frequency ω_0 with a full width at half-maximum of

$$(\Delta\omega)_{\text{laser}} = \frac{\mu}{2\bar{n}t_c} \qquad (10.7\text{-}29)$$

Recalling that the total power emitted by the electrons is $P = \bar{n}\hbar\omega_0/t_c$ and defining the passive resonator linewidth $\Delta\nu_{1/2} = (2\pi t_c)^{-1}$, we can rewrite (10.7-29) using (10.7-11a) as

$$(\Delta\nu)_{\text{laser}} = \frac{(\Delta\omega)_{\text{laser}}}{2\pi} = \frac{2\pi h\nu_0(\Delta\nu_{1/2})^2\mu}{P} \tag{10.7-30}$$

which, recalling the definition (10.7-8) of μ, is identical to (10.6-13).

Numerical Example: Linewidth of a He–Ne Laser and a Semiconductor Diode Laser

To obtain an order of magnitude estimate of the linewidth $(\Delta\nu)_{\text{laser}}$ predicted by (10.7-30), we will calculate it in the case of two largely different types of CW lasers: (1) a He–Ne laser and (2) a semiconductor GaInAsP laser.

1. *He–Ne laser.*

$\nu = 4.741 \times 10^{14}$ Hz ($\lambda = 6328$ Å)
l (distance between reflectors) = 100 cm
Loss = $(1 - R) = 1\%$ per pass

From these numbers we get

$$(\Delta\nu_{1/2}) = \frac{1}{2\pi t_c} \approx \frac{(1-R)c}{2\pi nl} \approx 5 \times 10^5$$

(i.e., $t_c = 3.2 \times 10^{-7}$ s) and from (10.7-30), assuming $\mu = 1$ (i.e., $N_1 \ll N_2$).

$$(\Delta\nu)_{\text{laser}} \cong 2 \times 10^{-3}\ \text{Hz}$$

at a power level $P = 1$ mW.

The predicted linewidth is thus so small as to be completely masked in almost all experimental situations by contributions due to extraneous causes, such as vibrations and temperature fluctuations.

2. *Semiconductor laser.* We use as a typical example the case of a GaInAsP ($\lambda = 1.55\ \mu$m) laser with the following pertinent characteristics:

$P = 3$ mW
$\nu = 1.935 \times 10^{14}$ ($\lambda_0 = 1.55\ \mu$m)

$$\Delta\nu_{1/2} = \frac{(1-R)c}{2\pi nl}$$

R (reflectivity) = 30%
$l = 300\ \mu$m
$n = 3.5$
$\mu = 3$ (at $T = 300$ K)

This results in $\Delta\nu_{1/2} \sim 3 \times 10^{10}$ (i.e., $t_c = 1/(2\pi\Delta\nu_{1/2}) = 5 \times 10^{-12}$ s) and

$$(\Delta\nu)_{\text{laser}} = 0.817 \times 10^6 \text{ Hz}$$

The experimental curve of Figure 10-12 shows the predicted [Equation (10.7-30)] P^{-1} dependence of $(\Delta\nu)_{\text{laser}}$, but the measured values of the linewidth are larger by a factor of ~70 than those predicted by the analysis. This discrepancy has been studied by a number of investigators [20–22], who have shown that the analysis leading to (10.7-30) ignores the modulation of the index of refraction of the laser medium, which is due to fluctuations of the electron density caused by spontaneous emission. When this effect is included, the result is to multiply Equation (10.7-30) by the factor

$$1 + \left(\frac{\Delta n'}{\Delta n''}\right)^2 \tag{10.7-31}$$

where $\Delta n'$ and $\Delta n''$ are, respectively, the changes in the real and imaginary parts of the index of refraction "seen" by the laser field due to some change in the electron density. The factor $1 + (\Delta n'/\Delta n'')^2$ can be calculated from measured parameters of the laser or measured directly [6]. Its value is ~30 in typical cases, enough to reconcile the observed data of Figure 10-12 and the prediction of Equation (10.7-30).

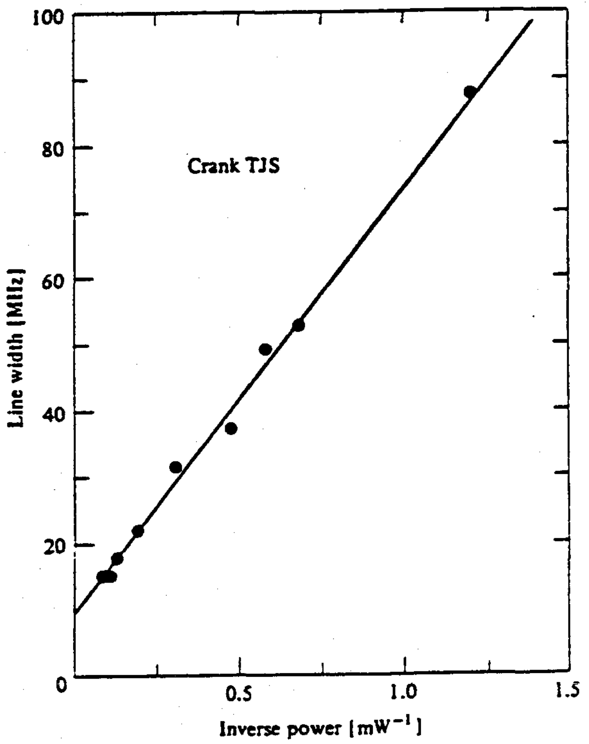


Figure 10-12 The measured dependence of the spectral linewidth of a semiconductor laser on the power output. (After Reference [19].)

The big difference, over nine orders of magnitude, between the limiting linewidth of conventional, say gas lasers and semiconductor lasers, is due mostly to the very short photon lifetime t_c in semiconductor laser resonators. At a given power output we have from (10.7-30) $(\Delta\nu)_{\text{laser}} \propto (\Delta\nu_{1/2})^2 \propto t_c^{-2}$. In the above examples we obtained $t_c = 3 \times 10^{-8}$ s in the case of the He–Ne laser, and $t_c = 5 \times 10^{-12}$ s in the semiconductor laser. Since $t_c \sim ln/c(1 - R)$, the main hope for increasing t_c in a semiconductor laser, thus decreasing the linewidth $(\Delta\nu)_{\text{laser}}$, is to increase l by placing the laser in an external resonator and by using high reflectance mirrors $R \sim 1$. Semiconductor laser linewidths in the kilohertz regime are obtainable.

An actual (measured) GaAs/GaAlAs semiconductor laser, Lorentzian field spectrum is shown in Figure 10-13.

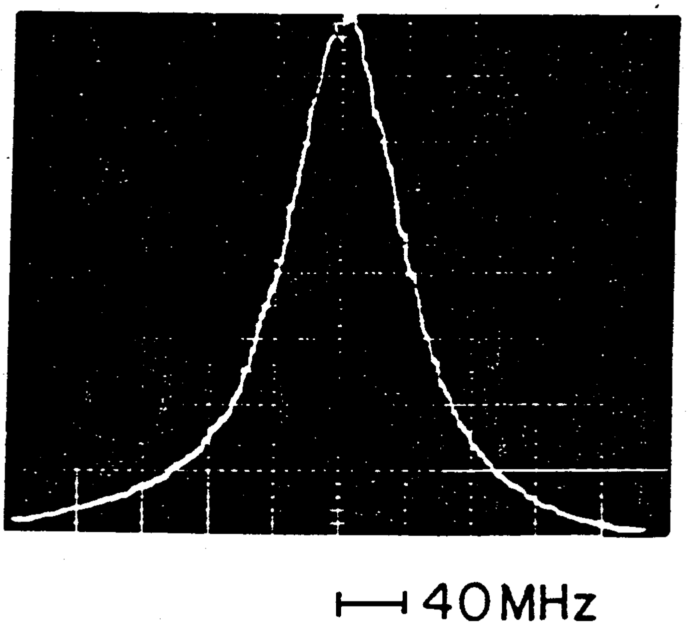


Figure 10-13 The measured Lorentzian field spectrum $S_e(\omega)$ of a semiconductor laser. (After Reference [19].)

10.8 COHERENCE AND INTERFERENCE

In Section 10.7 [Equation (10.7-25)] we have derived the following expression for the autocorrelation function of the single-mode laser field

$$C_{\mathscr{E}}(\tau) \equiv \langle \mathscr{E}(t)\mathscr{E}(t+\tau)\rangle \propto \cos \omega_0\tau e^{-\mu\tau/4\bar{n}_c}$$

$$= \cos \omega_0\tau e^{-\tau/\tau_c}$$

$$\tau_c = \frac{4\bar{n}t_c}{\mu} \qquad (10.8\text{-}1)$$

where $\bar{n}$ is the number of photons inside the resonator, $\mu = N_2/(N_2 - N_1)$ and t_c is the photon lifetime (the decay time constant for the mode optical energy if the gain mechanism were turned off).

The parameter τ_c is called the coherence time of the laser field. According to (10.7-29) it is equal to $2/(\Delta\omega)_{\text{laser}}$ where $(\Delta\omega)_{\text{laser}}$ is the laser output field linewidth. In practical terms it is the time duration during which we can count on the laser to act as a well-behaved sinusoidal oscillator with a well-defined phase. If we try and correlate (by means to be discussed below) the laser field with itself using a time delay exceeding τ_c, the result approaches zero. One form of a field $\mathscr{E}(t)$ that will display this behavior is shown in Figure 10-14. The field undergoes a phase memory loss on the average every τ_c seconds. It is intuitively clear that performing the autocorrelation operation as defined by the first equality of (10.8-1) will yield a result whose rough features agree with the form $(\cos \omega_0\tau)e^{-\tau/\tau_c}$.

Next we will consider how the autocorrelation function $C_{\mathscr{E}}(\tau)$ is obtained in practice. The configuration used most often is the Michelson interferometer illustrated in Figure 10-15. An input field $\mathscr{E}_i(t)$ is split into two components. One of these fields is delayed relative to the second by a time delay

$$\tau = \frac{2(L_1 - L_2)}{c} \tag{10.8-2}$$

The two fields are then incident on a square-law detector whose current constitutes the useful output of the experiment.

Assuming equal division of power, the total optical field at the detector plane is

$$\mathscr{E}_d(t) = \mathscr{E}(t) + \mathscr{E}(t + \tau) \tag{10.8-3}$$

According to the discussion of Section 11.1, which the student is advised to preview at this point, the output current of the detector is

$$i_d = a\overline{\mathscr{E}_d^2(t)} \tag{10.8-4}$$

a is some constant that is irrelevant in the present discussion, and the bar

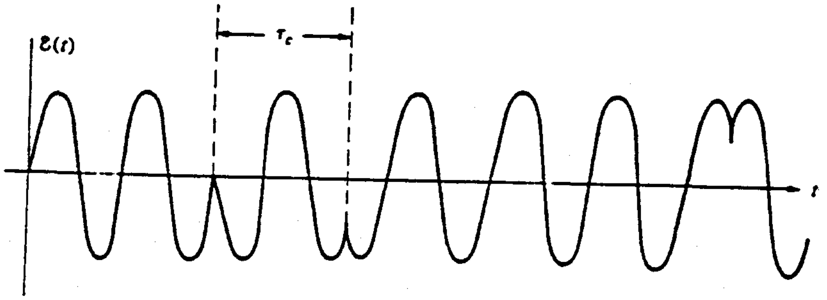


Figure 10-14 A sinusoidal field whose phase coherence is interrupted on the average every τ_c seconds.

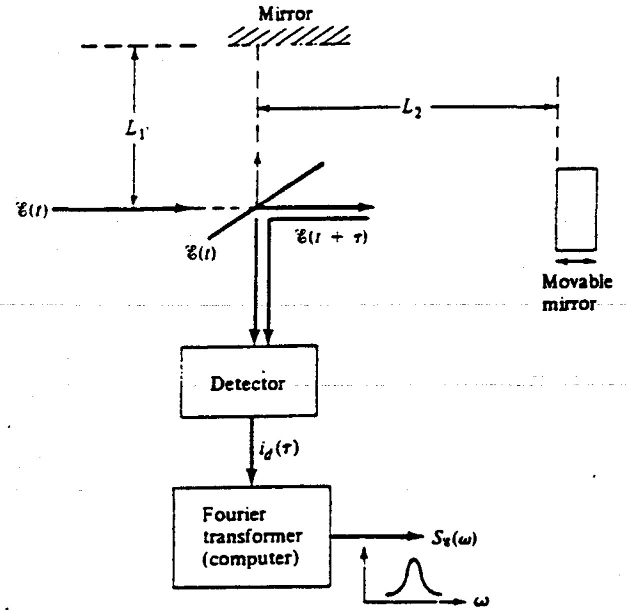


Figure 10-15 A Michelson interferometer "splits" an input beam into a two-component beam and then recombines them with a controlled time delay $\tau = 2(L_1 - L_2)/c$.

indicates, as it does throughout this book, time-averaging. The duration of this averaging depends on the detector and its associated electrical circuitry and in the very fastest detectors may be as short as 10^{-11} s. It is thus *always* very long compared to the optical field period which is $\sim 10^{-15}$ s.

The detector output is then

$$\begin{aligned} i_d &= a[\overline{\mathscr{E}^2(t)} + \overline{\mathscr{E}^2(t + \tau)} + \overline{2\mathscr{E}(t)\mathscr{E}(t + \tau)}] \\ &= 2a[\overline{\mathscr{E}^2} + \overline{\mathscr{E}(t)\mathscr{E}(t + \tau)}] \end{aligned} \tag{10.8-5}$$

since $\overline{\mathscr{E}^2(t)} = \overline{\mathscr{E}^2(t + \tau)} \equiv \overline{\mathscr{E}^2}$. The output current from the detector is thus made up of a dc component $2a\overline{\mathscr{E}^2}$ and a component $2a\overline{\mathscr{E}(t)\mathscr{E}(t + \tau)}$. The ratio of these two current components is, according to (10.2-9), the (normalized) autocorrelation function of the optical field $\mathscr{E}(t)$.

$$\gamma(\tau) \equiv \frac{\tau_{\text{dependent part of } i_d}}{\tau_{\text{independent part}}} \propto C_{\mathscr{E}}(\tau) \tag{10.8-6}$$

The spectral density function $S_{\mathscr{E}}(\omega)$ is obtained, according to (10.2-14), by a Fourier transformation

$$S_{\mathscr{E}}(\omega) = \frac{1}{\pi} \int_{-\infty}^{\infty} C_{\mathscr{E}}(\tau) e^{-i\omega\tau}\, d\tau \tag{10.8-7}$$

The above scheme for obtaining the spectrum (spectral density function) of optical fields is termed Fourier transform spectroscopy, and the configuration of Figure 10-15 is representative of commercial instruments designed for this purpose. These instruments are popular especially in the far infrared (say $\lambda > 10$ μm), since the relative inefficiency of detectors in this wavelength region can be compensated to some degree by a slow scanning rate (of τ) that allows for long integration times and better noise averaging.

A basic result of the Fourier integral transform relationships (10.2-13) and (10.2-14) between $C_{\mathcal{E}}(\tau)$ and $S_{\mathcal{E}}(\omega)$ is that in order to resolve $S_{\mathcal{E}}(\omega)$ to within, say, $\delta\omega$, i.e., to discern structure in $S_{\mathcal{E}}(\omega)$ on the scale of $\delta\omega$, we need to employ time delays $\tau > \pi/\delta\omega$. If we were, as an example, to employ interference spectroscopy to measure the output spectrum of a commercial semiconductor laser with a linewidth of $(\Delta\omega)_{\text{laser}} = 2\pi \times 10^6$ Hz, we would need a delay time τ that could be varied from 0 to 5×10^{-7} s.

In the case of lasers the finite spectral width of the optical field is due predominantly to phase, rather than amplitude, fluctuations. In this case a rather simple technique that involves mixing (heterodyning) the laser field with a delayed version of itself is sufficient to obtain the laser spectrum. This method, which employs a fixed delay instead of the variable delay of the Fourier transform method, is described next.

Delayed Self-Heterodyning of Laser Fields

Consider the configuration of Figure 10-16. An optical field is split into two components that, after a relative path delay t_d, are recombined at a detector. The spectrum of the resulting photocurrent is displayed by a spectrum analyzer. This detection method is referred to as *delayed self-heterodyning* since it involves a "mixing" of the field with a delayed version of itself.

Since the main fluctuation of laser fields is that of the phase and not the amplitude (see comment following Equation 10.7-19), we can approximate the field at the detector by the (complex) phasor

$$E_{\text{total}} = \frac{1}{4} E_0 e^{i\theta(t)} + \frac{1}{4} E_0 e^{i[\omega_0 t_d + \theta(t+t_d)]} \qquad (10.8\text{-}8)$$

This field is illustrated in Figure 10-17. For delays t_d that are considerably shorter than the phase coherence time τ_c of the laser field (defined by Equation (10.7-24)), $\theta(t + t_d) = \theta(t)$ and the magnitude of the total field phasor is a constant as shown in Figure 10-17. Although the phase angle $\theta(t)$ varies randomly, the angle α that determines the magnitude of E_{total} depends only on the difference $\theta(t + t_d) - \theta(t)$ and, in the limit $t_d \ll \tau_c$, does not change with time. The output current from the detector is constant, and nothing can be learned from it about the laser field spectrum. It is clear that we need to

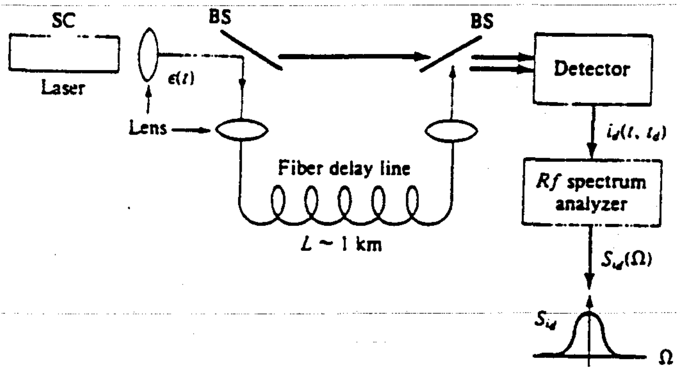


Figure 10-16 An interferometric arrangement employing a fiber delay for obtaining the spectrum $S_{\epsilon}(\omega)$ of the laser field. (After Reference [23].)

consider the case of $t_d \gg \tau_c$. In what follows we will consider the general case of arbitrary t_d.

The output current i_d is proportional to the time average of the square of the total optical field incident on the detector. It is thus proportional (see Equation 1.1-2) to the product of the complex amplitude of this field and its complex conjugate. Using (10.8-8) leads to

$$i_d = SE_0^2 \{e^{i\theta(t)} + e^{i[\omega_0 t_d + \theta(t+t_d)]}\} \times \{e^{-i\theta(t)} + e^{-i[\omega_0 t_d + \theta(t+t_d)]}\} \qquad (10.8\text{-}9)$$

$$= SE_0^2 \{2 + e^{i[\theta(t) - \omega_0 t_d - \theta(t+t_d)]} + e^{-i[\theta(t) - \omega_0 t_d - \theta(t+t_d)]}\} \qquad (10.8\text{-}10)$$

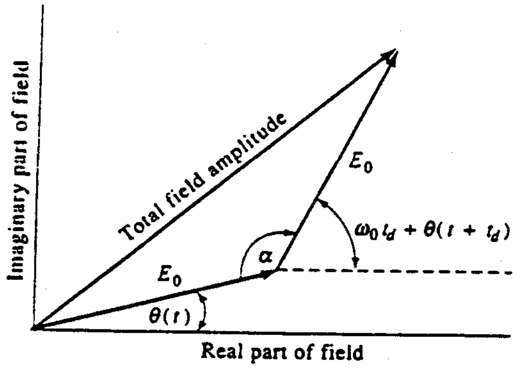


Figure 10-17 Construction showing the total optical field at the detector. For short delays, $t_d \ll \tau_c$, $\theta(t + t_d) = \theta(t)$ so that α and, consequently, the total field amplitude are constant.

where S is a constant depending on the detector. We will derive the spectrum of i_d by employing the Wiener–Khintchine theorem (Equation 10.2-14) so that first we need to obtain the autocorrelation function of $C_{i_d}(\tau)$ of the current i_d. Defining as in Equation (10.7-19)

$$\Delta\theta(t, \tau) \equiv \theta(t + \tau) - \theta(t) \tag{10.8-11}$$

We have reasoned in the last section (see discussion following Equation 10.7-21) that $\Delta\theta(t, \tau)$ is a random Gaussian variable. It follows that the difference $\Delta\theta(t, \tau) - \Delta\theta(t + t_d, \tau)$ is also a Gaussian variable so that, in a manner identical to that used to derive Equation (10.7-24), we obtain

$$\langle e^{i[\Delta\theta(t,\tau)-\Delta\theta(t+t_d,\tau)]}\rangle = e^{-1/2\langle[\Delta\theta(t,\tau)-\Delta\theta(t+t_d,\tau)]^2\rangle} \tag{10.8-12}$$

Now

$$\langle[\Delta\theta(t, \tau) - \Delta\theta(t + t_d, \tau)]^2\rangle = 2\langle[\Delta\theta(\tau)]^2\rangle - 2\langle\Delta\theta(t, \tau)\Delta\theta(t + t_d, \tau)\rangle \tag{10.8-13}$$

where we used

$$\langle[\Delta\theta(t, \tau)]^2\rangle = \langle[\Delta\theta(t + t_d, \tau)]^2\rangle \equiv \langle[\Delta\theta(\tau)]^2\rangle$$

From the equation preceding (10.7-10) and putting $t = \tau$

$$\langle[\Delta\theta(\tau)]^2\rangle = \frac{\mu\tau}{2\bar{n}t_c} = \frac{2\tau}{\tau_c} \qquad \tau_c = 4\bar{n}t_c/\mu \tag{10.8-14}$$

Using (10.8-13) and (10.8-14) in (10.8-12) and (10.8-10), we obtain

$$C_{i_d}(\tau) = S^2E_0^4\,[4 + 2e^{-\frac{|\tau|}{(\tau_c/2)}}\,e^{\langle\Delta\theta(t,\tau)\Delta\theta(t+t_d,\tau)\rangle}] \tag{10.8-15}$$

Special Case $t_d \gg \tau_c$

In the special, but important, long delay case $t_d \gg \tau_c$, we have

$$\lim_{t_d\to\infty} \langle\Delta\theta(t, \tau)\Delta\theta(t + t_d, \tau)\rangle \longrightarrow 0 \tag{10.8-16}$$

and

$$C_{i_d}(\tau)_{t_d\gg\tau_c} = S^2E_0^4\,(4 + e^{-\frac{|\tau|}{(\tau_c/2)}}) \tag{10.8-17}$$

Employing (10.7-14) or using directly the results of (10.7-28), we obtain the following expression for the spectral density of the current i_d

$$S_{i_d}(\Omega)_{t_d \gg \tau_c} = \frac{2S^2E_0^4}{\pi}\left[\frac{\left(\dfrac{4}{\tau_c}\right)}{\left(\dfrac{2}{\tau_c}\right)^2 + \Omega^2} + 4\pi\delta(\Omega)\right] \tag{10.8-18}$$

The spectrum thus consists of a dc, $4\pi\delta(\Omega)$, term plus a Lorentzian distribution centered (if we count negative frequencies $\Omega < 0$) on $\Omega = 0$ with a full width at half maximum of

$$(\Delta\Omega)_{\text{FWHM}} = \frac{4}{\tau_c} = 2(\Delta\omega)_{\text{laser}} \tag{10.8-19}$$

The last equality, derived from (10.7-29) states that the width of the spectrum of the photo-detected current in the limit $t_d \gg \tau_c$ is twice that of the laser field.

The rigorous treatment of the general case involving arbitrary values of the delay t_d is beyond the scope of this book, since it requires a knowledge of the function $\langle\Delta\theta(t, \tau)\Delta\theta(t + t_d, \tau)\rangle$. The derivation of this function involves the solution of the nonlinear, noise-driven laser equation. The result is (see Reference [22])

$$\langle\Delta\theta(t, \tau)\Delta\theta(t + t_d, \tau)\rangle = \frac{2\tau}{\tau_c} - \frac{2}{\tau_c}\min(\tau, t_d) \tag{10.8-20}$$

where $\min(\tau, t_d)$ signifies the smallest of τ and t_d. The last result together with (10.8-13, 10.8-20) when substituted in (10.8-15) give

$$C_{i_d}(\tau) \equiv \langle i_d(t)i_d(t + \tau)\rangle = S^2E_0^4\,[4 + 2e^{-(2\tau/\tau_c)\min(\tau, t_d)}] \tag{10.8-21}$$

$$\begin{aligned} S_{i_d}(\Omega) &= \frac{1}{\pi}\int_{-\infty}^{\infty} C_{i_d}(\tau)e^{-i\Omega\tau}\,d\tau \\ &= 8S^2E_0^4\,(1 + e^{-2t_d/\tau_c})\,\delta(\Omega) \\ &\quad + \left(\frac{8S^2E_0^4}{\pi\tau_c}\right)\frac{\left[1 - e^{-2t_d/\tau_c}\left(\cos\Omega t_d + \dfrac{2\sin\Omega t_d}{\Omega\tau_c}\right)\right]}{\left(\dfrac{2}{\tau_c}\right)^2 + \Omega^2} \end{aligned} \tag{10.8-22}$$

The integration leading to (10.8-22) is long but straightforward. Equation (10.8-22) reduces, as it should, to (10.8-18) when $t_d/\tau_c \rightarrow \infty$. In summation,

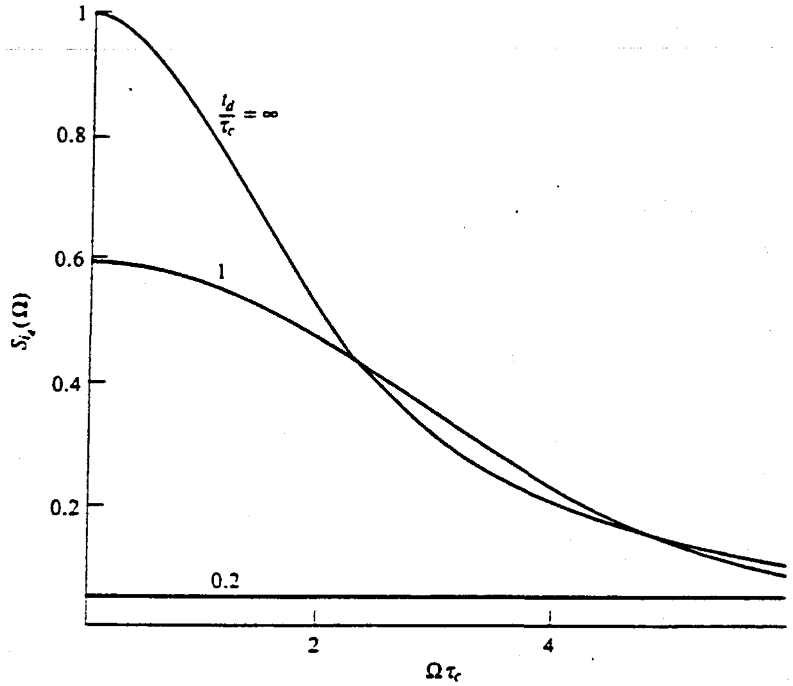


Figure 10-18 The spectral density $S_{i_d}(\Omega)$, as given by Equation (10.8-22) of the photocurrent in a delayed self-heterodyne detection of the output of a laser. The ratio of the delay time (t_d) to the laser field coherence time (τ_c) is a parameter. The frequency abcissa is in units of τ_c^{-1}. $\tau_c = (\Delta\nu)^{-1}_{\text{laser}}$.

we recall that only in the case $t_d/\tau_c \gg 1$, i.e., a long relative delay, is the spectrum $S_{i_d}(\Omega)$ a Lorentzian. A typical spectrum of a semiconductor laser obtained with a setup similar to that of Figure 10-16 is shown in Figure 10-13. A plot of the theoretical spectra of (10.8-22) for the cases $t_d/\tau_c = \infty, 1, 0.2$ is contained in Figure 10-18.

10.9 ERROR PROBABILITY IN A BINARY PULSE CODE MODULATION SYSTEM

The simplicity and reliability of digital processing by integrated electronic circuits has made it increasingly attractive to transmit information in the form of binary pulse trains. For optical communication systems, the analog data to be transmitted are coded into a train of 1 and 0 electrical pulses so that each pulse carries one bit of information. The electrical signal thus generated is impressed, say, by means of a modulator, on an optical beam. resulting in an optical train pulse. The optical signal having propagated

REF. S

PHYSICAL REVIEW A VOLUME 33, NUMBER 5 MAY 1986

Comparison between active- and passive-cavity interferometers

Alex Abramovici and Zeev Vager
Department of Nuclear Physics, The Weizmann Institute of Science, Rehovot 76100, Israel

(Received 11 October 1985)

Both active- and passive-cavity interferometers are considered at present for displacement sensing in broadband gravitational-radiation detectors. In spite of an apparent difference between the noise sources that limit their performance, we show that active- and passive-cavity interferometers are of the same strain (or displacement) sensitivity if identical resonators and stored fields of equal intensity are assumed.

I. INTRODUCTION

There is a strong belief that long-base-line laser interferometers will sooner or later achieve sensitivities that will allow detection of gravitational radiation (GR). Accordingly, large interferometers have been built and are currently undergoing a process of testing and upgrading[1-3] while very large ones are seriously being considered.[4,5] In order to achieve as long an effective base line as possible, the arms of these interferometers contain either optical delay lines or optical cavities. These are passive optical systems, since they receive light from an external source, usually an argon-ion laser.

Active-cavity systems that take advantage of the very high sensitivity of laser frequency to changes in resonator length are an alternative approach to GR detection.[6,7] A prototype gravitational-radiation detector employing an active-cavity displacement sensor has recently been constructed, with a noise level equivalent to displacements of 3×10^{-15} cm/Hz$^{1/2}$, above 2 kHz.[8]

In the early development stages, there was hope that active-cavity detectors could be made more sensitive than passive-cavity detectors.[6] On the other hand, since the phase noise of laser light, due to spontaneous emission, is higher than the one due to shot noise that limits the performance of passive interferometers, it has been argued that active-cavity systems are intrinsically less sensitive than passive ones. We show in what follows that for fields of equal intensities stored within identical resonators, active- and passive-cavity interferometers are of the same displacement sensitivity, although the types of noise which limit their performance are apparently of different nature.

II. ACTIVE-PASSIVE COMPARISON

Consider the interferometer geometry shown in Fig. 1, consisting essentially of two perpendicular Fabry-Perot cavities and a beam splitter. If an active medium is added to the resonators, they become lasers operating in a heterodyne configuration characteristic of an active-cavity interferometer.[8] If, on the other hand, the active medium is removed and the system is fed light from an external laser, the geometry of Fig. 1 corresponds to a passive-cavity interferometer of the type used for the gravitational-radiation detector[3] at the California Institute of Technology. For the sake of comparison, we shall assume a passive- and an active-cavity interferometer with identical Fabry-Perot cavities.

A parameter crucial to interferometer performance is the amount of light handled by the system. For convenience, we chose to describe this parameter by I_{st}, the intensity of the light stored within the resonators, integrated over the cross section of the beam. Active and passive systems will thus be compared under the assumption that they employ fields with the same I_{st}.

In an active-cavity interferometer, the laser beams of frequencies ω_1 and ω_2 combine at the beam splitter (see Fig. 1) and provide, after photodetection, a beat signal of frequency $\omega_B=\omega_2-\omega_1$. When the mirror spacings in the two lasers change by ΔL_1 and ΔL_2, the beat frequency changes by an amount $\Delta\omega_B=\omega(\Delta L_2-\Delta L_1)/L$, where ω is the mean value of ω_1 and ω_2, while L is the mean value of L_1,L_2, the optical length of the laser resonators. Changes in ω_B are monitored by frequency demodulation. The intrinsic noise that limits active-cavity interferometer performance consists of laser frequency fluctuations due

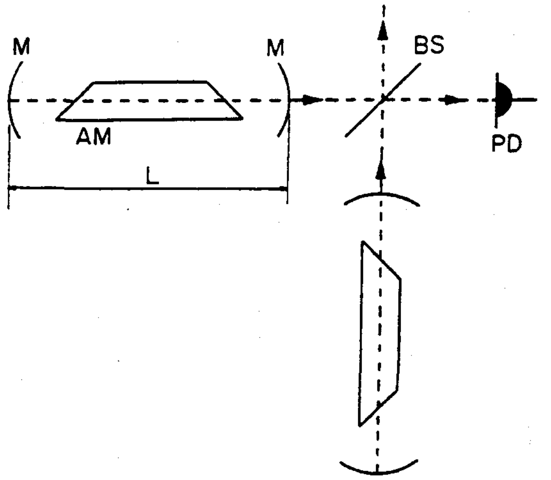


FIG. 1. Optical layout of the active-cavity interferometer. M, mirrors; AM, active medium; BS, beam splitter; PD, photodiode.

©1986 The American Physical Society

to spontaneous emission. The spectral density δL^2 of the smallest displacements detectable with an active-cavity interferometer[8] can be written as

$$\delta L_A^2=\frac{\hbar\omega L^2}{I_{st}Q^2(1-R_1R_2)}\,, \tag{1}$$

where the subscript A stands for the active cavity, Q is the quality factor of the resonators, and R_1,R_2 are the power reflectivities of the laser mirrors.

The limit to the measurement of small displacements with a Michelson interferometer is determined by the shot noise generated by the photon flux impinging on the photodetector. The spectral density of displacements equivalent to the shot noise is[9,10]

$$\delta L_P^2=\frac{\hbar\omega}{2\eta I(\phi')^2}\,, \tag{2}$$

where the subscript P stands for the passive system, η is the quantum efficiency of the photodetector, I is the total intensity of the light leaving the interferometer, and $\phi'=d\phi/dL$ is the sensitivity of the phase shift in each arm to changes in arm length.

Optimization of Eq. (2) is now carried out for an interferometer that contains a passive Fabry-Perot cavity in each arm. The sensitivity of this configuration is compared in Sec. III with that of a Michelson interferometer which employs delay lines.

Assume that one of the resonator mirrors has zero transmittance and an amplitude reflectivity r such that $R_1=r^2$. The fractional loss the beam experiences for each reflection on this mirror thus is $1-r^2$. Further, assume that the coupling mirror has amplitude reflectivity r_c, such that $R_2=r_c^2$ and that the losses are the same as for the high reflector. If we choose r and r_c real and negative, the amplitude transparency coefficient t has to be taken purely imaginary (see, e.g., Ref. 11). Coupler transmittance thus is $T_c=-t^2=r^2-r_c^2$. Under steady-state conditions, the incident amplitude A, the outcoming amplitude B, and the amplitude C of the stored field (see Fig. 2) are related as follows:

$$B=r_cA+tr\Phi C\,, \tag{3}$$

$$C=tA+r_cr\Phi C\,, \tag{4}$$

where $\Phi=\exp(i4\pi L/\lambda)$ is the phase factor corresponding to a full round trip in the resonator. Solving Eqs. (3) and (4) for B and C yields the output-to-input power ratio Θ, the sensitivity of the phase to changes in arm length ϕ', and the relation between the intensities of the output beam and the stored field for the resonator with losses[12] operated as a reflector:

$$\Theta=\frac{I}{I_0}=\frac{(r_c-r^3)^2}{(1-r_cr)^2}\,, \tag{5}$$

$$\phi'=\left[\frac{4\pi}{\lambda}\right]\frac{r(r_c^2-r^2)}{(r_c-r^3)(1-r_cr)}\,, \tag{6}$$

$$I=\frac{(r_c-r^3)^2}{2(r^2-r_c^2)}I_{st}\,. \tag{7}$$

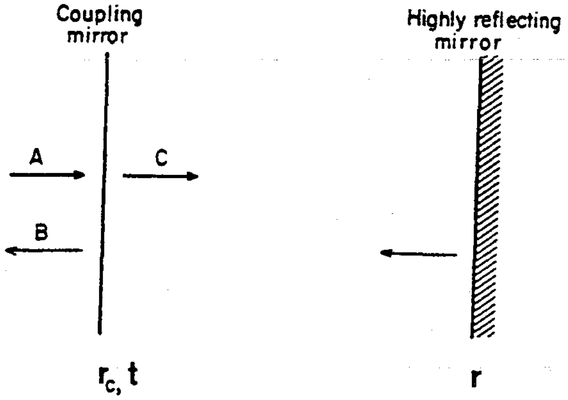


FIG. 2. Optical resonator geometry. A and B are the amplitudes of the incoming and outgoing fields, C is the amplitude of the field transmitted by the coupling mirror, t is the amplitude transmissivity of the coupling mirror, and r_c,r are amplitude reflectivities.

Equations (5)–(7) are evaluated for an optical cavity operated at resonance ($\Phi=1$), where ϕ' is the largest.

For given mirror losses, i.e., for given r, it is found that the coupler reflectivity that minimizes δL_P^2 is $r_c=r^3$. Using also the explicit form of the cavity quality factor $Q=2\pi L/\lambda(1-r_cr)$, the noise limit for an interferometer with arms consisting of passive optical resonators becomes

$$\delta L_P^2=\frac{\hbar\omega L^2}{4\eta r^4I_{st}Q^2(1-r^4)}\,. \tag{8}$$

Under the assumption that the active and the passive interferometer employ identical Fabry-Perot cavities and contain stored fields of the same intensity, comparison of Eqs. (1) and (8) yields

$$\frac{\delta L_A^2}{\delta L_P^2}=2\eta\,, \tag{9}$$

where the fact that $r^4\sim 1$ has been taken into account. In other words, the ultimate sensitivity of the interferometer is the same, irrespective of whether it employs active or passive optical resonators. The reason for this is that in both cases the noise level is determined by the stored energy, on one hand, and by the magnitude of the losses, on the other.

III. COMPARISON BETWEEN PASSIVE CAVITY AND DELAY LINE

In an optical delay line, the light beam is repeatedly bounced back and forth between two highly reflecting mirrors of power reflectivity $R=r^2$. The noise level for an interferometer with arms consisting of delay lines is obtained from Eq. (2) by replacing $I=\Theta I_0$ and taking into account the fact that for z reflections $\Theta=R^z$ and $\phi'=(2\pi/\lambda)(z+1)$:

$$\delta L_D^2=\frac{\hbar\omega}{2\eta I_0k^2R^z(z+1)^2}\,, \tag{10}$$

where the subscript D stands for the delay line and $k=2\pi/\lambda$.

For given R, the minimum value of δL_D^2 as a function of z is[13]

$$\delta L_D^2=\frac{1.85\hbar\omega(1-r^2)^2}{2\eta I_0 k^2}\ . \tag{11}$$

The optimum noise limit [Eq. (8)] for an interferometer with passive optical cavities is rewritten by using Eqs. (2), (5), and (6) and the optimum condition $r_c=r^3$:

$$\delta L_P^2=\frac{\hbar\omega(1-r^2)^2}{2\eta I_0 k^2 r^6}\ . \tag{12}$$

Comparison between Eqs. (11) and (12) yields

$$\frac{\delta L_D^2}{\delta L_P^2}=1.85\ , \tag{13}$$

since $r^6\sim 1$.

Equation (13) shows that the optimum noise limits are similar for interferometers using either passive Fabry-Perot cavities or optical delay lines.

We conclude this section by stressing that, for given mirror losses, the Fabry-Perot cavity is optimized by satisfying the condition $r_c=r^3$, while the delay line is optimized by an appropriate choice of z, the number of reflections.[13]

IV. DISCUSSION

The way to obtain the high strain sensitivity required by gravitational radiation detection is highlighted by evaluating the spectral density of the smallest detectable strain $\delta L^2/L^2$ by use of Eq. (8) [or of the equivalent Eq. (1)]:

$$\frac{\delta L^2}{L^2}=\text{const}\times\frac{1-r^4}{L^2 I_{st}}\ , \tag{14}$$

where the constant has dimensions energy×length^2. Equation (14) follows e.g., from Eq. (8) by replacing the explicit form of Q and the optimum coupler reflectivity $r_c=r^3$.

In an active-cavity system, I_{st} can be increased by employing a high-gain amplifying medium and high pumping levels. For passive-cavity devices, one has to increase the power of the laser beam injected into the cavities. It is also desirable to have mirrors of lowest possible losses and a long cavity. In the case of a passive resonator, the resulting narrow bandwidth sets critical frequency stability requirements upon the laser which provides the light to the interferometer. On the other hand, a long laser resonator means close mode spacing. The resulting large number of modes that may simultaneously oscillate is not an attractive possibility for an active-cavity interferometer. Thus, the natural way to improve the sensitivity of an active-cavity system is to increase both Q and I_{st} by use of top quality optics and by choosing a high-gain active medium, while keeping the resonators reasonably short.

It should be kept in mind that the active-passive cavity comparison in Sec. II has been made under the implicit assumption that the active medium of the laser does not affect resonator Q in any way. If very high quality mirrors and low output coupling are used, this requires the active medium itself to be of very low scattering and that virtually no absorption should take place, except for the laser transition itself. Moreover, worries have been expressed about the active medium contributing excess noise, e.g., gas pressure fluctuations and plasma noise in a He-Ne laser tube,[14] thus preventing operation at the spontaneous-emission noise limit. Nevertheless, we have been able to operate an active-cavity detector employing two low-power He-Ne lasers at the spontaneous-emission noise limit, whereas we measured a displacement noise level of 3×10^{-15} cm/Hz$^{1/2}$ in the kilohertz range.[8] While gas-pressure fluctuations and plasma noise might become a problem with long and/or high-power gas lasers, we expect that it will be possible to improve displacement sensitivity by 2—3 orders of magnitude by employing a properly selected solid-state amplifying medium.[8] Finally we note that, in the case of passive Fabry-Perot cavities, it might prove difficult to match the reflectivities of the mirrors as required by the optimum condition $r_c=r^3$.

V. CONCLUSIONS

It has been shown that for identical resonators and stored fields of equal intensity, active- and passive-cavity interferometers considered for gravitational-radiation detection are of the same sensitivity. However, the practical approach to high sensitivity is different for the two kinds of interferometers. Thus, high quality optics, high gain amplifying medium, and short resonators are the best way for active-cavity systems. For practical reasons, only a limited amount of optical power can presently be injected into passive-cavity systems. This is compensated for by an increase in arm length.

[1] H. Billing, K. Maischberger, A. Ruediger, R. Schilling, L. Schnupp, and W. Winkler, J. Phys. E 12, 1043 (1979).

[2] R. Schilling, L. Schnupp, D. H. Shoemaker, W. Winkler, K. Maischberger, and A. Ruediger, Max Planck Institute for Quantum Optics Report No. MPQ 88, 1984 (unpublished).

[3] R. W. P. Drever, in *Proceedings of the NATO Advanced Study Institute on Gravitational Radiation, Les Houches, 1982*, edited by N. Deruelle and T. Piran (North-Holland, Amsterdam, 1983).

[4] P. Linsay, P. Saulson, and R. Weiss (unpublished).

[5] K. Maischberger, A. Ruediger, R. Schilling, L. Schnupp, D. Shoemaker, and W. Winkler, Max Planck Institute for Quantum Optics Report No. MPQ 96, 1985 (unpublished).

[6] M. Weksler, Z. Vager, and G. Neumann, Appl. Opt. 19, 2717 (1980).

[7] S. N. Bagayev, V. P. Chebotayev, A. S. Dychkov, and V. G.

Goldort, Appl. Phys. 25, 161 (1981).

[8] A. Abramovici, Z. Vager, and M. Weksler, J. Phys. E (to be published).

[9] R. L. Forward, Phys. Rev. D 17, 379 (1978).

[10] W. A. Edelstein, J. Hough, J. R. Pough, and W. Martin, J. Phys. E 11, 710 (1978).

[11] A. E. Siegman, *An Introduction to Lasers and Masers* (McGraw-Hill, New York, 1971).

[12] P. Giacomo, Rev. Opt. 35, 442 (1956).

[13] R. Weiss, Quart. Prog. Rep. Res. Lab. Electronics MIT 105, 54 (1972).

[14] A. Brillet and P. Tourrenc, in *Proceedings of the NATO Advanced Study Institute on Gravitational Radiation, Les Houches, 1982,* edited by N. Deruelle and T. Piran (North-Holland, Amsterdam, 1983).

PHYSICAL REVIEW A VOLUME 35, NUMBER 6 MARCH 15, 1987

Passive versus active interferometers: Why cavity losses make them equivalent

J. Gea-Banacloche
Max-Planck Institut für Quantenoptik, D-8046 Garching bei München, West Germany
and Center for Advanced Studies, Department of Physics and Astronomy, University of New Mexico, Albuquerque, New Mexico 87131

(Received 16 September 1986)

Active and passive interferometers (as used, for example, for rotation-rate sensing or gravitational-wave detection) are known to have essentially the same ultimate sensitivity, although they appear to work differently, have signals of different sizes, and be limited by different kinds of noise (shot noise for the passive case, spontaneous emission for the active case). This paper explains this remarkable coincidence. The underlying physics common to both systems is brought forth and the role of the losses in limiting the sensitivity is clarified. The possibility of squeezing the field is explicitly considered; it is shown when it can or cannot help, and why.

I. INTRODUCTION

In the use of interferometry for certain high-precision measurements (e.g., rotation sensors[1] or gravitational-wave detectors[2]) it is common to distinguish between passive devices (when light from an external source is injected into an empty cavity) and active devices (where the light is internally generated by a gain medium inside the cavity). In both cases what is being measured is the detuning of the cavity from resonance, which is proportional to the signal (rotation rate, for example) that one is really interested in. In a passive cavity the detuning causes a phase shift; in an active cavity it causes a frequency shift, that is, a phase shift which grows linearly with time. For this reason active cavities appear to be potentially more sensitive than passive cavities over a sufficiently long measurement time.

On the other hand, when the noise limiting the performance of both kinds of systems is taken into account, it is found that the sensitivity (signal-to-noise ratio) is essentially the same in both. The signal-to-noise ratios for *optimized* passive and active systems differ only by numerical factors depending on the experimental arrangement, but the order of magnitude and the dependence on the cavity losses and the power is the same. The limit for the passive device is usually derived by considering shot noise at the photodetector; for the active device, instead, it is given by fluctuations in the laser phase—the fluctuations which give rise to the laser linewidth and which arise from spontaneous emission in the gain medium.[3]

From a fundamental point of view, this is a profoundly unsatisfactory result. It is as if two different systems and two different noise sources somehow conspired to produce the same result. Even more amazing is the fact that essentially the same limit recurs in every conceivable detection scheme, in very different experimental arrangements, from laser gyroscopes[4] to gravitational-wave detectors.[5] Yet no explanation for this remarkable coincidence appears to have been presented in the literature.

One wonders, of course, whether there is a sort of fundamental limit lurking in the background. Yet neither shot noise nor spontaneous emission noise are ultimate limits to signal processing. Shot noise can be reduced by squeezing the vacuum,[6,7] and phase-sensitive amplifiers may be conceived which need not degrade appreciably the signal-to-noise ratio of one quadrature (the phase, for instance) of the signal they amplify (in the language of Caves's classic paper,[8] they may have negligible added noise for that quadrature, although they still have to amplify the signal's inherent noise along with the signal itself). Such an amplifier, operating with a squeezed-state input, would have negligible spontaneous-emission-induced phase fluctuations.

Where, then, does the ultimate limit come from? A careful study of the problem reveals that the cavity losses play a crucial role, and this note explains why. The coincidence of the limits for active and passive devices is not, as it could not be, a coincidence at all: the differences between the two kinds of devices are not, in a way, as deep as one might have expected; and, from a certain point of view, it is the fluctuation-dissipation theorem which lies at the heart of the matter. In the process of reaching this conclusion, just about every fundamental problem in quantum optics, from squeezing to vacuum fluctuations and the laser linewidth, makes at least a cameo appearance.

II. PASSIVE CAVITIES

The first point that needs to be established is what is common to the response of both active and passive cavities to a cavity detuning, and we begin by showing that one can look at the passive cavity in a way that makes it look very similar to an active one, and which shows exactly what it is that the active one does that makes it different. All the discussions that follow will concentrate on the field inside the cavity only, in a single mode, and inquire as to how well its phase is defined; the problems associated with extracting the light and actually performing the measurement will be ignored, since the fundamental limit may be found in the intracavity field already.

Consider, then, first the response of a passive cavity to an elementary excitation of the field; specifically, consider the free decay of a mode of the electromagnetic field, of nominal frequency ω, inside a cavity which is slightly detuned (let the cavity resonant frequency be Ω and the de-

©1987 The American Physical Society

tuning $\delta\Omega=\omega-\Omega$). Semiclassically, the boundary conditions result in a difference equation which, for small losses, may be approximated by a differential equation for the (slowly varying) complex amplitude $\mathscr{E}(t)$,

$$\dot{\mathscr{E}}=(i\delta\Omega-\gamma)\mathscr{E}\ , \tag{1}$$

where γ is the decay rate due to losses. Writing $\mathscr{E}(t)=E(t)e^{-i\phi(t)}$, one sees from Eq. (1) that the phase does grow linearly with time

$$\dot{\phi}(t)=\delta\Omega\ , \tag{2}$$

but the amplitude is damped,

$$\dot{E}(t)=-\gamma E(t)\ . \tag{3}$$

Equation (2) expresses the essential similarity between the active and passive cavities, Eq. (3) their only essential difference; namely, that in the passive case the field dies away in a time of the order of γ^{-1}. It is important to realize, in particular, that the linear growth of the phase (2), which is usually said to be characteristic of the active systems, is actually already present in the passive cavity. The decay of the field, however, prevents one from observing it for times much longer than γ^{-1} [compare the discussion below, in terms of X_2; in particular Eq. (6)]. The contribution of the active medium in an active system, therefore, is only to keep the field from decaying by amplifying it (in a phase-preserving way, that is, coherently), thus making the phase growth (2) observable.

Accordingly, Eq. (2) might be derived by simply taking the phase and amplitude evolution equations for an active system (for example, the ring gyro equations from Ref. 1) and formally removing the active medium by setting all the gain coefficients equal to zero; the result is nothing but Eq. (1), which shows that (2) may indeed be regarded as a property of the passive cavity alone. [Equation (1) may, of course, also be established directly for a passive cavity, as mentioned above; for instance, one may take the evolution equations for an ordinary Fabry-Perot (see, for example, Ref. 9) and just set the injected field equal to zero: then (1) gives the free decay of the field in the cavity.]

The main point of this discussion is that it is legitimate to consider an active system as just a passive cavity with an amplifying medium inside. The consequences of this will be discussed in Sec. III.

Some passive schemes do actually exhibit a "growing phase" in a sense; for instance, the "delay line" or Michelson-type interferometers in gravity-wave detection (the phase difference between the two arms grows with every round trip of the light between the mirrors), and single-pass, many-turn optical fibers for rotation-rate sensing (the phase difference between the counterpropagating beams grows with every turn). Passive cavities of the Fabry-Perot—type, instead, are used most often with an injected field to keep the intensity inside constant. Then the phase of the intracavity field does not grow beyond a maximum value $\phi_{\max}=\delta\Omega/\gamma$, because the field that has been in the cavity for a long time (accumulating a large phase shift) dies away, and "fresh" light, with a constant phase, is continually coming in to replace it. We shall regard this system as being roughly equivalent to a continuous repetition of an "elementary measurement," in which some intracavity field is allowed to evolve freely, sample the cavity, and eventually die away; then another fresh field is allowed to do the same, then another, etc. This point of view gives the correct result for the sensitivity of a Fabry-Perot—type passive device [Eq. (12) below] aside from numerical factors which depend on the experimental setup, and measurement strategy.

To proceed with the study of one of these "elementary measurements," it is convenient to replace the phase ϕ, which is not a good observable, by something more suitable. We introduce the quadratures X_1 and X_2 of the electric field by the equation

$$E(t)=e^{-i\phi_0}(X_1+iX_2)\ . \tag{4}$$

Here X_1 and X_2 are real (or, as quantum operators, Hermitian) and ϕ_0 is the initial value of the phase, so that initially $X_2=0$; then, as the phase grows, we may take X_2 to be our signal (X_2 is the phaselike quadrature, X_1 the amplitudelike quadrature). Equations for X_1 and X_2 follow immediately from (1):

$$\dot{X}_1=-\gamma X_1-\delta\Omega X_2\ , \tag{5a}$$

$$\dot{X}_2=-\gamma X_2+\delta\Omega X_1\ . \tag{5b}$$

These equations are easily integrated. We shall consider only the case when the signal is very small, so that $\delta\Omega t$, and therefore X_2, is always much smaller than 1; then the term in X_2 may be neglected in (5a). The solution for X_2 grows at first linearly, and then it is damped,

$$X_2(t)=e^{-\gamma t}\delta\Omega X_1(0)t\ ; \tag{6}$$

it is maximum precisely when $t=\gamma^{-1}$, so that the maximum signal equals $X_{2\max}=e^{-1}\delta\Omega X_1(0)/\gamma$. Note that $X_1(0)$ is just the electric field amplitude at $t=0$. We shall use units such that, quantum mechanically, $X_1^2+X_2^2=n+\frac{1}{2}$, where n is the photon number operator. Then $\langle X_1(0)\rangle \simeq \langle n\rangle^{1/2}$, if $\langle n\rangle$ is large.

We need to enquire now about the precision with which the signal (6) can be known—that is, about the "noise." The quantum-mechanical operators for X_1 and X_2 have an intrinsic uncertainty expressed by the relation

$$\Delta X_1\Delta X_2\geq\tfrac{1}{4}\ , \tag{7}$$

but this by itself does not tell us how large or small ΔX_2 has to be. In particular, we might consider a squeezed state with negligible ΔX_2.

The crucial point, however, is that Eq. (6) has been obtained as the solution of a system of equations (5) for a *damped* field. Now, if this field is a quantum-mechanical one, the preservation of the commutation relations (ultimately, the uncertainty principle) requires that the damping mechanism (whatever it is) introduce noise, which will be represented by noise operators in the equations of motion. It is this noise that is going to determine the ultimate sensitivity.

The exact form of the noise operators is model dependent. Their correlations, which are all we need here, are determined by the fluctuation-dissipation theorem;[10] for definiteness, the reader may want to think of the classic model of damping by a bath of harmonic oscillators[11] (with the standard Markov approximation). At any rate, what we have to do is to rewrite Eqs. (5) as Langevin equations,

$$\dot{X}_1=-\gamma X_1+F_1(t)\,, \tag{8a}$$

$$\dot{X}_2=-\gamma X_2+\delta\Omega X_1+F_2(t) \tag{8b}$$

(as explained before, we have neglected the term proportional to X_2 in the equation for X_1). The noise operators F_1 are F_2 and Hermitian. They are uncorrelated in the sense that their Hermitian correlation function $\langle F_1(t)F_2(t')+F_2(t')F_1(t)\rangle$ vanishes [the non-Hermitian correlation function $\langle F_1(t)F_2(t')\rangle$ is purely imaginary]. Most importantly, they satisfy

$$\langle F_1(t)F_1(t')\rangle=\langle F_2(t)F_2(t')\rangle=\tfrac{1}{2}\gamma\delta(t-t')\,. \tag{9}$$

The crucial point, apparent from Eq. (9), is really that the losses are *phase insensitive:* the damping bath puts the same amount of noise in each quadrature. It is for this reason that squeezing is destroyed by losses (as already pointed out by Caves[12]); it is from this fact that a fundamental limit arises.

When the system (8) is integrated, one finds for the noise in X_2

$$\Delta X_2^2(t)\simeq e^{-2\gamma t}(\Delta X_2^2)_0+\tfrac{1}{4}(1-e^{-2\gamma t})\,, \tag{10}$$

aside from terms which are smaller than those kept by a factor of $\delta\Omega/\gamma$ (which was assumed earlier to be very small). Equation (10) shows that, regardless of what the initial noise in the quadrature X_2 is, the noise associated with damping will (because of its phase-insensitive nature) tend to put in X_2 the noise associated with vacuum fluctuations—that is, $\Delta X_1=\Delta X_2=\frac{1}{2}$.

It is now a simple exercise to use Eqs. (6) and (10) to calculate the maximum signal-to-noise ratio. The result depends somewhat, of course, on the initial amount of squeezing that is present [that is, the value of $(\Delta X_2)_0$], but not in order of magnitude: the maximum signal-to-noise ratio is always reached after a time of the order of γ^{-1}; by that time, the noise in X_2 is already of the order of magnitude of that for the unsqueezed vacuum (i.e., $\frac{1}{2}$), and the minimum detectable signal $\delta\Omega$ (defined as the value of $\delta\Omega$ giving a signal-to-noise ratio of unity) is, therefore, of the order of

$$\delta\Omega_{\min}\sim\gamma/X_1(0)\,. \tag{11}$$

We must take into account now the possibility mentioned earlier of repeating the measurement a large number of times. Since each elementary measurement lasts for a time of the order of γ^{-1} over a total measurement time t_m, we may perform $N=\gamma t_m$ elementary measurements, and the signal-to-noise ratio will improve by a factor of $N^{1/2}$. Then the minimum detectable $\delta\Omega$ becomes

$$\delta\Omega_{\min}\sim\left[\frac{\gamma}{\bar{n}t_m}\right]^{1/2}, \tag{12}$$

where we have replaced the $X_1(0)$ of Eq. (11) by $\sqrt{\bar{n}}$, $\bar{n}$ being the average number of photons in the cavity. This is indeed the result obtained for passive interferometers in which one is constantly injecting fresh light, as was said above.

We have seen now the two ways in which losses affect adversely the performance of a (passive) interferometer. First, because the field is damped, the "phaselike quadrature" which carries the information about the signal does not grow past a certain maximum value (reached after a time of the order of γ^{-1}). Second, the losses introduce some noise whose effect is to ensure that, after a time of the order of γ^{-1} again, the noise in that quadrature is the same as for unsqueezed vacuum fluctuations, regardless of whether one started with a squeezed state or not. It is precisely this latter effect which ensures that the "shot-noise limit" calculation gives the same order of magnitude as Eq. (12), since shot-noise may be related in various ways (depending on the experimental arrangement) to vacuum fluctuations at the photodetector.[7,13]

III. ACTIVE SYSTEMS

Since we have identified what limits the sensitivity of a passive interferometer, we might think of doing something about it in the following way (which, as discussed earlier, leads essentially to an "active" scheme): to counteract the damping of the field due to the losses, introduce a gain medium in the cavity which coherently regenerates the signal. Then, with the losses effectively gone from Eq. (8b) (and X_1, the "amplitudelike" quadrature, locked to some saturation value), X_2 would be free to grow linearly with time instead of eventually decaying as in Eq. (6).

When the operation of an active device is understood in this way, the reason why it does not work (better than the passive system, that is) is actually almost obvious: the gain medium cannot but amplify the signal *and* the noise *together.* The active system could not, therefore, have a larger signal-to-noise ratio than the underlying passive system.

This may be formally shown without much difficulty. Assume that the evolution of X_2 is given by Eq. (8b), without the losses, and with a constant $X_1=X_1(0)$. We are neglecting any "added" noise (in the terminology of Ref. 8) introduced by the amplifier which might, therefore, be a totally classical device, or a phase-sensitive amplifier[3] with negligible added noise for the quadrature X_2. It might seem at first sight that we are restricting ourselves to linear amplifiers only, but this is not so. We are only requiring that the amplifier's treatment *of the quadrature* X_2 be, to a good approximation, linear. This had better be the case, at any rate, since otherwise the relating of the output of the device to the signal of interest is not a trivial task; in any event, the validity of this assumption is practically guaranteed in all the cases of interest here (namely the detection of very weak signals, where $\delta\Omega t_m \ll 1$, so that $X_2 \ll 1$). The amplifier may (and, in the case of an ordinary laser medium, will) treat the quadrature X_1 nonlinearly, but the linear approximation will describe its processing of X_2 quite well.

The solution for $X_2(t)$ is then

$$X_2(t)=\delta\Omega X_1(0)t+\int_0^t F_2(t')dt' . \qquad (13)$$

We see that the noise in X_2 does indeed add up, undamped (unlike in the passive case), just like the variable X_2 itself. Using again Eq. (9), we find for the magnitude of this noise

$$\Delta X_2^2=(\Delta X_2^2)_0+\tfrac{1}{2}\gamma t . \qquad (14)$$

Again Eqs. (13) and (14) may be used to investigate the signal-to-noise ratio, and again, when the measurement time is long enough ($t_m>\gamma^{-1}$), the initial amount of squeezing in X_2 is found to make very little difference. Over a total measurement time t_m, the minimum detectable $\delta\Omega$ (with the signal-to-noise ratio equal to 1, as before) is

$$\delta\Omega_{\min}\sim\left[\frac{2\gamma}{\bar{n}t_m}\right]^{1/2} , \qquad (15)$$

which is essentially the same as Eq. (12). The only advantage over Eq. (11) is the one arising from a large number of independent measurements, which we might say the active cavity performs automatically for us, not surprisingly, since we are sustaining the field inside: the passive cavity with a constant injected field did the same. One might say that the only difference between the two systems is that the active cavity is an "integrator," in that we might think of it as adding up the results of all the elementary measurements [which accounts for the linear growth of $X_2(t)$]; each one, of course, with its corresponding noise. The result is, of course, neither more nor less precise (save, perhaps, for a numerical factor of the order of unity) than the "average" $X_{2\max}$ calculated by the passive device.

This continuous adding up of noise results in the diffusion process of Eq. (14), familiar indeed from discussions of the laser linewidth.[14] Note that, semiclassically, $X_2^2(t)\simeq n[\phi(t)-\phi_0]^2$, so that Eq. (14) does describe a phase-diffusion process.

What is the origin of this noise, when we have ignored the "added noise" introduced by the amplifier? Formally it comes from the noise operator F_2 associated with the damping of the field. But all that these operators did, in the passive case, was to restore the normal vacuum fluctuations. Thus the noise in (14) is, roughly speaking, amplified vacuum fluctuations. *Unsqueezed* vacuum—the phase-insensitive nature of the losses sees to that. Physically, one might think of the losses as letting unsqueezed vacuum "leak into" the cavity (just as they let the inside field "leak out"), with quotation marks to indicate that we are not in general thinking of transmission losses (which are essentially reversible, and can be counteracted in various ways) but of irreversible absorption (maybe also diffraction, etc.) losses.

The process (14) accounts for *one-half* of the phase diffusion in a laser, which is usually attributed entirely to spontaneous emission. It is somewhat odd to see the losses take half of the credit for it here, although this is the way it comes naturally from a Langevin approach (compare the discussion in Ref. 14, and the work of Lax in Ref. 15). In this context, it is well known that different orderings of the operators lead to different interpretations. In fact, with the choice we have made here of working consistently with *Hermitian* operators, it is not surprising to find one-half of the spontaneous emission to come from amplified vacuum fluctuations (the missing half would come from the amplifier's own added noise, which we have neglected); compare this with the results in Ref. 16.

One final comment may be made. It seems reasonable to assume, as we have done, that by neglecting the amplifier's added noise we are indeed looking at the most favorable scenario, from the point of view of keeping the signal-to-noise ratio as large as possible. We might, however, wonder about the possibility of the amplifier introducing some noise which might be *anticorrelated* with F_2 in the equation of motion for X_2. But, since F_2 is uncorrelated with any other noise in the problem (including, in the sense mentioned earlier, F_1), this could only happen through some kind of feedback of X_2 upon itself, that is, some nonlinearity in the amplifier's treatment of X_2, which we have already discarded as being negligible. The amplifier's added noise could therefore only make matters worse, as it does indeed in the case of ordinary laser media (by the factor of 2 mentioned above).

IV. CONCLUSIONS

All the foregoing, either as contained in the mathematics or in the simpler statement: "The active device sustains (against the cavity losses, i.e., by amplifying it) and adds up both the signal *and* the noise of the passive device," explains how "shot noise" and "spontaneous emission," apparently conspired to make active and passive systems equivalent. In reality, "passive" and "active" systems are only different ways to process a single elementary measurement—one whose maximum duration and associated noise is determined solely by the cavity losses.

The limit encountered here is "fundamental" only in as much as the losses are unavoidable. It would seem from what we have presented here that one always has to gain from increasing the measurement time t_m, even as to make $t_m\gg\gamma^{-1}$; if that were the case then all the systems would be "loss limited", as the ones discussed here. There are, however, cases where t_m cannot be increased beyond certain limits (in a gravity-wave detector, for instance, it should not be chosen larger than half the expected period of the wave; in ring laser gyros, there are other sources of error which degrade the performance for very large integration times). If the losses can be reduced to the point when γ^{-1} is greater than the allowed measurement time, the system is no longer loss limited. In this case ($t_m<\gamma^{-1}$), the active and passive devices are still equivalent [expand the exponentials in (6) and (7), and compare with (13) and (14)] but *now the initial amount of squeezing becomes relevant*, and can indeed increase the sensitivity substantially, as explained, e.g., in Ref. 12, for gravity-wave detectors.

Aside from this, of course, in a practical apparatus the passive and active devices will not in general be equivalent from an experimentalist's point of view, each one having other merits and problems of its own (in different contexts, these have been discussed in Refs. 1, 2, and 5, among many other places). It is in this context that all those "numerical factors of the order of unity" that we might afford to ignore in this paper will, of course, become relevant.

ACKNOWLEDGMENTS

My understanding of this subject has been sharpened through many conversations with L. M. Pedrotti, to whom I am very much indebted. I have also learned much from conversations with C. M. Caves. Thanks are also due to W. Schleich, for a critical reading of the manuscript, and to M. O. Scully, for the motivation and constant challenge.

[1] W. W. Chow, J. Gea-Banacloche, L. M. Pedrotti, V. E. Sanders, W. Schleich, and M. O. Scully, Rev. Mod. Phys. 57, 61 (1985).

[2] See, in particular, A. Brillet and P. Tourrenc, in *Proceedings of the NATO Advanced Study Institute on Gravitational Radiation, Les Houches, 1982,* edited by N. DeRuelle and T. Piran (North-Holland, Amsterdam, 1983).

[3] It is perhaps worth pointing out that we are not talking here about the so-called "standard quantum limit" (see, e.g., Ref. 12 below) which is independent of the cavity losses and therefore only obtainable when these are negligible. We have in mind a loss-limited operation, as will be explained later.

[4] For example, a laser gyro operating at this limit was reported by T. A. Dorschner, H. A. Haus, M. Holz, I. W. Smith, and H. Statz, IEEE J. Quantum Electron. QE-16, 1376 (1980), while essentially the same limit has been achieved in a passive system by J. L. Davies and S. Ezekiel, Opt. Lett. 6, 505 (1981), and predicted for yet a different kind of passive system by S. Ezekiel, J. A. Cole, J. Harrison, and G. Sanders, in *Laser Inertial Rotation Sensors,* edited by S. Ezekiel and G. E. Knausenberger (SPIE, Bellingham, WA, 1978), Vol. 157, p. 68.

[5] As observed in Ref. 2 above and most recently by A. Abramovici and Z. Vager, Phys. Rev. A 33, 3181 (1986).

[6] As has been recently shown experimentally by R. E. Slusher, L. W. Hollberg, B. Yurke, J. C. Mertz, and J. F. Valley, Phys. Rev. Lett. 55, 2409 (1985).

[7] H. P. Yuen and V. W. S. Chan, Opt. Lett. 8, 177 (1983); B. L. Schumaker, *ibid.* 9, 189 (1984).

[8] C. M. Caves, Phys. Rev. D 26, 1817 (1982).

[9] J. A. Goldstone and E. M. Garmire, IEEE J. Quantum Electron. QE-17, 366 (1981). (The nonlinearity considered by these authors is, of course, not needed here.)

[10] Early papers on the fluctuation-dissipation theorem are those by H. B. Callen and T. A. Welton, Phys. Rev. 83, 34 (1951); J. R. Senitzky, *ibid.* 119, 670 (1960); 124, 642 (1961). Closer in spirit to the present paper is the work of M. Lax, Phys. Rev. 145, 110 (1966).

[11] As used, for instance, by H. Haken, *Handbuch der Physik,* Vol. 25 of *Laser Theory* (Springer, Berlin 1970), Chap. 2; also in Ref. 14 below, and most recently by C. W. Gardiner and M. J. Collet, Phys. Rev. A 31, 3761 (1985).

[12] C. M. Caves, Phys. Rev. D 23, 1693 (1981).

[13] Since, by the fluctuation-dissipation theorem, the losses (at zero temperature as assumed here) only introduce the noise necessary to ensure that the attenuated signal still has its undiminished quantum fluctuations, and since it is these quantum fluctuations (together with the photodetector's partition noise) which give rise to the shot-noise limit (see Ref. 7), it is, in this author's opinion, misleading to treat the loss-related noise as a noise that would be present *in addition* to shot noise in a passive system, as was done recently by A. Abramovici, Opt. Commun. 57, 1 (1986).

[14] For instance, M. Sargent III, M. O. Scully, and W. E. Lamb, Jr., *Laser Physics* (Addison-Wesley, Reading, MA, 1974).

[15] M. Lax, in *Brandeis University Summer Institute Lectures, 1966,* edited by M. Chretien, E. P. Gross, and S. Deser (Gordon and Breach, New York, 1968).

[16] J. Dalibard, J. Dupont-Roc, and C. Cohen-Tannoudji, J. Phys. (Paris) 43, 1617 (1982).

Nonstationary shot noise and its effect on the sensitivity of interferometers

T. M. Niebauer, R. Schilling, K. Danzmann, A. Rüdiger, and W. Winkler
Max-Planck-Institut für Quantenoptik, D-8046 Garching bei München, Germany

(Received 30 August 1990)

We treat the shot noise of a light source modulated in power as a nonstationary random process. The spectrum of such modulated shot noise, although it is still white, is shown to contain correlations between different frequency components. In addition, the noise is not equally distributed in phase. These effects can deteriorate the shot-noise-limited sensitivity of modulated interferometers. Maximizing the signal-to-noise ratio (SNR) introduces constraints on both the modulation and demodulation waveforms. The sensitivities obtained with several commonly used modulation schemes are calculated, and new modulation strategies are proposed to realize good SNR. We apply the results to the case of laser interferometer gravitational wave detectors where it is essential to reach a shot-noise-limited sensitivity. By taking into account the additional noise contribution from the modulated shot noise, we reduce the 3-dB discrepancy between the measured sensitivity of the Garching prototype detector and the theoretical shot-noise limit to about 1.5 dB.

I. INTRODUCTION

The goal of gravitational-wave (GW) detection places extremely high demands on the sensitivity of interferometric measurements. The existing prototype detectors are able to measure fluctuations in the optical phase difference between two interfering beams with a sensitivity on the order of $10^{-8}\,\mathrm{rad/Hz^{1/2}}$ in linear spectral density.[1,2] Some of these measurements go down to about $10^{-9}\,\mathrm{rad/Hz^{1/2}}$ and have been within a few dB of the theoretical shot-noise-limited sensitivity of the optical setups.[3,4] Note that in the literature it is more common to specify the sensitivity of the prototypes to gravitational waves (strain in space) or mirror motions. However, in this paper we are more interested in the limit that shot noise places on the resolution of an optical fringe.

These highly sensitive arrangements employ internal phase modulation. As a consequence the output light power exhibits a time dependence containing the harmonics of the modulation frequency, and the associated shot noise is nonstationary. The standard shot-noise formula[5] assumes constant light power and is not suited without some modification if the detected light power is time dependent.[6] The object of this paper is to describe the effect of the modulation on the shot noise, derive its frequency spectrum, and apply the results to signal detection in modulated interferometers.

It would be tempting to assume that modulated shot noise can be described as a white-noise source with a variance proportional to the time-averaged light power. We will see, however, that the shot-noise characteristics derived by appropriate consideration of the nonstationary random process alter this conclusion. Although it will be shown that the noise spectrum is indeed white (frequency independent) with a variance given by the mean light power, this spectrum differs from stationary white noise in two important ways. First, the modulated shot noise contains correlations between different frequency components. Second, the noise is not equally distributed in phase. In fact, the noise for a modulated interferometer may be anomalously high in the signal quadrature. These subtle differences in the noise statistics significantly affect the optimal demodulation strategy.

The mathematics used in this paper can be generally applied to the problem of signal detection in any type of nonstationary noise. The noise power spectrum is derived directly from the time domain correlation function. We limit, however, the discussion to the special case of modulated laser light.

II. MODULATED INTERFEROMETERS

Interferometers with phase modulation of the interfering beams provide an example of oscillating output light power, and thus of time-dependent shot noise. Typical cases are two-beam interferometers, e.g., of the Michelson or Mach-Zehnder type, and Fabry-Pérot cavities used in the rf reflection locking technique.[7,8] For the latter it is not the interference inside the cavity that is interesting, but the interference between the phase modulated light reflected off the front mirror and the unmodulated light leaking out of the cavity.

Let us consider the simple case of a Michelson interferometer, as it is used in GW prototype detectors. A schematic diagram of the operational principle is shown in Fig. 1. The phase difference between the two arms is modulated with an electro-optical phase modulator (EOM_1) at a frequency much higher than any antici-

©1991 The American Physical Society

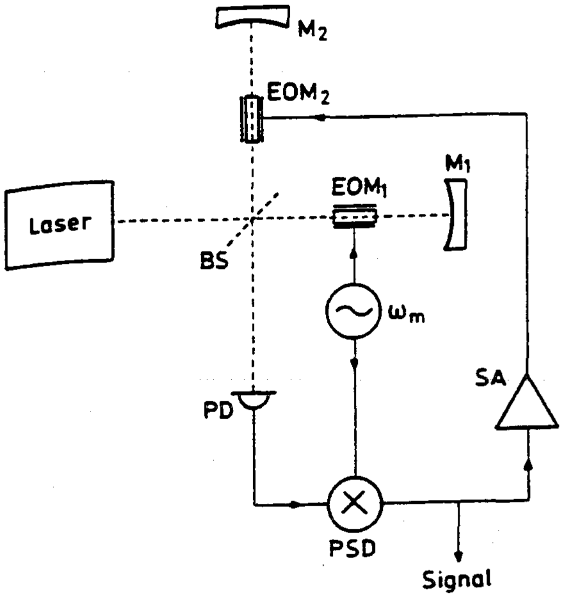


FIG. 1. Internal modulation in a simple Michelson interferometer, BS being the beam splitter, M the mirrors, EOM the electro-optic modulators, PD the photodiode, PSD the phase-sensitive demodulator, and SA the servo amplifier.

pated GW signal frequencies. The signal is recovered by phase-sensitive demodulation (PSD) and also is used as a feedback signal to lock the interferometer to a dark fringe.

The light power at the output of an ideal Michelson interferometer near the dark fringe is given by $P_0 \sin^2(\delta\phi/2)$, where P_0 is the input light power and $\delta\phi$ is the phase difference between the two arms. This phase difference is the sum of the internal phase modulation $m(t)$ and a signal $s(t)$. The function $s(t)$ represents the unmodulated phase difference between the interfering beams due to the signal to be measured, e.g., mirror motion or gravitational waves.

In addition to this ideal output, we also include a constant background light power $P_{\rm min}$ to describe the effect of imperfect fringe contrast. In the limit of small modulation and weak signals, $m(t) < 1$ and $s(t) \ll m(t)$, the light power at the interferometer output can be written as

$$\begin{aligned} P(t) &= \frac{P_0}{4}\left[m^2(t) + 2m(t)s(t)\right] + P_{\rm min} \\ &= \frac{P_0}{4}\left[m^2(t) + 2m(t)s(t) + b^2\,\overline{m^2(t)}\right]. \qquad (1) \end{aligned}$$

The first term describes the oscillating light power caused by the internally modulated path difference. The second, much smaller term is proportional to both the signal and the modulation. The signal is amplitude modulated at the modulation frequency ω_m, whereas the (much larger) first term is oscillating at twice this frequency. This makes it possible to separate the small signal contribution using phase-sensitive demodulation as will be described later. In the last term, b^2 is the ratio of the background light to the increment in average light power due to the modulation. This term introduces a constant noise background which for an ideal interferometer would be zero. However, in practical situations this term is never totally negligible, and the amplitude of the modulation is usually chosen relative to this background light power.

III. MODULATED SHOT NOISE

A. Time domain

Consider a measurement of the power $P(t)$ of a modulated light source over short time slices of length Δt during an observation time T. The time slices are understood to be short relative to the modulation period so that the average light power in each interval can be approximated by the value $P(t)$, constant during Δt. The observation time T is chosen to be an integer multiple of the modulation period and the length of the time slice. The shot noise during each time interval is found by considering the statistical fluctuation in the photon number. Assuming that the arrival times of all the photons are uncorrelated, the statistics for each time interval follow the Poisson distribution. The associated noise in each time slice is then described by a random variable $\mathcal{P}_{\rm n}(t)$ with zero mean and a variance proportional to the average photon count in each time interval. The correlation function of the noise can be written as

$$\langle \mathcal{P}_n(t)\,\mathcal{P}_n(t')\rangle = \frac{h\nu}{\Delta t}\,P(t)\,\delta_{t,t'}\,, \qquad (2)$$

where $h\nu$ is the energy of each photon and the Kronecker delta expresses the fact that photons in different time slices are uncorrelated. This treatment assumes that the light field is in a single-mode coherent state, in which case the second order coherence $g^{(2)}(t)$ is unity.[9]

The shot noise described by Eq. (2) is δ correlated in the time domain but is nonstationary because the variance is time dependent. We will see that this produces correlations in the frequency domain which are not present in the standard case of unmodulated shot noise.

B. Frequency domain

1. Amplitude dependence

The discrete Fourier transform of a single member $\mathcal{P}_n(t)$ of the ensemble of noise realizations over the finite observation time T is defined by

$$\tilde{\mathcal{P}}_n(\omega) = \frac{1}{T}\sum_{t=-T/2}^{T/2} \mathcal{P}_n(t)\,e^{-i\omega t}\,\Delta t\,. \qquad (3)$$

The variable t is an integer denoting the time slice (of length Δt) and ω is also understood to be an integer multiple of $2\pi/T$, running over positive and negative frequencies. In the limit of $\Delta t \to 0$, this equation is simply the complex Fourier series expansion for the noise realization for an observation time T.

The frequency components $\tilde{\mathcal{P}}_n(\omega)$ are also random variables with zero mean. The correlation function of these components can be derived using the definition Eq. (3), and the time domain correlation function Eq. (2) which collapses one of the sums:

$$\begin{aligned}\langle \tilde{\mathcal{P}}_n(\omega)\tilde{\mathcal{P}}_n^*(\omega')\rangle &= \frac{\Delta t^2}{T^2}\sum_t\sum_{t'}\langle \mathcal{P}_n(t)\mathcal{P}_n(t')\rangle e^{-i\omega t}e^{i\omega' t'} \\ &= \frac{h\nu}{T}\left(\frac{1}{T}\sum_t P(t)e^{-i(\omega-\omega')t}\Delta t\right). \qquad (4)\end{aligned}$$

These formulas can also be treated in the continuous sense by letting the time slices become infinitely thin. The term in the curly brackets is then the Fourier series expansion of the modulated light power $P(t)$. The noise correlations in the frequency and time domains can be written in the following compact forms:

$$\langle \tilde{\mathcal{P}}_n(\omega)\tilde{\mathcal{P}}_n^*(\omega')\rangle = \frac{h\nu}{T}\,\tilde{P}(\omega-\omega') \qquad (5)$$

and

$$\langle \mathcal{P}_n(t)\,\mathcal{P}_n(t')\rangle = h\nu\,P(t)\,\delta(t-t'). \qquad (6)$$

The frequency domain description is of course still discrete but the frequency resolution $\Delta f = 1/T$ can be made as high as desired by lengthening the observation time.

The expectation value of the squared noise spectrum is given by Eq. (5), setting $\omega = \omega'$,

$$\langle |\tilde{\mathcal{P}}_n(\omega)|^2\rangle = \frac{h\nu}{T}\,\tilde{P}(0) = \frac{h\nu}{T}\,\overline{P(t)}, \qquad (7)$$

where the bar denotes the average over the observation time. Thus, we see that the spectrum of the shot noise, Eq. (7), is frequency independent or white with a value proportional to the average light power. This result justifies the intuitive notion mentioned in the introduction that the average light power produces a shot noise with a white spectrum. However, Eq. (5) reveals that the modulation introduces correlations between different frequency components of the noise. The frequencies contained in the Fourier expansion of the time-dependent light power $P(t)$ give the separation between these correlated components. For example, a dc light source has white noise with uncorrelated frequency components, whereas a 100% modulated light source given by

$$P(t) = P_{\rm av}\,(1-\cos\omega_p t) \qquad (8)$$

introduces correlations between all noise components at frequencies separated by ω_p. The latter case occurs in an internally phase modulated interferometer with perfect fringe contrast, where ω_p equals twice the phase modulation frequency.

2. Phase dependence

Another interesting consequence of the modulation is that the shot noise may be unequally distributed in phase. To see this, we calculate the expectation value of the squared noise spectrum at a frequency ω with a phase angle θ. This can be found by evaluating the expectation value of the real part of $\tilde{\mathcal{P}}_n(\omega)$ in a reference frame rotated by θ:

$$\begin{aligned}\langle |\tilde{\mathcal{P}}_n(\omega,\theta)|^2\rangle &= \langle |\tfrac{1}{2}[\tilde{\mathcal{P}}_n(\omega)e^{-i\theta} + \tilde{\mathcal{P}}_n^*(\omega)e^{i\theta}]|^2\rangle \\ &= \frac{h\nu}{2T}\,\frac{1}{T}\int_{-T/2}^{T/2} dt\,P(t)\,[1+\cos(2\omega t+2\theta)]. \qquad (9)\end{aligned}$$

This equation shows that the noise contribution in two quadratures can be different. The deviation from a uniform distribution over phase angle θ can be seen by normalizing Eq. (9) to the average squared noise:

$$\frac{\langle |\tilde{\mathcal{P}}_n(\omega,\theta)|^2\rangle}{\langle |\tilde{\mathcal{P}}_n(\omega)|^2\rangle} = \frac{1}{2}\left(1+\frac{\overline{P(t)\cos(2\omega t+2\theta)}}{\overline{P(t)}}\right). \qquad (10)$$

Setting $\theta = 0$ in the above equation gives the noise in the cosine quadrature. For the case of a constant light power we recover the usual result that the shot noise is equally distributed in any two quadrature components separated by 90°. But for the example already mentioned above, i.e., a light source varying according to Eq. (8), we get

$$\begin{aligned}\frac{\langle |\tilde{\mathcal{P}}_n(\omega,\theta)|^2\rangle}{\langle |\tilde{\mathcal{P}}_n(\omega)|^2\rangle} &= \tfrac{1}{2} - \tfrac{1}{4}\,\overline{\cos[(\omega_p-2\omega)t-2\theta]} \\ &= \begin{cases} \tfrac{1}{2}-\tfrac{1}{4}\cos 2\theta & \text{for } \omega=\omega_p/2 \\ \tfrac{1}{2} & \text{otherwise.}\end{cases} \qquad (11)\end{aligned}$$

This shows that an unequal distribution of the noise occurs at the first subharmonic $\omega_p/2$ of the light power modulation frequency, where the squared noise in any one quadrature can vary between $\frac{1}{2}$ and $\frac{3}{2}$ times the usual mean value. This is particularly important for modulated interferometers, where the power oscillates at twice the frequency with which the signal is modulated. Unfortunately, the enhanced noise always appears in the signal quadrature. Thus, one will lose a factor $\sqrt{3/2}$ in signal-to-noise ratio (SNR) if one filters out the signal frequency only. However, one can almost fully recover the loss in SNR using a proper demodulation scheme.

C. General remarks

Summarizing the above: Modulated light power produces nonstationary shot noise. The spectrum of the noise is white but is no longer equally distributed in phase. In addition, different frequency components are correlated.

We note that the white power spectrum [Eq. (7)] is a direct consequence of the δ correlation assumed for the correlation function [Eq. (6)]. The time-dependent noise variance results in correlations between different components in the frequency domain, but is not evident in the power spectrum. One can see by analogy that correlations in the time domain would give rise to a frequency dependent or colored shot noise power spectrum. This may be important if one chooses a light source with a more complex second-order coherence function.

We have also mentioned that for the case of modulated interferometers, the shot noise is larger in the signal quadrature. We will later investigate demodulation schemes in which the SNR approaches that of the unmodulated case. This can only be achieved by utilizing the correlated noise components at the higher harmonics of the modulation frequency to reduce the overall noise contribution.

IV. DEMODULATION AND SIGNAL EXTRACTION

The effect of the modulation in a two-beam interferometer according to Eq. (1) is to produce an oscillating light power proportional to $m^2(t)$ and a signal which is amplitude modulated by $m(t)$. A modulation $m(t) = \sin\omega_m t$ shifts the signal to ω_m whereas the light power is modulated with $2\omega_m$. The signal is returned to dc by multiplying with a demodulation function $d(t)$ that is periodic with the same fundamental frequency ω_m. Since the modulation frequency is chosen much higher than the signal frequencies expected, the demodulated signal can be lowpass filtered with a cutoff frequency well below ω_m.

A. Demodulation of the signal

Demodulation of the signal produces a new function proportional to the product of the signal, modulation, and demodulation functions, $d(t)m(t)s(t)$. Since both modulation functions are assumed periodic, the product $q(t) = d(t)m(t)$ is also periodic with a Fourier series expansion containing a dc term and harmonics of the fundamental frequency ω_m. The signal $s(t)$, on the other hand, is assumed to have Fourier components $\tilde{s}(\omega)$ only at frequencies much lower than ω_m. The product $q(t)s(t)$ is most easily understood in the frequency domain as a convolution $\tilde{q}(\omega)\star\tilde{s}(\omega)$ in which the signal frequencies are located near dc and repeated at harmonics of ω_m. For frequencies less than $\omega_m/2$ the demodulated photodiode current can be expressed in the frequency domain as

$$\tilde{I}_s(\omega) = \frac{e\eta P_0}{2h\nu}\,\tilde{q}(0)\tilde{s}(\omega) \quad \text{for } \omega < \omega_m/2\,, \tag{12}$$

where $\tilde{q}(0) = \overline{m(t)d(t)}$, the quantum efficiency of the photodiode is η and the elementary charge is e.

B. Demodulation of the noise

In order to describe the demodulation of the noise we modify the time domain correlation function Eq. (6). Multiplication of the nonstationary shot noise by the function $d(t)$ does not alter the δ correlation in the time domain. The expectation value of the demodulated noise remains zero at any given instant, but the variance is multiplied by the square of the demodulation function. Thus, the correlation of the noise in terms of the demodulated photodiode current becomes

$$\langle \mathcal{I}_n(t)\mathcal{I}_n(t')\rangle = \frac{e^2\eta}{h\nu}\,d^2(t)\,P(t)\,\delta(t-t')\,. \tag{13}$$

This demodulated shot noise is also nonstationary and has the same form as the modulated shot noise given in Eq. (6). Thus, all the results derived for modulated shot noise are still valid with the simple replacement of $P(t)$ by $d^2(t)P(t)$ and a proper scale factor converting from light power to photodiode current. For example, the power spectrum of the demodulated noise is

$$\langle|\tilde{\mathcal{I}}_n(\omega)|^2\rangle = \frac{e^2\eta}{h\nu T}\,\overline{d^2(t)\,P(t)}\,. \tag{14}$$

An equation similar to Eq. (5) follows immediately which shows that correlations exist between different frequency components of the demodulated noise that are separated by frequencies contained in the Fourier series expansion of $d^2(t)P(t)$. More important, however, is that the lower frequencies of the demodulated noise which survive the lowpass filter stage are not correlated. This means that standard matched filter signal processing can proceed with the assumption of uncorrelated white noise.

The demodulated noise is particularly simple for the case of an ideal interferometer, $P_{\min} = 0$, with a small modulation index, and small signals $s(t) \ll m(t) \ll 1$. In this case, $P(t) = P_0 m^2(t)/4$ and the squared noise, given by Eq. (14), then is proportional to the time average of $q^2(t) = d^2(t)m^2(t)$. This can also be written as a sum of frequency components using Parseval's theorem:

$$\langle|\tilde{\mathcal{I}}_n(\omega)|^2\rangle = \frac{e^2\eta P_0}{4h\nu T}\,\overline{q^2(t)} \tag{15}$$

$$= \frac{e^2\eta P_0}{4h\nu T}\sum_{\omega}|\tilde{q}(\omega)|^2\,, \tag{16}$$

where the sum has to be taken over negative and positive frequencies.

V. SIGNAL-TO-NOISE RATIO IN MODULATED INTERFEROMETERS

For a sinusoidal signal with unknown phase we define a SNR as

$$N_{\rm SNR}(\omega) = \frac{|\tilde{I}_s(\omega)|}{\sqrt{\langle|\tilde{\mathcal{I}}_n(\omega)|^2\rangle}}\,, \tag{17}$$

where the numerator is the Fourier component of the signal and the denominator is the noise contribution that can be calculated quite generally using Eq. (14). Maximizing this ratio will constrain the optimal modulation and demodulation waveforms. We will investigate this formula in detail for the case of two-beam interferometers, and also briefly for Fabry-Pérot cavities used in the rf reflection locking technique.

For the following discussion it is convenient to split off from the SNR a factor F (≤ 1) that depends only on the modulation and demodulation waveforms:

$$N_{\mathrm{SNR}}(\omega) = F N_{\mathrm{SNR}_0}(\omega)\,, \tag{18}$$

where N_{SNR_0} is the maximum theoretical signal-to-noise ratio that could be obtained under ideal conditions. For two-beam interferometers we have

$$N_{\mathrm{SNR}_0}(\omega) = \left(\frac{\eta P_0 T}{h\nu}\right)^{1/2} |\tilde{s}(\omega)|\,. \tag{19}$$

When comparing this equation with shot-noise sensitivities quoted in the literature one should remember that the rms value of a narrow-band signal is $\sqrt{2}$ larger than the double-sided Fourier component $|\tilde{s}(\omega)|$ for frequencies $\omega \neq 0$.

A. Two-beam interferometers

For simplicity, we will still make the assumption of a quadratic response to phase differences, as was already done in Eq. (1).

1. Perfect fringe contrast

Let us first consider an interferometer with perfect fringe contrast ($P_{\min} = 0$) and small signals. Using Eqs. (12) and (15) we can write

$$F^2 = \frac{\overline{d(t)m(t)}^{\,2}}{\overline{d^2(t)m^2(t)}}\,. \tag{20}$$

It should be noticed that the fact that F is independent of the modulation amplitude originates from the approximation $s \ll m < 1$ made in Eq. (1).

The factor F is always less than or equal to unity. It becomes unity only when the product $m(t)d(t)$ is time independent. Thus, the optimum demodulation function for the case of modulated noise is $d(t) \propto 1/m(t)$. This condition is automatically met in the case of square wave modulation and demodulation. Sine modulation, on the other hand, would be best demodulated using an inverse sine (but such a waveform cannot be fully realized). This result is quite different from the usual notion that the best SNR would be obtained using identical waveforms for modulation and demodulation. Another interesting fact is that the SNR is symmetric with respect to the modulation and demodulation waveforms for interferometers with perfect fringe contrast.

2. Imperfect fringe contrast

In practical situations the assumption of a perfect contrast is not valid. Including the background light [$b^2 > 0$ in Eq. (1)] the factor F becomes

$$F^2 = \frac{\overline{d(t)m(t)}^{\,2}}{\overline{d^2(t)m^2(t)} + b^2\,\overline{d^2(t)}\;\overline{m^2(t)}}\,. \tag{21}$$

Now F is no longer independent of the modulation amplitude.

The second term in the denominator containing b^2 affects the SNR in two ways. First, it reduces the achievable value of F below unity for all choices of modulation waveforms. This is simply due to the fact that there is noise, but no signal contained in the background light. The second effect is more subtle. A poor fringe contrast introduces a nonmodulated noise component that is uncorrelated in the frequency domain. Thus, as the background light increases, we expect that the importance of the frequency correlations should decrease. In the limit of high background noise the optimal modulation and demodulation waveforms are identical instead of reciprocal as in the case of a perfect interferometer. The introduction of a minimum light power therefore changes the condition with respect to the optimal demodulation waveform.

B. Fabry-Pérot cavities

The case of a single Fabry-Pérot cavity used in the rf reflection locking technique is somewhat more difficult and we will present only the results for sine-wave modulation and demodulation here. For highly reflecting mirrors and assuming that only the carrier (of the phase modulated input light) enters the cavity, the time dependence of the light power hitting the photodiode can be written as

$$P(t) = P_0\left(1 - M(2A_c - A_c^2)J_0^2 - 4MA_cJ_0\sum_{k=1}^{\infty} J_{2k}\cos 2k\omega_m t\right), \tag{22}$$

where $M \leq 1$ is the mode-matching factor (for light power), A_c is the relative amplitude of the light leaking out of the cavity in resonance for perfect mode-matching and without modulation ($A_c = 1 \pm \sqrt{P_{\min}/P_0}$), and J are the Bessel functions of the first kind. The phase modulation of the input light is assumed to be $\phi(t) = \phi_m \sin\omega_m t$, where ϕ_m is the modulation index that has to be used as the argument of the Bessel functions.

The signal term in the above equation has been omitted. A deviation $\delta\nu(t)$ from the resonance leads to a phase shift $s(t) = 2\delta\nu(t)/\Delta\nu$ of the carrier leaking out of the cavity, where $\Delta\nu$ is the FWHM bandwidth of the cavity. For $s(t) \ll 1$, the signal term becomes

$$P_s(t) = 4s(t)P_0MA_cJ_0J_1\sin\omega_m t\,, \tag{23}$$

where the higher harmonics have been dropped since they do not contribute after sine-wave demodulation. We can calculate a factor F modifying the SNR where SNR_0 for the Fabry-Pérot cavity is two times larger than that found for the two-beam interferometer [see Eq. (19)]:

$$F^2 = \frac{2M^2A_c^2J_0^2J_1^2}{1 - M(2A_c - A_c^2)J_0^2 + 2MA_cJ_0J_2}\,. \tag{24}$$

This equation deviates from a treatment using the average light power in the standard shot-noise formula only by the addition of the third term in the denominator proportional to J_2.[10] Investigation of Eq. (24) shows that for an *undercoupled* Fabry-Pérot with optimal modulation index ($\phi_m \approx 1$) the correction due to this term is only a few percent if A_c stays below 0.5. In all other cases, the full equation must be used. For example, with $M = 1$, $A_c = 1$, and a small modulation index, the output light power has a form given by Eq. (8) and $F^2 = \frac{2}{3}$ as expected.

It is obvious that, also for the Fabry-Pérot, square-wave modulation avoids the time dependence of the light on the photodiode. Thus the corresponding optimal demodulation waveform would also be a square wave according to the matched filter theory.

VI. EVALUATION OF MODULATION AND DEMODULATION SCHEMES

In this section we will discuss different modulation and demodulation schemes that improve the SNR by appropriate utilization of the correlated noise in the harmonics. We will limit the discussion to two-beam interferometers and quadratic approximation of the phase response [see Eq. (1)], with emphasis on the typical case of sine-wave modulation.

A. Perfect fringe contrast

Let us now return to the formulas for perfect interference, $b^2 = 0$, and quote results for several realizable modulation and demodulation schemes. For square-wave modulation, demodulation using square or sine waveforms yields a correction factor F to SNR_0 of 1.0 and $\sqrt{8}/\pi = 0.900$, respectively. Using sine modulation, the F for square and sine demodulation is $\sqrt{8}/\pi = 0.900$ and $\sqrt{2/3} = 0.816$, respectively. We notice again that the results are symmetric with respect to the modulation and demodulation waveforms.

The case of sine modulation deserves special attention for two reasons. First, this is easiest to achieve experimentally and in fact is the dominant modulation used in existing setups. Secondly, from a theoretical point of view, sine modulation reveals a surprising effect. Consider the results quoted above which state that it is better to demodulate using a square wave than a sine wave. This result agrees with the graphical picture that a square wave better approximates the ideal inverse sine demodulation function. On the other hand, this result is surprising since one would expect the higher harmonics contained in the square wave to demodulate extra noise, but certainly not to increase the signal contribution. In the case of white uncorrelated noise, the square-wave demodulation would clearly be inferior for sine-wave modulation. The improvement is only possible in modulated shot noise because the additional noise components demodulated by the odd harmonics in the square wave lead to an overall reduction in noise. This is due to the correlation of noise components separated by twice the phase modulation frequency.

In order to clarify this statement let us assume sine-wave modulation and consider the effect of adding the third harmonic with amplitude α to the demodulation function:

$$m(t) = \sin\omega_m t$$
$$d(t) = \sin\omega_m t + \alpha \sin 3\omega_m t\,. \tag{25}$$

The product $q(t) = d(t)m(t)$ in the frequency domain has a dc component, a 2nd and a 4th harmonic. The noise power, calculated according to Eq. (15), is proportional to $\frac{3}{2}+\alpha^2-\alpha$ and obtains a minimum value when the even harmonics of $q(t)$ are of the same size, i.e. when $\alpha = \frac{1}{2}$. Thus, the addition of a third harmonic improves the F from 0.816 to 0.894.

The process of adding harmonics to the demodulation function can be extended in order to further improve the SNR. The optimum demodulation function containing N odd harmonics is given by

$$d(t) = \sum_{n=0}^{N-1}(1 - n/N)\sin[(2n+1)\omega_m t]\,. \tag{26}$$

The product function $q(t)$ contains a dc term and N even harmonics of equivalent strength. The squared noise contribution is proportional to $1 + 1/(2N)$ and approaches unity as the number of odd harmonics is increased. These relations show that better SNR can be obtained by selecting the optimum strength of the higher harmonics. This is desirable for designing waveforms which have both good SNR and relatively small bandwidth. However, we note that the energy, i.e., the time average of the squared function $d(t)$, increases proportional to $N/6 + 1/4 + 1/(12N)$ as more harmonics are added.

B. Imperfect fringe contrast

In the case of imperfect fringe contrast, the background light contributes uncorrelated noise reducing the advantage of adding higher harmonics (that do not contain any signal) to the demodulation function.

For the example of sine-wave modulation, there is a break-even point where sine-wave demodulation becomes superior to square wave. Equating the values of F^2 given by Eq. (21) for these two demodulation waveforms, we find that for relative background levels $b^2 > 1.14$ sine-wave demodulation is preferred. On the other hand, using the optimal modulation amplitude usually leads to a value for b^2 less than unity.

If only the third harmonic is added to the demodulation function [Eq. (25)] the optimal amplitude becomes $\alpha = 1/[2(1+b^2)]$ instead of $\frac{1}{2}$ as was found for perfect fringe contrast. Determining analytically the optimal amplitudes for a finite number of additional harmonics becomes increasingly difficult. It is more practical to first calculate the ideal demodulation function and truncate after the desired number of harmonics, accepting some deviation from the optimum SNR.

In order to arrive at an optimal SNR, the demodulation function in the time domain must be proportional to the signal modulation and inversely proportional to the time-dependent squared noise. For the case of a sine-wave modulation, this gives

$$d(t) = \frac{\sin\omega_m t}{\sin^2\omega_m t + \frac{1}{2}b^2} . \tag{27}$$

This equation shows explicitly that the optimal demodulation waveform is an inverse sine in the case of perfect fringe contrast ($b^2 = 0$), as mentioned earlier. We also recover the expected result that for poor fringe contrast ($b^2 \gg 1$) it is best to match the modulation function using sine-wave demodulation.

The Fourier series expansion of Eq. (27) consists of sinusoidal terms at the odd harmonics $(2k+1)\omega_m$ with amplitudes proportional to a^k, where $a = 1+b^2-\sqrt{2b^2+b^4}$. The rms value of this function, found by adding the squared frequency components, can be seen to be finite for all values $b^2 > 0$ in contrast to the case of perfect interference, where Eq. (26) gives a diverging series of harmonics for $N \rightarrow \infty$. The value of F^2 for sine-wave modulation and optimal demodulation can be written as

$$F^2 = 1 - \frac{|b|}{\sqrt{2+b^2}} . \tag{28}$$

VII. COMPARISON WITH THE STANDARD SHOT-NOISE FORMULA

A. Calculating a correction factor

We define the quantity R^2 as the ratio of the noise level calculated with the standard shot-noise formula (using the average light power hitting the photodiode) to the actual noise level (including the correlations in the modulated shot noise). This quantity is independent of the existence of a signal and gives a measure of the effect of modulation on the noise level. If one uses the power spectrum of modulated noise given by Eq. (7) and ignores correlations, the expected noise level, after demodulation, would just contain another normalization constant equal to the rms value of the demodulation function. The ratio R of the noise ignoring frequency correlations to the actual noise level is given by

$$R^2 = \frac{1+b^2}{\dfrac{\overline{d^2(t)m^2(t)}}{\overline{d^2(t)}\;\overline{m^2(t)}} + b^2} . \tag{29}$$

This ratio approaches unity as the stationary white noise contribution from the background light increases. One should note, however, that at the same time the SNR decreases.

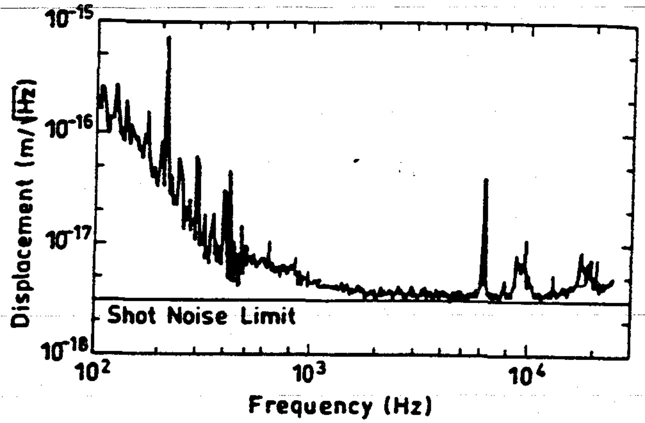


FIG. 2. Noise level of the Garching 30-m prototype detector. The straight line denotes the calculated shot-noise limit.

B. Application to a prototype GW detector

To our knowledge, the effect of correlations in modulated shot noise has not yet been included in published derivations of the shot-noise limits for gravitational wave detector prototypes. We find that the extra contribution from modulated noise can explain much of the discrepancy between the measured noise and the theoretical shot-noise limit reported by Shoemaker *et al.*[3] for the Garching prototype experiments. In these setups the modulation was sinusoidal. The demodulation waveform can also be taken to be sinusoidal since the higher harmonics were removed with a bandpass filter before demodulation.[11] For the experiment with the 30-m prototype a value of $b^2 \approx 0.22$ is estimated which gives a correction $R \approx 0.84$, or about 1.5 dB. The corrected shot-noise limit for the Garching prototype is graphed in Fig. 2 in comparison with the measured noise. The high power experiment described in Appendix C of Ref. 3 now shows excellent agreement between calculated and measured noise above 800 Hz.

VIII. CONCLUSION

We have shown that the modulation technique commonly used in making highly sensitive measurements with laser interferometry raises a special problem of detecting a signal in modulated, nonstationary noise. It is not sufficient to treat the resulting noise as a white distribution with a power spectrum equal to the average energy of the noise, because the modulation introduces correlations between various frequency components in the noise spectrum.

This can significantly alter the optimal demodulation scheme that one would follow if the noise were stationary. The usual procedure is to construct a demodulation function using only frequency components contained in the signal but to avoid including components that contain noise but no signal. These considerations lead to the conceptual picture that the optimal demodulation function should "match" the modulated signal. In nonstationary noise, however, we have seen that one can gain by including frequencies in the demodulation function that do not contain signal. These components contain correlated noise which tends to cancel that contained in the signal frequencies.

Maximizing the SNR for a modulated interferometer requires that the modulation and demodulation waveforms should be reciprocal in case of perfect fringe contrast. On the other hand, in the limit of very bad interference the optimal choice is to make both waveforms identical. Square-wave modulation and demodulation satisfies both these criteria simultaneously and from this point of view provides the ideal modulation technique. This is in some sense obvious since square-wave modulation produces a constant light output which gives stationary white noise. The practical disadvantage is that infinite bandwidth is needed for both the modulation and demodulation waveform generation. It is possible, however, to design modulation schemes in which odd harmonics are added to the modulation and/or demodulation waveforms in order to compromise between SNR and low bandwidth.

Furthermore, we want to emphasize that the effect described in this paper is caused by the time dependence of the output light power. In the case of two-beam interferometers, this results from internal phase modulation. Clearly, for schemes where the output power is not modulated, the usual shot-noise formula applies. For example, a Michelson interferometer with external modulation, as has been proposed for future GW detectors,[12] ideally will not have correlations in the shot noise. Also, for a sufficiently *undercoupled* Fabry-Pérot interferometer ($A_c < 0.5$), using the rf reflection locking technique with sinusoidal modulation and demodulation, the correction to the SNR is not higher than a few percent.

In this paper, modulated shot noise has been considered for interferometers. The results, however, are also relevant to optical experiments which produce modulated light *without* using an interferometer. For example, Carusotto *et al.*[13] have made accurate measurements of the magnetic birefringence (Cotton-Mouton effect) of gases by modulating the polarization state of the light and detecting the light after it passes an analyzer. The system is operated near extinction so that the light power oscillates with $2\omega_m$ and the signal is recovered at ω_m using a phase-sensitive demodulation. This leads to the same dependence of the SNR on the modulation and demodulation waveforms as was derived above for interferometers.

Finally, we note that shot noise formulas for the case of modulated light sources can be derived placing the average light power in the standard shot-noise formula and correcting this result with the factor $1/R$ which accounts for the nonstationarity. Applying this correction to the calculated shot-noise limit for the Garching prototype experiments improves the agreement with the measured sensitivity considerably.

ACKNOWLEDGMENTS

We wish to thank B. Meers, W. Schleich, L. Schnupp, and H. Ward for their helpful comments during the preparation of this manuscript. We also thank P. Das-Gupta and B. S. Sathyaprakash for useful discussions on the statistics of nonstationary random noise. This work was supported by the Bundesministerium für Forschung und Technologie (BMFT).

[1] H. Ward *et al.*, in *Experimental Gravitational Physics*, Proceedings of the International Symposium, Guangzhou, 1987, edited by P. F. Michelson, Hu En-Ke, and G. Pizzella (World Scientific, Singapore, 1988).

[2] R. Vogt, in *General Relativity and Gravitation*, Proceedings of the Twelfth International Conference on General Relativity and Gravitation, Boulder, 1989, edited by N. Ashby, D. F. Bartlett, and W. Wyss (Cambridge University Press, New York, 1990).

[3] D. Shoemaker, R. Schilling, L. Schnupp, W. Winkler, K. Maischberger, and A. Rüdiger, Phys. Rev. D **38**, 423 (1988).

[4] K. Maischberger, A. Rüdiger, R. Schilling, L. Schnupp, W. Winkler, and G. Leuchs, in *Experimental Gravitational Physics*, Proceedings of the International Symposium, Guangzhou, 1987, edited by P. F. Michelson, Hu En-Ke, and G. Pizzella (World Scientific, Singapore, 1988).

[5] W. Schottky, Ann. Phys. (Leipzig) **57**, 541 (1918); **68**, 157 (1922).

[6] The problem of modulated shot noise was treated in the time domain by Lise Schnupp (unpublished).

[7] R. V. Pound, Rev. Sci. Instrum. **17**, 490 (1946).

[8] R. W. P. Drever, J. L. Hall, F. V. Kowalski, J. Hough, G. M. Ford, A. J. Munley, and H. Ward, Appl. Phys. B **31**, 97 (1983).

[9] R. Loudon, *The Quantum Theory of Light* (Clarendon, Oxford, 1973).

[10] We have recently become aware that this term was already included in an unpublished derivation of the sensitivity for a GW prototype using Fabry-Pérot cavities by S. Whitcomb. However, this work did not discuss the properties of modulated shot noise nor the influence of the modulation and demodulation waveforms on the optimal SNR.

[11] Correcting Eqs. (A9), (A10), and (A12) in Ref. 3 according to Eq. (29) requires the term $2e(I_{dc}+I_{det})$ to be replaced by $2e\left(\frac{3}{2}I_{dc} - \frac{1}{2}I_{min} + I_{det}\right)$. The optimum modulation depth [Eq. (A13) in Ref. 3] becomes smaller by a factor of $\sqrt{2/3}$.

[12] See, e.g., J. Hough, B. J. Meers, G. P. Newton, N. A. Robertson, H. Ward, G. Leuchs, T. M. Niebauer, A. Rüdiger, R. Schilling, L. Schnupp, H. Walther, W. Winkler, B. F. Schutz, J. Ehlers, P. Kafka, G. Schäfer, M. W. Hamilton, I. Schütz, H. Welling, J. R. J. Bennet, I. F. Corbett, B. W. H. Edwards, R. J. S. Greenhalgh, and V. Kose, Max-Planck-Institut für Quantenoptik Report No. MPQ147 (1989).

[13] S. Carusotto, E. Iacopini, E. Polacco, F. Scuri, G. Stefanini, and E. Zavattini, J. Opt. Soc. Am. B **1**, 635 (1984); F. Scuri, G. Stefanini, E. Zavattini, S. Carusotto, E. Iacopini, and E. Polacco, J. Chem. Phys. **85**, 1789 (1986).

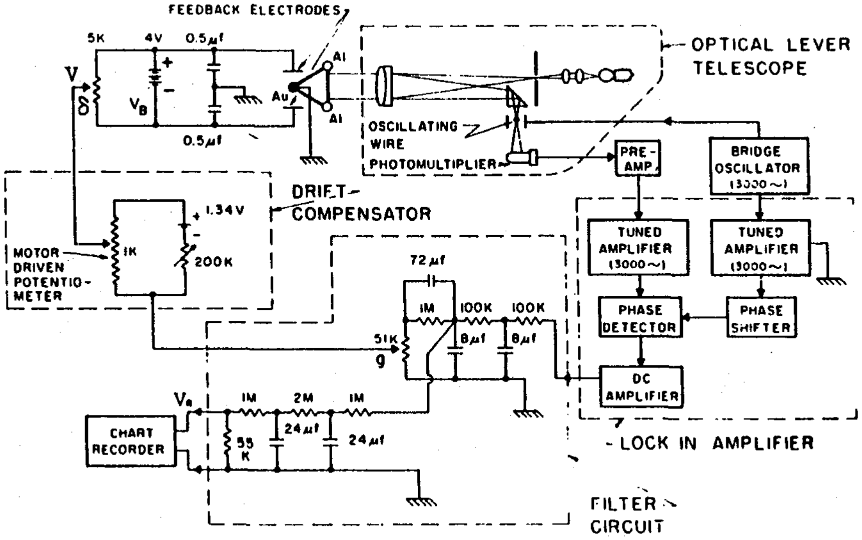


FIG. 6. Block diagram of the optical lever detection system

corded by the ionization gauge remained reasonably constant at about 1 × 10^{-8} mm. When the apparatus was removed from service 15 months later, the pressure, as carefully measured with the Bayard-Alpert gauge at reduced emission current, was still 10^{-8} mm Hg.

D. Optical Lever Detection System

At the heart of the experiment is the instrumentation for measuring very small rotations of the torsion balance. This equipment is shown schematically in Fig. 6. The source of light for the optical lever is a 6-volt flashlight bulb, operated at 5 volts from a regulated power supply to prolong the life of the bulb indefinitely. The light is focused through a 25μ slit, reflected from the aluminized flat on the quartz frame of the torsion balance, deflected off the telescope axis by a small prism, and the image of the slit focused on a 25μ-diam tungsten wire. By locating this wire in the field of a small magnet and connecting it in a balanced bridge oscillator circuit, it was made to oscillate at its mechanical resonance frequency of about 3000 cps and with an amplitude of 25 to 50μ. Mechanical and electronic details of the oscillating wire arrangement are shown in Fig. 7.

When the diffraction pattern of the 25μ slit produced by the 40 mm diameter telescope lens is centered exactly on the equilibrium position of the oscillating wire (Fig. 7(d)), the photomultiplier will detect only the even harmonics of the

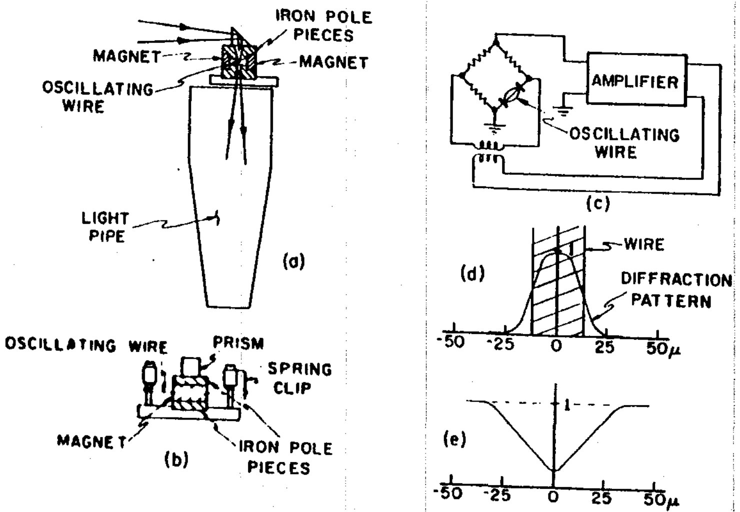


FIG. 7. Details of the oscillating wire light modulator. (a) Top view of the oscillating wire device, showing the magnet and pole piece assembly, prism, and light pipe. (b) Side view, showing the method of mounting the oscillating wire between the pole pieces. (c) Block diagram of the balanced bridge oscillator which drives the oscillating wire. (d) Sketch of the diffraction image of the slit focused and centered on the equilibrium position of the oscillating wire. As the wire oscillates about the position illustrated, the light received by the photomultiplier is modulated at the second harmonic of the wire frequency. Only when the diffraction image shifts off-center from the equilibrium position of the wire is the fundamental wire frequency detected by the photomultiplier. (e) Calculated fractional light intensity received by the photomultiplier as a function of displacement of the diffraction image of the slit from the center of the wire.

3000 cps fundamental frequency. As the torsion balance rotates slightly and shifts the diffraction pattern off center, the fundamental frequency will begin to appear in the photomultiplier output. The phase of the fundamental (0° or 180° relative to the oscillator signal driving the wire) indicates the direction of rotation of the pendulum, and its amplitude is proportional to the magnitude of the rotation for sufficiently small angular displacements.

The calculated fractional light intensity received by the photomultiplier is sketched in Fig. 7(e) as a function of displacement of the center of the diffraction pattern from the center of the wire. The full width at half maximum of this curve (the "line width" which must be split by the detection apparatus) is about 30μ or 3 × 10^{-5} rad.

Next, the photomultiplier output is increased by a preamplifier and an ampli-

REF. V

fier tuned to the fundamental frequency, and phase detected by mixing with the wire oscillator signal. The output of the phase detector is pulsating direct current which is filtered and further amplified. As indicated in Fig. 6, these processes are all performed by a lock-in amplifier. The output of the lock-in amplifier, then, is proportional in sign and magnitude to the angular displacement of the torsion balance from the position at which the slit image is centered on the wire.

This output goes to a filter circuit, whence part of it drives a chart recorder and part of it is fed back to the torsion balance. High frequency noise, including the 0.82 cps signal which appears as a result of the balance swinging in a plane including the telescope axis, is effectively removed from the chart recorder signal by the two 48 sec time constants in the recorder filter. The two $RC = 0.8$ sec sections in the feedback filter serve to remove high frequencies from the feedback electrodes. Following these two "integrating" filter sections is a "differentiating" circuit with a time constant of 3.6 sec. This provides the velocity feedback or damping which is necessary to prevent oscillation due to a Nyquist instability. Finally, the 51 kΩ potentiometer serves as a feedback gain and torque sensitivity control.

The time response of the "servo system" illustrated in Fig. 6 is determined by the feedback filter, and predominantly by the differentiating element in that circuit. Given an initial displacement, the fed-back torsion balance will oscillate with a complex frequency $\omega = (\omega_r + i\omega_i)$ which is a function of the open-loop feedback gain A. The cubic equation for $\omega(A)$, resulting from the feedback filter used, was solved in a straightforward but tedious manner to obtain the results shown in Fig. 8. Also indicated on this graph is the range of operating gain in which data was accumulated. As is desirable to minimize the effects of transient disturbances, the operating range was in or very near to the region of critical damping (vanishing real frequency). The effects of the magnitude and stability of A on the sensitivity of the torsion balance will be discussed in the next section.

Ignoring for the moment the drift compensator, a constant potential difference V_B of 4 volts is applied across the two feedback electrodes. However, the potential of each electrode relative to ground (the gold weight) is determined by the position of the movable contact on the 5 kΩ electrode bias potentiometer, and by the potential V applied to this movable contact from the feedback filter circuit. If the gold weight is approximately centered between the electrodes, it can be shown (see Appendix A) that the electrostatic torque on it is

$$L_e = cV_B^*[(1 - 2\delta)V_B + 2V], \tag{4}$$

where c is a constant involving the geometry of the arrangement, and δ is the electrode bias potentiometer setting ($0 \leqq \delta \leqq 1.0$, and $\delta = 0$ corresponds to the potentiometer slider connected to the negative battery terminal). Hence, the

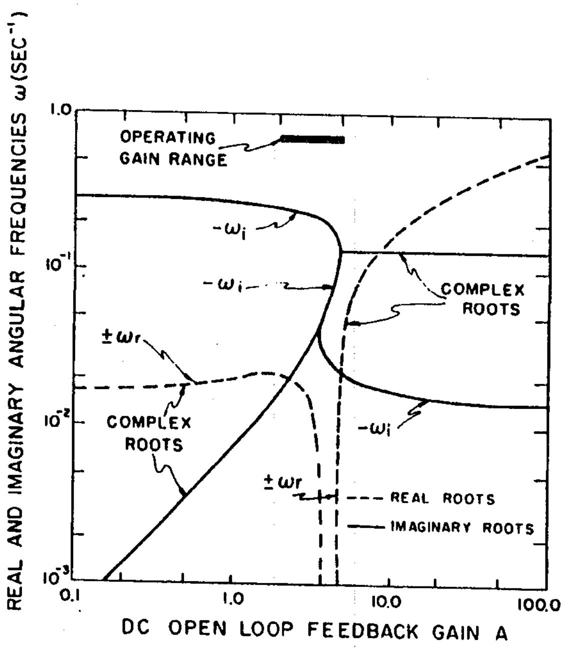


FIG. 8. Resonant angular frequencies of the torsion balance system as a function of the dc open loop feedback gain A. The dashed curves represent real parts of the resonance frequencies, while the solid curves represent imaginary parts which lead to damping of torsional oscillations. Also shown as solid curves are purely imaginary roots. The curves were calculated from the transfer characteristic of the feedback filter circuit shown in Fig. 6.

electrode bias potentiometer (actually a 10 kΩ four-decade precision voltage divider in parallel with a precision 10 kΩ resistor) provides a fine adjustment of the constant torque on the torsion balance, and therefore of its angular position.

As indicated in Fig. 6, V_B is the potential supplied to the feedback electrodes by the bias battery. The actual potential V_B^* between the surfaces of these electrodes, however, is the sum of V_B and the difference ΔV_B in the contact potentials between the surfaces:

$$V_B^* = V_B + \Delta V_B .$$

ΔV_B could amount to a volt or so, while V_B was 4 volts under operating conditions. In similar fashion, V should not be interpreted as the potential applied to the electrode bias potentiometer wiper, but the contact difference of potential between the gold weight and the copper electrodes should be added. However, as long as this contact potential difference does not vary, it is without effect on the experiment, and for simplicity it is ignored in the remainder of the discussion.

There is apparently more than one source of steady drift in the equilibrium angular position of the torsion balance. One well-established source of drift is the time rate of change of temperature. Since the apparatus is adequately shielded from 24-hr fluctuations in temperature, as will be discussed later, slow temperature drifts do not affect the experiment seriously. In addition to temperature-related drifts, there seems to be a small, steady drift which decreases with a time constant of months. Although the origin of this drift is not completely understood, it again does not affect the measurements of a 24-hr period, and may possibly be attributed to a slow relaxation of strains in the torsion fiber or other mechanical parts of the suspension. A third source of drift arises from the potential V_B of the batteries across the electrodes. These consist of three carefully selected RM 42-R mercury cells in series, enclosed in a temperature-regulated oven to be described later. As the batteries discharge through the 5 kΩ potentiometer, the total potential across the electrodes decreases at a rate of 0.5 mV/day or less, but in a linear fashion with no sign of 24-hr periodicities.

In order to keep the chart recorder on scale when operating at high sensitivities, it was necessary to remove as much of this drift as possible. Hence, a drift compensator was used to insert a voltage which increased linearly with time between the output of the feedback filter circuit and the electrode bias potentiometer. This device consisted of a 1 kΩ, 0.025% linearity, 25-turn Helipot driven by a synchronous motor at a rate of 6 rev/day, and in series with an RM-42R mercury cell and a large variable resistor. By adjusting the variable resistor to give 10 to 30 mV across the Helipot (generating a linear drift of 2.5 to 7.5 mV per day), depending on the weather (i.e., on the long-term behavior of the outdoor temperature), it was possible to keep the chart recorder on scale for reasonable periods of time. Nevertheless, it was still occasionally necessary to adjust the electrode bias potentiometer to discontinuously reposition the recorder trace on the chart. Since the torque sensitivity is independent of this potentiometer, the resulting discontinuity in the chart recorder trace could easily be removed by measuring and subtracting it from all subsequent data. The drift rate of the drift compensator was only infrequently changed, never during a run, so it could not introduce a spurious 24-hr period.

The potential of the drift compensator mercury cell was monitored with a potentiometer for one or two month periods a few times during the accumulation of data, and was found to fluctuate typically by less than 10 or 20 μV (close to the resolution of the potentiometer) from day to day, in a nonperiodic way. Since only 1 to 3% of the total cell voltage was applied to the bias potentiometer, such fluctuations were entirely negligible as far as the torsion balance output was concerned.

E. Torque Calibration and Sensitivity of the System

In order to obtain a quantitative measure of the external torque acting on the torsion balance, it was necessary to calibrate the apparatus. This was accomplished by using a micrometer head to rotate the telescope of the optical lever through a known angle relative to the balance. Such a rotation is fully equivalent to applying an external torque which would rotate the balance through the same angle in the opposite direction. The angle of rotation, together with the known torsion constant K of the fiber, gives the effective torque applied to the balance, which is then registered as a displacement of the trace on the chart recorder. (The torsion constant was obtained from the torsional oscillation period, knowing the moment of inertia of the torsion balance.)

Figure 9 illustrates the mechanical arrangement used in making this rotation. The clamping screw opposite the micrometer stem was loosened and the telescope gently pushed toward it with the micrometer, generally 1 or 2 × 10^{-3} in. at a time. The set of three clamping screws nearest the vacuum chamber served as the fulcrum for the rotation, and the weight of the telescope at the micrometer end was borne by the third clamping screw underneath the telescope (not visible

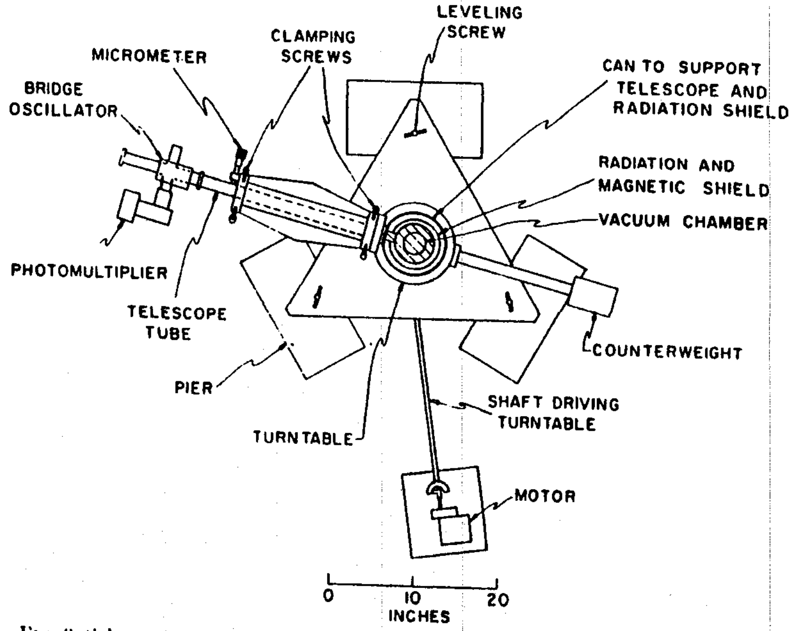


FIG. 9. Schematic top view of the torsion balance mounted in the instrument well

REF W

A Mode Selector to Suppress Fluctuations in Laser Beam Geometry

A. Rüdiger, R. Schilling, L. Schnupp, W. Winkler, H. Billing and K. Maischberger

Optica Acta, **28**, 641–658 (1981)

This reference turns out to be the same as Reference Q, which appears earlier in this volume. Therefore, we do not reproduce it here.

REF. X

Walter Koechner

Solid-State Laser Engineering

Second Completely Revised and Updated Edition

With 371 Figures

Springer-Verlag Berlin Heidelberg New York
London Paris Tokyo

Dr. Walter Koechner
Fibertek, Inc., 510-A Herndon Parkway,
Herndorn, VA 22070, USA

Editorial Board

Professor Koichi Shimoda, Ph. D.
Faculty of Science and Technology
Keio University, 3-14-1 Hiyoshi, Kohoku-ku
Yokohama 223, Japan

David L. MacAdam, Ph. D.
68 Hammond Street
Rochester, NY 14615, USA

Professor Anthony E. Siegman, Ph. D.
Electrical Engineering
E. L. Ginzton Laboratory, Stanford University
Stanford, CA 94305, USA

Arthur L. Schawlow, Ph. D.
Department of Physics, Stanford University
Stanford, CA 94305, USA

Theodor Tamir, Ph. D.
Polytechnic University
333 Jay Street, Brooklyn, NY 11201, USA

Managing Editor: Dr. Helmut K. V. Lotsch
Springer-Verlag, Tiergartenstrasse 17, D-6900 Heidelberg, Fed. Rep. of Germany

ISBN 3-540-18747-2 Springer-Verlag Berlin Heidelberg New York
ISBN 0-387-18747-2 Springer-Verlag New York Berlin Heidelberg

ISBN 3-540-90167-1 1. Auflage Springer-Verlag Berlin Heidelberg New York
ISBN 0-387-90167-1 1st edition Springer-Verlag New York Berlin Heidelberg

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in other ways, and storage in data banks. Duplication of this publication or parts thereof is only permitted under the provisions of the German Copyright Law of September 9, 1965, in its version of June 24, 1985, and a copyright fee must always be paid. Violations fall under the prosecution act of the German Copyright Law.

© Springer-Verlag Berlin Heidelberg 1976 and 1988
Printed in Germany

The use of registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Printing: Druckhaus Beltz, 6944 Hemsbach/Bergstr.
Binding: J. Schäffer GmbH & Co. KG, 6718 Grünstadt.
2154/3150-543210 – Printed on acid-free paper

1. Introduction

In this introductory chapter we shall outline the basic ideas underlying the operation of solid-state lasers. Assuming familiarity with laser physics [1.1] we shall sketch some of the principles governing the interaction of radiation with matter.

1.1 Optical Amplification

In this chapter we will outline the basic ideas underlying laser action. To understand the operation of a laser we have to know some of the principles governing the interaction of radiation with matter.

Atomic systems such as atoms, ions, and molecules can exist only in discrete energy states. A change from one energy state to another, called a transition, is associated with either the emission or the absorption of a photon. The wavelength of the absorbed or emitted radiation is given by Bohr's frequency relation

$$E_2 - E_1 = h\nu_{21} \quad , \tag{1.1}$$

where E_2 and E_1 are two discrete energy levels, ν_{21} is the frequency, and h is Planck's constant. An electromagnetic wave whose freqency ν_{21} corresponds to an energy gap of such an atomic system can interact with it. To the approximation required in this context, a solid-state material can be considered an ensemble of very many identical atomic systems. At thermal equilibrium, the lower energy states in the material are more heavily populated than the higher energy states. A wave interacting with the substance will raise the atoms or molecules from lower to higher energy levels and thereby experience absorption.

The operation of a laser requires that the energy equilibrium of a laser material be changed such that energy is stored in the atoms, ions, or molecules of this material. This is achieved by an external pump source which transfers electrons from a lower energy level to a higher one. The pump radiation thereby causes a "population inversion." An electromagnetic wave of appropriate frequency, incident on the "inverted" laser material, will be amplified because the incident photons cause the atoms in the higher level to drop to a lower level and thereby emit additional photons. As a result, energy is extracted from the atomic system and supplied to the radiation field. The release of the stored energy by interaction with an electromagnetic wave is based on stimulated or induced emission.

Stated very briefly, when a material is excited in such a way as to provide more atoms (or molecules) in a higher energy level than in some lower level, the material will be capable of amplifying radiation at the frequency corresponding to the energy level difference. The acronym "laser" derives its name from this process: "Light Amplification by Stimulated Emission of Radiation."

A quantum mechanical treatment of the interaction between radiation and matter demonstrates that the stimulated emission is, in fact, completely indistinguishable from the stimulating radiation field. This means that the stimulated radiation has the same directional properties, same polarization, same phase, and same spectral characteristics as the stimulating emission. These facts are responsible for the extremely high degree of coherence which characterizes the emission from lasers. The fundamental nature of the induced or stimulated emission process was already described by A. Einstein and M. Planck.

In solid-state lasers, the energy levels and the associated transition frequencies result from the different quantum energy levels or allowed quantum states of the electrons orbiting about the nuclei of atoms. In addition to the electronic transitions, multiatom molecules in gases exhibit energy levels that arise from the vibrational and rotational motions of the molecule as a whole.

1.2 Interaction of Radiation with Matter

Many of the properties of a laser may be readily discussed in terms of the absorption and emission processes which take place when an atomic system interacts with a radiation field. In the first decade of this century Planck described the spectral distribution of thermal radiation, and in the second decade Einstein, by combining Planck's law and Boltzmann statistics, formulated the concept of stimulated emission. Einstein's discovery of stimulated emission provided essentially all of the theory necessary to describe the physical principles of the laser.

1.2.1 Blackbody Radiation

When electromagnetic radiation in an isothermal enclosure, or cavity, is in thermal equilibrium at temperature T, the distribution of radiation density $\varrho(\nu)\,d\nu$, contained in a bandwidth $d\nu$, is given by Planck's law

$$\varrho(\nu)\,d\nu = \frac{8\pi\nu^2\,d\nu}{c^3}\frac{h\nu}{e^{h\nu/kT}-1} \quad , \tag{1.2}$$

where $\varrho(\nu)$ is the radiation density per unit frequency [Js/cm^3], k is Boltzmann's constant, and c is the velocity of light. The spectral distribution of thermal radiation vanishes at $\nu = 0$ and $\nu \rightarrow \infty$, and has a peak which depends on the temperature.

The factor

$$\frac{8\pi\nu^2}{c^3} = p_n \tag{1.3}$$

in (1.2) gives the density of radiation modes per unit volume and unit frequency interval. The factor p_n can also be interpreted as the number of degrees of freedom associated with a radiation field, per unit volume, per unit frequency interval. The expression for the mode density p_n [modes s/cm^3] plays an important role in connecting the spontaneous and the induced transition probabilities.

For a uniform, isotropic radiation field, the following relationship is valid

$$W = \frac{\varrho(\nu)c}{4} \quad , \tag{1.4}$$

where W is the blackbody radiation [W/cm^2] which will be emitted from an opening in the cavity of the blackbody. Many solids radiate like a blackbody. Therefore, the radiation emitted from the surface of a solid can be calculated from (1.4).

According to the Stefan-Boltzmann equation, the total black body radiation is

$$W = \sigma T^4 \quad , \tag{1.5}$$

where $\sigma = 5.68 \times 10^{-12}\,\mathrm{W/cm^2\,K^4}$. The emitted radiation W has a maximum which is obtained from Wien's displacement law

$$\frac{\lambda_{\max}}{\mu\mathrm{m}} = \frac{2893}{T/\mathrm{K}} \quad . \tag{1.6}$$

For example, a blackbody at a temperature of 5200 K has its radiation peak at 5564 Å, which is about the center of the visible spectrum.

A good introduction to the fundamentals of radiation and its interaction with matter can be found in [1.2].

1.2.2 Boltzmann Statistics

According to a basic principle of statistical mechanics, when a large collection of similar atoms is in thermal equilibrium at temperature T, the relative populations of any two energy levels E_1 and E_2, such as the ones shown in Fig. 1.1, must be related by the Boltzmann ratio

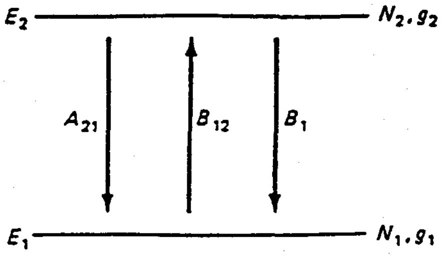


Fig. 1.1. Two energy levels with population N_1, N_2 and degeneracies g_1, g_2, respectively

$$\frac{N_2}{N_1} = \exp\left(\frac{-(E_2 - E_1)}{kT}\right) \quad , \tag{1.7}$$

where N_1 and N_2 are the number of atoms in the energy levels E_1 and E_2, respectively. For energy gaps large enough that $E_2 - E_1 = h\nu_{21} \gg kT$, the ratio is close to zero, and there will be very few atoms in the upper energy level at thermal equilibrium. The energy kT at room temperature ($T \approx 300\,\mathrm{K}$) corresponds to an energy gap $h\nu$ with $\nu \approx 6 \times 10^{12}\,\mathrm{Hz}$, which is equivalent in wavelength to $\lambda \approx 50\,\mu\mathrm{m}$. Therefore, for any energy gap whose transition frequency ν_{21} lies in the near-infrared or visible regions, the Boltzmann exponent will be $\gg 1$ at normal temperatures. The number of atoms in any upper level will then be very small compared to the lower levels. For example, in ruby the ground level E_1 and the upper laser level E_2 are separated by an energy gap corresponding to a wavelength of $\lambda \approx 0.69\,\mu\mathrm{m}$. Let us put numbers into (1.7). Since $h = 6.6 \times 10^{-34}\,\mathrm{Ws}^2$, then $E_2 - E_1 = h\nu = 2.86 \times 10^{-19}\,\mathrm{Ws}$. With $k = 1.38 \times 10^{-23}\,\mathrm{Ws\,K}$ and $T = 300\,\mathrm{K}$, it follows that $N_2/N_1 \approx 10^{-32}$. Therefore at thermal equilibrium virtually all the atoms will be in the ground level.

Equation (1.7) is valid for atomic systems having only non-degenerate levels. If there are g_i different states of the atom corresponding to the energy E_i, then g_i is defined as the degeneracy of the ith energy level.

We recall that atomic systems, such as atoms, ions, molecules, can exist only in certain *sationary states*, each of which corresponds to a definite value of energy and thus specifies an energy level. When two ore more states have the same energy, the respective level is called degenerate, and the number of states with the same energy is the multiplicity of the level. All states of the same energy level will be equally populated, therefore the number or atoms in levels 1 and 2 is $N_1 = g_1 N_1'$ and $N_2 = g_2 N_2'$, where N_1' and N_2' refer to the population of any of the states in levels 1 and 2, respectively. It follows then from (1.7) that the populations of the energy levels 1 and 2 are related by the formula

$$\frac{N_2}{N_1} = \frac{g_2}{g_1}\frac{N_2'}{N_1'} = \frac{g_2}{g_1}\exp\left(\frac{-(E_2 - E_1)}{kT}\right) \quad . \tag{1.8}$$

At absolute zero temperature, Boltzmann's statistics predicts that all atoms will be in the ground state. Thermal equilibrium at any temperature requires that a state with a lower energy be more densely populated than a state with a higher energy. Therefore N_2/N_1 is always less than unity for $E_2 > E_1$ and $T > 0$. This will turn out to mean that optical amplification is not possible in thermal equilibrium.

1.2.3 Einstein Coefficients

We can most conveniently introduce the concept of Einstein's A and B coefficients by loosely following Einstein's original derivation. To simplify the discussion, let us consider an idealized material with just two nondegenerate energy levels, 1 and 2, having populations of N_1 and N_2, respectively. The total number of atoms in these two levels is assumed to be constant

$$N_1 + N_2 = N_{\text{tot}} \quad . \tag{1.9}$$

Radiative transfer between the two energy levels which differ by $E_2 - E_1 = h\nu_{21}$ is allowed. The atom can transfer from state E_2 to the ground state E_1 by emitting energy; conversely, transition from state E_1 to E_2 is possible by absorbing energy. The energy removed or added to the atom appears as quanta of $h\nu_{21}$. We can identify three types of interaction between electromagnetic radiation and a simple two-level atomic system:

Absorption. If a quasimonochromatic electromagnetic wave of frequency ν_{21} passes through an atomic system with energy gap $h\nu_{21}$, then the population of the lower level will be depleted at a rate proportional both to the radiation density $\varrho(\nu)$ and to the population N_1 of that level

$$\frac{\partial N_1}{\partial t} = -B_{12}\varrho(\nu)N_1 \quad , \tag{1.10}$$

where B_{12} is a constant of proportionality with dimensions $\mathrm{cm}^3/\mathrm{s}^2\,\mathrm{J}$.

The product $B_{12}\varrho(\nu)$ can be interpreted as the probability per unit frequency that transitions are induced by the effect of the field.

Spontaneous Emission. After an atom has been raised to the upper level by absorption, the population of the upper level 2 decays spontaneously to the lower level at a rate proportional to the upper level population.

$$\frac{\partial N_2}{\partial t} = -A_{21}N_2 \quad , \tag{1.11}$$

where A_{21} is a constant of proportionality with the dimensions s^{-1}. The quantity A_{21}, being a characteristic of the pair of energy levels in question, is called the spontaneous transition probability because this coefficient gives the probability that an atom in level 2 will spontaneously change to a lower level 1 within a unit of time.

Spontaneous emission is a statistical function of space and time. With a large number of spontaneously emitting atoms there is no phase relationship between the individual emission processes; the quanta emitted are incoherent. Spontaneous emission is characterized by the lifetime of the electron in the excited state, after which it will spontaneously return to the lower state and radiate away the energy. this can occur without the presence of an electromagnetic field.

Equation (1.11) has a solution

$$N_2(t) = N_2(0)\exp\left(\frac{-t}{\tau_{21}}\right) \quad , \tag{1.12}$$

where τ_{21} is the lifetime for spontaneous radiation of level 2. This radiation lifetime is equal to the reciprocal of the Einstein's coefficient,

$$\tau_{21} = A_{21}^{-1} \quad . \tag{1.13}$$

In general, the reciprocal of the transition probability of a process is called its lifetime.

Stimulated Emission. Emission takes place not only spontaneously but also under stimulation by electromagnetic radiation of appropriate frequency. In this case, the atom gives up a quantum to the radiation field by "induced emission" according to

$$\frac{\partial N_2}{\partial t} = -B_{21}\varrho(\nu_{21})N_2 \quad , \tag{1.14}$$

where B_{21} again is a constant of proportionality.

Radiation emitted from an *atomic* system in the presence of external radiation consists of two parts. The part whose intensity is proportional to A_{21} is the spontaneous radiation; its phase is independent of that of the external radiation. The part whose intensity is proportional to $\varrho(\nu)B_{21}$ is the stimulated radiation; its phase is the same as that of the stimulating external radiation.

The probability of induced transition is proportional to the energy density of external radiation in contrast to spontaneous emission. In the case of induced transition there is a firm phase relationship between the stimulating field and the atom. The quantum which is emitted to the field by the induced emission is coherent with it.

But we shall see later, the useful parameter for laser action is the B_{21} coefficient; the A_{21} coefficient represents a loss term and introduces into the system photons that are not phase-related to the incident photon flux of electric field. Thus the spontaneous process represents a noise source in a laser.

If we combine absorption, spontaneous, and stimulated emission, as expressed by (1.10, 11, and 14), we can write for the change of the upper and lower level populations in our two-level model

$$\frac{\partial N_1}{\partial t} = -\frac{\partial N_2}{\partial t} = B_{21}\varrho(\nu)N_2 - B_{12}\varrho(\nu)N_1 + A_{21}N_2 \quad . \tag{1.15}$$

The relation

$$\frac{\partial N_1}{\partial t} = -\frac{\partial N_2}{\partial t} \tag{1.16}$$

follows from (1.9).

In thermal equilibrium, the number of transitions per unit time from E_1 to E_2 must be equal to the number of transitions from E_2 to E_1. Certainly, in thermal equilibrium

$$\frac{\partial N_1}{\partial t} = \frac{\partial N_2}{\partial t} = 0 \quad . \tag{1.17}$$

Therfore we can write

$$\underset{\substack{\text{Spontaneous}\\ \text{emission}}}{N_2A_{21}} \quad + \quad \underset{\substack{\text{Stimulated}\\ \text{emission}}}{N_2\dot{\varrho}(\nu)B_{21}} \;=\; \underset{\text{Absorption}}{N_1\varrho(\nu)B_{12}} \quad . \tag{1.18}$$

Using the Boltzmann equation (1.8) for the ratio N_2/N_1, we then write the above expression as

$$\varrho(\nu_{21}) = \frac{(A_{21}/B_{21})}{(g_1/g_2)(B_{12}/B_{21})\exp(h\nu_{21}/kT) - 1} \quad . \tag{1.19}$$

Comparing this expression with the black body radiation law (1.2), we see that

$$\frac{A_{21}}{B_{21}} = \frac{8\pi\nu^2 h\nu}{c^3} \quad \text{and} \quad B_{21} = \frac{g_1 B_{12}}{g_2} \quad . \tag{1.20}$$

The relations between the A's and B's are known as Einstein's relations. The factor $8\pi\nu^2/c^3$ in (1.20) is the mode density p_n given by (1.3).

In solids the speed of light is $c = c_0/n$, where n is the index of refraction and c_0 is the speed of light in vacuum.

For a simple system with no degeneracy, that is, one in which $g_1 = g_2$, we see that $B_{21} = B_{12}$. Thus, the Einstein coefficients for stimulated emission and absorption are equal. If the two levels have unequal degeneracy, the probablity for stimulated absorption is no longer the same as that for stimulated emission.

1.2.4 Phase Coherence of Stimulated Emission

The stimulated emission provides a phase-coherent amplification mechanism for an applied signal. The signal extracts from the atoms a response that is directly proportional to, and phase-coherent with, the electric field of the stimulating signal. Thus the amplification process is phase-preserving. The stimulated emission is, in fact, completely indistinguishable from the stimulating radiation field. This means that the stimulated emission has the same directional properties, same polarization, same phase, and same spectral characteristics as the stimulating emission. these facts are responsible for the extremely high degree of coherence which characterizes the emission from lasers. The proof of this fact is beyond the scope of this elementary introduction, and requires a quantum mechanical treatment of the interaction between radiation and matter. However, the concept of induced transition, or the interaction between a signal and an atomic system, can be demonstrated, qualitatively, with the aid of the classical electron-oscillator model.

Electromagnetic radiation interacts with matter through the electric charges in the substance. Consider an electron which is elastically bound to a nucleus. One can think of electrons and ions held together by spring-type bonds which are capable of vibrating around equilibrium positions. An applied electric field will cause a relative displacement between electron and nucleus from their equilibrium position. They will execute an oscillatory motion about their equilibrium position. Therefore, the model exhibits an oscillatory or resonant behavior and a response to an applied field. Since the nucleus is so much heavier than the electron, we assume that only the electron moves. The most important model for understanding the interaction of light and matter is that of the harmonic oscillator. We take as our model a single electron, assumed to be bound to its equilibrium position by a linear restoring force. We may visualize the electron as a point of mass suspended by springs. Classical electromagnetic theory asserts that any oscillating electric charge will act as a miniature antenna or dipole and will continuously radiate away electromagnetic energy to its surroundings.

A detailed description of the electric dipole transition and the classical electron-oscillator model can be found in [1.3].

1.3 Absorption and Optical Gain

In this section we will develop the quantitative relations that govern absorption and amplification processes in substances. This requires that we increase the realism of our mathematical model by introducing the concept of atomic lineshapes. Therefore, the important features and the physical processes which lead to different atomic lineshapes will be considered first.

1.3.1 Atomic Lineshapes

In deriving Einstein's coefficients we have assumed a monochromatic wave with frequency ν_{21} acting on a two-level system with an infinitely sharp energy gap $h\nu_{21}$. We will now consider the interaction between an atomic system having a finite transition linewidth $\Delta\nu$ and a signal with a bandwidth $d\nu$.

Before we can obtain an expression for the transition rate for this case, it is necessary to introduce the concept of the atomic lineshape function $g(\nu, \nu_0)$. The distribution $g(\nu, \nu_0)$, centered at ν_0, is the equilibrium shape of the linewidth-broadened transitions. Suppose that N_2 is the total number of ions in the upper energy level considered previously. The spectral distribution of ions per unit frequency is then

$$N(\nu) = g(\nu, \nu_0) N_2 \quad . \tag{1.21}$$

If we integrate both sides over all frequencies we have to obtain N_2 as a result:

$$\int_0^\infty N(\nu) d\nu = N_2 \int_0^\infty g(\nu, \nu_0) d\nu = N_2 \quad . \tag{1.22}$$

Therefore the lineshape function must be normalized to unity:

$$\int_0^\infty g(\nu, \nu_0) d\nu = 1 \quad . \tag{1.23}$$

If we know the function $g(\nu,\nu_0)$, we can calculate the number of atoms $N(\nu)d\nu$ in level 1 which are capable of absorbing in the frequency range ν to $\nu + d\nu$, or the number of atoms in level 2 which are capable of emitting in the same range. From (1.21) we have

$$N(\nu)\,d\nu = g(\nu,\nu_0)\,d\nu\,N_2 \quad . \tag{1.24}$$

From the foregoing it follows that $g(\nu,\nu_0)$ can be defined as the probability of emission or absorption per unit frequency. Therefore $g(\nu)\,d\nu$ is the probability that a given transition will result in an emission (or absorption) of a photon with energy between $h\nu$ and $h(\nu + d\nu)$. The probability that a transition will occur between $\nu = 0$ and $\nu = \infty$ has to be 1.

It is clear from the definition of $g(\nu,\nu_0)$ that we can, for example, rewrite (1.11) in the form

$$-\frac{\partial N_2}{\partial t} = A_{21}N_2 g(\nu,\nu_0)\,d\nu \quad , \tag{1.25}$$

where N_2 is the total number of atoms in level 2, and $\partial N_2/\partial t$ is the number of photons spontaneously emitted per second between ν and $\nu + d\nu$.

The linewidth and lineshape of an atomic transition depends on the cause of line broadening. Optical frequency transitions in gases can be broadened by lifetime, collision, or Doppler broadening, whereas transitions in solids can be broadened by lifetime, dipolar, thermal broadening, or by random inhomogeneities. All these linewidth-broadening mechanisms lead to two distinctly different atomic lineshapes, the homogeneously and the inhomogeneously broadened line [1.4].

The Homogeneously Broadened Line

The essential feature of a homogeneously broadened atomic transition is that every atom has the same atomic lineshape and frequency response, so that a signal applied to the transition has exactly the same effect on all atoms in the collection. This means that within the linewidth of the energy level each atom has the same probability function for a transition.

Differences between homogeneously and inhomogeneously broadened transitions show up in the saturation behavior of these transitions. This has a major effect on the laser operation. The important point about a homogeneous lineshape is that the transition will saturate uniformly under the influence of a sufficiently strong signal applied anywhere within the atomic linewidth.

Mechanisms which result in a homogeneously broadened line are lifetime broadening, collision broadening, dipolar broadening, and thermal broadening.

Lifetime Broadening. This type of broadening is caused by the decay mechanisms of the atomic system. Spontaneous emission or fluorescence has a radiative lifetime. Broadening of the atomic transition due to this process is related to the fluorescence lifetime τ by $\Delta\omega_a\tau = 1$, where ω_a is the bandwidth.

Actually, physical situations in which the lineshape and linewidth are determinded by the spontaneous emission process itself are vanishingly rare. Since the natural or intrinsic linewidth of an atomic line is extremely small, it is the linewidth that would be observed from atoms at rest without interaction with one another.

Collision Broadening. Collision of radiating particles (atoms or molecules) with one another and the consequent interruption of the radiative process in a random manner leads to broadening. As an atomic collision interrupts either the emission or the absorption of radiation, the long wave train which otherwise would be present becomes truncated. The atom restarts its motion after the collision with a completely random initial phase. After the collision the process is restarted without memory of the phase of the radiation prior to the collision. The result of frequent collisions is the presence of many truncated radiative or absorptive processes.

Since the spectrum of a wave train is inversely proportional to the length of the train, the linewidth of the radiation in the presence of collision is greater than that of an individual uninterrupted process.

Collision broadening is observed in gas lasers operated at higher pressures, hence the name pressure broadening. At higher pressures collisions between gas atoms limit their radiative lifetime. Collision broadening, therefore, is quite similar to lifetime broadening, in that the collisions interrupt the initial state of the atoms.

Dipolar Broadening. Dipolar broadening arises from interactions between the magnetic or electric dipolar fields of neighboring atoms. This interaction leads to results very similar to collision broadening, including a linewidth that increases with increasing density of atoms. Since dipolar broadening represents a kind of coupling between atoms, so that excitation applied to one atom is distributed or shared with other atoms, dipolar broadening is a homogeneous broadening mechanism.

Thermal Broadening. Thermal broadening is brought about by the effect of the thermal lattice vibrations on the atomic transition. The thermal vibrations of the lattice surrounding the active ions modulate the resonance frequency of each atom at a very high frequency. This frequency modulation represents a coupling mechanism between the atoms, therefore a homogeneous linewidth is obtained. Thermal broadening is the mechanism responsible for the linewidth of the ruby laser and Nd:YAG laser.

The lineshape of homogeneous broadening mechanisms lead to a Lorentzian lineshape for atomic response. For the normalized Lorentz distribution, the equation

$$g(\nu) = \left(\frac{\Delta\nu}{2\pi}\right)\left[(\nu-\nu_0)^2 + \left(\frac{\Delta\nu}{2}\right)^2\right]^{-1} \tag{1.26}$$

is valid. Here, ν_0 is the center frequency, and $\Delta\nu$ is the width between the half-power points of the curve. The factor $\Delta\nu/2\pi$ assures normalization of the area under the curve according to (1.23). The peak value for the Lorentz curve is

$$g(\nu_0) = \frac{2}{\pi\Delta\nu} \quad . \tag{1.27}$$

The Inhomogeneous Broadened Line

Mechanisms which cause inhomogeneous broadening tend to displace the center frequencies of individual atoms, thereby broadeing the overall response of a collection without broadening the response of individual atoms. Different atoms have slightly different resonance frequencies on the same transition, for example, owing to Doppler shifts. As a result, the overall response of the collection is broadened. An applied signal at a given frequency within the overall linewidth interacts strongly only with those atoms whose shifted resonance frequencies lie close to the signal frequency. The applied signal does not have the same effect on all the atoms in an inhomogeneously broadened collection.

Since in an inhomogeneously broadened line interaction occurs only with those atoms whose resonance frequencies lie close to the applied signal frequency, a strong signal will eventually deplete the upper laser level in a very narrow frequency interval. The signal will eventually "burn a hole" in the atomic absorption curve. Examples of inhomogeneous frequency-shifting mechanisms include Doppler broadening and broadening due to crystal inhomogeneities.

Doppler Broadening. The apparent resonance frequencies of atoms undergoing random motions in a gas are shifted randomly so that the overall frequency response of the collection of atoms is broadened. A particular atom moving with a velocity component ν relative to an observer in the z direction will radiate at a frequency measured by the observer as $\nu_0(1 + v/c)$. When these velocities are averaged, the resulting lineshape is Gaussian. Doppler broadening is one form of inhomogeneous broadening, since each atom emits a different frequency rather than one atom having a probability distribution for emitting any frequency within the linewidth. In the actual physical situation, the Doppler line is best visualized as a packet of homogeneous lines of width $\Delta\nu_n$, which superimpose to give the observed Doppler shape. The He-Ne laser has a Doppler-broadened linewidth. Most visible and near-infrared gas laser transitions are inhomogeneously broadened by Doppler effects.

Line Broadening Due to Crystal Inhomogeneities. Solid-state lasers may be inhomogeneously broadened by crystalline defects. This happens only at low temperatures where the lattice vibrations are small. Random variations of dislocations, lattice strains, etc., may cause small shifts in the exact energy level spacings and transition frequencies from ion to ion. Like Doppler broadening, these variations do not broaden the response on an individual atom, but they do cause the exact resonance frequencies of different atoms to be slightly different. Thus random crystal imperfection can be a source of inhomogeneous broadening in a solid-state laser crystal.

A good example of an imhomogeneously broadened line occurs in the fluorescence of neodymium-doped glass. As a result of the so-called glassy state, there are variations, from rare earth site to rare earth site, in the relative atomic positions occupied by the surrounding lattice ions. This gives rise to a random distribution of static crystalline fields acting on the rare-earth ions. Since the line shifts corresponding to such crystal-field variations are larger, generally speaking, than the width contributed by other factors associated with the transition, an inhomogeneous line results.

The inhomogeneous-broadened linewidth can be represented by a Gaussian frequency distribution. For the normalized distribution, the equation

$$g(\nu) = \frac{2}{\Delta\nu}\left(\frac{\ln 2}{\pi}\right)^{1/2} \exp\left[-\left(\frac{\nu-\nu_0}{\Delta\nu/2}\right)^2 \ln 2\right] \tag{1.28}$$

is valid. Where ν_0 is the frequency at the center of the line, and $\Delta\nu$ is the linewidth at which the amplitude falls to one-half. The peak value of the normalized Gaussian curve is

$$g(\nu_0) = \frac{2}{\Delta\nu}\left(\frac{\ln 2}{\pi}\right)^{1/2} \quad . \tag{1.29}$$

In Fig. 1.2 the normalized Gaussian and Lorentz lines are plotted for a common linewidth.

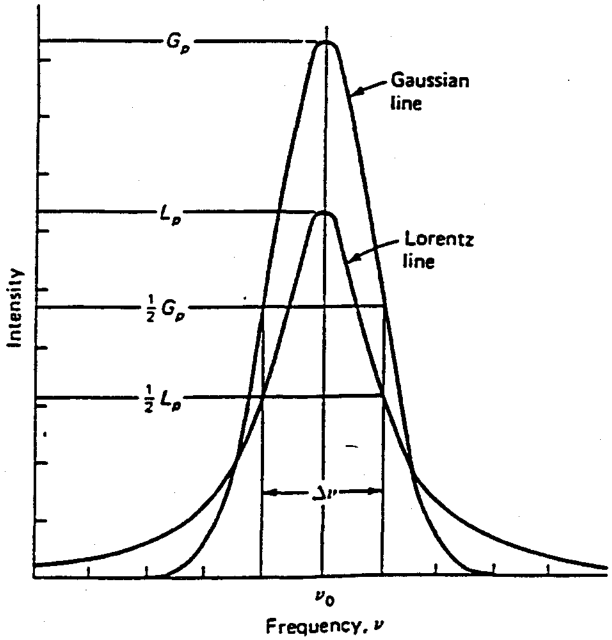


Fig. 1.2.
Gaussian and Lorentz lines of common linewidth (G_p and L_p are the peak intensities)

1.3.2 Absorption by Stimulated Transitions

We assume a quasicollimated beam of energy density $\varrho(\nu)$ incident on a thin absorbing sample of thickness dx; as before, we consider the case of an optical system that operates between only two energy levels as illustrated schematically in Fig. 1.1. The populations of the two levels are N_1 and N_2, respectively. Level 1 is the ground level and level 2 is the excited level. We consider absorption of radiation in the material and emission from the stimulated processes but neglect the spontaneous emission. From (1.15 and 1.20) we obtain

$$-\frac{\partial N_1}{\partial t} = \varrho(\nu) B_{21}\left(\frac{g_2}{g_1}N_1 - N_2\right). \tag{1.30}$$

As we recall, this relation was obtained by considering infinitely sharp energy levels separated by $h\nu_{21}$ and a monochromatic wave of frequency ν_{21}.

We will now consider the interaction between two linewidth-broadened energy levels with an energy separation centered at ν_0, and a half-width of $\Delta\nu$ characterized by $g(\nu, \nu_0)$ and a signal with center frequency ν_s and bandwidth $d\nu$. The situation is shown schematically in Fig. 1.3. The spectral width of the signal is narrow, as compared to the linewidth-broadened transition. If N_1 and N_2 are the total number of atoms in level 1 and level 2, then the number of atoms capable of interacting with a radiation of frequency ν_s and bandwidth $d\nu$ are

$$\left(\frac{g_2}{g_1}N_1 - N_2\right) g(\nu_s, \nu_0) d\nu \quad . \tag{1.31}$$

The net change of atoms in energy level 1 can be expressed in terms of energy density $\varrho(\nu)d\nu$ by multiplying both sides of (1.30) with photon energy $h\nu$ and dividing by the volume V. We will further express the populations N_1 and N_2 as population densities n_1 and n_2.

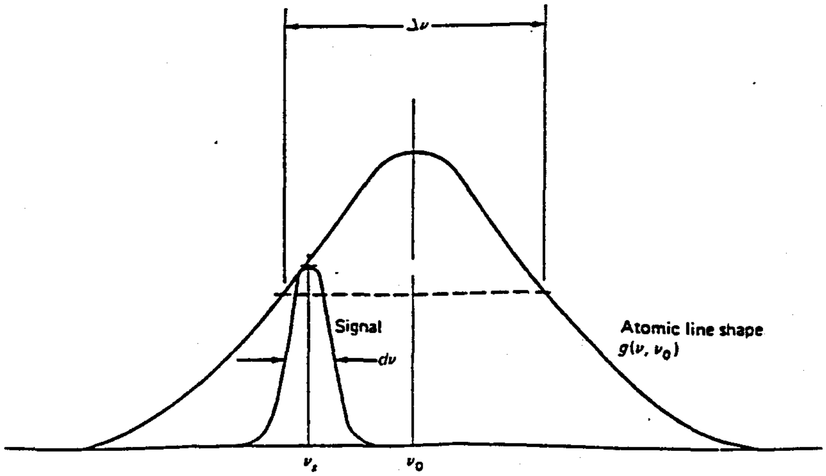


Fig. 1.3. Linewidth-broadened atomic transition line centered at ν_0 and narrow band signal centered at ν_s

Equation (1.30) now becomes

$$-\frac{\partial}{\partial t}[\varrho(\nu_s)d\nu] = \varrho(\nu_s)d\nu\, B_{21}h\nu g(\nu_s,\nu_0)\left(\frac{g_2}{g_1}n_1 - n_2\right) \quad . \tag{1.32}$$

This equation gives the net rate of absorbed energy in the frequency interval $d\nu$ centered around ν_s. In an actual laser system the wavelength of the emitted radiation, corresponding to the signal bandwidth $d\nu$ in our model, is very narrow as compared to the natural linewidth of the material. Ruby, for example, has a fluorescent linewidth of 5 Å, whereas the linewidth of the laser output is typically 0.1 to 0.01 Å. The operation of a laser, therefore, can be fairly accurately characterized as the interaction of linewidth-broadened energy levels with a monochromatic wave. The photon density of a monochromatic radiation of frequency ν_0 can then be represented by a delta function $\delta(\nu - \nu_0)$. After integrating (1.32) in the interval $d\nu$, we obtain, for a monochromatic signal of frequency ν_s and a linewidth-broadened transition,

$$-\frac{\partial \varrho(\nu_s)}{\partial t} = \varrho(\nu_s)B_{21}h\nu_s g(\nu_s,\nu_0)\left(\frac{g_2}{g_1}n_1 - n_2\right) \quad . \tag{1.33}$$

The signal will travel through the material of thickness dx in the time $dt = dx/c = (n/c_0)dx$. Then, as the wave advances from x to $x + dx$, the decrease of energy in the beam is

$$-\frac{\partial \varrho(\nu_s)}{\partial x} = h\nu_s\varrho(\nu_s)g(\nu_s,\nu_0)B_{21}\left(\frac{g_2}{g_1}n_1 - n_2\right)\frac{1}{c} \quad . \tag{1.34}$$

Integration of (1.34) gives

$$\frac{\varrho(\nu_s)}{\varrho_0(\nu_s)} = \exp\left[-h\nu_s g(\nu_s,\nu_0)B_{21}\left(\frac{g_2}{g_1}n_1 - n_2\right)\frac{x}{c}\right] \quad . \tag{1.35}$$

If we introduce an absorption coefficient $\alpha(\nu_s)$,

$$\alpha(\nu_s) = \left(\frac{g_2}{g_1}n_1 - n_2\right)\sigma_{21}(\nu_s) \quad , \qquad \text{where} \tag{1.36}$$

$$\sigma_{21}(\nu_s) = \frac{h\nu_s g(\nu_s,\nu_0)B_{21}}{c} \quad . \tag{1.37}$$

Then we can write (1.35) as

$$\varrho(\nu_s) = \varrho_0(\nu_s)\exp[-\alpha(\nu_s)x] \quad . \tag{1.38}$$

Equation (1.38) is the well-known exponential absorption equation for thermal equilibrium condition $n_1g_2/g_1 > n_2$. The energy of the radiation decreases exponentially with the depth of penetration into the substance. The maximum possible absorption occurs when all atoms exist in the ground state n_1. For equal population of the energy states $n_1 = (g_1/g_2)n_2$, the absorption is eliminated and the material is transparent. The parameter σ_{21} is the cross section for the radiative transition 2 → 1. The cross section for stimulated emission σ_{21} is related to the absorption cross section σ_{12} by the ratio of the level degeneracies,

$$\sigma_{21}/\sigma_{12} = g_1/g_2 \quad . \tag{1.39}$$

The cross section is a very useful parameter to which we will refer in the following chapters. If we replace B_{21} by the Einstein relation (1.20), we obtain σ_{21} in a form which we will find most useful:

$$\sigma_{21}(\nu_s) = \frac{A_{21}\lambda_0^2}{8\pi n^2} g(\nu_s, \nu_0) \quad . \tag{1.40}$$

As we will see later, the gain for the radiation building up in a laser resonator will be highest at the center of the atomic transitions. Therefore, in lasers we are mostly dealing with stimulated transitions which occur at the center of the linewidth.

If we assume $\nu \approx \nu_s \approx \nu_0$, we obtain, for the spectral stimulated emission cross section at the center of the atomic transition for a Lorentzian lineshape,

$$\sigma_{21} = \frac{A_{21}\lambda_0^2}{4\pi^2 n^2 \Delta\nu} \quad , \tag{1.41}$$

and for a Gaussian lineshape,

$$\sigma_{21} = \frac{A_{21}\lambda_0^2}{4\pi n^2 \Delta\nu} \left(\frac{\ln 2}{\pi}\right)^{1/2} \quad . \tag{1.42}$$

Here we have introduced into (1.40) the peak values of the lineshape function, as given in (1.27 and 1.29) for the Lorentzian and Gaussian curves respectively. For example, in the case of the R_1 line of ruby, where $\lambda_0 = 6.94 \times 10^{-5}$ cm, $n = 1.76, A_{21} = 2.5 \times 10^2 \, \mathrm{s}^{-1}$, and $\Delta\nu = 11.2 \, \mathrm{cm}^{-1} = 3.4 \times 10^{11}$ Hz at 300 K, one finds, according to (1.41), $\sigma_{21} = 2.8 \times 10^{-20} \, \mathrm{cm}^2$. The experimental value σ_{21} at the center of the R_1 line equals $2.5 \times 10^{-20} \, \mathrm{cm}^2$.

1.3.3 Population Inversion

According to the Boltzmann distribution (1.7), in a collection of atoms at thermal equilibrium there are always fewer atoms in a higher-lying level E_2 than in a lower level E_1. Therefore the population difference $N_1 - N_2$ is always positive, which means that the absorption coefficient $\alpha(\nu_s)$ in (1.36) is positive and the incident radiation is absorbed (Fig. 1.4).

Suppose that it were possible to achieve a temporary situation such that there are more atoms in an upper energy level than in a lower energy level. The normally positive population difference on that transition then becomes negative, and the normal stimulated absorption as seen from an applied signal on that transition is correspondingly changed to stimulated emission, or amplification of the applied signal. That is, the applied signal gains energy as it interacts with the atoms and hence is amplified. The energy for this signal amplification is supplied by the atoms involved in the interaction process. This situation is characterized by a negative absorption coefficient $\alpha(\nu_s)$ according to (1.36). From (1.34) it follows that $\partial\varrho(\nu)/\partial x > 0$.

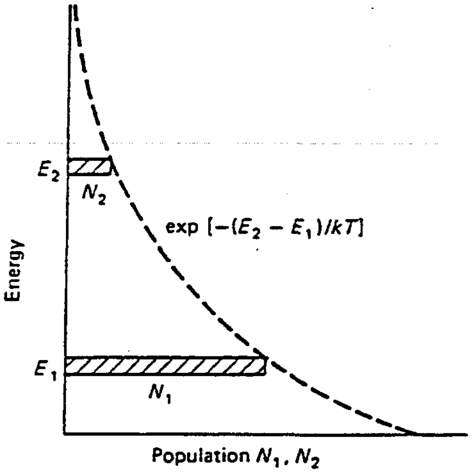


Fig. 1.4. Relative populations in two energy levels as given by the Boltzmann relation for thermal equilibrium

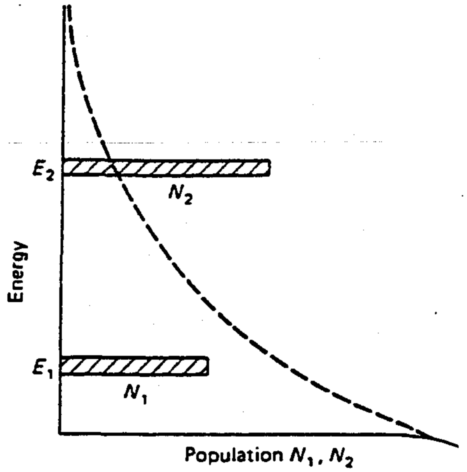


Fig. 1.5. Inverted population difference required for optical amplification

The essential condition for amplification is thus that somehow we must have, at a given instant, more atoms in an upper energy level than in a lower energy level; i.e., for amplification,

$$N_2 > N_1 \quad \text{if} \quad E_2 > E_1 \quad , \tag{1.43}$$

as illustrated in Fig. 1.5. The resulting negative sign of the population difference $(N_2 - g_2N_1/g_1)$ on that transition is called a population inversion. Population inversion is clearly an abnormal situation; it is never observed at thermal equilibrium. The point at which the population of both states is equal is called the "inversion threshold."

Stimulated absorption and emission processes always occur side by side independently of the population distribution among the levels. So long as the population of the higher energy level is smaller than that of the lower energy level, the number of absorption transitions is larger than that of the emission transitions, so that there is an overall attenuation of the radiation. When the numbers of atoms in both states are equal, the number of emissions becomes equal to the number of absorptions; the material is then transparent to the incident radiation. As soon as the population of the higher level becomes larger than that of the lower level, emission processes predominate and the radiation is enhanced collectively during passage through the material. In order to produce an inversion, we must have a source of energy to populate a specified energy level; we call this energy the pump energy.

In Sect. 1.4 we will discuss the type of energy level structure an atomic system must possess in order to make it possible to generate an inversion. Techniques by which the atoms of a solid-state laser can be raised or pumped into upper energy levels are discussed in Sect. 6.1. Depending on the atomic system involved, an inverted population condition may be obtainable only on a transient basis, yielding intermittent or pulsed laser action; or it may be possible to maintain the population inversion on a steady-state basis, yielding continuous-wave (cw) laser action.

The total amount of energy which is supplied by the atoms to the light wave is

$$E = \Delta N h\nu \quad , \tag{1.44}$$

where ΔN is the total number of atoms which are caused to drop from the upper to the lower energy level during the time the signal is applied. If laser action is to be maintained, the pumping process must continually replenish the supply of upper-state atoms. The size of the inverted population difference is reduced not only by the amplification process but also by spontaneous emission which always tends to return the energy level populations to their thermal equilibrium values.

1.4 Creation of a Population Inversion

We are concerned in this section with how the necessary population inversion for laser action is obtained in solid-state lasers. We can gain considerable understanding on how laser devices are pumped and how their population densities are inverted by studying some simplified but fairly realistic models.

The discussion up to this point has been based on a hypothetical 2 ↔ 1 transition and has not been concerned with how the levels 2 and 1 fit into the energy level scheme of the atom. This detached point of view must be abandoned when one tries to understand how laser action takes place in a solid-state medium. As already noted, the operation of the laser depends on a material with narrow energy levels between which electrons can make transitions. Usually these levels are due to impurity atoms in a host crystal. The pumping and laser processes in real laser systems typically involve a very large number of energy levels, with complex excitation processes and cascaded relaxation processes among all these levels. Operation of an actual laser material is properly described only by a many-level energy diagram. The main features can be understood, however, through the familiar three-level or four-level idealizations of Figs. 1.6 and 1.7. More detailed energy level diagrams of some of the most important solid-state laser materials are presented in Chap. 2.

1.4.1 The Three-Level System

Figure 1.6 shows a diagram which can be used to explain the operation of an optically pumped three-level laser, such as ruby. Initially, all atoms of the laser material are in the lowest level 1. Excitation is supplied to the solid by radiation of frequencies which produce absorption into the broad band 3. Thus, the pump light raises atoms from the ground state to the pump band, level 3. In general, the "pumping" band, level 3, is actually made up of a

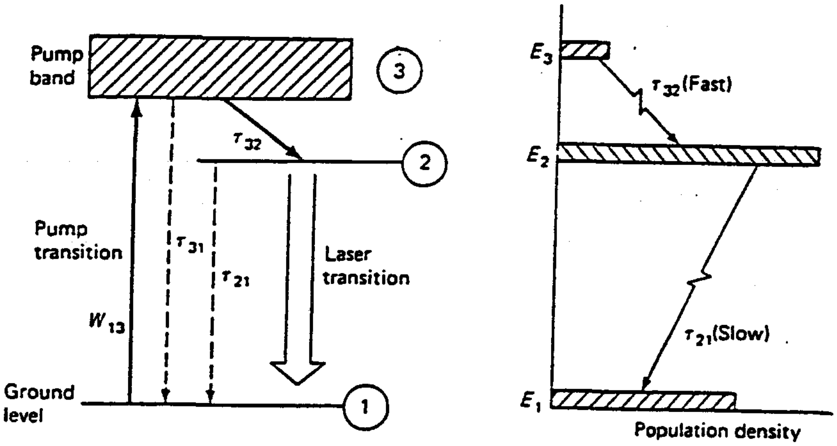


Fig. 1.6. Simplified energy level diagram of a three-level laser

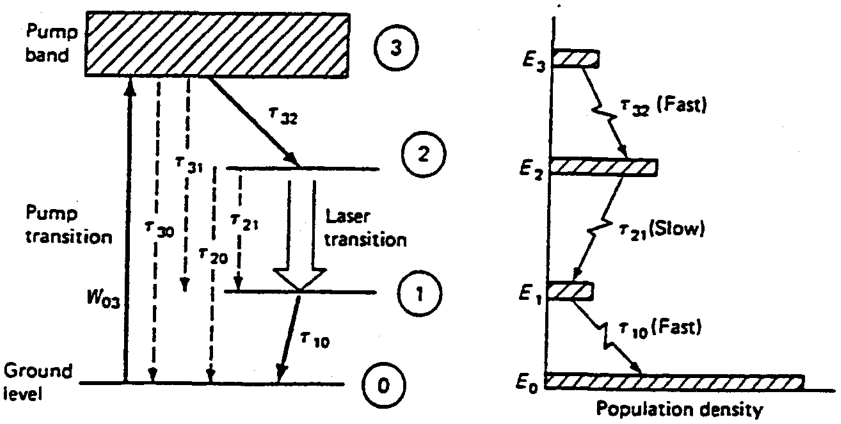


Fig. 1.7. Simplified energy level diagram of a four-level laser

number of bands, so that the optical pumping can be accomplished over a broad spectral range. In practice, xenon, krypton, mercury, and tungsten lamps are used for optically pumping solid-state lasers. Most of the excited atoms are transferred by fast radiationless transitions into the intermediate sharp level 2. In this process the energy lost by the electron is transferred to the lattice. Finally, the electron returns to the ground level by the emission of a photon. It is this last transition that is responsible for the laser action. If pumping intensity is below laser threshold, atoms in level 2 predominantly return to the ground state by spontaneous emission. Ordinary fluorescence acts as a drain on the population of level 2. After the pump radiation is extinguished, level 2 is emptied by fluorescence at a rate that varies from material to material. In ruby, at room temperature, the lifetime of level 2 is 3 ms. When the pump intensity is above laser threshold, the decay from the fluorescent level consists of stimulated as well as spontaneous radiation; the stimulated radiation produces the laser output beam. Since the terminal level of the laser transition is the highly populated ground state, a very high population must be reached in the E_2 level before the 2 → 1 transition is inverted.

It is necessary, in general, that the rate of radiationless transfer from the uppermost level to the level at which the laser action begins be fast compared with the other spontaneous transition rates in a three-level laser. Therefore, the lifetime of the E_2 state should be large in comparison with the relaxation time of the 3 → 2 transition, i.e.,

$$\tau_{21} \gg \tau_{32} \quad . \tag{1.45}$$

The number of atoms N_3 in level E_3 is then negligible compared with the number of atoms in the other two states, i.e., $N_3 \ll N_1, N_2$. Therefore,

$$N_1 + N_2 \approx N_{\text{tot}} \quad . \tag{1.46}$$

A vital aspect of the three-level system is that the atoms are in effect pumped directly from level 1 into the metastable level 2 with only a momentary pause as they pass through level 3. With these conditions, we can calculate as if only two levels were present. In order that an equal population is achieved between the E_2 and E_1 levels, one-half of all atoms must be excited to the E_2 level:

$$N_2 = N_1 = \frac{N_{\text{tot}}}{2} \quad . \tag{1.47}$$

In order to maintain a specified amplification, the population of the second level must be larger than that of the first level. In most cases which are of practical importance, however, the necessary inversion $(N_2 - N_1)$ is small compared with the total number of all atoms. The pump power necessary for maintaining this inversion is also small compared with the inversion power necessary for equal population of the level.

The disadvantage of a three-level system is that more than half of the atoms in the ground state must be raised to the metastable level E_2. There are thus many atoms present to contribute to the spontaneous emission. Moreover, each of the atoms which participate in the pump cycle transfer energy into the lattice from the $E_3 \rightarrow E_2$ transition. This transition is normally radiationless, the energy being carried into the lattice by phonons.

1.4.2 The Four-Level System

The four-level laser system, which is characteristic of the rare earth ions in glass or crystalline host materials, is illustrated in Fig. 1.7. Note that a characteristic of the three-level laser material is that the laser transition takes place between the excited laser level 2 and the final ground state 1, the lowest energy level of the system. This leads to low efficiency. The four-level system avoids this disadvantage. The pump transition extends again from the ground state (now level E_0) to a wide absorption band E_3. As in the case of the three-level system, the atoms so excited will proceed rapidly to the sharply defined level E_2. The laser transition, however, proceeds now to a fourth, terminal level E_1, which is situated above the ground state E_0. From here the atom undergoes a rapid nonradiative transition to the ground level. In a true four-level system, the terminal laser level E_1 will be empty. To qualifiy as a four-level system a material must possess a relaxation time between the terminal laser level and the ground level which is fast compared to the fluorescent lifetime, i.e., $\tau_{10} \ll \tau_{21}$. In addition the terminal laser level must be far above the ground state so that its thermal population is small. The equilibrium population of the terminal laser level 1 is determined by the relation

$$\frac{N_1}{N_0} = \exp\left(\frac{-\Delta E}{kT}\right) \quad , \tag{1.48}$$

where ΔE is the energy separation between level 1 and the ground state, and T is the operating temperature of the laser material. If $\Delta E \gg kT$, then $N_1/N_0 \ll 1$, and the intermediate level will always be relatively empty. In some laser materials the energy gap between the lower laser level and the ground state is relatively small and, therefore, they must be cooled to function as four-level lasers. In a four-level system an inversion of the 2 → 1 transition can occur even with vanishingly small pump power, and the high pump rate, necessary to maintain equilibrium population in the aforementioned three-level system, is no longer needed. In the most favorable case, the relaxation times of the 3 → 2 and 1 → 0 transitions in the four-level system are short compared with the spontaneous emission lifetime of the laser transition τ_{21}. Here we can also carry out the calculations as if only the E_1 and E_2 states were populated.

By far the majority of lasers materials operate, because of the more favorable population ratios, as four-level systems. The only laser of practical importance which operates as a three-level system is ruby. By a combination of favorable circumstances, it is possible in this unique case to overcome the disadvantages of the three-level scheme.

1.4.3 The Metastable Level

After this brief introduction to the energy level structure of solid-state lasers we can ask the question, "what energy level scheme must a solid possess to make it a useful laser?" As we have seen in the previous discussion, the existence of a metastable level is of paramount importance for laser action to occur. The relatively long lifetime of the metastable level provides a mechanism by which inverted population can be achieved. Most transitions of atoms show rapid nonradiative decay, because the coupling of the internal atomic oscillations to the surrounding lattice is strong. Nonradiative decay processes can occur readily, and characteristically have short lifetimes and broad linewidths. A few transitions of selected atoms in solids turn out to be decoupled from the lattice vibration. These transitions have a radiative decay which leads to relatively long lifetimes.

In typical laser systems with energy levels, such as illustrated by Fig. 1.6 and 7, the 3 → 2 transition frequencies, as well as the 1 → 0 transition frequencies, all fall within the frequency range of the vibration spectrum of the host crystal lattice. Therfore, all these transitions can relax extremely rapidly by direct nonradiative decay, i.e., by emitting a phonon to the lattice vibrations, with $\tau_{32}, \tau_{10} \approx 10^{-8}$ to 10^{-11} s. However, the larger 3 → 0, 3 → 1, 2 → 0, and 2 → 1 energy gaps in these atoms often correspons to transition frequencies that are higher than the highest possible vibration frequency of the crystal lattice. Such transitions cannot relax via simple single-phonon spontaneous emission, since the lattice simply cannot accept phonons at those high frequencies. These transitions must then relax either by radiative (photon) emission or by multiple-phonon processes. Since both these processes are relatively weak compared to direct single-phonon relaxation, the high-frequency transitions will have much slower relaxation rates ($\tau_{21} \approx 10^{-5}$ to 10^{-3} s in many cases). Therfore the various levels lumped into level 3 will all relax mostly into level 2 while level 2 itself is metastable and long-lived because there are no other levels located close below it into which it can decay directly.

The existence of metastable levels follows from quantum mechanical considerations that will not be discussed here. However, for completeness we will at least explain the term "forbidden transition". As we have seen in Sect. 1.2.4, the mechanism by which energy exchange takes place between an atom and the electromagnetic fields is the dipole radiation. As a consequence of quantum-mechanical considerations and the ensuing selection rules, transfer between certain states cannot occur due to forbidden transitions. The term "forbidden" means that a transition among the states concerned does not take place as a result of the interaction of the electric dipole moment of the atom with the radiation field. As a result of the selection rules, an atom may get into an excited state from which it will have difficulty returning to the ground state. A state from which all dipole transitions to lower energy states are forbidden is metastable; an atom entering such a state will generally remain in that state much longer than it would in an ordinary excited state from which escape is comparatively easy.

In the absence of a metastable level, the atoms which become excited by pump radiation and are transferred to a higher energy level will return either directly to the ground state by spontaneous radiation or by cascading down on intermediate levels, or they may release energy by phonon interaction with the lattice. In order for the population to increase at the metastable laser level, several other conditions have to be met. Let us consider the more general case of a four-level system illustrated in Fig. 1.7. (Note that a three-level system can be thought of as a special case of a four-level scheme where level 1 and level 0 coincide). Pumping takes place between two levels and laser action takes place between two other levels. Energy from the pump band is transferred to the upper laser level by fast radiative transitions. Energy is removed from the lower laser level again by fast radiationless transitions.

For electrons in the pump band at level 3 to transfer to level 2 rather than return directly to the ground state, it is required that $\tau_{30} \gg \tau_{32}$. For population to build up, relaxation out of the lower level 1 has to be fast, $\tau_{21} \gg \tau_{10}$. Thus, as a first conclusion, we may say that if the right relaxation time ratio exists between any two levels (such as 3 and 2) in an energy level system, a population inversion should be possible. If so, then obtaining a large enough inversion for successful laser operation becomes primarily a matter of the right pumping method. The optical pumping method is generally applicable only by the availability of systems which combine a narrow laser emission line with a broad absorption transition, so that a broad-band intense light source can be used as the pump source.

Having achieved population inversion in a material by correct combination of relaxation times and the existence of broad pump bands, the linewidth of the laser transition becomes very important. In the following chapter we will see that the optical gain for a given population inversion is inversely proportional to linewidth. Therfore, the metastable level should have a sufficiently narrow linewidth.

1.5 Laser Rate Equations

The dynamic behavior of a laser can be described with reasonable precision by a set of coupled rate equations [1.5]. In their simplest forms, a pair of simultaneous differential equations describe the population inversion and the radiation density within a spatially uniform laser medium. We will describe the system in terms of the energy-level diagrams shown in Figs. 1.6 and 1.7. As we have seen in the preceding discussions, two energy levels are of prime importance in laser action: the excited upper laser level E_2 and the lower laser level E_1. Thus for many analyses of laser action an approximation of the three- and four-level systems by a two-level representation is very useful.

The rate-equation approach used in this section involves a number of simplifying assumptions; in using a single set of rate equations we are ignoring longitudinal and radial variations of the radiation within the laser medium. In spite of these limitations, the simple rate-equation approach remains a useful tool and, properly used, provides a great deal of insight into the behavior of real solid-state laser deviced. We will derive from the rate equations the threshold condition for laser actions, and obtain a first-order approximation of the relaxation oscillations in a solid-state laser. Furthermore, in Chap. 4 we will use the rate equations to calculate the gain in a laser amplifier.

In general, the rate equations are useful in predicting the gross features of the laser output, such as average and peak power, Q-switched pulse-envelope shape, threshold condition, etc. On the other hand, many details of the nature of the laser emission are inacessible from the point of view of a simple rate equation. These include detailed descriptions of the spectral, temporal, and spatial distributions of the laser emission. Fortunately, these details can often be accounted for independently.

In applying the rate equations to the various aspects of laser operation, we will find it more convenient to express the probability for stimulated emission $\varrho(\nu)B_{21}$ by the photon density ϕ and the stimulated emission cross section σ.

With (1.37) we can express the Einstein coefficient for stimulated emission B_{21} in terms of the stimulated emission cross section $\sigma(\nu)$,

$$B_{21} = \frac{c}{h\nu g(\nu)}\sigma_{21}(\nu) \quad , \tag{1.49}$$

where $c = c_0/n$ is the speed of light in the medium. The energy density per unit frequency $\varrho(\nu)$ is expressed in terms of the lineshape factor $g(\nu)$, the energy $h\nu$, and the photon density ϕ [photons/cm^2] by

$$\varrho(\nu) = h\nu g(\nu)\phi \quad . \tag{1.50}$$

From (1.49 and 50) we obtain

$$B_{21}\varrho(\nu) = c\sigma_{21}(\nu)\phi \quad . \tag{1.51}$$

Three Level System

In order to approximate the three-level system with a two-level scheme, we assume that the transition from the pump band to the upper laser level is so fast that $N_3 \approx 0$. Therefore pumping does not affect the other processes at all except to allow a mechanism of populating the upper level and thereby obtaining population inversion ($N_2 > N_1$).

Looking at Fig. 1.6, this assumption requires that the relaxation time ratio τ_{32}/τ_{21} be very small. In solid-state lasers $\tau_{32}/\tau_{21} = 0$ is a good approximation. Spontaneous losses from the pump band to the ground state can be expressed by a pumping efficiency factor η_0. This parameter, defined as

$$\eta_0 = \left(1 + \frac{\tau_{32}}{\tau_{31}}\right)^{-1} \leq 1 \quad , \tag{1.52}$$

specifies what fraction of the total atoms excited to level 3 drop from there to level 2, thus becoming potentially useful for laser action. A small η_0 obviously requires a correspondingly larger pump power.

The changes in the electron population densities in a three-level system, based on the assumption that essentially all of the laser ions are in either level 1 or level 2, are

$$\frac{\partial n_1}{\partial t} = \left(n_2 - \frac{g_2}{g_1}n_1\right)c\phi\sigma + \frac{n_2}{\tau_{21}} - W_\mathrm{p}n_1 \tag{1.53}$$

and

$$\frac{\partial n_2}{\partial t} = -\frac{\partial n_1}{\partial t} \quad , \tag{1.54}$$

since

$$n_{\text{tot}} = n_1 + n_2 \quad , \tag{1.55}$$

where W_{p} is the pumping rate [s^{-1}].

The terms of the right-hand side of (1.53) express the net stimulated emission, the spontaneous emission, and the optical pumping.

The time variation of the population in both levels due to absorption, spontaneous, and stimulated emission is obtained from (1.15). Note that the populations N_1 and N_2 are now expressed in terms of population densities n_1 and n_2. To take into account the effect of pumping, we have added the term $W_{\text{p}}n_1$, which can be thought of as the rate of supply of atoms to the metastable level 2. More precisely, $W_{\text{p}}n_1$ is the number of atoms transferred from the ground level to the upper laser level per unit time per unit volume. The pump rate W_{p} is related to the pump parameter W_{13} in Fig. 1.6 by

$$W_{\text{p}} = \eta_0 W_{13} \quad . \tag{1.56}$$

The negative sign in front of $W_{\text{p}}n_1$ in (1.53) indicates that the pump mechanism removes atoms from the ground level 1 and increases the population of level 2.

If we now define the inversion population density by

$$n = n_2 - \frac{g_2 n_1}{g_1} \tag{1.57}$$

we can combine (1.53, 54, and 57) to obtain

$$\frac{\partial n}{\partial t} - \gamma n \phi \sigma c - \frac{n + n_{\text{tot}}(\gamma - 1)}{\tau_f} + W_{\text{p}}(n_{\text{tot}} - n), \tag{1.58}$$

where

$$\gamma = 1 + \frac{g_2}{g_1} \quad \text{and} \quad \tau_f = \tau_{21} \quad . \tag{1.59}$$

In obtaining (1.58) we have used the relations

$$n_1 = \frac{n_{\text{tot}} - n}{1 + g_2/g_1} \quad \text{and} \quad n_2 = \frac{n + (g_2/g_1)n_{\text{tot}}}{1 + g_2/g_1} \quad . \tag{1.60}$$

Another equation, usually regarded together with (1.58), describes the rate of change of the photon density within the laser resonator,

$$\frac{\partial \phi}{\partial t} = c\phi\sigma n - \frac{\phi}{\tau_c} + S, \tag{1.61}$$

where τ_c is the decay time for photons in the optical resonator and S is the rate at which spontaneous emission is added to the laser emission.

If we consider for the moment only the first term on the right, which is the increase of the photon density by stimulated emission, then (1.61) is identical to (1.33). However, for the time variation of the photon density in the laser resonator we must also take into account the decrease of radiation due to losses in the system and the increase of radiation due to a small amount of spontaneous emission which is added to the laser emission. Although very small, this term must be included because it provides the source of radiation which initiates laser emission.

An important consideration for initiation of laser oscillation is the total number p of resonant modes possible in the laser resonator volume V_R, since in general only a few of these modes are initiated into oscillations. This number is given by the familiar expression (1.3),

$$p = 8\pi\nu^2 \frac{\Delta\nu V_R}{c^3} \quad , \tag{1.62}$$

where ν is the laser optical frequency, and $\Delta\nu$ is the bandwidth of spontaneous emission. Let p_L be the number of modes of the laser output. Then S can be expressed as the rate at which spontaneous emission contributes to stimulated emission, namely,

$$S = \frac{p_L n_2}{p \tau_{21}} \quad . \tag{1.63}$$

The reader is referred to Chap. 3 for a more detailed description of the factor τ_c which appears in (1.61). For now we only need to know that τ_c represents all the losses in an optical resonator of a laser oscillator. Since τ_c has the dimension of time, the losses are expressed in terms of a relaxation time. The decay of the photon population in the cavity results from transmission and absorption at the end mirrors, "spillover" diffraction loss due to the finite apertures of the mirrors, scattering and absorptive losses in the laser material itself, etc. In the absence of the amplifying mechanism, (1.61) becomes

$$\frac{\partial\phi}{\partial t} = -\frac{\phi}{\tau_c} \quad , \tag{1.64}$$

the solution of which is $\phi(\tau) = \phi_0 \exp(-t/\tau_c)$.

The importance of (1.61) should be emphasized by noting that the right-hand side of this equation describes the net gain per transit of an electromagnetic wave passing through a laser material

Four-Level System

We will assume again that the transition from the pump band into the upper laser level occurs very rapidly. Therefore the population of the pump band is negligible, i.e., $n_3 \approx 0$. With this assumption the rate of change of the two laser levels in a four-level system is

$$\frac{dn_2}{dt} = W_p n_0 - \left(n_2 - \frac{g_2}{g_1} n_1\right)\sigma\phi c - \frac{n_2}{\tau_{21} + \tau_{20}} \quad , \tag{1.65}$$

$$\frac{dn_1}{dt} = \left(n_2 - \frac{g_2}{g_1} n_1\right)\sigma\phi c + \frac{n_2}{\tau_{21}} - \frac{n_1}{\tau_{10}} \quad , \tag{1.66}$$

$$n_{\text{tot}} = n_0 + n_1 + n_2 \quad . \tag{1.67}$$

From (1.65) follows that the upper laser level population in a four-level system increases due to pumping and decreases due to stimulated emission and spontaneous emissions into level 1 and level 0. The lower level population increases due to stimulated and spontaneous emission and decreases by a radiationless relaxation process into the ground level. This process is characterized by the time constant τ_{10}. In an ideal four-level system the terminal level empties infinitely fast to the ground level. If we let $\tau_{10} \approx 0$, then it follows from (1.66) that $n_1 = 0$. In this case the entire population is divided between the ground level 0 and the upper level of the laser transition. The system appears to be pumping from a large source that is independent of the lower laser level. With $\tau_{10} = 0$ and $n_1 = 0$, we obtain the following rate equation for the ideal four-level system

$$n = n_2 \quad \text{and} \tag{1.68}$$

$$n_{\text{tot}} = n_0 + n_2 \quad . \tag{1.69}$$

Therefore, instead of (1.58), we have

$$\frac{\partial n}{\partial t} = -n\sigma\phi c - \frac{n}{\tau_{\text{f}}} + W_{\text{p}}(n_{\text{tot}} - n). \tag{1.70}$$

The fluorescence decay time τ_{f} of the upper laser level is given by

$$\frac{1}{\tau_{\text{f}}} = \frac{1}{\tau_{21}} + \frac{1}{\tau_{20}} \quad , \tag{1.71}$$

where $\tau_{21} = A_{21}^{-1}$ is the effective radiative lifetime associated with the laser line. In the equation for the rate of change of the upper laser level we have again taken into account the fact that not all atoms pumped to level 3 will end up at the upper laser level. It is

$$W_{\text{p}} = \eta_0 W_{03} \quad , \tag{1.72}$$

where η_0 depends on the branching ratios which are the relative relaxation rates for the atoms along the various possible downward paths,

$$\eta_0 = \left(1 + \frac{\tau_{32}}{\tau_{31}} + \frac{\tau_{32}}{\tau_{30}}\right)^{-1} \leq 1. \tag{1.73}$$

The equation which describes the rate of change of the photon density within the laser resonator is the same as in the case of the three-level system.

Summary

The rate equation applicable to three-and four-level systems can be expressed by a single pair of equations, namely, (1.58 and 61), where $\gamma = 1 + g_2/g_1$ for a three-level system and $\gamma = 1$ for a four-level system. The parameters τ_f and W_p are defined by (1.56, 59, 72, and 73) for the different systems. The factor S in (1.61), which represents the initial noise level of ϕ due to spontaneous emission at the laser frequency, is small and needs to be considered only for initial starting of the laser action. It will be dropped from this point on.

A more detailed analysis of the laser rate equations can be found in [1.3,6].

REF

THE SIXTH

MARCEL GROSSMANN MEETING

On recent developments in theoretical and experimental general relativity, gravitation and relativistic field theories

Proceedings of the Meeting held at

Kyoto International Conference Hall
Kyoto, Japan
23 – 29 June 1991

PART A

Editors

Humitaka Sato and Takashi Nakamura
Department of Physics and Yukawa Institute
Kyoto University
Kyoto 606-01, Japan

World Scientific
Singapore • New Jersey • London • Hong Kong

Published by

World Scientific Publishing Co. Pte. Ltd.
P O Box 128, Farrer Road, Singapore 9128
USA office: Suite 1B, 1060 Main Street, River Edge, NJ 07661
UK office: 73 Lynton Mead, Totteridge, London N20 8DH

THE SIXTH MARCEL GROSSMANN MEETING

Copyright © 1992 by World Scientific Publishing Co. Pte. Ltd.

All rights reserved. This book, or parts thereof, may not be reproduced in any form or by any means, electronic or mechanical, including photocopying, recording or any information storage and retrieval system now known or to be invented, without written permission from the Publisher.

ISBN 981-02-0950-9
981-02-0951-7 (pbk)

Printed in Singapore by JBW Printers & Binders Pte. Ltd.

OPTICAL PROBLEMS IN INTERFEROMETRIC GRAVITATIONAL WAVE ANTENNAS

W. Winkler, K. Danzmann, A. Rüdiger and R. Schilling
Max-Planck-Institut für Quantenoptik, 8046 Garching, Germany

ABSTRACT

The specifications for the quality of the optical components in an interferometric gravitational wave detector are presented, based on a given power build-up using power recycling. The tolerable distortions of the surface shape are below $\lambda/100$ for spatial wavelengths considerably smaller than several centimeters; for longer wavelengths the demands are slightly less.

To keep the effects of local heating from affecting the sensitivity, the absorption at the optical components should not exceed the ppm level, and the substrate materials should have a low thermal expansion coefficient, a low temperature dependence of the index of refraction, and a *high thermal conductivity*.

1. Introduction

The design of gravitational wave antennas has to be oriented at the extremely small signal amplitudes expected by astrophysicists. The sensitivity of the planned beam detectors, large Michelson interferometers in essence, is planned to reach the strain level of 10^{-22}. These detectors are optimized by matching the effective length of the light path L to the wavelength of the gravitational waves and by minimizing the resolvable change δL in path difference.

An effective light path of 100 km or more is realized with multi-reflection schemes, either Fabry-Perot cavities or optical delay lines. For given mirror separation the effective light path is, in the case of cavities, determined by the reflectivity of the mirrors. In the case of delay lines the number of beams and thus the light path is selectable by choosing a proper relation between mirror separation and radius of curvature of the mirrors. For many practical reasons, such as thermally induced vibrations of the mirror substrates or finite quality of the mirror surfaces, the mirror separation has to be chosen quite large; the present proposals all assume 3 to 4 km. The radius of curvature of the mirrors will be of the same order of magnitude.

Besides the long light path a high effective light power is mandatory to achieve a high sensitivity, as can be seen from the limits set by the measuring process. If the resolution for changes in path difference is limited by the shot noise of the photocurrent at the output of the interferometer, the strain in space simulated by that noise is given by

$$\frac{\delta L}{L} = 10^{-22} \left(\frac{\lambda}{0.5\,\mu\mathrm{m}} \frac{50\,\mathrm{kW}}{\eta P} \frac{\Delta f}{1\,\mathrm{kHz}} \right)^{1/2} \frac{10^5\,\mathrm{m}}{L}. \tag{1}$$

Here L denotes the optical path-length, λ the wavelength of the light used, P the effective light power at the beamsplitter, η the quantum efficiency of the photodiode, and Δf the frequency bandwidth of observation. For still longer light paths, optimized for frequencies much lower than 1 kHz, the light power necessary to reach the 10^{-22} strain level is considerably lower than the 50 kW assumed in the formula. But in any case high light power is desirable in order to allow for the highest possible sensitivities.

2. Recycling of light

One will start out with high laser power. In addition, recycling schemes for the light will be implemented in order to enhance the effective light power. For that purpose the interferometer is operated at a dark fringe in the measurement output port; ideally all of the light leaving the interferometer (in the absence of a signal) retraces the input path – it runs back to the laser. This light, usually thrown away, can be recycled by insertion of a single mirror M_3 between laser and interferometer (see Fig. 1). As the interferometer is operated in a mode where it sends all light back to the input port, it acts like a mirror. Together with mirror M_3 a cavity is formed (the so called power recycling cavity), which in resonance may show a considerable power build-up compared with the unrecycled case. Let us call the ratio between the

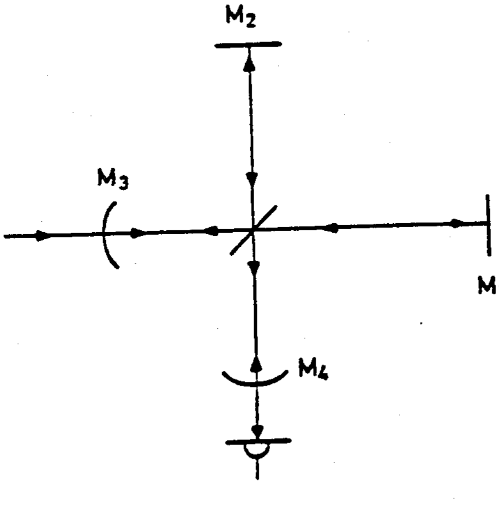


Figure 1: A Michelson interferometer with mirrors M_3 and M_4 inserted for recycling of the light power and the signal, respectively.

power P inside the cavity and the laser power P_{in} the power gain G. Maximizing this power gain requires minimizing the losses inside the power recycling cavity.

In addition, signal recycling will be implemented by inserting another partially transparent mirror M_4 in the signal output port in front of the photodiode. Light appearing in that output port because of a path difference is sent back into the interferometer and eventually comes out again through that output port. The signal recycling cavity thus formed can be tuned to a given frequency of observation by proper choice of the position of M_4, and its bandwidth is determined by the reflectivity of M_4. For simplicity let us restrict ourselves in the following to the case of power recycling.

The effectivity of the recycling schemes critically depends on the losses inside the recycling cavity. The power gain G can also be expressed in terms of the relative losses $\Delta P/P$:

$$G = \frac{P}{P_{\text{in}}} = \frac{P}{\Delta P}. \tag{2}$$

Losses occur because of many different reasons. First there is scattering – Rayleigh scattering, scattering due to inhomogeneities in the composition of the materials, and scattering because of micro-roughness (see below). In addition, there is absorption at the different kinds of impurities, at not perfectly oxidized metal in the coating, or in the vicinity of absorption bands of the materials used. Residual transmission of the mirrors also contributes to the losses. Finally there are losses due to a non-zero interference minimum, resulting from a mismatch between the interfering beams. The mismatch may occur either in amplitude, or, more severe, in the shape of the wavefronts. Wavefront deformations result from inhomogeneities of the index of refraction inside the components traversed by the light, and from imperfect surfaces hit by the light beam on its way through the interferometer.

In the following we will deduce specifications for the quality of the optical components required to reach a particular value for the envisaged power gain.

3. Surface quality

The ideal surface of the optical components for our purpose is either planar or spherical. Real surfaces deviate from that (see Fig. 2). The surface deformations can be characterized by their amplitude s and their spatial wavelength Λ. The effect of these deformations on the performance of the interferometer depends on the relation between that wavelength Λ and the beam diameter $2w$. Correspondingly, one has to consider three regions: surface deformations with $\Lambda < 2w$ (usually called "micro-roughness"), those with $\Lambda \approx 2w$ ("ripple"), and those with $\Lambda > 2w$ ("aberrations").

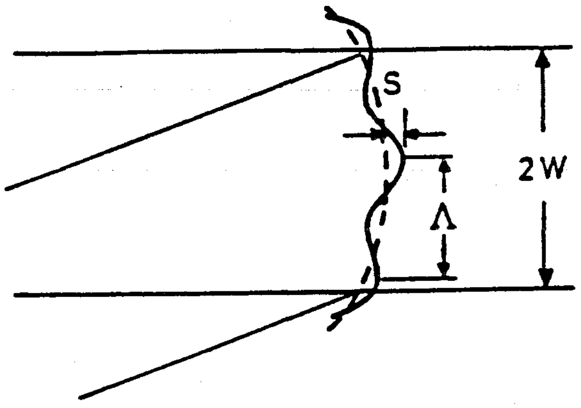


Figure 2: The deviations of a real surface (solid line) from its ideal shape (dashed line) are described by their spatial wavelengths Λ and their amplitude s. For the spatial wavelengths the quantity to compare with is the beam diameter $2w$.

3.1 Micro-roughness

Surface deformations with spatial wavelengths smaller than the beam diameter cause some light to leave the mode of light propagating through the interferometer. This process is called scattering and is one of the main reasons for losses. The relative power loss by scattering occurring for a reflection at a surface with a root mean square value s_{rms} of its micro-roughness amplitude is given by

$$\frac{\Delta P}{P} = \left(4\pi \frac{s_{\mathrm{rms}}}{\lambda}\right)^2 . \tag{3}$$

From that relation we can deduce a tolerable magnitude of s_{rms} if we ask for a particular value for the power gain in a setup with delay lines and N reflections:

$$N\frac{\Delta P}{P} < \frac{1}{G} . \tag{4}$$

For example, a power gain of 100 and 34 reflections in a delay line require the micro-roughness to stay below $\lambda/730$. For green light this is the quality known as *superpolish* with amplitudes smaller than a nanometer. But there is an important difference in comparison to the customary situation. Usually the micro-roughness is determined by averaging over spatial wavelengths up to a few mm – an upper limit for the beam size in almost all experiments. Here we have to average over all spatial wavelengths up to the beam diameter of about 5 cm for a 3 km interferometer and green light. It is well known that in all manufacturing procedures used today there is an increase in the spectral density of the amplitudes of surface deformations towards longer spatial wavelengths. Thus, scattering sets hard, but solvable requirements for the grinding and polishing process.

3.2 Ripple

Surface deformations with spatial wavelengths in the order of decimeters change the local radius of curvature of the mirrors. Correspondingly the curvature of the wavefront of the reflected beam is altered. Computer calculations show that the dark fringe of a perfect interference is deteriorated according to the following equation, when a surface deformation with an amplitude δs at one reflection anywhere inside the interferometer is introduced:

$$\frac{P_{\text{min}}}{P} = 10^{-3} \left(\frac{\delta s}{\lambda/100}\right)^2 . \qquad (5)$$

In a real interferometer there are many reflections at non-ideal components, for instance at the mirrors of an optical delay line. If we assume a Gaussian distribution for the amplitudes of the surface deformations with a standard deviation s_d, then on average the minimum of interference is deteriorated to

$$\frac{P_{\text{min}}}{P} = 5 \times 10^{-2} \frac{N}{34} \left(\frac{s_{\text{d}}}{\lambda/100}\right)^2 . \qquad (6)$$

If we shift N in this equation into the term in parentheses we see that the amplitude of the wavefront deformation grows proportional to the square root of the number N of reflections, leading to a degradation of the interference minimum proportional to N. This relation defines the tolerable amplitude of a ripple. To allow, for instance, a G of 100 in the case of 34 reflections, the rms amplitude of the ripple has to stay below $\lambda/230$. Again this is a hard demand, but seems to be within reach of present technology.

3.3 Aberrations

Finally we have to consider surface deformations with spatial wavelengths larger than the beam diameter. This is the range of the aberrations like astigmatism, spherical aberration or coma. The main effect of such deviations from perfect surfaces is a misorientation, and subsequently also a displacement of the beam. In Fabry-Perot cavities such displacements can be compensated by the alignment procedure, whereas in delay lines in general this is not possible. Therefore this section only deals with a setup with delay lines. The aberrations can be summarized in a value for the error in average slope of the surface. In a 3 km setup with 34 reflections and an envisaged power gain of 100, the error in slope has to stay below 10^{-7}. This requirement may be somewhat relaxed, if more than two mirrors in each interferometer arm are used. In that case position and orientation of the output beam can be adjusted by proper orientation of the mirrors, and perfect superposition is possible. One possibility are so called retro-mirrors: the beam leaves the delay line through a second hole in the near mirror, hits perpendicularly the retro-mirror and retraces its original path. This

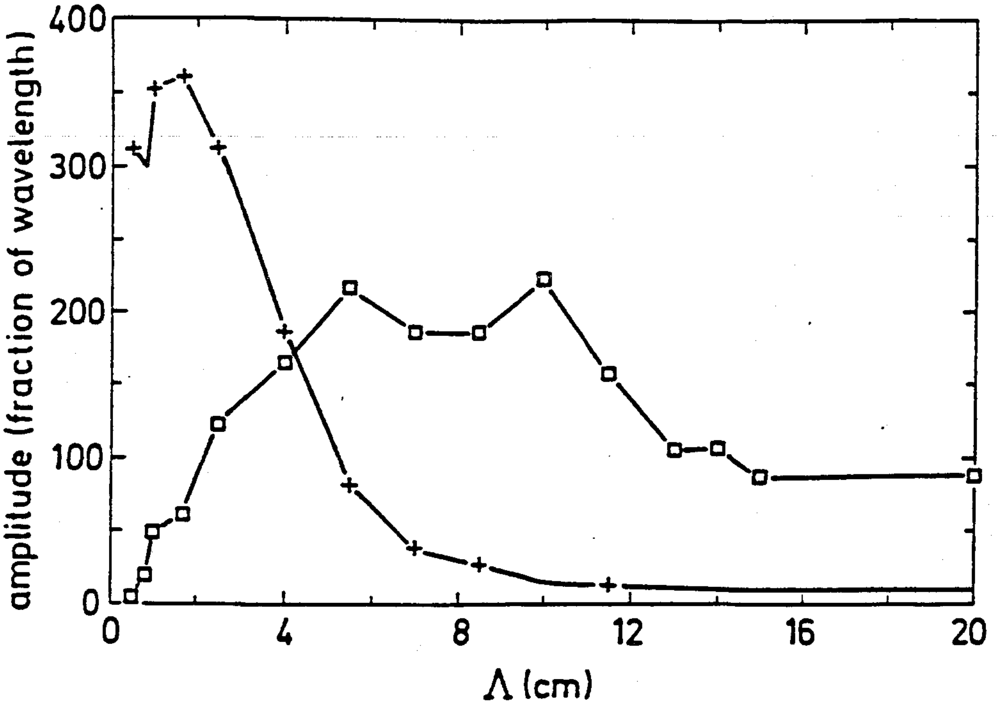


Figure 3: Tolerable surface deformation (in fractions of a wavelength of the light) as a function of its spatial wavelength Λ. Crosses: loss by scattering, squares: loss by bad interference.

arrangement has several advantages. The light path is doubled, recycling can be realized by insertion of only one extra mirror, and the specifications for the long wavelength surface distortions can be relaxed.

Fig. 3 shows the tolerable amplitude of surface deformations as a function of its spatial wavelength. Assumed were two-mirror delay lines with a mirror separation of 3 km, 26 reflections, a sinusoidal surface distortion (in two perpendicular directions) and a loss of 1%. The losses were either scattering (crosses) or bad interference (squares). At a spatial wavelength equal to the beam diameter (4.3 cm) both effects give the same contribution. One arm was assumed to have perfect mirrors; the overall demands may therefore be harder by a factor of $\sqrt{2}$.

3.4 Measurement of real surfaces

Present technology is able to reliably measure optical surfaces with the required precision. Determination of micro-roughness up to a spatial wavelength of about 5 mm and a vertical resolution of better than 0.1 nm is standard technology, obtained for instance by using a Mireau interferometer. This is by far sufficient for our purpose.

To determine surface deformations with adequate precision on a larger scale is more difficult. In order to satisfy our requirements, the optical company Zeiss (Oberkochen) has developed a measuring procedure[1] with a reproduciblity of 2 nm peak to valley and an rms value of 0.25 nm. This reproduciblity is achieved even after totally dismantling the measurement setup, building it up again and repeating the measurement. The procedure is based on some kind of a Michelson interferometer, where the contribution of each component to wavefront deformations is carefully determined, and each value is obtained by averaging over many single measurements.

Such measurements were taken on mirrors produced by Halle (Berlin) with a quoted tolerance in deviation from an ideal sphere of $\lambda/2$. In fact, the mirrors have been much better – the deviation was on the order of only $\lambda/50$. The mirrors were made from Herasil (fused silica), have a diameter of 240 mm, a substrate thickness of 75 mm and radius of curvature of 31.5 m. The measurement has been done in scan lines across the mirror surface with a lateral resolution of 1 to 2 mm. As an example one of the scan lines is plotted in Fig. 4, before and after coating, respectively. The coating was done with an ion sputtering technique. The totally different appearance is due to the fact that a highly reflecting concentric annular area was produced by rotating the mirror above a fixed off-axis evaporation source. The particular shape of the reflecting area matches the annular arrangement of reflection spots in a delay line.

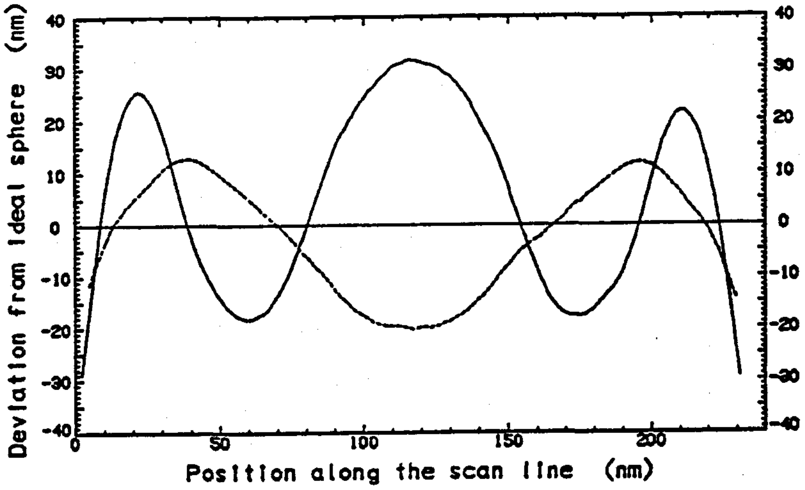


Figure 4: Surface profile of a mirror produced by conventional technique. Dashed line: before coating, solid line: after coating.

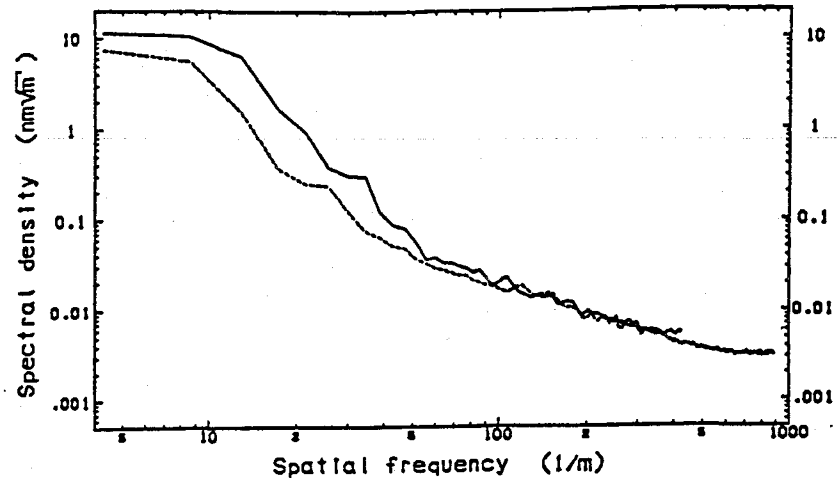


Figure 5: Linear spectral density of the surface deformations, before (dashed) and after coating (solid). The dashed line is an average over 140 scanlines, the solid one over 270 scanlines.

In Fig. 5 the spectral density of the surface deformations is plotted as a function of the spatial frequency. For wavelengths shorter than a few mm the measurement is limited by digitization noise. There is a striking similarity of the spectral density of the surface deformations between the coated and the uncoated case.

4. Absorption

Now let us consider the second main reason for losses, the absorption at coatings or inside components that are traversed by the light. The relative absorption losses allow an envisaged power gain only if

$$\frac{\Delta P}{P} < \frac{1}{NG} \tag{7}$$

for the case of N reflections in a delay line. For a recycling gain of 100 and 34 reflections, the relative absorption loss at each reflection has to stay below 300 ppm. This is no problem. The absorption inside the components traversed by the light is usually also small enough. More severe are the thermal effects related to the local heating by absorption: surface deformation and thermal lensing.

4.1 Surface deformation by local heating

In order to estimate the substrate deformation due to absorbed light power let us first consider the effect of absorption at the dielectric coatings. The heat is removed either

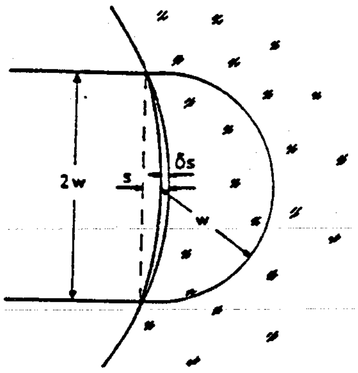


Figure 6: Surface deformation near a reflection spot heated by local absorption, as assumed in our simplified model.

by conduction or by radiation. For all relevant substrate materials heat conduction is dominant in the immediate vicinity of the reflection spot.

The steepest temperature gradient occurs in a hemisphere inside the substrate with its center at the beam center and its radius equal to the beam radius w (see Fig. 6). Outside this hemisphere the heat spreads over a rapidly increasing volume, leading to a weaker temperature gradient there. In equilibrium the entire substrate radiates all of the power away, as the heat conduction through the suspension wire is negligible.

The most important effect of local heating is a local change in the radius of curvature of the mirror at the reflection spot in question. In a first approximation the deformed mirror surface can be assumed to be locally spherical, just as the undeformed surface. This simple model is justified in retrospect by the good agreement with the elaborate analytical calculations of Hello and Vinet.[2]

The relevant thermal expansion caused by heating can be estimated by considering the heat transport through the hemisphere around the reflection spot

$$P_a = \kappa A \nabla T \approx 2\pi\kappa w^2 \frac{\delta T}{w}, \qquad (8)$$

leading to a change δs of the sagitta s (defined in Fig. 6) according to

$$\delta s \approx \alpha w \, \delta T/2 . \qquad (9)$$

Here P_a is the absorbed light power, κ the heat conductivity of the substrate material, A the area through which the heat is transported, δT the temperature drop across the hemisphere and α the thermal expansion coefficient.

Combining the last two equations we get

$$\delta s \approx \frac{\alpha}{4\pi\kappa} P_a . \qquad (10)$$

The relevant quantity for the magnitude of the effect is the ratio α/κ of thermal expansion to heat conductivity. Clearly this should be as small as possible.

The change δs in sagitta may be compared to the sagitta s itself (see Fig. 6). By simple geometrical considerations s can be estimated to be

$$s \approx \frac{\lambda}{2\pi} \tag{11}$$

in the near confocal case where the mirror separation equals the radius of curvature of the mirrors. The surface deformation δs certainly has to be kept far below that value, otherwise the beam would no longer be properly refocussed by the mirrors.

For quantitative estimates it may be useful to list α/κ for a few relevant materials (in units of 10^{-8} m/W): Fused silica 33, sapphire 28, silicon 1.67, ULE ± 2.3 (a material made to have a very low coefficient of thermal expansion), and diamond 0.13. The latter would in almost all respects be the optimal material for our purposes. Here its amazingly high thermal conductivity prevents a steep temperature gradient and thus a strong local deformation. Unfortunately the technique for growing single crystals of diamond is not yet well enough developed; only mm-thick plates of polycrystalline diamond have been produced so far. Another material with very low thermal expansion, Zerodur, is ruled out because of its high internal damping.[3,4]

4.2 Local heating in a delay line

In the just described treatment of the local deformations of optical components due to absorption of light power, it is its simplicity that allows to simulate analytically a variety of situations in real interferometers. An example is shown in Fig. 7. In an otherwise ideal interferometer with delay lines in its arms a constant absorption was assumed across the surface of one mirror. With increasing light power the local deformation grows correspondingly, leading to a deterioration of the interference quality at the output of the interferometer.

There are two possibilities to realize a particular number of reflections – mirror separation larger or smaller than the confocal arrangement. Striking is their different behaviour with respect to local heating by absorbed light power. The interference minimum – ideally zero at the dark fringe output – deteriorates with increasing light power in both cases in a nonlinear fashion. For the shorter mirror separation, with increasing light power the minimum deteriorates to about 10%, becomes perfect again, and keeps oscillating between good and bad interference in a fairly complicated way (Fig. 7a). The case of larger mirror separation shows a monotonic degradation of the interference quality with increasing absorbed light power.

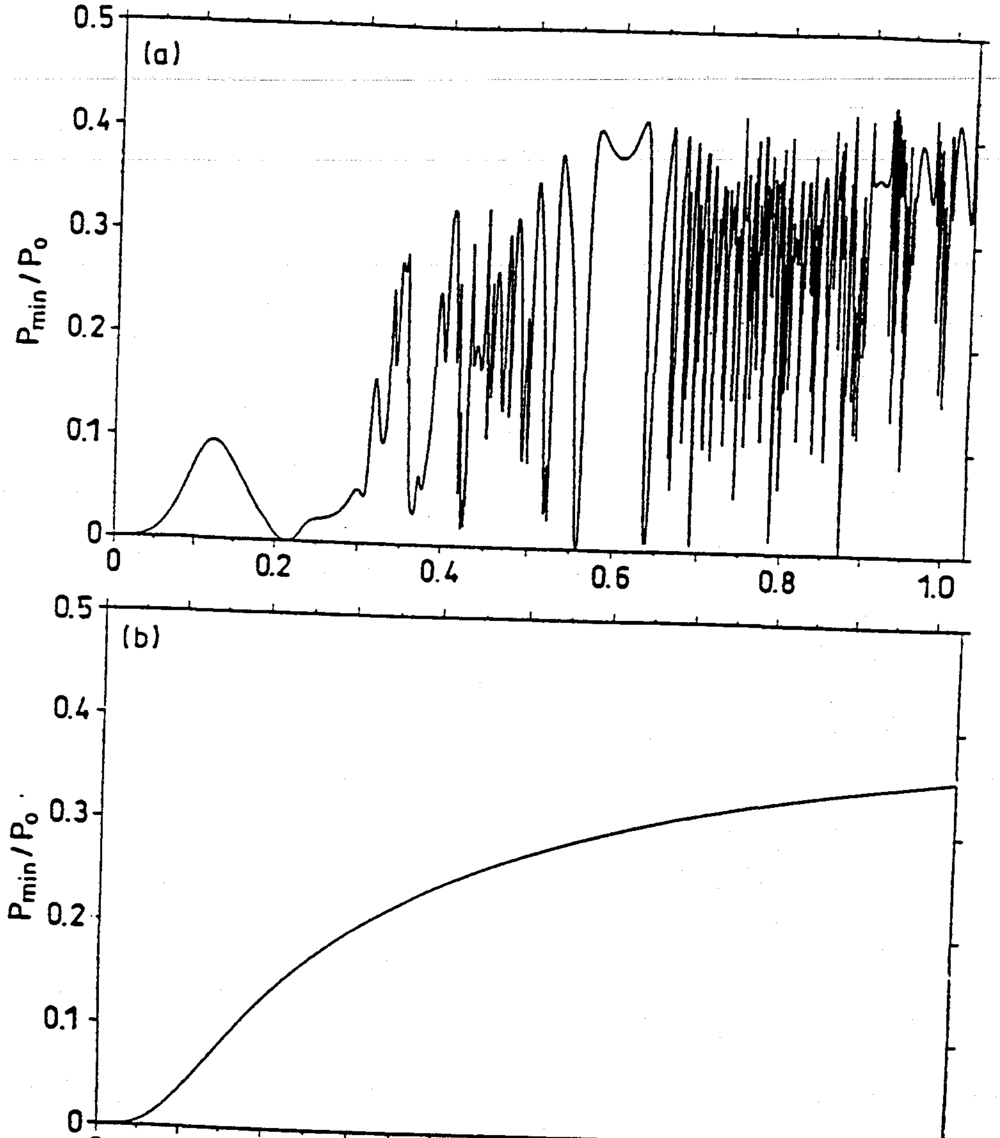


Figure 7: Deterioration of the interference quality of a delay line system as a function of local heating at the reflection spots.
(a) Mirror separation smaller than confocal.
(b) Mirror separation larger than confocal.

4.3 Thermal lensing

In addition to the deformation of a wavefront by a reflection at a locally deformed surface there is a deformation by an effect called thermal lensing. This effect occurs when the beam traverses a region with a temperature gradient because of a temperature dependence of the index of refraction, $\beta = \delta n/\delta T$. Such a temperature gradient may occur as a consequence of absorption of light at the surface coatings or inside the material traversed. What counts for the wavefront deformation is the path difference between the different parts of the beam, as introduced by the inhomogeneity in the index of refraction. For an estimate of the magnitude of this effect, let us again consider the heated hemisphere mentioned above and shown in Fig. 6, but now for the case of light passing through the substrate. The path difference δs between beam axis and outer parts of the beam, introduced by thermal lensing, is approximately

$$\delta s = \beta w \frac{\delta T}{P_a} \approx \frac{\beta}{4\pi\kappa} P_a . \tag{12}$$

This relation is very similar to the one derived above for the wavefront deformation occurring for reflections at deformed surfaces (Eq. 10). An expression similar to Eq. 12 holds for the case of absorption inside the substrate, with P_a now being the light power absorbed there. It is clear that one wants to keep the ratio β/κ small.

For the materials listed above the value for β/κ (in units of 10^{-8} m/W) is: fused silica 1000, ULE 850, sapphire 60 and diamond 1. Again diamond would be the optimum material because of its high thermal conductivity. On the other hand, fused silica, a material widely used for optical components, shows strong thermal lensing. But nowadays it can be provided with good purity and homogeneity and with low absorption. Therefore it is still considered to be used as substrate material. Silicon is not quoted here, since it is not transparent for the relevant wavelengths.

5. Measurement of absorption

Usually the absorption is measured calorimeterically, that is by monitoring the increase in temperature due to absorbed light power. This method is no longer appropriate for the low loss optical components available now. A more sensitive measurement was proposed by Olmstead *et al.*[5] and Boccara *et al.*[6] It is based on the effects of thermal expansion and thermal lensing treated in the previous chapter. For this purpose a strong laser beam with the wavelength in question is sent to the component to be investigated. The change in orientation of a weak and narrow probe beam, reflected at or transmitted through the locally heated substrate, is monitored with a position sensitive photodiode. Absorbed light powers on the order of 10^{-7} watts are detectable. This sensitivity is sufficient for our purposes.

It is well known that absorption losses usually are smaller in the infrared than in the visible, since absorption bands of the materials involved are further away. Measurements on high quality coatings on pure fused silica samples gave absorptions at the ppm level for the infrared.[7] A few months ago we have measured a few ppm absorption at coatings for green light.[8]

6. Power recycling in the presence of thermal distortion

The plots in Figs. 8 to 10 have been obtained in a computer simulation, taking into account the thermal effects in a real interferometer. As an example of the influence of thermal distortions on the performance of interferometric gravitational wave detectors, the light power P_{in} necessary to be delivered by a laser is plotted versus the effective light power P_{circ} circulating inside an interferometer with power recycling. The ratio $P_{\text{circ}}/P_{\text{in}}$ is equivalent to the power gain G defined above. .

For Figs. 8 and 9 the following parameters were assumed: 200 ppm Rayleigh scattering and 20 ppm absorption inside components traversed by the light (this absorption was assumed to differ between equivalent components by 10%), 34 reflections in the arms (in case of a Fabry-Perot system an equivalent finesse), 50 ppm and 20 ppm loss and absorption in the coatings, respectively. The losses in the coatings were as-

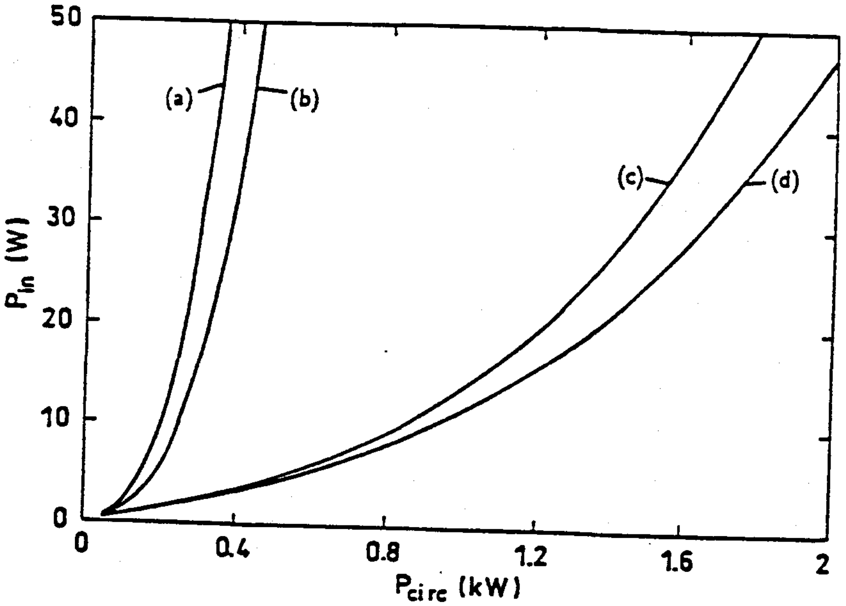


Figure 8: Power enhancement in an interferometer with power recycling and Fabry-Perot cavities in the arms.
(a) All components made of fused silica.
(b) Same as (a), but thermal lens in the beamsplitter compensated by the compensation plate down to 10%.
(c) All components made of sapphire.
(d) Same as (c), but compensation as in (b).

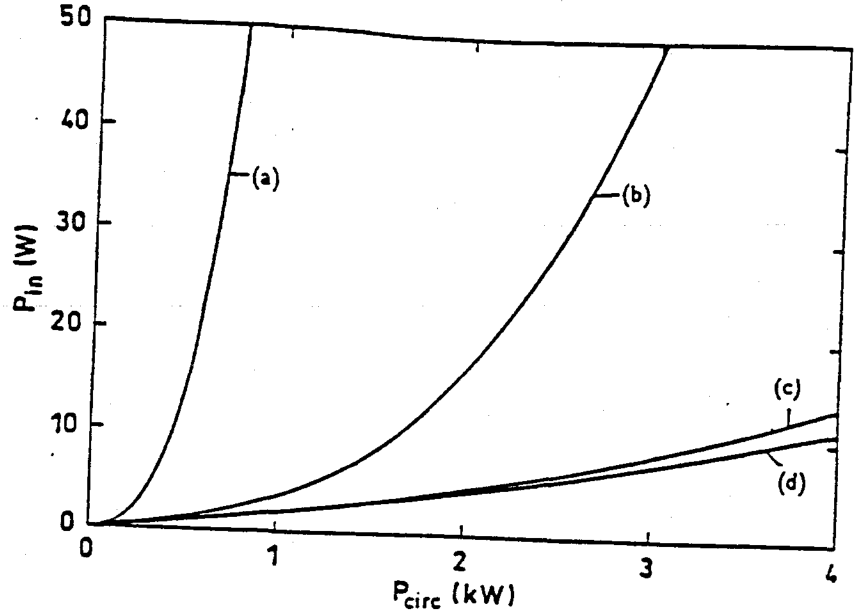


Figure 9: Same as in Fig. 8, but for delay-lines.
(a) All components made of fused silica, no compensation of thermal lensing.
For (b), (c) and (d) compensation of the beamsplitter lens as in Fig. 8, curve (b), was assumed:
(b) All components made of fused silica.
(c) Sapphire beamsplitter and fused silica mirrors.
(d) Sapphire beamsplitter and silicon mirrors.

sumed to vary between the different components by 20%, for the many reflections in the delay line a statistical fluctuation of the losses at the different reflections was assumed with a standard deviation equal to 20% of the average value. The wavelength of the light was assumed to be 514.5 nm.

In Fig. 8 the dominating effect for the rather poor performance of curve (a) is the thermal lens in the substrate of the coupling mirror in the Fabry-Perot cavities, introduced by absorption in the coating at the high power reflection spot. This lens reduces the coupling of the input beam to the mode inside the cavity. The thermal lens in the beamsplitter is of minor importance, as its reduction by a compensation plate improves the performance only from (a) to (b). This result is expected because of the lower intensity in the beamsplitter compared to the interior of the cavity. The striking improvement from (b) to (c) is due to the smaller thermal lens effect in sapphire compared to fused silica.

In Fig. 9 the dominating effect in curve (a) is the thermal lens in the beamsplitter, as its reduction gives the improvement from (a) to (b). In (b) the thermal lens in the beamsplitter is still dominating, as its further reduction by using sapphire

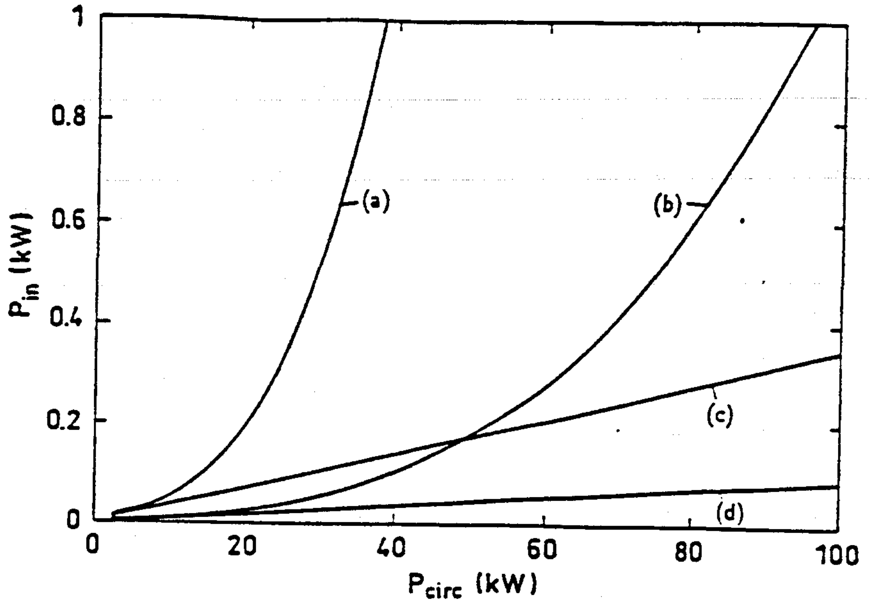


Figure 10: Power enhancement under optimistic conditions.
(a) A cavity system using components made of sapphire.
(b) A delay line system with a sapphire beamsplitter and silicon mirrors.
(c) A Fabry-Perot system using components made of diamond.
(d) A delay line system using components made of diamond.

as beamsplitter substrate gives the drastic improvement from (b) to (c). The thermal expansion is of minor importance, since in (a), (b) and (c) the mirrors have been assumed to be made of fused silica, whereas in curve (d) they have been assumed to be made of silicon with its low value for α/κ (see above).

For Fig. 10 substrate materials have been assumed that are not yet available in the required size and quality, but there are indications that this will change in the future. The calculations for Fig. 10 were based on lower losses than Figs. 8 and 9: substrate absorption 10 ppm, coating loss 20 ppm, coating absorption 2 ppm, 34 beams in each arm, and an unbalance in absorption (coating and substrate) of 10%, as defined for Figs. 8 and 9. The thermal lens in the beamsplitter was assumed to be compensated down to 10% by a compensation plate. Fig. 10 is certainly based on optimistic assumptions. But the low coating losses assumed have already been achieved for the infrared some time ago, and just recently we have measured a few ppm absorption and 25 ppm total loss on coatings for green light. Curves (c) and (d) show how perfect diamond would be as substrate material. The linear relation between $P_{\rm in}$ and $P_{\rm circ}$ in that particular case indicates that even at these high power levels there would be no thermal effects; the losses would be totally determined by scattering.

7. Conclusion

An interferometer with long baseline and high light power sets unusually high demands on the quality of the optical components. The rapid improvement of technology that took place recently already gives a smoothness of the surfaces in the 10^{-10} meter region over dimensions of millimeters. It is necessary to extend this scale to centimeters. On a larger scale the deviations of the surface shape with respect to an ideal sphere have to stay below $\lambda/100$. Research programs are going on, and the perspectives are encouraging.

Wavefront deformation because of thermal effects related to local heating by absorption limit the tolerable absorption level to the order of ppm. Such coatings are now available not only for the infrared, but also for the visible.

References

1. K. Freischlad, M. Küchel, W. Wiedmann, W. Kaiser, and M. Mayer, *Optical Testing and Metrology III: Recent Advances in Industrial Optical Inspection,* Proc. SPIE 1332 (1990) 8.

2. P. Hello and J. Y. Vinet, J. Phys. (France) **51** (1990) 2243.

3. R. Weiss, MIT Research Laboratory of Electronics Report No. 105, 1972 (unpublished).

4. W. Winkler, in *The Detection of Gravitational Radiation*, edited by D. Blair (Cambridge University Press, Cambridge, 1991).

5. M. A. Olmstead, N. M. Amer, and S. Kohn, Appl. Phys. A **32** (1983) 141.

6. A. C. Boccara, D. Fournier, W. Jackson, and N. M. Amer, Opt. Lett. **5** (1980) 377.

7. A. Brillet (private communication, 1990).

8. M. Engl, *Optische Verluste in Gravitationswellendetektoren,* Diplomarbeit, Aug. 1991 (unpublished).

Copyright © 1978 by John Wiley & Sons, Inc.

All rights reserved. Published simultaneously in Canada.

Reproduction or translation of any part of this work beyond that permitted by Sections 107 or 108 of the 1976 United States Copyright Act without the permission of the copyright owner is unlawful. Requests for permission or further information should be addressed to the Permissions Department, John Wiley & Sons, Inc.

Library of Congress Cataloging in Publication Data:

Optical shop testing.
(Wiley series in pure and applied optics)
Includes bibliographics and index.

1. Optical measurements. 2. Interferometer.
I. Malacara, Daniel, 1937–
QC367.059 681'.4'028 77-13222
ISBN 0-471-01973-9

Printed in the United States of America

10 9 8 7

367
059
1978

ASTR

Optical Shop Testing

Edited by

DANIEL MALACARA
Instituto Nacional de Astrofisica
Optica y Electrónica
Tonantzintla, Pue. México

John Wiley and Sons
New York / Chichester / Brisbane / Toronto

If the reference surface is spherical, and the surface under test is aspherical (hyperboloid or paraboloid), the ideal fringe patterns will be those of a Twyman Green interferometer for spherical aberration (see Chapter 2).

The reference surface may also be another aspherical surface that exactly matches the ideal configuration of the surface under test. This procedure is useful when a convex aspheric is to be made, since a concave aspheric can be made and tested more easily than a convex surface. The advantage of this method is that a null test is obtained. It has the disadvantage that the relative centering of the surfaces is very critical because both surfaces have well-defined axes and these must coincide while testing. This problem is not serious, however, because the centering can be achieved with some experience and with some device that permits careful adjustment.

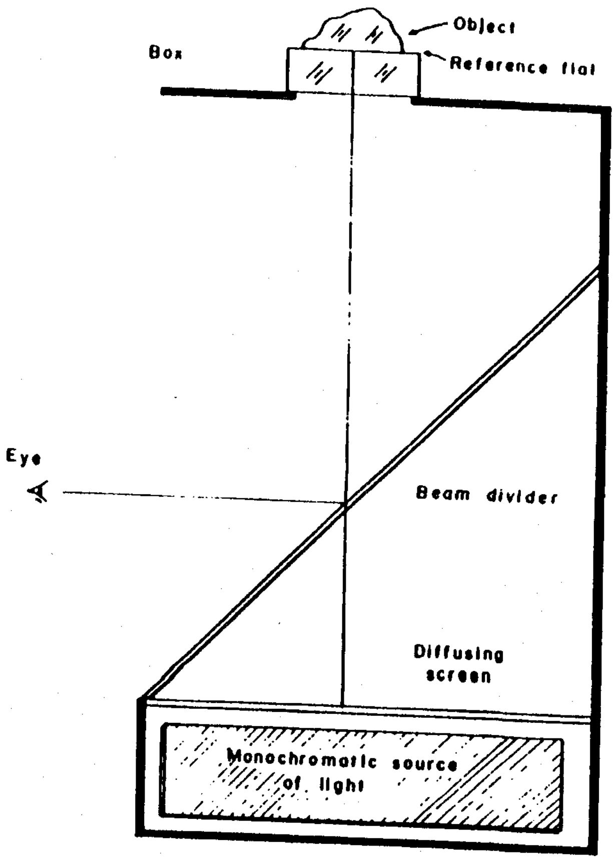


Figure 1.15. Schematic arrangement showing the method of testing opaque plane surfaces on irregular objects by placing them on top of the optical flats.

When mathematically interpreting the interferograms, it should be remembered that the OPD is measured perpendicularly to the surfaces whereas the surface sagitta z (see Appendix 1) is given along the optical axis. Therefore the OPD is given by $2(z_1 - z_2)\cos\theta$, where $\sin\theta = Sc$.

Measurement of Flatness of Opaque Surfaces. Sometimes we encounter plane surfaces generated on such metal substrates as steel, brass, and copper. An optical flat made of glass should be put on top of such objects for viewing Newton's fringes. It is not always the case that the metal object is in the form of a parallel plate. The plane surface may be generated on an otherwise irregular component, and hence some means of holding the component while testing becomes necessary. This can be avoided if we can put the object on top of the optical flat and observe the fringes through the bottom side of the flat. This sort of arrangement is shown in Fig. 1.15. Since most metal surfaces have reflectivities quite high compared to the value for a glass surface, the contrast of the fringes is not very good. To improve this situation, the optical flat is coated with a thin evaporated film of chromium or inconel having a reflectivity of about 30 to 40%. This brings about the formation of sharper, more visible fringes.

It is necessary to point out that, if the object is very heavy, it will bend the optical flat and the meaurement will not be accurate. Therefore this kind of arrangement is suitable for testing only small, light opaque objects. In dealing with heavy objects, it is preferable to place the optical flat on top of the object.

1.2. FIZEAU INTERFEROMETER

In the Newton interferometer the air gap between the surfaces is very small, of the order of a few wavelengths of light. Sometimes it is convenient to obtain fringes similar to the ones obtained in the Newton interferometer, but with a much larger air gap. When the air gap is larger, the surfaces need not be cleaned as thoroughly as they must be before being tested in the Newton interferometer. However, the surfaces may become scratched if not cleaned properly.

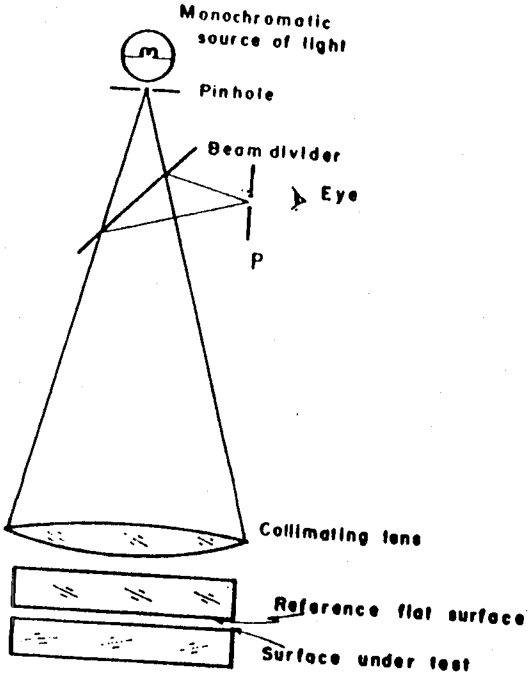
We showed earlier that the angular size of the source to be used depends on the air gap. If, for instance, the air gap between the flats is 5 mm, the permissible value of 2θ is given by Eq. 1.12 and is

$$2\theta \leqslant 10^{-2} \text{ radian,} \tag{1.20}$$

taking $\lambda = 5 \times 10^{-4}$ mm. Such a small angular source can be obtained generally by the use of a pinhole illuminated by a monochromatic source of light and located at the focus of a collimating lens or mirror. Thus, for example, a collimating lens of 250 mm focal length with a pinhole of 2.5 mm diameter will satisfy the above requirements. It can be seen that, as we increase the air gap more and more, the pinhole becomes smaller and smaller.

1.2.1. The Basic Fizeau Interferometer

From the foregoing considerations it is seen that we should have a collimating system in a Fizeau interferometer. Figure 1.16 shows the schematic arrangement of a Fizeau interferometer using a lens for collimation. The optical flat that serves as the reference is generally mounted along with the lens and is preadjusted so that the image of the pinhole reflected by the reference surface falls on the pinhole itself. Either the back side of the flat is antireflection coated, or (more conveniently) the reference optical flat is made in the form of a wedge (about 10 to 20 min of arc) so that the reflection from the back surface can be isolated. To view the fringes, a beam divider is located close to the pinhole. The surface under test is kept below the reference flat, and the air gap adjusted to the smallest value possible; then the air wedge is gradually reduced by manipulating the flat under test. When the air wedge is very large, two distinct images of the pinhole by the two surfaces can be seen in the plane *P* in Fig. 1.16. By making use of screws provided to tilt the flat under test, one can observe the movement of the image of the pinhole and can stop when it coincides with that of the reference flat. Then the observer places his eye at the plane *P* and sees, localized at the air gap, the fringes due to variation in the air gap thickness. Further adjustment, while looking at the fringes, can be made to alter the number and direction of the fringes. The interpretation of these fringes is exactly the same as that for Newton's fringes.



ure 1.16. Schematic arrangement of a Fizeau interferometer using a lens for collimation light.

Figure 1.17 is a schematic of a Fizeau interferometer using a concave

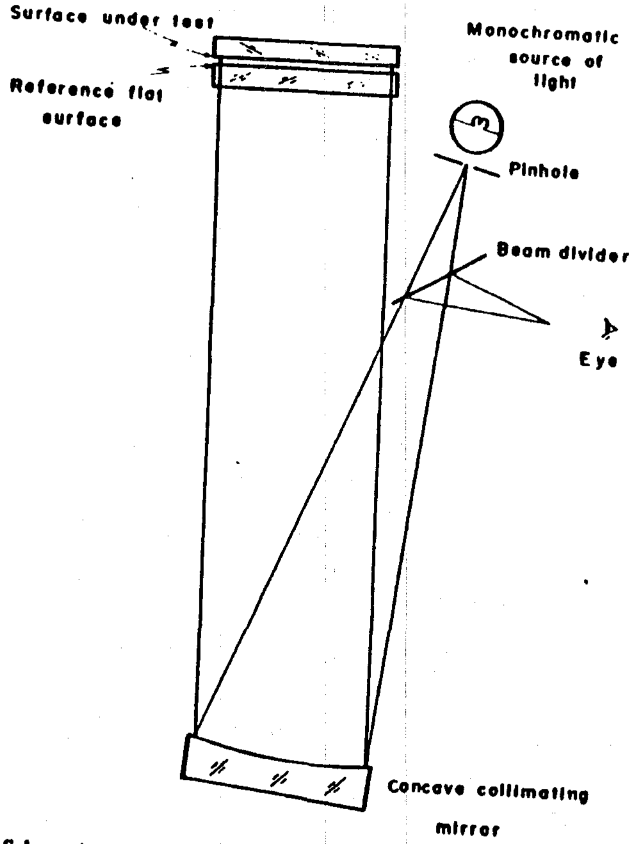


Figure 1.17. Schematic arrangement of a Fizeau interferometer using a concave mirror for collimation of light.

mirror as the collimating element. If a long focal length is chosen for the concave mirror, a spherical mirror can be used. For shorter focal lengths an off-axis paraboloidal mirror may be required. Both the schemes of Figs. 1.16 and 1.17 may be arranged in either a vertical (upright and inverted) or a horizontal layout. In the vertical situation the optical flats are horizontal, whereas in the horizontal layout the optical flats stand on their edges.

1.2.2. Liquid Reference Flats

It is well known that a liquid surface can be used as a reference flat. Basically the liquid surface has a radius of curvature equal to that of the earth. If the radius of the earth is taken as 6400 km, the sag of the surface is

$$\frac{y^2}{2R} = \frac{y^2}{2\times 6.4\times 10^9}\,(\text{mm}) \qquad (1.21)$$

where $2y$ is the diameter of the liquid surface considered. If we stipulate that this should not exceed $\lambda/100$ ($\lambda = 5\times 10^{-4}$ mm), then

$$y^2 \leqslant 6.4\times 10^4$$

or

$$2y \leqslant 512 \text{ mm}. \qquad (1.22)$$

Thus a liquid surface of about 0.5 meter diameter has a peak error of only $\lambda/100$ as compared to an ideal flat. Therefore it has been a very attractive proposition to build liquid flats as standard references. In practice, however, there are many problems, mainly in isolating the disturbing influence of vibrations. It is also necessary to exclude the region near the wall of the vessel which holds the liquid and to make sure that no dust particles are settling down on the surface. Possible liquids that can be useful for the purpose are those which are clear and viscous, such as glycerin, certain mineral oils, and bleached castor oil. Water is probably not suitable because of its low viscosity. Mercury may not be suitable because of its high reflectivity; the two interfering beams will have very unequal intensities, resulting in poor contrast of the fringes unless the surface under test is also suitably coated. However, mercury has been used as a true horizontal reference plane reflecting surface in certain surveying and astronomical instruments.

1.2.3. Testing Nearly Parallel Plates

In many applications glass plates having surfaces that are both plane and parallel are required. In such cases the small wedge angle of the plate can be determined by the Fizeau interferometer, and the reference flat of the

interferometer need not be used since the fringes are formed between the surfaces of the plate being tested. If A is the angle of the wedge and N is the refractive index of the glass, the angle between the front- and back-reflected wavefronts is given by $2NA$, and hence the fringes can be expressed as

$$2NA = \frac{\lambda}{d} \qquad (1.23)$$

where d is the distance between two consecutive bright or dark fringes. Hence the angle A is given by

$$A = \frac{\lambda}{2Nd}. \qquad (1.24)$$

To determine the thinner side of the wedge, a simple method is to touch the plate with a hot rod or even with a finger. Because of the slight local expansion, the thickness of the plate increases slightly. Hence a straight fringe passing through the region will form a kink pointing toward the thin side, as shown in Fig. 1.18. For instance, if we take $N = 1.5$, $\lambda = 5\times 10^{-4}$ mm, and $A = 5\times 10^{-6}$ (1 sec of arc), we get for d a value of about 33 mm. Hence a plate of 33 mm diameter, showing one fringe, has a wedge angle of 1 sec of arc. If the plate also has some surface errors, we get curved

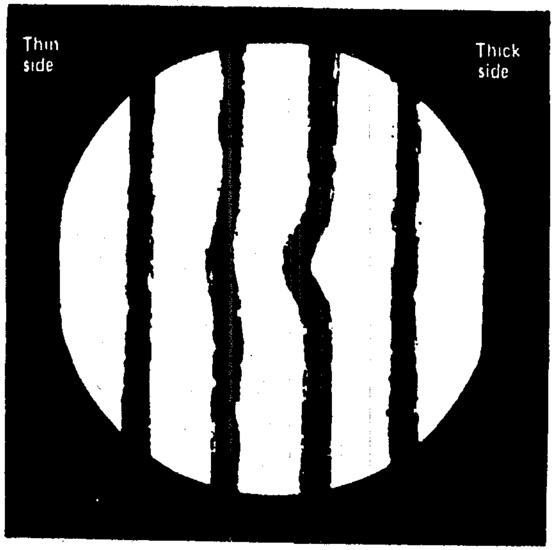


Figure 1.18. Kink formation in the straight Fizeau fringes of a slightly wedged plate, obtained by locally heating the plate. The kink is pointing toward the thin side of the wedge.

fringes, indicating both surface and wedge errors. If the surfaces are independently tested and found to be flat, and even in this situation one is getting curved fringes, these should be attributed to variation of the refractive index inside the plate in an irregular manner. In fact, by combining the tests on the Newton interferometer and the Fizeau interferometer for a parallel plate, it is possible to evaluate the refractive index variation (inhomogeneity) (Murty 1963, Murty 1964a, Forman 1964).

1.2.4. Fizeau Interferometer for Curved Surfaces

Just as collimated light is employed for testing optical flats on the Fizeau interferometer, it is possible to use either divergent or convergent light for testing curved surfaces. Figure 1.19 shows an arrangement for testing a concave surface against a reference convex surface. The point source of light is located at the center of curvature of the convex reference surface. The concave surface under test is adjusted until its center of curvature, too,

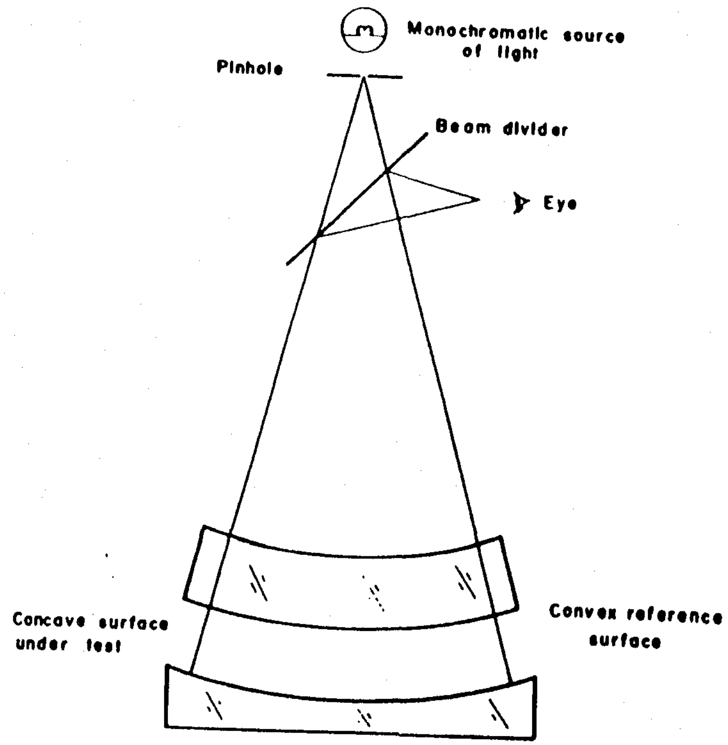


Figure 1.19. Fizeau interferometer setup for curved surfaces. Here the convex surface is the reference surface, and the concave surface is under test.

almost coincides with the point source of light. The procedure is exactly the same as before except that to achieve the uniform air gap we have to provide some translational motion also.

The same setup can be used very easily for checking the uniformity of thickness (concentricity) of spherical shells. In this case the interfering beams are obtained from the front and back of the two spherical concentric surfaces. Figure 1.20 shows this setup for testing the concentricity of a spherical shell. If the radii of curvature are correct but the shell has a wedge (the centers of curvature are laterally displaced), we get straight fringes characteristic of the wedge. The hot rod or finger touch procedure described in Section 1.2.3 can be adopted to determine which side is thinner. If the two radii are not of proper value ($r_1 - r_2 \neq t$, where r_1 and r_2 are the two radii and t is the center thickness), the value of t is not

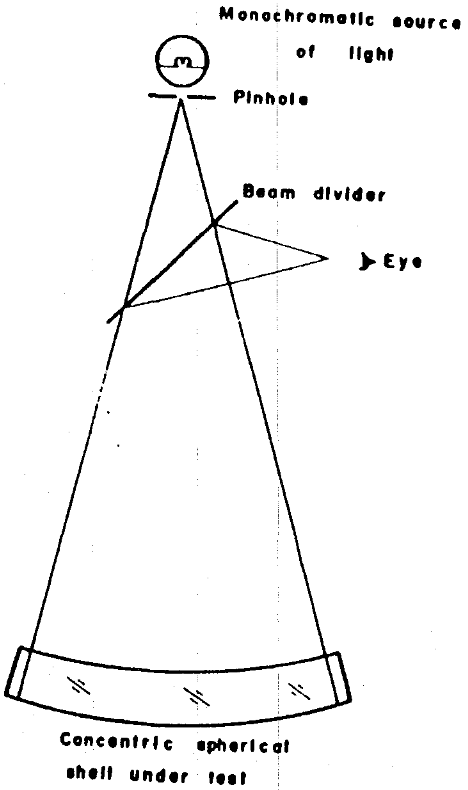

Monochromatic source of light
Pinhole
Beam divider
Eye
Concentric spherical shell under test

Figure 1.20. Fizeau interferometer setup for testing the concentricity of a spherical shell.

constant over the entire shell. Hence we get circular fringes like Newton's fringes. If in addition a wedge is present, the center of these circular fringes will be decentered with respect to the center of the shell. In this situation also we can adopt the hot rod or finger touch procedure to decide whether the shell is thin at the edge or at the center.

We can also have an arrangement for testing convex surfaces against a concave referenceç surface, as shown schematically in Fig. 1.21. Here we use a lens or a group of lenses at finite conjugate distances such that the point source of light is at one conjugate whereas the common center of curvature of the test surface and the reference surface is at the other conjugate. The concave reference surface is fixed to the instrument, while the convex surface under test is manipulated in the usual manner to obtain a uniform air gap.

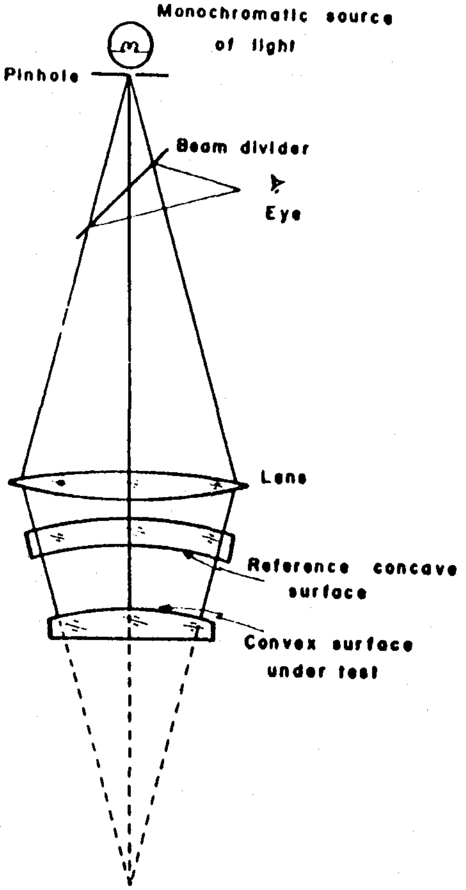

Monochromatic source of light
Pinhole
Beam divider
Eye
Lens
Reference concave surface
Convex surface under test

Figure 1.21. Fizeau interferometer setup for testing a convex surface against a concave reference surface.

1.2.5. Monochromaticity Requirement for the Source

In the Fizeau interferometer the air gap can be made quite small when plane surfaces are tested. Hence the total optical path difference involved may not exceed a few millimeters, and any small, low pressure mercury vapor lamp can be used with a green filter as the source of light. When testing for the wedge of thick plates of glass, there is a limitation on the thickness. For instance, a glass plate of 25 mm thickness gives rise to the equivalent of a 75 mm air gap between the front- and back-reflected wavefronts. For the lamp mentioned above, this OPD is probably the maximum we can use. For plates of greater thickness, the contrast of the interference fringes is greatly reduced because the lamp does not give a very sharp spectral line. Similarly, the same situation (low contrast) occurs when thick glass shells are tested or when spherical test plates are tested with one test plate, and hence the air gap can be large for certain situations. This limitation can be eliminated, however, if we can use a source of very high monochromaticity. Fortunately such a source, the laser, has recently become available. For our application the low power (2 mW) helium–neon gas laser operating in a single mode TEM_{00} and with a wavelength of emission at 6328 Å is ideal. With this as the source of light, we can tolerate an OPD of at least 2 meters and obtain Fizeau fringes of high contrast. Even larger OPDs are possible provided that a properly stabilized laser is chosen and vibration isolation is provided for the instrument.

1.2.6. Fizeau Interferometer with Laser Source

We shall now describe a Fizeau interferometer using a source such as the helium–neon gas laser of about 2 mW power lasing at 6328 Å in the single mode. Such an instrument, P. F. Forman kindly pointed out to the author, is manufactured by the Zygo Corporation, Middletown, Connecticut. A schematic diagram is shown in Fig. 1.22. A very well corrected objective serves to collimate the light from the pinhole, illuminated by a combination of the laser and a microscope objective. Between the collimating objective and the pinhole (spatial filter), a beam divider is placed so that the fringes can be observed from the side. It is also desirable to provide a screen, upon which the Fizeau fringes are projected, to avoid looking into the instrument, as is normally done when conventional light sources are used. The laser has a high radiance compared to other sources, and a direct view may be dangerous to the eye under some circumstances. The

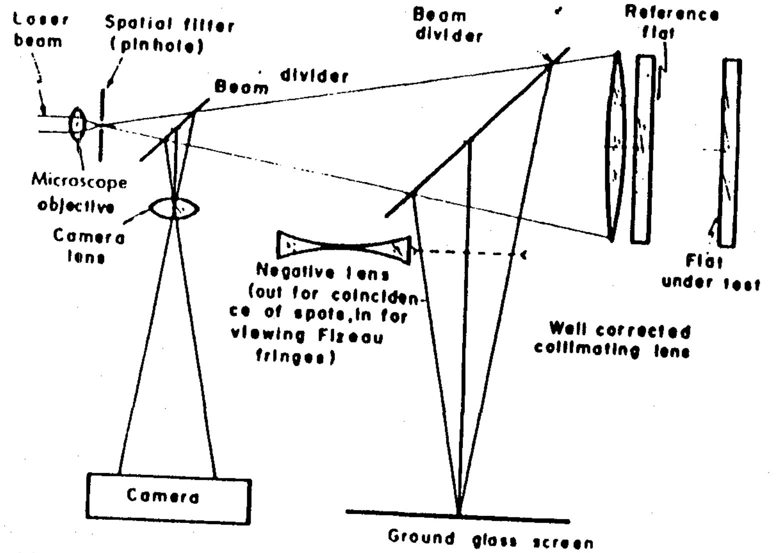


Figure 1.22. Schematic arrangement of a Fizeau interferometer using a laser source. The scheme shown here is for plane surfaces.

reference plane surface is permanently adjusted so that the reflected image of the pinhole is autocollimated. The surface under test is adjusted until the image reflected from it also comes into coincidence with the pinhole. To facilitate preliminary adjustment, the screen is used to project the two pinhole images from the two reflecting plane surfaces. This is accomplished by removing the negative lens between the beam divider and the ground-glass screen. The pinhole image from the reference surface is at the center of the screen, whereas the one from the surface under test is somewhere on the screen; by manipulation of this surface, the two spots of light on the screen can be brought into coincidence. Then the negative lens is inserted in the path, and the Fizeau fringes are projected on the screen. These fringes can be further adjusted in direction and number as required. By the use of another beam divider, it is possible to divert part of the beam to a camera for taking a photograph of the fringe pattern. The whole instrument must be mounted on a suitable vibration-isolated platform.

The instrument described above can be used for various other applications that are normally not possible with conventional sources of light. We describe some such applications in the sections that follow. In addition, many possibilities exist for other applications, depending on the particular situations involved.

1.2.7. Multiple-Beam Fizeau Setup

If, instead of two-beam fringes, multiple-beam fringes of very good sharpness are required, the reference optical flat and the optical flat under test are coated with a reflecting material of about 80 to 90% reflectivity (see Chapter 6), such as aluminum or silver. If higher reflectivities are required, multilayer dielectric coatings can be applied. In fact the instrument may be provided with several reference flats having coatings of different reflectivities.

1.2.8. Testing the Inhomogeneity of Large Glass or Fused Quartz Samples

The sample is made in the form of a parallel plate. The surfaces should be made as flat as possible with a peak error of not more than λ. Then the plate is sandwiched between two well-made parallel plates of glass with a suitable oil matching the refractive index of the sample. This will make the small surface errors of the sample negligible, and only straight fringe deformation due to the inhomogeneity of the sample will be seen. If the sandwich is kept in the cavity formed by the two coated mirrors, very sharp dark fringes on a bright background are obtained. If, for instance, the maximum fringe deviation from straightness is k and the distance between two fringes is d, the optical path difference is $(k/d)\lambda$. Now the OPD due to the inhomogeneity ΔN and thickness t of the sample is given

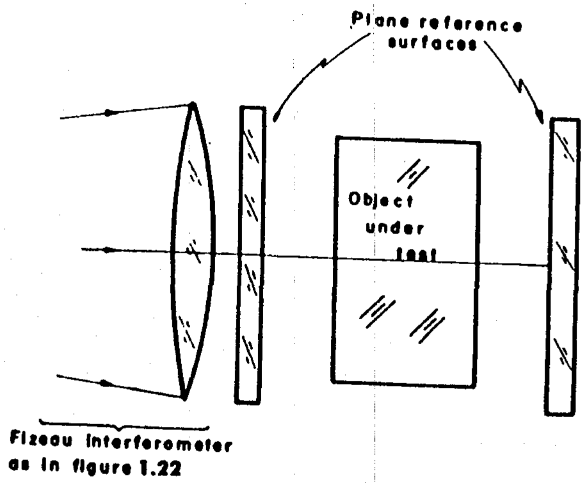


Figure 1.23. Schematic arrangement of a Fizeau interferometer for testing the homogeneity of solid samples of glass, fused quartz, and so on.

by $2\Delta N \cdot t$, and hence

$$\Delta N = \left(\frac{k}{d}\right)\left(\frac{\lambda}{2t}\right). \qquad (1.25)$$

As an example, if $k/d = 0.25$, $\lambda = 6328$ Å, and $t = 50$ mm, we have $\Delta N = 1.6 \times 10^{-6}$. Thus a maximum variation of 1.6×10^{-6} may be expected in the sample for the direction in which it has been tested. Figure 1.23 shows the schematic arrangement of the Fizeau interferometer for the method just described.

1.2.9. Testing Cube Corner and Right Angle Prisms

If the right angles of cube corner and right angle prisms are exact without any error, they reflect an incident plane wavefront as a single emerging plane wavefront. Otherwise the reflected wavefront consists of several plane wavefronts. Thus it is possible to see the errors in the wavefronts reflected from such prisms. Because of total internal reflection, the intensity of reflected light from these prisms is very high. The reference flat, if it is not coated, will give only a very low reflection, and hence fringes of poor contrast result. On the other hand, if the reference flat is coated, a confusing system of fringes will appear, because of multiple reflections, when there is any error in the right angle. Hence it is preferable to obtain effectively two-beam interference fringes. This can be done in two ways.

The reference flat, of course, is uncoated. To reduce the effective reflectivity of the right angle or cube corner prism, we can introduce a parallel plate of glass coated with a metallic film having a transmission between 20 and 30%. In this case the intensities of the two beams match reasonably well, and we get good-contrast two-beam fringes. The coated plate between the prism and the uncoated reference flat should be tilted sufficiently to avoid the directly reflected beam. This method is shown schematically in Fig. 1.24.

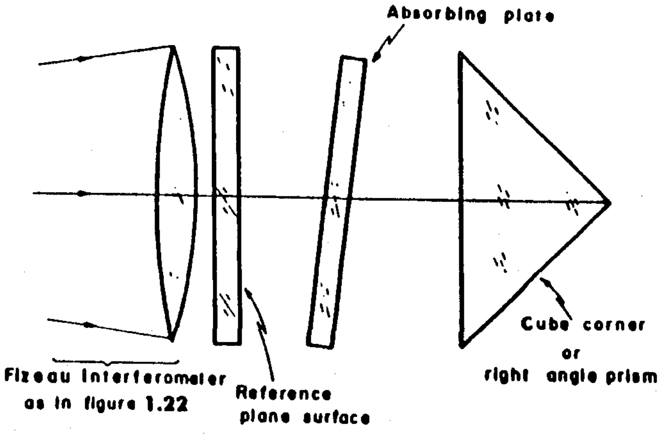


Figure 1.24. Schematic arrangement of a Fizeau interferometer for testing cube corner prisms and right angle prisms. Here an absorbing plate is inserted between the prisms and the reference flat surface to equalize the intensities of the two interfering beams.

Another possible method is to reduce the reflectivity of one of the total reflecting surfaces. This can be done by constructing a special cell in which the prism is mounted, and behind one reflecting surface a thin layer of water or some other suitable liquid is in contact with the surface. Thus, in effect, the refractive index difference is reduced at one total internal reflecting surface, and hence the intensity of the wavefront reflected from the prism matches that of an uncoated flat. This method is shown schematically in Fig. 1.25. A very good cube corner prism will give rise to an interferogram like that shown in Fig. 1.26. The fringes are straight throughout the aperture. A cube corner prism with angular errors produces an interferogram such as is shown in Fig. 1.27, in which the straight fringes abruptly change their direction. Figures 1.28 and 1.29 show similar situations for a right angle prism of no error and of some angular error, respectively. If, in addition to angle errors, the surfaces are not flat or the glass is not homogeneous, an interferogram with curved fringes is obtained.

We describe here a brief method for obtaining the angular error in a right angle prism. If the right angle has an error, the fringes look like those

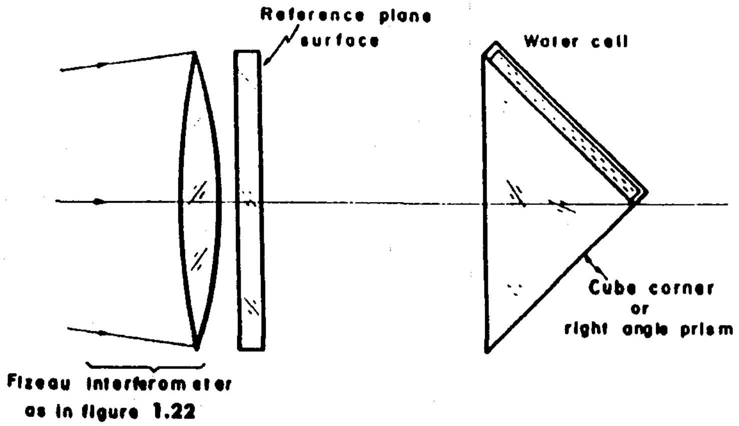


Figure 1.25. A scheme for reducing the intensity of reflected light from the cube corner prism and the right angle prism. One of the total internally reflecting faces is brought into contact with water or some other liquid by the use of a cell behind it.

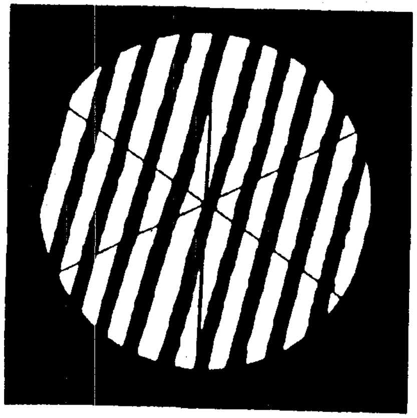


Figure 1.26. The interferogram of a very good cube corner prism. The reference flat surface is to be tilted slightly to obtain the straight fringes.

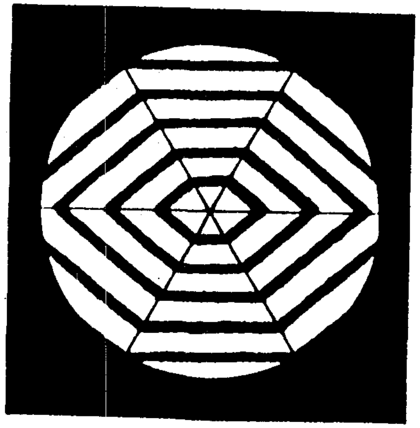


Figure 1.27. The interferogram of a cube corner prism with angular errors.

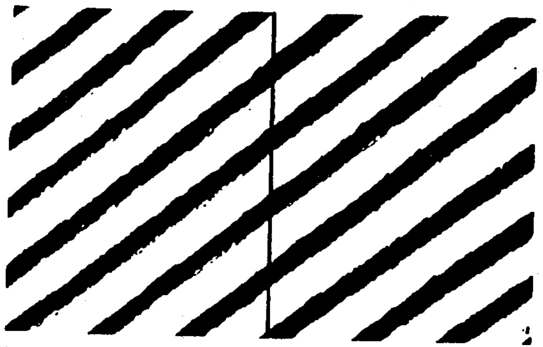


Figure 1.28. The interferogram of a good right angle prism.

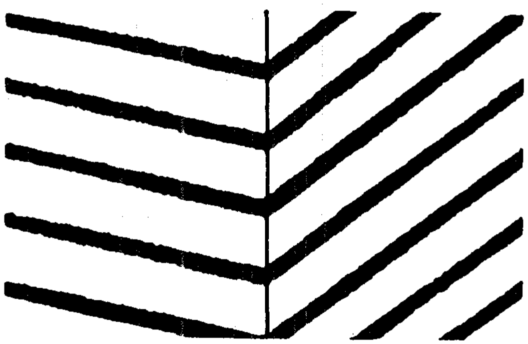


Figure 1.29. The interferogram of a right angle prism with a slight error in the 90° angle.

shown in Fig. 1.29 and can be manipulated to look like those in Fig. 1.30. If $2L$ is the width of the face of the prism, $\pi/2 \pm \epsilon$ is the angle of the prism, d is the distance between two successive fringes, k is the deviation of the fringe from the straight fringe after bending, N is the refractive index of the prism, and λ is the wavelength used, the error is given by

$$\epsilon = \left(\frac{k}{d}\right)\left(\frac{\lambda}{4NL}\right). \tag{1.26}$$

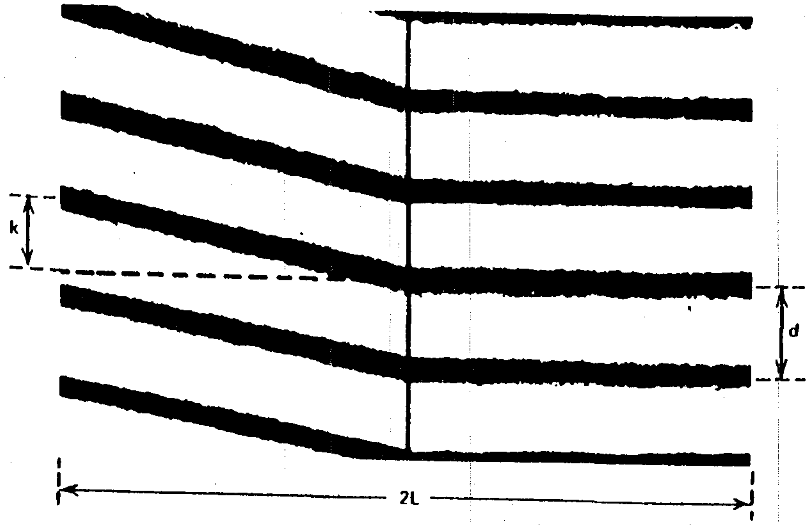


Figure 1.30. The interferogram of a right angle prism with a slight error in the right angle. The fringes are adjusted so that they are perpendicular to the roof edge on one side.

For example, for a prism of 100 mm face width and $k/d=0.25$, the error ϵ of the 90° angle is about 1 sec of arc. In regard to the sign of the error, the hot rod or finger procedure described in Section 1.2.3 can be used.

1.2.10. Testing Concave or Convex Surfaces

The reference surface is again the uncoated flat surface which is part of the Fizeau interferometer. The collimated light from the instrument, after passing through the optical flat, is again focused by the use of another highly corrected lens. If the surface is concave, it is set up as shown in Fig. 1.31; if convex, as shown in Fig. 1.32. When the surface is spherical and the center of curvature coincides with the focus of the lens, a plane wavefront is reflected back. Hence we should obtain straight fringes due to interference of the two beams. If the optical reference flat and the spherical surfaces are coated with high reflecting material, we can get very sharp, multiple-beam Fizeau fringes. If the surfaces are not spherical but are aspheric, appropriate null lenses must be used in the interferometer. This setup can also be used to measure the radius of curvature if a length-measuring arrangement is provided.

1.2.11. Quality of Collimation Lens Required

We shall briefly examine the quality of collimating lens required for the Fizeau interferometer. Basically we are interested in determining the variation in air-gap thickness. However, the OPD is a function of not only air-gap thickness but also the angle of illumination, and at a particular point this is $2t\cos\theta$. The air gap t varies because of the surface defects of the flats under test, while the variation of θ is due to the finite size of the source and to the aberration of the collimating lens.

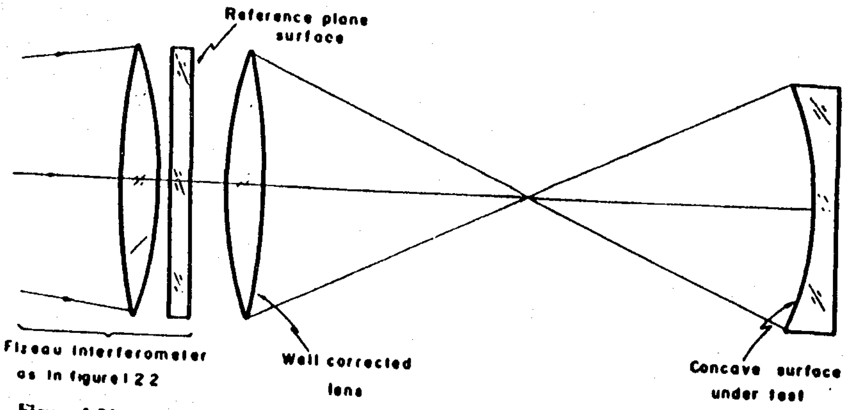

Reference plane surface
Fizeau interferometer as in figure 1.22
Well corrected lens
Concave surface under test

Figure 1.31. Schematic diagram of a Fizeau interferometer for testing a concave surface.

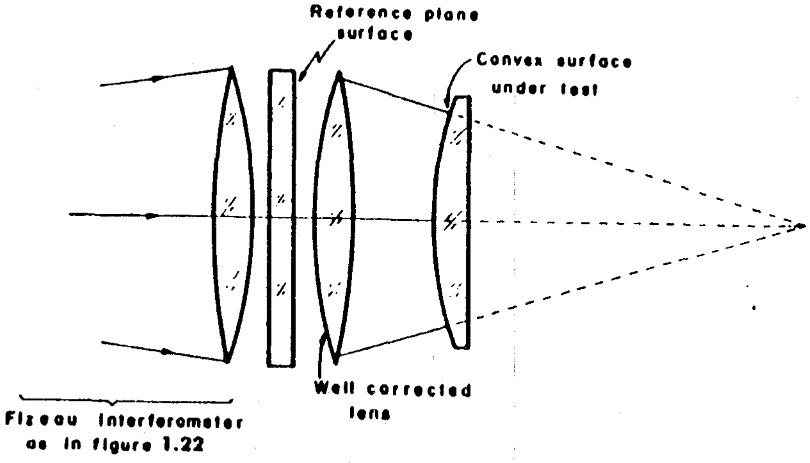

Reference plane surface
Convex surface under test
Well corrected lens
Fizeau interferometer as in figure 1.22

Figure 1.32. Schematic diagram of a Fizeau interferometer for testing a convex surface.

For Fizeau interferometers using conventional sources of light, the maximum air gap that is useful is 50 mm. Also, in this case we have to consider the size of the source and the aberration of the lens separately. The effect of the size of the source is mainly on the visibility of the Fizeau fringes. The excess optical path difference $t\theta^2$ should be less than $\lambda/4$ for good contrast of the Fizeau fringes, and the pinhole is chosen to satisfy this condition. The effect of the pinhole is uniform over the entire area of the Fizeau fringes. On the other hand, the effect of aberration in the collimating lens is not uniform. Thus we have to consider the angular aberration of the lens and its effect. If ϕ is the maximum angular aberration of the lens, then $t\phi^2$ should be less than $k\lambda$, where k is a small fraction that depends on the accuracy required in the instrument. Thus let us set $k=0.001$ so that the contribution of $t\phi^2$ is 0.001λ. Taking a maximum value of $t=50$ mm for the ordinary source situation, we have

$$\phi^2 < \frac{0.001\lambda}{t} \approx 10^{-8}$$

or

$$\phi < 10^{-4} \text{ radian.} \qquad (1.27)$$

This angular aberration is quite large, being of the order of 20 sec of arc. Hence suitable lenses or mirror systems can be designed for the purpose (Taylor 1957, Yoder 1957, Murty and Shukla 1970).

In the situation where the laser is the source of light, there is a much higher limit for the value of t. Even though several meters can be used for t, we shall set $t = 1000$ mm. In this case, using $\lambda = 6.328 \times 10^{-4}$ mm, we get for ϕ an upper limit of 5 sec of arc. Hence it is not difficult to design a collimating system to satisfy this condition.

Another aspect that is important, especially with large values of t, is the lateral shear one can get in the instrument. To avoid this, the autocollimated pinhole images must coincide with the pinhole itself. Similarly, if the collimating lens is not properly collimated, either a convergent or a divergent beam will emerge. The collimation may be accurately performed by using any of the various devices available, such as the plane parallel plate shearing interferometer (Murty 1964b).

1.3. HAIDINGER INTERFEROMETER

With the Newton and Fizeau interferometers, we were basically interested in finding the variation in the air-gap thickness. In these cases the fringes are referred to as fringes of equal thickness. If, however, the thickness of the air gap is uniform and it is illuminated by a source of large angular size, we get what are called fringes of equal inclination. These fringes are formed at infinity, and a suitable lens can be used to focus them on its focal plane. If the parallel gap is that of air, we have the simple relation $2t\cos\theta = n\lambda$, as given in Eq. 1.9, from which we can easily see that, for a constant value of t, we obtain fringes of equal inclination that are circles and are formed at infinity.

If the air gap is replaced by a solid plate such as a very good parallel plate of glass, Eq. 1.9 is modified slightly to include the effect of the refractive index N of the plate and becomes

$$2Nt\cos\theta' = n\lambda \tag{1.28}$$

where θ' is the angle of refraction inside the glass plate. For small values of θ', we may approximate this expression as

$$2Nt - \left(\frac{t}{N}\right)\theta^2 = n\lambda. \tag{1.29}$$

To see Haidinger fringes with simple equipment, the following method, illustrated in Fig. 1.33, may be adopted. A parallel plate of glass is kept on a black paper and is illuminated by the diffuse light reflected from a white card at 45°. At the center of the white card is a small hole through which we look at the plate. With relaxed accommodation our eyes are essentially focused at infinity, and we see a system of concentric circular fringes. For the light source we can use a sodium or even a fluorescent lamp.

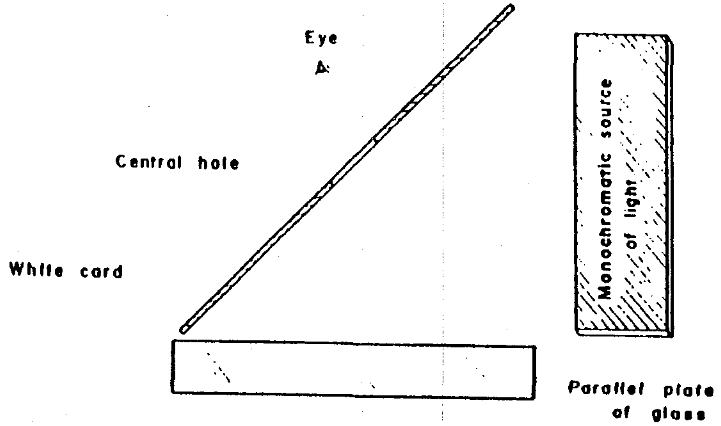


Figure 1.33. A simple arrangement to see Haidinger fringes for a nearly parallel glass plate.

A somewhat better method is to use a lens for focusing the system of Haidinger fringes on its focal plane. This requires a setup almost identical to that for the Fizeau interferometer. The only difference is that, instead of a pinhole, a wider aperture is used to have a large angular size for the source. The Haidinger fringes are then formed in the focal plane of the lens.

1.3.1. Applications of Haidinger Fringes

The Haidinger fringes may be used as a complementary test to that provided by the Fizeau interferometer. If we are testing a nearly parallel plate, we can find its wedge angle either by the Fizeau or by the Haidinger method. In the Haidinger method we look for the stability of the concentric fringes as we move our line of sight across the plate with a small aperture. If t is slowly varying, the center of the circular fringe system also appears to change. If t is decreasing, we are moving toward the thinner side of the wedge and in this case the circular Haidinger fringes appear to expand from the center. On the contrary, the fringes appear to converge to the center if we are moving toward the thick side of the wedge. If we note how many times the center of the fringe system has gone through bright and dark cycles, we can also estimate the wedge angle in the same manner as for the Fizeau situation.

1.3.2. Use of Laser Source for Haidinger Interferometer

A helium–neon laser source of low power is very useful for this interferometer, as it is for the Fizeau instrument. It enables the fringes to be projected on a screen. In this case the laser can be made to give effectively a point

© H A Macleod 1986

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the publisher.

British Library Cataloguing in Publication Data

Macleod, H. A.
Thin-film optical filters—2nd ed.
1. Light filters 2. Thin filters—Optical properties
I. Title
621.36 QC373.L5

ISBN 0-85274-784-5

Consultant Editor: **Professor W T Welford**
Imperial College, London

Published by Adam Hilger Ltd
Techno House, Redcliffe Way, Bristol BS1 6NX

Phototypeset by Macmillan India Ltd, Bangalore
Printed in Great Britain by J W Arrowsmith Ltd, Bristol

Thin-film optical filters

Second edition

H A Macleod

Professor of Optical Sciences
University of Arizona

Adam Hilger Ltd, Bristol

REF. A,

γ	*The equivalent phase thickness of a symmetrical assembly (p 192).*
Δ_q	(η_p/η_s) where η_p and η_s are modified admittances (p 316). This is a quantity used in the design of polarisation-free coatings. (p 339).
ε	Indicates a small error or a departure from a reference value of a number (p 74).
ε	The permittivity of a medium (p 12).
η	The tilted optical admittance (p 25).
η_m	The tilted admittance of the substrate. See y_m. (p 36).
η_E	The equivalent admittance of a symmetrical assembly. See also *E*. (p 120).
θ	The angle of incidence in a medium (p 17).
θ_0	The angle of incidence in the incident medium (p 17).
λ	The wavelength of the light, usually the wavelength in free space (p 14).
λ_0	The reference wavelength. See g. (p 76).
ν_0	The reference wavenumber. $\nu_0 = 1/\lambda_0$. See g.
ρ	The amplitude reflection coefficient (p 20).
ρ	The electric charge density (p 14).
τ	The amplitude transmission coefficient (p 20).
ϕ	The phase shift on reflection (p 37).
ψ	Potential transmittance. $\psi = T/(1-R)$. (p 39).
ψ	Used in some limited calculations on pages 199 and 200 to represent $2\delta_p/\delta_q$ (p 199).

1 Introduction

This book is intended to form an introduction to thin-film optical filters for both the manufacturer and the user. It does not pretend to present a detailed account of the entire field of thin-film optics, but it is hoped that it will form a supplement to those works already available in the field and which only briefly touch on the principles of filters. For the sake of a degree of completeness, it has been thought desirable to repeat again some of the information that will be found elsewhere in textbooks, referring the reader to more complete sources for greater detail. The topics covered are a mixture of design, manufacture, performance and application, including enough of the basic mathematics of optical thin films for the reader to carry out thin-film calculations. The aim has been to present, as far as possible, a unified treatment, and there are some alternative methods of analysis which are not discussed. For a much more complete study of thin-film calculations, the reader should consult the book by Knittl[1]. Similarly, some of the manufacturing methods are not dealt with in depth because there is an excellent textbook on the vacuum deposition of thin films by Holland[2] which, although it was written more than 20 years ago, is still relevant and topical. For further information on coatings involving a few layers, such as beam splitters, high reflectance coatings and metal–dielectric filters, together with information on alternative deposition techniques, the book by Anders[3] will be found useful. There is the well known book by Heavens[4] which deals principally with the basic optical properties, mainly of a single film, and includes much information on thin-film calculations. More recently, there has been a survey paper by Lissberger[5], the reports of two meetings devoted entirely to thin-film optical coatings[6,7], and an excellent review of the entire field of optical filters by Dobrowolski[8]. There is also the recent and useful book on filter design by Liddel[9].

In a work of this size, it is not possible to cover the entire field of thin-film optical devices in the detail that some of them may deserve. The selection of topics is due, at least in part, to the author's own preferences and knowledge. In this book, optical filters have been interpreted fairly broadly to include such items as antireflection and high-reflectance coatings.

EARLY HISTORY

The earliest of what might be called modern thin-film optics was the discovery, independently, by Robert Boyle and Robert Hooke of the phenomenon known as 'Newton's rings'. The explanation of this is nowadays thought to be a very simple matter, being due to interference in a single thin film of varying thickness. However, at that time, the theory of the nature of light was not sufficiently far advanced, and the explanation of this and a number of similar observations made in the same period by Sir Isaac Newton on thin films eluded scientists for almost a further 150 years. Then, on 12 November 1801, in a Bakerian Lecture to the Royal Society, Thomas Young enunciated the principle of the interference of light and produced the first satisfactory explanation of the effect. As Henry Crew[10] has put it, 'This simple but tremendously important fact that two rays of light incident upon a single point can be added together to produce darkness at that point is, as I see it, the one outstanding discovery which the world owes to Thomas Young.'

Young's theory was far from achieving universal acceptance. Indeed Young became the victim of a bitter personal attack, against which he had the greatest difficulty defending himself. Recognition came slowly and depended much on the work of Augustin Jean Fresnel[11] who, quite independently, also arrived at a wave theory of light. Fresnel's discovery, in 1816, that two beams of light which are polarised at right angles could never interfere, established the transverse nature of light waves. Then Fresnel combined Young's interference principle and Huygens' ideas of light propagation into an elegant theory of diffraction. It was Fresnel who put the wave theory of light on such a firm foundation that it has never been shaken. For the thin-film worker, Fresnel's laws, governing the amplitude and phase of light reflected and transmitted at a single boundary, are of major importance. It has been pointed out recently by Knittl[12], that it was Fresnel who first summed an infinite series of rays to determine the transmittance of a thick sheet of glass and that it was Simeon Denis Poisson, in correspondence with Fresnel, who included interference effects in the summation to arrive at the important results that a half-wave thick film does not change the reflectance of a surface, and that a quarter-wave thick film of index $(n_0 n_s)^{1/2}$ will reduce to zero the reflectance of a surface between two media of indices n_s and n_0. Fresnel died in 1827, at the early age of 39.

In 1873, the great work of James Clerk Maxwell, *A Treatise on Electricity and Magnetism*[13], was published, and in his system of equations we have all the basic theory for the analysis of thin-film optical problems.

Meanwhile, in 1817, Joseph Fraunhofer had made what are probably the first ever antireflection coatings. It is worth quoting his observations at some length because they show the considerable insight which he had, even at that early date, into the physical causes of the effects which were produced. The following is a translation of part of the paper as it appears in the collected works[14].

> Before I quote the experiments which I have made on this I will give the method which I have made use of to tell in a short time whether the glass will withstand the influence of the atmosphere. If one grinds and then polishes as finely as possible, one *surface of glass which has become etched* through long exposure to the atmosphere, then wets one part of the surface, for example half, with concentrated sulphuric or nitric acid and lets it work on the surface for twenty-four hours, one finds after cleaning away the acid that that part of the surface on which the acid was, reflects much less light than the other half, that is it shines less although it is not in the least etched and still transmits as much light as the other half, so that one can detect no difference on looking through. The difference in the amount of reflected light will be most easily detected if one lets the light strike approximately vertically. It is the greater the more the glass is liable to tarnish and become etched. If the polish on the glass is not very good this difference will be less noticeable. On glass which is not liable to tarnish, the sulphuric and nitric acid does not work. . . . Through this treatment with sulphuric or nitric acid some types of glasses get on their surfaces beautiful vivid colours which alter like soap bubbles if one lets the light strike at different angles.

Then, in an appendix to the paper added in 1819:

> Colours on reflection always occur with all transparent media if they are very thin. If for example, one spreads polished glass thinly with alcohol and lets it gradually evaporate towards the end of the evaporation, colours appear as with tarnished glass. If one spreads a solution of gum-lac in a comparatively large quantity of alcohol very thinly over polished warmed metal the alcohol will very quickly evaporate, and the gum-lac remains behind as a transparent hard varnish which shows colours if it is thinly enough laid on. Since the colours, in glasses which have been coloured through tarnishing, alter themselves if the inclination of the incident light becomes greater or smaller, there is no doubt that these colours are quite of the same nature as those of soap bubbles, and those which occur through the contact of two polished flat glass surfaces, or generally as thin transparent films of material. Thus there must be on the surface of tarnished glass which shows colours, a thin layer of glass which is different in refractive power from the underlying. Such a situation must occur if a component is partly removed from the surface of the glass or if a component of the glass combines at the surface with a related material into a new transparent product.

It seems that Fraunhofer did not follow up this particular line into the development of an antireflection coating for glass, perhaps because optical components were not, at that time, sufficiently complicated for the need for antireflection coatings to be obvious. Possibly the important point that not only was the reflectance less but the transmittance also greater had escaped him.

In 1886, Lord Rayleigh reported to the Royal Society an experimental verification of Fresnel's reflection law at near-normal incidence[15]. In order to attain a sufficiently satisfactory agreement between measurement and prediction, he had found it necessary to use freshly polished glass because the reflectance of older material, even without any visible signs of tarnish, was too low. One possible explanation which he suggested was the formation, on the surface, of a thin layer of different refractive index from the underlying material. He was apparently unaware of the earlier work of Fraunhofer.

Then, in 1891, Dennis Taylor published the first edition of his famous book *On the Adjustment and Testing of Telescopic Objectives* and mentioned that[16]

> As regards the tarnish which we have above alluded to as being noticeable upon the flint lens of an ordinary objective after a few years of use, we are very glad to be able to reassure the owner of such a flint that this film of tarnish, generally looked upon with suspicion, is really a very good friend to the observer, inasmuch as it increases the transparency of his objective.

In fact, Taylor went on to develop a method of artificially producing the tarnish by chemical etching[17]. This work was followed up by Kollmorgen, who developed the chemical process still further for different types of glasses[18].

At the same time, in the nineteenth century, a great deal of progress was being made in the field of interferometry. The most significant development, from the thin-film point of view, was the Fabry–Perot interferometer[19], described in 1899, which has become one of the basic types of structure for thin-film filters.

Developments became much more rapid in the 1930s, and indeed it is in this period that we can recognise the beginnings of modern thin-film optical coating. In 1932, Rouard[20] observed that a very thin metallic film reduced the internal reflectance of a glass plate, although the external reflectance was increased. In 1934, Bauer[21], in the course of fundamental investigations of the optical properties of halides, produced reflection-reducing coatings, and Pfund[22] evaporated zinc sulphide layers to make low-loss beam splitters for Michelson interferometers, noting, incidentally, that titanium dioxide could be a better material. In 1936, John Strong[23] produced antireflection coatings by evaporation of fluorite to give inhomogeneous films which reduced the reflectance of glass to visible light by as much as 89 %, a most impressive figure. Then, in 1939, Geffken[24] constructed the first thin-film metal–dielectric interference filters.

The most important factor in this sudden expansion of thin-film optical coatings was the manufacturing process. Although sputtering was discovered about the middle of the nineteenth century, and vacuum evaporation around the beginning of the twentieth, they were not considered as useful manufacturing processes. The main difficulty was the lack of really suitable pumps, and it was not until the early 1930s that the work of C R Burch on diffusion pump oils made it possible for this process to be used satisfactorily.

Since then, tremendous strides have been made, particularly in the last few years. Filters with perhaps one hundred layers are not uncommon and uses have been found for them in almost every branch of science and technology.

THIN-FILM FILTERS

To understand in a qualitative way the performance of thin-film optical devices, it is necessary to accept several simple statements. The first is that the

amplitude reflectance of light at any boundary between two media is given by $(1-\rho)/(1+\rho)$, where ρ is the ratio of the refractive indices at the boundary (the intensity reflectance is the square of this quantity). The second is that there is a phase shift of 180° when the reflectance takes place in a medium of lower refractive index than the adjoining medium and zero if the medium has a higher index than the one adjoining it. The third is that if light is split into two components by reflection at the top and bottom surfaces of a thin film, then the beams will recombine in such a way that the resultant amplitude will be the difference of the amplitudes of the two components if the relative phase shift is 180°, or the sum of the amplitudes if the relative phase shift is either zero or a multiple of 360°. In the former case, we say that the beams interfere destructively and in the latter constructively. Other cases where the phase shift is different will be intermediate between these two possibilities.

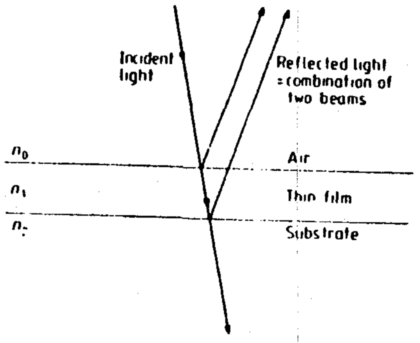


Figure 1.1 A single thin film.

The antireflection coating depends for its operation on the more or less complete cancellation of the light reflected at the upper and lower of the two surfaces of the thin film. Let the index of the substrate be n_2, that of the film n_1, and that of the incident medium, which will in almost all cases be air, n_0. For complete cancellation of the two beams of light, the intensities of the light reflected at the upper and lower boundaries of the film should be equal, which implies that the ratios of the refractive indices at each boundary should be equal, i.e. $n_0/n_1 = n_1/n_2$ or $n_1 = (n_0 n_2)^{1/2}$, which shows that the index of the thin film should be intermediate between the indices of air, which may be taken as unity, and of the substrate, which may be taken as at least 1.52.

Part of the incident light will be reflected at the top and bottom surfaces of the antireflection film, and in both cases the reflection will take place in a medium of lower refractive index than the adjoining medium. Thus, to ensure that the relative phase shift is 180°, the optical thickness of the film should be made one quarter wavelength when the total difference in phase between the two beams will correspond to twice one quarter wavelength, that is 180°.

A simple antireflection coating should, therefore, consist of a single film of refractive index equal to the square root of that of the substrate, and of optical thickness one quarter of a wavelength. As will be explained in the chapter on antireflection coatings, there are other improved coatings covering wider wavelength ranges involving greater numbers of layers.

Another basic type of thin-film structure is a stack of alternate high- and low-index films, all one quarter wavelength thick (see figure 1.2). Light which is reflected within the high-index layers will not suffer any phase shift on reflection, while those beams reflected within the low-index layers will suffer a change of 180°. It is fairly easy to see that the various components of the incident light produced by reflection at successive boundaries throughout the assembly will reappear at the front surface all in phase so that they will recombine constructively. This implies that the effective reflectance of the assembly can be made very high indeed, as high as may be desired, merely by increasing the number of layers. This is the basic form of the high-reflectance coating. When such a coating is constructed, it is found that the reflectance remains high over only a limited range of wavelengths, depending on the ratio of high and low refractive indices. Outside this zone, the reflectance changes abruptly to a low value. Because of this behaviour, the quarter-wave stack is used as a basic building block for many types of thin-film filters. It can be used as a longwave-pass filter, a shortwave-pass filter, a straightforward high-reflectance coating, for example in laser mirrors, and as a reflector in a thin-film Fabry–Perot interferometer (figure 1.3), which is another basic filter type described in some detail in chapters 5 and 7. Here, it is sufficient to say that it consists of a spacer layer which is usually half a wavelength thick, bounded by two high-reflectance coatings. Multiple-beam interference in the spacer layer causes the transmission of the filter to be extremely high over a narrow band of wavelengths around that for which the spacer is a multiple of one half wavelength thick. It is possible, as with lumped electric circuits, to couple two or more Fabry–Perot filters in series to give a more rectangular pass band.

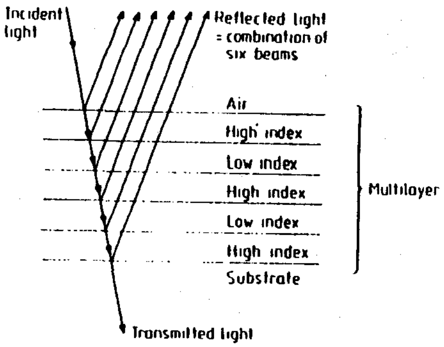


Figure 1.2 A multilayer.

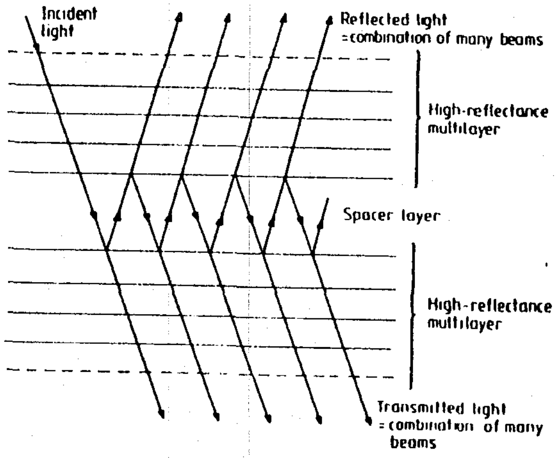

Incident light
Reflected light =combination of many beams
High-reflectance multilayer
Spacer layer
High-reflectance multilayer
Transmitted light = combination of many beams

Figure 1.3 A Fabry–Perot filter showing multiple reflections in the spacer layer.

In the great majority of cases the thin films are completely transparent, so that no energy is absorbed. The filter characteristic in reflection is the complement of that in transmission. This fact is used in the construction of such devices as dichroic beam splitters for colour primary separation in, for example, colour television cameras.

This brief description has neglected the effect of multiple reflections in most of the layers and, for an accurate evaluation of the performance of a filter, these extra reflections must be taken into account. This involves extremely complex calculations and an alternative, and more effective, approach has been found in the development of entirely new forms of solution of Maxwell's equations in stratified media. This is, in fact, the principal method used in chapter 2 where basic mathematics are considered. The solution appears as a very elegant product of 2×2 matrices, each matrix representing a single film. Unfortunately, in spite of the apparent simplicity of the matrices, calculation by hand of the properties of a given multilayer, particularly if there are absorbing layers present and a wide spectral region is involved, is an extremely tedious and time-consuming task. The advent of the electronic pocket calculator has greatly reduced the necessary labour, and programmable versions can be used for calculations of multilayers, provided the number of layers and of wavelength or frequency points is limited. The preferred method, when layers are numerous and especially absorbing or when the calculation is over a wide spectral region, is to use a computer. A particularly significant advantage of the matrix method is that it has made possible the development of exceedingly powerful design techniques based on the algebraic manipulation of the matrices themselves.

In the design of a thin-film multilayer, we are required to find an arrangement of layers which will give a performance specified in advance, and this is much more difficult than straightforward calculation of the properties of a given multilayer. There is no analytical solution to the general problem. The normal method of design is to arrive at a possible structure for a filter, using techniques which will be described, and which consist of a mixture of analysis, experience and the use of well known building blocks. The evaluation is then completed by calculating the performance on a computer. Depending on the results of the computations, adjustments to the proposed design may be made, then recomputed, until a satisfactory solution is found. This adjustment process can itself be undertaken by a computer and is often known by the term refinement.

The successful application of refinement techniques depends largely on a starting solution which has a performance close to that required. Under these conditions it has been made to work well. Methods of completely automatic synthesis of designs without any starting solution are less frequently used, although great progress has been made. In common with refinement techniques, they operate to adjust the parameters of the system to minimise a merit coefficient representing the gap between the performance achieved by the design at any stage and the desired performance. A major problem is the enormous number of parameters which can potentially be involved. Refinement is kept within bounds by limiting the search to small changes in an almost acceptable starting design, but with no starting design the possibilities are virtually infinite, and so the rules governing the search procedure have to be very carefully organised. Automatic design synthesis is undoubtedly increasing in importance with developments in computers, but this branch of the subject is much more a matter of computing techniques rather than fundamental to the understanding of thin-film filters, and so it has been considered outside the scope of this book. The recent book by Liddell[9] gives a full account of the various methods. The real limitation to what is, at the present time, possible in optical thin-film filters and coatings, is the capability of the manufacturing process to produce layers of precisely the correct optical constants and thickness, rather than any deficiency in design techniques.

The basic manufacturing method for the construction of thin-film filters is that of vacuum deposition and this is the principal process described in this book. Further information on the vacuum deposition process is given in the book by Holland[2], now over twenty years old and unfortunately out of print but still very useful. Then, while this book was in the proof stage, the exceptionally detailed book on the physics and chemistry involved in thin-film coatings on glass by Pulker[25] appeared. It also contains much information on cleaning techniques, alternative deposition methods, and testing methods.

Optical filters are used in almost every conceivable field. To cover the whole range of applications would clearly be impossible, so several typical uses have been selected and described along with some of the important points to watch in designing such applications.

REFERENCES

1 Knittl Z 1976 *Optics of Thin Films* (London: Wiley)

2 Holland L 1956 *Vacuum Deposition of Thin Films* (London: Chapman and Hall)

3 Anders H 1965 *Dünne Schichten für die Optik* (Stuttgart: Wissenschaftliche Verlagsgesellschaft) (Engl. transl. 1967 *Thin films in optics* (London: Focal))

4 Heavens O S 1955 *Optical Properties of Thin Solid Films* (London: Butterworths). Reprinted 1965 (New York: Dover)

5 Lissberger P H 1970 Optical applications of dielectric thin films *Rep. Prog. Phys.* **33** 197–268

6 DeBell G W and Harrison D H eds 1974 Optical Coatings, Applications and Utilization *Proc. Soc. Photo-optical Instrum. Eng.* **50**

7 DeBell G W and Harrison D H eds 1978 Optical Coatings II (Applications and Utilization) *Proc. Soc. Photo-optical Instrum. Eng.* **140**

8 Dobrowolski J A 1978 *Coatings and Filters* in *Handbook of Optics* ed. Driscoll W G (New York: McGraw-Hill) pp. 8-1–8-124

9 Liddell H M 1981 *Computer-aided Techniques for the Design of Multilayer Filters* (Bristol: Adam Hilger)

10 Crew H 1930 Thomas Young's place in the history of the wave theory of light *J. Opt. Soc. Am.* **20** 3–10

11 de Senarmont H, Verdet E and Fresnel L 1866–70 *Oeuvres complètes d'Augustin Fresnel* (Paris: Impériale)

12 Knittl Z 1978 Fresnel historique et actuel *Opt. Acta* **25** 167–73

13 Maxwell J C 1873 *A Treatise on Electricity and Magnetism*. First edition published in 1873. The third edition, originally published by the Clarendon Press in 1891, was republished in unabridged form in 1954 (New York: Dover)

14 von Fraunhofer J 1817 *Versuche über die Ursachen des Anlaufens und Mattwerdens des Glases und die Mittel denselben zuvorzukommen.* Taken from *Joseph von Fraunhofer's Gesammelte Schriften* 1888 (Munich). The extracts appear on pages 35 and 46

15 Lord Rayleigh 1886 On the intensity of light reflected from certain surfaces at nearly perpendicular incidence *Proc. R. Soc.* **41** 275–94

16 Taylor H D 1983 *On the adjustment and testing of Telescopic Objectives.* First published in 1891. The quotation is taken from the second edition, p 62, published by T Cooke & Sons in 1896. The third edition, 1921, was later republished unchanged by Sir Howard Grubb, Parsons & Company Limited in 1946. The quotation may now be found on p 59 of the fifth edition (1983) (Bristol: Adam Hilger)

17 Taylor H D 1904 Lenses *UK Patent Specification* 29 561

18 Kollmorgen F 1916 Light transmission through telescopes *Trans. Am. Illum. Eng. Soc.* **11** 220–8

19 Fabry C and Perot A 1899 Théorie et applications d'une nouvelle méthode de spectroscopie interférentielle *Ann. Chim. Phys., Paris* 7th series **16** 115–44

20 Rouard P 1932 Sur le pouvoir réflecteur des métaux en lames très minces *C. R. Acad. Sci., Paris* **195** 868–71

21 Bauer G 1934 Absolutwerte der optischen Absorptionskonstanten von Alkalihalogenidkristallen im Gebiet ihrer ultravioletten Eigenfrequenzen *Ann. Phys. Lpz.* 5th series **19** 434–64

22 Pfund A H 1934 Highly reflecting films of zinc sulphide *J. Opt. Soc. Am.* **24** 99–102

23 Strong J 1936 On a method of decreasing the reflection from non-metallic substances *J. Opt. Soc. Am.* **26** 73–4

24 Geffken W 1939 Interferenzlichtfilter *Deutsches Reich Patentschrift* 716 153

25 Pulker H K 1984 *Coatings on Glass* (Oxford: Elsevier)

2 Basic theory

The next part of the book is a long and rather tedious account of some basic theory which is necessary in order to make calculations of the properties of multilayer thin-film coatings. It is perhaps worth reading just once, or when some deeper insight into thin-film calculations is required. In order to make it easier for those who have read it to find the basic results, or, for those who do not wish to read it at all, to proceed with the remainder of the book, the principal results are summarised, beginning on page 40.

MAXWELL'S EQUATIONS AND PLANE ELECTROMAGNETIC WAVES

For those readers who are still with us we begin our attack on thin-film problems by solving Maxwell's equations together with the appropriate material equations. In isotropic media these are

$$\operatorname{curl} \boldsymbol{H} = \boldsymbol{j} + \partial \boldsymbol{D}/\partial t \tag{2.1}$$

$$\operatorname{curl} \boldsymbol{E} = -\partial \boldsymbol{B}/\partial t \tag{2.2}$$

$$\operatorname{div} \boldsymbol{D} = \rho \tag{2.3}$$

$$\operatorname{div} \boldsymbol{B} = 0 \tag{2.4}$$

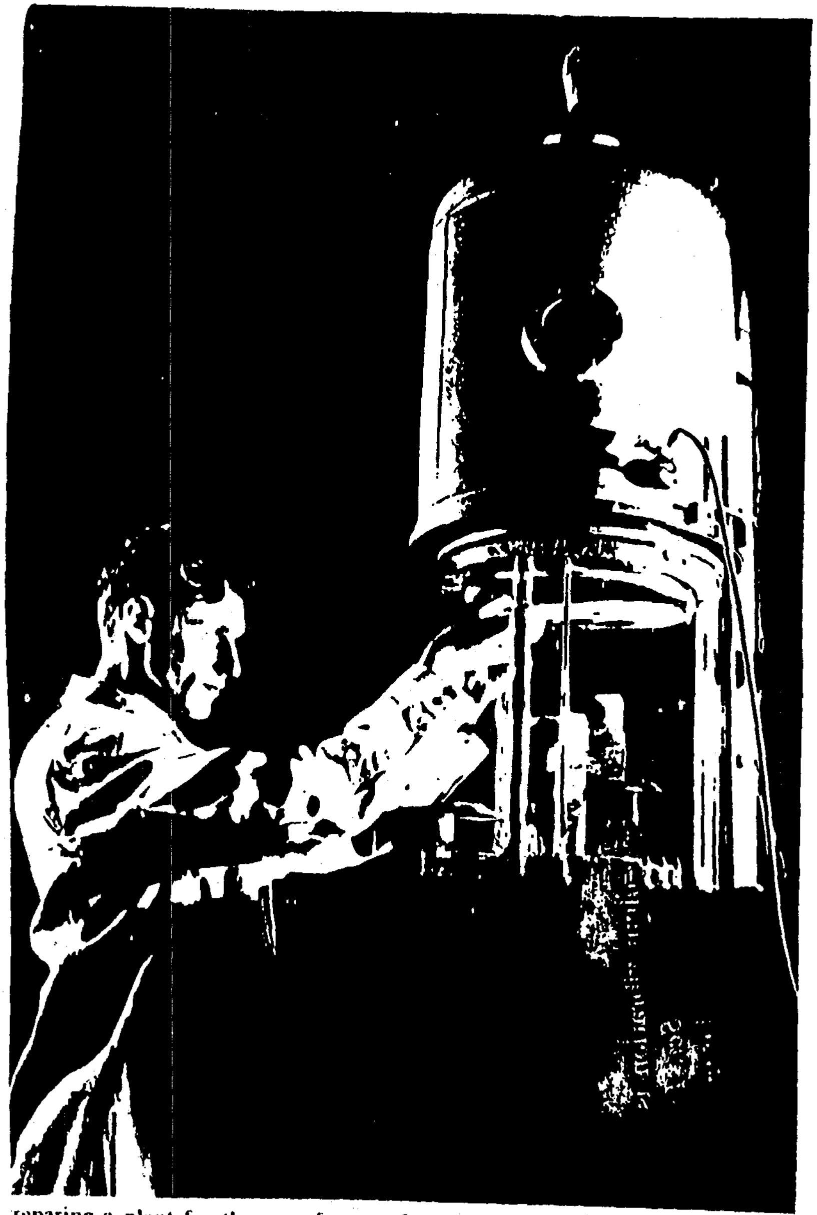
$$\boldsymbol{j} = \sigma \boldsymbol{E} \tag{2.5}$$

$$\boldsymbol{D} = \varepsilon \boldsymbol{E} \tag{2.6}$$

$$\boldsymbol{B} = \mu \boldsymbol{H}. \tag{2.7}$$

In anisotropic media, equations (2.5)–(2.7) become much more complicated with σ, ε and μ being tensor rather than scalar quantities.

The International System of Units (SI) is used as far as possible throughout this book. Table 2.1 shows the definitions of the quantities in the equations together with the appropriate SI units.



reparing a plant for the manufacture of narrowband filters. (Courtesy of Walter urnberg FIEP FRPS, the editors of *Engineering*, and Sir Howard Grubb, Parsons & o Ltd.)

Thin-film optical filters

Second edition

H A Macleod

Professor of Optical Sciences
University of Arizona

Adam Hilger Ltd, Bristol

Ph 103c: The Physics of LIGO 29 April 1994

LECTURE 10.

Control Systems for Test-Mass Position and Orientation

Lecture by Seiji Kawamura

Assigned Reading:

BB. S. Kawamura and M. E. Zucker, *Applied Optics*, in press. [This paper explains the influence of angular mirror orientation errors on the length of a Fabry-Perot resonator.]

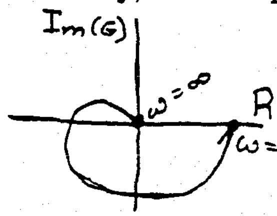
Read either item CC. below or item DD. [Item DD. is highly recommended, since feedback loops are so important; but for some students it may entail a fair amount of work, and CC. might be preferred.]

CC. M. Stephens, P. Saulson, and J. Kovalik, "A double pendulum vibration isolation system for a laser interferometric gravitational wave antenna," *Rev. Sci. Instrum.*, **62**, 924–932 (1991). [Here you are asked to focus on the control of the pendulum, rather than on the penedulum's role in vibration isolation.]

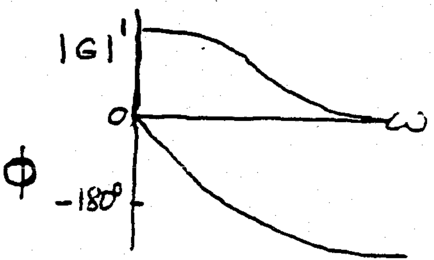
DD. Read, in your favorite control theory book [e.g., R. C. Dorf, *Modern Control Systems* 5th editon (Addison-Wesley, 1989), cited as *Dorf* below] or elsewhere, about the following issues:

a. The relationship of Laplace transforms to Fourier transforms [e.g., *Dorf* pp. 264–266]. Control theory is often formulated in terms of Lapace transforms rather than Fourier transforms because Lapace transforms are more naturally suited to describing the transient response of a system to some input; the reason is that they entail only the behavior of the system between some initial time $t = 0$ and $t = \infty$, by contrast with Fourier transforms which involve the behavior over all time. In this course we will probably *not* deal with any issues where the Laplace transform has an advantage; and we will most always discuss things in terms of Fourier transforms and thus in terms of the response of a system at some frequency ω. However, in order to read control theory books on these issues, it is necessary to understand Laplace transforms and their relation to Fourier transforms. [Note that, although theoretical physicists normally use the form $e^{-i\omega t}$ for the time dependence of a Fourier component of frequency ω, engineers, and control theorists normally use $e^{+j\omega t}$ (where $i = j = \sqrt{-1}$). In this course we shall use the engineers' conventions.]

b. The use of complex frequency-response plots to describe the ratio of the output amplitude $V_{\rm out}$ of a linear system such as a control loop, to its input amplitude $V_{\rm in}$, when the input and output have frequency ω [e.g., read *Dorf*, pp. 266–283]. In these plots, $V_{\rm out}/V_{\rm in} \equiv G(\omega)$, which is a complex quantity, is plotted as a curve in the complex plane parametrized by ω, for real ω. Such a plot contains the same information as a Bode diagram, in which one gives two plots, one of $|G(\omega)|$ plotted upward and ω horizontally; the other of the phase $\phi(\omega)$ of G plotted upward and ω horizontally; for example:



Frequency - Response Plot



Bode Di

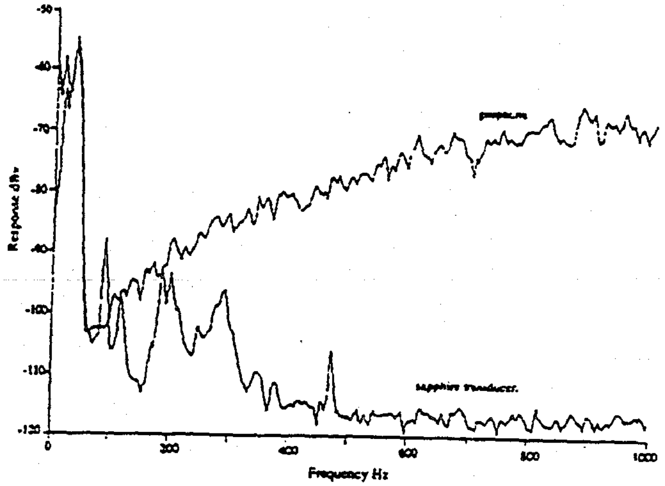


Figure 7. Comparison of the performance of the geophone and the sapphire transducer. The rise of the geophone response at high frequencies is due to electrical pick-up from the high-power drive signal.

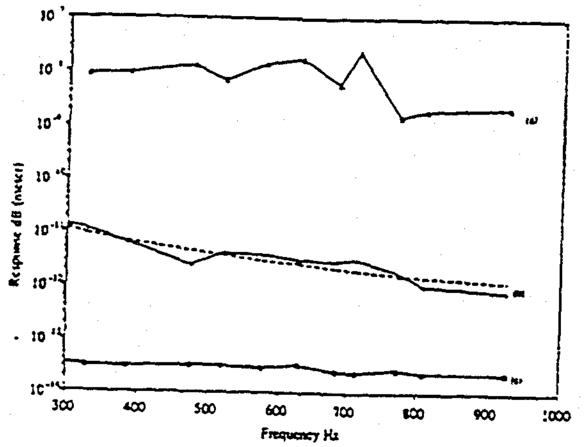


Figure 8. Result of measuring noise floor of the isolator. (a) The measured applied excitation at geophone 1. (b) The normal level of seismic noise compared with the theoretical curve $10^{-6} f^{-2}$ (broken line). (c) The observed noise floor of the sapphire transducer.

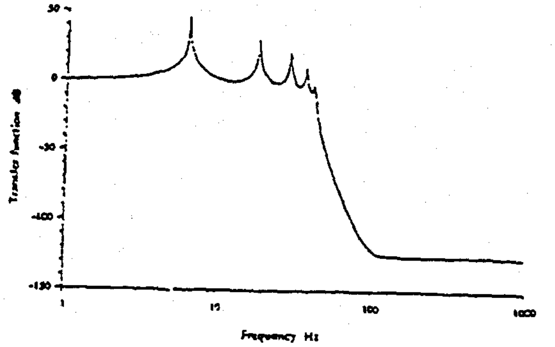


Figure 9. Theoretical transfer function of a five-stage isolator at room temperature, limited to 125 dB attenuation above 100 Hz due to thermal noise.

900 Hz. This is set by the available drive amplitude which is reduced at 900 Hz. Theoretically the attenuation should be 125 dB above 200 Hz, limited by thermal noise as shown in figure 9. Therefore, so far we have no evidence that the actual performance of the isolator is less than its theoretical value.

4.3. Active damping of the low-frequency normal modes

To damp the low-frequency normal modes, we applied active feedback in the vertical direction to the first element of a four-stage isolator. We successfully attenuated the third and fourth response peaks by 15 dB by feeding back the broad band signal from the first element to a mass loaded loudspeaker on the same element as shown in figure 4. The response curves with and without feedback are shown in figure 10(*a*). Here, the velocity output from the geophone has been converted to an amplitude response. For comparison, the response curves of the computer model with damping on first element are shown in figure 10(*b*). Clearly there is good agreement between the experimental result and the computer model.

Since the isolator is normally driven by seismic noise, it is the lowest response peak that has the highest amplitude. This mode can be damped by adding a second single-stage narrow band feedback loop tuned to the first normal mode of the isolator. We applied the feedback to the fourth element in the five-stage isolator. With this double feedback loop, we are able to attenuate the first normal mode by 7 dB, the third mode by 4 dB and the fourth and fifth modes by 14 dB, as shown in figure 11. However, it has proved difficult to reduce all normal modes simultaneously, due to interactions

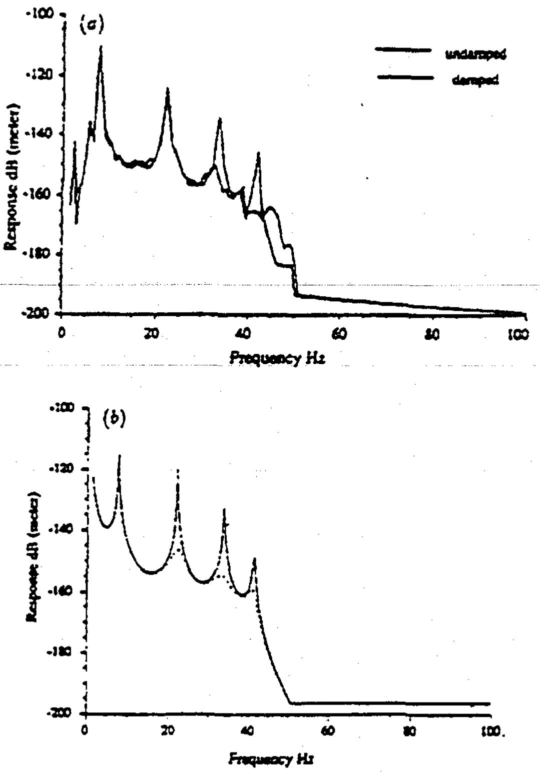


Figure 10. Result of a single-stage feedback applied to the first element of a five-stage isolator. (*a*) Experimental geophone response curve (the result has been converted from the velocity response to amplitude response). (*b*) Theoretical amplitude response curve.

between modes and probably due to the reaction feedback system employed. We are planning to model the full feedback system to define the stability criteria more clearly.

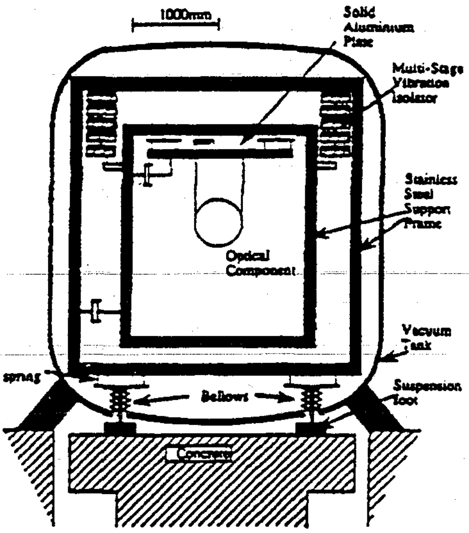


Figure 12. Proposed isolation platform.

5. Proposed full scale structure

As part of the proposed AIGO project [11], we plan to construct a general purpose isolation platform. The isolators described above will be used in a complex isolator shown in figure 12.

The whole system is supported by massive concrete foundations. A pair of sand-filled stainless steel (vacuum tight) box girder frames provide the first and eighth stages of isolation. The first is supported on springs, the eighth is hung by four 1-metre pendulums, each incorporating the isolator designed here. From the eighth element a solid aluminium plate is hung. This provides the suspension base for conventional pendulum supports for the optical components. The isolator is therefore

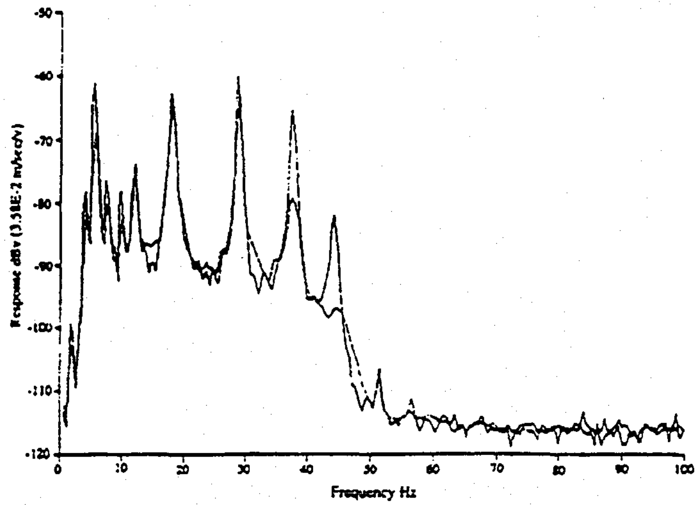


Figure 11. Response curves of a five-stage isolator using a double feedback loop to damp the low-frequency normal modes. First stage, broad band feedback (bandwidth 0–100 Hz): fourth stage, 6.4 Hz narrow band feedback.

designed to provide a general purpose ultralow vibration environment for frequencies above 50 Hz. An important part of this system is the use of active damping at the points shown schematically in figure 12. This will allow the use of a suspension point servo to reduce the amplitude of the seismic noise peak at 0.2 Hz. The full scale system will be modelled using three-dimensional finite element analysis.

6. Conclusion

The measurements reported here cover a dynamic range from 10^{-6} to 10^{-14} m. The one-dimensional computer model agrees with the experiments in vertical direction quite well. The high-frequency tests using a sapphire transducer show that this isolator works as expected even in the presence of large excitation. In addition, there is no indication of any nonlinear upconversion at the amplitudes of excitation achieved here. Active feedback using reaction force drive works well on higher normal modes but higher performance would be desirable for the lower modes.

The horizontal vibration isolation of this isolator has not been well investigated. From figure 6 we can see that there exists a low level of cross coupling between orthogonal directions. In the next stage of this work, we will use a horizontally mounted sapphire transducer to test the high-frequency performance in the horizontal direction.

It is possible to trade off the maximum load on the isolator against its corner frequency. It is reasonably easy to see in this way how to construct an isolator of this type with a corner frequency as low as 10 Hz vertically and 3 Hz horizontally.

Acknowledgments

We thank the Workshop of the Department of Physics, especially Gary Light, for their willing help in the construction of the isolator and its test system. Other members of the gravitational wave group were very helpful both in discussion and heavy labour work. This work was supported by the Australia Research Council.

References

[1] Weber J 1969 Evidence for discovery of gravitational radiation *Phys. Rev. Lett.* 22 1320

[2] Del Fabbro R, Di Virgilio A, Giazotto A, Kautzky H, Montelatici V and Passuello D 1988 Performance of a gas spring harmonic oscillator *Rev. Sci. Instrum.* 59 293

[3] Veitch P J 1991 Isolation of distributed mechanical oscillators by mechanical suspensions with application to resonant-mass gravitational radiation antennae *Rev. Sci. Instrum.* 62 140

[4] Stephens M, Saulson P and Kovalik J 1991 A double pendulum vibration isolation system for a laser interferometric gravitational wave antenna *Rev. Sci. Instrum.* 62 924

[5] Del Fabbro R, Di Virgilio A, Giazotto A, Kautzky H, Montelatici V and Passuello D 1987 Three-dimensional seismic super-attenuator for low frequency gravitational wave detection *Phys. Lett.* A 124 253

[6] Linthorne N P, Veitch P J and Blair D G 1990 Interaction of a parametric transducer with a resonant bar gravitational radiation detector *J. Phys. D: Appl. Phys.* 23 1

[7] Blair D G, Linthorne N P, Mann A G, Peng H, Sebo K M, Tobar M E and Turner P J 1991 Progress in optimizing a high *Q* resonant bar antenna *Gravitational Astronomy: Instrument Design and Astrophysical Prospects* ed D E McClell and H-A Bachor (Singapore: World Scientific)

[8] Blair D G (ed) 1991 *The Detection of Gravitational waves* part III (Cambridge: Cambridge University Press)

[9] Blair D G and Peng H 1991 Sapphire microwave transducer for ultra-high sensitive displacement measurement *Gravitational Astronomy: Instrument Design and Astrophysical Prospects* ed D E McClelland and H-A Bachor (Singapore: World Scientific)

[10] Blair D G, Peng H and Ivanov E N 1991 Theory and application of the sapphire dielectric resonator transducer *The Sixth Marcel Grossmann Conf. on General Relativity (Kyoto 1991)* Singapore: World Scientific)

[11] Sandeman R J and McClelland D E 1990 Laser interferometer gravitational-wave observatories: an overview *J. Mod. Opt.* 37 1747

PHYSICAL REVIEW VOLUME 83, NUMBER 1 JULY 1, 1951

Irreversibility and Generalized Noise*

HERBERT B. CALLEN AND THEODORE A. WELTON†
Randal Morgan Laboratory of Physics, University of Pennsylvania, Philadelphia, Pennsylvania

(Received January 11, 1951)

A relation is obtained between the generalized resistance and the fluctuations of the generalized forces in linear dissipative systems. This relation forms the extension of the Nyquist relation for the voltage fluctuations in electrical impedances. The general formalism is illustrated by applications to several particular types of systems, including Brownian motion, electric field fluctuations in the vacuum, and pressure fluctuations in a gas.

I. INTRODUCTION

THE parameters which characterize a thermodynamic system in equilibrium do not generally have precise values, but undergo spontaneous fluctuations. These thermodynamic parameters are of two classes: the "extensive" parameters,[1] such as the volume or the mole numbers, and the "intensive" parameters[2] or "generalized forces," such as the pressure or chemical potentials.

An equation relating particularly to the fluctuations in voltage (a "generalized force") in linear electrical systems was derived many years ago by Nyquist,[3] and such voltage fluctuations are generally referred to as Nyquist or Johnson "noise." The voltage fluctuations are related, not to the standard thermodynamic parameters of the system, but to the electrical resistance. The Nyquist relation is thus of a form unique in physics, correlating a property of a system in *equilibrium* (i.e., the voltage fluctuations) with a parameter which characterizes an irreversible process (i.e., the electrical resistance). The equation, furthermore, gives not only the mean square fluctuating voltage, but provides, in addition, the frequency spectrum of the fluctuations. The proof of the relation is based on an ingenious union of the second law of thermodynamics and a direct calculation of the fluctuations in a particular simple system (an ideal transmission line).

It has frequently been conjectured that the Nyquist relation can be extended to a general class of dissipative systems other than merely electrical systems. Yet, to our knowledge, no proof has been given of such a generalization, nor have any criteria been developed for the type of system or the character of the "forces" to which the generalized Nyquist relation may be applied. The development of such a proof and of such criteria is the purpose of this paper (Secs. II, III, and IV). The general theorem thus establishes a relation between the "impedance" in a general linear dissipative system and the fluctuations of appropriate generalized "forces."

Several illustrative applications are made of the general theorem. The viscous drag of a fluid on a moving body is shown to imply a fluctuating force, and application of the general theorem immediately yields the fundamental result of the theory of Brownian motion. The existence of a radiation impedance for the electromagnetic radiation from an oscillating charge is shown to imply a fluctuating electric field in the vacuum, and application of the general theorem yields the Planck radiation law. Finally, the existence of an acoustic radiation impedance of a gaseous medium is shown to imply pressure fluctuations, which may be related to the thermodynamic properties of the gas. The theorem thus correlates a number of known effects under one general principle and is able to predict a class of new relations.

In the final section of the paper, we discuss an intuitive interpretation of the principles underlying the theorem.

It is felt that the relationship between equilibrium fluctuations and irreversibility which is here developed provides a method for a general approach to a theory of irreversibility, using statistical ensemble methods. We are currently investigating such an approach.

II. THE DISSIPATION

A system may be said to be dissipative if it is capable of absorbing energy when subjected to a time-periodic perturbation (as an electrical resistor absorbs energy from an impressed periodic voltage). The system may be said to be linear if the power dissipation is quadratic in the magnitude of the perturbation. For a linear dissipative system, an impedance may be defined, and the proportionality constant between the power and the square of the perturbation amplitude is simply related to the impedance [in the electrical case, Power $=(\text{voltage})^2\cdot R/|Z|^2$].

In the present section we treat the applied perturbation by the usual quantum mechanical perturbation

* This work was supported in part by the ONR.

† Now at Oak Ridge National Laboratory, Oak Ridge, Tennessee.

[1] For the theory of fluctuations of extensive parameters see Fowler, *Statistical Mechanics* (Cambridge University Press, London, 1936), second edition; or Tolman, *Principles of Statistical Mechanics* (Oxford University Press, London, 1938). A recent development of the theory is given by M. J. Klein and L. Tisza, Phys. Rev. 76, 1861 (1949).

[2] A statistical mechanical theory of fluctuations of intensive parameters will be given in a subsequent paper by R. F. Greene and H. B. Callen.

[3] H. Nyquist, Phys. Rev 32, 110 (1928). A very neat derivation and an interesting discussion is given by J. C. Slater, Radiation Laboratory Report; "Report on Noise and the Reception of Pulses," February 3, 1941, unpublished.

methods and thus relate the power dissipation to certain matrix elements of the perturbation operator. We thereby show that for small perturbations, a system with densely distributed energy levels is dissipative and linear, and we obtain certain pertinent information relative to the impedance function.

Let the hamiltonian of the system in the absence of the perturbation be H_0, a function of the coordinates $q_1\cdots q_K\cdots$ and momenta $p_1\cdots p_K\cdots$ of the system. In the presence of the perturbation, the hamiltonian is

$$H=H_0(\cdots q_K\cdots p_K\cdots)+VQ(\cdots q_K\cdots p_K\cdots). \quad (2.1)$$

where Q is a function of the coordinates and momenta, and V is a function of time which measures the instantaneous magnitude of the perturbation.

Again invoking the electrical case as a clarifying example, we may have V as the impressed voltage and $Q=\sum e_i x_i/L$, where e_i is the charge on the ith particle, x_i is its distance from one end of the resistor, and L is the total length of the resistor.

If the applied perturbation varies sinusoidally with time, we have

$$V=V_0\sin\omega t. \quad (2.2)$$

We may now employ standard time-dependent perturbation theory to compute the power dissipation. Let $\psi_1, \psi_2\cdots\psi_n\cdots$ be the set of eigenfunctions of the unperturbed hamiltonian H_0, so that

$$H_0\psi_n=E_n\psi_n, \quad (2.3)$$

and let the true wave function be ψ. Expanding ψ in terms of the ψ_n,

$$\psi=\sum_n a_n(t)\psi_n, \quad (2.4)$$

and substituting into the Schroedinger equation for ψ,

$$H_0\psi+V_0\sin\omega t Q\psi=i\hbar\partial\psi/\partial t, \quad (2.5)$$

one obtains a set of first-order equations for the coefficients $a_n(t)$, which may readily be integrated. If the energy levels of the system are densely distributed, one thus finds that the total induced transition probability of a system initially in the state ψ_n is

$$\tfrac{1}{2}\pi V_0^2\hbar^{-1}\{|\langle E_n+\hbar\omega|Q|E_n\rangle|^2\rho(E_n+\hbar\omega) + |\langle E_n-\hbar\omega|Q|E_n\rangle|^2\rho(E_n-\hbar\omega)\}, \quad (2.6)$$

where the symbol $\langle E_n+\hbar\omega|Q|E_n\rangle$ indicates the matrix element of the operator corresponding to Q between the state with eigenvalue $E_n+\hbar\omega$ and the state with eigenvalue E_n. The symbol $\rho(E)$ indicates the density-in-energy of the quantum states in the neighborhood of E, so that the number of states between E and $E+\delta E$ is $\rho(E)\delta E$.

Each transition from the state ψ_n to the state with eigenvalue $E_n+\hbar\omega$ is accompanied by the absorption of energy $\hbar\omega$, and each transition from ψ_n to the state with eigenvalue $E_n-\hbar\omega$ is accompanied by the emission of energy $\hbar\omega$. Thus the rate of absorption of energy by a system initially in the nth state is

$$\tfrac{1}{2}\pi V_0^2\omega\{|\langle E_n+\hbar\omega|Q|E_n\rangle|^2\rho(E_n+\hbar\omega) - |\langle E_n-\hbar\omega|Q|E_n\rangle|^2\rho(E_n-\hbar\omega)\}. \quad (2.7)$$

To predict the behavior of a real thermodynamic system, we must average over-all initial states, weighting each according to the Boltzmann factor $\exp(-E_n/kT)$. Let the weighting factor be $f(E_n)$, so that

$$f(E_n+\hbar\omega)/f(E_n)=f(E_n)/f(E_n-\hbar\omega) =\exp(-\hbar\omega/kT). \quad (2.8)$$

The power dissipation is, then,

$$\text{Power}=\tfrac{1}{2}\pi V_0^2\omega\sum_n\{|\langle E_n+\hbar\omega|Q|E_n\rangle|^2\rho(E_n+\hbar\omega) - |\langle E_n-\hbar\omega|Q|E_n\rangle|^2\rho(E_n-\hbar\omega)\}f(E_n). \quad (2.9)$$

The summation over n may be replaced by an integration over energy

$$\sum_n(\quad)\rightarrow\int_0^\infty(\quad)\rho(E)dE, \quad (2.10)$$

whence

$$\text{Power}=\tfrac{1}{2}\pi V_0^2\omega\int_0^\infty\rho(E)f(E) \times\{|\langle E+\hbar\omega|Q|E\rangle|^2\rho(E+\hbar\omega) - |\langle E-\hbar\omega|Q|E\rangle|^2\rho(E-\hbar\omega)\}dE. \quad (2.11)$$

We thus find that a small periodic perturbation applied to a system, the eigenstates of which are densely distributed in energy, leads to a power dissipation quadratic in the perturbation. For such a linear system it is possible to define an impedance $Z(\omega)$, the ratio of the force V to the response $\dot{Q}$, where all quantities are now assumed to be written in standard complex notation,

$$V=Z(\omega)\dot{Q}. \quad (2.12)$$

The instantaneous power is $V\dot{Q}R/|Z|$, and the average power is

$$\text{Power}=\tfrac{1}{2}V_0^2R(\omega)/|Z(\omega)|^2, \quad (2.13)$$

where $R(\omega)$, the resistance, is the real part of $Z(\omega)$.

If the applied perturbation is not sinusoidal, but some general function of time $V(t)$, and if $v(\omega)$ and $\dot{q}(\omega)$ are the fourier transforms of $V(t)$ and $\dot{Q}(t)$, the impedance is defined in terms of the fourier transforms:

$$v(\omega)=Z(\omega)\dot{q}(\omega). \quad (2.14)$$

In this notation we then obtain, for the general linear dissipative system,

$$R/|Z|^2=\pi\omega\int_0^\infty\rho(E)f(E)\{|\langle E+\hbar\omega|Q|E\rangle|^2\rho(E+\hbar\omega) - |\langle E-\hbar\omega|Q|E\rangle|^2\rho(E-\hbar\omega)\}dE. \quad (2.15)$$

III. THE FLUCTUATION

We have, in the previous section, considered a system to which is applied a force V, eliciting a response $\dot{Q}$. We now consider the system to be left in thermal equilibrium, with no applied force. We may expect, even in this isolated condition, that the system will exhibit a spontaneously fluctuating $\dot{Q}$, which may be associated with a spontaneously fluctuating force. We shall see, in this section, that such a spontaneously fluctuating force does in fact exist, and we shall find its magnitude.

Let $\langle V^2\rangle$ be the mean square value of the spontaneously fluctuating force, and let $\langle\dot{Q}^2\rangle$ be the mean square value of the spontaneously fluctuating $\dot{Q}$. Although we shall be primarily interested in $\langle V^2\rangle$, we shall find it convenient to compute $\langle\dot{Q}^2\rangle$ and to obtain $\langle V^2\rangle$ from Eq. (2.14).

Consider that the system is known to be in the nth eigenstate. The hermitian property of H_0 causes the expectation value of $\dot{Q}$, $\langle E_n|\dot{Q}|E_n\rangle$, to vanish. The mean square fluctuation of $\dot{Q}$ is therefore given by the expectation value of $\dot{Q}^2$ or $\langle E_n|\dot{Q}^2|E_n\rangle$. Then

$$\begin{aligned}\langle E_n|\dot{Q}^2|E_n\rangle &= \textstyle\sum_m\langle E_n|\dot{Q}|E_m\rangle\langle E_m|\dot{Q}|E_n\rangle\\ &= \hbar^{-2}\textstyle\sum_m\langle E_n|H_0Q-QH_0|E_m\rangle\\ &\qquad\times\langle E_m|H_0Q-QH_0|E_n\rangle\\ &= \hbar^{-2}\textstyle\sum_m(E_n-E_m)^2|\langle E_m|Q|E_n\rangle|^2. \end{aligned}\qquad(3.1)$$

Introducing a frequency ω by

$$\hbar\omega=|E_n-E_m|, \qquad (3.2)$$

the summation over m may be replaced by two integrals over ω (one for $E_n<E_m$ and one for $E_n>E_m$):

$$\begin{aligned}\langle E_n|\dot{Q}^2|E_n\rangle &= \hbar^{-2}\int_0^\infty(\hbar\omega)^2|\langle E_n+\hbar\omega|Q|E_n\rangle|^2\\ &\quad\times\rho(E_n+\hbar\omega)\hbar d\omega+\hbar^{-2}\int_0^\infty(\hbar\omega)^2\\ &\quad\times|\langle E_n-\hbar\omega|Q|E_n\rangle|^2\rho(E_n-\hbar\omega)\hbar d\omega.\\ &=\int_0^\infty\hbar\omega^2\{|\langle E_n+\hbar\omega|Q|E_n\rangle|^2\rho(E_n+\hbar\omega)\\ &\quad+|\langle E_n-\hbar\omega|Q|E_n\rangle|^2\rho(E_n-\hbar\omega)\}d\omega. \end{aligned}\qquad(3.3)$$

The fluctuation actually observed in a real thermodynamic system is obtained by multiplying the fluctuation in the nth state by the weighting factor $f(E_n)$, and summing

$$\begin{aligned}\langle\dot{Q}^2\rangle=\sum_n f(E_n)\int_0^\infty\hbar\omega^2\{|\langle E_n+\hbar\omega|Q|E_n\rangle|^2\rho(E_n+\hbar\omega)\\ +|\langle E_n-\hbar\omega|Q|E_n\rangle|^2\rho(E_n-\hbar\omega)\}d\omega.\end{aligned}\qquad(3.4)$$

As in Eq. (2.10), the summation over n may be replaced by an integration over the energy spectrum if we introduce the density factor $\rho(E)$. Thus we finally obtain

$$\begin{aligned}\langle\dot{Q}^2\rangle=\int_0^\infty\hbar\omega^2\bigg[\int_0^\infty\rho(E)f(E)\{|\langle E+\hbar\omega|Q|E\rangle|^2\rho(E+\hbar\omega)\\ +|\langle E-\hbar\omega|Q|E\rangle|^2\rho(E-\hbar\omega)\}dE\bigg]d\omega,\end{aligned}\qquad(3.5)$$

or, utilizing the definition (2.14) of the impedance,

$$\begin{aligned}\langle V^2\rangle=\int_0^\infty|Z|^2\hbar\omega^2\bigg[\int_0^\infty\rho(E)f(E)\\ \times\{|\langle E+\hbar\omega|Q|E\rangle|^2\rho(E+\hbar\omega)\\ +|\langle E-\hbar\omega|Q|E\rangle|^2\rho(E-\hbar\omega)\}dE\bigg]d\omega.\end{aligned}\qquad(3.6)$$

IV. THE GENERALIZED NYQUIST RELATION

In the two previous sections we have computed $R/|Z|^2$ and $\langle V^2\rangle$. These quantities involve the constructs

$$\begin{aligned}\int_0^\infty\rho(E)f(E)\{|\langle E+\hbar\omega|Q|E\rangle|^2\rho(E+\hbar\omega)\\ \pm|\langle E-\hbar\omega|Q|E\rangle|^2\rho(E-\hbar\omega)\}dE,\end{aligned}\qquad(4.1)$$

the negative sign being associated with $R/|Z|^2$ and the positive sign with $\langle V^2\rangle$. We shall now see that the two values of (4.1) are simply related, and thus establish the desired relation between $\langle V^2\rangle$ and $R(\omega)$.

Consider first the value of (4.1) corresponding to the negative sign, which we denote by $C(-)$.

$$\begin{aligned}C(-)=\int_0^\infty f(E)|\langle E+\hbar\omega|Q|E\rangle|^2\rho(E+\hbar\omega)\rho(E)dE\\ -\int_0^\infty f(E)|\langle E-\hbar\omega|Q|E\rangle|^2\rho(E)\rho(E-\hbar\omega)dE.\end{aligned}\qquad(4.2)$$

In the second integral we note that $\langle E-\hbar\omega|Q|E\rangle$ vanishes for $E<\hbar\omega$, and making the transformation $E\rightarrow E+\hbar\omega$ in the integration variable, we obtain

$$\begin{aligned}C(-)=\int_0^\infty|\langle E+\hbar\omega|Q|E\rangle|^2\rho(E+\hbar\omega)\rho(E)f(E)\\ \times\{1-f(E+\hbar\omega)/f(E)\}dE.\end{aligned}\qquad(4.3)$$

By Eq. (2.8) this becomes

$$\begin{aligned}C(-)=\{1-\exp(-\hbar\omega/kT)\}\int_0^\infty|\langle E+\hbar\omega|Q|E\rangle|^2\\ \times\rho(E+\hbar\omega)\rho(E)f(E)dE.\end{aligned}\qquad(4.4)$$

If $C(+)$ denotes the value of (4.1) corresponding to the positive sign, we obtain, in an identical fashion,

$$C(+)=\{1+\exp(-\hbar\omega/kT)\}\int_0^\infty|\langle E+\hbar\omega|Q|E\rangle|^2 \times\rho(E+\hbar\omega)\rho(E)f(E)dE. \quad (4.5)$$

With these alternative expressions for (4.1), we can write, from Eq. (2.15),

$$R(\omega)/|Z(\omega)|^2=\pi\omega\{1-\exp(-\hbar\omega/kT)\} \times\int_0^\infty|\langle E+\hbar\omega|Q|E\rangle|^2\rho(E+\hbar\omega)\rho(E)f(E)dE, \quad (4.6)$$

and from Eq. (3.6),

$$\langle V^2\rangle=\int_0^\infty|Z|^2\hbar\omega^2\{1+\exp(-\hbar\omega/kT)\} \times\int_0^\infty|\langle E+\hbar\omega|Q|E\rangle|^2\rho(E+\hbar\omega)\rho(E)f(E)dEd\omega. \quad (4.7)$$

Comparison of these equations yields directly our fundamental theorem:

$$\langle V^2\rangle=(2/\pi)\int_0^\infty R(\omega)E(\omega,T)d\omega, \quad (4.8)$$

where

$$E(\omega,T)=\tfrac{1}{2}\hbar\omega+\hbar\omega[\exp(\hbar\omega/kT)-1]^{-1}. \quad (4.9)$$

It may be recognized that $E(\omega, T)$ is, formally, the expression for the mean energy at the temperature T of an oscillator of natural frequency ω.

At high temperatures, $E(\omega, T)$ takes its equipartition value

$$E(\omega,T)\simeq kT, \quad (kT\gg\hbar\omega) \quad (4.10)$$

and the generalized Nyquist relation takes its most familiar form

$$\langle V^2\rangle\simeq(2/\pi)kT\int R(\omega)d\omega. \quad (4.11)$$

To reiterate then, a system with a generalized resistance $R(\omega)$ exhibits, in equilibrium, a fluctuating force given by Eq. (4.8) or, at high temperature, by Eq. (4.11).

We shall now consider a few specific applications of this theorem. The application to the electrical case is obvious, the general Eq. (4.8) being identical with the Nyquist relation if the force V is interpreted as the voltage. The content of the general theorem is, however, clarified by considering certain less trivial applications.

V. APPLICATION TO BROWNIAN MOTION

The fundamental result of the theory of the Brownian motion of a small particle immersed in a fluid is that the particle moves in response to a randomly fluctuating force $F(t)$ (with components F_x, F_y, F_z) such that

$$\langle F_x^2\rangle=(2/\pi)kT\eta\int d\omega. \quad (5.1)$$

Here η is a frictional constant, so defined that the viscous drag on a particle moving with velocity v is

$$\text{Frictional force}=-\eta v. \quad (5.2)$$

(If, in particular, the particle is spherical, η is known by Stokes' law as $6\pi\cdot$(viscosity)$\cdot$(radius).)

It is interesting to recall briefly the rather complicated and circuitous chain of reasoning by which the above result is obtained. One first makes the *assumption* that the particle moves in response to a randomly fluctuating force which has a *constant*, but unknown, spectral density. (The spectral density is, in actuality, not constant, and Eq. (5.1) is not valid at high frequencies.) By application of the theory of stochastic processes, one is then able to predict the distribution functions for either the displacement or the velocity of the particle.[4] The distribution function for displacement yields the diffusion constant, which in turn may be related by the Einstein relation[5] to the frictional constant η, thus evaluating the spectral density.[6] Alternatively, the distribution function for velocity yields the energy, which is known by the equipartition theorem and which therefore evaluates the spectral density, yielding Eq. (5.1).

We now apply our general formalism to the Brownian motion. We assume the existence of a viscous force as given by Eq. (5.2). The system of a particle in a fluid, the particle being acted on by an external force, is then dissipative and linear. The real part of the impedance is simply η (the inertial mass of the particle giving a pure reactance of $m\omega$). We conclude immediately, in accordance with Eq. (4.8), that a particle in a fluid is acted upon by a spontaneously fluctuating force for which

$$\langle F_x^2\rangle=(2/\pi)\eta\int_0^\infty E(\omega,T)d\omega. \quad (5.3)$$

For high temperatures or low frequencies, ($\hbar\omega\ll kT$); this reduces to Eq. (5.1).

VI. ELECTRIC DIPOLE RADIATION RESISTANCE AND ELECTRIC FIELD FLUCTUATIONS IN THE VACUUM

An oscillating electric charge radiates energy, leading to a radiation resistance. We shall see that this radiation resistance implies a fluctuating electric field as given by the Planck radiation law.

[4] See M. C. Wang and G. E. Uhlenbeck, Revs. Modern Phys. 17, 323 (1945); and J. L. Doob, Ann. Math. 43, 351 (1942).

[5] See A. Einstein, *Investigations on the Theory of the Brownian Movement* (Dutton and Company, New York); or A. Einstein, Ann. Physik 17, 549 (1905).

[6] A similar analysis has been applied to the flow of heat by L. S. Ornstein and J. M. W. Milatz, Physica 6, 1139 (1939).

Consider a dipole, of charge e, displacement x, and dipole moment $p=ex$. Let one charge be fixed and let the other oscillate so that

$$P=P_0\sin\omega t. \tag{6.1}$$

It is well known that the electric dipole radiation leads to a dissipative force[7]

$$F_d=-\tfrac{2}{3}e^2c^{-3}d^2v/dt^2, \tag{6.2}$$

where v is the velocity of the moving charge. The equation of motion of this charge is

$$mdv/dt+m\omega_0{}^2x+F_d=F, \tag{6.3}$$

where F is the applied force, and ω_0 is the natural frequency associated with the intra-dipole binding force. Inserting (6.1) in (6.3) we get

$$F=mP_0e^{-1}(\omega_0{}^2-\omega^2)\sin\omega t+\tfrac{2}{3}e\omega^3c^{-3}P_0\cos\omega t. \tag{6.4}$$

One may note that the average rate of radiation of energy $\langle Fv\rangle$ is

$$\langle Fv\rangle=\tfrac{1}{2}(\tfrac{2}{3}e\omega^3c^{-3}P_0)(\omega P_0e^{-1})=\tfrac{1}{3}\omega^4c^{-3}P_0{}^2. \tag{6.5}$$

The real part of the impedance is obtained by taking the ratio of the in-phase component of F to v. Thus

$$R(\omega)=(\tfrac{2}{3}e\omega^3c^{-3}P_0)/(\omega P_0e^{-1})=\tfrac{2}{3}e^2c^{-3}\omega^2. \tag{6.6}$$

According to our general theorem, we now deduce that there exists a randomly fluctuating force $e\mathcal{E}_x$ on the charge, and hence a randomly fluctuating electric field $\mathcal{E}_x$, such that

$$\langle e^2\mathcal{E}_x{}^2\rangle=(2/\pi)\int_0^\infty E(\omega,T)\tfrac{2}{3}e^2c^{-3}\omega^2d\omega,$$

or

$$\langle\mathcal{E}_x{}^2\rangle=(4/3)\pi^{-1}c^{-3}$$
$$\times\int_0^\infty\{\tfrac{1}{2}\hbar\omega+\hbar\omega[\exp(\hbar\omega/kT)-1]^{-1}\}\omega^2d\omega. \tag{6.7}$$

This conclusion can be put into a more familiar form by utilizing the fact that the energy density in an isotropic radiation field is simply

$$\text{Energy density}=\langle\mathcal{E}^2\rangle/4\pi=3\langle\mathcal{E}_x{}^2\rangle/4\pi \tag{6.8}$$

whence

Energy density

$$=\pi^{-2}c^{-3}\int_0^\infty\{\tfrac{1}{2}\hbar\omega+\hbar\omega[\exp(\hbar\omega/kT)-1]^{-1}\}\omega^2d\omega. \tag{6.9}$$

The first term in this equation gives the familiar infinite "zero-point" contribution, and the second term gives the Planck radiation law.[8]

VII. ACOUSTIC RADIATION RESISTANCE AND PRESSURE FLUCTUATIONS IN A GAS

We now consider the acoustic radiation from a small oscillating sphere in a gaseous medium. This radiation leads to a radiation impedance which, in accordance with our general theorem, implies a fluctuating pressure in the gas.

The wave equation for the propogation of pressure waves in the gas is

$$\nabla^2P=c^{-2}\partial^2P/\partial t^2, \tag{7.1}$$

where c is the velocity of sound in the gas. Let the radius of the sphere be a, and let

$$a=a_0+e^{-i\omega t}\delta a \tag{7.2}$$

so that the sphere expands and contracts sinusoidally. The boundary condition to be satisfied by the pressure waves at $r=a_0$ is

$$\rho\partial^2a/\partial t^2=-\partial P/\partial r\quad\text{at}\quad r=a_0, \tag{7.3}$$

where ρ is the equilibrium value of the density. The solution of these equations is readily found to be

$$P=r^{-1}P_0\exp(iKr-i\omega t), \tag{7.4}$$

where

$$K=\omega/c \tag{7.5}$$

and

$$P_0=-\rho\omega^2a_0{}^2\delta a[1+iKa_0]\cdot[1+(Ka_0)^2]^{-1}$$
$$\times\exp(-iKa_0). \tag{7.6}$$

Thus, the compressive force acting on the surface of the sphere is

$$F=4\pi a_0P_0\exp(iKa_0-i\omega t), \tag{7.7}$$

and defining the radiation impedance as the ratio of complex force to complex velocity, we find

$$\begin{aligned}Z&=F/[-i\omega\exp(-i\omega t)\delta a]\\ Z&=[4\pi a_0{}^2\rho c(Ka_0)^2-i4\pi a_0{}^2\rho cKa_0]/[1+(Ka_0)^2].\end{aligned} \tag{7.8}$$

The generalized Nyquist relation now states that a sphere immersed in a gas will experience a fluctuating compressive force, such that

$$\langle F^2\rangle=(2/\pi)\int E(\omega,T)4\pi a_0{}^2\rho c(\omega a_0/c)^2$$
$$\times[1+(\omega a_0/c)^2]^{-1}d\omega. \tag{7.9}$$

The fluctuating pressure is the compressive force per unit area on a vanishingly small sphere.

$$\langle P^2\rangle=\lim_{a_0\to0}\langle F^2\rangle/(4\pi a_0{}^2)^2, \tag{7.10}$$

or

$$\langle P^2\rangle=\tfrac{1}{2}\pi^{-2}c^{-1}\rho\int E(\omega,T)\omega^2d\omega. \tag{7.11}$$

This result may be checked by a direct computation paralleling the standard derivation of the Planck radiation law for the electromagnetic modes in a

[7] W. Heitler, *The Quantum Theory of Radiation* (Oxford University Press, London, 1936).

[8] The interaction of free electron and radiation field has been discussed from a somewhat different point of view by W. Pauli, Z. Physik 18, 272 (1923); A. Einstein and P. Ehrenfest, Z. Physik 19, 301 (1923).

vacuum. The number of acoustic modes with frequency between ω and $\omega+d\omega$ is $\frac{1}{2}\pi^{-2}c^{-3}\omega^2 d\omega$, and the acoustic energy density is

$$\text{Energy density}=\int E(\omega,\,T)\tfrac{1}{2}\pi^{-2}c^{-3}\omega^2 d\omega. \quad (7.12)$$

Employing the relation that the acoustic energy density is proportional to the mean square excess pressure

$$\text{Energy density}=\rho^{-1}c^{-2}\langle P^2\rangle, \quad (7.13)$$

we again obtain Eq. (7.11).

It is interesting to compare the above result with the pressure fluctuations at a boundary of the gas. The proximity to the boundary, and the shape of the boundary, may be expected to influence the radiation impedance and hence the pressure fluctuations. We consider the pressure fluctuations immediately contiguous to a plane rigid boundary, and we shall find that for this simple case, the mean square pressure fluctuation is just twice that in the volume of the gas.

Consider a plane wall bounding a semi-infinite region containing the gas. If the wall contains a circular piston of radius a_0, the radiation resistance is[9]

$$R=\pi a_0^2\rho c[1-ca_0^{-1}\omega^{-1}J_1(2\omega a_0/c)], \quad (7.14)$$

where J_1 indicates the first order bessel function. The fluctuating force acting on a circular area in a plane boundary is therefore

$$\langle F^2\rangle=(2/\pi)\int E(\omega,\,T)\pi a_0^2\rho c$$
$$\times[1-ca_0^{-1}\omega^{-1}J_1(2\omega a_0/c)]d\omega, \quad (7.15)$$

and the fluctuating pressure is

$$\langle P^2\rangle=\lim_{a_0\to 0}\langle F^2\rangle/(\pi a_0^2)^2 \quad (7.16)$$

or

$$\langle P^2\rangle=\rho\pi^{-2}c^{-1}\int E(\omega,\,T)\omega^2 d\omega. \quad (7.17)$$

Thus the mean square fluctuating wall pressure, as given by (7.17), is just twice the mean square fluctuating volume pressure, as given by (7.11). This factor of two clearly arises from the fact that the pressure waves in the gas must have velocity nodes at the wall. Fluctuations in the neighborhood of the wall may be found by treating the radiation from an oscillating sphere near a reflecting boundary.

Finally, it will be noted that the above equations for pressure fluctuations involve the velocity of sound in the gas which is not a usual thermodynamic parameter. This quantity may, however, be expressed in terms of standard thermodynamic quantities. Thus for frequencies which are sufficiently high that the compressions in the acoustic waves may be considered to be adiabatic, we have[9]

$$c^2=C_PC_V^{-1}\rho^{-1}\mathcal{K}_T^{-1}, \quad (7.18)$$

where C_P and C_V are the specific heats at constant pressure and volume, ρ is the density, and $\mathcal{K}_T$ is the isothermal compressibility. For these frequencies, the pressure fluctuations in the volume of the gas are thus given by

$$\langle P^2\rangle=\tfrac{1}{2}\pi^{-2}\rho^2\mathcal{K}_TC_VC_P^{-1}\int E(\omega,\,T)\omega^2 d\omega. \quad (7.19)$$

VIII. CONCLUSION

The generalized Nyquist relation establishes a quantitative correlation between dissipation, as described by the resistance, and certain fluctuations. It seems to be possible to give an intuitive interpretation of such a connection.

A dissipative process may be conveniently considered to involve the interaction of two systems, which we characterize as the "source system" and the "dissipative system." The dissipative system, explicitly considered in Secs. II and III, is necessarily a system with densely distributed energy levels and is capable of absorbing energy when acted upon by a periodic force. The source system is the system which provides this periodic force and which delivers energy to the dissipative system.

Assume the source system to be first isolated from the dissipative system and to be given some internal energy. If the source system is a simple dynamical system, its subsequent dynamics will be periodic (as, for instance, the oscillations of a pendulum or of a polyatomic molecule). The system may be thought of as possessing a sort of internal coherence. If, now, the source system is allowed to act on the dissipative system, this internal coherence is destroyed, the periodic motion vanishes and the energy is sapped away, and the source system is left at last with only the random disordered energy ($\simeq kT$) characteristic of thermal equilibrium. This loss of coherence within the source system may be thought of as being *caused* by the random fluctuations generated by the dissipative system and acting on the source system. The dissipation thus appears as the macroscopic manifestation of the disordering effect of the Nyquist fluctuations and, as such, is necessarily quantitatively correlated with the fluctuations.

An analogy which is perhaps useful is provided by the historical development of the theory of spontaneous radiation from excited atoms. After the initial development of quantum mechanics, it was found impossible to compute the spontaneous transition probabilities for an isolated excited atom, and this dissipative process appeared to be outside the existing structure of dynamics. With the development of quantum electrodynamics, however, the dissipation could be computed,

[9] P. M. Morse, *Vibration and Sound* (McGraw-Hill Book Company, Inc., New York, 1936).

and it was found that the "spontaneous" transitions could be consistently considered to be induced by the random fluctuations of the electromagnetic field in the vacuum. In this case, of course, the excited atom plays the role of the source system, and the "vacuum" plays the role of the dissipative system.

It would thus appear that a reasonable approach to the development of a theory of linear irreversible processes is through the development of the theory of fluctuations in equilibrium systems. Certain results in this connection will be given in subsequent papers by Richard F. Greene and one of the authors (H.B.C.).

PHYSICAL REVIEW VOLUME 83, NUMBER 1 JULY 1, 1951

The Disintegration of Nd^{147}

W. S. Emmerich and J. D. Kurbatov
Ohio State University, Columbus, Ohio

(Received March 12, 1951)

Three groups of monochromatic electrons corresponding to gamma-rays of 91.5±1.0 kev, 320±3 kev, and 534±4 kev have been identified in the disintegration of Nd^{147}. These gamma-rays are ascribed to transitions in Pm^{147}. Evidence has been obtained for a complex beta-spectrum of Nd^{147}. On the basis of coincidences, a partial scheme of disintegration for Nd^{147} is proposed. The total energy decrease from the ground state of Nd^{147} to the ground state of Pm^{147} nuclei is 1.425±0.015 Mev.

I. INTRODUCTION

SINCE the original observation of an 11-day period radioactive neodymium[1] several papers have been published on the radiations emitted by this species.[2-6] When fission products became available, it was possible to identify the 11-day period as mass number 147. By absorption technique, the beta-emission was found to be 0.9 Mev and ~0.4 Mev with intensities of 60 and 40 percent, respectively.[4] Low energy electrons, x-rays, and gamma-rays of ~0.58 Mev with an intensity of 40 percent were also observed. Coincidences were obtained between high energy betas and x-rays, also between lower energy betas and gammas. In a recent letter[6] beta-energies of 780 kev and 175 kev were reported, complex beta-gamma-coincidences were found, and the absence of gamma-gamma-coincidences noted.

II. PROCEDURE

For the present study, commercially supplied neodymium in the form of neodymium oxide was irradiated with neutrons at the Oak Ridge National Laboratory. The activated material was aged to allow for the decay of 12-min Pm^{151} and 47-hr Pm^{149}. Corrections were applied in various phases of this investigation for the growth of 0.22-Mev beta-rays of the daughter product, Pm^{147}. Spectrometer sources were prepared on Cellophane tape. They were not covered, and a radiogram showed that the distribution of the activity was practically uniform. Since neutron-activated material was used, inert neodymium was present, and a method of obtaining a correction for scattering in the source (at low energies) is mentioned below in connection with the correction for counter window absorption. Sources for the coincidence counter were mounted on Cellophane and covered with zapon.

Measurements were carried out with the aid of a permanent magnet electron spectrograph, a thick lens beta-spectrometer, and coincidence counters.

The electron spectrograph is of the semicircular type using photographic plates as detectors. Although large sources had to be used to attain sufficient intensity for the measurement of internal conversion electrons in this instrument, a resolution of ~ one percent could be obtained for energies over 100 kev. The magnetic field strength between the pole pieces was determined with internal conversion electrons of I^{131} and Cs^{137}.

The ring focusing was efficiently attained in the beta-spectrometer by using a coil that extends along the total path length of the electrons. The position of the defining baffle was found from electron trajectories as determined by the empirically measured magnetic field distribution inside the coil. Additional baffles were installed to minimize scattering from the walls and to eliminate the higher order focusing of slow electrons. A 2-mg/cm^2 mica window G-M counter was used as detector. The spectrometer was operated with a resolution of ~4

TABLE I. Internal conversion electrons in Nd^{147}.

Energy of electrons (kev)	Estimated intensity	Conversion shell	Energy of gamma-ray (kev)
46.0±0.5	strong	*K*	91.5
84.5±0.5	strong	*L*	
89.9±0.5	medium	*M*	
275 ±3	weak	*K*	320
315 ±4	very weak	*L*	
489 ±4	medium	*K*	534
528 ±5	weak	*L*	

[1] Law, Pool, Kurbatov, and Quill, Phys. Rev. 59, 936 (1941).
[2] W. Bothe, Z. Naturforsch. 1, 179 (1946).
[3] Cork, Shreffler, and Fowler, Phys. Rev. 74, 240 (1948).
[4] Marinsky, Glendenin, and Coryell, J. Am. Chem. Soc. 69, 2781 (1947).
[6] C. E. Mandeville and E. Shapiro, Phys. Rev. 79, 391 (1950).

REF. 00

PHYSICAL REVIEW D

PARTICLES AND FIELDS

THIRD SERIES, VOLUME 42, NUMBER 8 15 OCTOBER 1990

Thermal noise in mechanical experiments

Peter R. Saulson*

Joint Institute for Laboratory Astrophysics, National Institute of Standards and Technology, and University of Colorado, Boulder, Colorado 80309-0440

(Received 8 June 1990)

The fluctuation-dissipation theorem is applied to the case of low-dissipation mechanical oscillators, whose losses are dominated by processes occurring inside the material of which the oscillators are made. In the common case of losses described by a complex spring constant with a constant imaginary part, the thermal noise displacement power spectrum is steeper by one power of ω than is predicted by a velocity-damping model. I construct models for the thermal noise spectra of systems with more than one mode of vibration, and evaluate a model of a specific design of pendulum suspension for the test masses in a gravitational-wave interferometer.

I. INTRODUCTION

Thermal noise is one of the fundamental limits to the precision of mechanical measurements. Its importance in high-sensitivity galvanometers is well studied.[1] It is also one of the dominant noise sources in resonant-mass detectors of gravitational waves and a major reason that such detectors operate at cryogenic temperatures.[2] In both of these instruments, what is observed is that the thermal noise excites the mechanical resonance with a root-mean-square level that corresponds to an energy of $k_B T$.

In many experiments, it is the thermal noise far from the resonant frequency that is most important. In laser interferometer gravitational-wave detectors, for example, resonant mechanical systems are employed, but mainly in the role of vibration isolators, with the resonant frequencies lying below the signal band.[3] Thermal noise motion of the test masses in the nearly free regime above the resonances is expected to be an important noise source, and thermal noise in the signal band from high-frequency internal resonances in the test masses may also be important. Other gravitational experiments employ delicate torsion balances. These are typically used in a mode where the signal frequency is well below the fundamental resonance.[4] Thermal noise sets a significant noise floor in these measurements as well.

Models of thermal noise almost invariably assume that the dissipative force is proportional to velocity. (Notable exceptions are the work of Speake and of Chan and Paik.[5]) However, in the low-loss oscillators typically used in sensitive gravitational experiments, the dependence of the dissipation on frequency seldom obeys this expected behavior. Calculations of thermal noise based on velocity-damping models can be seriously in error. In this paper I will discuss more realistic models of mechanical oscillators with small dissipation.

II. BROWNIAN MOTION

Brownian motion of a particle of mass m, subject to a frictional force of the form $F_{\text{friction}} = -fv$, is described by the Langevin equation[6]

$$m\ddot{x} + f\dot{x} = F_{\text{th}} , \tag{1}$$

where F_{th} is a random force with a white spectral density:

$$F_{\text{th}}^2(\omega) = 4k_B T f . \tag{2}$$

(Throughout this paper, I will use angular frequencies, with dimensions of rad/s, but will give power spectral densities referred to the customary 1-Hz bandwidth.) As is well known, the fluctuating force F_{th} comes about because of the randomness of the individual impacts from the molecules that make up the medium responsible for the deterministic force F_{friction}.

A damped harmonic oscillator [like the one shown in Fig. 1(a)] can be described by adding a term representing a Hooke's law restoring force $F_{\text{spring}} = -kx$, giving

$$m\ddot{x} + f\dot{x} + kx = F_{\text{th}} . \tag{3}$$

This equation of motion is easy to solve in the frequency domain, by replacing $x(t)$ with $x(\omega)e^{i\omega t}$. Then the power spectral density of the position of the mass can be shown to be

©1990 The American Physical Society

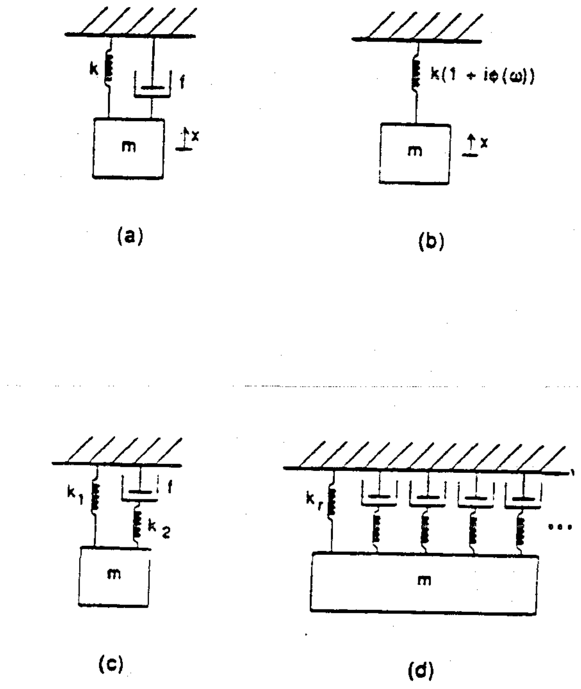


FIG. 1. (a) Schematic diagram of a mechanical oscillator consisting of a mass m, a spring of real spring constant k, and a dashpot with velocity coefficient f. (b) An oscillator consisting of a mass m and a spring with complex spring constant $k[1+i\phi(\omega)]$. (c) Schematic diagram of a standard anelastic solid. An ideal spring is connected in parallel with a spring-dashpot combination called a Maxwell unit. (d) Schematic model of an oscillator with an arbitrary frequency-dependent spring constant, constructed from a single ideal spring and many Maxwell units.

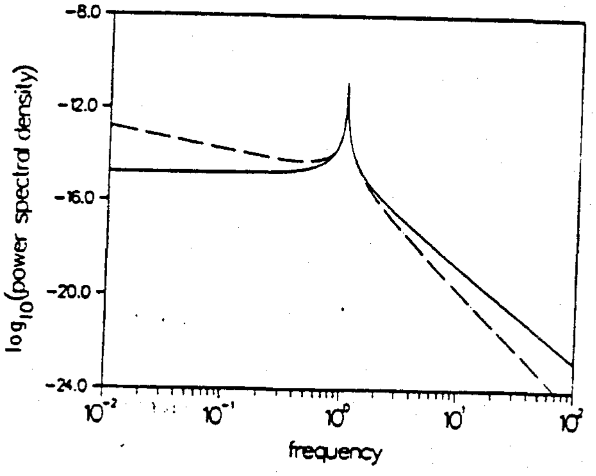


FIG. 2. *Thermal noise power spectra for two mechanical oscillators*, each with $m = 1$ g, resonant frequency $\omega_0 = 1\ \mathrm{s}^{-1}$, and $Q = 100$. The solid line shows the spectrum for an oscillator with damping proportional to velocity. The dash-dotted line shows the spectrum for an oscillator with internal damping characterized by constant $\phi(\omega)$. The units of the power spectral density are cm^2/Hz and of the frequency axis are s^{-1}.

$$x^2(\omega)=\frac{4k_BTf}{(k-m\omega^2)^2+f^2\omega^2}\ . \tag{4}$$

A graph of this power spectrum (for a representative set of the parameters k, m, and f) is shown in Fig. 2. If f is small, then the response of the particle is sharply peaked near $\omega_0=\sqrt{k/m}$. It is customary to denote the sharpness of the resonance by $Q\equiv\omega_0/\Delta\omega$, where $\Delta\omega$ is the full width measured at the half-power points. For the velocity-damped harmonic oscillator, $Q=m\omega_0/f$.

Predictions for the thermal noise in many delicate mechanical experiments have been made based on such models. In the next section I will set up a framework for more realistic models.

III. FLUCTUATION-DISSIPATION THEOREM

The fluctuations analogous to Brownian motion in any system with dissipation may be found using the fluctuation-dissipation theorem of Callen *et al.*[7] The spectral density of the thermal driving force is given by

$$F_{\mathrm{th}}^2(\omega)=4k_BTR(\omega)\ , \tag{5}$$

where $R(\omega)$ is the mechanical resistance, the real part of the impedance $Z\equiv F/v$ at the mass. Equivalently, the power spectrum of the motion of the mass is given by

$$x^2(\omega)=\frac{4k_BT\sigma(\omega)}{\omega^2}\ , \tag{6}$$

with $\sigma(\omega)$ denoting the mechanical conductance, the real part of the admittance $Y(\omega)\equiv Z^{-1}(\omega)$.

For the simple oscillator described above, the impedance is

$$Z=f+i\omega m+\frac{k}{i\omega}\ . \tag{7}$$

The admittance is

$$Y=\frac{\omega^2f+i(\omega k-m\omega^3)}{(k-m\omega^2)^2+\omega^2f^2}\ . \tag{8}$$

Substituting the real part of Eq. (8) into Eq. (6), we obtain the same result for the displacement power spectrum as we did using the Langevin equation directly.

IV. EXTERNAL VELOCITY DAMPING

There are several common sources of damping that give forces proportional to velocity. The classic example is the viscous drag on a Brownian particle suspended in a liquid. A few high-precision experiments operate at high enough pressure so that the drag from the residual gas is in the viscous regime.[8]

Most gravitational experiments are performed at low pressures (around 10^{-6} Torr or lower). When the mean free path is large compared to a characteristic dimension of the test object, a description in terms of viscosity is no longer applicable. Instead, one must calculate the sum of the momentum transfer between the test object and each of the gas molecules that collide with it. It can be shown that the oscillator Q can be estimated as[9–12]

$$Q_{\text{gas}}=Ch\frac{\rho\omega_0}{n\sqrt{m_{\text{mol}}k_BT}} . \tag{9}$$

Here ρ is the density of the oscillator mass, n is the number density of gas molecules, each of which has mass m_{mol}, h is a characteristic dimension of the oscillator, and C is a dimensionless constant of order unity that depends on the shape of the oscillator.

For a 1-Hz pendulum of mass 10 kg, operating at pressures below 10^{-6} Torr, values of Q_{gas} in excess of 10^9 should be readily attainable. This means that gas damping can be made negligible compared to the internal damping mechanisms described in Sec. V. Torsion balances, on the other hand, typically have much smaller values of h and ω_0, and so gas damping is often an important source of dissipation for them.[4,10]

Eddy currents in moving conductors also give a damping force that is proportional to velocity.[10] Good magnetic shielding, and use of nonconductors wherever possible, can reduce this to small levels.

V. INTERNAL DAMPING

Internal damping in materials has been found[13] to obey an extension of Hooke's law, which can be approximated by

$$F=-k[1+i\phi(\omega)]x . \tag{10}$$

If the force F is sinusoidal, the response x of the spring will lag the force by the angle $\phi(\omega)$. The time average of the product $F\dot{x}$ is proportional to ϕ (as long as $\phi \ll 1$). A fraction $2\pi\phi$ of the energy stored in the oscillatory motion is being dissipated during each cycle. Thus a complex spring constant is inevitably associated with damping. In turn, the fluctuation-dissipation theorem guarantees that damping generates mechanical noise.

It is instructive to study a simple mathematical model of an oscillator, substituting a general spring "constant" of the form of Eq. (10) for the velocity-damping term [see Fig. 1(b)]. The equation of motion becomes

$$m\ddot{x}=-k(1+i\phi)(x-x_g)+F . \tag{11}$$

The vibration transfer function is

$$\frac{x}{x_g}=\frac{\omega_0^2(1+i\phi)}{\omega_0^2-\omega^2+i\phi\omega_0^2} ; \tag{12}$$

By comparison with Eq. (4), it is easy to see that an oscillator of this sort has a quality factor given by

$$Q=\frac{1}{\phi(\omega_0)} . \tag{13}$$

The mechanical impedance at the mass is

$$Z=i\omega m+\frac{k}{i\omega}+\frac{k\phi}{\omega} , \tag{14}$$

and so the thermal noise force spectral density is proportional to the quantity $k\phi(\omega)/\omega$ in place of the velocity coefficient f. The admittance is

$$Y=\frac{\omega k\phi+i(\omega k-m\omega^3)}{(k-m\omega^2)^2+k^2\phi^2} . \tag{15}$$

The thermal noise power spectral density is given, according to the fluctuation-dissipation theorem, by

$$x^2(\omega)=\frac{4k_BTk\phi(\omega)}{\omega[(k-m\omega^2)^2+k^2\phi^2]} . \tag{16}$$

4kT ω_0^2 ϕ

$\omega\{(\omega_0^2$ [illegible]

VI. FORMS OF INTERNAL DAMPING

By far the most common functional form for $\phi(\omega)$ in materials of many kinds is ϕ approximately constant over a large band of frequencies.[14] [The lag function $\phi(\omega)$ can be any odd function of frequency.[15] Constant $\phi(\omega)$ is consistent with this condition as long as ϕ does not remain constant all the way to zero frequency.] In spite of the ubiquity of constant $\phi(\omega)$, there does not seem to be a simple model that gives a general explanation of the phenomenon. In some cases, a frequency-independent ϕ has been attributed to friction from dislocations.[16]

Sometimes, the damping exhibits a broad maximum at a characteristic frequency τ^{-1}. This is the classic phenomenon named "anelasticity" by Zener.[17] Such behavior is caused by the functional dependence of some internal degree of freedom of the system upon the stress. For oscillatory stresses applied near τ^{-1}, the response of the material can lag substantially because of the finite time it takes for the internal degree of freedom (and consequently the strain of the material) to come to equilibrium.

A simple model, called the standard anelastic solid, can be used to represent the relaxation process described in the previous paragraph. One way to represent this model is by an arrangement of two springs and a dashpot, as shown in Fig. 1(c). The spring constant k_1 is called the "relaxed spring constant," and the sum k_1+k_2 is called the "unrelaxed spring constant." (If the losses are small, then k_2 is much smaller than k_1.) Zener showed that this model predicts that the loss angle ϕ depends on frequency with the characteristic form

$$\phi=\Delta\frac{\omega\tau}{1+\omega^2\tau^2} , \tag{17}$$

as long as there are no other mechanisms with nearby relaxation times and $\phi \ll 1$. $\Delta \equiv k_2/k_1$ is called the "relaxation strength," while $\tau \equiv f/k_2$ is the "relaxation time."

VII. EQUIPARTITION THEOREM

The power spectrum of thermal noise will not in general have the functional form given in Eq. (4) for the case of velocity damping. For example, an oscillator with losses characterized by constant $\phi(\omega)$ has thermal noise whose power spectral density declines more rapidly with frequency (by one power of ω) than an oscillator subject to velocity damping (see the graph in Fig. 2). This means that if one had erroneously assumed velocity damping in a system with constant internal damping, one would have overestimated the thermal noise density for frequencies above the resonant frequency, but would have underestimated the noise in the region below the resonance. Relaxation damping, such as that described by the standard anelastic solid, also gives more noise at frequencies below $\omega_r \equiv 1/\tau$ than at higher frequencies.

The integral of Eq. (4) over all frequencies gives a mean-square displacement $\bar{x}^2_{\rm th}=k_B T/k$. This is, of course, consistent with the equipartition theorem, which states that each quadratic term in the energy has a mean value of $\frac{1}{2}k_B T$. Contrast the case of Eq. (16) for the case of $\phi(\omega)$ a constant. Here the integral diverges at $\omega=0$.

It is instructive to explore the relation between the equipartition theorem and thermal noise power spectra in general. First, it is important to point out that a spring constant of the form given in Eq. (10) is usually an excellent approximation, but it cannot be exact. For the standard anelastic solid of Fig. 1(c), a direct calculation gives a spring constant of

$$F=k_1 x\left[1+\frac{\Delta\omega^2\tau^2}{1+\omega^2\tau^2}+i\Delta\frac{\omega\tau}{1+\omega^2\tau^2}\right] . \qquad (18)$$

Note the additional frequency-dependent real term. For the common case of small Δ, this term is usually negligible compared to the constant term. The heuristic interpretation is that at high frequencies the effective spring constant is the unrelaxed spring constant k_1+k_2, because the dashpot appears rigid. At low frequencies, the dashpot is free to move, and so the effective spring constant is just the relaxed spring constant k_1.

This is a special case of a theorem usually attributed to Bode,[18] stating that there is a unique relationship between the phase of a network characteristic (such as a transfer function, impedance, admittance, or spring constant) and the functional form of its magnitude, as long as it has no poles or zeros in the right half of the complex plane. For example, such a function with magnitude proportional to ω^n has constant phase of $n\pi/2$. Applied to a spring with constant phase ϕ, the theorem requires that the magnitude is proportional to $\omega^{2\phi/\pi}$. For a low-loss spring, this is an extremely weak dependence on frequency, which is why it is usually neglected.

In order to understand mean-square thermal noise displacements, it is important to keep in mind the weak variation of a spring constant with frequency. Again, let us consider first the standard anelastic solid. Direct integration of the thermal noise power spectrum gives the result $\bar{x}^2_{th}=k_B T/k_1$. This is also the result expected from the equipartition theorem, since the displacement of the mass is equal to the extension of the energy storage element k_1, the relaxed spring constant.

A similar explanation can be given for other forms of the frequency-dependent spring constant. It is always possible to represent an arbitrary lossy spring with a model such as the one shown in Fig. 1(d). Here the single spring-dashpot element of the standard anelastic solid is replaced by a spectrum of such elements, whose spring constants and relaxation times are adjusted to give the observed frequency dependence. (Various methods to construct such a model are discussed by Nowick and Berry.[13]) The spectrum contains a longest relaxation time $\tau_{\rm max}$. At frequencies below $1/\tau_{\rm max}$, the spring behaves like the ideal spring k_r. [The requirement that there exist a longest relaxation time is a restatement of the realizability condition that $\phi(\omega)$ be an odd function of ω.] The equipartition theorem then states that $\bar{x}^2_{\rm th}=k_B T/k_r$.

A real experimental measurement of the mean-square displacement cannot integrate all the way to $\omega=0$, but only down to a frequency $\omega\approx 1/\tau_{\rm int}$, where $\tau_{\rm int}$ is the duration of the measurement. Thus, if $\tau_{\rm max}>\tau_{\rm int}$, one should not expect the equipartition theorem to hold exactly. Instead, we expect the approximate relation $\bar{x}^2_{{\rm th},\tau}\approx k_B T/k(1/\tau_{\rm int})$, where $\bar{x}^2_{{\rm th},\tau}$ is the mean-square displacement as measured in the finite integration time, and $k(1/\tau_{\rm int})$ is the magnitude of the spring constant at a frequency $\omega=1/\tau_{\rm int}$. This can be interpreted in light of a model of the form shown in Fig. 1(d). For measurements extending only to $\tau_{\rm int}$, all of the spring-dashpot elements that have $\tau>\tau_{\rm int}$ behave as if their spring constants were added to k_r. Thus the effective relaxed spring constant is approximately $k(1/\tau_{\rm int})$.

The contribution of the formally divergent part of the integral (for ϕ constant) will be quite small in most experiments. This is because in a lightly damped oscillator, most of the power is in the resonant peak itself. In the case of velocity damping, only roughly $1/Q$ of the mean-square displacement comes from frequencies below the resonance. For damping with constant ϕ, the highest octave below the resonance thus contains approximately the fraction ϕ of the total thermal noise power. Each octave lower in frequency contains the same power, since the power spectral density is proportional to $1/\omega$. In particular, to obtain a power comparable to that in the resonant peak, this behavior must extend down in frequency for ϕ^{-1} octaves below the resonance. At such a low frequency, the magnitude of the spring constant has declined by about a factor of 2 below its unrelaxed value k_u. Thus the integral of the power spectrum down to such a frequency is consistent with the prediction $\bar{x}^2_{\rm th}=k_B T/(k_u/2)$.

A numerical example will help to put this issue in perspective. Consider a 1-Hz oscillator that has damping characterized by frequency independent $\phi=10^{-3}$. In order to measure a mean-square displacement twice the velocity-damping prediction, it would be necessary to use an integration time of roughly 10^{300} s. Clearly, the low-frequency divergence of Eq. (16) is of more formal than practical concern.

VIII. THERMOELASTIC DAMPING

As an example of anelasticity, consider the mechanism known as thermoelastic damping. This can be an important source of losses for thin samples in flexure. The internal degree of freedom involved is the temperature, which couples to the strain because materials have nonzero coefficients of thermal expansion. As a wire is flexed, one side heats and the other cools. Heat flows to attempt to restore equilibrium, causing the restoring force from the wire to relax from its initial value to a smaller equilibrium value.

The theory of this mechanism was given by Zener.[19]

He showed that this mechanism is well described by a model of the form described above, with the parameters

$$\Delta=\frac{E\alpha^2 T}{c}\ , \tag{19}$$

and

$$f_0=\frac{1}{2\pi\tau}=2.16\frac{D}{d^2}\ . \tag{20}$$

Here E is the (unrelaxed) Young's modulus of the material, α is the linear coefficient of thermal expansion, and c is the specific heat per unit volume. In Eq. (20), d is the diameter of the wire, and the thermal diffusion coefficient D is given by $D=\kappa/c$, where κ is the thermal conductivity. Note that this damping mechanism depends only on the sort of properties of a material that are tabulated in handbooks, and not on details of its structure or composition.

Zener gave the solution not only for wires, but also for ribbons of rectangular cross section. The sole difference is that the characteristic frequency is given by

$$f_0=\frac{\pi}{2}\frac{D}{t^2}\ , \tag{21}$$

where t is the thickness of the ribbon. Thus, if the characteristic frequency is larger than the frequencies of interest, the thermoelastic damping effect can be reduced by flattening the suspension member. This occurs at the expense, of course, of introducing an anisotropy into the compliance of the suspension.

I have given prominent treatment to the thermoelastic relaxation mechanism because it sets a fundamental limit beyond which the losses cannot be reduced, given a choice of wire material and geometry. Note that thermoelastic relaxation is of no consequence for the longitudinal modes of wires (vertical modes of a pendulum), since the relevant length scale is not the thickness of the wire but the acoustic wavelength in the wire. (Note also that if the oscillator in question was a torsion pendulum, then thermoelastic damping cannot apply, since torsional motion involves only shear, nowhere expansion or contraction.)

Other relaxation mechanisms depend on much more obscure properties of a specimen. Nowick and Berry[13] stress the use of experiments on anelastic behavior as a probe of the structure of solids.

IX. PENDULUM

The universal choice of a pendulum as the final suspension stage in gravitational-wave interferometers is based on the desire to minimize thermal noise. In a pendulum, the primary "spring" for horizontal motion is the gravitational field, with only a small amount of restoring force coming from flexure of the wire that supports the mass against gravity. The gravitational spring is free of loss, and so the only mechanical loss is the fraction $2\pi\phi(\omega)$ per cycle of the mechanical energy stored in the flexing wire. That is, the relationship between the pendulum loss ϕ_p and the loss in the wire ϕ_w is given by

$$\phi_p=\phi_w\frac{E_{\rm el}}{E_{\rm grav}+E_{\rm el}}\approx\phi_w\frac{E_{\rm el}}{E_{\rm grav}}\ , \tag{22}$$

where $E_{\rm el}$ and $E_{\rm grav}$ represent, respectively, the energy stored in the flexing wire and in the gravitational field.[12] Thus a pendulum can have much lower loss than the material of which it is made.

This calculation can be made more explicit by remembering that $E_{\rm el}/E_{\rm grav}=k_{\rm el}/k_{\rm grav}$. The gravitational spring constant is of course $k_{\rm grav}=mg/l$ for a pendulum of length l. The elastic spring constant for a pendulum in which the mass is supported by n wires is $k_{\rm el}=n\sqrt{TEI}/2l^2$, where T is the tension in each wire, E is the Young's modulus, and I is the moment of inertia of the wire cross section. Substituting into Eq. (22), we find that

$$\phi_p(\omega)=\phi_w(\omega)\frac{n\sqrt{TEI}}{2mgl}\ . \tag{23}$$

It is interesting to consider how the thermal noise in a pendulum scales with the mass. The explicit dependence, taking the high-frequency limit of Eq. (16), is $x^2(\omega)\propto m^{-1}$. But $\phi(\omega)$ [here $\phi_p(\omega)$] also depends on the suspended mass. In addition to the explicit dependence displayed in Eq. (23), remember that T is proportional to m, and if the wires are kept at a fixed fraction of their breaking stress, then $I\propto m^2$. Thus $\phi_p\propto m^{1/2}$, and so the thermal noise in a pendulum scales as[20]

$$x_p^2(\omega)\propto m^{-1/2}\ . \tag{24}$$

A similar analysis shows that $x_p^2(\omega)$ also scales as $n^{-1/2}$.

X. MULTIMODE OSCILLATORS

The remainder of this paper is devoted to applications of the fluctuation-dissipation theorem to systems more complicated than a damped harmonic oscillator. A two-mode oscillator is perhaps the simplest of such systems. (Time-domain treatments of the problem have been given by Wang and Uhlenbeck[6] and by Paik.[9])

Consider the system shown schematically in Fig. 3. The equations of motion are

$$\begin{aligned} m_1\ddot{x}_1&=-k_1x_1-f_1\dot{x}_1-k_2(x_1-x_2)-f_2(\dot{x}_1-\dot{x}_2)\ ,\\ m_2\ddot{x}_2&=-k_2(x_2-x_1)-f_2(\dot{x}_2-\dot{x}_1)+F\ . \end{aligned} \tag{25}$$

It is useful to define the quantities $\omega_1^2\equiv k_1/m_1$, $\omega_2^2\equiv k_2/m_2$, $\beta_1\equiv f_1/m_1$, $\beta_2\equiv f_2/m_2$, and $\mu\equiv m_2/m_1$. Transforming the equations of motion into the frequency domain, we obtain the 2×2 matrix equation

$$\begin{bmatrix}\omega_1^2+i\omega\beta_1-\omega^2+\mu(\omega_2^2+i\omega\beta_2) & -\mu(\omega_2^2+i\omega\beta_2)\\ -(\omega_2^2+i\omega\beta_2) & \omega_2^2+i\omega\beta_2-\omega^2\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix}=\begin{bmatrix}0\\ F/m_2\end{bmatrix}\ , \tag{26}$$

or

$$\mathbf{D}\mathbf{x}=\mathbf{F} \, . \tag{27}$$

Then it is easy to show that the impedance at the mass m_2 is

$$Z=\frac{m_2\det(\mathbf{D})}{i\omega[\omega_1^2+i\omega\beta_1-\omega^2+\mu(\omega_2^2+i\omega\beta_2)]} \, , \tag{28}$$

where det(D) is the determinant of the matrix D. Its real part is

$$R=\frac{m_2[\omega^6\beta_2+\omega^4(\beta_1^2\beta_2+\mu\beta_1\beta_2^2-2\beta_2\omega_1^2)+\omega^2(\beta_2\omega_1^4+\mu\beta_1\omega_2^4)]}{\omega^6+\omega^4(\beta_1^2-2\omega_1^2+2\mu(\beta_1\beta_2-\omega_2^2)+\mu^2\beta_2^2)+\omega^2(\omega_1^4+2\mu\omega_1^2\omega_2^2+\mu^2\omega_2^4)} \, . \tag{29}$$

This last expression, multiplied by $4k_BT$, gives the power spectral density of the thermal noise driving force applied to m_2. In the limit of large frequency, the real part of the impedance approaches $m_2\beta_2=f_2$, and so only the damping applied directly to m_2 matters. If that damping should be vanishingly small, the dominant term is $\mu^2 f_1\omega_2^4/\omega^4$.

For the case of $\mu=1$, $\omega_1=\omega_2=1$, $\beta_1=10^{-2}$, and $\beta_2=10^{-6}$, a graph of the thermal noise power spectrum is shown in Fig. 4. Note that both of the normal modes have a low Q, since both modes involve substantial motion of the more highly damped m_1. Yet the thermal noise motion of m_2 in the limit of high frequency is determined only by the damping coefficient β_2. This is in accord with the intuitive picture that the thermally driven fluctuations of m_1 can be thought of as an input to the lower oscillator that is filtered in the same way that the lower oscillator acts as a low-pass filter for vibration of any sort.

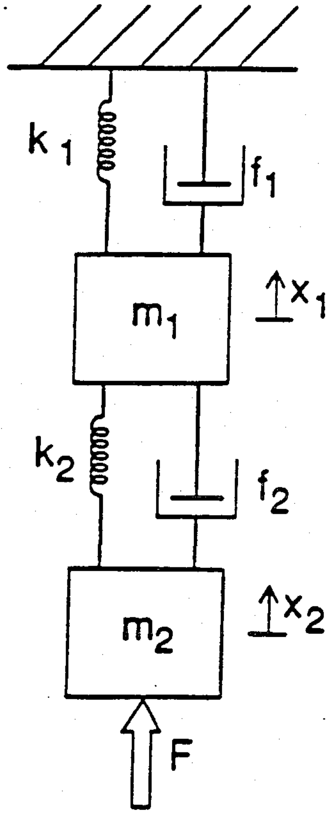


FIG. 3. Schematic diagram of a double oscillator. A force F may be applied to the second mass m_2.

XI. MODES OF CONTINUOUS SYSTEMS

It is sometimes necessary to take account of the fact that real oscillators are distributed systems, not point masses and massless springs. A pendulum exhibits transverse vibrational modes in its wire(s), as well as longitudinal modes of its mass. This means that Eq. (16) applied to the fundamental mode of the pendulum will cease to apply at a high enough frequency, since eventually the thermal noise from another mode of higher resonant frequency will dominate.

The character of the solution is especially clear in the admittance formulation of the fluctuation-dissipation theorem. The expansion theorem[21] states that the response of a system to an applied force is equal to the superposition of the responses of each of the normal modes of the system. Consider, for simplicity, a one-dimensional system with linear mass density $\rho(x)$. It has modes $\psi_n(x)$, which are normalized according to the relation

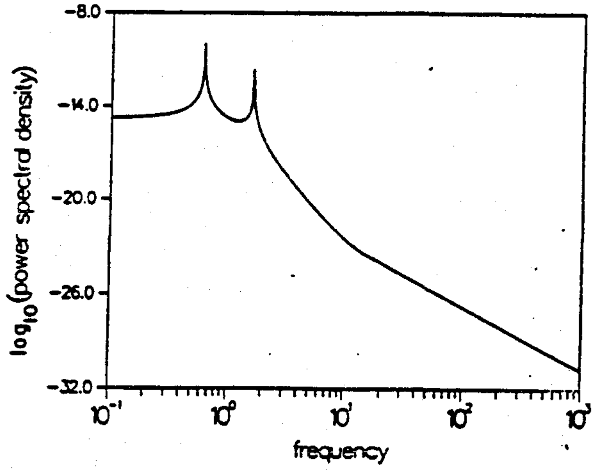


FIG. 4. Thermal noise power spectrum for a double oscillator. Each mass has $m=1$ g. The other parameters of the oscillator are as given in the text.

$$\int_0^L \rho(x)\psi_m(x)\psi_n(x)dx=\delta_{mn} \ . \qquad (30)$$

The normal-mode expansion of a particular displacement $y(x,t)$ is given by

$$y(x,t)=\sum_{n=1}^{\infty}\psi_n(x)q_n(t) \ , \qquad (31)$$

where $q_n(t)$ is the generalized coordinate of mode n. Its equation of motion has the form

$$\ddot{q}_n(t)+\omega_n^2 q_n(t)=Q_n(t) \ . \qquad (32)$$

Q_n is the nth generalized force, given by

$$Q_n(t)=\int_0^L f(x,t)\psi_n(x)dx \ , \qquad (33)$$

with $f(x,t)$ being the force density applied to the system.

In particular, a force F applied at the end of the system $x=L$ is represented by generalized forces

$$Q_n=F\psi_n(L) \ . \qquad (34)$$

Then we have, from Eq. (31) (after switching to the frequency domain and explicitly including a damping term),

$$q_n=\frac{F\psi_n(L)}{\omega_n^2-\omega^2+i\phi_n(\omega)\omega_n^2} \ . \qquad (35)$$

Substituting into Eq. (31), we find

$$y(L)=\sum_{n=1}^{\infty}\frac{F\psi_n^2(L)}{\omega_n^2-\omega^2+i\phi_n(\omega)\omega_n^2} \ . \qquad (36)$$

Thus the admittance, $Y=v/F$, is given by

$$Y=\sum_{n=1}^{\infty}\frac{i\omega\psi_n^2(L)}{\omega_n^2-\omega^2+i\phi_n(\omega)\omega_n^2} \ . \qquad (37)$$

(This is just the superposition of the admittances of each of the normal modes.)

From the fluctuation-dissipation theorem [Eq. (16)], we can now find the thermal noise displacement at $x=L$:

$$x^2(\omega)=4k_BT\sum_{n=1}^{\infty}\frac{\psi_n^2(L)\phi_n(\omega)\omega_n^2}{\omega[(\omega_n^2-\omega^2)^2+\phi_n^2(\omega)\omega_n^4]} \ . \qquad (38)$$

This equation can be applied to the internal oscillations of the test mass in a gravitational-wave interferometer. The normal modes of a cylinder with an aspect ratio of order unity have a complicated mode shape.[22] The problem can be treated as one dimensional, weighting each mode by a factor that represents the mean motion of the central part of the front surface of the cylinder along the optic axis. All of the modes with circumferential order greater than zero get zero weight (if the optical axis is aligned with the center of mass), while several of the gravest modes have weights of about unity. The factor $\psi_n^2(L)\approx 2/M$, where M is the mass of the test mass. By design the resonant frequencies ω_n are usually large compared to the frequency of interest, and so we can write

$$x^2(\omega)\approx\frac{8k_BT}{\omega}\sum_{n=1}^{\infty}\frac{\phi_n(\omega)}{M\omega_n^2} \ . \qquad (39)$$

If, for example, $\phi(\omega)$ is constant, then the power spectral density is proportional to ω^{-1}.

It is interesting to consider again how the noise scales with the mass of the pendulum, as we did above for the fundamental mode. Here, in addition to the explicit factor of M^{-1}, the resonant frequencies have an implicit dependence on the mass. For a particular mode in a set of masses of the same aspect ratio, the quantity $\omega a/c$ is a constant, where a is the radius of the mass and c is the speed of sound. Since $M\propto a^3$ for any given material, then $\omega_n^2\propto M^{-2/3}$. Thus we find $x^2(\omega)\propto M^{-1/3}$.

This argument assumes that the loss function $\phi_n(\omega)$ does not itself depend on the size of the mass. That assumption could be false if the dominant loss mechanism were some process involving only the surface of the mass.[11] In such a case, one might expect the loss to decrease as the mass increased, giving a stronger mass dependence to the thermal noise motion.

The transverse modes of a pendulum wire ("violin modes") can also be modeled in this way. If we treat the pendulum as a wire of constant linear mass density ρ with a point mass M attached to the end at $x=L$, then the normalization equation, Eq. (30), becomes

$$\rho\int_0^L\psi_n^2(x)dx+M\psi_n^2(L)=1 \ . \qquad (40)$$

If we neglect the small stiffness of the wire, treating it as a perfectly flexible string under tension Mg, then the squared resonant frequencies are

$$\omega_n^2=\frac{\pi^2 Mg}{\rho L^2}n^2 \ , \qquad (41)$$

while the squared amplitudes are

$$\psi_n^2(L)=\frac{2\rho L}{\pi^2M^2}\frac{1}{n^2} \ . \qquad (42)$$

Thus the thermal noise power spectrum is

$$x^2(\omega)=\frac{8k_BT\rho^2L^3}{\pi^4M^3g}\frac{1}{\omega}\sum_{n=1}^{\infty}\frac{\phi_n(\omega)}{n^4} \ . \qquad (43)$$

Here, as for the fundamental mode of the pendulum, the loss $\phi_n(\omega)$ is only a small fraction of the loss of the wire material itself.

XII. RECOIL LOSSES

If a low-loss oscillator is suspended from a structure with low-Q resonances, then the loss at the resonant frequency may be substantially degraded. This effect may be analyzed in an approximate way by treating the system as a two-mode oscillator of the form shown in Fig. 3, with the resonant mode of the structure nearest in frequency to the mode of the sample playing the role of the upper oscillator. This is the same model discussed above, but here we are interested in what happens to the Q of the lightly damped resonance, instead of in the high-frequency behavior of the mechanical conductance. Thus, for the case of recoil damping as well as for a multimass oscillator, the thermal noise far from resonance may be substantially smaller than would be indicated by a naive interpretation of the damping of the resonance.

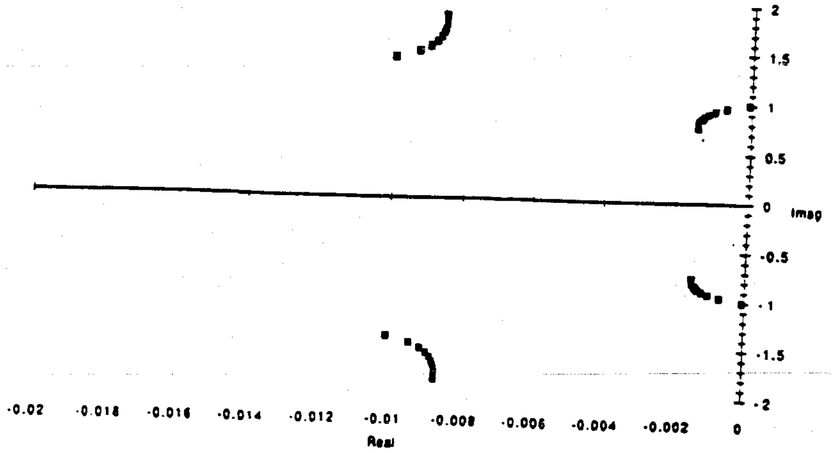


FIG. 5. Portion of the root locus for a double-oscillator model for recoil damping of a high-Q oscillator by resonances in its support structure. The high-Q oscillator has $\omega_2^2=1\ \mathrm{s}^{-2}$ and $\beta_2=10^{-4}$. The structure is modeled as a mode with $\omega_1^2=2\ \mathrm{s}^{-2}$ and $\beta_1=0.02$. The poles of the system are plotted for mass ratio $\mu=0$ (infinite mass structure) to $\mu=1$ (oscillator and structure of equal mass). Adjacent points are separated by $\Delta\mu=0.1$. For clarity, the real and imaginary axes have been drawn to different scales.

The root locus method, a standard tool for the analysis of servomechanisms, is useful for the study of the dependence of the poles of any system on the values of parameters of the system.[23] The two-mode oscillator, analyzed with Laplace transform methods, is characterized by a transfer function

$$\frac{x_2}{F}=\frac{\omega_1^2+\beta_1 s+s^2+\mu(\omega_2^2+\beta_2 s)}{(\omega_1^2+\beta_1 s+s^2)(\omega_2^2+\beta_2 s+s^2)+\mu s^2(\omega_2^2+\beta_2 s)} . \tag{44}$$

The denominator has the same form as that of a servomechanism of loop transfer function

$$G(s)=\mu\frac{s^2(\omega_2^2+\beta_2 s)}{(\omega_1^2+\beta_1 s+s^2)(\omega_2^2+\beta_2 s+s^2)} , \tag{45}$$

where the mass ratio $\mu\equiv m_2/m_1$ plays the role of an adjustable gain. We are interested primarily in how the Q of the lightly damped resonance ω_2 is changed by the recoil of the structure, as parametrized by μ. Figure 5 shows the locus of roots of the system as a function of μ, for one choice of the resonant frequencies and damping parameters.

It is possible to obtain a simple analytic expression for the recoil-damped Q of the oscillator, valid when the structure mass is much greater than the oscillator mass and when the structure has much more damping than the oscillator. The method sketched here makes use of the rules that govern the shape of the root locus in the vicinity of the zero-recoil ("open loop") poles.[23] The interesting features are the departure angle of the locus from the oscillator poles, and the relationship between the parameter μ and the distance traveled along the locus. The locus leaving one of the high-Q poles points almost directly toward the real axis, but has a fractional component of increasing real part of magnitude $\beta_1\omega_2/(\omega_1^2-\omega_2^2)$. The distance traveled along the locus is given by $\mu\omega_2^3/2(\omega_1^2-\omega_2^2)$. Combining these two results, one can show that

$$Q_{2,\mathrm{recoil}}^{-1}\approx Q_2^{-1}+Q_1^{-1}\mu\frac{\omega_1\omega_2^3}{(\omega_1^2-\omega_2^2)^2} . \tag{46}$$

Thus recoil damping is most important when a support structure resonance is close in frequency to the sample resonance.

XIII. MODEL PENDULUM

In this section I estimate the thermal noise displacement power spectrum for a pendulum of a type that might be used as the suspension for the test masses in a gravitational-wave interferometer. A graph is shown in Fig. 6.

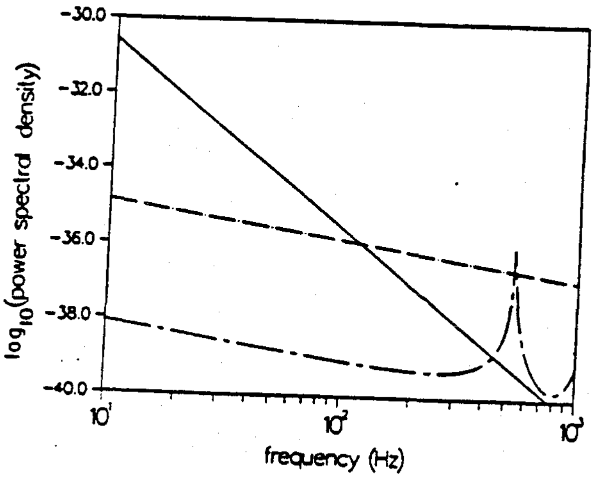


FIG. 6. Thermal noise power spectrum for the model pendulum described in the text. The solid line shows the thermal noise of the fundamental pendulum mode. The dash-dotted line shows the noise from the internal modes of the test mass. The third curve shows the noise from the modes of the pendulum wires. Note that for this graph the frequency axis is given in Hz.

A test mass in such an interferometer might have a mass of 10 kg, supported by four tungsten wires having a length $l=30$ cm. Each wire has a diameter of 1.2×10^{-2} cm, so that it supports half of its breaking stress.[24] Measurements made by Kovalik and Saulson indicate that $\phi_W=1\times10^{-3}$ (roughly independent of frequency) is an upper bound on the losses in tungsten wires.[25] The pendulum should then be characterized by $\phi_p=\phi_w(k_{el}/k_{\text{grav}})=5\times10^{-7}$. The resonant frequency is $\omega_0\approx\sqrt{g/l}=2\pi\times0.9$ Hz. The thermal noise power spectral density, for frequencies large compared to the resonant frequency, is then $x^2(\omega)=(2.7\times10^{-26}$ cm^2/Hz$)(2\pi\text{ s}^{-1}/\omega)^5$.

The test mass could be made of fused silica, with a radius of 10 cm and thickness of 16 cm. This aspect ratio is chosen to make equal the resonant frequencies of the two gravest internal modes (of the required symmetry). These will lie at $\omega_{1,2}=2\pi\times15.4$ kHz. Because the other modes fall at substantially higher frequencies, we can approximate the sum in Eq. (38) by its two lowest terms. If the loss factor appropriate to these resonances is a constant, 2.5×10^{-7} (Ref. 26), then the thermal noise is $x^2(\omega)=(1.4\times10^{-34}$ cm^2/Hz$)(2\pi\text{ s}^{-1}/\omega)$.

The wires of this pendulum have their lowest transverse resonance at about 540 Hz. Below this frequency, the thermal noise from the wires is dominated by the contribution of this resonance. A calculation of the ratio of elastic energy to gravitational energy gives $\phi=\phi_w\times10^{-4}$. The net thermal noise from the wires is $x^2(\omega)=(6.7\times10^{-38}$ cm^2/Hz$)(2\pi\text{ s}^{-1}/\omega)$.

As Fig. 6 shows, the fundamental mode of the pendulum is the dominant source of thermal noise at low frequencies. Above about 100 Hz, the strongest noise is the off-resonant thermal excitation of the lowest resonances of the test mass. The high-Q peaks from the wire resonances will also be visible. In a real gravitational-wave interferometer, seismic noise will probably dominate the noise budget at sufficiently low frequencies. Photon shot noise will be more important than thermal noise at the highest frequencies.[3]

ACKNOWLEDGMENTS

For stimulating discussions, I thank Alex Abramovici, David Bartlett, Peter Bender, Paul Boynton, David Cannell, Ron Drever, A. V. Granato, Joe Kovalik, Bob Krotkov, Peter Nelson, Riley Newman, Ho Jung Paik, Norna Robertson, Clive Speake, Bob Spero, Tuck Stebbins, Robbie Vogt, and Rai Weiss. This work was begun while the author was at the Department of Physics, Massachusetts Institute of Technology, and was supported in part by National Science Foundation Grant No. PHY-8803557. I thank the Joint Institute for Laboratory Astrophysics for financial support while this paper was being completed.

[*] Address starting January 1, 1991: Department of Physics, Syracuse University, Syracuse, NY 13244-1130.

[1] R. V. Jones and C. W. McCombie, Philos. Trans. R. Soc. A 244, 205 (1952).

[2] P. F. Michelson, J. C. Price, and R. C. Taber, Science 237, 150 (1987), and references therein.

[3] R. Weiss, in *Sources of Gravitational Radiation*, edited by L. Smarr (Cambridge University Press, New York, 1979).

[4] P. G. Roll, R. Krotkov, and R. H. Dicke, Ann. Phys. (N.Y.) 26, 442 (1964).

[5] C. C. Speake, Proc. R. Soc. London A414, 333 (1987); H. A. Chan and H. J. Paik, Phys. Rev. D 35, 3551 (1987).

[6] M. C. Wang and G. E. Uhlenbeck, reprinted in *Selected Papers on Noise and Stochastic Processes*, edited by N. Wax (Dover, New York, 1954).

[7] H. B. Callen and R. F. Greene, Phys. Rev. 86, 702 (1952); H. B. Callen and T. A. Welton; *ibid.* 83, 34 (1951).

[8] C. W. Stubbs, Ph.D. thesis, University of Washington, 1988; P. Boynton, in *5th Force and Neutrino Physics*, proceedings of the Twenty-Third Rencontre de Moriond (Eighth Workshop), Les Arcs, France, 1988, edited by O. Fackler and J. Tran Thanh Van (Editions Frontières, Gif-sur-Yvette, France, 1988), pp. 431–44.

[9] H. J. Paik, Ph.D. thesis, Stanford University, 1974.

[10] V. B. Braginsky and A. B. Manukin, *Measurement of Weak Forces in Physics Experiments* (University of Chicago Press, Chicago, 1977).

[11] V. B. Braginsky, V. P. Mitrofanov, and V. I. Panov, *Systems with Small Dissipation* (University of Chicago Press, Chicago, 1985).

[12] R. Weiss, P. S. Linsay, and P. R. Saulson, report, 1983 (unpublished).

[13] A. S. Nowick and B. S. Berry, *Anelastic Relaxation in Crystalline Solids* (Academic, New York, 1972).

[14] A. L. Kimball and D. E. Lovell, Phys. Rev. 30, 948 (1927); W. P. Mason, in *Physical Acoustics: Principles and Methods*, edited by W. P. Mason and R. N. Thurston (Academic, New York, 1971), Vol. VIII, and references therein; C. C. Speake and T. J. Quinn, Report No. BIPM-87/3, 1987 (unpublished).

[15] L. D. Landau and E. M. Lifshitz, *Statistical Physics* (Pergamon, New York, 1980).

[16] J. L. Routbort and H. S. Sack, J. Appl. Phys. 37, 4803 (1966).

[17] C. Zener, *Elasticity and Anelasticity of Metals* (University of Chicago Press, Chicago, 1948).

[18] H. W. Bode, *Network Analysis and Feedback Amplifier Design* (Krieger, Huntington, NY, 1975).

[19] C. Zener, Phys. Rev. 52, 230 (1937); 53, 90 (1938).

[20] R. Weiss (private communication).

[21] L. Meirovitch, *Elements of Vibration Analysis* (McGraw-Hill, New York, 1975). The derivation below closely follows Meirovitch's treatment.

[22] G. W. MacMahon, J. Acoust. Soc. Am. 36, 85 (1964).

[23] G. F. Franklin, J. D. Powell, and A. Emami-Naeini, *Feedback Control of Dynamic Systems* (Addison-Wesley, Reading, MA, 1986).

[24] *American Institute of Physics Handbook*, 3rd ed., edited by D. E. Gray (McGraw-Hill, New York, 1972).

[25] J. Kovalik and P. R. Saulson (unpublished).

[26] Mitrofanov and Frontov (1974), cited by Braginsky, Mitrofanov, and Panov (Ref. 11).

REF. ΓΓ

Physics Letters A 178 (1993) 357–363
North-Holland

PHYSICS LETTERS A

Thermal noise in the test mass suspensions of a laser interferometer gravitational-wave detector prototype

A. Gillespie and F. Raab
LIGO Project, California Institute of Technology, Pasadena, CA 91125, USA

Received 1 April 1993; revised version received 19 May 1993; accepted for publication 19 May 1993
Communicated by J.P. Vigier

The thermal noise of the test mass suspensions of a prototype gravitational-wave interferometer was calculated and found to be in agreement with the measured noise near the resonant frequencies of the suspensions. The damping mechanism of the suspension modes was characterized and found to be nearly independent of frequency.

1. Introduction

A laser interferometer gravitational-wave observatory (LIGO) interferometer [1] consists of a Michelson interferometer with each mirror replaced by a long (4 km) Fabry–Perot optical cavity (see fig. 1). To approximate a free test mass, each mirror of the interferometer is suspended by wires as a pendulum inside of a vacuum vessel. Laser interferometry is used to detect differences between the lengths of the two optical cavities induced by gravitational waves. The principal sources of noise expected to limit the sensitivity of the interferometer to gravitational waves are seismic noise that is transmitted to the test masses, photon shot noise (the uncertainty in the measurement of the differential length due to the quantum nature of light), and thermal noise.

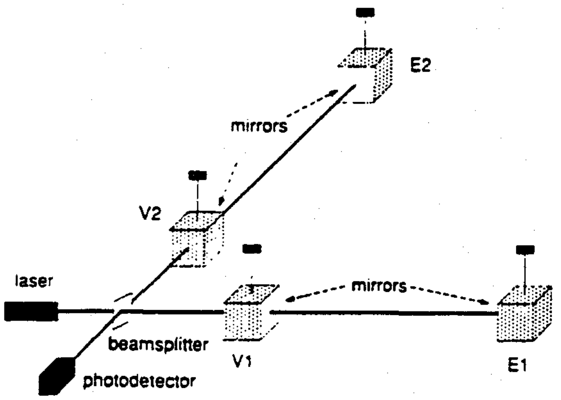


Fig. 1. Schematic view of a LIGO interferometer. E1 (end mirror 1), E2, V1 (vertex mirror 1), and V2 are arbitrary labels for the four mirrors.

Thermal noise in a gravitational-wave detector was first detected in an early resonant-mass detector by Weber [2]. Associated with each mode of oscillation of a physical system in equilibrium with a thermal reservoir is $k_B T$ of thermal energy, of which half will, on average, be kinetic energy and half will be potential energy (k_B and T are the Boltzmann constant and the temperature of the reservoir, respectively). In order to minimize the effects of this thermal noise in a laser interferometer, it is desirable to concentrate the energy in a very narrow frequency band centered on the resonant frequency (that is, to have a large quality factor, Q) of each mode which couples to the motion of the mirrors of the interferometer. These narrow frequency bands can then be filtered out of the gravitational-wave spectrum with negligible loss of observing bandwidth.

The suspension wires of a gravitational-wave detector have several classes of modes which may contribute thermal noise to the interferometer output. We consider here the double wire-loop suspension of the test masses in our 40 m arm length prototype interferometer [3], as shown in fig. 2. In the first class are the pendulum modes. These include a motion along the axis of the incident light, a transverse motion, and a torsional mode. In the next class of modes are the vertical spring modes, where each wire may be thought of as a spring. These include a common mode vertical motion of the mass, a tilt mode, and a roll mode. Of these six modes only the pendulum mode, moving along the axis of the length, will produce an interferometer signal directly; the other modes contribute only if the resonant optical mode is misaligned from the central axis of the mirror [4].

In this Letter we will concentrate on the modes of the suspension in which the wire vibrates like a violin string, henceforth referred to as "violin modes", investigating their contribution to the thermal noise and their role as a diagnostic for the thermal noise in the pendulum mode. The violin modes have two polarizations per wire. These are weakly coupled to the interferometer output in the sense that there is a large mechanical impedance mismatch between the wires and the test mass, but their resonant frequencies lie in the region of several hundred hertz (an important region of the interferometer observational bandwidth), making them both a significant source of thermal noise and a diagnostic for the pendulum thermal noise at frequencies far from the pendulum mode's resonant frequencies.

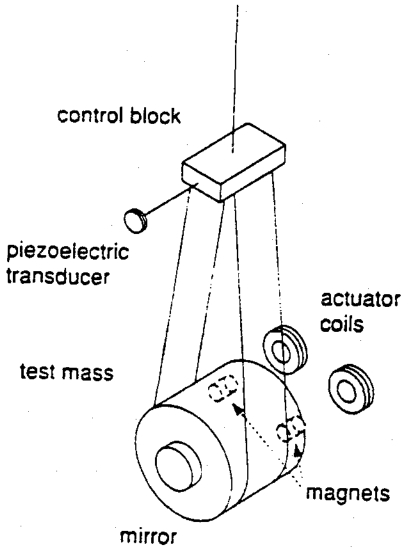


Fig. 2. Test mass suspension details.

2. Violin resonances and thermal noise

By examining the violin resonances in detail one can probe the specific lineshapes of the resonances and test thermal noise models. A general model to describe damping in a harmonic oscillator is a form of Hooke's law where the spring constant is taken to be complex [5],

$$F=-k[1+\mathrm{i}\varphi(\omega)]x\,. \qquad (1)$$

F, x, ω and $k[1+\mathrm{i}\varphi(\omega)]$ are the force, displacement, angular frequency and complex spring constant. The specific damping mechanism is parametrized by the frequency dependence of the imaginary part of the spring constant, $\varphi(\omega)$. Applying the fluctuation-dissipation theorem [6] to this model, a general form for the spectral density of displacement due to thermal noise for a simple harmonic oscillator can be derived [7],

$$\tilde{x}^2(f)=\frac{4k_\mathrm{B}T}{\omega}\frac{k\varphi(\omega)}{(k-m\omega^2)^2+k^2\varphi^2(\omega)}\,. \qquad (2)$$

m and f are the mass and frequency, respectively; $\tilde{x}(f)$ has the dimensions of distance×$\mathrm{Hz}^{-1/2}$.

For the system of a point mass suspended by a single finite mass wire and constrained to move in one dimension (which is mathematically simpler than our four wire system but contains the relevant modes), there are a large number of modes which contribute to thermal noise. The thermal fluctuations of the point mass can be described by a multimode expansion [7,8],

$$\tilde{x}^2(f)=\sum_n\frac{4k_\mathrm{B}T}{\omega\mu_n}\frac{\omega_n^2\varphi_n(\omega)}{(\omega_n^2-\omega^2)^2+\omega_n^4\varphi_n^2(\omega)}\,. \qquad (3)$$

ω_n is the angular resonant frequency of the nth mode; μ_n is the corresponding reduced mass. μ_n is approximately equal to m, the mass, for the pendulum mode ($n=0$), and $\frac{1}{2}m(\omega_n/\omega_\mathrm{p})^2$ for the violin modes ($n>0$); ω_p is the pendulum angular resonant frequency. For the remainder of this section we will discuss the violin modes; the pendulum mode will be discussed in section 4.

It is instructive to examine this model for the familiar case of viscous damping (where the damping acceleration is proportional to the velocity with a proportionality constant γ), commonly used to describe mechanical systems damped by external forces, such as in eddy current damping of moving conductors or as in gas damping of a pendulum. In this case $\varphi_n(\omega)=\gamma\omega/\omega_n^2$ and the resulting lineshape of the nth mode is

$$\tilde{x}_n^2(f)=\frac{4k_B T}{\mu_n}\frac{\gamma}{(\omega_n^2-\omega^2)^2+(\omega\gamma)^2}. \tag{4}$$

A damping mechanism which may be appropriate for the violin modes in a gravitational-wave interferometer corresponds to φ_n being independent of the frequency [9]. We add the assumption that the damping mechanism is the same for each violin mode, which we believe to be approximately true in the case of the harmonics of a thin violin wire. "Thin" means that the potential energy stored in the bending of the wire along its length is negligible compared to the energy in the bending at the endpoints and in the tension of the wire. In the case of a thin wire, the distribution of energy between bending near the ends and tension is the same for all harmonics, and therefore the damping, if its origin is internal to the wire, ought to be the same. In this case, $\varphi=1/Q$ (Q is the mechanical quality factor of a violin mode; all violin modes have the same Q) and the thermal noise lineshape is

$$\tilde{x}_n^2(f)=\frac{4k_B T}{\omega\mu_n}\frac{\omega_n^2/Q}{(\omega_n^2-\omega^2)^2+\omega_n^4/Q^2}. \tag{5}$$

Notice that when ω is near a particular ω_n the two lineshapes given by formulae (4) and (5) are approximately equal ($Q_n=\omega_n/\gamma$ for viscous damping). In fact in the limit that Q_n becomes very large, all lineshapes converge to the same shape near ω_n for any choice of $\varphi_n(\omega)$ which varies smoothly over the bandwidth ($\Delta\omega=\omega_n/Q_n$) of the system. With the 40 m interferometer the thermal noise of the violin modes dominated other noise only in a narrow band centered on the violin resonance frequencies, and therefore the noise spectrum of the interferometer did not directly indicate the thermal noise lineshapes or the nature of the damping mechanism. However the Q of the lower few harmonics of the violin modes could be measured; by examining the frequency dependence of the Q of those harmonics, one can assign some frequency dependence to φ.

3. Measurements on the 40 m interferometer

The Q of individual suspension wires in the 40 m interferometer were measured by exciting the violin resonance, turning off the excitation, and measuring the decay times ($Q_n=\frac{1}{2}\omega_n\tau$; τ is the measured amplitude decay time). Magnets and actuator coils were attached to the end test masses (shown in fig. 2) for applying calibration signals and to maintain resonance between the cavities and the light. On all masses a piezoelectric transducer was normally used to damp the residual pendulum motion. For measurements on the end masses the driving signal was applied to the actuator coils. The vertex masses were driven using the damping transducers on their control blocks. To verify that neither the pendulum damping transducers nor the interferometer control servomechanisms damped the wire resonances, the decays were measured with and without servomechanism signals applied to the actuators.

Table 1
The measured Q of the violin resonances.

Test mass	Frequency (Hz)	Q	Q of nth harmonic; n
E1	319.65	13000	
	324.90	16000	19000; 2
	326.08	19000	
	328.45	15000	16000; 2
V1	594.35	240000	
	596.68	280000	260000; 2
			220000; 3
			220000; 4
	598.15	43000	
	605.02	110000	
E2	505.85	66000	
	506.88	120000	110000; 2
	512.85	23000	23000; 2
	514.90	16000	
V2	592.70	295000	
	592.80	295000	
	596.42	356000	
	600.22	163000	

The frequencies and the Q of the fundamental violin modes of the suspension systems of the 40 m interferometer are given in table 1. The uncertainties in the Q measurements are approximately 1%. The differences in frequencies among the different masses are explained by differences in the suspensions (which arose through various modifications to the interferometer). One end mass (E1) uses 150 μm diameter wire. The other end mass (E2) uses 100 μm diameter wire. Both of the vertex masses use 75 μm diameter wire. Steel music wire is used for all masses; the length of all the suspension wires is 25 cm, and each test mass is 1.5 kg. For a thin wire the resonant frequency of the nth mode is given by

$$f_n = n/ld\sqrt{T/\pi\rho}\,,$$

where l and d are the length and diameter of the wire, T is the tension, and ρ is the density of the wire material.

The differences in the Q of the violin modes for the different masses might partly depend on the wire diameters, which affect both the tensile stress and the stiffness of the wires. However there were also significant variations in the connections of the wires to the masses. In all cases a small glass wedge was used to define the point where the wire connects to the test mass, and a similar metal wedge was used where the wire meets the control block. Unfortunately fine details (such as how the wedges and wires were bonded to the mass) varied among the wires. Furthermore, the end masses and the vertex masses had different types of control blocks, for historical reasons.

The Q of the second harmonics of five wires were measured and are also shown in table 1 – two wires from E1, two wires from E2 including both a high and a low Q wire, and a high Q wire from the vertex mass V1. In addition two additional higher order harmonics of the V1 wire mere measured. The Q of all harmonics were the same as the corresponding fundamental resonance Q within 25%. This result is consistent with a damping model which is independent of frequency in the range of hundreds of hertz. We will adopt this model for the remainder of this Letter.

4. Estimate of the damping and noise of the pendulum mode

A full description of the noise of the suspension must include the thermal noise from the pendulum mode. Due to the impedance mismatch between the wires and the test mass, the violin resonances contribute to the overall noise spectrum only in narrow bands (tenths of hertz) centered at their resonant frequencies. The thermal noise of the pendulum mode, which couples directly to the interferometer output, can be the principal noise at frequencies of order 100 Hz, a region of peak interest for LIGO interferometers. Predicting this thermal noise requires knowing the damping of the pendulum mode, $\varphi_{\rm p}(\omega)$ at these frequencies. Since direct measurement of th losses of a high Q mechanical system far from its resonant frequency can be extremely difficult [5], we use the Q of the violin modes to estimate the damping of the pendulum at frequencies of hundreds of hertz.

If we assume that all of the losses in both the pendulum and the violin modes are concentrated near the endpoints of the wire, where the bending is most severe, we can estimate the pendulum damping directly from the violin damping. These endpoint losses need not be restricted to intrinsic bending losses in the wire itself, but may also include losses due to flexing or friction in the clamps or the points of attachment at either the top or the bottom of the wire. The model only requires that the losses be associated with the motion of the wire in the region near the clamping through some angle θ from the vertical equilibrium position. This calculation assumes that there is some loss of energy per cycle $\Delta E(\theta)$ whenever the wire is bent through some angle θ. The strategy is to calculate the total energy of both the pendulum ($E_{\rm p}$) and the violin ($E_{\rm v}$) modes as functions of θ, and by comparing $E_{\rm p}(\theta)/\Delta E(\theta)$ with $E_{\rm v}(\theta)/\Delta E(\theta)$ to deduce the relationship between their respective damping.

The energy in the pendulum mode is, to second order in θ,

$$E_{\rm p}(\theta) \approx \tfrac{1}{2} mgl\theta^2\,. \qquad (6)$$

g is the acceleration due to gravity and l is the equilibrium length of the loaded wire. If we approximate the wire as having no stiffness, which is reasonable for the parameters of the 40 m interferometer test mass suspensions as far as the total energy of the lower order violin modes is concerned, the potential energy of the violin modes is stored in the tension of the wire, and the shape of the modes is sinusoidal. In this case, the energy of the violin mode is, to second order in θ,

$$E_v(\theta) \approx \tfrac{1}{4} T l \theta^2 . \qquad (7)$$

T is the tension of the wire. This result is independent of which harmonic is chosen. The pendulum has the same amount of energy at a given angle regardless of the number of wires. Since each wire in a four wire system is only supporting one quarter of the mass, the tension of each wire in our system is $\frac{1}{4}mg$, and $E_v = \frac{1}{16} mgl\theta^2$. Therefore for a given angle θ, the pendulum mode has eight times as much energy as a violin mode, $E_p = 8E_v$.

For the violin modes, the loss of energy per cycle, ΔE_v, is given by

$$\Delta E_v(\theta) = 2\pi E_v(\theta)\varphi_v . \qquad (8)$$

If the losses for both the violin mode and the pendulum mode are primarily in the bending of the wire, then the losses in the pendulum would equal the losses in the four wires,

$$\Delta E_p(\theta) = \Delta E_{v1}(\theta) + \Delta E_{v2}(\theta) + \Delta E_{v3}(\theta) + \Delta E_{v4}(\theta) . \qquad (9)$$

We have assumed that the mass is constrained by the four wires so that it does not rotate and the wire bends at both ends in the pendulum mode. We can then estimate the damping of the pendulum as

$$\varphi_p = \tfrac{1}{8}(\varphi_{v1} + \varphi_{v2} + \varphi_{v3} + \varphi_{v4}) . \qquad (10)$$

With this estimate of the damping, the thermal noise due to the pendulum mode at frequencies near the violin mode resonant frequencies can be derived by substituting $\varphi(\omega) = \varphi_p$ into the general thermal noise lineshape (eq. (2)), giving

$$\tilde{x}_p^2(f) = \frac{4k_B T}{m\omega} \frac{\omega_p^2 \varphi_p}{(\omega_p^2 - \omega^2)^2 + \omega_p^4 \varphi_p^2} . \qquad (11)$$

This result depends on the wires being sufficiently thin that essentially all of the bending occurs near the ends, an approximation which is true for the physical parameters of this prototype interferometer. LIGO interferometers will use larger masses which will require correspondingly thicker wires; these wires may have non-negligible stiffness. In this case a more detailed analysis treating the finite stiffness of the wire is required. but the principle of estimating the total thermal noise of the suspension system from the Q of the violin modes can still be applied [10].

5. Comparison of estimated suspension thermal noise with experiment

The lineshapes of the violin resonances of mass E1 were compared with the thermal noise prediction. The thermal noise lineshape of a pendulum with a finite mass single wire is given in eq. (3). For a four-wire system the lineshape of an individual violin mode becomes

$$\tilde{x}_n^2(f) = \frac{2k_B T}{\omega m} \frac{\omega_n^2/Q}{(\omega_n/\omega_p)^2[(\omega_n^2 - \omega^2)^2 + \omega_n^4/Q^2]} . \qquad (12)$$

This comparison is shown in fig. 3. Next to two of the resonances are smaller peaks. We believe that they are polarization modes of the wires which are nearly but not completely orthogonal to the optical axis. A

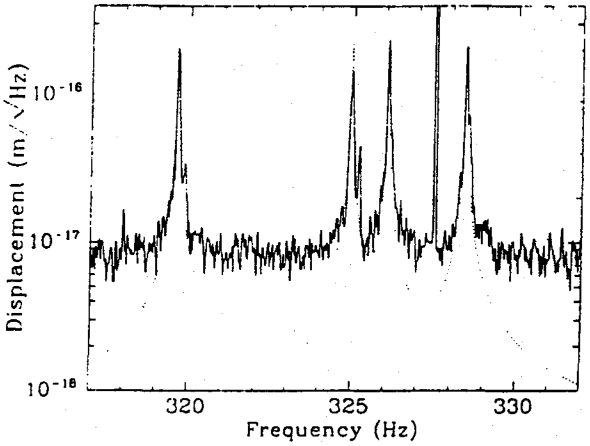


Fig. 3. E1 violin resonances. The solid line is the interferometer noise spectrum; the dashed line is the thermal noise prediction. The large peak at 327.5 Hz is a calibration signal. The thermal lineshape has been averaged over each 0.025 Hz bandwidth channel to produce the prediction.

finer comparison between theory and experiment is given in fig. 4 where the data points (for the middle resonances in fig. 3) are plotted on a linear scale with their associated error bars. The uncertainties in the data points are statistical errors due to the limited number of averages in the power spectrum. The systematic uncertainty of the calibration is 10%. The uncertainty in the theoretical curve due to the uncertainties in the Q is smaller. The agreement with the thermal noise calculated using the measured resonant frequencies and the Q is quite good. The predicted rms fluctuations of the test mass corresponding to a single violin resonance is about 0.06 fm.

Using the measured Q and assuming that the pendulum mode is damped as derived above, the total thermal noise of the wire suspensions can be predicted,

$$\tilde{x}^2(f) = \sum_{4\ \text{masses}} \left(\tilde{x}_{\mathrm{p}}^2(f) + \sum_{4\ \text{wires}} \sum_{n} \tilde{x}_n^2(f) \right) . \qquad (13)$$

Figure 5 shows this thermal noise prediction and compares it to the measured interferometer noise of June 1992. As can be seen from the figure, thermal noise contributed to the 40 m interferometer by the suspension wires is much smaller than other noise sources except in narrow frequency bands centered on the resonant frequencies of the violin modes. The increase in the predicted suspension noise below 200 Hz is due to the pendulum mode. The resonances near 320 Hz are excited above thermal noise by a factor of two by excess noise in the current version of the pendulum damping control electronics; when that control system is turned off (which degrades the overall performance of the interferometer) those peaks are at thermal noise (see fig. 3). The peaks in the noise between 1800 and 2000 Hz which appear to coincide with the violin resonances are due to residual frequency noise from the laser, not due to the violin resonances.

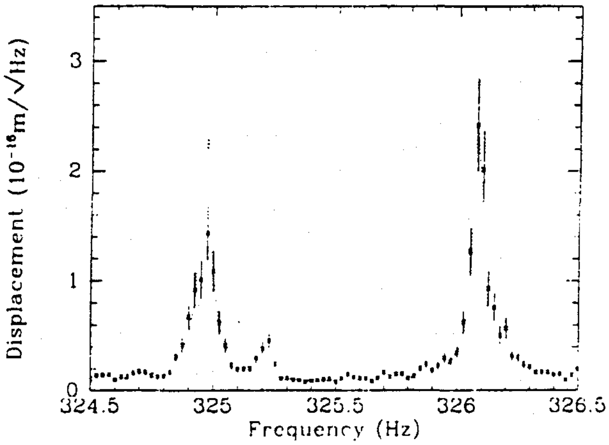


Fig. 4. This spectrum is an expansion of fig. 3 plotted on a linear scale to emphasize the peaks and show the detail.

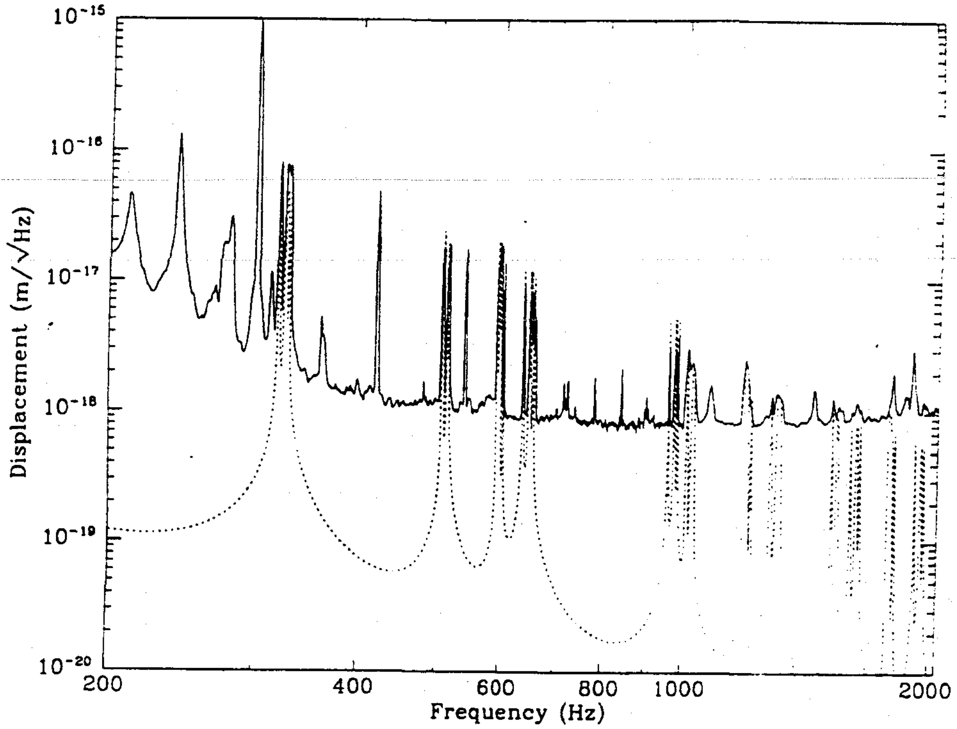
6. Conclusion

We have observed thermal noise of the violin modes near resonance in a 40 m interferometer. W have used the measured Q of the harmonics of th violin modes to conclude that the primary damping mechanism of the violin modes is frequency independent. Taking that conclusion and the assumption that the losses in the suspension occur primarily near the endpoints of the wires, we have estimated the total suspension thermal noise spectrum in the region of hundreds of hertz of the LIGO 40 m prototype interferometer.

The estimated suspension thermal noise background for the 40 m interferometer was dominated by the low Q wires. Our data indicate that wire Q of at least 3×10^5 are achievable using the basic design of the current suspension. If this Q were achieved on all wires, the resulting thermal noise would be $\sim2\times10^{-19}$ m/$\sqrt{\text{Hz}}$ at 100 Hz, comparable to the goal for initial LIGO interferometers [1]. Understanding the variations in the wire Q and the scaling with tension and wire diameter is the subject of an ongoing investigation. Similar studies can be used to characterize the expected thermal noise from candidate suspension materials which have shown much higher Q values in low frequency oscillators [11].

Acknowledgement

We thank R. Vogt and the entire LIGO team for making this work possible. We are especially grateful to P. Saulson, R. Weiss and S. Whitcomb for stimulating conversations and M. Zucker and S. Kawamura for their help in setting up the measurements. This work was supported by the National Sci-



ig. 5. The solid line is the interferometer noise; the dashed line is the thermal noise prediction. The data were taken at two different andwidths: 1.25 Hz from 100 to 1000 Hz, and 6.25 Hz from 1000 to 2000 Hz. The thermal noise prediction was generated by averaging e lineshape over the appropriate bandwidth. Due to this averaging the violin modes appear broadened.

nce Foundation cooperative agreement no. PHY-210038.

eferences

1] A. Abramovici, W.E. Althouse, R.W.P. Drever, Y. Gursel, S. Kawamura, F.J. Raab, D. Shoemaker, L. Sievers, R.E. Spero, K.S. Thorne, R.E. Vogt, R. Weiss, S.E. Whitcomb and M.E. Zucker, Science 256 (1992) 325.

2] J. Weber, Phys. Rev. Lett. 17 (1966) 1224.

3] M.E. Zucker, in: Proc. sixth Marcel Grossman Meeting on general relativity, ed. H. Sato (World Scientific, Singapore, 1993) Vol. 1, p. 224.

[4] S. Kawamura and M.E. Zucker, Mirror orientation noise in a Fabry–Perot interferometer gravitational wave detector, private communication (1993).

[5] A.S. Nowick, and B.S. Berry, Anelastic relaxation in crystalline solids (Academic Press, New York, 1972).

[6] H.B. Callen and T.A. Welton, Phys. Rev. 83 (1951) 34.

[7] P. Saulson, Phys. Rev. D 42 (1990) 2437.

[8] N. Mio, Japan. J. Appl. Phys. 31 (1992) 1243.

[9] J. Kovalik and P. Saulson, Mechanical loss in fibers for low noise pendulums, private communication (1993).

[10] G. Gonzalez and P. Saulson, Brownian motion of a mass suspended by an anelastic wire, private communication (1993).

[11] V.B. Braginsky, V.P. Mitrofanov and O.A. Okhrimenko, Phys. Lett. A 175 (1993) 82.

Ref. QQ

UNPUBLISHED DRAFT

Thermally excited vibrations of the mirrors of laser interferometer gravitational-wave detectors

Aaron Gillespie and Fred Raab

May 10, 1994

LIGO Project, California Institute of Technology, Pasadena, CA 91125, USA

Abstract

The effect of thermally excited mirror vibrations on length measurements using laser interferometers is calculated, and the number of vibrational modes which must be included to predict the total thermal noise in an interferometer is estimated. The vibrational modes are found to be more strongly coupled to the length measurements and more modes are found to be needed than previously thought based on simpler models, resulting in thermal noise estimates which are larger than previous estimates made using simpler models. The thermal noise due to mirror vibrations is estimated for interferometers with parameters relevant to the LIGO (Laser Interferometer Gravitational-wave Observatory) interferometers, and is found likely to be a significant contributor to the background noise.

I. Introduction

Gravitational-wave observatories, such as the LIGO [1] and VIRGO [2] facilities now being built, will use laser interferometry to sense the small motions of suspended test masses induced by gravitational waves. To observe signals from astrophysical sources, background motion of the test masses caused by local environmental forces must be minimized. Once background noise from seismic, acoustic, and electromagnetic forces are sufficiently well suppressed, the main background arises from thermally excited motion of the masses. Motion of the center of mass of the test mass, typically referred to as suspension thermal noise, has been treated elsewhere [3]. Here we concentrate on the internal vibrations of the test mass, which cause motion of the mirrored surfaces relative to the centers of mass of the test masses.

We present calculations of the optical path length changes induced in an interferometer by mirror vibrations and predict the spectral distribution of the noise in the interferometer due to thermally driven vibrations. Previous estimates [4,5] of the thermal noise in laser interferometers have only included the lowest few vibrational modes of the mirrors and have estimated the coupling between the vibrations of the mirror and the optical path length by treating the mirrors as vibrating in only one dimension (essentially equivalent to treating the mirrors like long, thin rods). In the calculations presented below, the mirrors are treated as three dimensional bodies, and the number of modes needed to describe the thermal noise accurately are determined. The vibrational thermal noise of the Laser Interferometer Gravitational-Wave Observatory (LIGO) project's first generation interferometer and 40–meter prototype interferometer are investigated as examples; however the method is general and can be applied to any interferometer with suspended mirrors. We find both that more modes must be considered to estimate accurately the thermal noise and that the modes are generally more strongly coupled to the interferometer than was previously thought. The result is that the thermal noise due to mirror vibrations is larger than previously estimated and may limit the sensitivity of advanced gravitational-wave detectors.

Our calculation of thermal noise also has one assumption which differs from the assumptions which were used in many previous thermal noise estimates. That assumption concerns the frequency dependence of the loss function of the test mass, discussed in reference [5]. Current experimental results [6] indicate that the loss function which is currently achievable in fused silica, the preferred test mass material for both optical and mechanical reason, may be approximately independent of frequency. Many previous calculations of thermal noise assumed that the loss function in fused silica was viscous in nature. Our method of calculation is independent of the specific loss function chosen, but the absolute level of the estimate of thermal noise is sensitive to the loss function. A frequency independent loss function leads to larger estimates of thermal noise in the vibrational modes of the test mass below the resonant frequencies of the modes than a viscous loss function.

The effect of mirror vibrations on the optical mode of an interferometer and the implications for length measurements are treated in Section II. We define a single parameter, the effective mass coefficient, which can be used to parametrize this interaction for any vibrational mode of the test mass. Our method for calculating these coefficients was verified as described in Section III. Section IV describes how thermal noise in an interferometer is calculated as a summation of test mass modes and Section V explains how physical parameters of the interferometer affect the convergence of this sum. The sensitivity of these thermal noise estimates to the exact centering of the light on the test mass is investigated in Section VI. Section VII discusses how the results of these calculations can be used to estimate thermal noise in real mirrors, where deviations from perfect symmetry may cause mixing of the vibrational modes. The implications of this work for gravitational-wave detectors are discussed in Section VIII.

II. Effect of mirror vibrations on optical modes of an interferometer

A LIGO interferometer [1], shown schematically in figure 1, will consist of suspended test masses forming Fabry-Perot optical cavities arranged along orthogonal axes which sense the strain induced by a passing gravitational-wave. The strain is manifested as apparent displacements of the cavity mirrors, detected by laser light which resonates in the optical cavities. Vibrations of the mirrors constitute a noise background which could mask or mimic a gravitational-wave signal.

We wish to find the displacements detected by the laser light when the mirror surfaces are vibrating. To solve this problem, we must consider the interaction between two different types of modes: the mechanical modes of the vibrating test masses, and the optical modes of the electromagnetic field resonating in the Fabry-Perot cavities.

The vibrational eigenmodes and eigenfrequencies of a free right solid cylinder can be found by solving the equations of elasticity using an analytic series solution [7]. From the results of that solution, the amplitude of the displacement of the mirror surface at each point, $\vec{u}_n(\rho,\theta)$, can be calculated for the n$^{\text{th}}$ mode normalized to a fixed energy, U.

The optical modes of the interferometer can be described by Hermite-Gaussian functions, ψ_{lm} [8]. The interferometer typically operates with the TEM$_{00}$ mode, ψ_{00}, on resonance. To avoid interference between modes, the length of the interferometer and the curvature of the mirrors are chosen such that the TEM$_{00}$ mode and other higher order modes cannot resonate simultaneously [9].

The optical mode experiences a phase shift upon reflection from a mirror excited in a particular vibrational mode:

$$\psi_{00}(\rho,\theta,z) \rightarrow \psi_{00}(\rho,\theta,z)e^{i2\vec{k}\cdot\vec{u}_n(\rho,\theta)} \approx \psi_{00}\left[1 + i2\vec{k}\cdot\vec{u}_n - 2\left(\vec{k}\cdot\vec{u}_n\right)^2\right]. \quad (1)$$

$\vec{k}$ is the wave vector, and $\left|\vec{k}\cdot\vec{u}_n\right|$ is taken to be much less than unity for all points on the mirror surface. This new perturbed mode in the interferometer can be described in terms of the unperturbed modes:

$$\psi = \sum_i \sum_j c_{ij}\psi_{ij} \quad (2)$$

$$c_{ij} = \int_S \psi^*_{ij}\psi_{00}e^{i2\vec{k}\cdot\vec{u}_n(\rho,\theta)}d\sigma \quad (3)$$

The integral is an area integral over the mirror surface, S. Since only the TEM$_{00}$ component of the perturbed light resonates, the new resonating mode can be written as

$$\psi \approx \psi_{00}\left[1 + i2\int_S \psi^*_{00}\psi_{00}\vec{k}\cdot\vec{u}_n d\sigma - 2\int_S \psi^*_{00}\psi_{00}\left(\vec{k}\cdot\vec{u}_n\right)^2 d\sigma\right]. \quad (4)$$

Each term of this expression can be easily interpreted. The imaginary term contains the phase shift from which the apparent length change Δl_n can be determined,

$$\Delta l_n = \frac{\int_S \psi^*_{00}\psi_{00}\vec{k}\cdot\vec{u}_n(\rho,\theta)d\sigma}{\left|\vec{k}\right|}. \quad (5)$$

The second integral term describes the light which is scattered out of the TEM$_{00}$ mode. The apparent motion, Δl_n, is generally different for different vibrational modes of the mirror with the same energy.

To remove the dependence of the amplitude of the displacements, $\vec{u}$, on the energy normalization, it is convenient to parametrize the coupling of each mode in terms of an effective mass coefficient, α_n, defined as

$$\alpha_n = \frac{U}{\frac{1}{2}m\omega_n^2\Delta l_n^2}. \quad (6)$$

m and ω_n are the actual mass of the mirror and the angular resonant frequency of the vibrational mode, respectively. With this parametrization, the apparent motion of the illuminated mirror surface oscillating in a particular vibrational mode can be modeled as if it were a point mass of magnitude $\alpha_n m$ vibrating with a resonant frequency ω_n.

As an example, the mode shapes, resonant frequencies, and effective mass coefficients of the first six axisymmetric modes of a "prototype" mirror are shown in figure 2. A prototype mirror (so designated because such a mirror is used in the LIGO project's 40–meter prototype interferometer) is a fused silica cylinder with a diameter of 10 cm, a length of 8.8 cm, and a mass of 1.6 kg. The effective mass coefficients depend not only on the parameters of the mirror, but also on the geometry of the optical mode on the mirror surface. For all the calculations of this paper, except in Section V where it is explicitly stated otherwise, the laser parameters used with the prototype mirror are those relevant to that prototype, which has a spot size (the radius at which the intensity is $1/e^2$ of its maximum) of 0.22 cm. The beam is assumed to be centered on the mirror and in the TEM$_{00}$ optical mode, so that non-axisymmetric vibrational modes do not contribute apparent motion to the mirror surface.

The effective mass coefficients of the first 100 axisymmetric modes of the prototype mirror are plotted against their respective resonant frequencies in figure 3. The effective mass coefficients vary by several orders of magnitude, reflecting the wide variety of modes shapes. There is, however, a general trend toward lower effective masses at larger resonant frequencies. To draw attention to this trend, a dashed line representing $\alpha_n \propto f_n^{-1}$ is drawn. This line is approximately the median effective mass coefficient in a given bandwidth. The significance of this line will become apparent in Section V. This general trend arises from the cylindrical symmetry of the axisymmetric modes, which dictates that the largest antinode of motion in the axial direction is at the center of the mirror, the position being sampled by the laser. Hence the apparent motion of the mirror can be relatively larger for the higher frequency modes, and the effective mass can be correspondingly smaller. To illustrate this point, the shapes of the mirror surface for four modes with small effective masses are shown in figure 4 (these modes are shown as dots in figure 3); notice that a relatively small portion of the mirror around the center of the mirror surface moves much more than any other spot. The benefit of this choice of beam location is that the center of the mirror is a node for all non-axisymmetric modes, making calculations much simpler by decreasing the number of modes involved. The implications of moving the beam spot off center are discussed in section VI.

Previous estimates of the coupling between the vibrational modes of the mirror and the optical path length of the interferometer were equivalent to treating the mirror as a long, thin rod, making the problem one dimensional (the axial dimension). In this case the mode shapes can be described simply by

$$u_n(z) = u_0 sin\left(\frac{n\pi z}{2h}\right), \tag{7}$$

where h is the thickness of the mirror. The resulting approximation predicts an effective mass coefficient of 0.5 for the lowest longitudinal mode (the 30 kHz, α = 0.34 mode in figure 2). The first drum mode (the 30 kHz, α = 0.59 mode in figure 2), which does not exist in a one dimensional system, is then assumed to have the same effective mass coefficient of 0.5. The one dimensional approximation gives a reasonable estimate of the effective mass coefficients of these low frequency modes. However, such a model fails to predict the much lower effective masses of some of the higher frequency modes that are shown in figure 3, and therefore can underestimate the total noise contributed by mirror vibrations.

To check the self-consistency of this model, the approximations that went into its formulation must be examined. The first assumption is that $\vec{k}\cdot\vec{u}_n << 1$. This is approximately the same as requiring that $\left|\vec{k}\right|\cdot\Delta l_n << 1$. Since we are supposing that the vibrations are thermally excited, it follows from the equipartition theorem that

$$\Delta l_n \approx \sqrt{\frac{k_B T}{\alpha_n m\omega_n^2}} \tag{8}$$

where k_B and T are Boltzman's constant and the temperature, respectively. Hence at a temperature of 300 K, $\vec{k}\cdot\vec{\mu}_n$ is of order 10^{-10} for the lower frequency modes shown in figure 2, a phase shift which can be detected by precision interferometry.

The second integral term of equation (4) which describes the scattered light is of order 10^{-20} for the modes of figure 2. This is small compared to the stationary scattering due to microroughness and figure errors in the mirror, which will be of order 10^{-4}. Since the scattering from mirror vibrations is frequency dependent, a comparison can be made between the uncertainty in the amplitude of the light in Fabry-Perot cavities due to photon shot noise and the scattering from mirror vibrations. An order of magnitude estimate of the bandwidth of the mirror modes is 1 mHz, and available laser light (including the gain from the Fabry-Perot optical resonator) is of order 1000 Watts. The corresponding shot noise in the light power is of order 10^{-12}, eight orders of magnitude larger than the light scattered by mirror vibrations. Clearly the scattering term and all higher order terms of equation (4) can safely be ignored.

Another assumption is that the TEM$_{00}$ component of the distorted optical mode reflected from the vibrating mirror surface still resonates in the Fabry-Perot cavity, and that the light scattered into other modes does not. This is true if the change in the resonant frequencies of the optical modes of the cavity due to the vibrations of the mirror is less than the linewidth of the cavity. The frequency shift, $\Delta f = f\Delta l/l$, is of the order of 10^{-3} Hz for the modes of figure 2, and the linewidth of the optical cavity under consideration is of order 10^3 Hz, so the light continues to resonate in the primary mode, and modes which were previously separate from the primary mode do not resonate.

III. Verification of the effective mass coefficients

The numerical code used to calculate the effective mass coefficients was subject to a number of consistency checks. The resonant frequencies were calculated using a FORTRAN code largely provided by J.R. Hutchinson [7]. The calculated eigenfrequencies agreed with that theoretical work and also the experimental work of McMahon [10], for the appropriate cylinder materials and dimensions. The mode shapes were checked for self consistency by comparing the elastic energy of deformation in the mode to the kinetic energy in the mode one quarter of a cycle later:

$$\int_M \left[\frac{1}{2}Ku_{ll}^2 + \mu\left(u_{ik} - \frac{1}{3}\delta_{ik}u_{ll}\right)^2\right]dV = \int_M \frac{1}{2}\rho\omega_n^2\vec{u}\cdot\vec{u}dV. \tag{9}$$

K, μ, δ_{ik}, and ρ are the bulk modulus, shear modulus, Kronecker delta, and the density of the fused silica. The integral is over the mirror volume, M, and the displacement vector, $\vec{u}$, is now evaluated over the entire mirror volume. u_{ik} is the strain tensor, defined in terms of the displacement vector as

$$u_{ik} = \frac{1}{2}\left(\frac{du_i}{dx_k} + \frac{du_k}{dx_i}\right). \tag{10}$$

Repeated indices are summed. The effective mass coefficients could then be checked in the one dimensional approximation where the mirror is made very long and thin (equation (7)). The effective mass coefficients of all modes in this approximation are 0.5. The numeric code passed these basic checks.

A simple experiment can be used to verify directly the calculation of the effective mass coefficients for modes which have acoustic wavelengths much larger than the beam spot size. These modes can be driven by gluing a small magnet (with dimensions much smaller than the acoustic wavelength of the mode) to the center of the back of the mirror, which by symmetry has the same effective mass coefficient as the front. Using a Michelson interferometer to measure the mirror's response to forces applied by a current in a coil near the magnet, one can experimentally infer the effective mass of the mode. The apparent motion on resonance is

$$\bar{x} = \frac{Q_n}{\alpha_n m \omega_n^2} \bar{F} \tag{11}$$

where $\bar{x}$ and $\bar{F}$ are the root mean squared displacement and force, and Q_n is the mechanical quality factor of the resonance. The force is proportional to the current I through the drive coils; therefore

$$\alpha_n \propto \frac{Q_n}{m\omega_n^2} \frac{\bar{I}}{\bar{x}}. \tag{12}$$

Such an experiment was carried out using the prototype mirror, investigating the first five modes of figure 2. The resonant frequencies were found to agree with the calculation within 2%; the mirror had a wedged shape for optical reasons which made the thickness of the mirror ill-defined at the 1% level. Figure 5 shows a comparison of the measurement of the effective mass coefficients for these five modes to the values of α_n calculated using equation (12). The line on the figure indicates a fit to a direct proportionality (the current to force ratio was not calibrated independently). Figure 5 indicates that the effective mass coefficients are an accurate way to model the vibrational modes of a mirror, and that the numeric code used in the calculations was functioning properly.

IV. Spectral Density of the Thermally Excited Motion

The root mean squared motion of a mode of the thermally excited mirror can be calculated using the equipartition theorem as in equation (8). At a temperature of 300 K, Δl_n is of order 10^{-16} meters for the lower frequency modes shown in figure 2. As a first approximation, most of the energy of the motion occurs within a bandwidth around the resonant frequency defined by the Q ($\Delta f_n = f_n/Q_n$). To get a more complete prediction of the spectral density of the motion, the Fluctuation-Dissipation theorem must be used [11]. The vibrational mode is modeled as a harmonic oscillator with a complex spring constant:

$$-\alpha_n m w^2 \tilde{x} + \alpha_n m \omega_n^2 [1 + i\varphi_n(\omega)]\tilde{x} = \tilde{F} \tag{13}$$

where the dissipation is parametrized by the imaginary part of the spring constant and can, in general, be frequency dependent. We refer to $\varphi_n(\omega)$ as the loss function because the fraction of energy lost in one cycle of oscillation at frequency ω is $2\pi\varphi(\omega)$.. The loss function at the resonant frequency is related to the Q of the mode by $\varphi_n(\omega_n) = 1/Q_n$. From the equation of motion (13) and the Fluctuation-Dissipation theorem, the general spectral density of displacement due to thermal excitation of the n$^{\text{th}}$ mode can be derived [5]:

$$S_{xn}(f) = \frac{4k_B T}{\alpha_n m \omega}\left[\frac{\omega_n^2 \varphi_n(\omega)}{(\omega^2 - \omega_n^2)^2 + \omega_n^4 \varphi_n^2(\omega)}\right]. \tag{14}$$

In this form, the spectral density is implicitly a function of the natural frequency, f in that the motion is described over the usual one hertz bandwidth; however the function is written explicitly in terms of the angular frequency, ω. This notation, although perhaps initially confusing, simplifies the equation in that it eliminates an abundance of 2π's. The frequency band of interest for earth based gravitational-wave detectors (1 Hz to 10 kHz) is generally well below the resonant frequency of the lowest vibrational mode of the mirror, in which case the total thermal noise can be approximated by

$$S_x(f) \approx \sum_n \frac{4k_B T}{\alpha_n m \omega_n^2}\frac{\varphi_n(\omega)}{\omega}. \tag{15}$$

All of the parameters of equations (14) and (15) can be readily measured or calculated except the loss function. Measuring the loss function for a system with low dissipation can be extremely difficult. What one generally relies on are measurements of the loss function at the resonant frequency and some dissipation model which predicts the general frequency dependence. Q values of order 10^7 have been measured for the vibrational resonances of the "prototype" mirror.

Furthermore fused silica oscillators with Q's of order 10^7 and resonant frequencies ranging from 1 Hz to 10 kHz have been built [6], suggesting a φ which is of order 10^{-7} and approximately independent of frequency is achievable in fused silica mirrors. This estimate should not be taken as either the fundamental level or frequency dependence of the dissipation of fused silica, which is unknown, but rather a summary of what has been observed thus far. The dissipation depends on both the purity and the preparation of the sample, and there is no strong reason to believe that the dissipation could not be reduced if sufficient care were taken.

Many previous estimates of thermal noise from the vibrational modes of the test mass have assumed that the dissipation were viscous in nature. In this case the loss function is

$$\varphi(\omega) = \frac{1}{Q}\frac{\omega}{\omega_0}. \tag{16}$$

We do not believe that current data support this model, which would predict larger Q values for lower frequency resonances. However, it should be noted with caution that the lower frequency resonators described in reference [6] were torsional fiber resonators. These had much larger surface area to volume ratios than the higher frequency resonators which consisted of compressional resonances in bulk material. If the losses in the torsion resonators are limited by surface effects, it may be inappropriate to use those measured values in an estimate of the loss function of fused silica test masses. Nevertheless, these data are the best which are currently available, and we therefore adopt as a loss function for the purposes of estimating thermal noise $\varphi = 10^{-7}$, independent of frequency. We believe this to be a conservative estimate.

V. Addition of modes

The total thermally excited motion of the mirror surface at a given frequency can be predicted from the sum of equation (15). This raises the question of how many modes must be counted in the sum to achieve an accurate estimate of the thermal noise. Figure 6 shows the cumulative contribution to the thermal motion at 100 Hz from the first one hundred axisymmetric modes for the prototype mirror, taking $\varphi(100\, Hz)$ to be 10^{-7} for all modes. The line indicates the cumulative contribution, and the x's indicate the individual contribution of each mode. The contribution of each individual mode decreases with resonant frequency due to the ω_n^2 term in the denominator of equation (15), but the mode density increases linearly with frequency (the axisymmetric modes form a two dimensional system). Also, there is a general decrease in the α_n terms in the denominator in equation (15). The result is a cumulative contribution which increases almost linearly with the maximum resonant frequency included. The approximate dependence of the effective mass coefficients on the resonant frequency needed to give this relation is $1/f$, plotted as the dashed line in figure 3. More than 100 modes were not calculated because of numerical precision errors in the series solving the equations of elasticity.

The thermal motion of the mirrors does not actually diverge, but convergence of the series depends upon the beam spot size. As an example, the initial LIGO interferometers will have a mirror with a diameter of 25 cm, a thickness of 10 cm, a mass of 10 kg, and a spot size of 2.2 cm. Figure 7 shows the convergence of the initial LIGO configuration at 100 Hz with the same dissipation as in figure 6. Figure 7 also shows the convergence of the prototype mirror with the same 2.2 cm size spot. The dot-dash vertical lines on the figure indicate the frequencies at which half an acoustic wavelength becomes equal to the diameter at which the laser beam intensity is $1/e$ of its maximum. At this frequency the displacement of the modes, $\vec{u}(\rho, \theta)$, begins to cancel in the integral of equation (4), resulting in larger effective masses, and the thermal noise sum begins to level off. The first vertical line corresponds to transverse acoustic waves, and the second corresponds to longitudinal acoustic waves. Each mode is a combination of both transverse and longitudinal motion, so there is no clear cut off frequency.

The thermal noise has the following pattern of convergence: the contribution increases approximately linearly with the largest resonant frequency included until the laser beam spot diameter becomes comparable to half an acoustic wave length. At that frequency the thermal noise sum levels off. Once this pattern of convergence was established, it was verified numerically for several mirror and spot size combinations.

Figure 7 also indicates that mirrors of different geometries can have significantly different thermal noise levels. Typically more modes can contribute to thermal noise for larger mirrors and fewer modes can contribute for larger spot sizes. One should be cautioned, however, that the prototype mirror would be inappropriate for use in a full scale (km length) interferometer. Since the radius of the prototype mirror is only twice the spot size, the optical loss at the edge of the mirror would be too great. Optical requirements thus give additional constraints on the geometry of the mirrors.

The convergence pattern of the sum in equation (13) is independent of the loss function as long as all of the modes have the same loss function. This is because the relevant parameter for predicting the thermal noise is the loss function evaluated at the frequency of interest, in our case 100 Hz, and not the Q of the resonance, which could depend on the resonant frequency for an arbitrary loss function. The convergence pattern does depend somewhat on the geometry of the mirror in that the axisymmetric modes were taken to be a two dimensional system. To meet this criterion, both the mirror diameter and its thickness must be greater than the laser spot size. This criterion is met in most realistic applications; it is not met in the one dimensional approximation (e.g. equation (7)).

VI. Effects of moving the beam off center

The numerical calculations described in the previous sections assumed that the laser was in the TEM_{00} mode and was centered on the mirror. This configuration puts the laser at an antinode of all axisymmetric modes and a node of all non-axisymmetric modes. By moving the beam off center, the contribution of the axisymmetric modes is decreased and the contribution of the non-axisymmetric modes is increased. Figure 8 shows the effect of moving the beam off center on the prototype mirror. 200 modes with circumferential order (number of nodal lines across the center; axisymmetric modes are order 0) of up to four were included, 40 modes each of circumferential order 0, 1, 2, 3, and 4. The thermal noise is computed using the same parameters as in section V (the laser beam spot size was 2.2 mm). The largest frequency mode included had an acoustic wavelength which was much larger than the laser spot size. Numeric precision errors prevented more modes from being included; only 40 axisymmetric modes were included to avoid giving those easy to calculate modes too much weight. The figure shows that except for spots quite near the edge, the thermal noise estimate is relatively insensitive to the exact location of the beam. It is not clear whether the variation seen across the mirror surface is real or whether it is an effect of including only a finite number of modes. We hypothesize that if all modes were included, then once the laser beam is more than a few spot diameters (or a few acoustic wavelengths of the highest frequency relevant modes) from the edge of the mirror, the mirror surface would effectively appear to the laser to be an infinite plane with all points having the same motion. Other interferometer details, such as noise in the alignment of the mirror [12], are likely to put much more stringent requirements on the centering of the beam on the mirror.

VII. Applicability of results to real mirrors

Actual interferometer mirrors differ from ideal right solid cylinders. As mentioned in Section III, the mirrors are wedge shaped for optical reasons. The surface might also be slightly curved. The mirrors are not completely free but are suspended as pendulums by fibers. To define the point of contact between the fiber and the mirror and to prevent rubbing, there may be attachments glued to the mirror, shown in figure 9. Furthermore, in order to apply forces to control the orientation of the mirror or to keep the TEM_{00} optical mode on resonance in the interferometer, there may be magnets attached to the mirrors.

These differences can be described as perturbations to shape and boundary conditions of a right solid cylinder. Such perturbations can cause mixing of the modes [13], making the effective mass coefficients difficult to calculate. This problem becomes particularly severe for modes with high resonant frequencies, where the mode density is large (the mode density increases as the resonant frequency squared for a three dimensional system). Nevertheless, it is reasonable to assume that the effects of such perturbations cancel when summed over a large number of modes so that the unperturbed calculation is expected to give a good approximation to the total noise at frequencies far from the mechanical resonances.

Attachments can also add dissipation to the system and may decrease the Q's of the mirror by large factors. This additional dissipation may affect each mode differently (affecting the results of section V) and can be frequency dependent. An example of such a dissipation mechanism which has been studied in detail is resonant coupling between a suspended cylinder and its suspension wires [14]. The additional dissipation of attachments can likely be minimized by using careful experimental technique. These and other surface losses will generally cause less additional dissipation in mirrors which have larger ratios of volume to surface (or contact) area, such as those to be used in the full scale interferometers.

VIII. Conclusion

We have developed a detailed model of the thermally excited vibrations of an interferometer mirror. The coupling of each mode to the interferometer has been explicitly calculated, and the number of modes which must be included in the model has been estimated. This calculation predicts a larger level of vibrational thermal noise than previous estimates for two reasons: first, the coupling between the mirror vibrations and the optical path length is generally stronger than previously estimated, and second, more modes must be included in calculations of the thermal noise than previously thought. The assumption that the loss function is independent of frequency also affects the magnitude of our thermal noise estimate as compared to other models; our assumption leads to larger thermal noise estimates than previous estimates using a viscous loss function.

This new estimate of the thermal noise level indicates that vibrational thermal noise may affect the sensitivity of advanced laser interferometer gravitational-wave detectors. The noise level shown in figure 7 for the LIGO mirror, corresponding to a displacement noise of $6 \times 10^{-20} m/\sqrt{Hz}$ or a strain noise of $1.5 \times 10^{-23} Hz^{-1/2}$ at 100 Hz with all four mirrors included, lies between the "initial" and "advanced" detector noise levels of reference [1]; vibrational thermal noise was not included in reference [1] because it was previously estimated to be below other noise sources in the detectors.

The noise level of figure 7 should not be taken as the final estimate of thermal noise, but it does indicate that the noise goals of the "advanced" interferometers likely cannot be met using present technology. The estimate of the dissipation, $\varphi(\omega)$, of 10^{-7} at 100 Hz was based on several recent experiments and reflects the levels of dissipation which have already been achieved. Data obtained so far probably do not represent the fundamental level of dissipation in fused silica, which is unknown. Dissipation in bulk fused silica can depend on both the levels of impurities and the preparation of the sample. Meeting the thermal noise goals of advanced gravitational-wave detectors will require a better understanding of the dissipation in fused silica (or finding a better mirror material) and an improvement in the loss over what has been achieved thus far. Also the dependence of the noise on mirror geometry indicated in figure 7 suggests that some additional improvement in the thermal noise level can be made by optimizing the mirror size and shape.

Acknowledgment

We thank the entire LIGO team for their encouragement and cooperation.

In particular T. Lyons, L. Sievers, and S. Whitcomb provided useful insights. We are especially grateful to J. Hutchinson, who provided the software which calculated the mode resonant frequencies and served as the foundation for these calculations. This work was supported by the National Science Foundation Cooperative Agreement PHY-9210038.

References

[1] LIGO collaboration, A. Abramovici *et al.*, Science **256**, 325 (1992).

[2] VIRGO collaboration, C. Bradaschia *et al.*, Nuc. Instr. Meth. Phys. Res. **A289**, 518 (1990).

[3] A. Gillespie and F. Raab, submitted to Phys. Lett. A (1994), and references therein.

[4] D. Shoemaker, R. Schilling, L. Schnupp, W. Wrinkler, K. Maischberger, and A. Rüdiger, Phys. Rev. D **38**, 423 (1988).

[5] P.R. Saulson, Phys. Rev. D **42**, 2437 (1990).

[6] V.B. Braginsky, V.P. Mitrofanov, and O.A. Okhrimenko, JETP Lett. **55**, 432 (1992), and references therein.

[7] J.R. Hutchinson, J. Appl. Mech. **47**, 901 (1990).

[8] H. Kogelnik and T. Li, Appl. Opt. **5**, 1550 (1966).

[9] A. Rüdiger, R. Schilling, L. Schnupp, W. Winkler, H. Billing, and K. Maischberger, Opt. Acta **28**, 641 (1980).

[10] G.W. McMahon, J. Acoust. Soc. Am. **36**, 85 (1964).

[11] H.B. Callen and T.A. Welton, Phys. Rev. **83**, 34 (1951).

[12] S. Kawamura and M.E. Zucker, accepted for publication in Appl. Opt. (1994).

[13] J.E. Logan, N.A. Robertson, J. Hough, and P.J. Veitch, Phys. Lett. A **161**, 101 (1991).

[14] J.E. Logan, N.A. Robertson, and J. Hough, Phys. Lett. A **170**, 352 (1992).

Figure Captions

FIG. 1: Schematic view of a LIGO interferometer.

FIG. 2: Mode shapes, resonant frequencies, and effective mass coefficients of a "prototype" mirror.

FIG. 3: The effective mass coefficients as a function of mode resonant frequency for a "prototype" mirror. The modes shown in figure 2 are plotted as squares; the modes shown in figure 4 are plotted as dots. The dashed line corresponds to $1/f$.

FIG. 4: The shape of the mirror surface for four higher frequency modes with small effective masses. a) f = 143 kHz, α = 0.013; b) f = 173 kHz, α = 0.014; c) f = 200 kHz, α = 0.021; d) f = 262 kHz, α = 0.009.

FIG. 5: Comparison between calculated effective mass coefficients and the measured coefficients. The solid line is a direct proportionality.

FIG. 6: The contribution to the low frequency thermal noise of the modes of the "prototype" mirror with a 0.22 cm beam spot. The line is the cumulative contribution and the x's are the individual contribution of the modes.

FIG. 7: The cumulative contribution to thermal noise of both the "prototype" (dotted line) and the LIGO (solid line) mirrors with a 2.2 cm beam spot. The vertical dot-dashed lines indicate the frequencies at which an acoustic wavelength is of order the beam spot size for both transverse (left line) and longitudinal (right line) waves.

FIG. 8: The thermal motion of the "prototype" mirror as a function of position on the mirror.

FIG. 9: Schematic view of a mirror with fiber attachments and magnets.

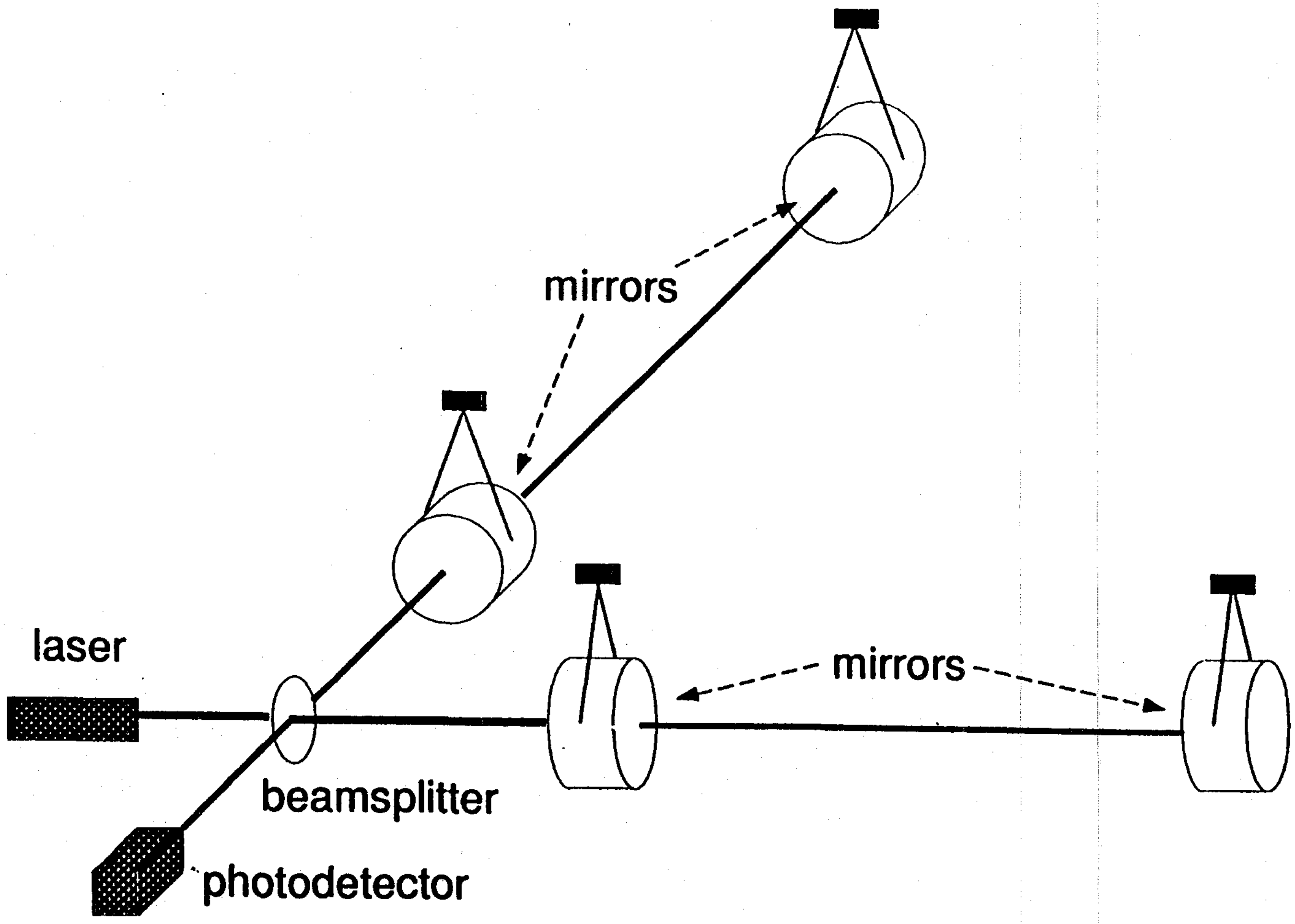
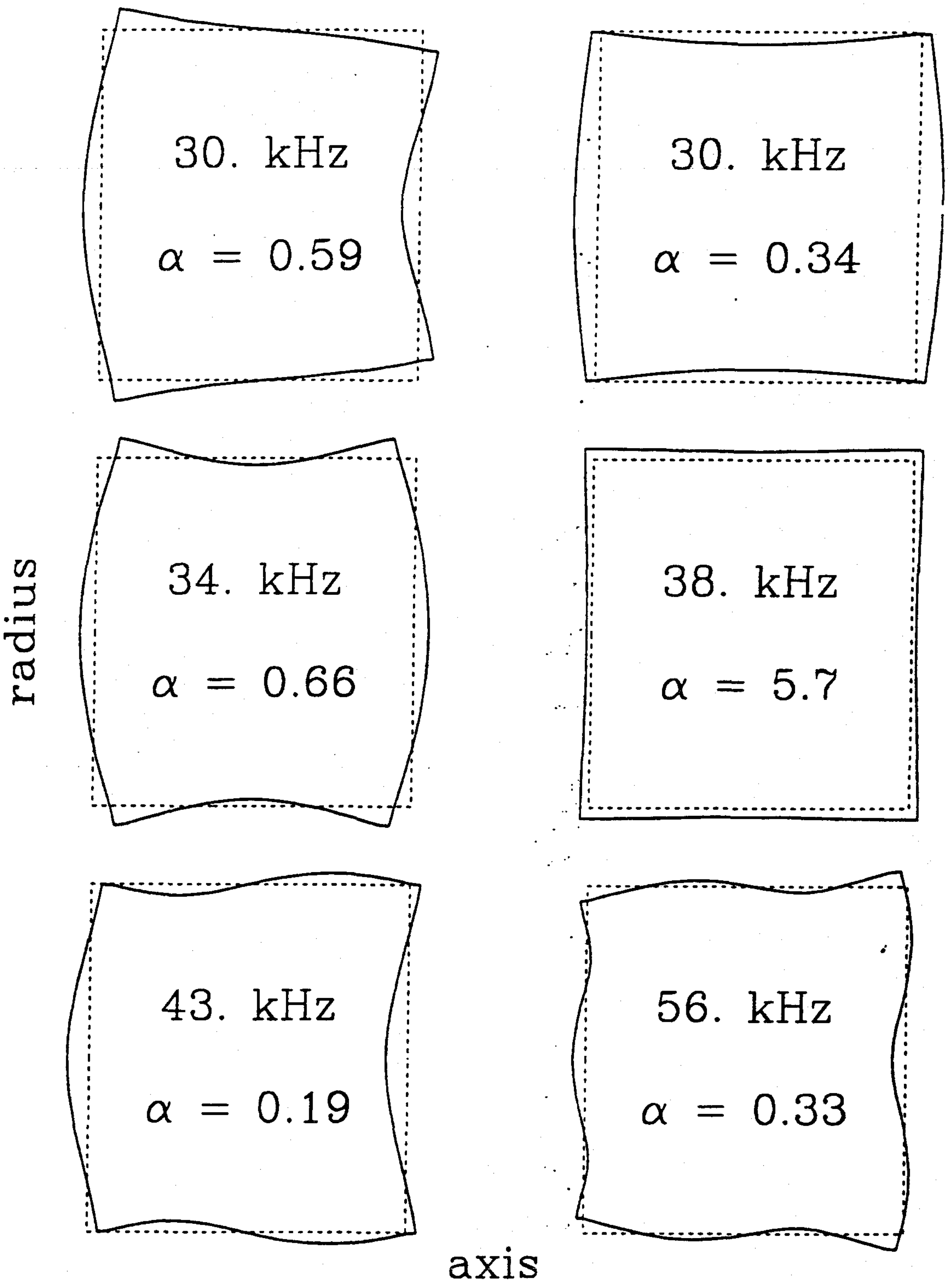


Figure 1

Figure 2


30. kHz
α = 0.59
30. kHz
α = 0.34
34. kHz
α = 0.66
38. kHz
α = 5.7
43. kHz
α = 0.19
56. kHz
α = 0.33
radius
axis

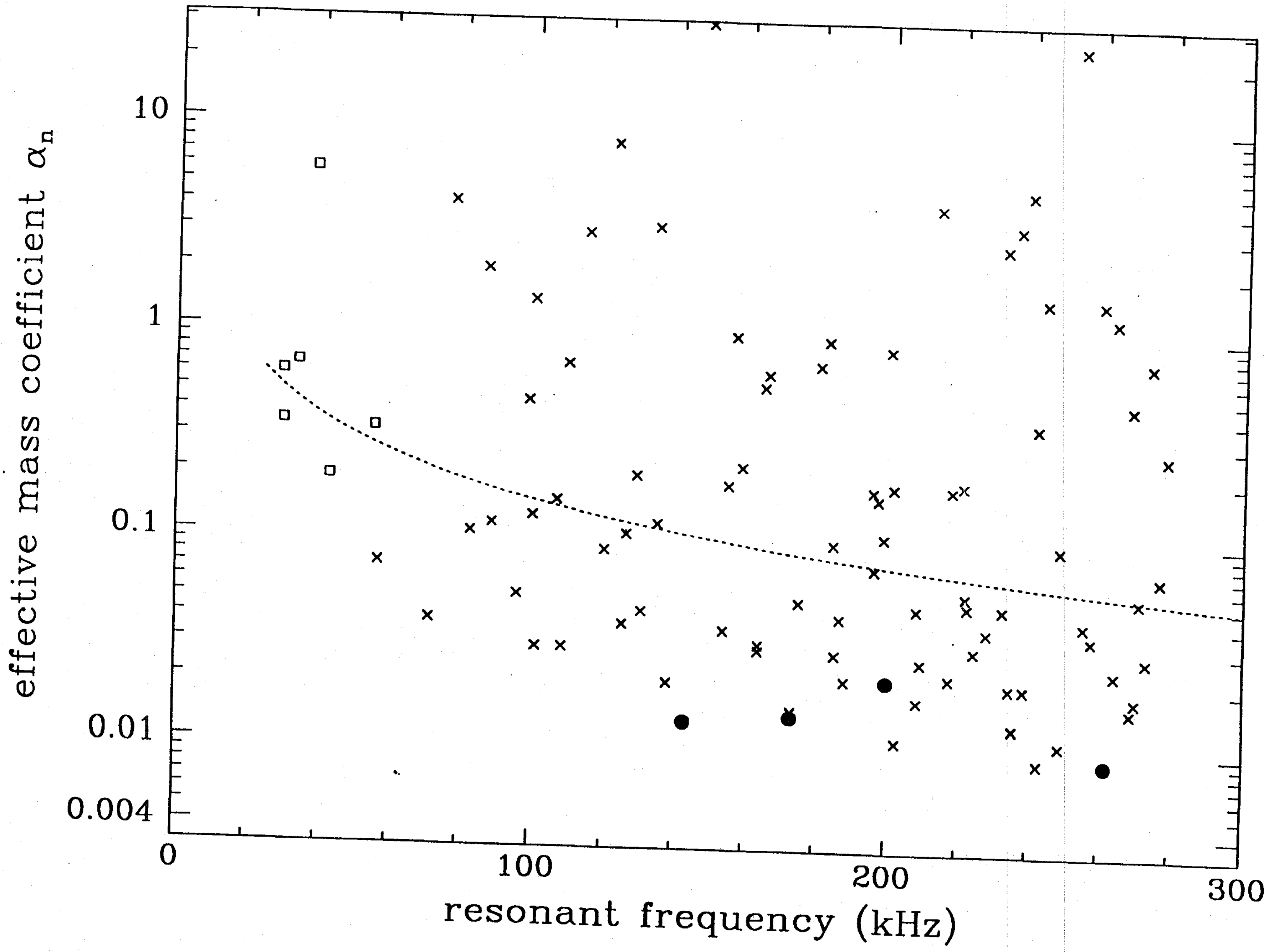
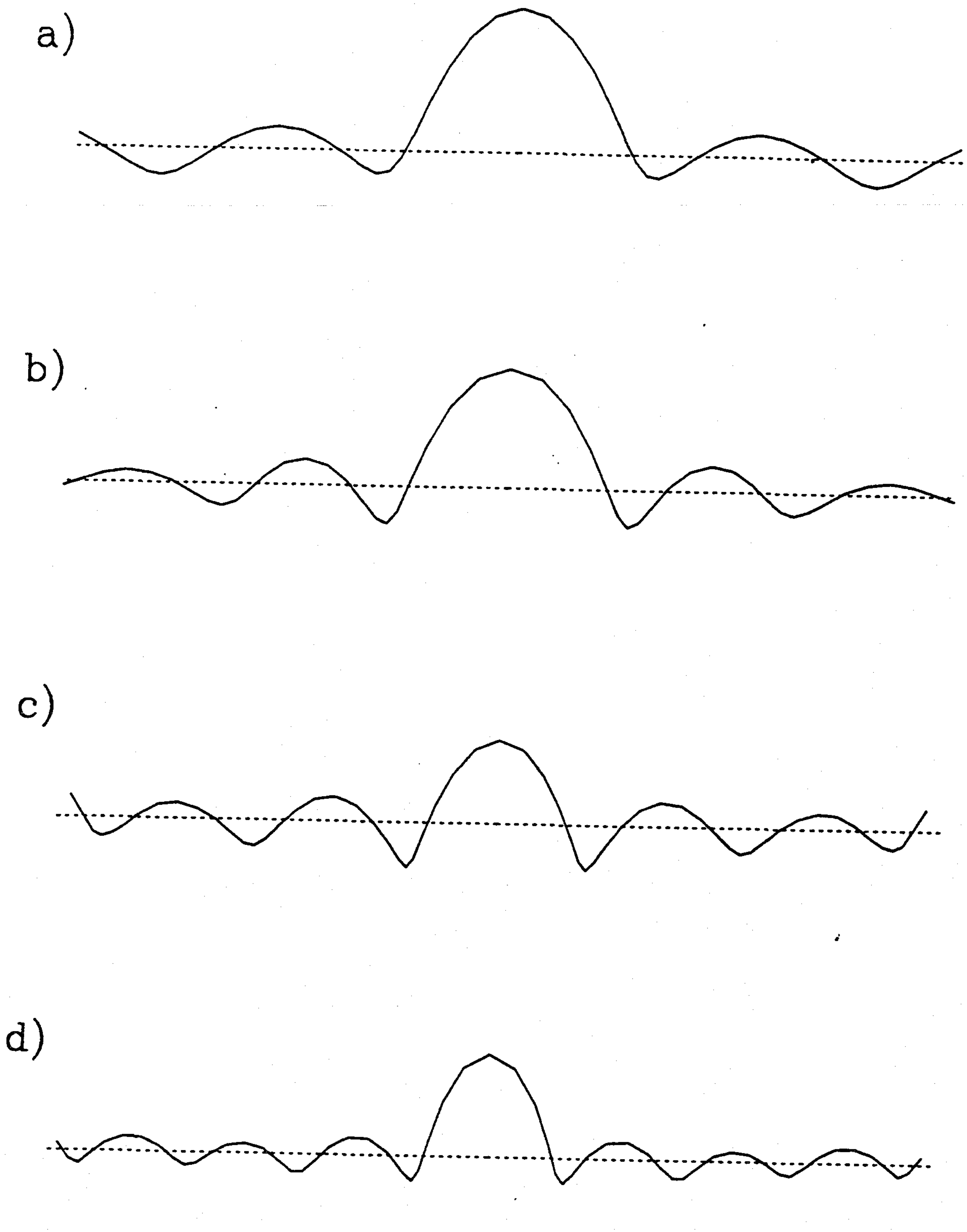

10
1
0.1
0.01
0.004
0
100
200
300
effective mass coefficient α_n
resonant frequency (kHz)

Figure 3

Figure 4



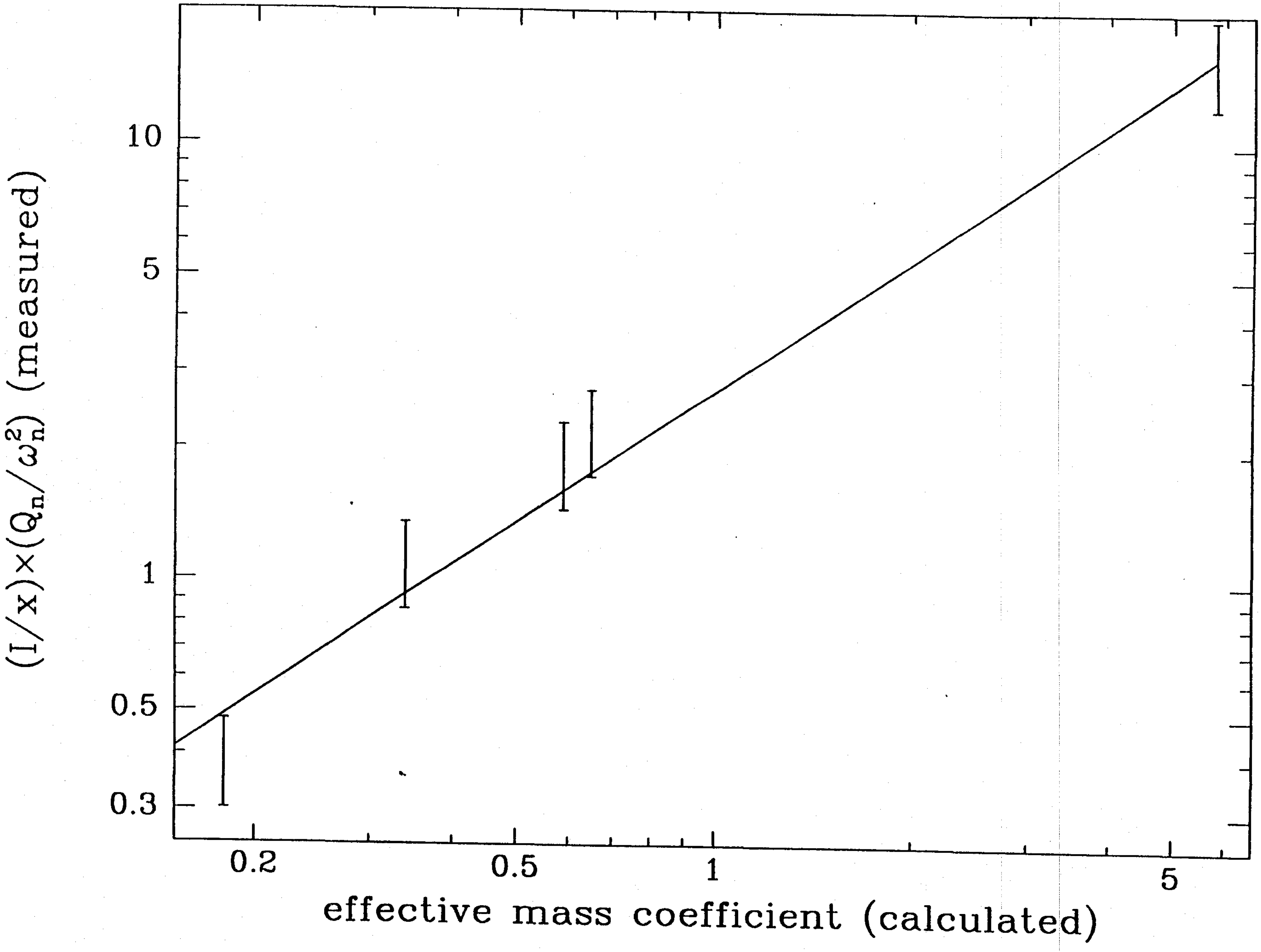
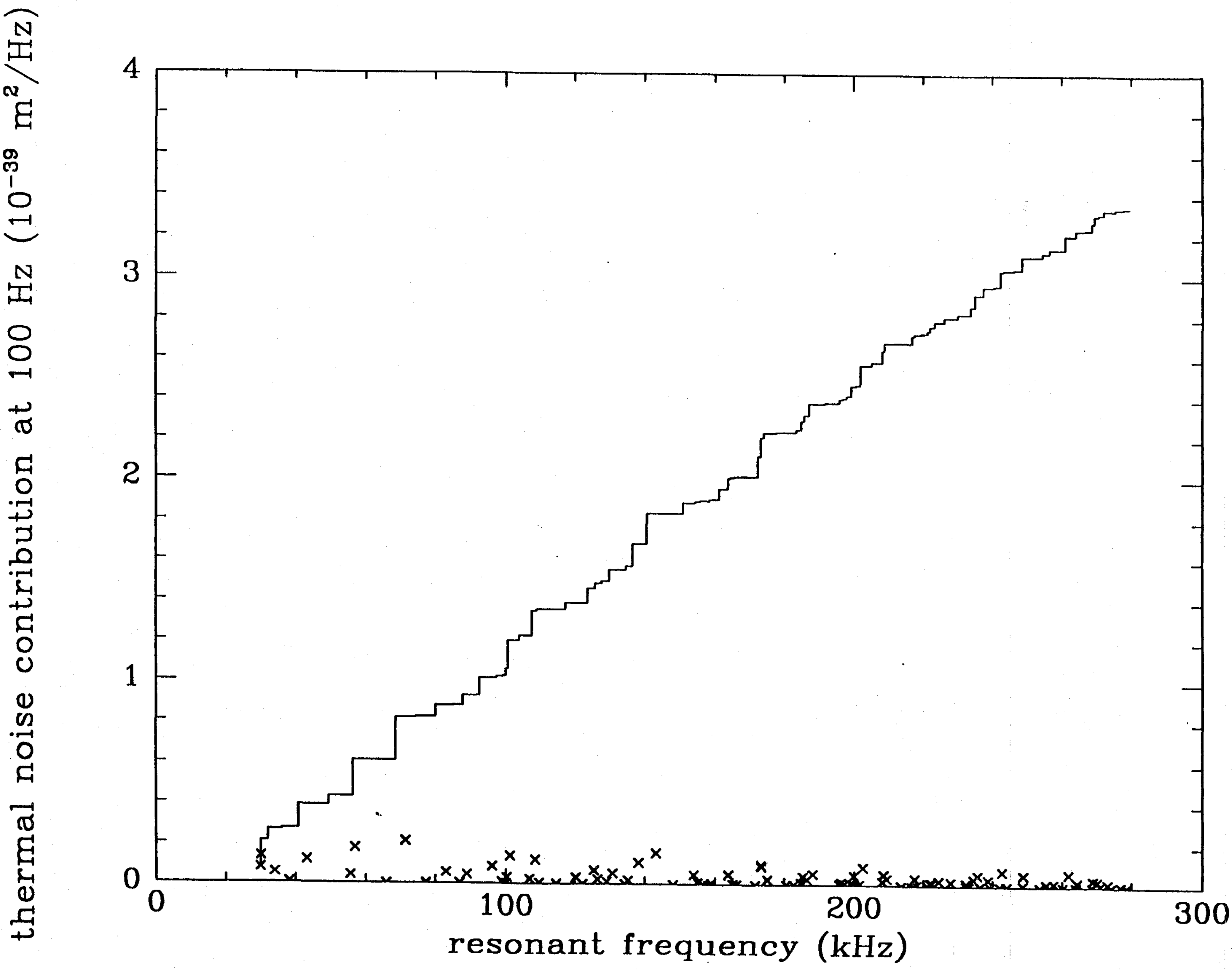

(I/x)×(Q$_n$/ω$_n^2$) (measured)
10
5
1
0.5
0.3
0.2
0.5
1
5
effective mass coefficient (calculated)

Figure 5


thermal noise contribution at 100 Hz (10^{-39} m^2/Hz)
4
3
2
1
0
0
100
200
300
resonant frequency (kHz)

Figure

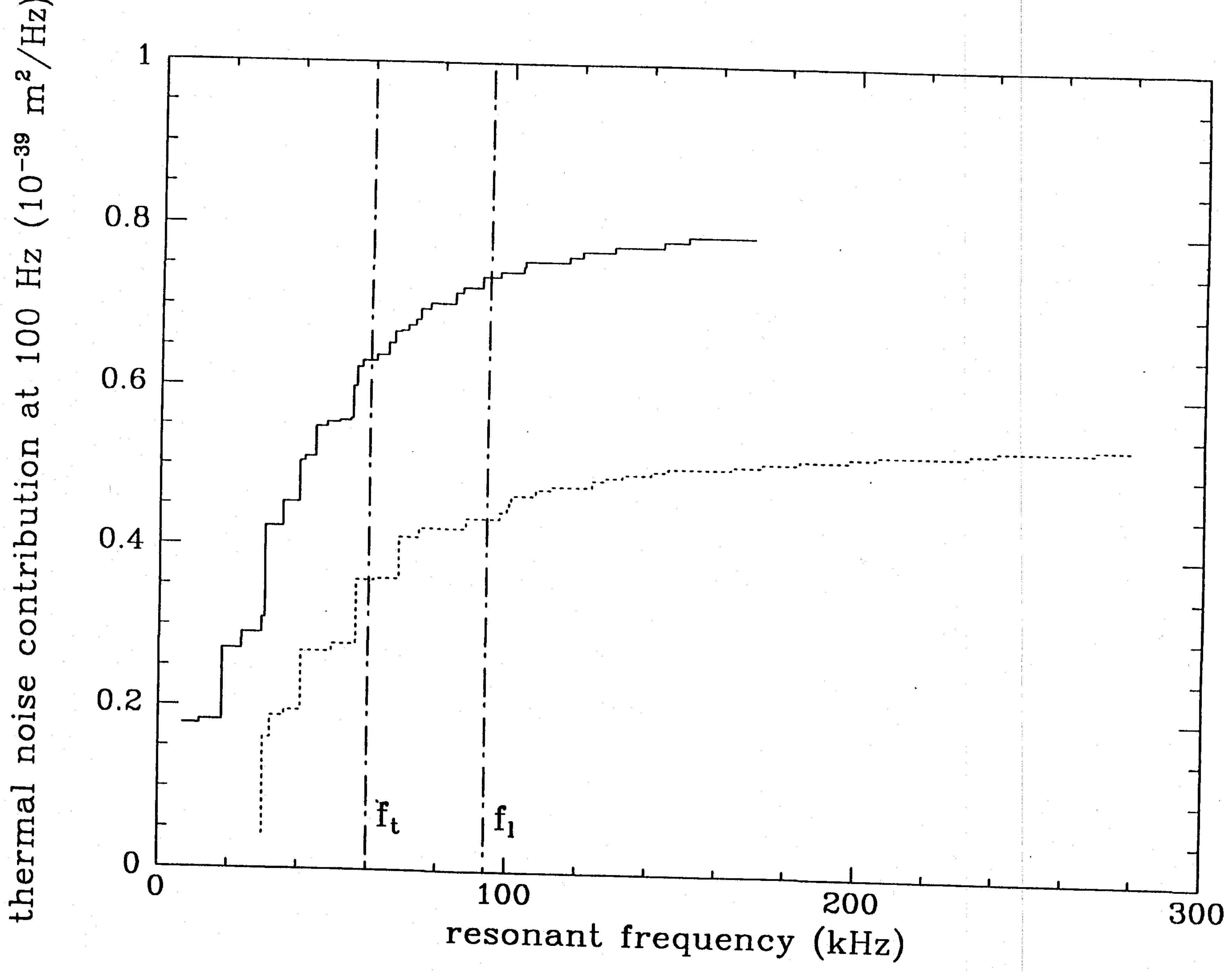

thermal noise contribution at 100 Hz (10^{-39} m^2/Hz)
1
0.8
0.6
0.4
0.2
0
0
100
200
300
resonant frequency (kHz)
f_t
f_l

Figure 7

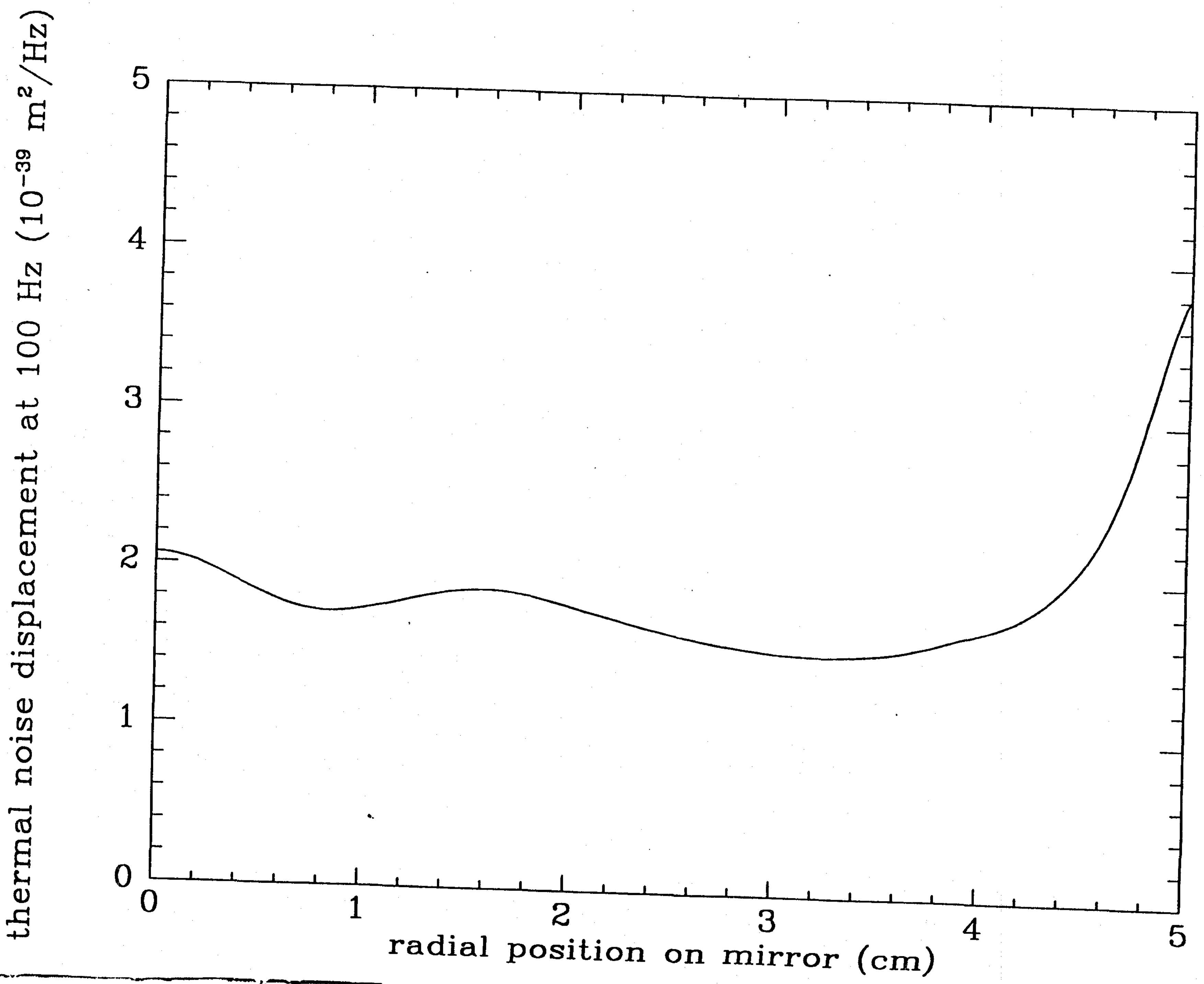
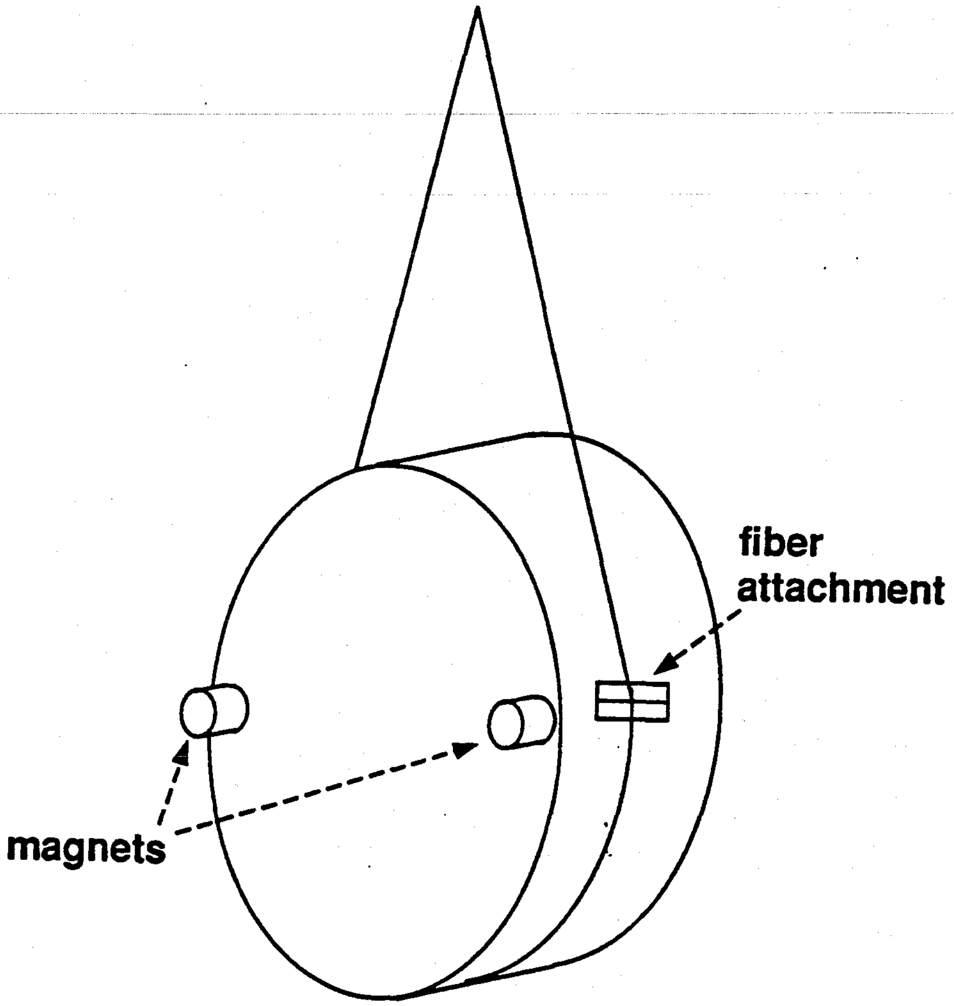

thermal noise displacement at 100 Hz (10^{-39} m^2/Hz)
5
4
3
2
1
0
0
1
2
3
4
5
radial position on mirror (cm)

Figure 8

Figure 9



REF. RR

Suspension Losses in the Pendula of Laser Interferometer Gravitational-Wave Detectors

A. Gillespie and F. Raab

LIGO Project, California Institute of Technology, Pasadena, CA 91125, USA

Abstract

We have experimentally tested models currently in use to estimate the mechanical losses and thermal noise of the test mass suspensions of laser interferometer gravitational-wave detectors. Observed losses are approximately independent of frequency from 1 Hz to 2 kHz, resulting in lower thermal noise estimates than with some previous models.

1. Introduction

The impending construction of large scale laser interferometer gravitational-wave detectors, such as the American LIGO [1] and the Italian/French VIRGO [2] projects, has increased interest in developing and understanding low loss pendulum suspensions. The losses of the suspension systems of the mirrors of these interferometers must be minimized in order to reduce the contribution of thermal noise to the overall noise spectrum [3]. The thermal noise as a function of frequency can be predicted if the specific frequency dependence of the damping mechanism is known. A number of assumptions have been used to predict this frequency dependence for a test mass suspension.

We present here results of an experimental study of the losses in an actual test mass suspension. The data support key assumptions made in recent models used to predict the thermal noise and allow us to estimate the general frequency dependence of the losses over the entire frequency band relevant to earth-based laser interferometer gravitational-wave detectors (1 Hz to 10 kHz). We find that the losses of the suspension system under study are nearly independent of frequency, a result which differs from the assumption of viscous losses that went into the formulation of the original thermal noise estimates for the LIGO detectors [1]. This result leads to generally lower estimates for thermal noise generated in the test mass suspension.

2. Model

The method which we use to predict the general frequency dependence of the thermal noise from the suspension system has been explained in detail by Saulson [3], and the model by which we relate the losses of the violin modes to the losses in the pendulum mode was derived in a previous letter [4]. In this section we highlight the results of those papers which are relevant to this work.

To model the damping of the suspension system, we adopt a general form of Hooke's law in which the spring constant is taken to be complex [5,3]:

$$F = -k[1 + i\varphi(\omega)]x. \tag{1}$$

F, x, ω, and $k[1 + i\varphi(\omega)]$ are the force, displacement, angular frequency, and complex spring constant, respectively. The specific damping mechanism is parametrized by the imaginary part of the spring constant, $\varphi(\omega)$, referred to as the loss function. Once $\varphi(\omega)$ is known, it is straightforward to calculate the spectral density of the displacement due to thermal excitations, $\tilde{x}^2(f)$, using the Fluctuation-Dissipation theorem [6]:

$$\tilde{x}^2(f) = \frac{4k_BT}{m}\frac{\omega_0^2\varphi(\omega)}{\omega\left[\left(\omega_0^2 - \omega^2\right)^2 - \omega_0^4\varphi^2(\omega)\right]}. \tag{2}$$

k_B, T, m, and ω_0 are Boltzman's constant, the temperature, the mass, and the resonant angular frequency, respectively.

Measuring $\varphi(\omega)$ of a low loss mechanical system can be exceedingly difficult. The Q of the system is related to the loss function at the resonant frequency by

$$\varphi(\omega_0) = \frac{1}{Q}, \tag{3}$$

but this yields no direct indication of the frequency dependence of φ. For a system with multiple resonances, the Q's of several resonant modes can be measured at their respective resonant frequencies. If a relationship among the loss functions for these modes can be obtained, one can then interpolate between the resonances to estimate the general frequency dependence of φ.

We are specifically interested in a pendulum consisting of a test mass supported by finite mass wires. Here we concentrate on two types of suspension resonances, the pendulum resonance, which is at approximately 1 Hz, and the violin resonances, which form a harmonic series starting at several hundred Hz. We have obtained a relationship between the losses of these modes using arguments relating the geometry and energy of the violin and pendulum modes and by assuming that the losses occur near the ends of the suspension wires [4]. For a double loop (4 wire) suspension system,

$$\varphi_p(\omega) = \frac{1}{8}[\varphi_{v1}(\omega) + \varphi_{v2}(\omega) + \varphi_{v3}(\omega) + \varphi_{v4}(\omega)]. \tag{4}$$

φ_p and φ_{vi} are the loss functions of the pendulum mode and of the violin modes of the i$^{\text{th}}$ wire, respectively. The assumption that the losses occur near the endpoints of the wires should be valid when the losses arise either in the attachment of the wire to the support structure or the test mass, or in the bending of the wire, which is most severe near the ends.

A consequence of the losses being concentrated at the endpoints is that the losses are a function not only of frequency, but also of the length of the suspension wires. The explicit length dependence can be obtained by combining equations (7) and (8) of reference [4], giving

$$\varphi(\omega, l) = \frac{1}{l}\varphi(\omega), \tag{5}$$

where l is the length of the wire.

Gonzalez and Saulson have derived the losses in the violin modes by solving for the exact mechanical transfer function through the suspension system and assuming the losses occur in the wire material itself [7]. The loss function of the n$^{\text{th}}$ violin mode is then

$$\varphi_v(\omega) \approx \frac{2}{l}\sqrt{\frac{EI}{t}}\left[1 + \frac{1}{2l}\sqrt{\frac{EI}{t}}(n\pi)^2\right]\varphi_0(\omega). \tag{6}$$

$E, I, t, n,$ and $\varphi_0(\omega)$ are the Young's modulus, the area moment of inertia, the tension, the harmonic number, and the losses in the wire material, respectively. The second term in the brackets is the correction for the losses due to the bending which occurs along the length of the wire. For the parameters appropriate to the first generation of laser interferometer gravitational-wave detectors, that term is negligible for the first several harmonics of the wire. In this limit, the losses in the pendulum mode can be predicted by equation (4), and the length dependence reduces to equation (5).

Logan, Hough, and Robertson [8] have produced a similar result for the losses in a pendulum with a finite mass wire by assuming that all of the losses occur near the ends of the wire and then developing an analogy to an electrical transmission line. This model also has equation (4) as its primary result, and can be tested by experimentally verifying equation (5).

3. Experiment

The losses in the suspension system were obtained from the Q's of the resonances, which were determined by exciting a particular mode, turning off the excitation, and measuring the decay time of the oscillation ($Q = \omega_0\tau/2$, where ω_0 is the resonant frequency and τ is the measured amplitude decay time). The apparatus used for the measurements is shown in figure 1.

The suspension system consisted of a 1.6 kg cylinder of fused silica (with a diameter of 10 cm and a length of 8.8 cm) suspended by two loops of 75 μm diameter steel music wire. For each measurement, new wire was taken directly from the spool and wiped with acetone to remove residual oils. The wire tension was approximately one half of its breaking strength. Two different types of clamps were used at the top. In one type the wire was simply clamped between two aluminum bars; the corners of the bars had approximate radii of curvature of 100 μm and the faces of the bars where the wire left were flush to approximately that level. For the other clamping arrangement the wire left a 45° wedge with a 25 μm radius edge. No difference in the Q's was found between the two clamping methods. At the test mass the wires were looped over a standoff to maintain a significant pressure on the wire and thus avoid rubbing at the point of contact. Either fused silica prisms (13 mm long by 2 mm equilateral triangle) or rods (13 mm long by 1 mm in diameter) were used as standoffs. The standoffs were held in place by one of three methods: held simply by the pressure of the wire with no glue, glued on the cylinder with a cyanoacrylate based glue, ?nd attached to the cylinder with a vacuum sealant epoxy. There was no observable difference between the Q's of either type of standoff even though the sharp edge of the prism exerted more contact pressure on the wire than the gentler curve of the rod, and the means of holding the standoff in place had no effect. However when no standoffs were used, the Q's of the violin modes were degraded by a factor of approximately 30.

The violin resonances were excited by a piezoelectric transducer glued to the support bar near the clamping point at the top of the suspension. The motion was measured by focusing a HeNe laser onto the wire and monitoring the diffraction pattern on a split photodiode. To damp the seismic motion of the mass while the violin mode Q's were being measured, magnets were attached to the mass and current was fed back from an active damping circuit to nearby coils. The pendulum mode was excited by physically pushing the mass just prior to pumping the vacuum system. The measurement was begun when the vacuum system reached its operating pressure ($<10^{-4}$ Torr). The test mass motion in the pendulum mode was monitored by an edge sensor (an LED which cast a shadow of the corner of the test mass on a photodiode).

4. Minimizing losses that primarily affect the pendulum mode

To compare the losses of the violin modes to the losses of the pendulum mode using equation (4), other possible losses due to forces which are external to the suspension but act directly on the test mass must be minimized. Due to the large mismatch between the mechanical impedances of the mass and the wires, these forces primarily affect the pendulum mode and not the violin modes. These are discussed briefly here because of their importance to the design of the experiment.

One such loss mechanism is eddy current damping which is important if magnets are used as actuators on the mass. Another mechanism is damping by residual gas in the vacuum system. Both of these damping mechanisms are viscous; that is, their damping forces are proportional to the velocity of the test mass. Because of the frequency dependence of these damping mechanisms ($\varphi(\omega) \propto \omega$), such losses could be important at frequencies of hundreds of Hz, while having little effect on the measured pendulum Q. However, these losses can be readily estimated and reduced by adjusting the relevant parameters of the system (the pressure in the vacuum system or the geometry of conductors near the magnets).

The results of such measurements are given in figure 3. Each cross represents a measurement done on a separate wire. As in figure 2 the reproducibility of a measurement made on a single wire was approximately five percent; the spread in the points at a given length indicates the variation in the Q's from wire to wire. The line in the figure represents the model of equation (11). The data support the hypothesis that the losses occur at the endpoints. Note that due to the different wire lengths, the data were taken at different frequencies. The observed weak frequency dependence of the losses can give systematic departures from the model of equation (11) of the order of 10% over this frequency band (180–800Hz).

Two possible sources of loss at the endpoints have been identified. One is damping in the clamps connecting the wire to either the test mass or the support structure. Since modeling the losses due to imperfect clamps is difficult, two different types of clamps were used at both the top and bottom of the wire (described in section 3). No difference in the measured Q's were found for any clamp combination. We interpret this result as indicating that the losses were not due to the clamps. The remaining possibility is that the losses were in the wire material itself.

6. Wire material losses

To compare the losses of the violin modes to the intrinsic losses of the wire material, the wire material Q's were measured using a method analogous to that of Kovalik and Saulson [13]. Pieces of wire were clamped at one end using the same aluminum bar clamp which was used for the suspension system, and the other end was left free. The modes were excited using the piezoelectric transducer glued to that clamp, and the motion was measured in the same manner as for violin modes.

The frequency dependence of the losses of the wire material is shown in figure 4. Three different wire lengths, 4, 5, and 6 cm, were used to arrive at the density of points shown. The error bars indicate the statistical reproducibility of the measurements on a single wire. The smoothness of the data gives an indication of the reproducibility from wire to wire. The dashed line on the figure is the loss due to thermoelastic damping.

A prediction of the losses in the violin modes of the test mass suspension due to the wire losses, using the model of equation (6), is represented by the x's in figure 2. Note that although the results of section 5 suggest that the losses may be due to the wire material itself, the measured test mass suspension losses exceed the losses predicted from the measurements on unloaded wires. One possible explanation for this discrepancy is that the losses in the wire material are a function of the stress in the material. A careful study of the stress dependence of the losses of some candidate suspension wire materials is being undertaken by Huang [14].

7. Implications for gravitational-wave detectors

The observed behavior of suspension losses has significant implications for the design of laser interferometer gravitational-wave detectors. For example, consider the initial LIGO detectors [1], where thermal noise is expected to be the principal noise contribution at frequencies near 100 Hz. For pendula consisting of 10 kg test masses with a 1 s period, thermal noise displacements of $10^{-19} m/\sqrt{Hz}$ have been estimated, corresponding to $\varphi(100Hz) = 10^{-5}$. If the losses were viscous as was assumed in reference [1], $\varphi(\omega) = (1/Q)(\omega/\omega_p)$, this would require a pendulum mode Q of 10^7 ($\varphi = 10^{-7}$ at 1 Hz). Viscous damping from residual gas or eddy currents can be reduced below this level using adequate precautions. The losses internal to the suspension observed here were nearly independent of frequency below a few kilohertz. Thus interpolated losses near 100 Hz satisfy thermal noise requirements even though the pendulum Q was significantly below 10^7.

The fact that the suspension losses are concentrated near the ends of the suspension wires suggests that the suspension length can be adjusted to optimize the interferometer sensitivity. Making the pendulum longer increases its Q and therefore decreases the thermal noise of the pendulum while increasing the number of violin modes in the frequency band of the detector. The relative importance of the thermal noise contributed by these two types of modes is determined by other noise sources and the signal that one is attempting to measure.

8. Summary

We have presented data which support an emerging model for predicting the suspension losses for a pendulum. The losses of this suspension were found to be nearly independent of frequency at frequencies where the losses were not limited by thermoelastic damping, and inversely proportional to the length, consistent with losses concentrated at the endpoints. The data do not agree with models currently being used to predict the losses in high Q pendula from the intrinsic losses in the wire material [7], possibly due to stress dependent effects. These results apply only to steel music wire; other candidate wire materials, such as tungsten [3], niobium [15], or fused silica [16], should be tested using similar techniques.

We thank R. Vogt and the entire LIGO team for making this work possible. We also thank P. Saulson, who made several useful comments while the work was in progress. This work was supported by the National Science Foundation cooperative agreement number PHY-9210038.

References

[1] A. Abramovici, W.E. Althouse, R.W.P. Drever, Y. Gursel, S. Kawamura, F.J. Raab, D. Shoemaker, L. Sievers, R.E. Spero, K.S. Thorne, R.E. Vogt, R. Weiss, S.E. Whitcomb, and M.E. Zucker, Science 256 (1992) 325.

[2] C. Bradaschia, R. Del Fabbro, A. Di Virgilio, A. Giazotto, H. Kautzky, V. Montelatici, D. Passuello, A. Brillet, O. Cregut, P. Hello, C.N. Man, P.T. Manh, A. Marraud, D. Shoemaker, J.Y. Vinet, F. Barone, L. Di Fiore, L. Milano, G. Russo, J.M. Aguirregabiria, H. Bel, J.P. Duruisseau, G. Le Denmat, Ph. Tourrenc, M. Capozzi, M. Longo, M. Lops, I. Pinto, G. Rotoli, T. Damour, S. Bonazzola, J.A. Marck, Y. Gourghoulon, L.E. Holloway, F. Fuligni, V. Iafolla, and G. Natale, Nuc. Instr. Meth. Phys. Res. A289 (1990) 518.

[3] P.R. Saulson, Phys. Rev. D, 42 (1990) 2437.

[4] A. Gillespie and F. Raab, Phys. Lett. A 178 (1993) 357.

[5] A.S. Nowick and B.S. Berry, Anelastic relaxation in crystalline solids (Academic Press, New York, 1972).

[6] H.B. Callen and T.A. Welton, Phys. Rev. 83 (1951) 34.

[7] G. Gonzalez and P.R. Saulson, to be published in J. Acoust. Soc. Am. (1994).

[8] J.E. Logan, J. Hough, and N.A. Robertson, Phys. Lett. A 183 (1993) 145.

[9] C. Zener, Phys. Rev. 52 (1937) 230; Phys. Rev. 53 (1938) 90.

[10] International Critical Tables of Numerical Data, Physics, Chemistry, and Technology, ed. E. Washburn (McGraw-Hill Book Co., Inc., New York, 1929).

[11] A.L. Kimball and D.E. Lovell, Phys. Rev. 30 (1927) 948.

[12] P.R. Saulson, R.T. Stebbins, F.D. Dumont, and S.E. Mock, Rev. Sci. Instrum. 65 (1994) 182.

[13] J. Kovalik and P.R. Saulson, Rev. Sci. Instrum. 64 (1993) 2942.

[14] Y. Huang, private communication (1994).

[15] D.G. Blair, L. Ju, and M. Notcutt, Rev. Sci. Instrum. 64 (1993) 1899.

[16] V.B. Braginsky, V.P. Mitrofanov, and O.A. Okhrimenko, Phys. Lett. A 175 (1993) 82.

Figure Captions

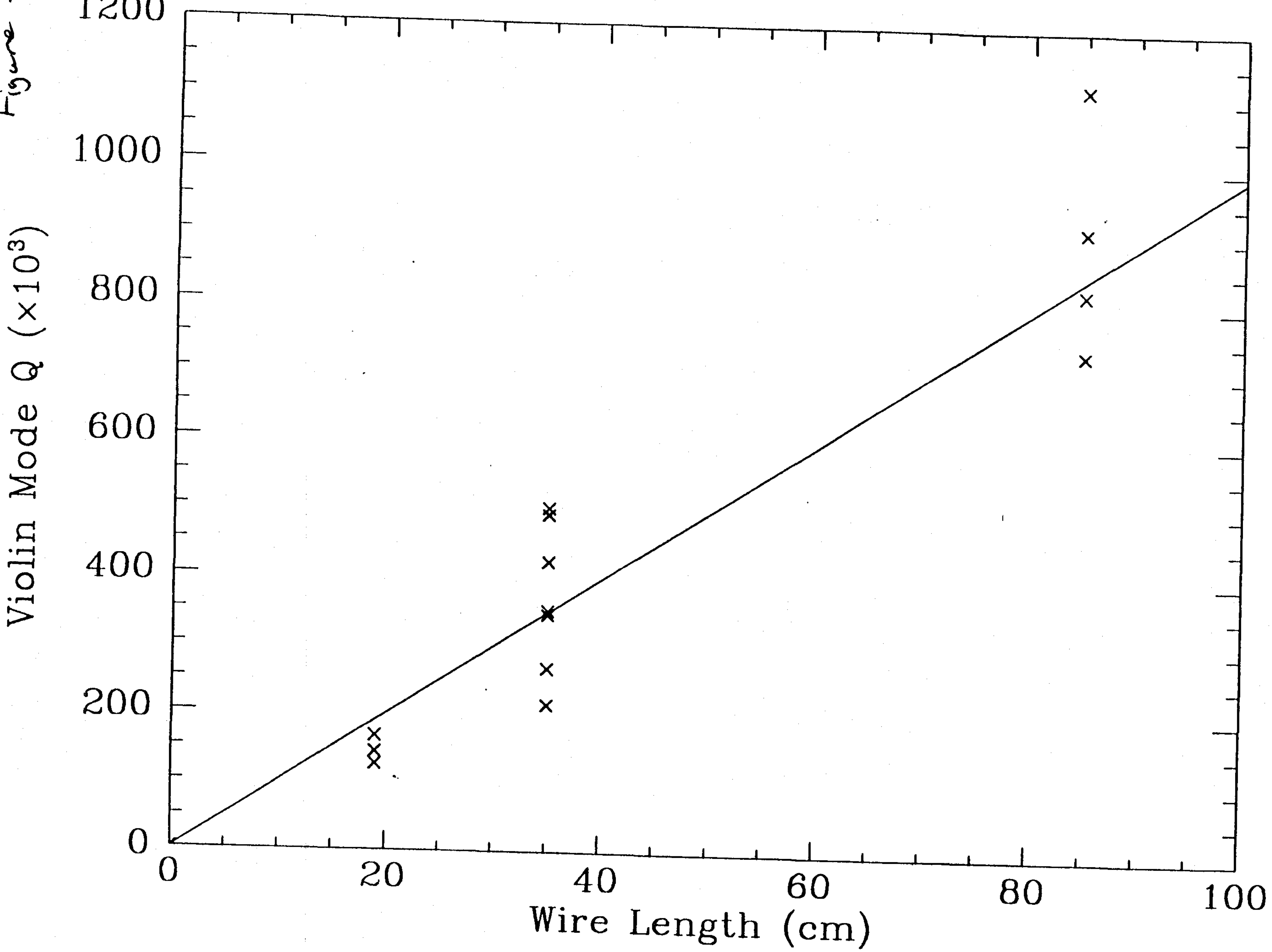
Figure 1: The experimental apparatus.

Figure 2: Frequency dependence of the violin mode losses. The solid squares represent the average losses of several wires at both the fundamental violin mode resonant frequency and at the pendulum resonant frequency; the open circles represent measurements on different harmonics of a single wire. The x's are the predicted losses from the wire material loss measurements described in section 6. The dashed line is the predicted loss due to thermoelastic damping.

Figure 3: Measured Q as a function of wire length.

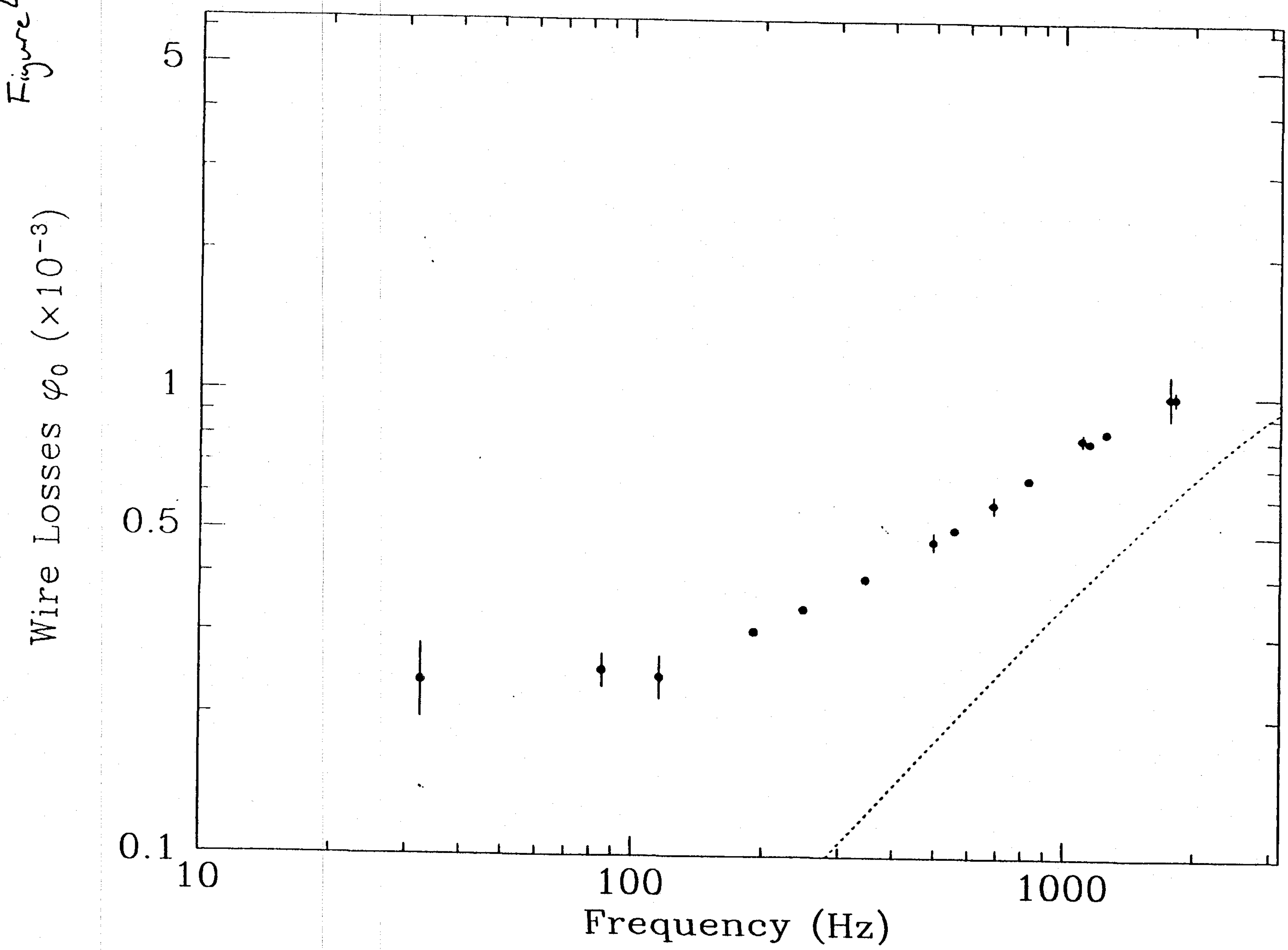
Figure 4: Wire material losses. The dashed line is the predicted loss due to thermoelastic damping. The error bars for the points in the mid frequency range are comparable to the size of the points.

Figure 3



Violin Mode Q (×10³)
1200
1000
800
600
400
200
0
0
20
40
60
80
100
Wire Length (cm)

Figure 4



Wire Losses φ_0 ($\times 10^{-3}$)
5
1
0.5
0.1
10
100
1000
Frequency (Hz)

BATCH START

Lecture 10: Light Scattering and its Control

STAPLE
OR
DIVIDER

Ph 103c: The Physics of LIGO

18 May 1994

LECTURE 15

Light Scattering and its Control

Lecture by Kip Thorne

Assigned Reading:

G. Chapter 7, "Diffraction" from the manuscript *Applications of Classical Physics* by Roger Blandford and Kip Thorne. [This material is needed as the foundation for the scattering analyses of Kip's lecture and for the Suggested Problem 2. at the end of this assignment. Sections 7.2 and 7.5 were assigned previously, in Lecture 4, and the manuscript was passed out then. If you have mastered the theory of diffraction, in some other course, in comparable detail to that given in this chapter, then you do not need to do this reading.]

Suggested Supplementary Reading:

SS. J. M. Elson, H. E. Bennett, and J. M. Bennett, "Scattering from Optical Surfaces," in *Applied Optical Engineering*, Vol. VII (Academic Press 1979), Chapter 7, page 191. [This was suggested previously, in Lecture 9. It is a review with few equations and with many references to the literature. The focus is on scattering from surfaces that are quite smooth (rms fluctuations in height much less than the wavelength of light, e.g., the LIGO mirrors).

l. Petr Beckmann and André Spizzichino, *The Scattering of Electromagnetic Waves from Rough Surfaces* (Macmillan/Pergamon, New York, 1963). [This is the classic treatise on the subject, with extensive equations. It deals with scattering from rough surfaces (rms fluctuations in height larger than a wavelength) as well as smooth ones. Unfortunately, it is written in such a way that one cannot readily understand later chapters without reading earlier ones.]

m. Kip S. Thorne, *Light Scattering and Proposed Baffle Configuration for the LIGO*, preprint GRP-200, available upon request from Kip. [This was the original, analytic calculation of the "gravity-wave" noise $\tilde{h}(f)$ caused by light scattering in LIGO both with and without baffles. It has two defects that make it not directly useful: (i) subsequent analytic calculations by Jean-Yves Vinet of the VIRGO Project ferreted out a serious error (a missing factor B inside the square brackets of Eqs. (4.6) and (4.7), which then propagated throughout GRP-200); and (ii) the final LIGO baffle configuration is rather different from the one in GRP-200. Kip's lecture is based on GRP-200, with the error corrected and the baffle configuration changed to the new one. The resulting noise spectrum $\tilde{h}(f)$, as discussed in Kip's lecture, is in good agreement with numerical simulations by the Breault Research Organization (BRO), under contract from LIGO. Eanna Flanagan and Kip are in the process of a final, thorough analytic reanalysis, which they plan to publish.]

A Few Suggested Problems:

1. *Backscatter off Baffles.* The dominant scattered-light noise source, according to the calculations by Kip, by Eanna Flanagan, by Jean-Yves Vinet, and by BRO, is backscatter off vibrating baffles; see the last of Kip's lecture transparencies.
   a. Give a list of factors that make the backscattered light coming from different directions superpose incoherently.
   b. Compute the "gravity-wave" noise $\tilde{h}(f)$ due to baffle backscatter, assuming incoherent superposition. Kip gives the answer (accurate to within a factor ~ 2) on his last transparency, when the mirrors are as close to the vacuum pipe wall as we expect them ever to be, $Y \simeq 20$ cm. You may prefer, for simplicity, to treat the case of mirrors centered in the beam tube, which has a radius $R = 60$ cm. [*Note:* In Kip's answer on his last transparency, $\alpha \lesssim 10^{-6}$ is the mirror's light-scattering coefficient (the probability for a photon to scatter from the main beam into a unit solid angle is $dP/d\Omega = \alpha/\theta^2$); $L =$ 4km is the length of the beam tube, l_1 =100m is the distance from the mirror to the nearest baffle, $\lambda = 0.4\mu$m is the wavelength of the laser light, $d\sigma/dAd\Omega \simeq 10^{-2}$ is the baffle's differential scattering cross section per unit area of baffle per unit solid angle into which the light goes (equivalently it is the probability that a photon, hitting the baffle at an angle of a few tens of degrees from its normal, gets backscattered into the direction from which it came); and $\tilde{\xi}(f)$ is the square root of the spectral density of the baffle's seismically induced displacement.

2. *Diffraction Off Baffles.* Consider the "gravity-wave" noise produced by diffraction of scattered light off vibrating baffles (the first process on Kip's next-to-the-last transparency.
   a. Compute $\tilde{h}(f)$ for the extreme worst-case scenario in which coherence increases the noise: Place the mirrors precisely at the center of the beam tube, assume the baffles are perfectly round and not serrated, and assume for each baffle that all points on the baffle's edge vibrate radially in phase with each other. Then light from all points on any chosen baffle will superpose coherently in $\tilde{h}(f)$. Give arguments why the various baffles should contribute incoherently with respect to each other. [Hint: one factor is the speed of sound along the vacuum pipe, which is $\sim$ 0.4km/sec; another deals with the baffle spacings.] Your final answer for $\tilde{h}(f)$ should be somewhat worse than the baffle backscatter noise of problem 1.
   b. The following factors mitigate the noise due to diffraction. For each factor make an estimate of the resulting reduction in $\tilde{h}(f)$. [Note that these mitigating factors do not act multiplicatively; the reduction in $\tilde{h}(f)$ is not equal to the product of the reductions due to the various factors. However, the net reduction makes $\tilde{h}(f)$ much less than baffle backscatter.] (i) The baffles will be serrated (jagged) with peak-to-valley serration heights of 3.5mm, which is somewhat larger than the width of a Fresnel zone (so as a baffle vibrates, some locations are alternately covering and uncovering an even numbered Fresnel zone, thereby producing phase shifts of one sign, while other locations are alternately covering and uncovering an odd numbered Fresnel zone, producing phase sifts of the opposite sign, and the two effects tend to cancel). There will be a $\sim$ 5 per cent irregularity in the

serrations on scales $\sqrt{\lambda L} \sim 4$cm. (ii) The mirrors will generally not be centered in the vacuum pipe, but rather will be off center by $\gtrsim 10$; and as a result, different regions of a baffle will intercept different Fresnel zones. (iii) The various points on a baffle do not vibrate in phase with each other. (iv) Each baffle will be out of round by a few millimeters in some random way.

LECTURE 16

Squeezed Light and its Potential Use in LIGO

Lecture by H. Jeff Kimble

Assigned Reading:

TT. C. M. Caves, "Quantum mechanical noise in an interferometer," *Phys. Rev. D*, **23**, 1693–1708 (1981).

UU. D. F. Walls, "Squeezed states of light," *Nature*, **306**, 141–146 (1983).

VV. M. Xiao, L. A. Wu, and H. J. Kimble, "Precision measurement beyond the shot-noise limit," *Phys. Rev. Lett.*, **59**, 278–281 (1987).

Suggested Supplementary Reading:

l. H. J. Kimble, "Quantum fluctuations in quantum optics—Squeezing and related phenomena," in *Fundamental Systems in Quantum Optics*, eds. J. Dalibard, J. M. Raimond, and J. Zinn-Justin, (Elsevier, Amsterdam, 1992), pp. 545–674.

m. "Squeezed States of the Electromagnetic Field," Feature Issue, *J. Opt. Soc. Amer.*, B4, 1450–1741 (1987).

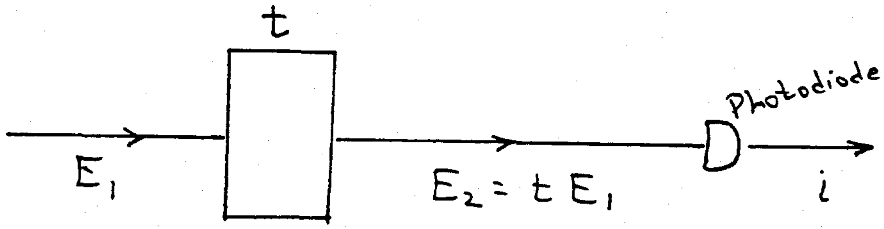
n. "Squeezed Light," Special Issue, *J. Modern Optics*, **34**, 709–1020 (1987).

o. "Quantum Noise Reduction," Special Issue, *Appl. Phys. B*, **55**, 189ff. (1992).

p. S. Reynaud, A. Heidman, E. Giacobino, and C. Fabre, "Quantum fluctuations in optical systems," in *Progress in Optics*, XXX, ed. E. Wolf (Elsevier, 1992), pp. 1–85.

A Few Suggested Problems:

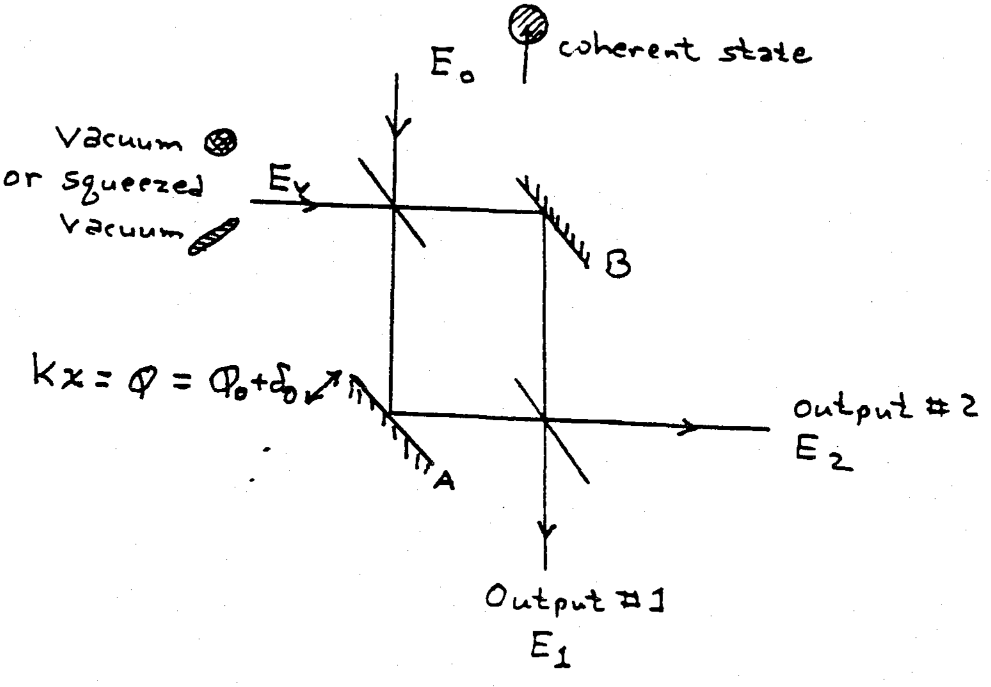
1. *Detection of Modulation in a Squeezed State.* An electromagnetic field propagates through a medium whose transmission coefficient is given by $t = t_0 e^{-\gamma(t)}$, where $\gamma(t) \equiv \gamma_0 \cos(\Omega_0 t)$ (i.e., sinusoidally modulated absorption with amplitude γ_0 and frequency Ω_0).
   a. Assuming that $\gamma_0 \ll 1$ and that the input field is in a coherent state (with frequency $\gg \Omega_0$), derive an expression for the minimum detectable value of γ_0, for a fixed input energy flux $\langle |E_1|^2 \rangle$ and a fixed bandwidth $B \equiv \Delta f$ (corresponding to a photodiode integration time $\hat{\tau} = 1/B$).
   b. If the input field instead is in a squeezed state, derive an expression for the minimum detectable amplitude γ_0. Illustrate in a "ball-and-stick" sketch the dependence of your answer on the orientation of the squeezing ellipse.



2. *Squeezed Vacuum in an Interferometer.* In Part IV of Kimble's lecture transparencies, he sketches a calculation of the minimum detectable phase deviation δ_0 when a coherent state is put into one port of the Mach-Zehnder interferometer shown below, and either the vacuum state or the squeezed vacuum state is put into the other port. His answer was $\delta_0 = 1/\sqrt{N}$ for the vacuum state, and $\delta_0 = (1+\xi S)^{1/2}/\sqrt{N}$ for the squeezed vacuum, where N is the total number of available photons, S is the squeeze factor $(-1 < S \leq 0)$, and $\xi < 1$ is the efficiency of the squeezing. Show, in a phasor diagram, the relative phase relationships for the fields that emerge from the outputs, and from your diagrams infer that to achieve the above optimal sensitivities with readout at output #1, the unperturbed position of mirror A should be adjusted so that the phase difference between the two paths along the two arms is $\phi_0 = \pi/2$. More specifically:

   a. Show the orientation of the squeezing ellipses relative to the coherent amplitudes for each of the two fields E_a, E_b that contribute to the total field E_1 at the output #1.

   b. Show how these two fields with their fluctuations sum to give a resultant E_1 that (for $\phi_0 = \pi/2$) produces noise in the photodetector below the standard shot-noise level $1/\sqrt{N}$ and a signal proportional to the phase deviation δ_0.



c. Note that for an efficiency $\xi \to 1$ and for perfect squeezing $S \to -1$, the above analysis and diagrams predict that the minimum detectable phase deviation becomes arbitrarily small, $\delta_0 \to 0$. Show that, in fact, if the interferometer system is perfectly lossless, and δ_0 is modulated so $\delta_0 = \Delta_0 \cos(\Omega_0 t)$, the minimum detectable modulation amplitude Δ_0 is actually $\Delta_0 \sim 1/N$. Calculate the corresponding length sensitivity Δx for the displacement of mirror A. Estimate the laser power required to achieve the sensitivity of the advanced LIGO, *if* this limit could be achieved.

d. In the above discussion it was tacitly assumed that the interferometer mirrors are so massive that light pressure fluctuations do not disturb them significantly. Suppose now that mirror A has a finite, small mass and is free to move in response to light pressure, and that we apply a feedback force to the back of the mirror, to counteract the time-averaged light-pressure force on its front. Show, using the phasor diagrams of parts a. and b., that when we improve our measurement of δ_0 (and hence of the mirror position x) by increasing the amount of squeezing, we increase the random light-pressure perturbations of the mirror, thereby enforcing the uncertainty principle. Relate this result to the standard quantum limit for sensing the position of the small mass, and thence to the curve labeled "Quantum Limit" in the plots of LIGO noise sources that were shown in earlier lectures. [For a quantitative analysis, in the context of a Michelson interferometer, see C. M. Caves, *Phys. Rev. D*, **23**, 1693 (1981). In this problem you are supposed to be ignoring the possibility of going beyond the standard quantum limit as discussed by Jackel and Reynaud, *Europhys. Lett.*, **13**, 301 (1990).]

LECTURE 17

The LIGO Vacuum System

Lecture by Jordan Camp

Assigned Reading:

WW. J. H. Moore, C. C. Davis, M. A. Coplan, *Building Scientific Apparatus* (Addison-Wesley, 1983), Chapter 3. "Vacuum Technology." A good overview of the basic issues involved in vacuum system design, including gas kinetics, pressure measurement, and pumping.

Suggested Supplementary Reading:

q. F. Reif, *Fundamentals of Statistical and Thermal Physics* (McGraw-Hill), Chapter 7: "Kinetic Theory of Dilute Gases in Equilibrium." A discussion of basic kinetic theory, including molecular flux, effusion, and pressure and momentum transfer.

r. J. O'Hanlon, *A User's Guide to Vacuum Technology* (John Wiley and Sons). Considerably more detailed than the assigned reading. Includes material vapor pressures and outgassing, calculation of conductances and a chapter on residual gas analyzers.

s. K. Welch, "The pressure profile in a long outgassing vacuum tube", *Vacuum*, **23**, 271–276.

A Few Suggested Problems

1. *Residual gas damping of test mass*: In Lecture 13, the following expression for the losses due to residual gas damping was given:

$$\phi(\omega) \sim \frac{2AP}{M}\sqrt{\frac{\mu}{kT}}\,\frac{\omega}{\omega_o^2}$$

where A, M are the test mass area and mass, P is the residual gas pressure, μ is the mass of a gas molecule, and the gas molecules are thermalized at temperature T. (Recall that $\phi = \gamma\omega/\omega_o^2$, where γv is the acceleration due to gas damping and v is the test mass velocity.) Derive this expression. For simplicity, assume that the gas molecules are of uniform velocity and normally incident on the test mass.

2. *Pressure in LIGO beam tubes:* The final pressure achieved in the LIGO beam tubes will depend on the outgassing rates, conductances and pumping speeds, and available budget.

a. Conductance of an orifice: the ion pumps, which will provide quiet, high vacuum pumping for the beam tubes, will be connected to the tubes through 25 cm diameter orifices. The conductance of an orifice of area A for molecular nitrogen (atomic weight=28) is given by C (in Liter/sec) = 11.6 A (in cm^2). How does this value scale with the molecular mass, and what is the conductance for hydrogen (atomic weight=2)? (Hint: the conductance is linearly related to the flux of molecules across the aperture). Assuming that the ion pumps have pumping speeds of 10000 L/sec, what is the combined pumping speed of the orifice and pump?

b. Conductance of a beam tube: in paper 4 of the suggested reading K. Welch derives the following expression for the average pressure of a long outgassing tube of diameter D and tube length l:

$$P_{av} = P_p + \frac{\pi q l^2}{3kD^2}$$

Here q is the outgassing rate (torr l/(sec-cm^2)) and k is a function of temperature and molecular weight (k=45 for hydrogen at room temp). The first term, P_p, is the pressure at the ion pump, while the second term accounts for the finite conductance of the 1.2 m diameter beam tube.

1) for the special LIGO low-outgassing steel, q ~ 1.0 x 10^{-13} torr l/(sec-cm^2). Assume an initial pumping configuration of 2 end pumps per each 2 km long beam tube module. What is the total outgassing flux (torr l/sec) seen by each of the pumps? Using the earlier calculation of pumping speed, find P_p. What is P_{av}? This number should be close to the goal of 1.0 x 10^{-9} torr for the advanced interferometer.

2) assume that unprocessed steel with a higher outgassing rate is used, where q ~ 1.0 x 10^{-12} torr l / (sec cm^2). What is P_{av} for this beam module? How many additional equally spaced pumps would be necessary to recover the desired value of P_{av}? (The 2 pieces of P_{av} scale differently with the # of pumps.) With a cost of $35 K per additional pump station and a total of 8 beam tube modules for the two sites, how much additional cost would be incurred if this steel were used?

LECTURE 18

The 40 Meter prototype Interferometer as an Example of Many of the Issues Studied in this Course

Lecture by Robert Spero

Assigned Reading:

XX. R. Weiss, *Quart. Prog. Rep. Res. Lab. Electron. M.I.T.* **105**, 54 (1972). [The seminal paper presenting in detail how laser interferometers can be used for gravitational wave detection, including a comprehensive analysis of noise sources.]

YY. Robert L. Forward, "Wideband laser-interferometer gravitational-radiation experiment," *Phys. Rev. D* **17**(2), 379–390 (1977). [A description of the first interferometric gravitational-wave detector to be built (a 2 m Michelson interferometer) and the first search for gravitational waves using such a detector (a coincidence run conducted in 1972 between the interferometer and several bar detectors).]

A Few Suggested Problems: *Note: Your homework for Lectures 17 and 18 is to be turned in to Shirley Hampton in room 151 Bridge Annex before 1:00PM Friday June 3*

1. In this course you have encountered all the significant noise sources for interferometric detectors that the LIGO team is now aware of. Which of these were unanticipated by Weiss in Ref. 1 above; and are they "fundamental" or are they "technical"?
2. The interferometer described in Weiss's paper has a different optical configuration from the 40 m interferometer, but the shot noise calculated by Weiss is similar to that achieved in the 40 m. Why? Compare the shot noise limited sensitivity calculated by Weiss with the sensitivity achieved by Forward, and account for the difference.
3. The thermal noise calculated by Weiss is based on viscous damping [$\phi(\omega) \propto \omega$]. How does the thermal noise prediction change when if the damping is structural [$\phi(\omega)$ independent of ω]? cf. Lectures 13 and 14.
4. Weiss calculated noise from laser intensity fluctuations acting on the test masses via radiation pressure. Intensity fluctuations also result in noise (in a manner discussed in Spero's lecture) if the interferometer's operating point is offset from a "dark fringe" at the photodiode. Under what circumstances will this noise be larger than that due to radiation-pressure fluctuations.

CHAPTER 7

Scattering from Optical Surfaces

J. M. ELSON, H. E. BENNETT, and J. M. BENNETT

Michelson Laboratories, Naval Weapons Center
China Lake, California

I. INTRODUCTION

The presence of stray light is a continuing problem in the design and performance of optical systems. If the systems are well designed most of it

REF. 33

Copyright © 1979 by Academic Press, Inc.
All rights of reproduction in any form reserved
ISBN 0-12-408607-1

comes from the optical components themselves. By proper baffling, it can be reduced in sensitive areas by orders of magnitude, but ultimately it is necessary to improve the components themselves. Most of the light is scattered by the component surfaces. Bulk scattering can also occur in windows or lenses, but it is typically one to two orders of magnitude less than surface scattering (Kozawa, 1962).

Surface scattering may arise from (1) irregularities such as scratches, digs, or particulates which are large relative to the wavelength of the incident light, (2) isolated irregularities which are comparable to or smaller than the wavelength of incident light, and (3) irregularities which are small in one or more dimensions relative to the wavelength but which are so closely spaced that they cannot be treated as independent scattering centers. The effect of each center is correlated with that of its neighbor's; scattering from such ensembles is often called "microirregularity scattering."

The appropriate theoretical treatment of scattering depends on which of the above categories the scattering centers fall into. Historically, scattering from scratches and other macroscopic defects was probably recognized first. It is probably the easiest of the three to visualize and can be handled by using geometrical optics. Oddly enough, it is the most difficult one to handle quantitatively since the shapes of the macroirregularities must often be known in great detail. When the irregularities become comparable to or smaller than the wavelength of the incident light, the scattering problem becomes a diffraction problem, and geometrical optics is no longer adequate. A great simplification occurs, however, if the scattering centers can be considered to act independently. Mie scattering theory, of which Rayleigh scattering is a special case, can then be used. Historically, these theories were initially worked out near the turn of the century, although modifications and improvements are continually being made. Latest to be developed is the correlated scatterers theory. The first hint of such a theory of which we are aware was given by Chinmayanandam (1919). However, most of the work in this area has been done in the past 20 years and it is still continuing.

In the visible and ultraviolet regions of the spectrum, microirregularities are the principal source of scattered light for most optically polished surfaces. The average height of these irregularities is only a few nanometers and low scatter surfaces for these wavelength regions are often specified in terms of their rms roughness. Typical glass optical flats have rms roughnesses of 25 to 30 Å; good polished metal surfaces typically run from 30 to 50 or 60 Å; and, any surface under 15 Å rms is often called "superpolished." Roughnesses as low as 5 Å rms have been achieved; since the scatter goes as the square of the roughness, these surfaces scatter 25–35 times less than a well-figured conventional glass optical flat.

Microirregularity scattering drops exponentially with wavelength and in the near infrared scattering from surface blemishes, scratches, dust, etc., typically becomes the limiting factor. It is often nearly independent of wavelength. Pitting of the surface by sand or rain erosion increases the importance of defect scattering and it may become the dominant effect even in the visible region. Finally, scattering from isolated particulates such as dust must be considered. Usually dust is not nearly as much of a problem on optics as the damage which is frequently done to the surface in trying to remove it. It does affect the optical performance of low scatter components even when used in the visible and ultraviolet regions, however, and since dust particles have diameters of about 1 μm, it becomes particularly important in the near infrared.

In this chapter we will discuss the general methods which have been used to calculate light scattering. These include, principally, the scalar scattering theory and vector scattering theory, both of which treat scattering from correlated surface microirregularities. Mie theory deals with scattering from isolated surface defects such as dust or other particulates and can explain the experimentally observed increase in total integrated scattering (TIS) near a wavelength of 1 μm. A detailed discussion of TIS includes microirregularity scattering, surface plasmon excitation, scattering in the infrared, and scattering from scratches and other defects. The section on angular dependence of scattering deals primarily with scattering from correlated surface microirregularities. In the simplest case one can consider angular scattering from a sinusoidal grating, since a randomly rough surface can be considered to be composed of a two-dimensional superposition of sinusoidal gratings having different amplitudes, periods, and phases. The complete vector scattering theory will be discussed and it will be shown how the theory can be modified to use measured surface statistics to calculate the angular dependence of scattering. Scattering from dielectric multilayers with rough interfaces is a considerably more difficult problem. Polarization effects can cause large differences in the scattered light depending on whether or not the surfaces of the layers replicate the roughness of the substrate. Angular scattering curves for surfaces with replicating and nonreplicating roughness will be presented. The conclusion will summarize the present status of scattering from optical surfaces both from a theoretical and an experimental standpoint.

II. METHODS FOR CALCULATING LIGHT SCATTERING

Light scattering calculations have been handled using various approaches. These include geometrical optics, scalar theory, vector theory, particulate scattering from polarizable particles (Mie theory), and numerical calculational methods. In this section different approaches will be briefly discussed and then compared.

A. Geometrical Optics

Geometrical optics, or ray optics, can be applied to light scattering when the dimensions of the surface roughness are large compared to the wavelength of light. A typical geometrical optics approach assumes that the scattering surface consists of plane, flat facets having lateral dimensions that are large relative to the wavelength and which reflect like plane mirrors. Each facet has a surface normal and there will be a statistical distribution of the directions of the surface normals. Thus, scattering from a geometrical optics type of rough surface consists of adding up the contributions from each tilted facet of the surface. In order to actually treat a surface using this approach, one needs to have a characterization of the surface: either a model of the shapes and distribution of surface facets or an equivalent statistical characterization. A statistical model of a facet surface generated by a Markov chain has been given by Beckmann and Spizzichino (1963, see especially pp. 109–114). However, surface characterizations of geometrical optics-type surfaces are not easy to obtain. Furthermore, geometrical optics fails to describe the effects of interference, diffraction, and polarization. Fortunately, these phenomena are included in scalar and vector scattering theories, and scalar scattering theory can be extended to the geometrical optics limit to describe scattering from surfaces whose roughness is much larger than a wavelength of light. For this reason, geometrical optics calculations are rarely used in treating optical scattering problems.

B. Scalar Theory

The starting point of scalar theory treating scattering from rough surfaces is the Helmholtz–Kirchhoff diffraction integral.* This integral, and the resulting diffraction formula, is based on the fact that the solution to the wave equation at some point in space surrounded by an arbitrary closed surface (a mathematical construction) may be obtained if the solution is known at all points on the surface. When applied to scattering from rough surfaces, strictly speaking, the fields must be known at the rough surface itself. However, these fields are not generally known and to overcome this problem certain approximations, known as the Kirchhoff boundary conditions (Jackson, 1962, p. 282), are made. These approximations limit the validity of the scalar theory to scattering near the specular direction; polarization properties of the scattered light are not included. The major problem in scattering theory is to find the relationship between the statistics of the

scattering surface, those of the scattered wave front leaving the surface, and the far-field statistics of the scattered light. Surfaces whose roughness is much less than the wavelength are sometimes termed "weak scatterers." They produce phase variations in the near field of less than 2π. On the other hand, surfaces whose rms roughness is greater than a wavelength ("strong scatterers") produce phase variations in the near field much greater than 2π. Different types of statistical treatments are necessary for the two types of scatterers. An excellent review article on this subject has been written by Welford (1977) and the book by Beckmann and Spizzichino (1963) is also a primary reference.

Scalar theories have been applied to surfaces whose roughness is much less than the wavelength of light (Davies, 1954; Bennett and Porteus, 1961; Porteus, 1963). Davies (1954) has given a relation predicting the angular dependence of scattered light from rough surfaces at angles near the specular direction. He assumed that the surfaces were perfectly conducting, i.e., that the specular reflectance of a perfectly smooth surface of the material or the total reflectance of a rough surface of the same material, would be 100%. The surfaces were also assumed to have Gaussian height distribution functions and Gaussian autocovariance functions. Bennett and Porteus (1961) modified Davies' relation to include the actual reflectance of the material and then experimentally verified that the relations were valid for predicting the specular reflectance of rough surfaces when the roughness was much smaller than the wavelength. Porteus (1963) expanded the relation of Bennett and Porteus to include arbitrary height and autocovariance functions. The restriction that the roughness be much smaller than the wavelength was also removed.

Scalar theories based on the Helmholtz–Kirchhoff diffraction integral have also been applied to surfaces whose roughness is much greater than the wavelength of light (Davies, 1954; Chandley and Welford, 1975; Leader and Fung, 1977; Holzer and Sung, 1976, 1977; Chandley, 1976). Chandley (1976), Eastman and Baumeister (1974), and Leader (1971a) have measured surface statistics and angular scattering from rough surfaces, and have compared these measurements with predictions from theory. In the limit when the dimensions of the surface facets become very large relative to the wavelength, there is a transition between the scalar diffraction theory and geometrical optics. Hagfors (1966) has shown that when a rough surface is assumed to have a Gaussian autocovariance function, the limit of scalar diffraction theory yields the correct geometrical optics scattering result. On the other hand, Fung and Moore (1966) showed that when an exponential autocovariance function is assumed for the surface, the limit of the scalar diffraction theory is in disagreement with geometrical optics. This latter result is reasonable since an exponential autocovariance function has an

* Several standard books on electromagnetic theory discuss this topic. See, for example, Jackson (1962).

unphysical discontinuous nonzero slope at the origin, which implies that the rms slope of such a surface would be infinite. Any real surface that approximated this condition would have a very jagged surface profile and would not appear facet- or mirror-like in the geometrical optics regime.

C. Vector Theory

To include the vector nature of the scattered field requires somewhat different techniques. Two dominant methods are considered here. The first is analogous to the Helmholtz–Kirchhoff scalar integral and is the vector equivalent or Stratton–Chu–Silver (SCS) integral (Silver, 1947). The other is a perturbation method, of which there are several variations. The SCS integral was originally applied to diffraction problems and is based on the calculation of radiation from a distribution of currents over a surface. The integral fully contains the vector nature of the currents (based on the surface fields) and yields the radiation from the distribution of surface currents. In principle, the integral can yield the scattering from any magnitude of surface roughness or shape, but in practice the results are often limited to roughnesses that are small compared to the wavelength. One major reason for this limitation is that the actual fields at the rough surface are not known (as is necessary to properly evaluate the integral) and, consequently, approximations are used. A number of authors have utilized the SCS integral as a starting point to calculate the scattering from rough surfaces, including Leader (1971b) and Fung and Chan (1969).

The perturbation technique, which was first used by Rayleigh for acoustical scattering (Rayleigh, 1945), depends on the surface roughness having a weak influence on the perfectly smooth situation. Under these assumptions it is justifiable to let the surface fields be approximated by the smooth surface fields. This assumption is generally made when the rms roughness δ is much less than the wavelength λ, so that only a small fraction of the total incident power is scattered out of the specular beam. Under these conditions the theory is termed first order because scattered field terms proportional to $(\delta/\lambda)^2$ or higher are dropped. Since the scattered fields are retained to an accuracy of δ/λ, then the boundary conditions need to be of similar accuracy, and the scattered power (Poynting vector) will be proportional to $(\delta/\lambda)^2$. In principle, perturbation theory may be used to calculate successively higher-order solutions, each of which are corrections to the zero-order or unperturbed solution. Normally, solutions beyond the first order are quite complicated.

There have been several variations of perturbation techniques. These include a plane wave expansion method by Peake (1959), a Dirac δ-function current model by Kröger and Kretschmann (1970), and Maradudin and Mills (1975), and a coordinate transformation method by Elson and Ritchie (1974) and Elson (1975). Done properly, all these variations lead to the same results. Quantum mechanical perturbation methods are given and discussed by Elson and Ritchie (1971) and Celli *et al.* (1975).

D. Particulate and Resonance Scattering

Another possible contribution to scattering arises from surface contaminants such as dust particles or particulates. In the simplest case, the particulates may be approximated as dielectric spheres with no interaction between the particulates and the surface on which they rest. Mie (1908; see also Stratton, 1941; van de Hulst, 1957) has solved the problem of electromagnetic radiation interacting with spherical scatterers. Mie theory has also been used to calculate absorption by spheres at a metal interface by Beaglehole and Hunderi (1970). The calculation of absorption and scattering cross sections is a complicated boundary value problem soluble only for particle shapes with a high degree of symmetry. This problem, in addition to having been solved for spheres (Mie's solution), has also been solved for infinite cylinders and ellipsoids including as limiting cases thin disks and needles (van de Hulst, 1957, p. 70). When the particle dimensions are much smaller than the wavelength, the particle shape becomes less important. The particles then tend to behave like oscillating dipoles and produce dipole radiation characteristic of Rayleigh scattering (Jackson, 1962, pp. 573, 603).

In modeling a particulate-covered surface, one can consider that the particles are distributed in a random manner on a plane surface. Scattering may then be calculated by adding the phases of the light backscattered by the particles to the phases of forward scattered light which is subsequently reflected by the plane surface. In general, the contribution of the forward scattered light is much larger than the contribution of the backscattered light.

Spherical particles exhibit reasonances in absorption (and also scattering) which may be distinguished as dielectric resonances and size resonances. Dielectric resonances may occur when the size of the particle is much less than the wavelength. In this Rayleigh region, the resonances are dipolelike, occurring when the real part of the dielectric constant of the sphere equals -2. The dipole resonance causes much higher intensity fields to be generated within the sphere and results in increased absorption and scattering. Size resonances may occur when the physical size of the scatterer and the wavelength relate in such a way as to produce internal standing waves similar to a resonant cavity. The standing waves constructively interfere to yield enhanced surface fields and hence scattered radiation.

Other efforts to explain scattering and related optical effects of particulates have considered ellipsoidal-shaped objects. The ellipsoids are assumed to model aggregates in metal films such as silver, where the aggregate sizes may vary from 100 to 4000 Å. There have been a limited number of theoretical and experimental investigations of scattering from such aggregated metal films (Truong and Scott, 1976, and references cited therein).

E. Numerical Methods

All of the surface scattering results discussed previously yield a solution in closed form. To obtain these solutions it is necessary to make various approximations and specify the ratio of the rms roughness to the wavelength. With the advent of high-speed computers, numerical solutions are often practical. The usual method involves a solution to the wave equation as an infinite series expansion. The coefficients are determined numerically by solving equations for the appropriate boundary conditions. The majority of the solutions that have been obtained by numerical methods are for scattering from grating surfaces. This is because a known surface profile is needed in order to specify the regions across which the electromagnetic theory boundary conditions must be satisfied. Numerical techniques are especially useful when applied to problems of high efficiency gratings where the grating amplitude is not small compared to the wavelength. It is usually possible to truncate the numerical algorithm after the desired accuracy has been attained. Some authors have calculated the diffraction expected from a colinear array of line currents (Zaki and Neureuther, 1971). Several different methods (Rayleigh, integral, differential) are outlined by Petit (1975). Berreman (1970) has calculated the expected change in coherent reflectance caused by a surface with hemispherical pits or bumps. The technique uses the computer to obtain a solution for general shapes; however, some simple examples are given in closed form. Because of the problem of specifying the surface profile, numerical methods have not been applied to scattering from surfaces having random roughness.

F. Comparison of A–E Methods

No one method can be singled out as best for calculating scattering from optical surfaces. When the surface roughness is much larger than the wavelength, geometrical optics may be applicable if the exact dimensions of the surface features can be specified. The Helmholtz–Kirchhoff diffraction integral provides the basis for a scalar theory which has been widely used to calculate scattering from surfaces which have correlated microirregularities. The theory does not provide polarization information, but it may be applied to situations where the roughness is much less than or much greater than the wavelength. It is most useful for predicting total integrated scatter (TIS), i.e., scattering in all directions, but has also been used in a limited way to predict the angular dependence of scattering. The accuracy of the angular dependence prediction is greatest for scattering near the specular direction and is most useful for highly reflecting metals where polarization effects are minimal. A widely used application of scalar theory has been the determination of the relation between the surface roughness and the decrease in specular reflectance of the rough surface (Bennett and Porteus, 1961; Porteus, 1963).

Vector theory is generally an improvement over scalar theory because polarization effects are included. The SCS integral is somewhat similar to the Helmholtz–Kirchhoff diffraction integral. The former is a surface integral which sums, at an observation point, the radiation from vector surface currents, while the latter sums the scalar phases from different points on the surface. When surface roughness is present, incoherent or scattered fields are produced. The SCS integral is formally exact but in practice cannot yield an exact solution. This is because the fields on the surface are not precisely known beyond reasonable approximations.

Perturbation methods also yield vector solutions for the incoherent field. These methods can, in principle, yield high-order solutions by iteration. Solutions beyond first order are generally not given because of complexity.

Single particle scattering, handled by the Mie theory, is especially useful for calculating scattering from a particulate-covered surface. Mie theory can handle absorption as well as scattering and treats both dielectric and metal particles. It is most easily applied to spherical objects and cannot treat scattering from correlated surface microirregularities.

While all of the previous approaches may offer closed form solutions, numerical approaches do not. In some cases the accuracy of solution may be quite good although the convergence of the solution may be poor. This could cause excessive computer run times. Usually the numerical method is especially chosen for the type of scattering surface to be evaluated. Random roughness is not considered in numerical techniques.

In this chapter, a perturbation method will be used to explore various aspects of angular scattering from a rough surface and from multilayers covering a rough surface. The results will be limited to first order, and are valid when the roughness is much less than the wavelength. The first topic to be discussed in detail is TIS from surfaces whose roughness is small compared to the wavelength. This will be handled principally by scalar scattering theory.

III. TOTAL INTEGRATED SCATTER

Total integrated scatter refers to the integrated sum of light scattered into all directions within the scattering hemisphere. It is not necessary to be aware of the angular distribution of the scattered light—only the fraction of the reflected light which is scattered out of the specular beam. This scattering may arise from (1) irregularities such as scratches, digs, or particulates which are large relative to the wavelength of light, (2) isolated irregularities which are comparable in size or smaller than the wavelength, and (3) correlated irregularities which have heights which are small relative to the wavelength but which cover the entire surface. Scattering from these correlated irregularities is often termed "microirregularity scattering," and is usually the dominant source of scattered light from optical components used in the near infrared, visible, and ultraviolet. At longer wavelengths scattering from scratches or particulates is often dominant.

A. Microirregularity Scattering

Figure 1 shows a micrograph of a well-polished metal surface with a measured value of the rms microroughness, $\delta = 35.5$ Å. The microirregularities are revealed by differential interference contrast (Nomarski) microscopy, an interferometric technique which is sensitive to height differences of fractions of a nanometer if the surfaces have jagged irregularities with steep slopes. These microirregularities cover the entire surface and are the dominant source of scattered light in the ultraviolet, visible, and near infrared regions of the spectrum.

FIG. 1. Differential interference contrast micrograph of a polished copper surface with an rms roughness of 35.5 Å. (Bennett, 1978.)

Statistically, the surface microirregularities can be described by a height distribution function and an autocovariance function. It is only necessary to understand what the minimum effective lateral separation between surface features will be. For reasons to be described later, the minimum effective lateral distances are equal to the wavelength of the light incident on the surface. Irregularities having lateral separations smaller than this minimum value are averaged with their neighbors to produce an area having an average height with respect to the mean surface level. It is thus perfectly possible to have these small "building-block" areas differ in average height by a fraction of an angstrom, even though atomic separations are ~4 Å and the crystallites in the surface may have steps of several hundred angstroms when viewed under an electron microscope.

The heights of these building-block areas above and below the mean surface level can be described by a "height distribution function." Experimentally, it is found that in nearly all cases the height distribution functions are Gaussian to a very good approximation. (Examples of some measured height distribution functions will be shown later in this section.) We may then describe the surface heights by an "rms roughness" as defined for a Gaussian height distribution function. The fraction of the total reflected light (specular plus nonspecular) scattered away from the specular direction by microirregularities is described by a simple scalar scattering theory based on the Kirchhoff diffraction integral. The dependence of TIS on wavelength is given by (Bennett and Porteus, 1961)

$$\mathrm{TIS} \equiv 1 - (R/R_0) = 1 - \exp[-(4\pi\delta\cos\theta_0/\lambda)^2] \cong (4\pi\delta\cos\theta_0/\lambda)^2 \qquad (1)$$

where R_0 is the fraction of the incident light which is reflected into all angles including the specular direction, R is the fraction which is specularly reflected at an angle θ_0, the angle of incidence, and δ is the rms height of surface microirregularities. Since the TIS is a function of the height but not the lateral separations of the microirregularities (so long as they fall within an appropriate range) and since real optical surfaces typically have Gaussian height distribution functions with irregularities which are symmetric about the mean surface level, it is possible to determine the rms value of the microroughness uniquely from TIS measurements. The rms microroughness is thus a simple method for specifying the quality of an optical surface with regard to scattering by microirregularities.

Some materials can be polished more easily than others and yield surfaces with lower microroughness which scatter less light for an equivalent amount of polishing effort. Techniques can often be developed to reduce the surface microroughness for a given material, so there is no single roughness value which a given material polishes to. However, optical glasses typically can be more easily polished to low rms roughness values than can most metals and other crystalline materials. Neither glasses nor crystalline materials automatically polish to very low rms roughness values, say under 15 Å rms. Special techniques are required to produce such surfaces, which are frequently referred to as being "superpolished." Recently, techniques have been developed to polish some surfaces in such a way that the scattering from them is optically equivalent to microroughness of 5 Å rms or less. Such surfaces, which scatter only about $\frac{1}{25}$th as much as a typical polished glass surface, are sometimes called ultralow scatter surfaces.

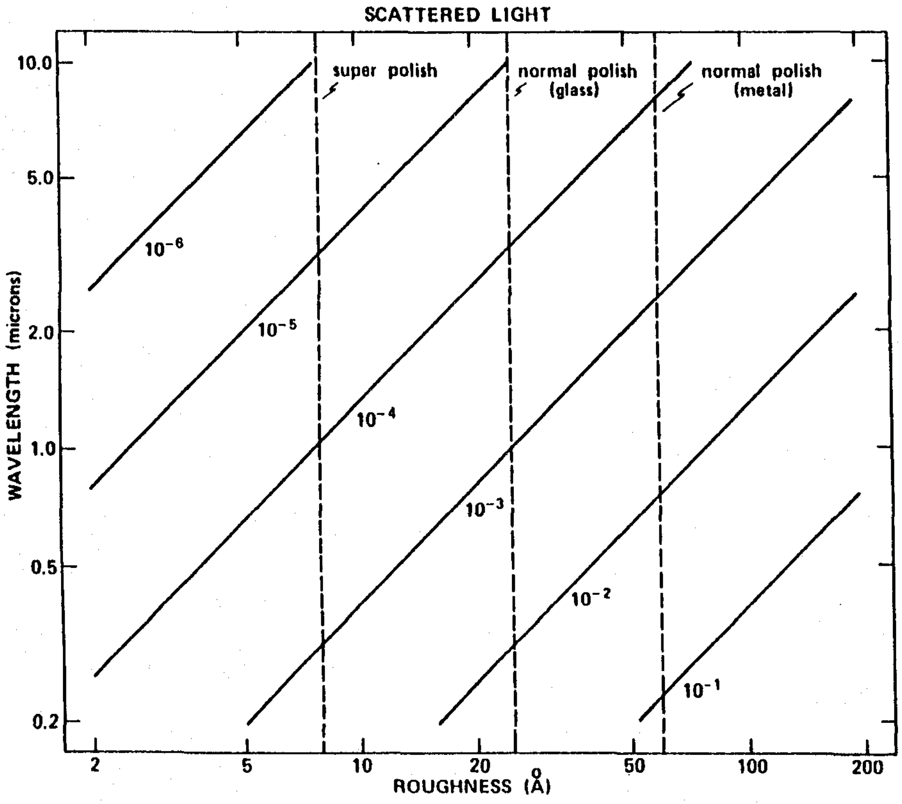


FIG. 2. Scattered light levels predicted theoretically (diagonal lines) for surfaces having rms roughnesses from 2 to 200 Å: The dashed lines indicate typical roughnesses of various kinds of polished surfaces. Wavelengths from the ultraviolet to the infrared are plotted logarithmically on the ordinate. (Bennett, 1978.)

If the rms roughness in Eq. (1) is plotted as a function of wavelength on log–log paper for a constant scattering level, a straight line results. A series of these scattering levels is shown in Fig. 2. By running a horizontal line across the graph at the wavelength of interest, it is possible to determine for a given roughness surface what the approximate scattering level will be. For example, at a wavelength of 5000 Å a superpolished surface will scatter less than 0.1%, whereas a conventionally polished glass surface will scatter several tenths of a percent and a typical polished metal surface will scatter over 1%. Recall that these percent values refer to the total reflected light which can vary markedly between dielectrics and metals (dielectrics can transmit much of the incident light).

In most cases, the scattered light obeys Eq. (1) quite well in the visible, ultraviolet, and near infrared regions of the spectrum. Figure 3 shows a typical plot of total scattered light as a function of wavelength in this wavelength region. In some cases, agreement is not so good. For example, if the sample is heavily scratched or coated with dust or particulates, the height distribution function is no longer Gaussian. Examples of measured height distribution functions are shown in Fig. 4. Many polished surfaces have

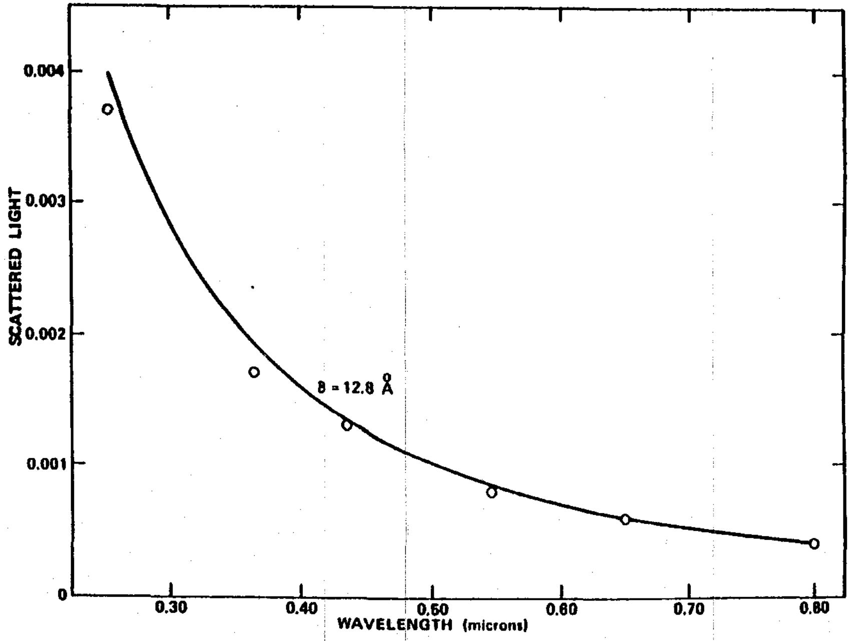


FIG. 3. Wavelength dependence of the total scattered light from an aluminum-coated superpolished fused quartz sample: (O) measured values; ——— calculated from Eq. (1) assuming a value of 12.8 Å for δ. (Bennett, 1978.)

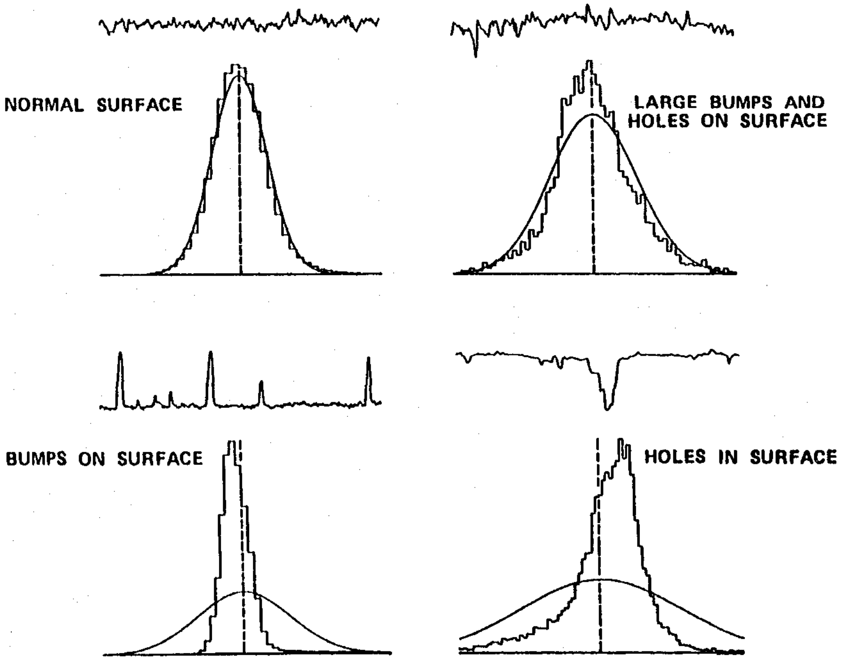


FIG. 4. Effect of the surface profile on the height distribution function for four types of surfaces. The surface profiles are shown above the corresponding histograms.

Gaussian height distribution functions similar to that shown in the upper left of the figure. (In this type of presentation the histogram is obtained from measured height data. The length of each bar represents the fraction of the total number of surface features which have heights equal to the value given on the abscissa, which is measured relative to the mean surface level (dashed vertical line). The smooth Gaussian curve encloses the same total area and has a half width derived from the measured rms roughness of the surface.) Some very soft materials such as polished KCl have proportionately too many large bumps or deep scratches. This distorts the measured histogram (upper right of Fig. 4), putting too much contribution in the tails. The influence of the extrema on the surface makes the rms roughness proportionately too large, so the half width of the Gaussian is also too large. TIS-derived values of the rms roughness tend to be higher than measured values for this type of surface. If a surface is extremely smooth but has large dust particles on it, a distorted histogram is obtained with the maximum shifted to negative values (because the particles have influenced the placement of the mean surface level). Similarly, an etched or pitted surface which is nominally smooth but has deep holes in it can have a distorted histogram with the peak shifted in the other direction. Scattering from particulate-covered or etched surfaces would not be expected to follow the relation in Eq. (1). Scattering from scratches also tends to be nearly independent of wavelength. If the autocovariance length of the surface is small enough so that it becomes shorter than some of the wavelengths tested, discrepancies will arise. It is thus useful when specifying quantitative surface microroughness values to specify the wavelength range in which the "equivalent surface microroughness" was measured. Convenient wavelengths are the visible mercury lines 4358 and 5461 Å, and the red HeNe laser line 6328 Å, or the red krypton laser line at 6471 Å.

Figure 5 shows typical roughness and the range of roughnesses, mostly derived from TIS measurements which were made at the Naval Weapons Center (NWC) on various well-polished substrate materials. The substrates were either polished in-house or were sent to NWC for evaluation by commercial vendors. In all cases, it is easy to obtain larger roughnesses, but reducing the lower limit is much more difficult. The average values shown in Fig. 5 are not representative of the quality of surface finish one would obtain by purchasing a polished mirror from any optical firm. For example, the copper mirror roughnesses are averages of samples polished by the Northrop Corporation, Battelle Northwest Laboratory, Perkin-Elmer Corporation, Spawr Optical, and NWC. They are not conventional copper mirrors but represent the best samples produced by these firms. The best bulk copper sample we have yet measured had an rms roughness of 15 Å rms, scattered only one quarter as much as the average good sample, and

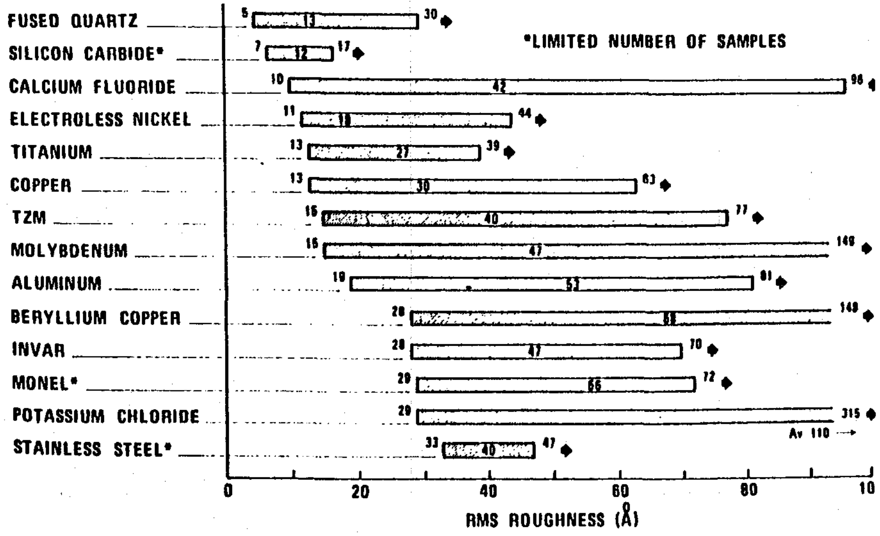


FIG. 5. Average rms roughness and range of roughnesses measured for polished optical surfaces.

over an order of magnitude less than conventionally polished mirrors. Because such a low microroughness has been achieved on copper is not a reason to specify bulk copper mirrors with an rms roughness of 15 Å unless the purchaser is ready to support a major development effort to push a technique from the laboratory one-of-a-kind stage to a production process. The same argument applies to the roughness values listed for the other materials.

B. Surface Plasmon Excitation

We have assumed in the above discussion that the optical constants of the coating do not affect the fraction of the reflected light which is scattered. This assumption is usually valid but is not justified in wavelength regions where surface plasmon excitations are important (Bennett and Stanford, 1976). Figure 6 shows the fraction of the incident light which is scattered from a polished glass surface coated with CaF_2 (to increase its microroughness) and then with silver. Dielectric overcoats of MgF_2 were then deposited on the silver. The long-dashed line shows the scattering level which is predicted from the scalar scattering theory, Eq. (1). The solid line shows the difference between the scattering level predicted from the simple theory and

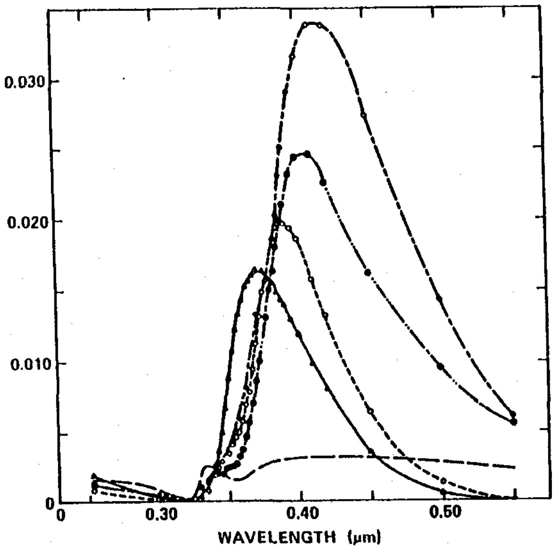


FIG. 6. Light scattered near the band edge in silver deposited on CaF_2-coated glass substrates: (- - -) scattering predicted from scalar theory; other curves show the increase in scattering caused by surface plasmon effects in (· ---) bare silver and silver coated with films of MgF_2 [(- ---) 150 Å MgF_2, (- - - -) 700 Å MgF_2, (--- · · ·) 120 Å MgF_2]. (Bennett and Stanford, 1976.)

that actually observed for the silver surface. The other dashed lines show the increase in scattering observed when the silver was overcoated with films of MgF_2 of various thicknesses. The maximum scattering level observed is an order of magnitude higher than would be predicted from the simple theory. If the metal coating had been aluminum, for example, instead of silver, scattering would have more nearly followed the simple theory. The character of the microroughness in this case made the effect unusually large; most silver-coated mirrors show much less dramatic effects in this wavelength region. The point, however, is that the optical constants of the surface can affect the scattered light levels observed in some cases and some of these are important technologically.

C. Scattering in the Infrared

In the infrared region scattering is usually higher than would be predicted from Eq. (1) for surfaces whose equivalent microroughness values are determined from visible measurements. A typical example of this effect is seen in Fig. 7. The scattering from an aluminized polished dense flint glass fits Eq. (1), which plots as a straight line on log–log paper, very well in the ultraviolet, visible, and near infrared. At a wavelength of 10 μm, however, the scattered light is an order of magnitude higher than predicted by the

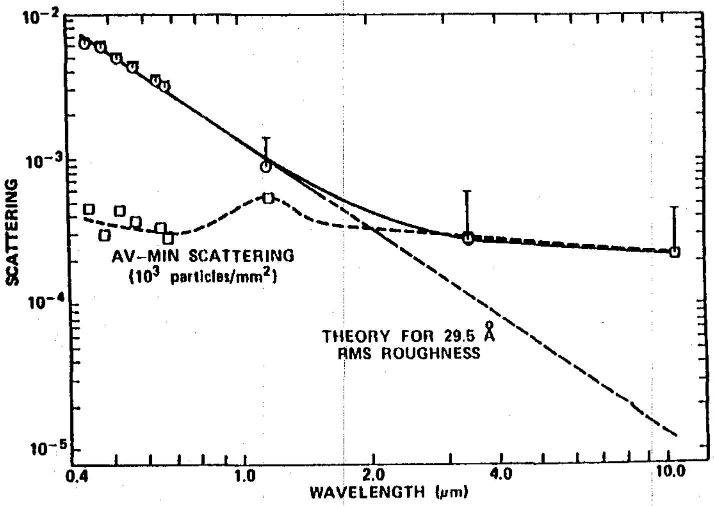


FIG. 7. Scattering from aluminized, polished dense flint glass. The diagonal line gives the contribution predicted for microirregularity scattering by a 29.5 Å rms roughness surface: (○) minimum scattering observed; bars and squares indicate the difference between average and minimum scattering observed at several points on the surface (10^3 particles/mm^2). (Bennett, 1978.)

theory. This discrepancy, which is commonly observed for surfaces in the infrared, may arise from several sources. The scattering angle as measured from the specular direction increases with increasing wavelength for correlated surface features having a given lateral size and separation (as can be seen from the grating equation). Long-range features with small average slopes, which at shorter wavelengths would produce light reflected in nearly the specular direction, will at longer wavelengths deflect light into larger angles where it will be detected as scattered light. A second and more important source of additional scattered light in the infrared is the presence of particulates, scratches, and various imperfections on the surface. The amount of light scattered from many of these imperfections is nearly independent of wavelength and since microirregularity scattering decreases exponentially with the square of the wavelength, at some point scattering from the larger surface imperfections will become the dominant source of scattered light. This effect is shown in Fig. 7. The squares are the differences between the average scattered light values at a given wavelength for a series of small areas on the surface and the minimum scattered light value observed from any of those small areas. These differences are nearly independent of wavelength. When the average scattering level reaches the difference value it stops decreasing and also becomes nearly independent of wavelength. This difference in scattering from point to point on the surface is believed to be caused by particulates and surface defects which are large compared to the wavelength and thus should be nearly independent of wavelength. A resonance should occur near a wavelength of 1 μm as a result of scattering by dust particles, which have a diameter of about 1 μm on the average. The surface tested was slightly dusty; dark field illumination revealed about 10^3 scattering sites/mm^2 of various sizes. It is thus not surprising that an increase in the difference values was observed in the 1-μm wavelength region.

D. Scratches and Other Macroscopic Defects: The Scratch/Dig Specification

Scratches and other macroscopic defects are widely recognized sources of scattered light from optical surfaces. The only official military specification on optical surface quality, MIL-13830A, refers to the allowed magnitude and density of these defects. As it is now formulated, both scratches and digs are defined by their width in micrometers. A dig having a diameter of 40 μm is a No. 40 dig, but a No. 40 scratch has a width of only 4 μm. Scratch and dig values are conveniently measured using a traveling microscope with Nomarski or dark field optics. Although digs (i.e., pits in the optical surface) have always been specified in terms of their diameter in

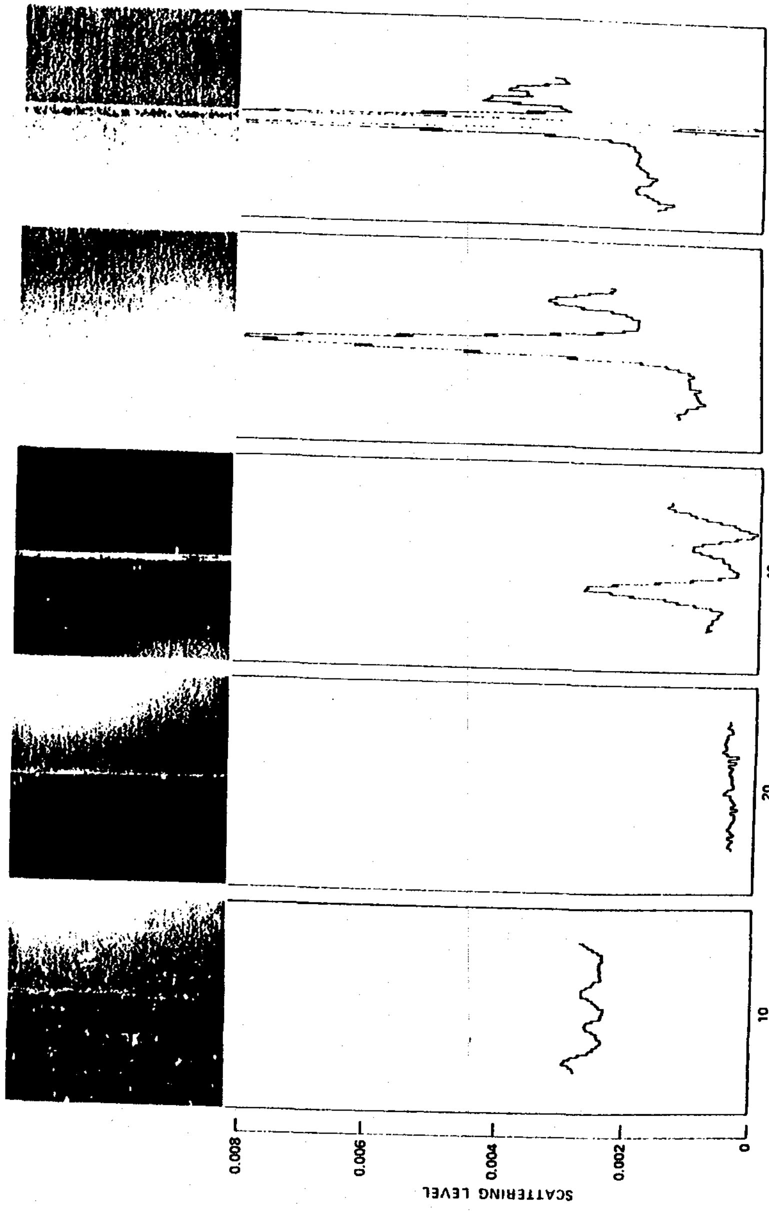


FIG. 8. Differential interference contrast micrographs and scattered light scans at 6328 Å for Frankford Arsenal Scratch Standard Serial No.F-66-1268. The apparent reversal at the center of the No.80 scratch is caused by overloading the electronics. The anomalously high scattering level [illegible]

micrometers, it is only recently that the width of scratches has been used to define them. Previously, sets of standard scratches were furnished. By visual comparison the inspector matched these standard scratches to scratches on the component to be tested. Although this test is quite subjective, it has been more or less successfully used for many years in the optical industry. Scattering at 0.6328 μm from a set of standard scratches furnished by Frankford Arsenal is shown in Fig. 8. The anomalously high scattering level observed for the No. 10 scratch sample is caused by microirregularity scattering, which completely masks the scattering from the fine scratch on this particular scratch standard.

Defining the scratch in terms of the scratch width makes the scratch specification significantly more quantitative than has been true before. There is not a unique relation between a scratch's width and its scattering behavior, however, as is illustrated by Fig. 9. The top set of scratches is an early standard set from Frankford Arsenal, which had cognizance of this particular military specification. These scratches have widths which are approximately in agreement with the new scratch identification scheme. The lower set of scratches are those on an experimental set of scratch standards produced under Frankford Arsenal control but having widths nearly an order of magnitude larger than those of the earlier standard. Scattering levels are similar to those found for the upper set of scratches, mainly because most of the scattered light comes from the sides of the scratches; the scratch bottoms are quite smooth. In practice the bottoms of scratches are not normally smooth and wide scratches tend to scatter more than narrow scratches. Since the width and scattering level of scratches can be adjusted separately, it would be possible to develop a set of scratch standards which not only had the specified widths but which scatter according to a predetermined scattering ratio. The narrow standard scratches would then look like weaker scatterers than the wider scratches. This scheme, in effect, was used in the original scratch standards; inspectors visually matched the appearance of the standard scratches to those of the sample tested. Unfortunately, visual observation is not a good way to detect differences in intensity. By using a more accurate detector it would be possible to quantify this procedure and arrive at a useful secondary test for surface quality which would be particularly useful for infrared optics. It has merit in the visible region also. Laser damage often occurs at scratches (Bloembergen, 1973), which must thus be eliminated for laser optics. In addition, in the process of polishing out the scratches a surface finish on glass surfaces is usually generated which is adequate for most noncritical applications. The main advantage of the scratch/dig specification is that it lends itself to volume production of parts and when properly set up is quick to use. Its main disadvantages are (1) it is not at present satisfyingly quantitative as far as scattering level goes and (2) it is not directly related to the source of most of the scattered light from optical surfaces in the visible, ultraviolet, and near infrared regions of the spectrum.

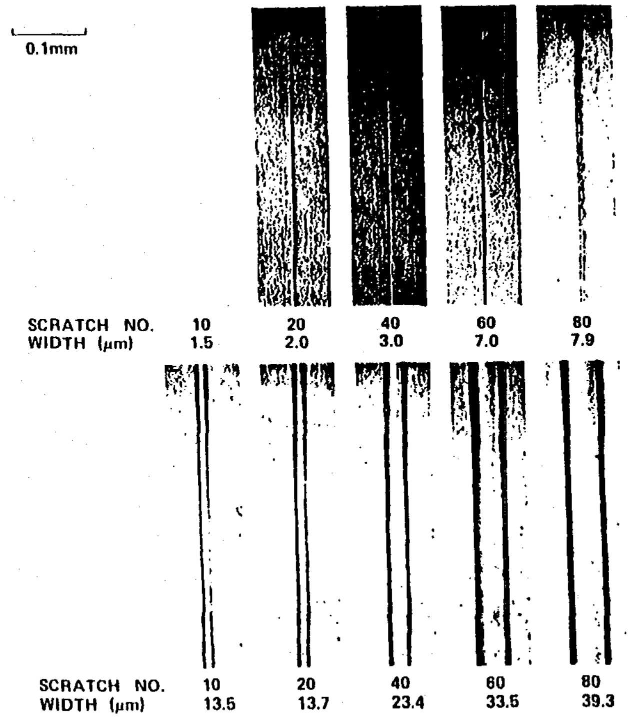

0.1mm
SCRATCH NO. 10 20 40 60 80
WIDTH (μm) 1.5 2.0 3.0 7.0 7.9
SCRATCH NO. 10 20 40 60 80
WIDTH (μm) 13.5 13.7 23.4 33.5 39.3

FIG. 9. Differential interference contrast micrographs of two sets of standard scratches. Although their widths differ by as much as an order of magnitude, the scattering levels from each pair are nearly equal. (Bennett, 1978.)

The reason scratches are so apparent visually is that scattering from them is localized in a small area of the surface where the scratch is. The total amount of light scattered per unit area of the surface by scratches is often negligible even though the scratches themselves are quite apparent, which gives rise to the opticians' admonition "optical components are made to be looked through, not at." For scratch dimensions large relative to the wavelength of light, scattering will be determined by geometrical optics and will be nearly independent of wavelength. The same is true of digs. Figure 10 shows the scattering observed for the standard and experimental No. 40 scratches as a function of wavelength. The wider experimental scratch shows almost no wavelength dependence. The narrower No. 40 standard scratch shows some decrease in scattering with increasing wavelength. This decrease results in part from the decrease in microirregularity scattering of the surface on which the scratch is made.

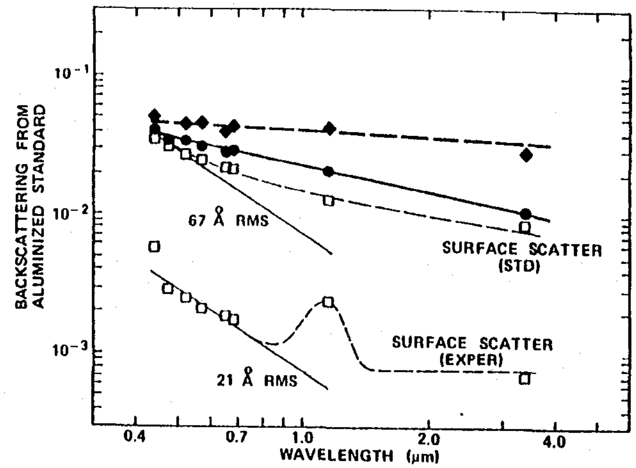


FIG. 10. Comparison of the scattering levels observed from: (◆) experimental 23-μm-wide No.40 scratch; (●) standard 3-μm wide No.40 scratch; (□) surface scattering adjacent to each scratch; the solid diagonal lines are predicted microirregularity scattering for roughnesses of 67 and 21 Å. The measuring beam diameter was 2 mm. (Bennett, 1978.)

Sand or rain erosion increases the number of pits on the surface and can significantly increase scattered light. Figure 11 shows the scattered light levels at 3.39 μm observed in the forward direction for two missile domes, one of which had been flown on an aircraft. Differences between the two were not apparent to a casual observer, but the scattering levels of the dome which had been flown were higher than those of the unused dome by over a factor of 2. If a rain field had been encountered, the differences would have been much more pronounced. Dust or sand erosion has a similar effect on exposed optics. These scattering measurements were made on the NWC Optical Evaluation Facility (Bennett and Stanford, 1976; Archibald and Bennett, 1978), a schematic of which is shown in Fig. 12. Either forward or backscattering can be measured at over 20 laser wavelengths extending from the ultraviolet to the infrared. Although the instrument is designed primarily to measure TIS, the angular dependence of scattering can also be investigated (Elson and Bennett, 1979). Reflectance and transmittance can be determined using this instrument, which is computer controlled so that measurements at several hundred points uniformly distributed over the sample surface can be taken automatically. The absorption of small samples which is sometimes related to surface effects can be determined in an adiabatic calorimeter, which has a sensitivity of about 3 × 10^{-5}/W of input power, and which is also part of this instrument.

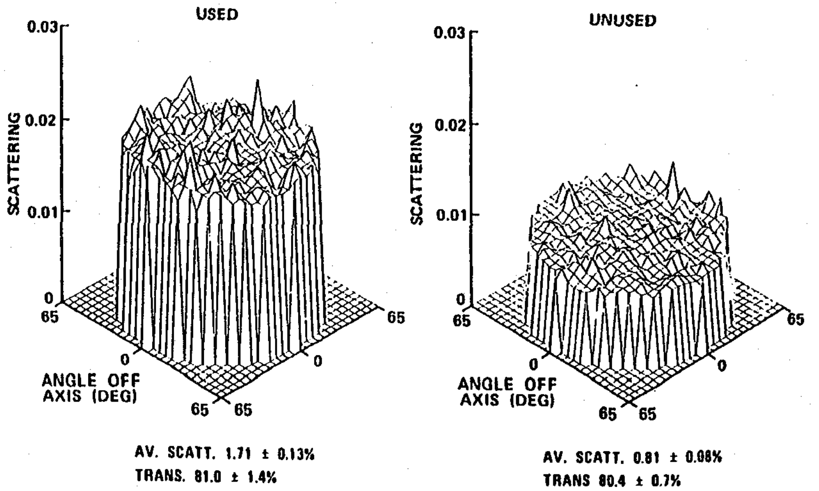


FIG. 11. Light scattered at a wavelength of 3.39 μm as a function of position for a missile dome which had been flown and one which had not. (Bennett, 1978.)

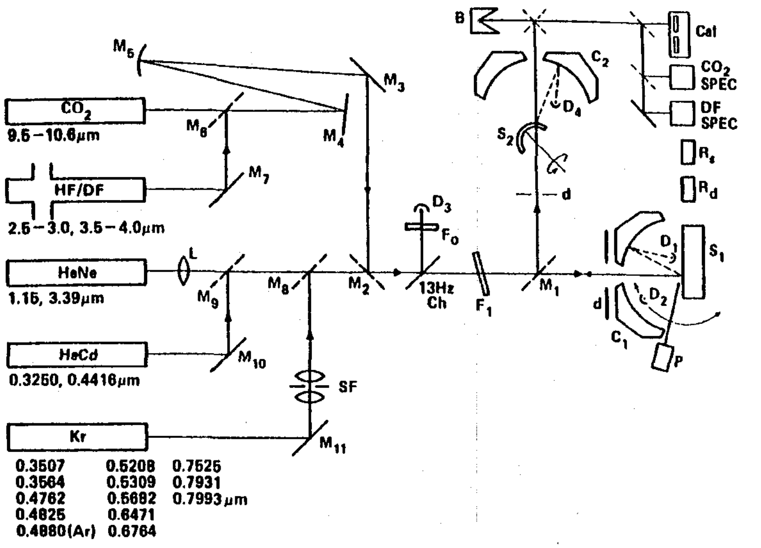


FIG. 12. *Optical evaluation facility*: mirrors M, lens L, spatial filter SF, chopper Ch, filters F, detectors D, calorimeter Cal, beam dump B, samples S, reference surfaces R, spectrometer monitors SPEC, Coblentz spheres C, and diaphragms d. Forward or back TIS, angular backscatter in plane of incidence, reflectance at normal incidence or 45°, transmission, and calorimetric measurements can be made with this instrument. (Bennett, 1978.)

A comparison of forward and backscattering from an etched potassium chloride sample as a function of wavelength is shown in Fig. 13. The forward scattering level drops about 15 times and the backscattering level nearly 40 times in going from the visible to the infrared for this sample, and the

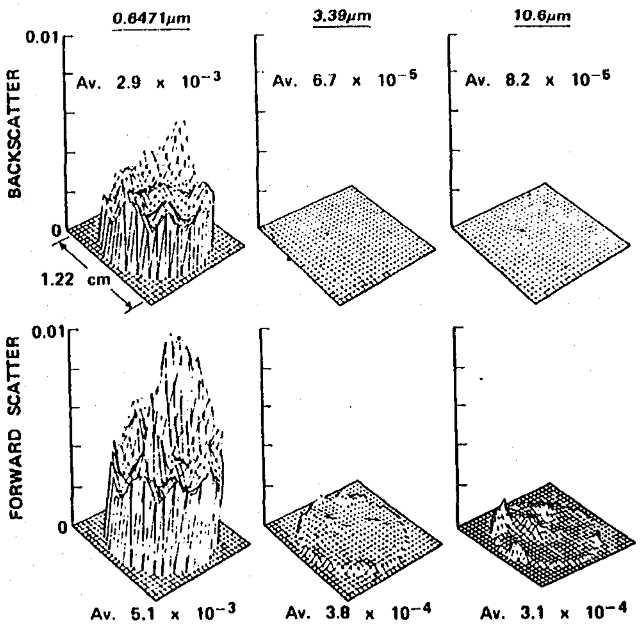


FIG. 13. Light scattered from etched potassium chloride as a function of wavelength and position on the surface. (Bennett, 1978.)

positions where maximum scattering occurs also shift somewhat between the visible and the infrared. The forward scattering level is always higher than the backscattering level, as would be expected for a dielectric material if the scattering were caused by particulates (van de Hulst, 1957, p. 145).

E. PARTICULATE SCATTERING FROM ISOLATED MICROIRREGULARITIES

When the surface irregularities become comparable in size to the wavelength of light scattered from the surface, the scattering process can no longer be described by geometrical optics. It is convenient in this case to separate the scattering into two categories: scattering from isolated particulates and the previously described scattering from correlated microirregularities. In particulate scattering each scattering center behaves independently of all others, so that the total scattered light arising from particulate scattering is the sum of the contributions from each particle. We can then use Mie scattering theory to describe the resultant scattered light. These scattering functions become very complicated, but if we consider only dipole scattering from transparent spherical particles and disregard the angular dependence of the scattered light, the TIS from isolated particulates is described approximately (Orr and Dallavalle, 1959) by the "universal scattering curve" shown in Fig. 14. It can also be applied to slightly absorbing particles. The essential features of this particulate scattering are seen to be:

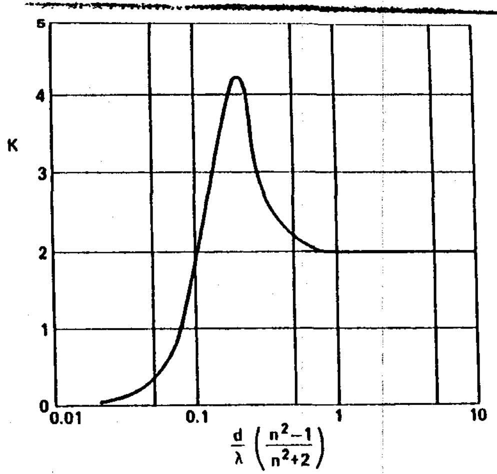


FIG. 14. "Universal" scattering curve for nonabsorbing spherical particles of refractive index n and diameter d at wavelength λ. The scattering coefficient is K. (Bennett, 1978.)

(a) When the particle size is large compared to the wavelength λ, the scattering cross section is independent of wavelength and is twice the geometrical cross section of the particle (the extinction paradox) (van de Hulst, 1957, p. 107);

(b) As the wavelength approaches the particle diameter, the scattering cross section k increases abruptly and a resonance occurs near the point at which the two are equal; and

(c) As the wavelength continues to increase the scattering cross section falls rapidly and for diameters much smaller than a wavelength we have Rayleigh scattering, which decreases as the inverse fourth power of the wavelength.

Another characteristic of particulate scattering is that forward scattering is always larger than backscattering for transparent particles (van de Hulst, 1957, p. 145). The two approach each other in the Rayleigh limit. If the particles were metallic, the reverse would be true and forward scattering would be less than backscattering.

The increased scattering predicted by Fig. 14 in the neighborhood of $d/\lambda = 1$ is probably the explanation for the increased scattering observed near a wavelength of 1 μm in Fig. 7. An increase of this kind is frequently seen for smooth surfaces and, as mentioned earlier, is thought to be caused by dust which has an average diameter of about 1 μm.

IV. ANGULAR DEPENDENCE OF SCATTERING

Although TIS measurements are relatively simple to make and are generally satisfactory for assessing optical surfaces, some applications require a knowledge of scattering into a particular angle or range of angles. For example, in a laser gyro system where light is incident on mirrors at 30 or 45°, the retroscattered light, i.e., light that is backscattered along the incident beam direction, is of crucial importance because it causes the gyro to malfunction. In trying to distinguish weak objects near a very bright object, such as in a reflecting solar coronograph, scattered light within a few degrees or even a fraction of a degree of the specular direction can completely mask the desired image. In high power laser systems where the light from the laser goes through a series of amplifiers, any light backscattered into the laser can depump it and cause catastrophic damage. Finally, in many types of optical systems it is not possible to baffle the system adequately to prevent all angles of scattered light from reaching the detector and thus degrading the image quality.

Scattering, whether from the viewpoint of angular scattering or TIS, can be caused by (1) scratches, digs, and other surface features whose dimensions are large compared to the wavelength of light, (2) isolated particulates on a surface, or (3) microroughness whose heights are much smaller than the wavelength. Angular scattering from the first type of surface features can, in principle, be calculated using the laws of geometrical optics. In practice, however, this is quite difficult because of the lack of a good model for the shapes and sizes of the facets making up the surface features. Scratches, for example, are often cut on glass surfaces with a diamond scribe and their edges consist of a series of conchoidal fractures. The steeply sloped facets on many scratches produce a considerable amount of large angle scattering, causing the scratches to appear distinctly against the smaller background scattering from the microirregularities. Ground or matte surfaces, whose facets are also large compared to the wavelength, produce a considerable amount of large angle scattering. Ditchburn (1964) has shown that if a ground glass surface has a Gaussian autocorrelation function, its scattering is independent of angle, i.e., it is a perfect diffuser. Such a surface is called a Lambertian surface since it follows Lambert's law (Ditchburn, 1964).

Angular dependence of scattering from dust or other isolated particulates on a surface can, in principle, be handled by Mie theory (Mie, 1908; see also Stratton, 1941; Kerker, 1969). In practice, this is also difficult because one

must know (1) the distribution of sizes and shapes of the particulates, (2) their (complex) refractive indices, and (3) the appropriate interaction terms for the effect of the surface on which the particles are sitting. In a simplified experiment where the angular scattering was measured from a low scatter mirror contaminated with spherical silver particles whose diameters were in the 9–18 μm range, Young (1976) found good agreement between the predictions of Mie scattering theory and the measured angular scattering at wavelengths of 0.6328 and 10.6 μm. Most of the observed scattering could be explained by assuming that the radiation scattered in the forward direction was reflected from the mirror surface unaffected by the presence of the particle. The scattering calculations assumed that the particles were in free space and were illuminated by polarized radiation. While the results of this study are very encouraging, they do not duplicate actual laboratory conditions where the dust particles are more likely in the size range around 1 μm, probably not spherical, and have optical constants more like slightly absorbing dielectrics, such as impure SiO_2. A general rule of thumb for angular scattering from particulates is that they can produce much more large angle scattering than the surface on which they are resting. For this reason one can see glints from very small dust particles on a surface when that surface is illuminated at oblique incidence by a microscope illuminator or laser beam.

Angular scattering that is most tractable to handle theoretically is the third type mentioned above, i.e., scattering from correlated surface microirregularities whose heights are much smaller than a wavelength of light. Even in this case, however, there is no simple relation giving the angular distribution of scattered light as a function of surface roughness, similar to the TIS relation in Eq. (1). The reason is that the angular distribution of the scattering depends not only on the rms height of the surface irregularities but also on their lateral separations. A convenient measure of the separation of surface features is the autocorrelation or autocovariance function which will be defined rigorously later in this section. For the moment we will assume an autocovariance length a (which is the standard deviation of the Gaussian) at which value the autocovariance has dropped to $1/e$ of the maximum value.

The scalar scattering theory can give information about the angular scattering at angles close to the specular direction. From this theory it can be shown that the scattering into an angle α measured from the specular direction is (Porteus, 1963) $16\pi^4\delta^2a^2\alpha^2/\lambda^4$, where δ is the rms roughness and λ the wavelength. This expression gives the scattering (ratio of scattered light to total reflected light from the surface) into a cone of half angle α. (The scattering expressions $dP/d\Omega$ given later in this section are for scattering into a unit solid angle $d\Omega$ when the incident intensity is unity and include the total reflectance of the surface.) The simple expression above holds only for small scattering angles since it is the first term of an expansion (Porteus, 1963).

The small angle approximation makes the expression independent of the functional form assumed for the autocovariance function, so it should work for real surfaces which do not have Gaussian autocovariance functions.

In order to understand angular scattering from an isotropic rough surface having random roughness, we can consider that the surface is made up of a superposition of sinusoidal gratings having different amplitudes, phases, and periods (Church *et al.*, 1977; Stover, 1975). In this approach one loses the phase information about how the various sinusoidal gratings are superimposed, but if the rough surface can be adequately characterized in this manner, one can obtain a correct expression for the angular scattering. Another approach is to actually measure the autocovariance function for the rough surface and use the vector scattering theory given later in this section to determine the angular scattering. Although in principle it is possible to calculate the two-dimensional angular scattering from the surface autocovariance function and to obtain the autocovariance function from the measured angular scattering, one cannot unambiguously obtain a representation of the actual surface profile because of the missing phase information mentioned above.

In this section we will first discuss the angular scattering from a sinusoidal grating (one component of surface roughness). Then we will discuss the complete vector theory which can be used to calculate the angular scattering, given the surface autocovariance function. This vector includes polarization effects and can be used for all angles of incidence. In the experimental part of this section we will describe three methods for obtaining surface statistics and will give the advantages and disadvantages of each. We will then show some autocovariance functions measured for actual optical surfaces, including a diamond-turned surface. Finally, we will compare angular scattering calculated from a measured autocovariance function with the actual angular scattering measured for the same surface.

A. Angular Scattering from a Sinusoidal Grating

Since a surface having random roughness can be considered to be composed of a two-dimensional superposition of sinusoidal gratings having different amplitudes, periods, and phases, an insight into angular scattering can be obtained by considering the angular scattering from a single sinusoidal grating having an amplitude A, measured from the mean surface level, and grating spacing d. For normally incident light the angle of diffraction (scattering) is given by the grating equation:

$$\sin \theta = m(\lambda/d) \tag{2}$$

where m is the order of interference and λ the wavelength. The intensity I of light diffracted into the mth order is proportional to (Church and Zavada, 1975)

$$I \sim (A/\lambda)^{2|m|} \tag{3}$$

For microrough surfaces, $A/\lambda \ll 1$, so that only the first diffracted order need be considered. It is clear from Eq. (2) that the *angle* into which light is diffracted or scattered depends *only* on the lateral spacing between the grooves and the scattered *intensity* depends primarily on the groove depth, which can be related to the rms surface roughness δ. A plot of the geometry for Eqs. (2) and (3) and the relation between d and θ is given in Fig. 15. Irregularities which have separations which are large compared to a wavelength will scatter light near the specular direction; those with smaller separations will scatter at larger angles and those whose separations approach the wavelength will scatter at angles approaching 90 . Figure 16 is another way of showing the relation between the wavelength, scattering angle, and separation of surface irregularities for normally incident light on a rough surface. The dashed lines are an example of how to use the nomogram. For example, for a wavelength of 0.5 μm, scattering at an angle of 5 from the specular direction is produced by surface features having a lateral separation of about

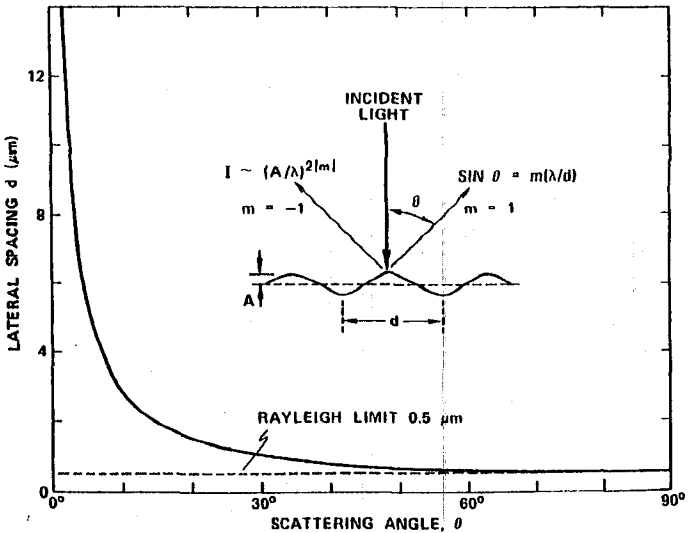


FIG. 15. Scattering angle from a coherent scatterer at normal incidence as a function of grating groove separation for light having a wavelength of 5000 Å. (Bennett, 1978.)

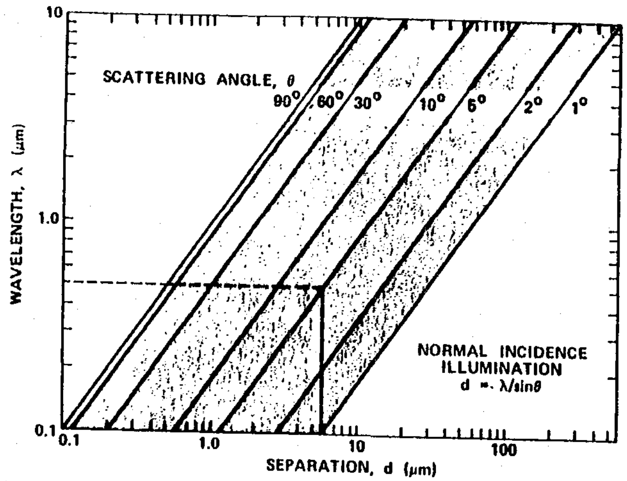


FIG. 16. Nomogram giving separation of surface features that produce scattering at a particular angle and wavelength. The dashed lines are an illustrative example. (Elson and Bennett, 1979.)

5.7 μm. Again, note that scattering into the maximum angle, 90°, is accomplished by microirregularities having a minimum separation, equal to the wavelength, and that irregularities having smaller separations cannot produce visible scattering when light is normally incident on a surface. Therefore, evaporated or sputtered films that consist of particles having diameters in the 400- 1000 Å range should not produce any scattering in the visible spectral region. Films tend to contour the underlying substrate so that the scattering is produced by features on the substrate having larger lateral separations. This situation for *correlated* microirregularities is in marked contrast to that for particulate scattering, where particles whose diameters are much smaller than a wavelength can produce scattered light.

If surfaces are anisotropic, such as diamond-turned surfaces, the light scattering depends on the polarization of the incident beam and the orientation of the surface features. Returning to the one-dimensional case of angular scattering from a sinusoidal grating, the complete expression for the diffracted (or scattered) intensity at an angle θ assuming a normally incident beam polarized parallel to the plane of diffraction (perpendicular to the grating grooves) is (Church and Zavada, 1975)

$$I_p/R_0 = (2\pi A/\lambda)^2 \sec\theta \qquad (4)$$

and for a beam polarized perpendicular to the plane of diffraction (parallel to the grating grooves) is

$$I_s/R_0 = (2\pi A/\lambda)^2 \cos\theta \qquad (5)$$

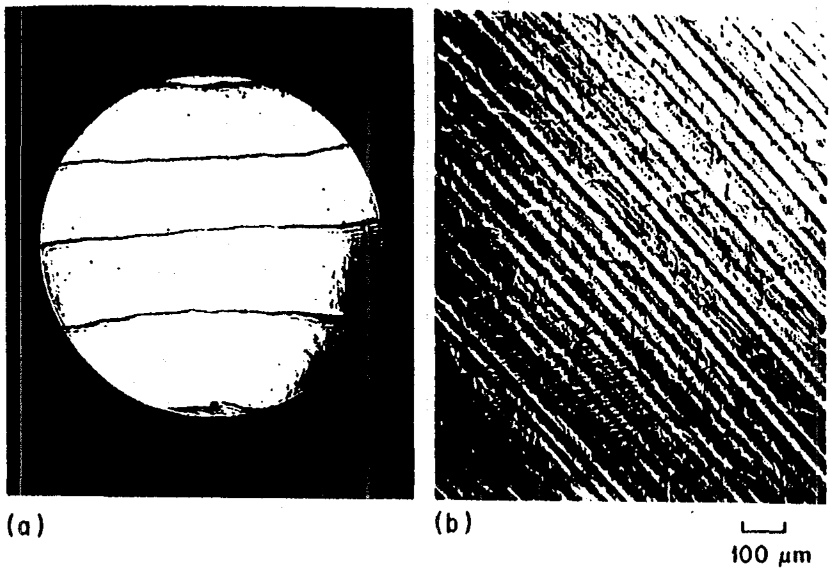


FIG. 17. (a) Optical figure (1/20th wave) and (b) surface finish (21.0 Å rms) of a 4-in. diameter, diamond-turned copper mirror produced by the Y-12 Plant at Oak Ridge, Tenn. (Bennett, 1978.)

Note that Eqs. (4) and (5) give the *intensity* of light diffracted into the first order at *angles* specified by the grating equation, Eq. (2), so that the θ's apply only to *specific angles*. When the groove spacing d approaches the wavelength λ (θ approaches 90°), the intensities of the p and s diffracted components are very different because of the sec θ and cos θ terms, and depend strongly on d as well as A. However, if the incident light is unpolarized, this effect is negligible except for grating spacings which are almost identical to the wavelength. The diffraction effects are particularly important for diamond-turned surfaces which exhibit shallow parallel grooves cut by the diamond point. A Nomarski micrograph of the surface of one such copper mirror is shown on the right side of Fig. 17. This mirror, which was produced by the Oak Ridge Y-12 Plant, has an excellent figure in spite of the grooves, as is shown on the left side of the figure. The light scattered from this mirror is comparable to that from a good, conventionally polished, glass optical flat.

B. Angular Scattering from Nonsinusoidal Surfaces

In the preceding section scattering from a surface having a simple sinusoidal profile was discussed. Although very few optical surfaces are of this type, surfaces having nonsinusoidal profiles may be reconstructed as a Fourier superposition of sinusoidal profiles. Hence, the concept of diffraction from a single sinusoidal profile may be used to give a qualitative picture of scattering from a more complex surface.

Surfaces having nonsinusoidal profiles may be divided into two types: those having periodic and random surface irregularities. Any general surface profile $z(\boldsymbol{\rho})$ can be considered as a Fourier superposition of sinusoidal profiles as follows:

$$z(\boldsymbol{\rho}) = \int d^2k\, Z(\mathbf{k}) e^{i\mathbf{k}\cdot\boldsymbol{\rho}} \tag{6}$$

The quantity $\boldsymbol{\rho} = (x, y)$ is a position vector of a point in the (x, y) plane from which the height $z(\boldsymbol{\rho})$ is measured. Since the surface is Fourier-analyzed in two dimensions, $\mathbf{k} = (k_x, k_y)$ is the wavevector of the Fourier component whose amplitude is $Z(\mathbf{k})$. Equation (6) may represent random or periodic roughness. If $z(\boldsymbol{\rho})$ refers to random (periodic) roughness, then $Z(\mathbf{k})$ is generally a continuous (discrete) function of $\mathbf{k}$. Light incident on a surface having a nonsinusoidal profile will scatter according to the sum of the scattering from the sinusoidal components making up its profile. Thus, scattering from periodic roughness components will be in discrete directions, while scattering from random roughness will be continuously distributed in angle because the distribution of sinusoidal surface components is continuous. To see how the sinusoidal distribution enters into the scattering formulas, the equations obtained from first-order perturbation theory will now be considered.

C. Angular Scattering Theory for Polished Surfaces

Polished surfaces whose rms roughnesses δ are much less than the wavelength λ and gratings whose amplitudes are small compared to λ are ideally suited for scattering calculations using first-order perturbation theory. A number of authors (Silver, 1947; Leader, 1971b; Fung and Chan, 1969; Rayleigh, 1945) have treated the problem of scattering from rough surfaces with no overcoats in the case where $\delta \ll \lambda$ and vector properties of the scattered light are included. The differential fraction dP of the incident energy scattered per unit solid angle $d\Omega = \sin\theta\, d\theta\, d\phi$ may be written as (Elson, 1975)

$$\frac{dP}{d\Omega} = \frac{(\omega/c)^4}{\pi^2} \cos\theta_0 \cos^2\theta |1 - \varepsilon|^2 \frac{|Z(\mathbf{k} - \mathbf{k}_0)|^2}{L^2} \left[\frac{|\chi_\theta|^2}{|\nu - iq\varepsilon|^2} + \frac{|\chi_\phi|^2}{|\nu - iq|^2} \right] \tag{7}$$

where

$$\chi_\theta = \frac{(\nu\nu_0 \cos\phi + kk_0\varepsilon)\cos\phi'}{\nu_0 - iq_0\varepsilon} - \frac{i(\omega/c)\nu \sin\phi \sin\phi'}{\nu_0 - iq_0} \tag{8a}$$

and

$$\chi_\phi = \frac{\omega}{c}\left[\frac{(\omega/c)\cos\phi \sin\phi'}{\nu_0 - iq_0} - \frac{i\nu_0 \sin\phi \cos\phi'}{\nu_0 - iq_0\varepsilon}\right] \tag{8b}$$

The light is scattered into a direction specified by (θ, ϕ). This expression can be applied to scattering from randomly rough surfaces or diffraction from low efficiency gratings. The component of the wave vector parallel (perpendicular) to the mean surface is $k_0 = (\omega/c)\sin\theta_0$ $[q_0 = (\omega/c)\cos\theta_0]$, and $k = (\omega/c)\sin\theta$ $[q = (\omega/c)\cos\theta]$ for the incident and scattered light, respectively. Also, $\nu_0 = [k_0{}^2 - \varepsilon(\omega/c)^2]^{1/2}$ and $\nu = [k^2 - \varepsilon(\omega/c)^2]^{1/2}$, where ε is the complex dielectric constant of the scattering medium for angular frequency ω. The quantity $Z(\mathbf{k} - \mathbf{k}_0)$ is the Fourier transform of the surface profile. The electric field polarization angle ϕ', measured relative to the plane of incidence, may be set equal to 0 (or $\pi/2$) to specialize to the case of p-polarized (s-polarized) incident light. Averaging over ϕ' from 0 to 2π yields the result for unpolarized incident light. The first and second terms in the bracket of Eq. (7) represent scattered light which is p- and s-polarized, respectively (measured relative to the plane of scattering). Since $\omega/c = 2\pi/\lambda$, it is seen that Eq. (7) varies as λ^{-4}. The distribution of Fourier components enters as the magnitude squared of $Z(\mathbf{k} - \mathbf{k}_0)$, as given in Eq. (6). Recall that $Z(\mathbf{k} - \mathbf{k}_0)$ may represent either periodic or random surface irregularities. In the case of periodic profiles, the surface shape is known *a priori*, and consequently $Z(\mathbf{k} - \mathbf{k}_0)$ may be readily calculated. In the case of random roughness, $z(\boldsymbol{\rho})$ is not known and statistical properties for the surface based on ensemble averages must be used.

There are several observations that can be made from Eq. (7). These include the relation of Eq. (7) to the expression for TIS, Eq. (1), the bidirectional reflectance distribution function (BRDF), diffuse scattering from surfaces having random roughness, and diffraction from low efficiency gratings. First we will show how Eq. (7) can be used to obtain the expression for TIS as given in Eq. (1). The following assumptions are made: (1) the rough surface is assumed to be perfectly reflecting, i.e., the limit of $|\varepsilon|$ approaches ∞; (2) the angle of incidence θ_0 is 0; (3) the scattering angle θ is limited to angles near the specular direction; and (4) the autocovariance function is assumed to have zero slope at the origin. A convenient form is a Gaussian, which yields, for normal incidence $(\mathbf{k}_0 = 0)$,

$$\langle |Z(\mathbf{k})|^2 \rangle / L^2 = \pi a^2 \delta^2 \exp(-k^2 a^2/4) \tag{9}$$

where a is the autocovariance length. Equation (9) is obtained by assuming that an ensemble average, denoted by $\langle\ \ \rangle$, yields a Gaussian autocovariance function of the form $G(\tau) = \delta^2 \exp(-\tau^2/a^2)$. The Fourier transform of $G(\tau)$ yields Eq. (9). We have assumed that the autocovariance length a is much greater than the wavelength, so the surface is gently undulating and has a region about the origin comparable to λ, where the autocovariance function would be quite flat, as is a Gaussian. Physically, this means that the surface height does not vary much over a distance of a wavelength or less. This assumption is also consistent with the Kirchhoff boundary conditions where the surface is assumed to be locally flat to make the Fresnel reflection coefficients applicable.

To calculate the TIS from Eqs. (7) and (9), an integration over the total scattering hemisphere must be performed. Letting $\theta_0 = 0, \theta \ll 1$, and $|\varepsilon| \to \infty$, Eq. (7) becomes

$$\frac{dP}{d\Omega} = \frac{(\omega/c)^4}{\pi^2} a^2\delta^2 \exp\left(\frac{-k^2a^2}{4}\right) \tag{10}$$

which may be integrated to yield

$$P = (4\pi\delta/\lambda)^2 \tag{11}$$

Equation (11) is the same as Eq. (1) for a perfectly conducting surface ($R_0 = 1$, $|\varepsilon| \to \infty$). Note that because of the condition $a/\lambda \gg 1$ the range of integration on θ has been extended to include the range $(0, \infty)$ because the exponential term becomes negligible several degrees away from the specular direction.

The BRDF is sometimes used in connection with scattering measurements or calculations. It is simply related to Eq. (7), being obtained by dividing Eq. (7) by $\cos\theta$, i.e., BRDF $= (dP/d\Omega)/\cos\theta$. The BRDF is completely symmetrical with respect to interchanging θ_0 and θ. This means that a surface which scatters light incident at θ_0 into an angle θ also scatters reciprocally, so that light incident at an angle θ is scattered into θ_0. For example, s-polarized light incident at θ_0 is scattered into p-polarized light at θ, on the one hand, and p-polarized light incident at θ is scattered into s-polarized light at θ_0. The BRDF is defined by Nicodemus (1970) and is proposed to be a function from which various scattering relationships can be obtained.

Diffuse scattering occurs because the polished surface is not perfectly smooth. The residual microroughness is random and the profile shape is not known in detail *a priori*. To handle random roughness, the surface profile must be measured directly or treated statistically. One statistical method of treating random processes, in principle, is to employ ensemble averages. The angular scattering from random roughness is ensemble averaged and applying this to Eq. (7) leads to the abbreviation

$$\langle|Z(\mathbf{k} - \mathbf{k}_0)|^2\rangle/L^2 = \delta^2 g(\mathbf{k} - \mathbf{k}_0) \tag{12}$$

where $\langle\ \rangle$ denote an ensemble average and $g(\mathbf{k} - \mathbf{k}_0)$, the spectral density function, is a continuous function of $\mathbf{k}$. Equation (12) is the Fourier transform

of the autocovariance function, δ^2 is the mean value of the square of the roughness, and L^2 is the area of illumination. If the spectral density function is not evaluated by experimental measurement, one can assume a form for the autocovariance function (such as a Gaussian or exponential) and take the Fourier transform, or assume a form for the spectral density function itself. Experimentally, the spectral density function may be determined from a measurement of the angular distribution of scattered light using Eq. (7), or by measuring surface height data directly (see Section D, following).

Diffraction from low efficiency gratings is much easier to handle analytically because one can calculate $Z(\mathbf{k} - \mathbf{k}_0)$ directly from the periodic profile $z(\boldsymbol{\rho})$, and an ensemble average is not required. The height profile may be expanded in a Fourier series. Rather than keeping the plane of incidence perpendicular to the direction of the grating grooves as is usually done, more generality is obtained if the angle between the plane of incidence and groove direction is allowed to vary, as shown in Fig. 18. When the area of illumination on the surface is much larger than the incident wavelength,

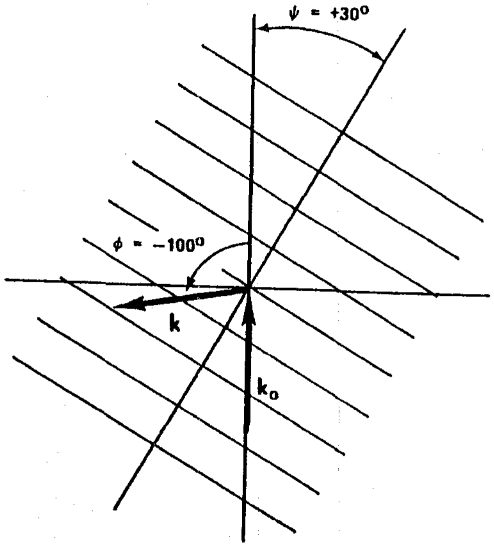


FIG. 18. Schematic representation of diffraction from a grating when there is an angle Ψ between the direction perpendicular to the grooves (shown as parallel lines) and the incident wave-vector component $\mathbf{k}_0$ parallel to the surface. The plane of the paper is the plane of the scattering surface. The wave-vector component $\mathbf{k}$, parallel to the surface, of the -1 diffracted order is at an angle Φ relative to the plane of incidence. The angles Φ and Ψ are positive in the clockwise direction. In this example the grating spacing is 1.1λ, θ_0 (angle of incidence) is 45° and θ (angle of diffraction) is 27.5°. (Elson, 1977.)

the distribution of surface wave vector components is given by

$$\frac{|Z(\mathbf{k}-\mathbf{k}_0)|^2}{L^2} = \pi^2 \sum_{n=-\infty}^{\infty}{}' |C_n|^2 \delta(k_x - k_0 - k_n \cos\psi)\delta(k_y - k_n \sin\psi) \quad (13)$$

where $\delta(x)$ are Dirac δ functions, $k_0 = (\omega/c)\sin\theta_0$ (the plane of incidence is the x-z plane), $k_n = 2\pi n/d$, and d is the grating spacing. The coefficient of expansion in the Fourier series of the periodic profile is C_n and L^2 is the area of illumination. The angle ψ is the angle between the plane of incidence and the direction perpendicular to the grating grooves, as shown in Fig. 18. In Eq. (13) the wave vector component parallel to the surface of the diffracted light is $k = (k_x, k_y)$, where $k_x = (\omega/c)\sin\theta\cos\phi$ and $k_y = (\omega/c)\sin\theta\sin\phi$. The angle θ is the polar diffraction angle (measured relative to the mean surface normal) and ϕ is the azimuthal diffraction angle (measured from the incident plane, the clockwise direction being positive). It is seen that

$$\cos\phi = (k_0 + k_n \cos\psi)/k \quad (14a)$$

$$\sin\phi = k_n \sin\psi/k \quad (14b)$$

where $k = [(k_0 + k_n \cos\psi)^2 + k_n{}^2 \sin^2\psi]^{1/2}$ and $\theta = \sin^{-1}(kc/\omega)$. Using Eq. (13), Eq. (7) may be integrated over the scattering hemisphere to yield the result for the fractional amount of incident energy diffracted into the nth grating order:

$$P^{(n)} = \frac{16\pi^2}{\lambda^2} |C_n|^2 \cos\theta\cos\theta_0 |1-\varepsilon|^2 \left[\frac{|\chi_0|^2}{|\nu - iq\varepsilon|^2} + \frac{|\chi_0|^2}{|\nu - iq|^2}\right] \quad (15)$$

This equation is applicable to low efficiency gratings where the grating amplitude is much less than the incident wavelength. The direction of the diffracted beam (θ, ϕ) is given by Eqs. (14), and the first and second terms in the bracket of Eq. (15) denote diffracted light polarized parallel (p) and perpendicular (s) to the plane of diffraction.

D. Determination of the Spectral Density Function

As discussed in the next section, interferometric or profilometer techniques may be used to measure surface height data directly. These height data yield autocovariance functions $G(\tau)$ and spectral density functions $g(\mathbf{k} - \mathbf{k}_0)$. To see how measured height data are used, we first assume that a continuous one-dimensional record of surface height data $z(\rho)$ is available and that the mean surface level is zero, i.e., $\langle z(\rho)\rangle = 0$, where $\langle\ \ \rangle$ denote an ensemble average (Lee, 1960). The autocovariance function $G(\tau)$ may be defined as (Lee, 1960)

$$G(\tau) = \langle z(\rho)z(\rho + \tau)\rangle \quad (16)$$

or, under the assumption that the data are stationary and ergodic (Lee, 1960), equivalently as

$$G(\tau) = \lim_{L\to\infty} \frac{1}{L}\int_{-L/2}^{L/2} z(\rho)z(\rho+\tau)\,d\rho \quad (17)$$

From Eq. (7) it is seen that the basic quantity relating to the surface factor is $Z(k)$, which is the Fourier transform of the surface profile, defined as

$$Z(k) = \int d\rho\, z(\rho)e^{-ik\rho} \quad (18)$$

Only one dimension is considered in Eq. (18). Forming the expression $\langle|Z(k)|^2\rangle$ yields, after some algebraic manipulation, the surface factor $g(k)$ as

$$\delta^2 g(k) = \frac{\langle|Z(k)|^2\rangle}{L} = \int G(\tau)e^{ik\tau}\,d\tau \quad (19)$$

where L results from integration over ρ. We now need to relate the continuous height data, assumed above, to experimental height data which are discrete and are measured over a finite length interval. Discrete data imply that there are discrete units of lag length τ such that $\tau = m\tau_0 \le M\tau_0$. Here m is an integer ranging from 0 to M, and τ_0 is the digitization interval. Also, the height data are given as $z(n)$ [corresponding to $z(\rho)$], where $\rho = n\tau_0$. The total number of data points is $N > M$, where M indicates the maximum lag length chosen. The discrete counterpart to Eq. (17) is (Bendat and Piersol, 1971)

$$G_M(m) = \frac{1}{N-m}\sum_{n=1}^{N-m} z(n)z(n+m) \quad (20)$$

and the discrete counterpart to Eq. (19) becomes

$$g(k) = \tau_0\left(G_M(0) + 2\sum_{m=1}^{M-1} D(m)G_M(m)\cos km\tau_0\right) \quad (21)$$

where $D(m)$ is a lag window such that $D(0) = 1$ and $D(M) = 0$. Also, the wave vector k, because of the digitization of the data, has a maximum allowable magnitude $k_{\max} = \pi/\tau_0$. It is convenient to choose $k = Kk_{\max}/M$, where $K = 0, \pm 1, \pm 2, \ldots, \pm M$. The product $D(m)G_M(m)$ is often called a modified autocovariance function. The advantage of modifying the autocovariance function by introducing the window function $D(m)$ is so that the autocovariance function can be defined over the infinite range of its argument in order to permit its Fourier transform to be taken. Complete details of analyzing discrete data records of finite length are given elsewhere (Elson and Bennett, 1979). Examples of measured autocovariance functions are given in the next section.

E. Experimental Measurements of Angular Scattering and Surface Statistics

Angular scattering from surfaces whose roughness is small compared to the wavelength is frequently measured by illuminating the surface at normal incidence or near normal incidence and moving a detector in a plane at a constant distance from the surface. If the source is polarized, s- or p-polarized scattered light can be measured, depending on whether the source is polarized perpendicular or parallel to the plane of incidence. However, to avoid surface plasmon coupling at the surface, s-polarized light is more straightforward to measure and interpret. Elson and Bennett (1979), Young (1976), and Stover (1975) have described instruments of the type mentioned above to measure angular scattering. Church *et al.* (1977) have an arrangement in which the angle between the source and detector is fixed at 90° and the sample rotates, causing the angles of incidence and scattering to vary between 0 and 90°. When the angle of incidence is 0°, the scattering angle is 90°, and vice versa. Other scattering experiments are referenced in the paper by Elson and Bennett (1979). In all scattering experiments it is important to have a well-defined incident beam with a small divergence angle, either focused on the sample or collimated. If the beam profile is Gaussian, it may have to be measured with the sample removed, and then subtracted from the intensity profile of the scattered light. The collection angle should also be well defined and the detector response needs to be linear over several decades. When measuring very low intensities of scattered light, it is imperative to track down and eliminate spurious scattered light from other sources. Since dust in the air is a very effective scatterer, it is desirable to make scattering measurements in a clean room or similar environment.

Statistical properties of optical surfaces may be measured using stereo electron microscopy (Dancy and Bennett, 1976), multiple-beam interferometry (Bennett, 1976), or a surface profiling instrument (Elson and Bennett, 1979). Stereo electron microscopy involves making a replica of the surface for observation in a transmission electron microscope, taking stereo pairs of electron micrographs, and then measuring heights of surface features using a stereo viewer and parallax bar. Figure 19 shows examples of one of a pair of electron micrographs used for stereo measurements. The surfaces have been shadowed obliquely with a platinum–carbon mixture to make surface detail such as the fine scratch visible. Since the surfaces have an overall texture, it is not possible to determine heights of surface features by measuring shadow lengths. Further, the heights of the features are too small to be observable in the scanning electron microscope. Heights of features are measured relative to a reference level, which must also be chosen on the micrograph, and under favorable conditions can be determined with an uncertainty of ±20 Å. In order to obtain surface slope information from stereo electron micrographs, heights of surface features need to be measured, preferably at equally spaced intervals, height differences calculated, and divided by the separations between points. This type of data is tedious to obtain since all measurements must be made visually. Fortunately, surface features which scatter light in the visible and infrared spectral regions have lateral separations which can be detected with a profilometer or interferometer. Thus, stereo electron microscopy is only needed for determining surface statistics for vacuum ultraviolet or x-ray scattering.

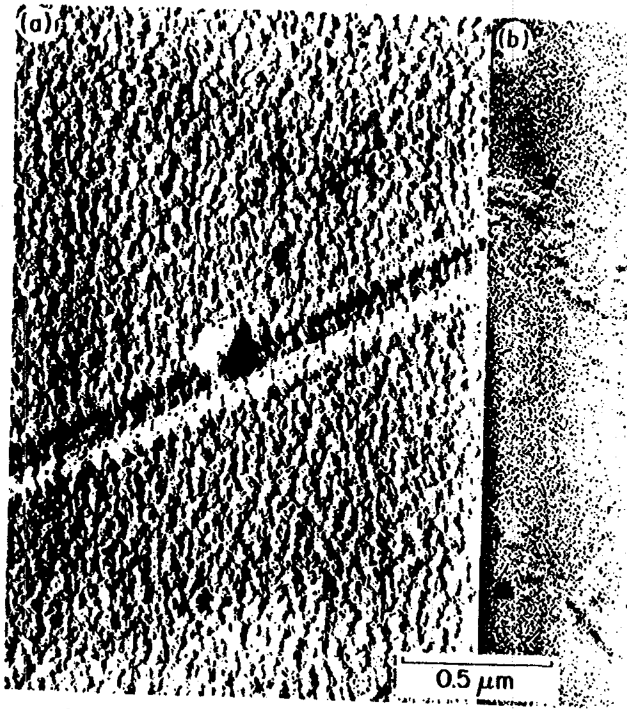


FIG. 19. One print of a stereo pair showing single-stage replicas of (a) silver rapidly sputtered onto a room temperature fused quartz substrate and (b) a bare, bowl feed polished fused quartz substrate. Original total magnification is 60,000X. The scratch in (a) is approximately 0.18-*μm* wide and 215-Å deep, and could not be seen in an optical microscope.

Multiple-beam interferometry is a method for obtaining statistics of surface features which have lateral separations of the order of several micrometers. Although the nominal resolution of one such system, a scanning interferometer employing multiple-beam fringes of equal chromatic order (FECO) (Bennett, 1976), was about 2 μm, the measured lateral resolution was somewhat larger (Elson and Bennett, 1979). FECO interferometers have the advantage that all portions of the surface can be observed interferometrically, so a representative place can be selected for the statistical measurements. The height sensitivity obtainable with the FECO scanning interferometer is about 3 Å rms. Figure 20 shows an autocovariance function of a polished fused quartz surface obtained with the scanning interferometer. The curve is an average of 51 scans taken at different places on the surface. It can be compared with Fig. 21 taken on the same surface with a profilometer, as described in a following paragraph. A scanning Fizeau interferometer has also been used to measure surface statistics (Eastman and Baumeister, 1974). Its lateral resolution, which was determined by the effective detector size was 7 μm and the height sensitivity was ~20 Å rms. Fizeau interferometers can give information about the surface roughness along the length of the interference fringe in the field of view. In order to inspect other portions of the surface, the wedge angle of the interferometer must be adjusted, which is a tedious operation.

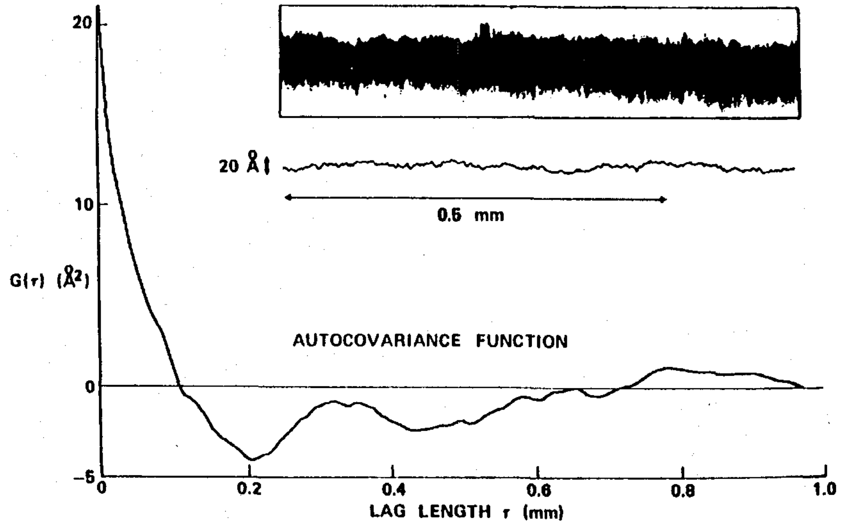


FIG. 20. Autocovariance function obtained from interferometric data for an 18.6 Å rms roughness polished fused quartz surface. Curve is an average of 51 scans. A photograph of a FECO interference fringe and TV scanning camera trace are also shown. (Bennett, 1976.)

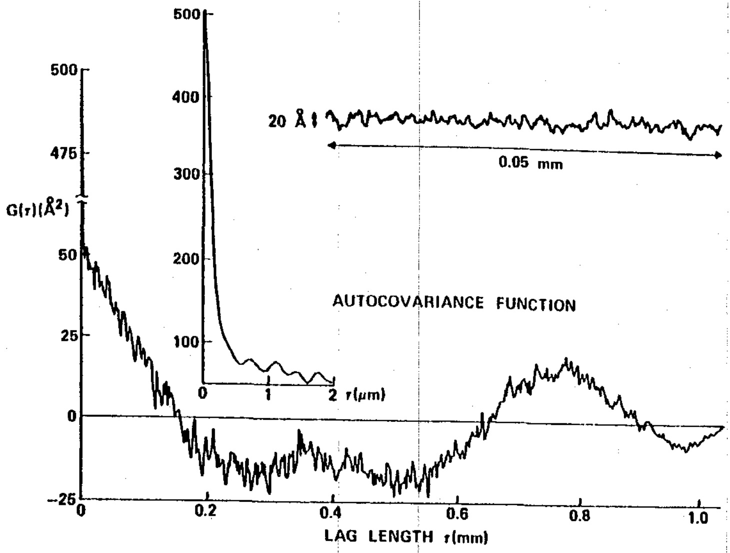


FIG. 21. Autocovariance function and surface scan obtained from profilometer data on the same surface as in Fig. 20. Data are for a single surface scan. (Elson and Bennett, 1979.)

A surface profiling instrument is probably the best method for obtaining surface statistics to be used in the vector scattering theory for calculating angular scattering because of its excellent height sensitivity and good lateral resolution. One such instrument (Elson and Bennett, 1979), whose output is digitized and fed into a minicomputer for statistical calculations, has a height sensitivity of about 1 Å rms and a lateral resolution of about 0.1 μm. The lateral resolution, which is a small fraction of the stylus radius, is possible because the slopes on the polished surfaces are small. The loading on the stylus is adjustable and can be set so that no permanent marks are made on the surface (Dancy and Bennett, 1977). The main drawback to all stylus-type instruments is that one cannot preselect a typical area on the sample except by visual observation using a low power microscope. Table I

TABLE I
COMPARISON OF METHODS FOR OBTAINING STATISTICS OF OPTICAL SURFACES

	Height sensitivity	Lateral resolution	Maximum length
FECO interferometer	~3 Å rms	~2.0 μm	1 mm
Talystep profilometer	~1 Å rms	~0.1 μm	2 mm
Stereo microscopy	±20 Å	~5 Å	~1 μm

summarizes the characteristics of the three methods of determining surface statistics.

The digitized surface scan data obtained with the profilometer are used to calculate height distribution functions, such as those shown in Fig. 4, slope distribution functions, rms roughnesses, rms slopes, and autocovariance functions. An autocovariance function obtained from a single 1-mm long scan on a polished fused quartz surface is shown in Fig. 21. (This is the same surface whose autocovariance function obtained with the FECO scanning interferometer was shown in Fig. 20.) The profilometer shows fine structure in the autocovariance function, such as the initial sharp spike produced by closely spaced scratches on the surface that was not resolved by the scanning interferometer. However, the general shape of both curves is similar for larger values of the lag length τ. The initial spike in the data in Fig. 21 has an approximately exponential shape because the surface data points were not taken close enough together to show the continuous nature of the surface. Exponential autocovariance functions are suggestive of random surface data (Elson and Bennett, 1979). Figure 22 shows the autocovariance function of a chemically polished sapphire sample that was locally very smooth but had a long-range waviness. In this case, the initial portion of the autocovariance function is a very good Gaussian, as indicated by the dashed curve which almost completely duplicates the measured curve in the insert. This is one of a very few samples we have found with Gaussian autocovariance functions. The initial portions of most of the curves are closer to exponentials, but generally cannot be fit to any simple analytic function. Figure 23 shows a curve of this type for silicon carbide. This material can be polished to an extremely low scatter surface devoid of pits and scratches, similar to the surface on bowl feed polished fused quartz (Dietz and Bennett, 1966), and is a promising material for ultraviolet and x-ray mirrors (Rehn *et al.*, 1977; Choyke *et al.*, 1977). The shape of the autocovariance function is similar to that for other scratch-free but slightly wavy surfaces.

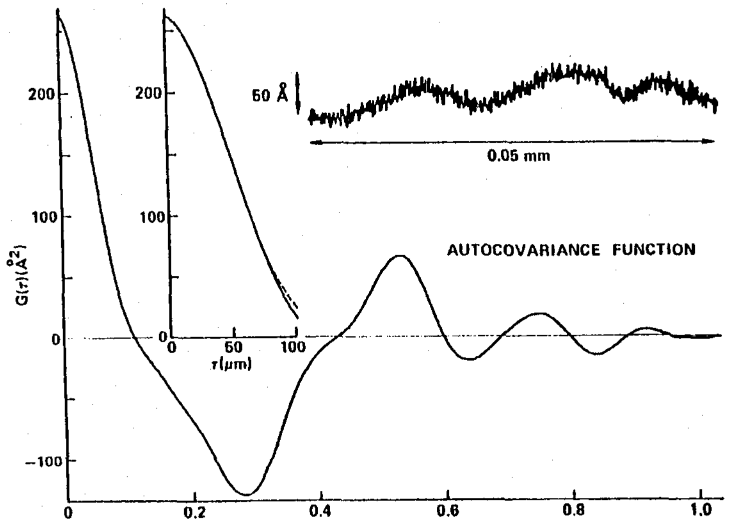


FIG. 22. Autocovariance function and surface scan obtained from profilometer data on a chemically polished sapphire surface: initial portion of the curve is an excellent fit to a Gaussian (- - -). Data are for a single surface scan. (Elson and Bennett, 1979.)

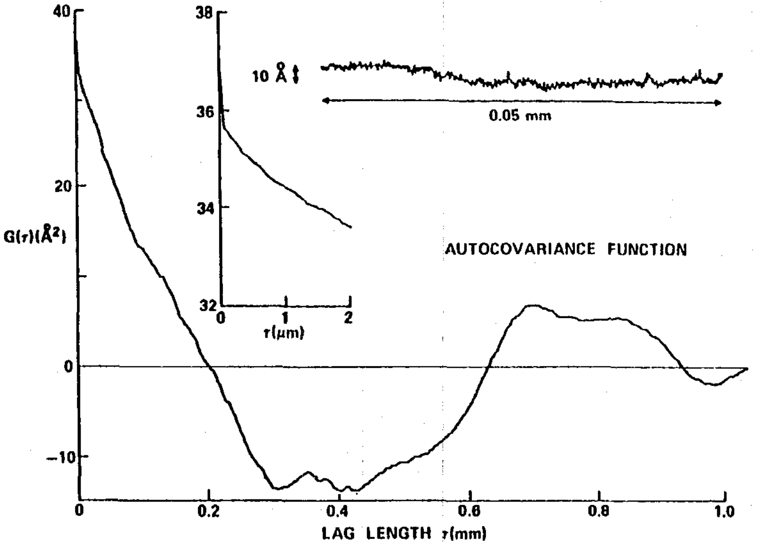


FIG. 23. Autocovariance function and surface scan obtained from profilometer data on a polished silicon carbide surface. Data are for a single surface scan. (Elson and Bennett, 1979.)

Surfaces with periodic components, such as diamond-turned surfaces or gratings have oscillatory autocovariance functions. Figure 24 shows the initial portion of a curve for one such surface, a diamond-turned copper sample made at the Lawrence Livermore Laboratory. The periodicity of about 14 μm appearing on this surface is not the basic groove separation, but is a longer-range effect probably caused by vibration between the tool and workpiece. An excellent article on scattering from diamond-turned surfaces has recently been published by Church *et al.* (1977), who show that the angular scattering profile consists of scattering peaks caused by periodic components of surface roughness superimposed on a continuous scattering background produced by the residual random roughness on the surface.

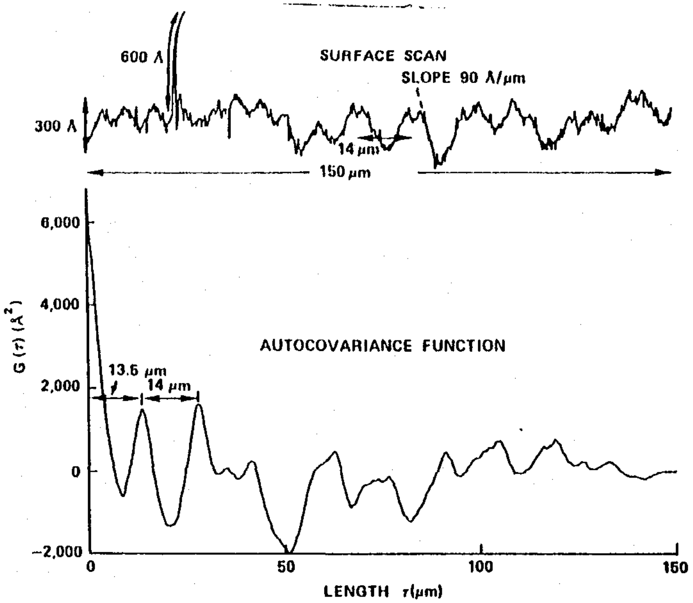


FIG. 24. Initial portion of autocovariance function and surface scan obtained from profilometer data on a diamond-turned copper sample made at the Lawrence Livermore Laboratory. Data are for a single surface scan. (Becker *et al.*, 1978.)

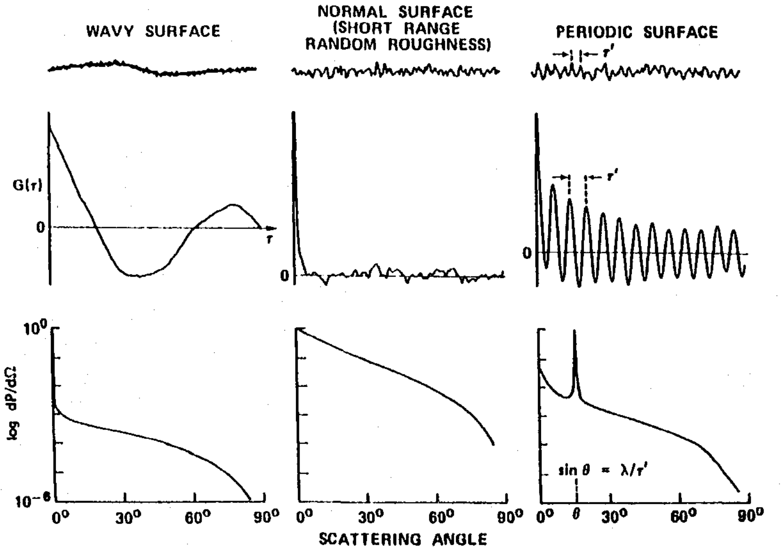


FIG. 25. Three basic types of surface profiles and their corresponding autocovariance functions and angular scattering curves. (Elson and Bennett, 1979.)

By examining the autocovariance functions in the preceding four figures, one can see common components of surface structure (Elson and Bennett, 1979). These are summarized in Fig. 25 and consist of (1) long-range waviness, (2) short-range random roughness, and (3) periodicity. The autocovariance functions and angular scattering curves for these components are also shown. Surfaces having long-range waviness scatter light predominantly very close to the specular direction, while those with short-range random roughness (most polished optical surfaces) scatter over a wider range of angles. Surfaces having only discrete periodic components will scatter light only at angles given by the grating equation, Eq. (2), as discussed in Section IV.A. To calculate angular scattering from surfaces that have more than one component of surface structure, each type of roughness can be treated separately, the angular scattering calculated, and then the scattering curves superposed (to first order). This type of approach is possible because any rough surface can be treated as a superposition of different types of surface structure.

Figure 26 shows two angular scattering curves calculated from measured surface statistics compared with angular scattering curves measured on the same samples (Elson and Bennett, 1979). The measured (dashed) curve for the rougher surface was matched to the calculated curve in the 20–40 range of scattering angles, but no other adjustments were made. The agreement between theory and experiment is encouraging, but not yet quantitative.

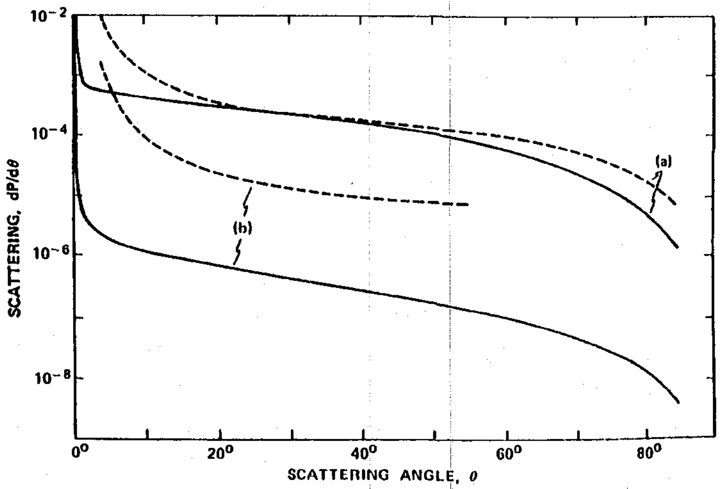


FIG. 26. (——) Calculated and (– – –) measured curves showing the angular scattering for polished fused quartz surfaces whose rms roughnesses are (a) 18.6 Å and (b) 5.2 Å. The measured curve for the rougher surface was matched to the calculated curve in the 20 40 range of scattering angles; no other adjustments were made. (Elson and Bennett, 1979.)

In particular, it appears that there may be another unaccounted for scattering mechanism operating in the case of the smoother sample. One possibility is dipole scattering from isolated particulates which may increase the entire scattering level without affecting the shape of the curve.

V. SCATTERING FROM DIELECTRIC MULTILAYERS

Optical surfaces having one or more overcoats are very common. These include tarnish or oxide layers, protective coatings, antireflection coatings, reflection increasing coatings, band pass or band elimination filters, polarizing beam splitters, etc. Along with the increased versatility to be gained by using multilayers comes the possibility of increased scattering, either from within the individual layers or at the interfaces between the layers. Although scattering within a film may be important for certain types of materials, it generally can be neglected relative to the scattering at the interfaces. For this reason we will consider only scattering at the interfaces of the multilayers. Possibilities of complicated interference effects in the scattered light from the multilayers arise because (1) the optical thicknesses of the multilayers can enhance or reduce the scattered light by interference effects just as they affect the specularly reflected or transmitted light and (2) correlation effects of the roughness across the surface of a layer or cross correlation effects of the roughness between layers can modify the angular distribution of the scattered light. Effects of correlated roughness on a single surface have already been discussed in Section IV, where we have shown how different autocovariance functions for rough surfaces produce different angular distributions of the scattered light.

Only a few people have considered scattering from surfaces having a single dielectric overlayer (Elson, 1976b, Mills and Maradudin, 1975) and even fewer have considered scattering from multilayers. Eastman (1974) used a scalar theory to investigate the TIS from surfaces covered with multilayers and Elson (1976a, 1977) used a vector perturbation method which yields the angular distribution of the scattered light. Details of the vector theory to be used here have already been published (Elson, 1977; Scott and Elson, 1978) and will not be repeated. The multilayer scattering formulas are rather complicated algebraically but are in closed form. The only assumption is that the roughness of any interface must be much less than the wavelength. In the theory the dielectric constants of the substrate and/or multilayers may be complex and the thicknesses and number of layers, the incident beam polarization, and the angles of incidence and scattering can all be specified as inputs.

We have applied the multilayer scattering formulas to two situations: (1) diffraction from a low efficiency grating covered with a reflectance enhancing dielectric stack and (2) diffuse scattering from a metallic mirror covered with a dielectric multilayer stack. In the first case we will consider the polarization properties of the light diffracted into the -1 grating order and in the second case we will compare the angular dependence of the light scattered by a single rough surface. Autocorrelation and cross correlation effects of the roughness, polarization of the incident beam, and the thicknesses of the layers will be shown to be important.

A. POLARIZATION EFFECTS IN LOW EFFICIENCY GRATINGS

The periodic rough surfaces are assumed to have a rectangular profile with the groove width equal to half of the groove spacing (duty cycle 0.5), and the groove depth H is assumed to be much less than the wavelength. The grating surface is covered with silver, with dielectric constant $\varepsilon_s = -16.4 + i0.53$. The dielectric stack on top of the silver is assumed to be composed of alternating layers of $\mathrm{MgF_2}$ with $\varepsilon_l = 1.9$ and ZnS with $\varepsilon_h = 5.29$, each of $\lambda/4$ optical thickness at the design wavelength of 6328 Å. The grating surface has a spacing d of 1.1λ, or 6961 Å, and the physical thicknesses of the $\mathrm{MgF_2}$ and ZnS layers are 1148 and 673 Å, respectively. Thus, a 10 (20) layer dielectric stack would have a total physical thickness of 0.91 μm (1.82 μm). Since the physical thicknesses of the mutilayer coatings are so large, there is a question of whether the rectangular profile on the silver grating surface replicates through the multilayers. Fortunately, there is experimental evidence which indicates that the degree of replication in multilayers on grating surfaces may be quite good (Elson, 1976b, 1977). Thus, we will assume that replication does occur, so that there is perfect correlation between all the surfaces of the multilayers, as well as perfect correlation between the grooves on any one surface.

In the following example we will consider the polarization ratio of the p- and s-reflected components in the -1 diffracted order, assuming unpolarized incident light. This information is of importance if, for example, one is using a low efficiency grating as a beam sampler and wants the sampled light in the -1 diffracted order to be independent of the polarization of the incident beam. Although the following calculations are for a grating with a rectangular profile, calculated polarization ratios (to first order) also hold for other shape profiles such as triangular and sinusoidal. However, the *intensities* of the diffracted orders do depend on the profile shape.

Figure 27 shows the calculated polarization ratio in the -1 diffracted order as a function of angle of incidence for a value of $\psi = 20$, where ψ is the angle between the plane of incidence and the groove direction, as shown in Fig. 18. We are assuming that the light is p polarized (s polarized) when the electric field is parallel (perpendicular) to the plane of incidence or diffraction. The plane of incidence (diffraction) is defined by the incoming (diffracted)

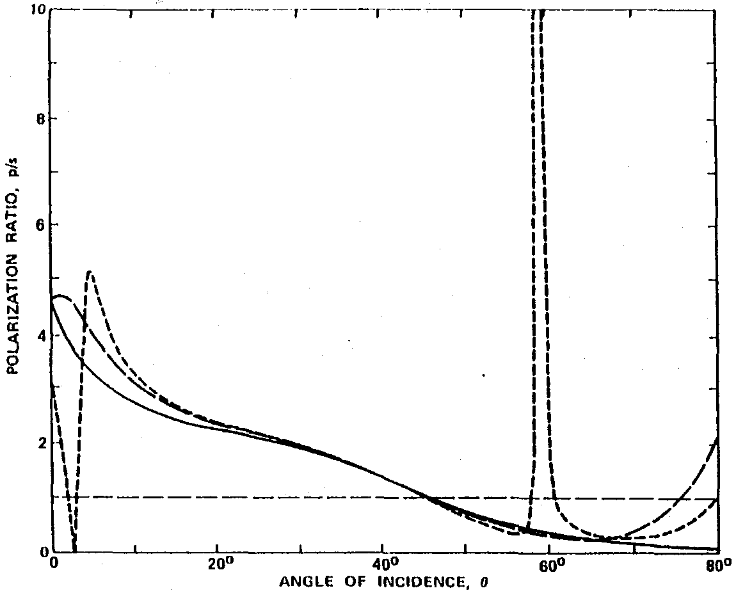


FIG. 27. Calculated polarization ratio in the -1 diffracted order of a low efficiency grating with a rectangular profile as a function of angle of incidence, for $\Psi = 20°$: (— –) bare silver grating surface; the dashed curves are for a silver grating surface overcoated with (- —) 10 or (-- –) 20 alternating $\lambda/4$ layers of Zns and MgF_2. Both sets of multilayers perfectly replicate the grating groove profile. The horizontal dashed line indicates a polarization ratio of unity. Other conditions are given in the text.

wavevector and the normal to the grating surface. The polar and azimuthal angles θ and ϕ of the diffracted beams in the -1 order can be calculated from Eqs. (14). In Fig. 27 are shown curves for the polarization ratio of (1) a bare silver grating surface, (2) a silver grating surface overcoated with 10 alternating layers of ZnS and MgF_2, and (3) a silver surface overcoated with 20 alternating layers of ZnS and MgF_2. All dielectric layers are assumed to have $\lambda/4$ optical thickness at normal incidence. It is clear from Fig. 9 that the polarization ratios in the 20-layer example behave quite differently from the other two cases, exhibiting anomalies at angles of incidence θ_0 of 2.4 and 59.5°. For an explanation of this anomalous behavior, consider for the moment the polarizing angle for a simple dielectric stack on a transparent substrate without any grating grooves present (planar interfaces). For p-polarized light incident at the polarizing angle, 59.5°, the specular reflectance of the 10 (20) layer stack is about 54% (0.025%), while both stacks have nearly 100% reflectance for s-polarized light. Hence, the 20-layer stack could serve as an excellent polarizer with the p-polarized light being transmitted and the

s-polarized light reflected. At the polarizing angle the 20-layer stack does not allow p-polarized light to be specularly reflected. If the transparent substrate were replaced by a metallic substrate with grating grooves present, then light incident in the 2.0–3.0° range yields a -1 diffracted order which coincides with the polarizing angle. Thus, the dip in the p/s ratio at $\theta_0 = 2.4$ in Fig. 27 is caused by the polarizing angle effect. The anomaly appearing at $\theta_0 = 59.5°$ is also a direct result of the polarizing angle effect. The reason for this anomalous increase in p-polarized light stems from the deep penetration of the p-polarized field into the stack when incident at the polarizing angle. The large p-polarized field strength yields large scattering currents which consequently generate an anomalous increase in the -1 order. A more detailed discussion of these effects is given elsewhere (Elson, 1979).

Two other important points are illustrated in Fig. 27. First, there are situations where it is possible to obtain polarization ratios of unity and hence the -1 diffracted order of the grating can be made polarization independent for an unpolarized incident beam. This result is also true when the optical thicknesses of the layers in the stack are $\lambda/4$ at the angle of incidence at which the grating is being used instead of at normal incidence, as is assumed in Fig. 27. Second, the choice of the incident wavelength does not affect the appearance of anomalies or the ψ dependence if the layer thicknesses are adjusted appropriately. In other words, the anomalies illustrated in Fig. 27 will also be seen at 10.6 μm at angles related to the correct polarizing angle for 10.6 μm.

B. ANGULAR SCATTERING FROM SURFACES HAVING RANDOM ROUGHNESS

The anomalies noted in the preceding section for gratings coated with multilayer dielectric films have analogues in the case of scattering from rough silver surfaces coated with multilayer dielectric films, as will be discussed in this section. Here we will consider a rough polished surface coated with silver and then overcoated with 20 alternating layers of ZnS and MgF_2, each of $\lambda/4$ optical thickness at the design wavelength, 6328 Å. Two cases will be considered: (1) uncorrelated roughness where the surfaces of the multilayers are all rough and have the same autocovariance function, but the roughness does not replicate through the stack, and (2) correlated roughness where it does. An exponential autocovariance function is assumed for each surface in both cases, with an autocovariance length of 0.35 μm. Autocovariance functions of similar form have been measured for polished quartz surfaces. The situation for real multilayer films is probably somewhere between the two extremes of correlated and uncorrelated roughness. There is evidence that evaporated silver films can replicate steps on surfaces (Koehler and Eberstein, 1953) and multilayer films can replicate grating profiles (Elson, 1976b, 1977), so presumably the substrate roughness will be replicated by multilayer dielectric films evaporated on it. However, for very smooth substrates, i.e., those having roughnesses ~5 Å rms or less, the crystallites of the evaporated or sputtered films which generally have lateral dimensions of 1000 Å or less, may modify the roughness of the substrate and produce a partially correlated situation.

Figure 28 shows $dP/d\Omega$, the scattered light per unit solid angle (incident intensity assumed to be unity), plotted as a function of polar scattering angle θ for (1) a rough silver surface, (2) a rough silver surface coated with 20 dielectric layers having correlated roughness, and (3) a rough silver surface coated with 20 dielectric layers having no roughness correlation. The light is assumed to be normally incident and the scattered light is p polarized, i.e., polarized in the plane defined by the incident wave vector and the polarization of the incident beam. In this case the azimuthal angle ϕ is 0° for positive values of θ, denoting scattering in the forward quadrant, and 180° for negative values of θ, denoting scattering into the backward quadrant. It is seen that there is very little difference in the general shapes of the three curves; all are peaked in the specular direction. However, the magnitude of the scattering for the uncorrelated roughness is much lower than for the correlated roughness. Evidently the $\lambda/4$ optical thicknesses of the dielectric films suppress the scattered light when the roughnesses on the surfaces are uncorrelated since the scattered light from the bare silver surface is almost identical to the scattered light from the correlated case. The anomalies associated with the polarizing angle that were discussed in the preceding section are also present here, but to a much lesser extent. It is seen that there is an abrupt discontinuity in the scattered light for the correlated case at scattering angles near ±60°. The effect is shown to a lesser extent for the uncorrelated case. It appears that it is possible that a $\lambda/4$ dielectric stack may not only enhance the specular reflectance of a metal surface, but at the same time decrease the amount of scattering if the stack can be made so that the roughnesses at the interfaces of the layers are as small and as uncorrelated as possible. The subject of scattering from multilayers having uncorrelated and correlated roughnesses is discussed further by Elson (1979) and Scott and Elson (1978). They show that the anomalous effects at scattering angles of ±60° do not appear for s-polarized scattered light and that the corresponding s-polarization curves are nearly identical to the curves in Fig. 28 except for the anomalies.

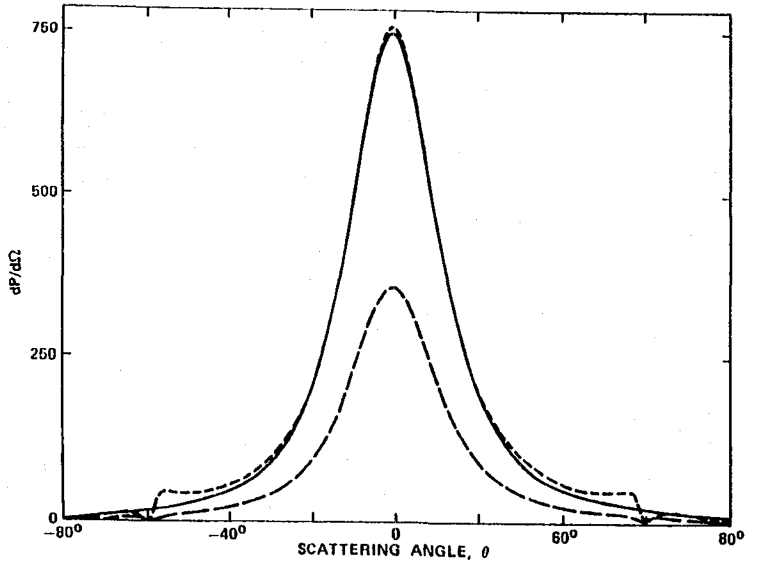


FIG. 28. Calculated scattered light per unit solid angle $dP/d\Omega$ plotted as a function of polar scattering angle θ assuming normally incident light and p-polarized scattered light: (——) rough bare silver surface; the dashed curves are for a rough silver surface overcoated with 20 alternating $\lambda/4$ layers of ZnS and MgF_2 whose roughnesses at the film interfaces are (— —) correlated or (- - -) uncorrelated. Other conditions are given in the text.

VI. SUMMARY

In this chapter we have described the types of defects that generally occur on optical surfaces and the kinds of scattering they produce. Theoretical treatments of scattering are grouped into categories, depending on the nature of the scatterers (isolated particulates, scratches, correlated random microroughness, or periodic surface profiles), and their sizes (large or small compared to the wavelength of light).

(1) The scalar theory relating TIS to rms microroughness for Gaussian distributed surface irregularities whose heights are small compared to the wavelength is in good agreement with experiment for most polished optical surfaces. As would be expected, predictions based on the scalar theory disagree with experiment in cases where the surface is particulate covered, heavily scratched, etched, or pitted. These surfaces do not have Gaussian height distributions. There is also disagreement between theory and experiment for many diamond-turned surfaces whose rms slopes are quite different from those on polished glass surfaces. Surfaces on which the predominant lateral separation of microirregularities is smaller than a wavelength also scatter less than is predicted by a simple application of the scalar theory.

(2) Vector scattering theory can, in principle, predict the angular distribution of scattered light from a surface having correlated microroughness provided that the two-dimensional spectral density function for the surface

is known. Experimentally, only one-dimensional *is known. Experimentally, only one-dimensional surface statistics* have been measured and there are mathematical difficulties in using these data to accurately calculate two-dimensional scattering from areas on surfaces. Qualitative agreement between theory and experiment has, however, been obtained.

(3) Mie theory can, in principle, be used to calculate scattering from isolated, uncorrelated particulates. In practice, this theory is only tractable when the particles are assumed to have simple shapes, such as spheres, ellipsoids, platelets, etc., sizes in a limited range, and known dielectric constants. There are difficulties therefore in using it to calculate accurate scattering levels from real surfaces covered by dust particles whose shapes, size distribution, and dielectric constants are largely unknown. Also, interaction effects between the particulates and the surface on which they rest are difficult to handle theoretically. Qualitative agreement between theory and experiment for dust-covered surfaces is, however, probable.

(4) Scattering from scratches, digs, and other surface defects whose dimensions are large compared to the wavelength can, in principle, be handled by geometrical optics. However, in practice this is difficult because the exact shapes, sizes and orientations of facets making up the surface defect must be known in detail. Typically, there are not enough of these defects on the surface to be able to apply statistics with any degree of confidence. The TIS from scratches and digs is in many cases related approximately to their widths, making some quantification of the scratch–dig standards that are used for inspection and quality control of optical surfaces possible.

(5) The scalar theory predicting the scattering from densely packed, correlated surface microirregularities whose dimensions and separations are comparable to or larger than a wavelength is in qualitative agreement with experiment, but the height distribution function and autocovariance function for the rough surface must both be known.

(6) Finally, both TIS and the angular dependence of scattering from microrough surfaces covered with dielectric multilayers can now be handled theoretically for any angle of incidence provided that the roughness is small compared to the wavelength, scattering within the layers is neglected, the autocovariance function for the surface roughness at the interfaces is known, and the correlation between the roughnesses of the different multilayers is either total or absent. Predictions from this multilayer vector scattering theory have not yet been verified experimentally.

REFERENCES

Archibald, P. C., and Bennett, H. E. (1978). *Opt. Eng.* **17**, 480.

Beaglehole, D., and Hunderi, O. (1970). *Phys. Rev. B* **2**, 309.

Becker, D. L., Bennett, J. M., Foileau, M. J., Porteus, J. O., and Bennett, H. E. (1978). Surface and optical studies of diamond turned and other metal mirrors, *Opt. Eng.* **17**, 160.

Beckmann, P., and Spizzichino, A. (1963). "The Scattering of Electromagnetic Waves From Rough Surfaces." Macmillan, New York.

Bendat, J. S., and Piersol, A. G. (1971). "Random Data: Analysis and Measurement Procedures," p. 311. Wiley (Interscience), New York.

Bennett, H. E. (1978). Scattering characteristics of optical materials, *Opt. Eng.* **17**, 480.

Bennett, H. E., and Porteus, J. O. (1961). *J. Opt. Soc. Am.* **51**, 123.

Bennett, H. E., and Stanford, J. L. (1976). *J. Res. Natl. Bur. Stand., Sect. A* **80**, 643.

Bennett, J. M. (1976). *Appl. Opt.* **15**, 2705.

Berreman, D. W. (1970). *Phys. Rev. B* **1**, 381.

Bloembergen, N. (1973). *Appl. Opt.* **12**, 661.

Celli, V., Marvin, A., and Toigo, F. (1975). *Phys. Rev. B* **11**, 1779.

Chandley, P. J. (1976). *Opt. Quantum Electron.* **8**, 323, 329.

Chandley, P. J., and Welford, W. T. (1975). *Opt. Quantum Electron.* **7**, 393.

Chinmayanandam, T. K. (1919). *Phys. Rev.* **13**, 96.

Choyke, W. J., Partlow, W. D., Supertzi, E. P., Venskytis, F. J., and Brandt, G. B. (1977). *Appl. Opt.* **16**, 2013.

Church, E. L., and Zavada, J. M. (1975). *Appl. Opt.* **14**, 1788.

Church, E. L., Jenkinson, H. A., and Zavada, J. M. (1977). *Opt. Eng.* **16**, 360.

Dancy, J. H., and Bennett, J. M. (1976). *In* "High Energy Laser Mirrors and Windows," Semiannu. Rep. Nos. 7 and 8, pp. 166–176. NWC TP 5845 (July). Naval Weapons Center, China Lake, California.

Dancy, J. H., and Bennett, J. M. (1977). *In* "High Energy Laser Mirrors and Windows," Annu. Rep. No. 9, pp. 133–144. NWC TP 5988 (Nov.). Naval Weapons Center, China Lake, California.

Davies, H. (1954). *Proc. Inst. Electr. Eng.* **101**, 209.

Dietz, R. W., and Bennett, J. M. (1966). *Appl. Opt.* **5**, 881.

Ditchburn, R. W. (1964). "Light," 2nd Ed., pp. 209–210, 394. Wiley (Interscience), New York.

Eastman, J. M. (1974). Ph D Thesis, Univ. of Rochester, Rochester, New York.

Eastman, J. M., and Baumeister, P. W. (1974). *Opt. Commun.* **12**, 418.

Elson, J. M. (1975). *Phys. Rev. B* **12**, 2541.

Elson, J. M. (1976a). "Low Efficiency Diffraction Grating Theory," AFWL-TR-75-210 (Mar). Kirtland Air Force Base, Albuquerque, New Mexico.

Elson, J. M. (1976b). *J. Opt. Soc. Am.* **66**, 682.

Elson, J. M. (1977). *Appl. Opt.* **16**, 2872.

Elson, J. M. (1979). *J. Opt. Soc. Am.* **69**, 48.

Elson, J. M., and Bennett, J. M. (1979). *J. Opt. Soc. Am.* **69**, 31.

Elson, J. M., and Ritchie, R. H. (1971). *Phys. Rev. B* **4**, 4129.

Elson, J. M., and Ritchie, R. H. (1974). *Phys. Stat. Solidi B* **62**, 461.

Fung, A. K., and Chan, H. (1969). *IEEE Trans. Antennas Propag.* AP-17, 590.

Fung, A. K., and Moore, R. K. (1966). *J. Geophys. Res.* **71**, 2939.

Hagfors, T. (1966). *J. Geophys. Res.* **71**, 379.

Holzer, J. A., and Sung, C. C. (1976). *J. Appl. Phys.* **47**, 3363.

Holzer, J. A., and Sung, C. C. (1977). *J. Appl. Phys.* **48**, 1739.

Jackson, J. D. (1962). "Classical Electrodynamics." Wiley, New York.

Kerker, M. (1969). "The Scattering of Light and Other Electromagnetic Radiation." Academic Press, New York.

Koehler, W. F., and Eberstein, A. (1953). *J. Opt. Soc. Am.* **43**, 747.

Kozawa, S. (1962). *In* "Proceedings of the Conference on Optical Instruments and Techniques" (K. J. Habell, ed.), pp. 410–428. Chapman and Hall, Ltd., London.

Kröger, E., and Kretschmann, E. (1970). *Z. Phys.* **237**, 1.

Leader, J. C. (1971a). "Spectral and Angular Dependence of Specular Scattering from Rough Surfaces," MCAIR 71-013 (Apr). McDonnell Aircraft Company, Saint Louis, Missouri.

Leader, J. C. (1971b). *J. Appl. Phys.* **42**, 4808.

Leader, J. C., and Fung, A. K. (1977). *J. Appl. Phys.* **48**, 1736.

Lee, Y. W. (1960). "Statistical Theory of Communication," pp. 200–218. Wiley, New York.

Maradudin, A. A., and Mills, D. L. (1975). *Phys. Rev. B* **11**, 1392.

Mie, G. (1908). *Ann. Phys.* (*Leipzig*) **25**, 377.

Mills, D. L., and Maradudin, A. A. (1975). *Phys. Rev. B* **12**, 2943.

Nicodemus, F. E. (1970). *Appl. Opt.* 9, 1474.

Orr, C., Jr., and Dallavalle, J. M. (1959). "Fine Particle Measurement," p. 119. Macmillan, New York.

Peake, W. H. (1959). *IRE Trans. Antennas Propag.* **AP-7**, Spec. Suppl., S324.

Petit, R. (1975). *Nouv. Rev. Opt.* **6**, 134.

Porteus, J. O. (1963). *J. Opt. Soc. Am.* **53**, 1394.

Rayleigh, Lord (1945). "The Theory of Sound," Vol. 2. Dover, New York.

Rehn, V., Stanford, J. L., Baer, A. D., Jones, V. O., and Choyke, W. J. (1977). *Appl. Opt.* **16**, 1111.

Scott, M. L., and Elson, J. M. (1978). *Appl. Phys. Lett.* **32**(3), 158.

Silver, S. (1947). "Microwave Antenna Theory and Design," p. 161. McGraw-Hill, New York.

Stover, J. C. (1975). *Appl. Opt.* **14**, 1796.

Stratton, J. A. (1941). "Electromagnetic Theory," pp. 415, 566. McGraw-Hill, New York.

Truong, V. V., and Scott, G. D. (1976). *J. Opt. Soc. Am.* **66**, 124.

van de Hulst, H. C. (1957). "Light Scattering by Small Particles." Wiley, New York.

Welford, W. T. (1977). *Opt. Quantum Electron.* **9**, 269.

Young, R. P. (1976). *Opt. Eng.* **15**, 516.

Zaki, K. A., and Neureuther, A. R. (1971). *IEEE Trans. Antennas Propag.* **AP-19**, 208.

APPLIED OPTICS AND OPTICAL ENGINEERING, VOL. VII

CHAPTER 8

Adaptive Optical Techniques for Wave-Front Correction

JAMES E. PEARSON, R. H. FREEMAN,
and
HAROLD C. REYNOLDS, JR.

United Technologies Research Center
Optics and Applied Technology Laboratory
West Palm Beach, Florida

Copyright © 1979 by Academic Press, Inc.
All rights of reproduction in any form reserved
ISBN 0-12-408607-1

J Kible REF. TT

Copyright © 1979 by [illegible]
All rights of reproduction in any form reserved
ISSN 0012-4086/07

PHYSICAL REVIEW D VOLUME 23, NUMBER 8 15 APRIL 1981

Quantum-mechanical noise in an interferometer

Carlton M. Caves
W. K. Kellogg Radiation Laboratory, California Institute of Technology, Pasadena, California 91125

(Received 15 August 1980)

The interferometers now being developed to detect gravitational waves work by measuring the relative positions of widely separated masses. Two fundamental sources of quantum-mechanical noise determine the sensitivity of such an interferometer: (i) fluctuations in number of output photons (photon-counting error) and (ii) fluctuations in radiation pressure on the masses (radiation-pressure error). Because of the low power of available continuous-wave lasers, the sensitivity of currently planned interferometers will be limited by photon-counting error. This paper presents an analysis of the two types of quantum-mechanical noise, and it proposes a new technique—the "squeezed-state" technique—that allows one to decrease the photon-counting error while increasing the radiation-pressure error, or vice versa. The key requirement of the squeezed-state technique is that the state of the light entering the interferometer's normally unused input port must be not the vacuum, as in a standard interferometer, but rather a "squeezed state"—a state whose uncertainties in the two quadrature phases are unequal. Squeezed states can be generated by a variety of nonlinear optical processes, including degenerate parametric amplification.

I. INTRODUCTION

The task of detecting gravitational radiation is driving dramatic improvements in a variety of technologies for detecting very weak forces.[1] These improvements are forcing a careful examination of quantum-mechanical limits on the accuracy with which one can monitor the state of a macroscopic body on which a weak force acts.[2] One promising technology uses an interferometer to monitor the relative positions of widely separated masses. This paper analyzes the quantum-mechanical limits on the performance of interferometers, and it introduces a new technique that might lead to improvements in their sensitivity.

The prototypal interferometer for gravitational-wave detection is a two-arm, multireflection Michelson system, powered by a laser (see Fig. 3 below). The intensity in either of the interferometer's output ports provides information about the difference $z \equiv z_2 - z_1$ between the end mirrors' positions relative to the beam splitter, and changes in z reveal the passing of a gravitational wave. The first interferometer for gravitational-wave detection was built and operated at the Hughes Research Laboratories in Malibu, California, in the early 1970's (Ref. 3); this first effort was small-scale and had modest sensitivity. Now several groups around the world are developing interferometers of greatly improved sensitivity.[4-6] A long-range goal is to construct large-scale interferometers, with baselines $l \sim 1$ km, in order to achieve a strain sensitivity $\Delta z/l \sim 10^{-21}$ for frequencies from about 30 Hz to 10 kHz. This sensitivity goal is based on estimates for the strength of gravitational waves that pass the Earth reasonably often.[1]

It has been known for some time that quantum mechanics limits the accuracy with which an interferometer can measure z—or, indeed, the accuracy with which any position-sensing device can determine the position of a free mass.[2,5,7] In a measurement of duration τ, the probable error in the interferometer's determination of z can be no smaller than the "standard quantum limit":

$$(\Delta z)_{\rm SQL} = (2\hbar\tau/m)^{1/2}\,, \tag{1.1}$$

where m is the mass of each end mirror $[(\Delta z)_{\rm SQL} \sim 6\times 10^{-18}$ cm for $m \sim 10^5$ g, $\tau \sim 2\times 10^{-3}$ sec]. The validity of the standard quantum limit is unquestionable, resting as it does solely on the Heisenberg uncertainty principle applied to the quantum-mechanical evolution of a free mass.

The standard quantum limit for an interferometer can also be obtained from a more detailed argument[5,8-10] that balances two sources of error: (i) the error in determining z due to fluctuations in the number of output photons (photon-counting error) and (ii) the perturbation of z during a measurement produced by fluctuating radiation-pressure forces on the end mirrors (radiation-pressure error). As the input laser power P increases, the photon-counting error decreases, while the radiation-pressure error increases. Minimizing the total error with respect to P yields a minimum error of order the standard quantum limit and an optimum input power[9,11]

$$P_0 \simeq \tfrac{1}{2}(mc^2/\tau)(1/\omega\tau)(1/b^2) \tag{1.2}$$

at which the minimum error can be achieved. Here ω is the angular frequency of the light, and b is the number of bounces at each end mirror.

© 1981 The American Physical Society

At the optimum power the photon-counting and radiation-pressure errors are equal.

The original argument[8] leading to the optimum power (1.2) attributed the radiation-pressure fluctuations that perturb z to fluctuations in input power. Some later versions of the argument were unclear about the source of the relevant radiation-pressure fluctuations.[5,9] As a result, the argument had always been under suspicion,[8] because fluctuations in input power should divide equally at the beam splitter and, therefore, should have no effect on z. This suspicion led to a "lively but unpublished controversy"[6] over the existence of a radiation-pressure force that affects z and, consequently, over the existence of an optimum laser power.

In a recent paper I resolved this controversy.[11] There I pointed out that the relevant radiation-pressure force has nothing to do with input power fluctuations; instead, it can be attributed to vacuum (zero-point) fluctuations in the electromagnetic field, which enter the interferometer from the unused input port (direction of dashed arrow in Fig. 3 below). When superposed on the input laser light, these fluctuations produce a fluctuating force that perturbs z. (An alternative and equivalent point of view attributes the relevant radiation-pressure fluctuations to random scattering of the input photons at the beam splitter.[10,11]) In the same paper I claimed that these vacuum fluctuations (not input power fluctuations) are also responsible for the unavoidable fluctuations in number of output photons.

A reasonable set of values for the interferometer's parameters, which I shall use as a fiducial set throughout this paper, is $m \sim 10^5$ g, $\tau \sim 2\times 10^{-3}$ sec, $\omega \sim 4\times 10^{15}$ rad sec^{-1} (wavelength $\lambda \sim 5000$ Å), and $b \sim 200$. For these values P_0 is approximately 8×10^3 W – a power far higher than powers of present continuous-wave lasers. The low available input power means that the interferometers now planned for use as gravitational-wave detectors will be limited not by the standard quantum limit, but rather by $1/\sqrt{N}$ photon-counting statistics. In this paper I address this problem by introducing a new technique, which in principle allows an interferometer to achieve the quantum-limited position sensitivity (1.1) for input powers far less than P_0.

Perhaps suprisingly, this new technique does not require modifying the input laser light; instead, it requires modifying the light entering the normally unused input port. Specifically, the unused port must see not the vacuum (ground) state of the electromagnetic field, but rather a "squeezed state"—a state whose fluctuations in one quadrature phase are less than zero-point fluctuations (or the fluctuations in any coherent state), and whose fluctuations in the other phase are greater than zero-point fluctuations. This technique works because one of the two phases is responsible for the fluctuations in number of output photons, while the other is responsible for the radiation-pressure fluctuations that perturb z, Thus, by "squeezing the vacuum" before it can enter the normally unused input port, one can reduce the photon-counting error (i.e., beat $1/\sqrt{N}$ photon-counting statistics) at the expense of increasing the radiation-pressure error, or vice versa.

In practice, this squeezed-state technique is not likely to allow gravitational-wave interferometers to operate at the standard quantum limit. However, it might allow a given underpowered interferometer to achieve a somewhat better sensitivity, without changes in its input power or any of its other parameters. Unfortunately, the usefulness of the squeezed-state technique is likely to be severely limited by the losses in real mirrors, which destroy the crucial feature of the technique—the reduced noise in one of the two quadrature phases. The technique can be useful only in interferometers whose performance is not limited by losses in the mirrors.

This paper extends and refines the analysis given in Ref. 11, and it introduces the new squeezed-state technique. Section II gives a detailed analysis of the quantum-mechanical noise in an interferometer, with emphasis on the theoretical capabilities of the squeezed-state technique. Section II A presents some formal considerations that facilitate handling various states of the electromagnetic field, including squeezed states. Section II B begins by presenting an idealized model of an interferometer and an outline of the procedures used in the subsequent analysis, and it then proceeds to that analysis. Specifically, the radiation-pressure fluctuations, the output fluctuations in number of photons, the optimum sensitivity, and the optimum power are analyzed for an interferometer that has either vacuum or a squeezed state incident on the normally unused input port. Along the way the intensity-correlation properties of the light in the two arms of the interferometer are investigated. Section III focuses on more practical matters, with emphasis on application of the squeezed-state technique to real interferometers, which are limited by photon-counting statistics. Section III reviews a method for generating squeezed states using an optical degenerate parametric amplifier, it investigates the limitations imposed by mirror losses, and it proposes a method for doing the photon counting that can realize the potential reduction in photon-counting error even with inefficient photodetectors. Section IV comments briefly on the results and their relevance to gravitational-wave detection.

II. ANALYSIS OF QUANTUM-MECHANICAL NOISE IN AN INTERFEROMETER

A. Formal considerations

Before turning to a detailed analysis of an interferometer, it is useful to review some properties of various special states of a harmonic oscillator. These states play a crucial role in the subsequent analysis.

Consider a single mode of the electromagnetic field with angular frequency ω, and let a and $a^\dagger$ be its annihilation and creation operators ($[a, a^\dagger]=1$). Then the operator for the number of photons in the mode is

$$N=a^\dagger a\,, \tag{2.1}$$

and a dimensionless complex-amplitude operator for the mode is

$$X_1+iX_2=a\,. \tag{2.2}$$

The Hermitian operators X_1 and X_2 (real and imaginary parts of the complex amplitude) give dimensionless amplitudes for the mode's two quadrature phases. The commutation relation for a and $a^\dagger$ impiies a corresponding commutation relation for X_1 and X_2: $[X_1,X_2]=i/2$. The resulting uncertainty principle is $\Delta X_1 \Delta X_2 \geq \frac{1}{4}$.

The complex amplitude of a single mode is a constant of the motion—i.e., it is constant in the Heisenberg picture. Thus, in Eq. (2.2), X_1 and X_2 are Heisenberg-picture operators, whereas a is a Heisenberg-picture operator evaluated at a particular time (or, equivalently, a Schrödinger-picture operator). There is a phase ambiguity in the relation between the complex amplitude and the annihilation operator; this phase ambiguity corresponds to the freedom to make rotations in the complex-amplitude plane (or to freedom in the choice of fiducial time in the relation between X_1+iX_2 and a).

A particularly useful set of states for the electromagnetic field is the set of coherent states introduced by Glauber.[12] These states are most easily generated using the unitary *displacement operator*[12]:

$$D(\alpha)\equiv \exp(\alpha a^\dagger - \alpha^* a)=e^{-|\alpha|^2/2}e^{\alpha a^\dagger}e^{-\alpha^* a}\,, \tag{2.3}$$

where α is an arbitrary complex number. Note that $D^\dagger(\alpha)=D^{-1}(\alpha)=D(-\alpha)$. The most useful property of the displacement operator is the way it transforms a and $a^\dagger$:

$$\begin{aligned} D^\dagger(\alpha)aD(\alpha)&=a+\alpha\,,\\ D^\dagger(\alpha)a^\dagger D(\alpha)&=a^\dagger+\alpha^*\,. \end{aligned} \tag{2.4}$$

By displacing the vacuum (ground) state $|0\rangle$, one obtains the *coherent state* $|\alpha\rangle$:

$$|\alpha\rangle \equiv D(\alpha)|0\rangle = e^{-|\alpha|^2/2}e^{\alpha a^\dagger}|0\rangle\,. \tag{2.5}$$

The expectation values and variances of X_1, X_2, and N in a coherent state are given by

$$\begin{aligned} \langle X_1+iX_2\rangle&=\alpha\,, \quad \Delta X_1=\Delta X_2=\tfrac{1}{2}\,,\\ \langle N\rangle&=|\alpha|^2\,, \quad \Delta N=|\alpha|\,. \end{aligned} \tag{2.6}$$

The coherent state $|\alpha\rangle$ has mean complex amplitude α, and it is a minimum-uncertainty (Gaussian) state for X_1 and X_2, with equal uncertainties in the two quadrature phases. A coherent state is conveniently represented by an "error circle" in a complex-amplitude plane whose axes are X_1 and X_2 [see Fig. 1(a)]. The center of the error circle lies at $\langle X_1+iX_2\rangle=\alpha$, and the radius $\Delta X_1 = \Delta X_2 = \frac{1}{2}$ accounts for the uncertainties in X_1 and X_2.

Squeezed states constitute another useful set of states. They are conveniently generated by using the unitary *squeeze operator*[13-15]:

$$S(\zeta)\equiv \exp[\tfrac{1}{2}\zeta^* a^2 - \tfrac{1}{2}\zeta(a^\dagger)^2]\,, \quad \zeta = re^{i\theta}\,, \tag{2.7}$$

where ζ is an arbitrary complex number. The squeeze operator was introduced by Stoler,[13] and the name was coined by Hollenhorst.[15] Note that $S^\dagger(\zeta)=S^{-1}(\zeta)=S(-\zeta)$. The most useful unitary transformation properties of the squeeze operator are

$$\begin{aligned} S^\dagger(\zeta)aS(\zeta)&=a\cosh r - a^\dagger e^{i\theta}\sinh r\,,\\ S^\dagger(\zeta)a^\dagger S(\zeta)&=a^\dagger\cosh r - ae^{-i\theta}\sinh r\,,\\ S^\dagger(\zeta)(Y_1+iY_2)S(\zeta)&=Y_1e^{-r}+iY_2e^{r}\,, \end{aligned} \tag{2.8}$$

where

$$Y_1+iY_2=(X_1+iX_2)e^{-i\theta/2} \tag{2.9}$$

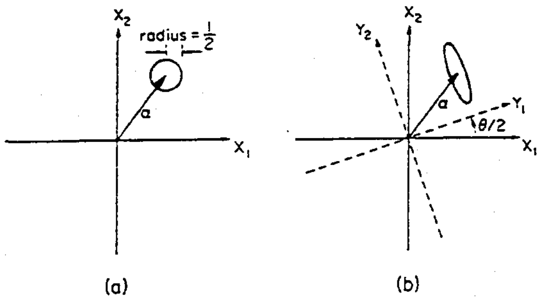


FIG. 1. (a) Error circle in complex-amplitude plane for coherent state $|\alpha\rangle$. (b) Error ellipse in complex-amplitude plane for squeezed state $|\alpha, re^{i\theta}\rangle$ ($r>0$).

is a rotated complex amplitude. The squeeze operator attenuates one component of the (rotated) complex amplitude, and it amplifies the other component. The degree of attenuation and amplification is determined by $r = |\zeta|$, which therefore will be called the *squeeze factor*.

The *squeezed state* $|\alpha,\zeta\rangle$ is obtained by first squeezing the vacuum and then displacing it:

$$|\alpha,\zeta\rangle \equiv D(\alpha)S(\zeta)|0\rangle . \tag{2.10}$$

Note that $|\alpha,0\rangle = |\alpha\rangle$. The most important expectation values and variances for a squeezed state are

$$\begin{aligned}
&\langle X_1 + iX_2\rangle = \langle Y_1 + iY_2\rangle e^{i\theta/2} = \alpha ,\\
&\Delta Y_1 = \tfrac{1}{2}e^{-r}, \quad \Delta Y_2 = \tfrac{1}{2}e^{r} ,\\
&\langle N\rangle = |\alpha|^2 + \sinh^2 r ,\\
&(\Delta N)^2 = |\alpha \cosh r - \alpha^* e^{i\theta} \sinh r|^2\\
&\qquad\qquad + 2\cosh^2 r \sinh^2 r .
\end{aligned} \tag{2.11}$$

The squeezed state $|\alpha,\zeta\rangle$ has the same expected complex amplitude as the corresponding coherent state $|\alpha\rangle$, and it is a minimum-uncertainty (Gaussian) state for Y_1 and Y_2. The difference lies in its unequal uncertainties for Y_1 and Y_2. In the complex-amplitude plane, the coherent-state error circle has been "squeezed" into an "error ellipse" of the same area [see Fig. 1(b)]. The principal axes of the ellipse lie along the Y_1 and Y_2 axes, and the principal radii are ΔY_1 and ΔY_2.

Squeezed states were introduced by Stoler,[13,16] who called them "minimum-uncertainty packets." They have since been considered by Lu[14,17] ("new coherent states"), by Yuen[18,19] ("two-photon coherent states"), and by Hollenhorst[15] ("wave-packet states"). Yuen, in particular, has examined in detail the properties of squeezed states.[19] The reduced uncertainty in one quadrature phase of a squeezed state is obviously attractive for optical communications purposes; in a recent series of papers,[20-22] Yuen and his collaborators have considered this application of squeezed states and have given detailed analyses of several photodetection techniques applied to squeezed states.

Squeezed states have also found application in the theory of mechanically resonant gravitational-wave detectors. A detector of this type[1] is a macroscopic mechanical system (usually a massive cylinder of aluminum); a gravitational wave betrays its presence by changing the complex amplitude of oscillation of some mode of the mechanical system (usually the fundamental mode). A fundamental theoretical problem has been how to monitor the mode of interest in a way that allows detection of gravitational waves so weak that they change the complex amplitude of that mode by less than the width of a coherent state. This problem has become known as the "quantum nondemolition" problem, and solutions to it are known as quantum nondemolition measurement techniques. The quantum nondemolition problem and its solutions have been analyzed extensively (see Ref. 2 and references cited therein). It should not be surprising that squeezed states, with their reduced uncertainty in one component of the complex amplitude, play a key role in one quantum nondemolition technique, the "back-action-evading" method of making quantum nondemolition measurements.[2] In the back-action-evading method one designs a measuring device that monitors only one component of the relevant mode's complex amplitude; this device automatically forces that mode into a state similar to a squeezed state.

To better visualize the properties of coherent and squeezed states, it is perhaps useful to consider the time dependences of the expectation value and variance of a field quantity, such as the electric field $E(t)$. These time dependences are easily read off the complex-amplitude plots in Fig. 1; a complex-amplitude plane whose axes are E and $\dot{E}$ must rotate with angular velocity ω relative to the (X_1,X_2) phase plane, in order to produce a sinusoidal oscillation of the expectation value of E. For a coherent state the rotation of the error circle leads to a constant value for the variance of the electric field [see Fig. 2(a)]. For a squeezed state, however, the rotation of the error ellipse leads to a variance that oscillates with frequency 2ω. This situation is depicted in Fig. 2 for two cases: the case where the coherent excitation of the mode appears in the quadrature phase that has reduced noise [Fig. 2(b)], and the case where the coherent excitation appears in the quadrature phase that has increased noise [Fig. 2(c)].

B. Detailed analysis of an interferometer.

1. Model interferometer and outline of analysis

A typical interferometer for gravitational-wave detection is a multireflection Michelson system of the sort sketched in Fig. 3 (Refs. 4-6,8). An idealized version of such an interferometer works as follows. Light enters the interferometer from a laser, is split at a lossless, 50-50 beam splitter, bounces back and forth many times between perfectly reflecting mirrors in the nearly equal-length arms, and finally is recombined at the beam splitter. The number of bounces at each end mirror is denoted by b. The end mirrors are attached to large masses, each of mass [illegible] The beam splitter and the inner mirrors [illegible]

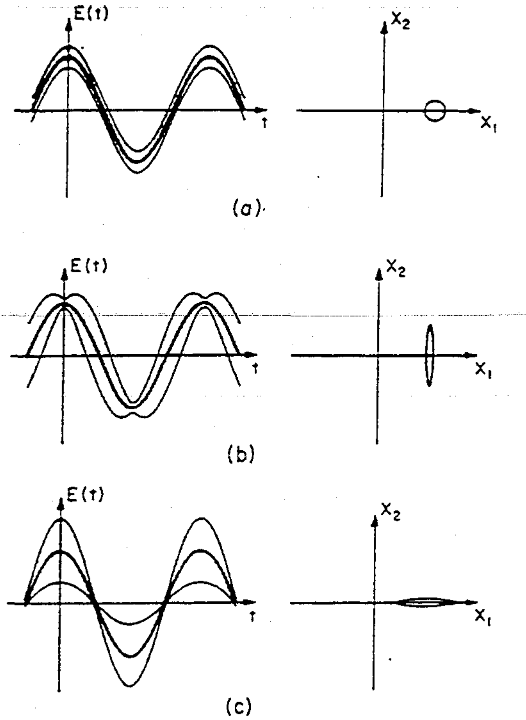


FIG. 2. Graphs of electric field versus time for three states of the electromagnetic field. In each graph the dark line is the expectation value of the electric field, and the shaded region represents the uncertainty in the electric field. To the right of each graph is the corresponding "error box" in the complex-amplitude plane. (a) Coherent state $|\alpha\rangle$ (α real). This state exhibits neither bunching nor antibunching ($g_{12}^{(2)}=1$). (b) Squeezed state $|\alpha,r\rangle$ (α real) with $r>0$. This state exhibits antibunching ($g_{12}^{(2)}<1$) as long as $0<r\le\frac{1}{4}\ln(8\alpha^2)$. (c) Squeezed state $|\alpha,r\rangle$ (α real) with $r<0$. This state exhibits bunching.

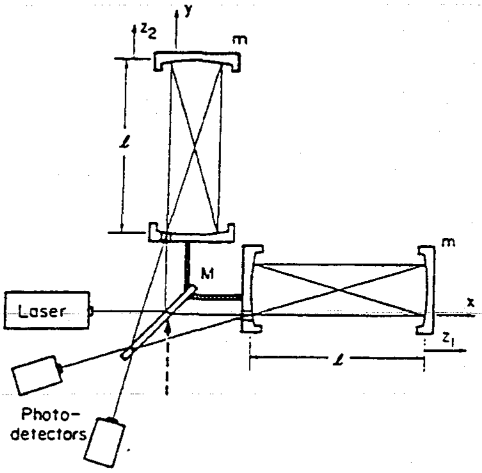


FIG. 3. Schematic diagram of Michelson interferometer ($b=2$) described in text.

rigidly attached to one another and to a mass M. For simplicity I assume $M\gg m$, so that the radiation-pressure-induced motion of the beam splitter can be ignored and the beam splitter can be regarded as at rest. Each arm of the interferometer has a fiducial length l, and the displacements of the end mirrors from their fiducial positions are denoted z_1 and z_2.

The intensity in one—or perhaps both—of the output ports is measured by an ideal photodetector (quantum efficiency one), and this measurement provides information about the difference $z\equiv z_2-z_1$ between the positions of the end mirrors. The information about z is not the instantaneous value; rather, it is some sort of average of z over the storage time—the time $\tau_s=2bl/c$ the light spends in each arm. Thus the storage time defines the interferometer's time resolution; the best sensitivity is achieved when the measurement time τ— the time over which one averages the output to get a value for z—is approximately equal to τ_s. For a baseline $l\sim1$ km and $b\sim200$, $\tau_s\sim10^{-3}$ sec. Throughout the following analysis I assume $\tau_s\lesssim\tau$.

The most important idealizations in this model are the assumptions of lossless mirrors and ideal photodetectors. The consequences of relaxing these two assumptions are considered in Sec. III.

The goal of this section is to analyze the quantum-mechanical limits on the performance of an interferometer. The philosophy is ruthless simplification: throw away all details not necessary for understanding the fundamental limits. The quantum-mechanical uncertainty in the interferometer's determination of z can be thought of as coming from three sources: (i) the intrinsic quantum-mechanical uncertainties in the end mirrors' positions and momenta; (ii) the perturbations of the end mirrors' positions by radiation-pressure fluctuations (radiation-pressure error); and (iii) the fluctuations in number of photons at the output ports (photon-counting error). In reality, all three sources of error manifest themselves in the same way—by feeding into the interferometer's output and producing fluctuations in that output. A complete analysis must consider all three simultaneously.[23] Nonetheless, the division of the total uncertainty is a useful conceptual device, and it serves as the basis of the simplified approach adopted here: calculate the error produced by each type of uncertainty separately, and then assume that the total error is the quadrature sum of the separate errors.

The intrinsic quantum-mechanical uncertainties in the end mirrors' positions and momenta can be dealt with most easily. They feed into the interferometer's output and degrade its determination of z. In the best case these uncertainties enforce a minimum error in z given by the standard quantum limit (1.1).[2,5,7] These uncertainties and the limits they impose on measurements of z are well understood; consequently, they need not be considered here—i.e., the end mirrors are treated as classical, rather than quantum-mechanical, objects.

The radiation-pressure error is obtained using the following simple procedure. The momentum transferred to the end masses is calculated assuming the end masses remain at rest throughout the measurement—i.e., assuming $m \to \infty$. Since the end masses really do move, the actual momentum transferred will be slightly different because of the Doppler shift of the reflected radiation; this difference is negligible for the cases of interest. The perturbation of z produced by the momentum transferred is estimated by reverting to finite masses m and allowing the end mirrors to move. This perturbation of z feeds into the interferometer's output and produces a comparable error in determining z.

In much the same way the fluctuations in number of output photons are determined assuming the end mirrors are at rest. These fluctuations are then converted into the photon-counting error by considering differences between neighboring $z =$const configurations. This procedure ignores the complicated averaging produced by the storage time, but it retains the essential features of the output fluctuations in number of photons.

The above-outlined assumptions allow a further drastic simplification. Instead of dealing with beams of finite size and finite duration, I can restrict attention to only a small number of plane-wave modes of the electromagnetic field. In addition, I ignore the small angular deviations in the directions of the beams—i.e., I assume that the plane-wave modes propagate precisely along the directions of the x and y axes of Fig. 3.

2. *Radiation-pressure error and second-order coherence*

With the above assumptions one can calculate the momentum transferred to each end mirror quite simply. One finds the momentum carried by the light in each arm of the interferometer; the momentum transferred at each bounce is twice this amount.

Carrying out this procedure requires only four modes of the electromagnetic field in the presence of the beam splitter ("beam-splitter modes"). The first two modes of interest are "in" modes—modes appropriate for constructing precollision wave packets that scatter off the beam splitter. Thus they are in states in the sense of scattering theory. The first mode of interest (mode 1$^+$) describes light incident from the input (laser) port. It consists of an incident plane wave with angular frequency ω, propagating inward along the x axis, and scattered waves propagating along the two arms of the interferometer. The second mode (mode 2$^+$) is the corresponding in mode that describes light incident from the normally unused input port (light incident along the y axis; direction of dashed arrow in Fig. 3). Outside the beam splitter the electric fields of these two modes have the forms

$$E_1^+ = \begin{cases} \mathcal{Q} e^{i(kx-\omega t)} - 2^{-1/2}\mathcal{Q} e^{i(\Delta-\mu)} e^{i(ky-\omega t)}, & y>x \\ 2^{-1/2}\mathcal{Q} e^{i\Delta} e^{i(kx-\omega t)}, & y<x \end{cases}$$

$$E_2^+ = \begin{cases} 2^{-1/2}\mathcal{Q} e^{i\Delta} e^{i(ky-\omega t)}, & y>x \\ \mathcal{Q} e^{i(ky-\omega t)} + 2^{-1/2}\mathcal{Q} e^{i(\Delta+\mu)} e^{i(kx-\omega t)}, & y<x \end{cases} \tag{2.12}$$

where it is assumed that the electric fields are polarized out of the page. Here $k=\omega/c$ is the wave number, $\mathcal{Q}$ is a real constant determined by one's choice of normalization, and the overall phase shift Δ and the relative phase shift μ are properties of the beam splitter. The relation between the phase shifts the two modes suffer at the beam splitter is dictated by the assumed symmetries of the beam splitter—time-reversal invariance and reflection symmetry through the plane $x=-y$. (The further and unnecessary assumption of reflection symmetry through the plane $x=y$ was made in Ref. 11; this assumption implies $\mu=\pi/2$.)

The other two modes of interest (modes 1$^-$ and 2$^-$) are "out" modes (time reversed "in" modes). Out modes are appropriate for constructing post-collision wave packets. Modes 1$^-$ and 2$^-$ are the out modes whose exiting plane waves propagate along the x and y axes, respectively. The symmetries of the beam splitter allow one to relate the electric fields of the out modes to those of the in modes:

$$\begin{aligned} E_1^- &= 2^{-1/2} e^{-i\Delta}(E_1^+ + e^{-i\mu} E_2^+), \\ E_2^- &= 2^{-1/2} e^{-i\Delta}(E_2^+ - e^{i\mu} E_1^+). \end{aligned} \tag{2.13}$$

Now let the creation and annihilation operators for modes 1$^+$ and 2$^+$ be denoted by $a_1^\dagger$, a_1 and $a_2^\dagger$, a_2; similarly, for modes 1$^-$ and 2$^-$, $b_1^\dagger$, b_1 and $b_2^\dagger$, b_2. Equation (2.13) implies

$$\begin{aligned} b_1 &= 2^{-1/2} e^{i\Delta}(a_1 + e^{i\mu} a_2), \\ b_2 &= 2^{-1/2} e^{i\Delta}(a_2 - e^{-i\mu} a_1). \end{aligned} \tag{2.14}$$

The operator $\mathcal{P}$ that specifies the difference between the momenta transferred to the end masses is proportional to the difference between the number of photons in modes 1^- and 2^-:

$$\begin{aligned}\mathcal{P} &\equiv (2b\hbar\omega/c)(b_2^\dagger b_2 - b_1^\dagger b_1)\\ &= -(2b\hbar\omega/c)(e^{i\mu}a_1^\dagger a_2 + e^{-i\mu}a_2^\dagger a_1)\,. \end{aligned} \tag{2.15}$$

Note that, when written in terms of operators for the in modes, $\mathcal{P}$ is clearly due solely to the interference of modes 1^+ and 2^+. This is merely a restatement of the fact, true classically, that the difference in the intensity in the two arms of an interferometer is produced solely by the interference of light coming from the two input ports.

Now assume the state of the electromagnetic field is

$$|\Psi\rangle = S_2(\zeta)D_1(\alpha)|0\rangle, \quad \alpha \text{ real}, \quad \zeta = -re^{-2i\mu}\,, \tag{2.16}$$

where D_1 is the displacement operator for mode 1^+ and S_2 is the squeeze operator for mode 2^+ [see Eqs. (2.3) and (2.7)]. Mode 1^+ is in a coherent state with complex amplitude α (the choice of α real is merely a choice of phase for mode 1^+), and mode 2^+ is in a squeezed state with zero expected complex amplitude. Note that the squeeze factor r can be either positive or negative. The phase with which mode 2^+ is squeezed is chosen carefully so that, in the arms of the interferometer, the reduced-noise quadrature phase of mode 2^+ is either in phase or 90° out of phase with the coherent excitation of mode 1^+. The numbers of photons in modes 1^+ and 2^+ and their variances are given by

$$\begin{aligned}(N_1)_{\text{in}} &= \langle a_1^\dagger a_1\rangle = \alpha^2\,,\\ (\Delta N_1)_{\text{in}}{}^2 &= \alpha^2\,,\\ (N_2)_{\text{in}} &= \langle a_2^\dagger a_2\rangle = \sinh^2 r\,,\\ (\Delta N_2)_{\text{in}}{}^2 &= 2\cosh^2 r \sinh^2 r\end{aligned} \tag{2.17}$$

[cf. Eqs. (2.6) and (2.11)]. To relate to the case where the light has a finite duration τ, one uses the mean numbers of photons in the two modes to define a mean power P into the laser port and a mean power P_2 into the normally unused input port:

$$\begin{aligned}P &= \hbar\omega\alpha^2/\tau\,,\\ P_2 &= \hbar\omega\sinh^2 r/\tau\,.\end{aligned} \tag{2.18}$$

For reasonable values of r, P_2 is an extremely small power ($\hbar\omega/\tau \sim 2\times 10^{-16}$ W).

It is now a simple matter, using Eqs. (2.4) and (2.8), to evaluate the expectation value and variance of $\mathcal{P}$:

$$\langle\mathcal{P}\rangle = 0\,, \tag{2.19a}$$

$$(\Delta\mathcal{P})^2 = (2b\hbar\omega/c)^2(\alpha^2 e^{2r} + \sinh^2 r)\,. \tag{2.19b}$$

Both terms in Eq. (2.19b) come from the interference of modes 1^+ and 2^+—the first term from the superposition of the coherent excitation of mode 1^+ on the fluctuations in mode 2^+ and the second from the interference of the fluctuations in the two modes. There is *no* contribution to Eq. (2.19b) from the superposition of the coherent excitation of mode 1^+ on fluctuations in mode 1^+; these input power fluctuations perturb only the *sum* of the end masses' momenta and, therefore, do not affect the interferometer's performance. Equation (2.19b) also displays the effect of putting mode 2^+ in a squeezed state. In the arms of the interferometer one quadrature phase of mode 2^+ is in phase with the coherent excitation of mode 1^+—but with opposite sign in the two arms. This phase is entirely responsible for the first term in Eq. (2.19b). By attenuating or amplifying the noise in this phase, one can reduce ($r<0$) or increase ($r>0$) $\Delta\mathcal{P}$.

In a time τ the disturbance (2.19b) of the difference between the end masses' momenta perturbs z by an amount

$$(\Delta z)_{\text{rp}} \simeq (\Delta\mathcal{P})\tau/2m = (b\hbar\omega\tau/mc)(\alpha^2 e^{2r} + \sinh^2 r)^{1/2}\,. \tag{2.20}$$

This is the *radiation-pressure* (rp) *error* in z. Below, $(\Delta z)_{\text{rp}}$ is used in an analysis of the optimum sensitivity and optimum power.

Before going on to a consideration of the photon-counting error, it is perhaps useful to look in a more general way at the intensity correlation of the light in the two arms of the interferometer. The beam-splitter modes of Eqs. (2.12) and (2.13) are ideally suited to a discussion of a characteristic intensity-correlation experiment in quantum optics—an experiment of the type pioneered by Hanbury Brown and Twiss.[24, 25] In such an experiment one counts the number of photons in the two beams emerging from a beam splitter and looks at the cross correlation of the number of counts. The quantity that characterizes this photon-number correlation is the *second-order coherence*,[12] which, for the simple case of two out modes considered here, is given by

$$g_{12}^{(2)} \equiv \frac{\langle b_1^\dagger b_1 b_2^\dagger b_2\rangle}{\langle b_1^\dagger b_1\rangle\langle b_2^\dagger b_2\rangle}\,. \tag{2.21}$$

If $g_{12}^{(2)} > 1$, the intensities in the two output beams are correlated—a phenomenon known as photon bunching. If $g_{12}^{(2)} < 1$, the intensities are anticorrelated—a situation referred to as antibunching. Antibunching is generally considered to be an intrinsically quantum-mechanical property of light; a great deal of effort has gone into trying to produce and detect antibunched light.[26]

To see how squeezed states fit into this picture, I consider a more general state of the field than $|\Psi\rangle$. This state,

$$|\Psi'\rangle = S_2(\zeta_2)D_1(\alpha)S_1(r_1)|0\rangle\,,$$
$$\alpha \text{ and } r_1 \text{ real}\,,\quad \zeta_2 = -r_2 e^{-2i\mu}\,, \qquad (2.22)$$

differs from $|\Psi\rangle$ in that mode 1^+ is in an excited *squeezed* state with complex amplitude α (S_1 is the squeeze operator for mode 1^+). The squeezing of mode 1^+ is of the sort depicted in Figs. 2(b) and 2(c).

Using Eqs. (2.14), (2.4), and (2.8), one can evaluate the second-order coherence of $|\Psi'\rangle$:

$$g_{12}^{(2)} = 1 + \frac{Q}{(\alpha^2 + \sinh^2 r_1 + \sinh^2 r_2)^2}\,, \qquad (2.23a)$$

$$Q = \alpha^2 e^{-2r_1} - \alpha^2 e^{2r_2} + 2\cosh^2 r_1 \sinh^2 r_1 + 2\cosh^2 r_2 \sinh^2 r_2 - (\sinh r_1 \cosh r_2 - \cosh r_1 \sinh r_2)^2\,. \qquad (2.23b)$$

The terms in Q have simple physical interpretations: the first and second terms come from interference of the coherent excitation of mode 1^+ with the fluctuations in modes 1^+ and 2^+; the third and fourth terms come from the fluctuations in the two modes separately; and the last term comes from interference of the fluctuations in the two modes. The terms arising from each in mode separately (terms 1, 3, and 4) always make a positive (correlated) contribution to Q, whereas the interference terms (terms 2 and 5) always make a negative (anticorrelated) contribution.

Only one quadrature phase of each in mode is in phase with the coherent excitation of mode 1^+. These in-phase quadratures are responsible for the first two terms in Q. Thus, by squeezing the in modes to put more or less noise in the in-phase quadratures, one can make the light bunched ($Q > 0$) or antibunched ($Q < 0$). Rewriting Q in a different form makes it clear that only the first two terms in Q can lead to antibunching:

$$Q = -2\alpha^2 e^{r_2 - r_1} \sinh(r_1 + r_2) + \sinh^2(r_1 + r_2)[1 + 2\sinh^2(r_1 - r_2)]\,. \qquad (2.24)$$

The pure-fluctuation terms in Q always make a net positive contribution.

It is useful now to look at some special cases of the above. The simplest is the case where mode 1^+ is in a coherent state ($r_1 = 0$) and mode 2^+ is in the vacuum state ($r_2 = 0$). Then one obtains the well-known result that $g_{12}^{(2)} = 1$. This lack of correlation results from the precise cancellation of the first two terms in Q [Eq. (2.23b)]; the correlated contribution due to fluctuations in the input coherent state is canceled by the anticorrelated contribution due to vacuum fluctuations entering the other input port.

The second case is the one where mode 1^+ is in an excited squeezed state and mode 2^+ is again vacuum ($r_2 = 0$). The intensity correlations of this type of light have been investigated by several authors.[14,19,27,28] For this case Q takes the form

$$Q = \alpha^2(e^{-2r_1} - 1) + \sinh^2 r_1(1 + 2\sinh^2 r_1)\,. \qquad (2.25)$$

The light is bunched if one increases the noise in the quadrature phase of mode 1^+ that carries the coherent excitation [$r_1 < 0$; Fig. 2(c)]. It can be antibunched if one decreases the noise in that quadrature phase [$r_1 > 0$; Fig. 2(b)]. Note, however, that as one increases the value of r_1, the pure-fluctuation terms in Eq. (2.25) eventually dominate and the light becomes bunched. If $|\alpha| \gg 1$ this transition from antibunching to bunching occurs when $r_1 \simeq \frac{1}{4}\ln(8\alpha^2)$.

The last case is the one where mode 1^+ is in a coherent state ($r_1 = 0$), but mode 2^+ is in a squeezed state. Then Q takes the form

$$Q = \alpha^2(1 - e^{2r_2}) + \sinh^2 r_2(1 + 2\sinh^2 r_2)\,. \qquad (2.26)$$

This case is similar to the previous one. The output light is bunched if one decreases the noise in the quadrature phase of mode 2^+ that gives rise to the anticorrelated contribution in Eq. (2.26) ($r_2 < 0$). The output light can be antibunched if one increases the noise in that quadrature phase [$r_2 > 0$; it is antibunched light of this sort that increases the radiation-pressure error (2.20) in an interferometer]. As in the previous case, the light eventually becomes bunched as one increases r_2, the transition occurring at $r_2 \simeq \frac{1}{2}\ln(8\alpha^2)$ if $|\alpha| \gg 1$. That this case can produce antibunched light is perhaps not very surprising, because if one allows light to enter both input ports of a beam splitter, anticorrelation can be obtained using classical light. On the other hand, this case should not be dismissed too quickly, because the squeezed state of mode 2^+, with its zero expected complex amplitude, is certainly not like a classical radiation field.

The antibunched light discussed in this section is different from the antibunched light that has been produced and detected in a recent series of experiments.[29,30] In these experiments the goal has been to produce light that is closer to a photon-number eigenstate than a coherent state is. Such light is antibunched by virtue of its reduced amplitude fluctuations, which are purchased at the price of an ill-defined phase. For the case of a squeezed mode 1^+ ($r_2 = 0$, $r_1 > 0$), the light considered here is also antibunched because of decreased (correlated) amplitude fluctuations, and the phase is correspondingly less well defined than for a coherent state. For the case of a squeezed mode 2^+ ($r_1=0$, $r_2>0$), however, the antibunching is due to increased (anticorrelated) amplitude fluctuations, and the phases of the two output beams are better defined than for a coherent state. The difference between the light used in the current experiments and the light considered here could be observed using phase-sensitive detection of the two beams from the beam splitter.

3. Photon-counting error

For the case of ideal photodetectors, the photocount statistics of the photodetectors are the same as the photon statistics, so the photon-counting error can be determined simply by calculating the fluctuations in number of output photons. Just as for the case of radiation-pressure fluctuations, this calculation requires only four modes of the electromagnetic field, but now they must be modes for the entire interferometer ("interferometer modes").

The in modes of interest are again called modes 1^+ and 2^+, with mode 1^+ describing light *incident* from the input (laser) port and mode 2^+ describing light incident from the other input port. Outside the arms of the interferometer the electric fields of the two in modes have the forms

$$E_1^+=\begin{cases}\mathfrak{A}e^{i(kx-\omega t)}-i\mathfrak{A}e^{i(\Phi-\mu)}\sin(\phi/2)e^{-i(kx+\omega t)}, & y>x\\ \mathfrak{A}e^{i\Phi}\cos(\phi/2)e^{-i(ky+\omega t)}, & y<x\end{cases} \tag{2.27}$$

$$E_2^+=\begin{cases}\mathfrak{A}e^{i\Phi}\cos(\phi/2)e^{-i(kx+\omega t)}, & y>x\\ . & \\ \mathfrak{A}e^{i(ky-\omega t)}-i\mathfrak{A}e^{i(\Phi+\mu)}\sin(\phi/2)e^{-i(ky+\omega t)}, & y<x\end{cases}$$

where $\mathfrak{A}$ is a (real) normalization constant. The phases ϕ and Φ can be defined precisely as follows: let light enter the interferometer from the laser port, and consider the output light in the bottom output port of Fig. 3 ($-y$ direction); then ϕ is the phase difference between the light from the two arms, and Φ is the mean phase. The two phases are related to the (constant) positions of the end mirrors by

$$\phi=2b\omega z/c+\pi-2\mu\ , \tag{2.28a}$$

$$\Phi=2b\omega Z/c+\Phi_0\ , \tag{2.28b}$$

where $Z\equiv\frac{1}{2}(z_1+z_2)$ and Φ_0 is a constant.

The two out modes, again called modes 1^- and 2^-, are the time reverses of the in modes. Mode 1^- is the time reverse of mode 1^+ and, therefore, describes light leaving the interferometer along the $-x$ axis. Mode 2^-, the time reverse of mode 2^+, describes light exiting along the $-y$ axis. The fields of the two sets of modes are related by

$$E_1^- = e^{-i\Phi}[\,ie^{i\mu}E_1^+\sin(\phi/2)+E_2^+\cos(\phi/2)]\ ,$$
$$E_2^-=e^{-i\Phi}[\,E_1^+\cos(\phi/2)+ie^{-i\mu}E_2^+\sin(\phi/2)]\ . \tag{2.29}$$

The creation and annihilation operators for the interferometer in modes are denoted $a_1^\dagger$, a_1 and $a_2^\dagger$, a_2; similarly, for the interferometer out modes, $c_1^\dagger$, c_1 and $c_2^\dagger$, c_2. Equation (2.29) implies

$$c_1=e^{i\Phi}[\,-ie^{-i\mu}a_1\sin(\phi/2)+a_2\cos(\phi/2)]\ ,$$
$$c_2=e^{i\Phi}[\,a_1\cos(\phi/2)-ie^{i\mu}a_2\sin(\phi/2)]\ . \tag{2.30}$$

The operators of interest are the photon-number operators for the two out modes and the difference between these two. These are given by

$$c_2^\dagger c_2-c_1^\dagger c_1=(a_1^\dagger a_1-a_2^\dagger a_2)\cos\phi$$
$$-i\sin\phi(e^{i\mu}a_1^\dagger a_2-e^{-i\mu}a_2^\dagger a_1), \tag{2.31}$$

$$c_1^\dagger c_1=a_1^\dagger a_1\sin^2(\phi/2)+a_2^\dagger a_2\cos^2(\phi/2)$$
$$+i\sin(\phi/2)\cos(\phi/2)(e^{i\mu}a_1^\dagger a_2-e^{-i\mu}a_2^\dagger a_1)\ , \tag{2.32}$$

where $c_2^\dagger c_2$ can be obtained from $c_1^\dagger c_1$ by the transformation $\phi-\phi+\pi$.

Now assume, as before, that the electromagnetic field is in the state $|\Psi\rangle$ of Eq. (2.16). For this state the expectation values and variances of the operators (2.31) and (2.32), evaluated using Eqs. (2.4) and (2.8), are

$$n_{\rm out}\equiv\langle c_2^\dagger c_2-c_1^\dagger c_1\rangle=\cos\phi(\alpha^2-\sinh^2 r), \tag{2.33a}$$

$$\Delta n_{\rm out}{}^2=\alpha^2\cos^2\phi+2\cos^2\phi\cosh^2 r\sinh^2 r$$
$$+\sin^2\phi(\alpha^2e^{-2r}+\sinh^2 r)\ , \tag{2.33b}$$

$$(N_1)_{\rm out}\equiv\langle c_1^\dagger c_1\rangle=\alpha^2\sin^2(\phi/2)+\cos^2(\phi/2)\sinh^2 r\ , \tag{2.34a}$$

$$(\Delta N_1)_{\rm out}{}^2=\alpha^2\sin^4(\phi/2)+2\cos^4(\phi/2)\cosh^2 r\sinh^2 r$$
$$+\sin^2(\phi/2)\cos^2(\phi/2)(\alpha^2e^{-2r}+\sinh^2 r)\ . \tag{2.34b}$$

Equations (2.34) characterize the output (signal and noise) of an ideal photodetector in one of the output ports (mode 1^-), and Eqs. (2.33) characterize the differenced output of two ideal photodetectors, one in each output port.

Changes in z are detected by looking at changes in n_{out} or $(N_1)_{\text{out}}$. Taking the case of differenced photodetectors, one finds that a change δz produces a change

$$\delta n_{\text{out}} = -(2b\omega/c)\alpha^2 \sin\phi \, \delta z \tag{2.35}$$

in n_{out}, where I assume $|\alpha| \gg |\sinh r|$ in order to neglect the second term in Eq. (2.33a), and where Eq. (2.28a) is used to convert ϕ to z. Using Eq. (2.35) to transform Δn_{out} into a corresponding error in z, one obtains a *photon-counting* (pc) *error:*

$$(\Delta z)_{\text{pc}} \simeq \frac{c}{2b\omega}\left(\frac{\cot^2\phi}{\alpha^2} + \frac{2\cot^2\phi \cosh^2 r \sinh^2 r}{\alpha^4} + \frac{e^{-2r}}{\alpha^2} + \frac{\sinh^2 r}{\alpha^4}\right)^{1/2} . \tag{2.36}$$

For the case of a single photodetector [Eqs. (2.34)], a similar calculation yields the following *photon-counting error:*

$$(\Delta z)_{\text{pc}} \simeq \frac{c}{2b\omega}\left[\frac{\tan^2(\phi/2)}{\alpha^2} + \frac{2\cot^2(\phi/2)\cosh^2 r \sinh^2 r}{\alpha^4} + \frac{e^{-2r}}{\alpha^2} + \frac{\sinh^2 r}{\alpha^4}\right]^{1/2} , \tag{2.37}$$

where it is necessary to assume $|\alpha| \gg |\cot(\phi/2)\sinh r|$ in order to neglect the second term in Eq. (2.34a).

The terms in Eqs. (2.36) and (2.37) [or, alternatively, in Eqs. (2.33b) and (2.34b)] can be interpreted in ways familiar from the earlier discussions of $\Delta\mathcal{P}$ and Q. The first term in both equations comes from fluctuations in mode 1^+ superposed on the coherent excitation of mode 1^+ (input power fluctuations), the second term comes from fluctuations in mode 2^+, and the last two terms come from the interference of modes 1^+ and 2^+. For the case of differenced photodetectors [Eq. (2.36)], both of the first two terms can be made zero by operating at an appropriate place in the fringe pattern ($\cos\phi = 0$); hence, input power fluctuations can be made irrelevant. For a single photodetector [Eq. (2.37)], the contribution from input power fluctuations can be made negligible by operating near a null fringe [$\sin(\phi/2) \simeq 0$], and as long as $|\alpha| \gg |\cot(\phi/2)e^r \cosh r \sinh r|$, the second term in Eq. (2.37) is also negligible compared to the interference terms. Thus, in either case, *one can arrange that the dominant contribution to the photon-counting error comes from the interference of modes 1^+ and 2^+; since the interference terms in Eqs. (2.36) and (2.37) are independent of position in the fringe pattern (i.e., independent of ϕ), they make the truly* ***unavoidable*** *contribution to the photon-counting error.*

The third term in Eqs. (2.36) and (2.37) displays a result of a now-familiar sort: only one quadrature phase of mode 2^+ superposes on the coherent excitation of mode 1^+ to contribute to the photon-counting error. This quadrature phase is not the one that makes the same sort of contribution to the radiation-pressure error [cf. third term in Eqs. (2.36) and (2.37) and first term in Eq. (2.20)]. Consequently, squeezing mode 2^+ can reduce the photon-counting error while increasing the radiation-pressure error ($r > 0$), or vice versa ($r < 0$).

4. *Optimum sensitivity and optimum power*

The objective now is to investigate what happens to the interferometer's optimum sensitivity and optimum power as one squeezes mode 2^+. For the case of differenced photodetectors operating at $\cos\phi = 0$, the photon-counting error (2.36) and the radiation-pressure error (2.20) have the forms

$$(\Delta z)_{\text{pc}} \simeq (c/2b\omega)|\alpha|^{-1}e^{-r} , \tag{2.38}$$

$$(\Delta z)_{\text{rp}} \simeq (b\hbar\omega\tau/mc)|\alpha|e^{r} . \tag{2.39}$$

Here I assume that α^2 is large enough so that the last term in both Eq. (2.20) and Eq. (2.36) can be neglected—i.e., I assume that $|\alpha| \gg \sinh^2 r$, which is equivalent to $P_2 \ll (P\hbar\omega/\tau)^{1/2}$ [see Eq. (2.18)]. The total error is $\Delta z = [(\Delta z)_{\text{pc}}{}^2 + (\Delta z)_{\text{rp}}{}^2]^{1/2}$. If one minimizes the total error with respect to α^2, one finds a minimum error

$$(\Delta z)_{\text{opt}} \simeq (\hbar\tau/m)^{1/2} \simeq (\Delta z)_{\text{SQL}} \tag{2.40}$$

[see Eq. (1.1)] and an optimum value for α^2,

$$\alpha_{\text{opt}}{}^2 \simeq \alpha_0{}^2 e^{-2r} , \quad \alpha_0{}^2 \equiv \tfrac{1}{2}(mc^2/\hbar\omega)(1/\omega\tau)(1/b^2) . \tag{2.41}$$

The quantity $\alpha_{\text{opt}}{}^2$ is the optimum number of photons in mode 1^+ [see Eq. (2.17)], and it translates into an optimum input power

$$P_{\text{opt}} \simeq P_0 e^{-2r} , \tag{2.42}$$

where $P_0 \equiv \hbar\omega\alpha_0{}^2/\tau$ is the optimum power for a standard ($r = 0$) interferometer [Eq. (1.2)]. For the fiducial parameters, $\alpha_0{}^2 \sim 4 \times 10^{19}$ and $P_0 \sim 8 \times 10^3$ W.

Equation (2.42) displays the desired result: the optimum power can be adjusted by squeezing the vacuum before it can enter the normally unused input port. There are, however, limits to the validity of Eqs. (2.41) and (2.42)—limits imposed by the validity condition, $|\alpha| \gg \sinh^2 r$, for Eqs. (2.38) and (2.39). What happens, for example, if the squeeze factor r becomes so large that α_{opt} violates this condition? The answer is contained in the exact equation for the photon-counting error [Eq. (2.36)]. As the squeeze factor is increased past a value $r_{\max} \sim \frac{1}{3}\ln\alpha_0$, the last term in Eq. (2.36) begins to dominate the optimum sensitivity. For $r \gtrsim r_{\max}$ the optimum sensitivity becomes rapidly worse than the standard quantum limit. As a result, the maximum useful value of r for this case of differenced photodetectors is $r_{\max}$, which corresponds to a minimum optimum power $(P_{\rm opt})_{\min} \sim (P_0{}^2\hbar\omega/\tau)^{1/3}$ $[(\alpha_{\rm opt}{}^2)_{\min} \sim \alpha_0{}^{4/3}]$ and a power into the other port $P_2 \sim [P_0(\hbar\omega/\tau)^2]^{1/3}$ [Eq. (2.18)].

For the case of a single photodetector [Eq. (2.37)] operated near a null fringe [$\sin(\phi/2) \simeq 0$], one obtains the same results (2.38)–(2.42), but with more stringent validity conditions, $|\alpha| \gg \sinh^2 r$ and $|\alpha| \gg e^{2r}\sinh^2 r$, because of the second term in Eq. (2.37). These more stringent validity conditions mean that the optimum power cannot be reduced as much as in the case of differenced photodetectors (see, however, the discussion at the end of Sec. III).

One interesting question not investigated in the above analysis is whether there is some reason to squeeze the light in mode 1^+. Doing so can, for example, reduce the size of the first term in Eqs. (2.36) and (2.37), but there is little point in doing this, because this term can be made negligible by operating at an appropriate place in the fringe pattern. There is, however, another reason for squeezing the input laser light. Recall that, for the case of differenced photodetectors, the term that limits the reduction of the optimum power is the last term in the photon-counting error (2.36)—a term that arises from the interference of fluctuations in modes 1^+ and 2^+. For $r \gg 1$ the noise in mode 2^+ is concentrated in one quadrature phase, so one can reduce the size of this term by squeezing mode 1^+ with appropriate phase. In doing so, however, one inevitably increases the term of the same type in the radiation-pressure error [second term in Eq. (2.20)]. In particular, if one puts the field in the state $|\Psi'\rangle$ of Eq. (2.22) with $r_2 = -r_1 = r$ (squeezed light into both input ports), one finds that the last term in Eq. (2.36) disappears, but the last term in Eq. (2.20) becomes $\sinh^2 2r$ rather than $\sinh^2 r$. The result is a mimimum optimum power

$$(P_{\rm opt})_{\min} \sim (P_0\hbar\omega/\tau)^{1/2} \sim P_2 \qquad (2.43)$$

$[r_{\max} \sim \frac{1}{2}\ln\alpha_0,\ (\alpha_{\rm opt}{}^2)_{\min} \sim \alpha_0]$. Equations (2.18) and (2.42) imply that, for $r \gg 1$, $P_{\rm opt}\, P_2 \sim P_0\hbar\omega/\tau$, so Eq. (2.43) is the best one can do in reducing the total power $P_{\rm opt} + P_2$ required to run an interferometer at the standard quantum limit.

Squeezing the input laser light finds another application in an interferometer that measures $Z = \frac{1}{2}(z_1 + z_2)$ rather than z. The intensity in an interferometer's output ports is determined by z, but the phase of the output is determined by Z. A standard interferometer, in which one monitors the output intensity, is sensitive only to changes in z, but an interferometer in which one monitored the output light using phase-sensitive detection would be sensitive to changes in Z. For the case of a coherent state in mode 1^+ and vacuum in mode 2^+, a Z interferometer would have an optimum sensitivity $(\Delta Z)_{\rm opt} \simeq (\hbar\tau/m)^{1/2}$ and an optimum power P_0. By squeezing the light in mode 1^+, one could adjust the optimum power just as in Eq. (2.42).

III. PRACTICAL CONSIDERATIONS RELATED TO SQUEEZED-STATE TECHNIQUE

In this section I turn from the abstract analysis of Sec. II to somewhat more practical matters related to implementing the squeezed-state technique. In particular, I focus on the situation relevant for real interferometers, which are limited by photon-counting statistics. The strain sensitivity of such an interferometer is given by

$$\frac{(\Delta z)_{\rm pc}}{l} \simeq \frac{c}{2b\omega l}\,\frac{e^{-r}}{|\alpha|} = \frac{c}{2b\omega l}\left(\frac{\hbar\omega}{P\tau}\right)^{1/2} e^{-r}$$

$$= \frac{1}{\omega\tau_s}\left(\frac{\hbar\omega}{P\tau}\right)^{1/2} e^{-r} \qquad (3.1)$$

[Eqs. (2.38) and (2.18)]. For a given measurement time τ, one can improve the strain sensitivity by increasing b, ω, l, P, or r. Thus, the squeezed-state technique provides an additional option for improving the strain sensitivity. The availability of this option might be important, because changes in the interferometer's other parameters might be precluded by practical limitations—e.g., unavailability of cw lasers of higher power or higher frequency, unavailability of optical components to handle higher powers or higher frequencies, or limitations on the size of high-quality mirrors.

There is one situation in which the squeezed-state technique might be especially important. The improvement in strain sensitivity afforded by increasing the number of bounces b or the baseline l is limited by the condition that the storage time $\tau_s = 2bl/c$ be less than the measurement time τ. For $\tau_s > \tau$ the sensitivity does not improve beyond the value attained at $\tau_s \simeq \tau$; this limits the strain sensitivity of a standard ($r = 0$) interferometer to $(\Delta z)_{\rm pc}/l \simeq (\hbar/P\omega\tau^3)^{1/2}$. The greatest potential usefulness of the squeezed-state technique probably lies in its ability to improve the sensitivity of an interferometer for which $\tau_s \simeq \tau$, without increasing the input power P.

In considering the design of a squeezed-state interferometer, the first question to be addressed is how to generate the required squeezed states. One way of generating squeezed states is to use a degenerate parametric amplifier. An *optical parametric amplifier*[31] is an optical component in which one pumps a nonlinear medium, which has a nonvanishing second-order nonlinear susceptibility, with an electromagnetic wave whose angular frequency is denoted ω_p. The nonlinearity of the medium couples this pump wave to two other wave modes, called the signal and the idler, whose frequencies ω_s and ω_i satisfy $\omega_s+\omega_i=\omega_p$. If the wave vectors of the three waves in the medium satisfy, or nearly satisfy, $\bar{k}_s+\bar{k}_i=\bar{k}_p$ (phase-matching condition), then the signal and idler are amplified (neglecting losses) as they propagate through the medium. A *degenerate parametric amplifier* is a parametric amplifier for which the signal and idler coincide ($\omega_s=\omega_i=\frac{1}{2}\omega_p$, $\bar{k}_s=\bar{k}_i=\frac{1}{2}\bar{k}_p$).

A degenerate parametric amplifier is a phase-sensitive device: it amplifies one quadrature phase of the signal mode, and it attenuates the other. Takahasi[32] was the first to point out that this behavior applies to the quantum-mechanical fluctuations in the mode. He considered a simple model of a degenerate parametric amplifier, a harmonic oscillator (the signal mode) whose spring constant is modulated classically (the pump modulation) at twice the oscillator's frequency, and he showed that an initial coherent state for the oscillator is transformed into a state whose uncertainties in the two quadrature phases are unequal. Since Takahasi's work, there have been several quantum-mechanical analyses[14, 27, 33-35] of the light generated by an optical degenerate parametric amplifier. These analyses vary in complexity, some including the effects of losses in the nonlinear medium[33, 35] and the effects of jitter in the amplitude and phase of the pump.[34]

The basic conclusion to be drawn from these analyses is that an ideal degenerate parametric amplifier generates squeezed states.[14, 27] Specifically, the state of the signal mode at the output of a degenerate parametric amplifier is obtained by applying the squeeze operator (2.7) to the state at the input. The phase of the squeezing [θ in Eq. (2.7)] is determined by the phase of the pump, and the squeeze factor r can be read off the results of a classical analysis[31]:

$$r=\left(\frac{4\pi\omega_s L}{c n_s}\right)d\,|E_p| = \left(\frac{4\pi\omega_s L}{c n_s}\right)d\left(\frac{8\pi P_p}{c n_p A}\right)^{1/2} \quad \text{(cgs units).} \tag{3.2}$$

Here I assume there is perfect phase matching at degeneracy, and I use the notation of Ref. 31: d is the effective nonlinear susceptibility, E_p is the amplitude of the pump wave's electric field, L is the length over which the interaction takes place in the nonlinear medium, and n_s and n_p are the values of the index of refraction at the signal and pump frequencies. In the second part of Eq. (3.2) I have converted the pump electric field into a pump power P_p distributed over an area A.

The best nonlinear optical media have $d/n^{3/2} \sim 10^{-8}$ in cgs units.[31] Thus, for a pump power $P_p \sim 100$ mW $=10^6$ erg sec^{-1}, a beam area $A\sim 3\times10^{-2}$ cm^2, an interaction length $L\sim 10$ cm, and a signal frequency $\omega_s\sim 4\times10^{15}$ rad sec^{-1}, the squeeze factor is $r\sim 0.03$. This does not look very encouraging, and one must remember that the estimate is overly optimistic because losses in the medium have been neglected. There is, however, a way to increase the achievable squeeze factor. If the single-pass loss through the medium does not exceed the single-pass gain, one can increase the squeeze factor (gain) by enclosing the medium in an optical cavity that resonates at the signal frequency. This increases the effective interaction length, because the signal wave passes through the nonlinear medium many times before it leaves the cavity. The resulting device is called an *optical parametric oscillator*.[31] For pump powers of 10–100 mW, parametric oscillators at optical frequencies have achieved a signal-mode output of a few milliwatts in a bandwidth of about an angstrom[31]; this corresponds to a squeeze factor $e^r \sim 200$ ($r\sim 5$). The bandwidth here is huge compared to that necessary in a gravitational-wave interferometer; nonetheless, these results hint at the possibility of achieving reasonable squeeze factors.

Even given a reasonable squeeze factor, there are still stringent demands on the operation of the degenerate parametric amplifier in an interferometer. The critical demands are that the amplifier be pumped at exactly twice the frequency of the laser ($\omega_p=2\omega$) and that the pumping be done with just the right phase so that the vacuum is squeezed with a phase that is properly matched to the phase of the laser [see Eq. (2.16)]. To satisfy these demands in practice, one would probably extract a small fraction of the laser light at a beam splitter, run this light through a frequency doubler, and then use the doubled light (with just the right phase) to pump a degenerate parametric amplifier located in the normally unused input port. This scheme is shown in Fig. 4. If the nonlinear medium is pumped at exactly twice the laser frequency, it is *not* necessary that the amplifier operate precisely at degeneracy. If it operates a small distance from degeneracy, then it produces two output waves, a signal and an idler, whose frequencies satisfy $\omega_s + \omega_i = \omega_p = 2\omega$; the fluctuations in these two waves are correlated in just such a way that their superposition mimics the behavior of a squeezed state at frequency ω.

There are a host of practical problems to be faced in implementing the squeezed-state technique using a degenerate parametric amplifier. Nonetheless, the brief discussion given here is perhaps sufficiently encouraging to motivate further investigation of the idea.

Degenerate parametric amplification is not the only optical process that generates squeezed states. Any phase-sensitive nonlinear process is a good candidate. For example, Yuen and Shapiro[36] have pointed out that a degenerate four-wave mixer generates output waves that can produce squeezed states when they are combined at a beam splitter. Four-wave mixers are now being vigorously developed because of their ability to produce phase-conjugated (time-reversed) light.[37] Yuen[18,19] has also suggested that an ideal two-photon laser would produce squeezed states.

The losses in real mirrors are likely to impose the most severe practical limitation on the usefulness of the squeezed-state technique. Losses destroy the crucial feature of the technique—the reduced noise in one of the two quadrature phases of mode 2^+. One can estimate the effect of losses from the following argument. Consider a nearly monochromatic beam of light bouncing back and forth between mirrors of reflectivity $\mathcal{R}$. If one identifies $\mathcal{R}$ as the probability that a given photon is reflected, then a simple "random-walk" argument yields the mean and variance of the number of photons in the beam after q reflections:

$$N_q = N_0 \mathcal{R}^q \; , \tag{3.3a}$$

$$(\Delta N)_q{}^2 = (\Delta N)_0{}^2 \mathcal{R}^{2q} + N_0 \mathcal{R}^q (1 - \mathcal{R}^q) \; . \tag{3.3b}$$

Here the subscript 0 designates the initial values. Since reflection is a linear process, Eqs. (3.3) suggest that, after q reflections, the X_1 and X_2 of the light beam have the following variances:

$$\begin{aligned}(\Delta X_1)_q{}^2 &= (\Delta X_1)_0{}^2 \mathcal{R}^q + \tfrac{1}{4}(1 - \mathcal{R}^q) \; , \\ (\Delta X_2)_q{}^2 &= (\Delta X_2)_0{}^2 \mathcal{R}^q + \tfrac{1}{4}(1 - \mathcal{R}^q) \; .\end{aligned} \tag{3.4}$$

Equations (3.3) and (3.4) have the same form as the equations for a damped harmonic oscillator in contact with a heat reservoir at zero temperature. The first term in Eqs. (3.3b) and (3.4) represents the damping of the initial fluctuations, and the second term represents the fluctuations added

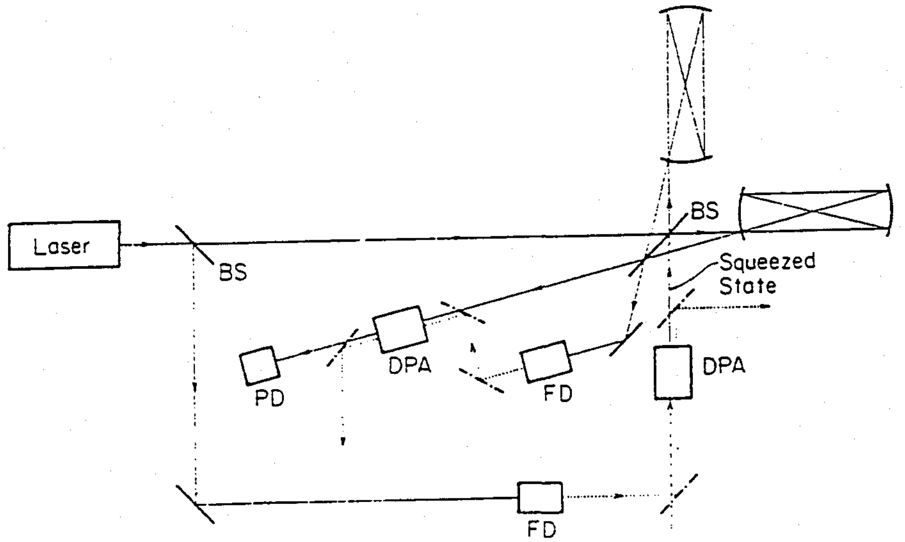


FIG. 4. Squeezed-state interferometer (abbreviations: BS = beam splitter; FD = frequency doubler; DPA = degenerate parametric amplifier; PD = photodetector). The crucial feature of the squeezed-state technique is the DPA located in the normally unused input port. This DPA takes the vacuum fluctuations incident on it (dashed arrow) and produces a squeezed state. To pump the DPA, one uses light that is extracted from the laser beam at a beam splitter and then doubled in frequency. There is another DPA in one of the output ports. This output DPA squeezes the light in that por which is near a null in the fringe pattern, and thereby matches the noise in the light to the shot noise in an inefficient PD. The output DPA is pumped by frequency-doubled light from the other output port. The laser operates at frequenc ω. Light beams at frequency ω are drawn with thin lines, and the components for handling them are drawn with heavy lines. The pump beams at frequency 2ω are drawn with dotted lines, and the mirrors for routing them are drawn with heavy, broken lines. These mirrors are assumed to transmit at frequency ω.

due to losses. The added fluctuations appear with random phase. Thus, an initial coherent state $[(\Delta X_1)_0 = (\Delta X_2)_0 = \frac{1}{2},\ (\Delta N)_0{}^2 = N_0]$ remains coherent as its mean amplitude damps away. An initial squeezed state, however, loses its squeezed nature; the losses randomize the phase of its fluctuations, and its initial error ellipse becomes round.

In a squeezed-state interferometer, the added random-phase noise must be small enough so that it does not greatly increase the fluctuations in the low-noise quadrature phase of mode 2^+. This requirement suggests that the number of bounces b must satisfy

$$b \leq b_0 e^{-2r} \equiv e^{-2r}/(1-\mathcal{R}) \tag{3.5}$$

[Eqs. (2.11), (2.16), and (3.4)], where I use the fact that the total number of reflections in each arm of the interferometer is $q = 2b - 1$, and where $b_0 \equiv (1-\mathcal{R})^{-1}$ is the optimum number of bounces in a standard $(r=0)$ interferometer.[6]

Equation (3.5) is a severe restriction. It implies that the squeezed-state technique can be used only if the interferometer is not limited by mirror losses (i.e., only if $b < b_0$). At the beginning of this section it was remarked that the most likely application of the squeezed-state technique is to interferometers whose number of bounces and baseline are large enough so that $\tau_s \simeq \tau$. The mirror-loss restriction (3.5) means that the technique can be used in this case only if $2b_0 l/c > \tau$—i.e., only if b_0 bounces correspond to a storage time longer than the desired measurement time (for $l \sim 1$ km and $\mathcal{R} \simeq 0.999$, $2b_0 l/c \sim 7\times10^{-3}$ sec). Thus, the potential usefulness of the squeezed-state technique is restricted to the case of a long baseline, high-reflectivity mirrors, and a short measurement time.

It is perhaps useful to emphasize the situation in which the squeezed-state technique is likely to become useful. Consider an interferometer operating with b_0 bounces and a measurement time τ. The interferometer's strain sensitivity can be improved by increasing its baseline, but this improvement continues only until $l \simeq c\tau/2b_0$, at which point the strain sensitivity is approximately $(\hbar/P\omega\tau^3)^{1/2}$. A further increase in length does not improve the strain sensitivity—unless one applies the squeezed-state technique as one decreases the number of bounces. Use of the squeezed-state technique allows the strain sensitivity to improve as $(\Delta z)_{pc}/l \simeq (\hbar/P\omega\tau^3)^{1/2}(c\tau/2b_0 l)^{1/2}$ for $l \gtrsim c\tau/2b_0$ [Eqs. (3.1) and (3.5)].

The analysis in Sec. II assumed ideal photodetectors—a case for which the photocount statistics at the output of the photodetectors coincide with the photon statistics of the light incident on the photodetectors. If a photodetector has a quantum efficiency ξ less than one, then the photocount statistics are a combination of the photon statistics and the shot noise in the photodetector. How does the shot noise change the previously obtained photon-counting error? For the case of a single photodetector in one of the output ports [Eq. (2.34)], the mean and variance of the number of photons counted by the photodetector are given by[38]

$$N_{pd} = \xi(N_1)_{out}\ , \tag{3.6a}$$

$$\Delta N_{pd}{}^2 = \xi^2(\Delta N_1)_{out}{}^2 + \xi(1-\xi)(N_1)_{out}\ , \tag{3.6b}$$

where $(N_1)_{out}$ and $(\Delta N_1)_{out}$ are the mean and variance due to the photon statistics alone [Eqs. (2.34)]. The second term in (3.6b) is the shot-noise contribution. Using the same procedure as in Sec. III B 3, one can convert ΔN_{pd} into a photon-counting error,

$$(\Delta z)_{pc} \simeq \frac{c}{2b\omega}\left[\frac{\tan^2(\phi/2)}{\alpha^2} + \frac{e^{-2r}}{\alpha^2} + \frac{1-\xi}{\alpha^2 \xi \cos^2(\phi/2)}\right]^{1/2} . \tag{3.7}$$

Here I retain only the terms that dominate when $|\alpha|$ is sufficiently large. Near a null the first term in Eq. (3.7) is negligible, but the shot-noise contribution completely swamps the remaining photon-statistics term unless $1-\xi \leq \xi e^{-2r}$. If the squeezed-state technique is to significantly improve the sensitivity, this requirement demands extraordinarily efficient photodetectors.

There is an alternative approach that avoids the requirement for high-efficiency photodetectors. The light emerging from the interferometer has an excellent signal-to-noise ratio near a null by virtue of its low noise in the quadrature phase that carries the signal. The photodetector ruins this good signal-to-noise ratio, because its shot noise is much larger than the noise in the light. To overcome this difficulty, one would like to amplify both the light signal and the noise in phase with the signal while keeping their ratio constant, thereby matching the noise in the light to the shot noise. This is precisely what would happen if one squeezed the output light before it reached the photodetector. The squeezing could be done by a degenerate parametric amplifier located in the appropriate output port (see Fig. 4).

Formally, the squeezing is described by introducing new creation and annihilation operators $\tilde{c}_1^\dagger, \tilde{c}_1$. These operators characterize the light emerging from the degenerate parametric amplifier, and they are related to the operators for interferometer mode 1^- by

$$\bar{c}_1 \equiv S_{c_1}^\dagger(\bar{\zeta})c_1S_{c_1}(\bar{\zeta})$$
$$= c_1\cosh r - c_1^\dagger e^{2i(\Phi-\mu)}\sinh r, \quad \bar{\zeta} = re^{2i(\Phi-\mu)}. \tag{3.8}$$

Here S_{c_1} is the squeeze operator for mode 1^-, and Eq. (2.8) is used to transform c_1. Note that the squeeze factor here is the same as the one used previously, and that the phase of the squeezing is carefully matched to the phase of the output light in mode 1^- [see Eqs. (2.28b) and (2.30)].

The mean and variance of the number of photons emerging from the degenerate parametric amplifier is obtained by evaluating the expectation value and variance of $\bar{c}_1^\dagger\bar{c}_1$ in the state $|\Psi\rangle$ of Eq. (2.16). Using Eqs. (3.8), (2.30), (2.4), and (2.8), one finds that

$$(\bar{N}_1)_{\text{out}} \equiv \langle \bar{c}_1^\dagger \bar{c}_1 \rangle = \sin^2(\phi/2)(\alpha^2e^{2r} + \sinh^2 r) , \tag{3.9a}$$

$$(\Delta\bar{N}_1)_{\text{out}}{}^2 = \sin^4(\phi/2)(\alpha^2e^{4r} + 2\cosh^2 r\sinh^2 r)$$
$$+ \sin^2(\phi/2)\cos^2(\phi/2)(\alpha^2e^{2r} + \sinh^2 r) . \tag{3.9b}$$

These equations now replace Eqs. (2.34); they characterize the light incident on the photodetector. The disadvantage of this approach is revealed by Eq. (3.9a): unless one operates very close to a null fringe [$|\sin(\phi/2)| \ll e^{-r}$], the power out of the degenerate parametric amplifier becomes comparable to or larger than the input power P—a situation clearly inconsistent with the operation of the parametric amplifier and with the desire to reduce the total power requirements. It is worth noting as a matter of principle that Eq. (3.9b), unlike Eq. (2.34b), imposes no restriction on the reduction of the optimum power. Thus, by using a degenerate parametric amplifier at the output, one can in principle achieve with a single photodetector the minimum total power of Eq. (2.43).

One can now obtain the mean and variance of the number of photons counted by the photodetector by applying Eqs. (3.6) to Eqs. (3.9). These results are then converted in the usual way into a photon-counting error in z:

$$(\Delta z)_{pc} \simeq \frac{c}{2b\omega}\left\{\frac{\tan^2(\phi/2)}{\alpha^2} + \frac{e^{-2r}}{\alpha^2}\left[1 + \frac{1-\xi}{\xi\cos^2(\phi/2)}\right]\right\}^{1/2}$$
$$\simeq (c/2b\omega)\,\xi^{-1/2}\,|\alpha|^{-1}e^{-r} \quad \text{for } \sin(\phi/2) \simeq 0. \tag{3.10}$$

Here I again retain only the terms that dominate when $|\alpha|$ is large. Equation (3.10) explicitly demonstrates that insertion of a degenerate parametric amplifier into the output port allows one to use an inefficient photodetector without a significant increase in the photon-counting error [cf. Eqs. (3.10) and (2.38)].

To make this approach work, one must find a way to pump the degenerate parametric amplifier. The pump must have just the right phase relative to the phase of the output light in interferometer mode 1^-, to ensure that the squeezing of the output light occurs with the right phase. There is only one available beam of light that carries the necessary phase information, and that is the light in the other output port (interferometer mode 2^-). Since one wants to work very close to a null, there is plenty of power available in the other output port. One would take the light in the other output port, double its frequency, and then use the doubled light to pump the degenerate parametric amplifier in the output port. This approach is sketched in Fig. 4.

IV. CONCLUSION

The squeezed-state technique outlined in this paper will not be easy to implement. A refuge from criticism that the technique is difficult can be found by retreating behind the position that the entire task of detecting gravitational radiation is exceedingly difficult. Difficult or not, the squeezed-state technique might turn out at some stage to be the only way to improve the sensitivity of interferometers designed to detect gravitational waves. As interferometers are made longer, their strain sensitivity will eventually be limited by the photon-counting error for the case of a storage time approximately equal to the desired measurement time. Further improvements in sensitivity would then await an increase in laser power or implementation of the squeezed-state technique. Experimenters might then be forced to learn how to very gently squeeze the vacuum before it can contaminate the light in their interferometers.

ACKNOWLEDGMENTS

I thank R.W.P. Drever for a series of conversations that clarified the details of the squeezed-state technique, and I thank K. S. Thorne for making available his notes on the properties of the squeeze operator. This work was supported in part by the National Aeronautics and Space Administration [NGR 05-002-256 and a grant from PACE] and by the National Science Foundation [AST79-22012].

[1] For reviews of efforts to detect gravitational waves and of theoretical estimates of gravitational-wave strengths, see K. S. Thorne, Rev. Mod. Phys. 52, 285 (1980) and references cited therein; see also the chapters by R. Weiss and by R. Epstein and J. P. A. Clark, in *Sources of Gravitational Radiation*, edited by L. Smarr (Cambridge University Press, Cambridge, 1979).

[2] C. M. Caves, K. S. Thorne, R. W. P. Drever, V. D. Sandberg, and M. Zimmermann, Rev. Mod. Phys. 52, 341 (1980).

[3] R. L. Forward, Phys. Rev. D 17, 379 (1978).

[4] H. Billing, K. Maischberger, A. Rüdiger, R. Schilling, L. Schnupp, and W. Winkler, J. Phys. E 12, 1043 (1979).

[5] R. W. P. Drever, J. Hough, W. A. Edelstein, J. R. Pugh, and W. Martin, in *Experimental Gravitation*, proceedings of a meeting held at Pavia, Italy, 1976, edited by B. Bertotti (Accademia Nazionale dei Lincei, Rome, 1977), p. 365.

[6] R. Weiss, in *Sources of Gravitational Radiation*, edited by L. Smarr (Cambridge University Press, Cambridge, 1979), p. 7.

[7] V. B. Braginsky and Yu. I. Vorontsov, Usp. Fiz. Nauk 114, 41 (1974) [Sov. Phys.—Usp. 17, 644 (1975)].

[8] R. Weiss, Quarterly Progress Report No. 105, Research Laboratory Electronics, MIT, 1972 (unpublished).

[9] W. Winkler, in *Experimental Gravitation*, proceedings of a meeting held at Pavia, Italy, 1976, edited by B. Bertotti (Accademia Nazionale dei Lincei, Rome, 1977), p. 351.

[10] W. A. Edelstein, J. Hough, J. R. Pugh, and W. Martin, J. Phys. E 11, 710 (1978).

[11] C. M. Caves, Phys. Rev. Lett. 45, 75 (1980).

[12] R. J. Glauber, Phys. Rev. 131, 2766 (1963).

[13] D. Stoler, Phys. Rev. D 1, 3217 (1970).

[14] E. Y. C. Lu, Lett. Nuovo Cimento 3, 585 (1972).

[15] J. N. Hollenhorst, Phys. Rev. D 19, 1669 (1979).

[16] D. Stoler, Phys. Rev. D 4, 1925 (1971).

[17] E. Y. C. Lu, Lett. Nuovo Cimento 2, 1241 (1971).

[18] H. P. Yuen, Phys. Lett. 51A, 1 (1975).

[19] H. P. Yuen, Phys. Rev. A 13, 2226 (1976).

[20] H. P. Yuen and J. H. Shapiro, IEEE Trans. Inf. Theory IT-24, 657 (1978).

[21] J. H. Shapiro, H. P. Yuen, and J. A. Machado Mata, IEEE Trans. Inf. Theory IT-25, 179 (1979).

[22] H. P. Yuen and J. H. Shapiro, IEEE Trans. Inf. Theory IT-26, 78 (1980).

[23] W. G. Unruh has sketched a procedure for doing a complete analysis of a *standard interferometer* (no squeezed states); see the chapter by Unruh in *Gravitational Radiation, Collapsed Objects, and Exact Solutions*, proceedings of the Einstein Centenary Summer School, Perth, Australia, 1979, edited by C. Edwards (Springer, Berlin, 1980), p. 385.

[24] R. Hanbury Brown and R. Q. Twiss, Nature 177, 27 (1956).

[25] R. Hanbury Brown and R. Q. Twiss, Proc. R. Soc. London A243, 291 (1958).

[26] For a review of the theory of bunched and antibunched light and of the efforts to detect antibunching, see R. Loudon, Rep. Prog. Phys. 43, 913 (1980), or D. F. Walls, Nature 280, 451 (1979).

[27] D. Stoler, Phys. Rev. Lett. 33, 1397 (1974).

[28] C. W. Helstrom, Opt. Commun. 28, 363 (1979).

[29] H. J. Kimble, M. Dagenais, and L. Mandel, Phys. Rev. Lett. 39, 691 (1977).

[30] M. Dagenais and L. Mandel, Phys. Rev. A 18, 2217 (1978).

[31] A review of optical parametric amplifiers and parametric oscillators is given by R. G. Smith, in *Laser Handbook*, edited by F. T. Arecchi and E. O. Schulz-Dubois (North-Holland, Amsterdam, 1972), Vol. I, p. 837.

[32] H. Takahasi, Adv. Commun. Syst. 1, 227 (1965), especially Sec. XI.

[33] M. T. Raiford, Phys. Rev. A 2, 1541 (1970).

[34] M. T. Raiford, Phys. Rev. A 9, 2060 (1974).

[35] L. Mišta, V. Peřinová, J. Peřina, and Z. Braunerová, Acta Phys. Pol. A 51, 739 (1977).

[36] H. P. Yuen and J. H. Shapiro, Opt. Lett. 4, 334 (1979).

[37] J. AuYeung and A. Yariv, *Laser Spectroscopy IV*, edited by H. Walther and K. W. Rothe (Springer, Berlin, 1979), p. 492.

[38] R. Loudon, *The Quantum Theory of Light* (Clarendon, Oxford, 1973), especially Chap. 9.

Reprinted from Nature, Vol. 306, No. 5939, pp. 141-146, 10 November 1983
© *Macmillan Journals Ltd., 1983*

REF. 424

Squeezed states of light

D. F. Walls

Physics Department, University of Waikato, Hamilton, New Zealand

The properties of a unique set of quantum states of the electromagnetic field are reviewed. These 'squeezed states' have less uncertainty in one quadrature than a coherent state. Proposed schemes for the generation and detection of squeezed states as well as potential applications are discussed.

THE electric field for a nearly monochromatic plane wave may be decomposed into two quadrature components with time dependence cos ωt and sin ωt respectively. In a coherent state, the closest quantum counterpart to a classical field, the fluctuations in the two quadratures are equal and minimize the uncertainty product given by Heisenberg's uncertainty relation. The quantum fluctuations in a coherent state are equal to the zero-point fluctuations and are randomly distributed in phase. These zero-point fluctuations represent the standard quantum limit to the reduction of noise in a signal. Even an ideal laser operating in a pure coherent state would still possess quantum noise due to zero-point fluctuations.

Other minimum uncertainty states are possible which have less fluctuations in one quadrature phase than a coherent state at the expense of increased fluctuations in the other quadrature phase. Such states, which have been called squeezed states[1-5] (other names include two photon coherent states, generalized coherent states), no longer have their quantum noise randomly distributed in phase. Such states offer intriguing possibilities. In the present optical communication systems which use coherent beams of laser light propagating in optical fibres, the ultimate limit to the noise is given by the quantum noise or zero-point fluctuations. If, instead, beams of squeezed light were used to transmit information in the quadrature phase that had reduced fluctuations the quantum noise level could be reduced below the zero-point fluctuations. Optical communication systems based on light signals with phase sensitive quantum noise have been proposed by Yuen and Shapiro[6,7].

The concept of squeezed states applies to other quantum mechanical systems. For example, they may have a role in increasing the sensitivity of a gravitational wave detector. A standard bar detector for gravitational radiation may be treated as a harmonic oscillator. The effect of the gravitational radiation is so weak that the expected displacement of the bar is of the order of 10^{-19} cm. This is the same order of magnitude as the quantum mechanical uncertainty of the bar's position in its ground state. Thus the signal from the gravitational wave detector may be obscured by the zero-point fluctuations of the detector. This is a striking example of the influence of quantum fluctuations on a macroscopic system. In principle, a way of beating this problem is clear. Instead of the ground state of the oscillator with its quantum noise randomly distributed in phase one prepares the oscillator in a squeezed state. One then measures the displacement due to the gravitational radiation in the quadrature with reduced fluctuations. In this way it should be possible to detect displacements less than the quantum mechanical uncertainty in the bar's position. Of course, this leaves a lot of technical questions unanswered. How does one prepare the bar in a squeezed state? How does one make a measurement on the bar's quadrature phase? These problems and suggested solutions are discussed elsewhere[8,9] in treatments of quantum non-demolition measurements.

The statistical properties of light fields such as coherent or thermal light may be calculated by techniques similar to classical probability theory using an expansion of the density operator in terms of coherent states, the Glauber–Sudarshan P representation[10,11]. Coherent light has poissonian photon counting statistics. Squeezed states of light on the other hand may have sub-poissonian photon counting statistics and have no nonsingular representation in terms of the Glauber–Sudarshan *P* distribution. The statistical properties of such fields cannot be calculated by techniques analogous to classical probability theory. Squeezed states are, therefore, an example of a nonclassical light field. To be precise we shall define a nonclassical light field as one that has no positive nonsingular Glauber–Sudarshan *P* function.

Another example of a nonclassical light field is a number state. This certainly has no nonsingular Glauber–Sudarshan *P* function and clearly has sub-poissonian photon statistics. Such nonclassical light fields with sub-poissonian photon statistics which exhibit photon antibunching have been observed experimentally[12,13,50]. A number state, however, has its quantum fluctuations randomly distributed in phase and hence does not exhibit squeezing. While a squeezed state may exhibit sub-poissonian photon statistics and hence photon antibunching it is not a necessity. Sub-poissonian statistics result if the quadrature phase with reduced fluctuations carries the coherent excitation. Using photon counting techniques direct measurements of the intensity fluctuations of a light field are possible. To determine the fluctuations in the quadrature phases a phase sensitive detection scheme is necessary. This can be achieved by homodyning or heterodyning the signal with a local oscillator followed by photon counting measurements. To generate a squeezed state a phase dependent nonlinear optical process is necessary.

Phase dependent correlation functions

Detection of a light signal with a photon counter yields a measurement of the light intensity $I(t)$ or photon number $n(t)$. Using electronic correlators one may then compute the intensity or photon number correlations of the light field. For example, one may measure the normalized second-order correlation function

$$g^{(2)}(0)=\frac{\langle :I^2: \rangle}{\langle I \rangle^2} \qquad (1)$$

where : : denotes normal ordering of the quantum mechanical operators. For sufficiently short counting times the variance $V(n)$ of the photon number distribution is related to $g^{(2)}(0)$ by

$$\frac{V(n)-\langle n \rangle}{\langle n \rangle}=g^{(2)}(0)-1 \qquad (2)$$

A coherent light field with poissonian statistics has $g^{(2)}(0)=1$. Thermal light which has increased intensity fluctuations has $g^{(2)}(0)=2$. Since $g^{(2)}(0)$ represents the probability of two photons arriving simultaneously this is referred to as photon

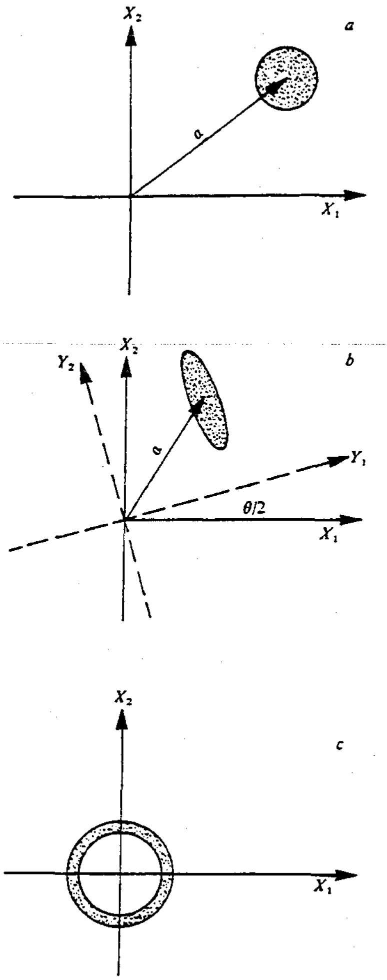


Fig. 1 Phase space plot showing the uncertainty in: *a*, a coherent state $|\alpha\rangle$; *b*, a squeezed state $|\alpha, r\,e^{i\theta}\rangle$ $(r>0)$; *c*, a number state $|n\rangle$.

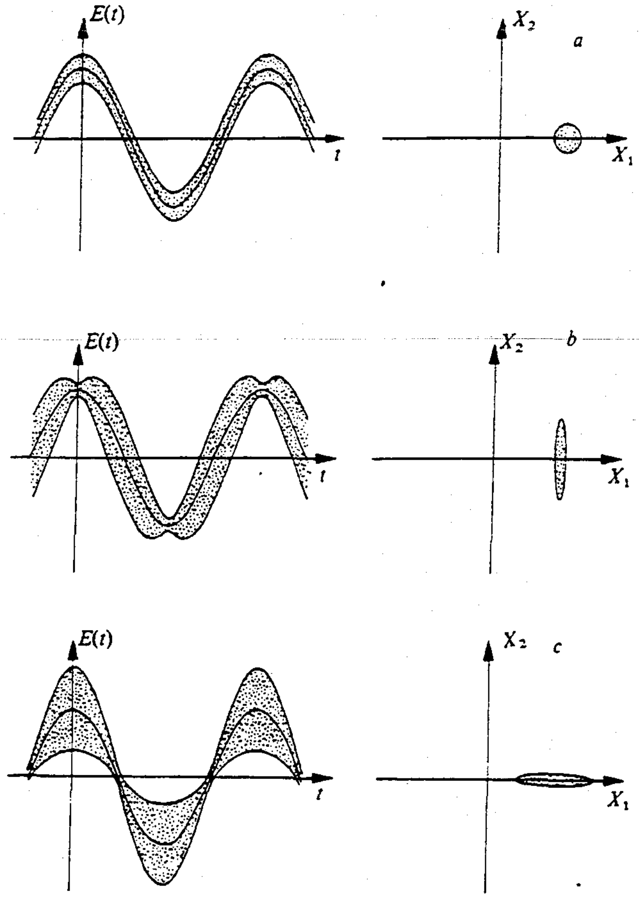


Fig. 2 Plot of electric field against time showing the uncertainty for: *a*, a coherent state $|\alpha\rangle$ (α real); *b*, a squeezed state $|\alpha, r\rangle$ with reduced amplitude fluctuations (α real, $r>0$); *c*, a squeezed state $|\alpha, r\rangle$ with reduced phase fluctuations (α real, $r<0$). Reproduced with permission from Caves[21].

bunching. A light field with sub-poissonian statistics will have $g^{(2)}(0)<1$, an effect known as photon antibunching. Photon antibunching is a quantum mechanical effect which may not be derived from a classical description of the field. Such fields do not have a positive nonsingular representation in terms of the Glauber–Sudarshan P distribution which expresses the density operator for a single mode field as[10,11]

$$\rho = \int P(\alpha)|\alpha\rangle\langle\alpha|\, d^2\alpha \qquad (3)$$

where $|\alpha\rangle$ is a coherent state. This representation has found considerable application in optics because the taking of quantum mechanical averages resemble classical averaging procedures provided $P(\alpha)$ exists as a positive nonsingular function. For fields which exhibit photon antibunching, however, the $P(\alpha)$ are highly singular functions. In this sense we say that such fields are nonclassical. The quantum theory of light received further verification when photon antibunching was observed experimentally in resonance fluorescence from a two level atom[12,13] in agreement with theoretical predictions[14–16] (for reviews see refs 17–19).

Our discussion of the properties of phase dependent correlation functions is illustrated with reference to a single mode field. We may write the electric field as

$$E(t)=\lambda(a\,e^{-i\omega t}+a^{\dagger}\,e^{i\omega t}) \qquad (4)$$

where λ is a constant including the spatial wave functions. In the quantum theory of radiation the amplitudes a and $a^{\dagger}$ are quantum mechanical operators which obey boson commutation relations. We may write

$$a=X_1+iX_2 \qquad (5)$$

where X_1 and X_2 are hermitian operators obeying the commutation relation

$$[X_1, X_2]=\frac{i}{2} \qquad (6)$$

In terms of X_1 and X_2 one may write $E(t)$ as

$$E(t)=\frac{\lambda}{2}(X_1 \cos\omega t+X_2 \sin\omega t) \qquad (7)$$

Thus X_1 and X_2 may be identified as the amplitudes of the two quadrature phases of the field.

From the commutation relation (6) we deduce the following relation for the uncertainties $\Delta X_i=\{V(X_i)\}^{1/2}$ in X_1 and X_2

$$\Delta X_1 \Delta X_2 \geqslant \tfrac{1}{4} \qquad (8)$$

A family of minimum uncertainty states is defined by taking the equal sign. One such class of minimum uncertainty states is the coherent states which have $V(X_1)=V(X_2)=\frac{1}{4}$. A broader class of minimum uncertainty states may have unequal variances in each quadrature. These are the so called squeezed states. The condition for squeezing is

$$V(X_i)<\tfrac{1}{4} \qquad i=1 \text{ or } 2 \qquad (9)$$

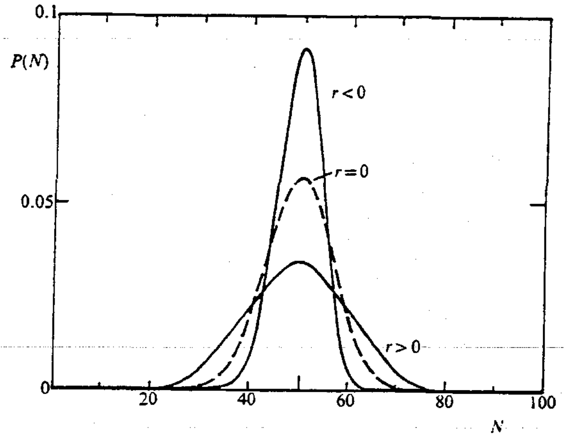


Fig. 3 Photon number distribution for a squeezed state $|\alpha, r\rangle$ ($\alpha = 7$, $r = \pm 0.5$) compared with a coherent state ($r = 0$).

It is sometimes convenient especially for multimode fields to write the condition in terms of the normally ordered variance

$$:V(X_i):<0 \qquad i=1 \text{ or } 2 \tag{10}$$

Figure 1 shows a phase space plot of the uncertainties in X_1 and X_2 for a coherent state, a squeezed state and a number state is shown. These error ellipses may be rigorously derived as the contours of the Q function[4].

The time dependence of $E(t)$ including the uncertainty $\Delta E(t)$ is shown in Fig. 2 for a a coherent state, b a squeezed state with reduced amplitude fluctuations, c a squeezed state with reduced phase fluctuations. The corresponding error box for these states at $t=0$ is also shown.

For a single mode field the variance in one quadrature may be calculated using the Glauber–Sudarshan P representation

$$V(X_1)=\tfrac{1}{4}\left\{1+\int P(\alpha)[(\alpha+\alpha^*)-(\langle\alpha\rangle+\langle\alpha^*\rangle)]^2\,\mathrm{d}^2\alpha\right\} \tag{11}$$

The condition for squeezing $V(X_1)<\frac{1}{4}$ requires that $P(\alpha)$ be a nonpositive definite function. In this sense squeezing like photon antibunching is a nonclassical property of the electromagnetic field. Note that to derive equation (11) the commutation relation $[a, a^\dagger]=1$ has been used. If a classical field is assumed from the outset arbitrary squeezing may be obtained in either quadrature. Thus squeezing has a non trivial significance only in the case of quantized fields. A distinction between classical and quantum fields may be obtained from the normally ordered correlation function $g^{(2)}(0)$ which is always ≥ 1 for classical fields (see ref. 20).

Properties of squeezed states

We shall now briefly describe the mathematical properties of squeezed states. A coherent state $|\alpha\rangle$ may be generated by the action of the displacement operator $D(\alpha)$ on the vacuum

$$|\alpha\rangle=D(\alpha)|0\rangle \tag{12}$$

where

$$D(\alpha)=\mathrm{e}^{-\frac{1}{2}|\alpha|^2}\,\mathrm{e}^{\alpha a^\dagger}\,\mathrm{e}^{-\alpha^* a}$$

A squeezed state $|\alpha, \zeta\rangle$ may be generated by first acting with the squeeze operator $S(\zeta)$ on the vacuum followed by the displacement operator $D(\alpha)$ (ref. 21)

$$|\alpha, \zeta\rangle=D(\alpha)S(\zeta)|0\rangle \tag{13}$$

where

$$S(\zeta)=\exp\left(\tfrac{1}{2}\zeta^* a^2-\tfrac{1}{2}\zeta a^{\dagger 2}\right)$$

and

$$\zeta=r\,\mathrm{e}^{i\theta}$$

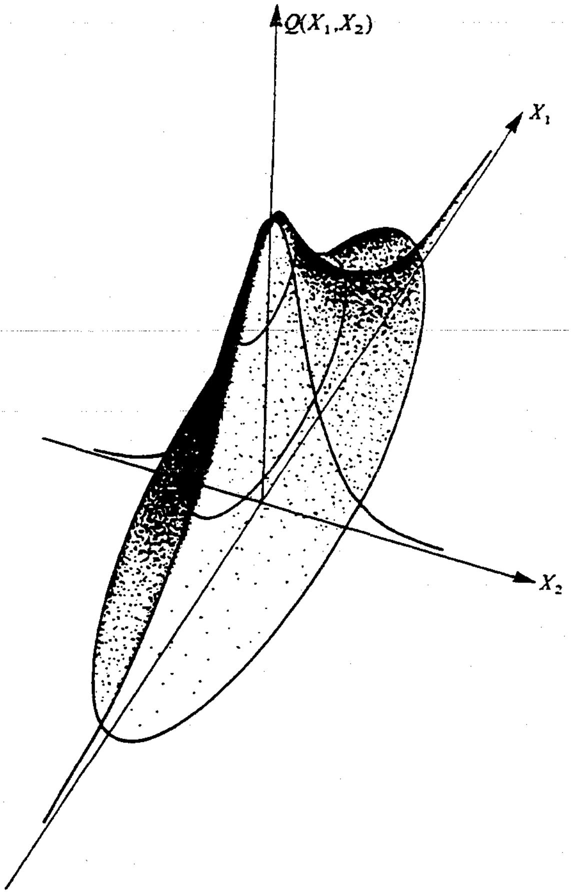


Fig. 4 Q function for a squeezed state. Reproduced with permission from Yuen[4].

An alternative but equivalent characterization of squeezed states has been given by Yuen[4]. We note that whereas a coherent state is generated by linear terms in a and $a^\dagger$ in the exponent a squeezed state requires quadratic terms.

The variances in squeezed state $|\alpha, \zeta\rangle$ are given by

$$\begin{aligned} V(Y_1)&=\tfrac{1}{4}\,\mathrm{e}^{-2r} \\ V(Y_2)&=\tfrac{1}{4}\,\mathrm{e}^{2r} \end{aligned} \tag{14}$$

where $Y_1+\mathrm{i}Y_2=(X_1+\mathrm{i}X_2)\,\mathrm{e}^{-i\theta/2}$ is a rotated complex amplitude so that $2\Delta Y_1$ and $2\Delta Y_2$ represent the length of the minor and major axes of the error ellipse. The mean photon number in the squeezed state $|\alpha, \zeta\rangle$ is

$$\langle n\rangle=|\alpha|^2+\sinh^2 r \tag{15}$$

Clearly, the variances $V(X_i)$ are independent of the field amplitude α. Thus squeezing is a quantum mechanical effect which may occur in fields with high intensity. In this sense one may say it is a macroscopic quantum effect. This is a significant difference from photon antibunching which is only appreciable for fields with low intensity. There is no general relation between photon antibunching and squeezing, however, we shall consider the limit where the coherent amplitude greatly exceeds the squeezing ($|\alpha|^2 \gg \sinh^2 r$). In this limit we find

$$:V(X_1):=\frac{\alpha^2}{4}\left(g^{(2)}(0)-1\right)\approx\tfrac{1}{4}\left(\mathrm{e}^{-2r}-1\right) \tag{16}$$

where we have chosen α real so that the amplitude is carried by X_1.

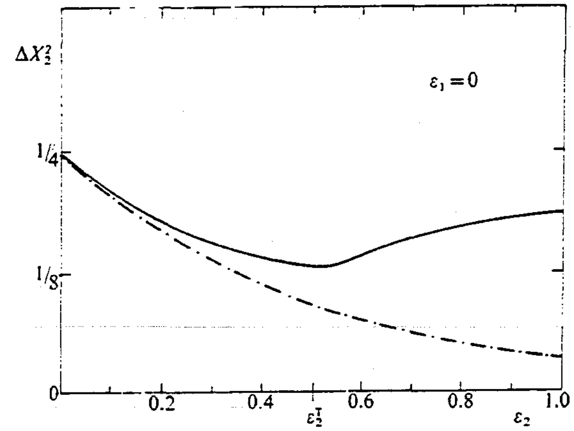


Fig. 5 Squeezing in the parametric oscillator (——) compared with an ideal parametric amplifier (–·–·–·–) as a function of the pump driving field.

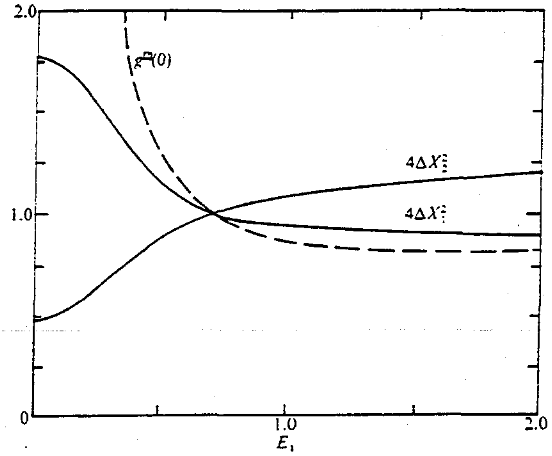


Fig. 6 Photon statistics $g^{(2)}(0)$ and variances ΔX_i^2 for the parametric oscillator as a function of the idler driving field. The pump driving field is held fixed at the threshold value.

For $r > 0$ we have a reduction in amplitude fluctuations $(:V(X_1):<0)$ and photon antibunching $(g^{(2)}(0)<1)$ whereas for $r<0$ we have an increase in amplitude fluctuations and photon bunching. Hence in this limit a squeezed state may show either photon bunching or antibunching depending on whether the amplitude fluctuations are increased or reduced. The photon number distribution[4] in a squeezed state $|\alpha, r\rangle$ is plotted in Fig. 3 for $\alpha = 7$, $r = \pm 0.5$. We can see that the photon statistics are sub- or super-poissonian depending on whether $r > 0$ or $r < 0$.

No such simple relation between antibunching and squeezing exists for all values of α. For example in the opposite limit of $\alpha \ll 1$, that is, a squeezed vacuum $|0, r\rangle$,

$$g^{(2)}(0) = 1 + \frac{\cosh 2r}{\sinh^2 r} \tag{17}$$

Thus the photons in a squeezed vacuum are always bunched irrespective of the sign of the squeeze parameter.

An example of a quantum state which exhibits photon antibunching but no squeezing is a number state $|n\rangle$ for which

$$V(X_1) = V(X_2) = (\tfrac{1}{4})(2n+1) \tag{18}$$

The complete absence of phase information in a number state is clear from the phase space annulus shown in Fig. 1c.

As the Glauber–Sudarshan P representation does not exist for a squeezed state we must consider an alternative representation such as the Wigner function, the Q function, or the generalized P function[22,23]. The Q function for a squeezed state is derived in ref. 4. The distribution $Q(X_1, X_2)$ plotted in Fig. 4 as a function of the amplitudes of the two quadratures clearly shows the unequal variances in X_1 and X_2.

Production of squeezed states

There has been no experimental manifestation of squeezed states of light. The requirement to produce a squeezed state may be simply expressed as follows. For a single mode field mix a part of the field with its phase conjugate to produce a new mode b such that

$$b = \mu a + \nu a^{+} \tag{19}$$

where $\mu^2 - \nu^2 = 1$. For mode a in a coherent state the mode b will be in a squeezed state[4]. Thus a scheme involving a phase conjugate mirror appears as one of the favourite candidates for a state squeezer[24]. The above prescription seems very simple, however, the phase conjugate mirror involves a nonlinear interaction, in this case a four-wave mixing interaction. Squeezed states may also be generated by a three-wave mixing interaction as for example in the parametric amplifier[25–27,51,52]. The prototype for these interactions is described by the Hamiltonian,

$$H = \hbar[\chi^{(n)*}(\varepsilon)a^2 + \chi^{(n)}(\varepsilon)a^{+2}] \tag{20}$$

where

$\chi^{(2)}(\varepsilon) = \varepsilon\chi^{(2)}$ (degenerate parametric amplifier)

$\chi^{(3)}(\varepsilon) = \varepsilon^2\chi^{(3)}$ (four-wave mixing)

$\chi^{(n)}$ is the *nonlinear susceptibility* of the optical medium and ε is the amplitude of the pump field which has been treated classically. This approximate form of the Hamiltonian generates squeezed states with a squeeze parameter given by $r = |2\chi^{(n)}(\varepsilon)t|$. Several objections to this ideal system can be raised. Fluctuations resulting from the quantization of the pump field and the nonlinear medium have been neglected as have vacuum fluctuations associated with any loss process. The vacuum fluctuations will tend to equalize the variances in the two quadratures and hence destroy the squeezing[28–29]. Thus the characteristic damping time of any loss mechanism should be long compared with the interaction time. Phase and amplitude fluctuations in the laser used for the pump may also degrade the squeezing[30]. Phase fluctuations may be compensated for by using part of the pump as the local oscillator in a homodyne detection scheme.

The magnitude of the squeezing is limited by the small values of the nonlinear susceptibility and the interaction time. To increase the interaction time the nonlinear crystal may be placed inside an optical cavity. There is a parametric oscillator configuration where the cavity modes are driven externally by classical fields. An analysis of the cavity must include the cavity losses which tend to destroy the squeezing. Thus there will be a competition between the squeezing produced by the nonlinear interaction and the degradation of the squeezing by the damping. This results in a limiting value to the squeezing attainable in the steady state. An analysis of the degenerate parametric oscillator including the quantization of the pump field has been carried out[31–33]. When only the pump mode is driven by an external field there exists a *threshold driving field* below which the semiclassical value of the mean field is zero. The squeezing in the idler mode as a function of the pump field amplitude is shown in Fig. 5. As the pump amplitude is increased from zero squeezing appears in the idler mode. However, the squeezing approaches a maximum value corresponding to $V(X_2)=1/8$ close to the threshold value of the pump field, then decreases as the pump power is increased above threshold.

The case where the driving field for the pump mode is held fixed at the threshold value and the driving field for the idler is increased from zero is shown in Fig. 6. For low values of the idler driving field the squeeze parameter is initially positive and amplitude fluctuations are increased, hence we have photon bunching ($g^{(2)}(0)>1$). As the idler driving field is increased the squeeze parameter goes to zero and then becomes negative. Thus we have reduced amplitude fluctuations and hence photon antibunching ($g^{(2)}(0)<1$). This is consistent with the general properties of squeezed states with $|\alpha|^2 \gg \sinh^2 r$ discussed above. This system provides a feasible scheme for detecting squeezed states by making photon correlation measurements directly on the squeezed field. Facility to change the sign of the squeeze parameter and observe the accompanying change of the photon statistics from bunching to antibunching would indicate the presence of a squeezed state.

Other nonlinear intracavity devices have been shown[34,35] to give a maximum squeezing factor not greatly exceeding 2. The coupling of the cavity modes to the vacuum fluctuations of the extracavity modes apparently acts as a counter to the squeezing produced by the nonlinear interaction in a steady-state configuration.

One possibility of avoiding the limitation to squeezing imposed by the vacuum fluctuations entering the cavity is to make one mirror perfectly reflecting. It has been claimed that since the vacuum fluctuations may no longer enter from the second port arbitrary squeezing is in principle attainable (B. Yurke, personal communication).

Another way to avoid the problem of vacuum fluctuations is to revert to the parametric amplifier configuration where the cavity losses no longer have a role. The parametric amplifier is a travelling wave phase matched interaction and the Hamilton equation (20) which only includes a single mode is not appropriate. A multimode analysis[36] of a travelling wave parametric amplifier indicates that a reduction in squeezing over the single mode case may occur for the non degenerate amplifier. This reduction in squeezing is caused by the contribution from non-resonant modes whose axes of squeezing become misaligned with respect to the resonant mode.

Another possible system for producing squeezed states is a two-photon laser due to the quadratic nature of the field interaction. A laser, however, is an active system in which the atoms are pumped to the excited state and may consequently decay by spontaneous emission. Calculations using a two-level model for the atomic medium reveal that any potential squeezing is destroyed by the fluctuations resulting from spontaneous emission[37,38].

It is clear, therefore, that a phase sensitive nonlinear interaction in a passive medium is required to produce squeezed states. Predictions of squeezing in a variety of nonlinear optical processes have now been made, for example the free electron laser[39], second harmonic generation[40,41], the single atom-single field mode interaction[42], and multiphoton absorption[43]. The prediction of squeezing in four wave mixing[24] has attracted the interest of experimentalists[42]. An analysis of the effect of atomic fluctuations in four-wave mixing based on a two-level atomic medium reveals that for the atoms driven near saturation or close to resonance the spontaneous emission will destroy the squeezing[45]. For significant squeezing the driving fields should be of low intensity and sufficiently far from resonance so as not to saturate the atoms.

Another system with somewhat different characteristics is squeezing in resonance fluorescence from a two-level atom[46]. Resonance fluorescence differs from many of the systems discussed above as it involves many modes of the radiation field. Resonance fluorescence deserves attention as it is the only system in which photon antibunching has been observed[12,13].

We consider a two-level atom driven by a coherent driving field. The product of the amplitude of the driving field and the dipole moment of the atom is characterized by the Rabi frequency. We denote the Rabi frequency normalized by the natural linewidth of the atom by Ω. The driving field may have a detuning with respect to the atomic transition. We shall use δ to characterize the detuning normalized by the natural linewidth.

The condition for squeezing in a field may best be expressed in terms of the normally ordered variances which do not include the contribution from the vacuum fluctuations. For squeezing in either quadrature ($E_1=(E^{(+)}+E^{(-)})/2$. $E_2=(E^{(+)}-E^{(-)})/2i$) of the field we require

$$:V(E_i):<0 \qquad i=1 \text{ or } 2 \tag{21}$$

We calculate the squeezing in the components of the fluorescent field in the direction along and perpendicular to the mean field. The variance in the component E_1' in the direction of the mean field is

$$:V(E_1'):=\frac{-\lambda}{4}\,\frac{(1+\delta^2-\Omega^2)}{1+\delta^2+\Omega^2}\,\Omega^2 \tag{22}$$

where λ is a constant.

Thus we find squeezing in this component provided $\Omega^2<1+\delta^2$. No squeezing occurs in the component orthogonal to the mean field. The reduced amplitude fluctuations occurring for $\Omega^2<1+\delta^2$ is consistent with the observed fact that the fluorescent light is antibunched. We note that the fluorescent light is also antibunched in the strong field limit $\Omega^2>1+\delta^2$ where there is no squeezing. In this limit the characteristics of the fluorescent light resemble a number state.

Detection of squeezed states

Proposals to measure the variances in the quadrature phases of a light field suggest homodyning or heterodyning the signal with a local oscillator which gives the necessary phase dependence followed by a photon counting measurement. Such measurements are feasible with existing technology. (For further details of such a measurement scheme see refs 6, 7, 20.)

The signal field is homodyned with a local oscillator which is assumed to be in a coherent state. The complex amplitude of the local oscillator may be written as $\varepsilon=|\varepsilon|\,e^{i\theta}$ where θ is the phase of the local oscillator with respect to the signal field. In the limit where the amplitude of the local oscillator greatly exceeds the amplitude of the signal field the photon statistics of the combined field are directly related to the normally ordered variance of the signal field. Assuming a perfect detector efficiency it may be shown that[6,7,20]

$$\begin{aligned} V(n)-\langle n\rangle &= 4|\varepsilon|^2{:}V(E_1){:} \quad \text{if } \theta=0 \\ &= 4|\varepsilon|^2{:}V(E_2){:} \quad \text{if } \theta=\pi/2 \end{aligned} \tag{23}$$

Thus by changing the phase of the local oscillator a measurement of the photon statistics yields the normally ordered variance in E_1 ($\theta=0$) and the normally ordered variance in E_2 ($\theta=\pi/2$). A change of photon statistics from sub- to super-poissonian as θ is varied will indicate the presence of squeezing.

Such measurements impose a stringent requirement on the relative phase stability between the local oscillator and the signal. Yuen and Chan[47] have recently suggested that photon number fluctuations in the local oscillator may be eliminated using a balanced detector scheme developed in the microwave region[48].

Another way to detect a squeezed state is by a direct photon correlation measurement if one has the facility to vary the sign of the squeeze parameter. The presence of a squeezed state is indicated by a change of photon statistics from bunching to antibunching as the squeeze parameter is varied. An example of such a system is the parametric oscillator with two driving fields discussed earlier. This method obviates the need for homodyning the signal with a local oscillator.

Applications of squeezed states

Squeezed states have several potential applications, one, for example, is in optical communication systems. In a proposed scenario information would be transmitted in the quadrature of the field with reduced quantum fluctuations. An enhanced signal-to-noise ratio could then be obtained in the quantum noise limited regime over information sent using coherent light beams. The application of squeezed states in optical communications systems is discussed in refs 6 and 7.

Similar considerations hold in the amplification of signals. Noise is necessarily added in the amplification process, however, if a suitable phase sensitive amplifier is used the noise may be added preferentially to the quadrature not carrying information. This leaves the amplification of the quadrature carrying the information essentially noise free.

Interferometric techniques to detect very weak forces such as gravitational radiation experience limitations on sensitivity due to quantum noise arising from photon counting and radiation pressure fluctuations. These sources of noise may be interpreted as arising from the beating of the input laser with the vacuum fluctuations entering the unused port of the interferometer. It turns out that these two different noise sources arise from fluctuations in the two different quadrature phases of the vacuum entering the unused input port. It has been suggested by Caves[21] that injecting a squeezed state into the unused input port will reduce one or other of the two sources of noise depending on which quadrature is squeezed.

Another intriguing application of squeezed states is in an optical waveguide tap. Shapiro has shown that a high signal-to-noise ratio may be obtained using a squeezed state in an optical waveguide to tap a signal carrying waveguide[49]. This may be achieved with very low energy loss from the signal thus offering the possibility of permitting optical data bus technology to reach multikilometre path lengths with many user sites but no repeaters.

Conclusions

The field of quantum optics has been an active field of research since the early 1960s. However, much which has been discussed under this heading could more correctly be described as nonlinear optics as no quantization of the electromagnetic field is necessary. Very few features which are explicitly a result of quantization of the field have been observed—photon antibunching being one exception. Squeezed states represent a class of quantum states for which no classical analogue exists, hence their detection would be of fundamental interest.

The achievements of quantum optics have been based on the measurement of photon correlation functions of the electromagnetic field. We now seem to be on the verge of an era where a new class of measurements on the phase dependent correlation functions of the electromagnetic field will be possible. This will enable information on the electromagnetic field to be obtained which was not accessible from photon correlation measurements. Such measurements based on homodyning or heterodyning the field with a local oscillator appear feasible with current technology. The presence of a squeezed state will be indicated by the observation of sub-poissonian photon statistics in such a phase sensitive detection process.

Present efforts are directed towards methods of generating a squeezed state. While a proof in principle of the existence of squeezed states seems possible in, for example, resonance fluorescence from a two-level atom or an intracavity nonlinear optical interaction the magnitude of the squeezing obtained in such systems is small. To obtain appreciable squeezing one must look to either a single pass device with a high nonlinearity and low losses or possibly to a cavity with a single input/output port which prevents the vacuum fluctuations entering as in the two port cavity.

Should a device be found to give a light field with significant squeezing the potential applications are attractive. These applications lie on the frontier of technology in quantum noise limited situations. For example, a squeezed light field could be used in an optical communication system where the information is carried by the quadrature with reduced quantum fluctuations. This would enable a better signal-to-noise ratio to be attained than using conventional laser sources which are limited by the quantum noise of a coherent state. The general concept of squeezed states with their phase dependence of quantum noise has important implications in quantum amplifier theory and ultrasensitive electronics such as required for the detection of gravitational radiation. While no experimental observation of squeezed states has yet been reported this is a goal well worth achieving both from a fundamental point of view and in consideration of the applications that will follow.

I thank H. J. Carmichael and G. J. Milburn for valuable suggestions. This work was supported in part by the United States Army through its European Research Office.

1. Robinson, D. R. *Commun. Math. Phys.* **1**, 159 (1965).
2. Stoler, D. *Phys. Rev.* **D1**, 3217 (1970); **D4**, 1925 (1971).
3. Lu, E. Y. C. *Lett. Nuovo Cimento* **2**, 1241 (1971); **4**, 585 (1972).
4. Yuen, H. P. *Phys. Rev.* **A13**, 2226 (1976).
5. Hollenhorst, H. N. *Phys. Rev.* **D19**, 1669 (1979).
6. Yuen, H. P. & Shapiro, J. H. *IEEE Trans. Inform. Theory* **IT24**, 657 (1978); **IT26**, 78 (1980).
7. Shapiro, J. H., Yuen, H. P. & Machado Mata, J. A. *IEEE Trans. Inform. Theory* **IT25**, 179 (1979).
8. Braginsky, V. B., Vorontsov, Yu. I. & Thorne, K. S. *Science* **209**, 547 (1980).
9. Caves, C. M., Thorne, K. S., Drever, R. W. P., Sandberg, V. D. & Zimmerman, M. *Rev. Mod. Phys.* **52**, 341 (1980).
10. Glauber, R. J. *Phys. Rev.* **131**, 2766 (1963).
11. Sudarshan, E. C. G. *Phys. Rev. Lett.* **10**, 277 (1963).
12. Kimble, H. J., Dagenais, M. & Mandel, L. *Phys. Rev.* **18**A, 201 (1978).
13. Leuchs, G., Rateike, M. & Walther, H. (in preparation).
14. Carmichael, H. J. & Walls, D. F. *J. Phys.* **9**B, 1199 (1976).
15. Kimble, H. J. & Mandel, L. *Phys. Rev.* **A13**, 2123 (1976).
16. Cohen Tannoudji, C. in *Frontiers in Laser Spectroscopy* (eds Balian, R., Haroche, S. & Liberman, S.) (North Holland, Amsterdam, 1977).
17. Walls, D. F. *Nature* **280**, 451 (1979).
18. Loudon, R. *Rep. Prog. Phys.* **43**, 58 (1958).
19. Cresser, J. D., Häger, J., Leuchs, G., Rateike, M. & Walther, H. in *Dissipative Systems in Quantum Optics* (ed. Bonifacio, R.) 21 (Springer, Berlin, 1982).
20. Mandel, L. *Phys. Rev. Lett.* **49**, 136 (1982).
21. Caves, C. M. *Phys. Rev.* **23**D, 1693 (1981).
22. Drummond, P. D. & Gardiner, C. W. *J. Phys.* **13**A, 211 (1980).
23. Drummond, P. D., Gardiner, C. W. & Walls, D. F. *Phys. Rev.* **24**A, 914 (1981).
24. Yuen, H. P. & Shapiro, J. H. *Opt. Lett.* **4**, 334 (1979).
25. Takahashi, H. *Adv. Commun. Syst.* **1**, 227 (1965).
26. Raiford, M. T. *Phys. Rev.* **A2**, 1541 (1970); **A9**, 2060 (1974).
27. Mista, L., Perinova, V., Perina, J. & Braunerova, Z. *Acta Phys. Pol.* **A51**, 739 (1977).
28. Hillery, M. & Scully, M. O. in *Quantum Optics, Experimental Gravitation and Measurement Theory* (eds Meystre, P. & Scully, M. O.) (Plenum, New York, in the press).
29. Milburn, G. & Walls, D. F. *Am. J. Phys.* (in the press).
30. Wodkiewicz K. & Zubairy, M. S. *Phys. Rev.* **27**A, 2003 (1983).
31. Milburn, G. J. & Walls, D. F. *Opt. Commun.* **39**, 401 (1981).
32. Lugiato, L. A. & Strini, G. *Opt. Commun.* **41**, 67 (1982).
33. Milburn, G. J. & Walls, D. F. *Phys. Rev.* **27**A, 392 (1983).
34. Walls, D. F. & Milburn, G. J. in *Quantum Optics, Experimental Gravitation and Measurement Theory* (eds Meystre, P. & Scully, M. O.) (Plenum, New York, in the press).
35. Lugiato, L. A. & Strini, G. *Opt. Commun.* **41**, 447 (1982).
36. Tombesi, P., Lane, A., Carmichael, H. J. & Walls, D. F. *Opt. Commun.* (in the press).
37. Lugiato, L. A. & Strini, G. *Opt. Commun.* **41**, 374 (1982).
38. Reid, M. D. & Walls, D. F. *Phys. Rev.* **28**A, 332 (1983).
39. Becker, W., Scully, M. O. & Zubairy, M. S. *Phys. Rev. Lett.* **48**, 475 (1982).
40. Mandel, L. *Opt. Commun.* **42**, 437 (1982).
41. Lugiato, L. A., Strini, G. & de Martini, F. *Opt. Lett.* **8**, 256 (1983).
42. Meystre, P. & Zubairy, M. S. *Phys. Lett.* **89**A, 390 (1982).
43. Zubairy, M. S., Razmi, M. S. K., Iqbal, S. & Idrees, M. (in preparation).
44. Kumar, P., Bondurant, R. S., Shapiro, J. H. & Salour, M. M. *Proc. 5th Rochester Conf. on Coherence and Quantum Optics* (eds Mandel, L. & Wolf, E.) (Plenum, New York, in the press).
45. Reid, M. D. & Walls, D. F. *Opt. Commun.* (in the press).
46. Walls, D. F. & Zoller, P. *Phys. Rev. Lett.* **47**, 709 (1981).
47. Yuen, H. P. & Chan, V. W. S. *Opt. Lett.* **8**, 177 (1983).
48. Dicke, R. H. *Rev. Scient. Instrum.* **17**, 268 (1946).
49. Shapiro, J. H. *Opt. Lett.* **5**, 351 (1980).
50. Short, R. & Mandel, L. *Phys. Rev. Lett.* **51**, 384 (1983).
51. Mollow, B. R. & Glauber, R. J. *Phys. Rev.* **160**, 1097 (1967).
52. Mollow, B. R. *Phys. Rev.* **162**, 1256 (1967).

Printed in Great Britain by Macmillan Production Ltd., Basingstoke, Hampshire

Precision Measurement beyond the Shot-Noise Limit

Min Xiao, Ling-An Wu, and H. J. Kimble
Department of Physics, University of Texas at Austin, Austin, Texas 78712

(Received 28 May 1987)

An improvement in precision beyond the limit set by the vacuum-state or zero-point fluctuations of the electromagnetic field is reported for the measurement of phase modulation in an optical interferometer. The experiment makes use of squeezed light to reduce the level of fluctuations below the shot-noise limit. An increase in the signal-to-noise ratio of 3.0 dB relative to the shot-noise limit is demonstrated, with the improvement currently limited by losses in propagation and detection and not by the degree of available squeezing.

PACS numbers: 42.50.Dv, 07.60.Ly, 42.50.Kb

The quantum nature of the electromagnetic field leads to limitations on the sensitivity of precision measurement of amplitude or phase changes of the field. The quantum fluctuations responsible for enforcing a lower limit on the "noise" in an optical experiment are succinctly expressed in terms of uncertainty products that follow from the commutation relations between conjugate field operators. The fundamental limit encountered in optical physics has been the so-called "shot-noise" limit (SNL), which represents a level of fluctuations for which the minimum uncertainty allowed by quantum mechanics is achieved for the uncertainty product and for which the variances for each of two conjugate operators are equal. This symmetric distribution of fluctuations is characteristic of the vacuum state of the field (zero-point fluctuations) or of a coherent state (approximated by a single-mode laser).

Although the vacuum fluctuations of the field have been the practical limit on precision optical measurement, these fluctuations are of course not a limit in principle since quantum states with variance less than that of the vacuum state can be employed. In particular, the use of squeezed states to circumvent the SNL has been discussed for many years in the theoretical literature.[1-8] Squeezed states are characterized by a phase-dependent distribution of quantum fluctuations such that the dispersion in one of two quadrature components of the field drops below the level set by the vacuum state. In a measurement with squeezed light, the signal that one wishes to detect is encoded on the field variable with reduced fluctuations. The detection scheme is arranged to be largely insensitive to the increased fluctuations in the conjugate variable that are required by the uncertainty relation.

In this Letter we report an experiment in which an improvement in the signal-to-noise ratio of 3.0 dB relative to the SNL has been achieved in an optical measurement with squeezed states. The experiment follows the work of Caves on precision interferometry[6] and employs squeezed light in a Mach-Zehnder interferometer for the detection of phase modulation in propagation along the two arms of the interferometer.[9] The observed increase in sensitivity in the experiment is currently limited by simple linear losses in propagation and detection, and not by the available degree of squeezing from our source. Thus one might anticipate that these rather modest initial results can be substantially improved as the losses in the experiment are reduced.

The experimental arrangement for the use of squeezed states in interferometry is shown in broad outline in Fig. 1. A Mach-Zehnder interferometer is formed by the two beam splitters (m_1, m_4) and the highly reflecting mirrors (m_2, m_3). A coherent field $\hat{E}_1$ at 1.06 μm is injected into the input port m_1, and the fields from the two paths through P_1 and P_2 are recombined at the output port m_4 to produce (complementary) interference fringes as a function of phase difference along the two arms. For the simplest case of 50-50 beam splitters, the fluxes incident upon the detectors D_1 and D_2 are

$$I_{1,2}=\tfrac{1}{2}\xi\langle\hat{E}_1^\dagger\hat{E}_1\rangle[1\pm\cos\phi], \qquad (1)$$

where ϕ is the phase difference for propagation along the

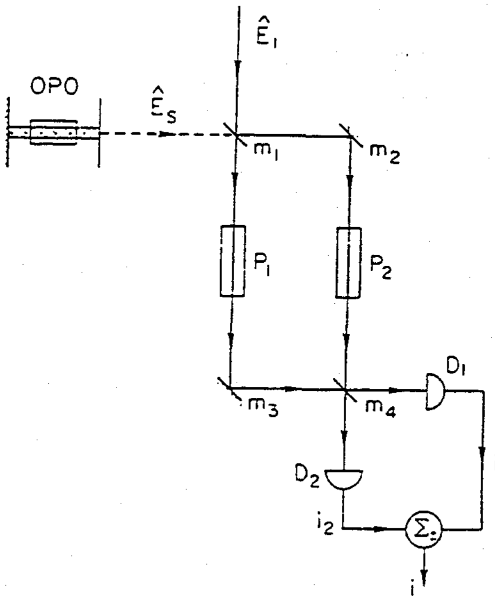


FIG. 1. Diagram of the principal elements of the apparatus for interferometry with squeezed states.

© 1987 The American Physical Society

two arms of the interferometer, ξ is an efficiency factor ($0 \leq \xi \leq 1$) that incorporates possible losses in propagation through the interferometer, and the units of $\hat{E}_1$ are chosen such that $I_{1,2}$ expresses the flux in photons/sec emerging from m_4. The low-frequency (dc–50 kHz) and high-frequency (>500 kHz) components of the photocurrents from the two photodiodes are amplified by separate electronics. For our measurements the operating point of the Mach-Zehnder interferometer is actively stabilized (servo gain unity at 1.2 kHz) near a point $\phi_0 = (2p+1)\pi/2$ ($p=0,1,2,\dots$), i.e., at the half-power point of the fringe, with use of an error signal derived from the low-frequency channels.

The elements (P_1,P_2) in the two arms of the Mach-Zehnder are deuterated potassium dihydrogen phosphate (KD*P) phase modulators, each with half-wave voltage $V' = 1800$ V at 1.06 μm. A voltage $V(t) = V\cos\Omega t$ is applied to the modulator P_1 to produce a phase dither at $\Omega/2\pi = 1.6$ MHz, with the phase $\phi(t) = \phi_0 + 2\delta\cos\Omega t$ and $\delta = \pi V/2V'$. Expanding (1) for small δ, we find a signal at frequency Ω in the difference photocurrent i of rms amplitude $i_s = \sqrt{2}e(\xi\alpha\langle\hat{E}_1^\dagger\hat{E}_1\rangle)\delta$. The noise current i_n against which this signal must be detected is the "shot noise" associated with the total power reaching the detectors (D_1,D_2). That is, $i_n^2 = 2e^2(\xi\alpha\langle\hat{E}_1^\dagger\hat{E}_1\rangle)B$, with B the detection bandwidth and α the detector quantum efficiency. The signal-to-noise ratio Ψ_v in the case of a coherent-state input $\hat{E}_1$ and a vacuum-state input $\hat{E}_s$ is thus

$$\Psi_v \equiv i_s^2/i_n^2 = \xi\alpha P\delta^2/B, \tag{2}$$

where $P \equiv \langle\hat{E}_1^\dagger\hat{E}_1\rangle$ is the power in photons/sec incident upon m_1. A signal-to-noise ratio of unity, $\Psi_v = 1$, implies $\delta_v = [B/\xi\alpha P]^{1/2} = 1/\sqrt{N}$, with N as the mean number of photoelectrons detected in the measurement interval B^{-1}. This limit on the minimum detectable phase change is the SNL and is the best sensitivity possible for inputs of a coherent state $\hat{E}_1$ and a vacuum field $\hat{E}_s$.

To achieve sensitivity beyond the SNL, a squeezed field $\hat{E}_s$ is injected in place of the vacuum field into the normally open input of m_1.[6] As indicated in Fig. 1, the squeezed light in our experiments is produced by an optical parametric oscillator (OPO), which is described in detail by Wu and co-workers.[10] For the present purposes, note that the field generated by the subthreshold OPO is characterized by a spectrum of squeezing $S(\nu,\theta)$, where ν is the frequency offset from the optical carrier, and θ selects the quadrature phase. The angle θ is defined relative to the phase of the coherent field $\hat{E}_1$ inside the interferometer, with $\theta=0$ corresponding to fluctuations in the amplitude quadrature $[S(\theta=0) \equiv S_+]$, and $\theta=\pi/2$ corresponding to fluctuations in the phase quadrature $[S(\theta=\pi/2) \equiv S_-]$. For the subthreshold OPO, the field is a squeezed vacuum state with

$$S_\pm(\bar{\nu}) = \pm 4r/[\bar{\nu} + (1 \mp r)^2], \tag{3}$$

where $\bar{\nu} \equiv \nu/\Gamma_0$ and $\Gamma_0/2\pi \sim 7$ MHz in our experiments.[11] The ratio $r = (p_2/p_0)^{1/2}$ with p_2 the pump power driving the OPO and p_0 the threshold pump power.

Although there are a number of issues of both fundamental and technical natures to be addressed for the use of squeezed states in precision interferometry,[6,12–14] we present only the simplest analysis that is approximately valid for the regime of operation of our interferometer. For the case of a modestly squeezed input $\hat{E}_s$, the calculation of the signal current proceeds exactly as before. However, the noise in the difference photocurrent is now given by $i^2(\nu,\theta) = i_n^2[1+\zeta S(\nu,\theta)]$, with corresponding signal-to-noise ratio

$$\psi(\nu,\theta) = \psi_v/[1+\zeta S(\nu,\theta)], \tag{4}$$

where the factor ζ expresses the efficiency with which the squeezed light is propagated and detected, with $\zeta = \rho T_0 \alpha\eta^2\xi$. Here ρ and T_0 are the efficiencies for escape from the cavity of the OPO and propagation to the beam splitter m_1, while η is the heterodyne efficiency. From Eq. (4) we see that noise reductions ($S<0$) below the level of fluctuations set by a vacuum-state input ($S=0$) lead to an improvement in signal to noise beyond the SNL by the factor $[1+\zeta S_-]^{-1} \gg 1$ for $\zeta = 1$ and $S_- \to -1$. On the other hand, enhanced fluctuations in the conjugate quadrature amplitude produce a large degradation in signal to noise by the factor $[1+\zeta S_+]^{-1} \ll 1$, emphasizing the need for precise control of θ.

Figures 2 and 3 document our observations of the detection of phase modulation $\phi(t)$ with and without squeezed light. In both figures the level of fluctuations Φ of the difference photocurrent i is displayed as a function of time for fixed analysis frequency $\nu/2\pi = 1.6$ MHz, analysis bandwidth $B = 100$ kHz, and two postdetection video filters of time constants $\tau_1 = 1.5\times10^{-4}$ sec and $\tau_2 = 5.0\times10^{-4}$ sec. Phase modulation at 1.6 MHz is applied with P_1 and is gated on and off with a square wave of repetition rate 50 Hz (Fig. 2) or 70 Hz (Fig. 3). The off state represents the noise level in the absence of the signal field with $\Phi = 10\log_{10}R$, where $R \equiv 1+\zeta S$. In Fig. 2(a) this is a noise level set by the laser input $\hat{E}_1$ together with a vacuum-state input $\hat{E}_s$ to the left port of m_1; in Fig. 2(b) it is a noise level set by a squeezed input to m_1 [$S<0$ in Eq. (4)]. The decrease in noise level shown in Fig. 2(b) of more than 3 dB relative to the vacuum or shot-noise level is apparent. The on states in Figs. 2(a) and 2(b) represent the total level of signal plus noise when the phase modulation at 1.6 MHz is gated on, with $\Phi = 10\log_{10}R'$, where R' includes the signal contribution from the phase modulation. The only difference between Figs. 2(a) and 2(b) is that in 2(b) squeezed light has been injected into the normally open input of m_1. In particular both the input power P and the phase modulation $V(t)$ are identical for the two traces. The figure thus represents an improvement in signal to noise of 3.0 dB relative to the SNL for the detection of differential phase changes in the Mach-Zehnder interferometer.

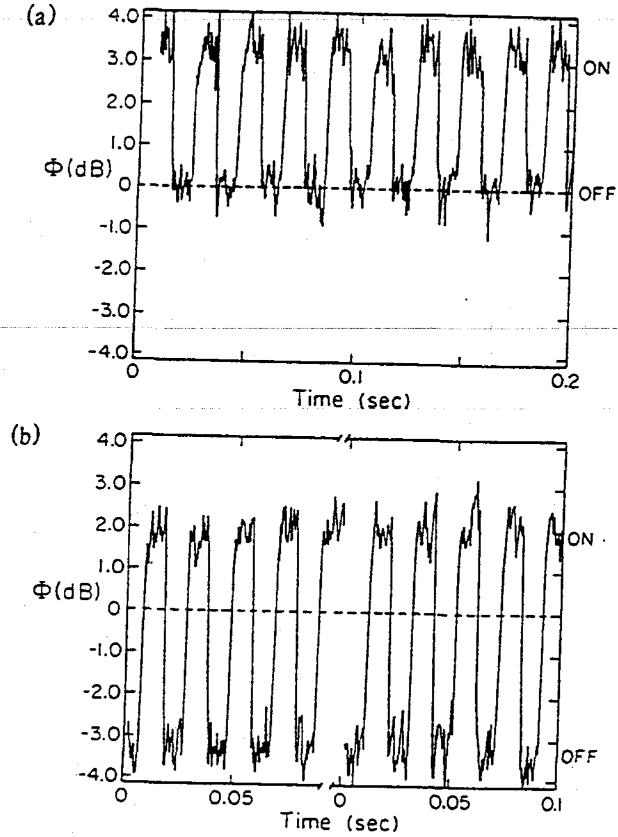


FIG. 2. Level of fluctuations Φ of the difference photocurrent i vs time for fixed analysis frequency $\nu/2\pi=1.6$ MHz, analysis bandwidth $B=100$ kHz, and two video filters of time constant $\tau=1.5\times10^{-4}$ sec and $\tau_2=5\times10^{-4}$ sec. The dashed line gives the vacuum level obtained with no phase modulation and with the squeezed input to m_1 blocked. The ON/OFF square-wave sequence is obtained by gating of the phase modulation to P_1 on and off at a 50-Hz repetition rate. That is, the ON level is obtained with both the modulation signal and noise present; the OFF level is with noise and no signal. The trace in (a) is taken with a vacuum-state input for the field $\hat{E}_s$; the trace in (b) is recorded with a squeezed-state input for $\hat{E}_s$ with the phase θ adjusted for a minimum noise level. An improvement of 3.0 dB in signal-to-noise ratio is achieved in (b) relative to (a). Note that (b) is composed of two traces which are part of the same record obtained during a search for the optimum in both the degree of squeezing S and the local oscillator phase θ.

This observed improvement is in reasonable agreement with that predicted by Eq. (4). Assuming operation at $r^2\sim0.4$, we find $10\log_{10}(\Psi/\Psi_r) = -3.7$ dB for the measured values $\xi=0.94$, $\rho=0.85$, $T_0=0.97$, $\alpha=0.89$, and $\eta=0.93$ appropriate to Fig. 2. While our measurements are conducted without phase-sensitive detection of the modulation at Ω, we note that coherent detection would provide an additional 3 dB of improvement in signal-to-noise ratio both for the case of a vacuum-state input and for a squeezed-state input.

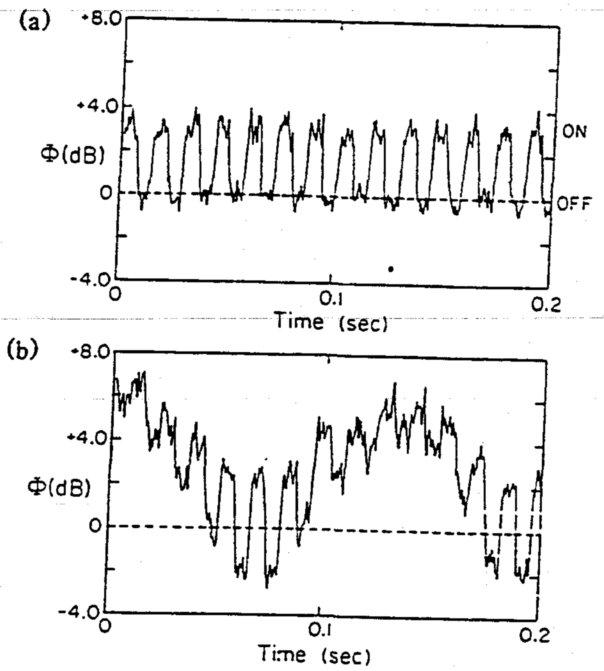


FIG. 3. Level of fluctuations Φ of the photocurrent i vs time. Parameters are similar to those in Fig. 2 with the exception that the phase angle θ is slowly swept with a linear ramp. (a) Vacuum-state input for the field $\hat{E}_s$; (b) Squeezed-state input $\hat{E}_s$. The variation of θ produces alternately a degradation and an improvement in signal to noise as first the increased $(S>0)$ and then the decreased $(S<0)$ fluctuations of the squeezed state are combined with the coherent field $\hat{E}_1$.

The dependence of the detected signal on the phase angle θ is displayed in Fig. 3, which is of the same format as Fig. 2 except that the phase between the fields $\hat{E}_1$ and $\hat{E}_s$ is slowly scanned instead of being held at a fixed value for minimum noise. Figure 3(a) is taken for a vacuum-state input; Fig. 3(b) is for a squeezed state. As expected from Eq. (4) for phase angles such that $S(\theta)<0$, the noise level is reduced, while for $S(\theta)>0$ there is a large degradation in signal-to-noise ratio. This degradation becomes even more pronounced for larger degrees of squeezing until the signal modulation is lost altogether near the maximum noise levels, which for operation closer to threshold are observed to be greater than 8 dB above the SNL. The measured efficiencies for Fig. 3 are $\xi=0.90$, $\rho=0.95$, $T_0=0.97$, $\alpha=0.89$, and $\eta=0.85$. For all our work the fringe visibility is approximately 0.96.

The dashed lines shown in Figs. 2 and 3 are obtained from multiple-trace averages and give the level of fluctuations for a vacuum-state input $\hat{E}_s$ with no phase modulation. That this is indeed the vacuum level is confirmed following the procedures discussed in Ref. 10. Operation at the half-power points of the output fringe has the advantage of preserving the virtues of the balanced homodyne scheme in suppressing excess local oscillator noise. In the current experiments the inherently quiet laser at 1.6 MHz together with the suppression provided by the subtraction arrangement (Σ_{-}) reduces the contribution of excess local oscillator noise to below 0.1%. The signal shown in Fig. 2(a) was generated by a modulation amplitude $\delta \simeq 7.6 \times 10^{-6}$, which agrees with the value $\delta_v = 7.8 \times 10^{-6}$ inferred from the parameters of the experiment ($P = 800\ \mu\text{W}$, $B = 100$ kHz).[15]

In conclusion, we have demonstrated an improvement in sensitivity for optical measurements beyond the limit set by the vacuum-state or zero-point fluctuations of the electromagnetic field. Phase modulation is detected in a Mach-Zehnder interferometer with an improvement in signal-to-noise ratio of 3.0 dB relative to the SNL. We note that this increase in sensitivity is currently limited by losses in propagation and detection and not by the degree of available squeezing. Indeed, the field emitted by the OPO exhibits a degree of squeezing $S^{\theta}_{-} = \rho S_{-} = -0.8$ for the conditions of Fig. 2, indicating the possibility of improvements in sensitivity of 7 dB if the efficiency factors (T_0, α, η, and ξ) can be increased toward unity. Although we have employed a Mach-Zehnder configuration, the results are generally applicable to other types of interferometers, in particular to the arrangements under development for gravity-wave detection.[16]

This research was supported by the U.S. Office of Naval Research and by the Venture Research Unit of British Petroleum North America. We gratefully acknowledge the contributions of C. M. Caves, J. L. Hall, and G. Leuchs, and the technical assistance of S. Pereira to this research program.

[1] V. Braginsky, Y. I. Vorontsov, and K. S. Thorne, Science **209**, 547 (1980).

[2] J. H. Hollenhorst, Phys. Rev. D **19**, 1669 (1979).

[3] H. Takahasi, in *Advances in Communication Systems: Theory and Applications,* edited by A. V. Balakrishnan (Academic, New York, 1965), Vol. 1, p. 227.

[4] H. P. Yuen, Phys. Rev. A **13**, 2226 (1974).

[5] H. P. Yuen and J. H. Shapiro, IEEE Trans. Inf. Theory **24**, 657 (1978), and **26**, 78 (1980).

[6] C. M. Caves, Phys. Rev. D **23**, 1693 (1981).

[7] D. F. Walls, Nature (London) **306**, 141 (1983).

[8] "Squeezed States of the Electromagnetic Field," J. Opt. Soc. Am. B **4** (to be published) (special issue).

[9] H. J. Kimble, Bull. Am. Phys. Soc. **32**, 1280 (1987).

[10] Ling-An Wu, H. J. Kimble, J. L. Hall, and H. Wu, Phys. Rev. Lett. **57**, 2520 (1986); Ling-An Wu, Min Xiao, and H. J. Kimble, in Ref. 8.

[11] M. J. Collett and C. W. Gardiner, Phys. Rev. A **30**, 1386 (1984); C. W. Gardiner and C. M. Savage, Opt. Commun. **50**, 173 (1984).

[12] R. S. Bondurant and J. H. Shapiro, Phys. Rev. D **30**, 2548 (1984).

[13] J. Gea-Banacloche and G. Leuchs, to be published, and in Ref. 8.

[14] B. Yurke, P. Grangier, and R. E. Slusher, in Ref. 8.

[15] Hewlett-Packard Application Note No. 150-4, Spectrum Analyzer Series (Hewlett-Packard Corp., Palo Alto, CA, 1974), Chap. 2.

[16] K. S. Thorne, in *300 Years of Gravitation,* edited by S. W. Hawking and W. Israel (Cambridge Univ. Press, Cambridge, 1987).

REF. WW

Building Scientific Apparatus

A Practical Guide to Design and Construction

John H. Moore ▪ Christopher C. Davis ▪ Michael A. Coplan

University of Maryland

1983

ADDISON-WESLEY PUBLISHING COMPANY
Advanced Book Program / World Science Division
London ▪ Amsterdam ▪ Don Mills, Ontario ▪ Sydney ▪ Tokyo

Figures 4.17, 4.18, 4.19(*a*), 4.45, and 4.119 are reprinted with permission from *Handbook of Lasers*, R. J. Pressley, Ed., CRC Press, Cleveland, 1971. Copyright The Chemical Rubber Co., CRC Press, Inc.

Illustrations by JAMES S. KEMPTON

Library of Congress Cataloging in Publication Data

Moore, John H., 1941–
Building scientific apparatus.

Includes index.
1. Scientific apparatus and instruments—Handbooks, manuals, etc. 2. Instrument manufacture—Handbooks, manuals, etc.
I. Davis, Christopher C., 1944– . II. Coplan, Michael A., 1938– . III. Title.
Q185.M66 1983 681′.75 82-8726
ISBN 0-201-05532-5 AACR2

Copyright © 1983 by Addison-Wesley Publishing Company, Inc.
Published simultaneously in Canada

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher, Addison-Wesley Publishing Company, Inc., Advanced Book Program/World Science Division, Reading, Massachusetts 01867, U.S.A.

Printed in the United States of America

ABCDEFGHIJ–HA–898765432

CHAPTER 3 VACUUM TECHNOLOGY

In the modern laboratory, there are many occasions when a gas-filled container must be emptied. Evacuation may simply be the first step in creating a new gaseous environment. In a distillation process, there may be a continuing requirement to remove gas as it evolves. Often it is necessary to evacuate a container to prevent air from contaminating a clean surface or interfering with a chemical reaction. Beams of atomic particles must be handled *in vacuo* to prevent loss of momentum through collisions with air molecules. A vacuum system is an essential part of laboratory instruments such as the mass spectrometer and the electron microscope. Many forms of radiation are absorbed by air and thus can propagate over large distances only in a vacuum. Far-IR, far-UV, and X-ray spectrometers are operated within vacuum containers. Simple vacuum systems are used for vacuum dehydration and freeze-drying. Nuclear particle accelerators and thermonuclear devices require huge, sophisticated vacuum systems.

3.1 GASES

The pressure and composition of residual gases in a vacuum system vary considerably with its design and history. For some applications a residual gas density of tens of billions of molecules per cubic centimeter is tolerable. In other cases no more than a few hundred thousand molecules per cubic centimeter constitutes an acceptable vacuum: "One man's vacuum is another man's sewer."[1] It is necessary to understand the nature of a vacuum and of vacuum apparatus to know what can and cannot be done, to understand what is possible within economic constraints, and to choose components that are compatible with each other as well as with one's needs.

3.1.1 The Nature of the Residual Gases in a Vacuum System

The pressure below one atmosphere is loosely divided into vacuum categories. The pressure ranges and number densities corresponding to these categories are listed in Table 3.1. As a point of reference for vacuum work it is useful to remember that the number density at 1 mtorr is about 3.5×10^{13} cm^{-3}, and at 1 Pa about 2.7×10^{14} cm^{-3}, and that the number density is proportional to the pressure.

The composition of gas in a vacuum system is modified as the system is evacuated because the efficiency of a vacuum pump is different for different gases. At low pressures molecules desorbed from the walls make up the residual gas. Initially, the bulk of the gas leaving the walls is water vapor; at very low pressures, in a container that has been baked, it is hydrogen.

Table 3.1 AIR AT 20°C

	Pressure (torr)[a]	*Number Density* (cm^{-3})	*Mean Free Path* (cm)	*Surface Collision Frequency* ($\text{cm}^{-2}\text{sec}^{-1}$)	*Times for Monolayer Formation*[b] (sec)
One atmosphere	760	2.7×10^{19}	7×10^{-6}	3×10^{23}	3.3×10^{-9}
Lower limit of:					
Rough vacuum	10^{-3}	3.5×10^{13}	5	4×10^{17}	2.5×10^{-3}
High vacuum	10^{-6}	3.5×10^{10}	5×10^{3}	4×10^{14}	2.5
Very high vacuum	10^{-9}	3.5×10^{7}	5×10^{6}	4×10^{11}	2.5×10^{3}
Ultrahigh vacuum	0				

[a] 1 torr = 132 Pa.

[b] Assuming unit adhesion efficiency and a molecular diameter of 3×10^{-8} cm.

3.1.2 Kinetic Theory

In order to understand mass flow and heat flow in a vacuum system it is necessary to appreciate the immense change in freedom of movement experienced by a gas molecule as the pressure decreases.

The average velocity of a molecule can be deduced from the Maxwell-Boltzmann velocity distribution law:

$$\bar{v} = \left(\frac{8kT}{\pi m}\right)^{1/2}.$$

For an air molecule (molecular weight of about 30) at 20°C,

$$\bar{v} = 5 \times 10^{4}\,\text{cm}\,\text{sec}^{-1} = \tfrac{1}{2}\,\text{km}\,\text{sec}^{-1}.$$

Each second a molecule sweeps out a volume with a diameter twice that of the molecule and a length equal to the distance traveled by the molecule in a second. As shown in Figure 3.1, this molecule collides with any of its neighbors whose center lies within the swept volume. The number of collisions per second is equal to the number of neighbors within the swept volume. On the average, the number of collisions per second (Z) is the number density of molecules (n) times the volume swept by a molecule of velocity $\bar{v}$ and diameter ξ. More precisely,

$$Z = \sqrt{2}\, n\pi\xi^{2}\bar{v},$$

where the $\sqrt{2}$ accounts for the relative motion of the molecules. The time between collisions is the reciprocal of this *collision frequency*, and the average distance between collisions, or *mean free path*, is

$$\lambda = \bar{v}Z^{-1}$$
$$= \frac{1}{\sqrt{2}\, n\pi\xi^{2}}.$$

The mean free path is inversely proportional to the pressure. For an N_2 or O_2 molecule, $\xi = 3 \times 10^{-8}$ cm. The number density at 1 mtorr is about 3.5×10^{13} cm^{-3}. Thus the mean free path in air at 1 mtorr is about 5 cm. Recalling the relationship between λ and P,

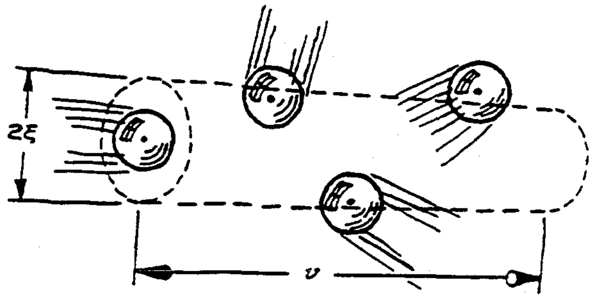


Figure 3.1 The volume swept in one second by a molecule of diameter ξ and velocity $\bar{v}$ ($\text{cm}\,\text{sec}^{-1}$). The molecule will collide with any of its neighbors whose center lies within the volume.

a valuable rule of thumb is that for air at 20°C

$$\lambda = \frac{5}{P(\text{mtorr})} \text{ cm.}$$

3.1.3 Surface Collisions

The frequency of collisions of molecules within a container with the surface of the container, per unit area, is

$$Z_{\text{surface}} = \frac{n\bar{v}}{4} \qquad (\text{sec}^{-1}\text{cm}^{-2}).$$

The sticking probability for most air molecules on a clean surface at room temperature is between 0.1 and 1.0. For water the sticking probability is about unity for most surfaces. Assuming unit sticking probability and a molecular diameter $\xi = 3 \times 10^{-8}$ cm, the time required to form a monolayer of adsorbed air molecules at 20°C is

$$t = \frac{2.5 \times 10^{-6}}{P(\text{torr})} \text{ sec.}$$

Thus to maintain a clean surface for a useful period of time may require a gas pressure over the surface less than 10^{-9} torr.

3.1.4 Bulk Behavior versus Molecular Behavior

The pressure of gas within a vacuum system may vary over ten or more orders of magnitude as the system is evacuated from atmospheric pressure to the lowest attainable pressure. At high pressures, when the mean free path is much smaller than the dimensions of the vacuum container, gas behavior is dominated by intermolecular interactions. These interactions, resulting in viscous forces, ensure good communication among all regions of the gas. At high pressures, a gas behaves as a homogeneous fluid. When the gas density is decreased to the extent that the mean free path is much larger than the container, molecules rattle around like the balls on a billiard table (not a pool table, where the ball density can be high), and gas behavior is determined by the random motion of the molecules as they bounce from wall to wall.

In the *viscous-flow* region where the mean free path is relatively small, gas flow improves with increasing pressure because gas molecules tend to queue up and push on their neighbors in front of them. Gas flow is impeded by turbulence and by viscous drag at the walls of the pipe that is conducting the gas. In the viscous region the coefficients of viscosity and thermal conductivity are independent of pressure.

When the mean free path far exceeds the dimensions of the container, the process of gas flow is called *molecular flow*. In this region momentum transfer occurs between molecules and the wall of a container, but molecules seldom encounter one another. A gas is not characterized by a viscosity. Gas flows from a region of high pressure to one of low pressure simply because the number of molecules leaving a unit of volume is proportional to the number of molecules within that volume. Gas flow is a statistical process. At very low pressure a molecule does not linger at a surface for a sufficient time to reach thermal equilibrium. Thus thermal conductivity at low pressures is a function of gas density, and the coefficient of thermal conductivity depends upon the pressure and upon the condition of the surface.

3.2 GAS FLOW

3.2.1 Parameters for Specifying Gas Flow

Before discussing vacuum apparatus it is necessary to define the parameters used by vacuum engineers to characterize gas flow.

The volume rate of flow through an aperture or across a cross section of a tube is defined as the *pumping speed* at that point:

$$S \equiv \frac{dV}{dt} \qquad \left[\text{liter sec}^{-1};\ \text{ft}^3\text{min}^{-1}\ (\text{cfm})\right].$$

The capacity of a vacuum pump is specified by the speed measured at its inlet:

$$S_p \equiv \frac{dV}{dt} \text{ (at pump inlet)}.$$

The mass rate of flow through a vacuum system is proportional to the *throughput*:

$$Q \equiv PS \qquad (\text{torr liter sec}^{-1};\ \text{torr cfm}).$$

To determine the throughput it is necessary that the pressure and speed be measured at the same place, since these quantities vary throughout the system.

The ability of a tube to transmit gas is characterized by its *conductance* C. The definition of conductance is analogous to Ohm's law for electrical circuitry. The throughput of a tube depends upon the conductance of the tube and the driving force, which in this case is the pressure drop across the tube:

$$Q = (P_1 - P_2)C \qquad (P_1 > P_2).$$

Notice that conductance has the same units as pumping speed.

3.2.2 Network Equations

A complicated network of tubes can be reduced to a single equivalent conductance for the purpose of analysis. In analogy to electrical theory once again, a number of tubes in series can be replaced by a single equivalent tube with conductance C_{series} given by

$$\frac{1}{C_{\text{series}}} = \frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3} + \cdots.$$

A parallel network can be replaced by an equivalent conductor with conductance

$$C_{\text{parallel}} = C_1 + C_2 + C_3 + \cdots.$$

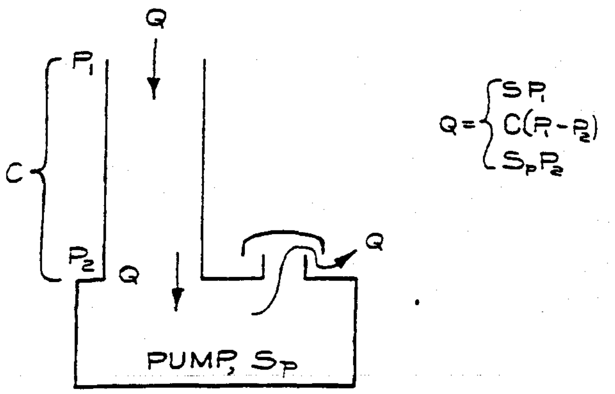


Figure 3.2 Analytical representation of a vacuum system.

3.2.3 The Master Equation

By use of the network equations given above, an entire vacuum system can be reduced to a single equivalent conductance leading from a gas source to a pump as shown in Figure 3.2. The net speed of the system is

$$S = \frac{Q}{P_1},$$

and the speed at the pump inlet is

$$S_p = \frac{Q}{P_2}.$$

Notice that the throughput is the same at every point in this system, since there is only one gas source. The throughput of the conductance is

$$Q = (P_1 - P_2)C.$$

Upon substitution for P_1 and P_2 from the previous two equations, this equation after some rearrangement becomes

$$\frac{1}{S} = \frac{1}{S_p} + \frac{1}{C}.$$

This is the master equation that relates the net speed of a system to the capacity of the pump and the conductors leading to the pump. From this equation we see that both the conductance of a system and the speed of the pump exceed the net speed of the system. In vacuum-system design, economics should be considered. The cost of a vacuum pump is usually greater than the cost of constructing a tube leading from a vacuum container to the pump. In order to achieve a given net speed for a system, it is usually most economical to design a system whose equivalent conductance exceeds the required net speed by a factor of three or four so that the speed of the pump need only exceed the net speed by a small amount.

3.2.4 Conductance Formulae

In the viscous flow region the conductance of a tube depends upon the gas pressure and viscosity. For a tube of circular cross section the conductance for air at 20°C is

$$C = 180\frac{D^4}{L}P_{av}\ \text{liter}\,\text{sec}^{-1},$$

when the diameter D and length L of the tube are in centimeters, and the average pressure P_{av} is in torr. This equation is useful for determining the dimensions of roughing and forepump lines. Elbows, bends and joints in a tube will reduce the conductance, but since conductors are usually overdesigned, it is sufficient to consider only the simple cylindrical tubes in a network and ignore the junctions between these tubes.

In the molecular-flow region conductance is independent of pressure. For air at 20°C the conductance of a tube of circular cross section is

$$C = 12\frac{D^3}{L}\ \text{liter}\,\text{sec}^{-1},$$

when the diameter and length are measured in centimeters. This equation is useful in determining the dimensions of a high-vacuum chamber and the size of the tubes leading from the chamber to a diffusion pump or ion pump. As an example of poor design consider the case of a vacuum chamber connected to a 100-liter sec^{-1} diffusion pump by a tube 2.5 cm in diameter and 10 cm long. The net speed of the pump and connecting tube is

$$S = \left(\frac{1}{100} + \frac{1}{19}\right)^{-1} = 16\ \text{liter}\,\text{sec}^{-1}.$$

The pump in this case is being strangled by the connecting tube. Consider also the pressure drop across the tube. Since the throughput is a constant, the pressure a the inlet of the tube P_1 is related to the pressure at the inlet of the pump P_2 by

$$SP_1 = S_pP_2;$$

thus

$$\frac{P_1}{P_2} = \frac{S_p}{S} = 6.$$

This result indicates the importance of proper location of a pressure gauge. In this case a gauge at the mouth of the pump would indicate a pressure six times lower than the pressure in the vacuum container.

In the molecular-flow region the conductance of an aperture for a gas of molecular weight M is

$$C = 3.7\left(\frac{T}{M}\right)^{1/2}A \qquad \text{liter}\,\text{sec}^{-1},$$

where A is the area in cm^2. This equation is useful for determining the rate of gas flow into a vacuum chamber through an aperture in a gas-filled collision chamber or ion source.[2]

There is a *transition region* between the viscous and molecular flow regions where the mean free path approximately matches the dimensions of the container. Any mathematical description of this region is difficult. In most vacuum systems the transition region is encountered only briefly during pumpdown from atmosphere to high vacuum. The formulae for conductance in the molecular flow region will give a conservative estimate of the conductance in the transition region.

3.2.5 Pumpdown Time

The time required to pump a chamber of volume V from pressure P_0 to P is

$$t = 2.3\frac{V}{S}\ln\frac{P_0}{P},$$

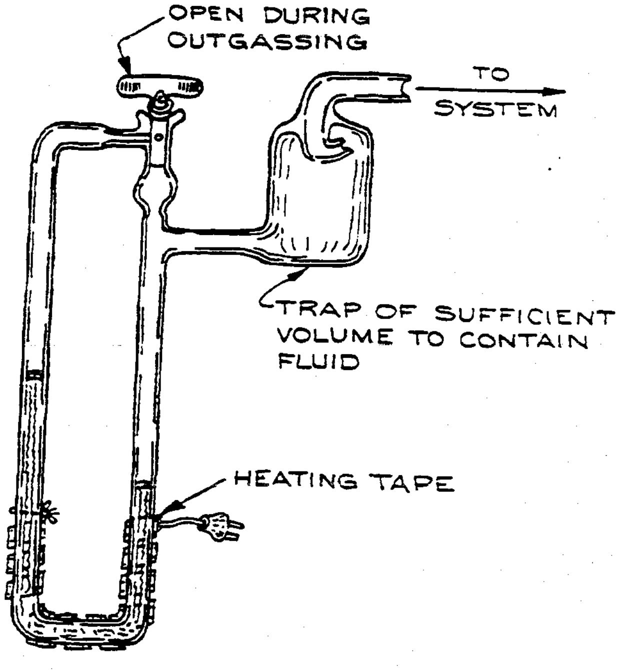
assuming a constant net pumping speed and assuming no additional gas is admitted to the system. A typical high-vacuum system is rough-pumped to about 5×10^{-2}

torr with a mechanical pump and then pumped to very low pressures with a diffusion pump or some other pump that works effectively in the molecular flow region. The pumpdown calculation is performed in two steps corresponding to these two operations.

3.2.6 Outgassing

The rate of evacuation of a chamber at pressures below 10^{-6} torr is not controlled by the pumping speed of a system but rather by the rate of evolution of gas from the walls of the chamber. Stainless steel outgasses at about 10^{-7} torr liter sec^{-1}cm^{-2} after an hour of pumping. More than a day of pumping may be required to reduce this rate below 10^{-8} torr liter sec^{-1}cm^{-2}. The rate of outgassing from metal surfaces can be considerably reduced by baking to drive out adsorbed gases. Plastics and elastomers outgas at as much as 10^{-5} torr liter sec^{-1}cm^{-2} after an hour of pumping, and as a result such materials are not suitable for use at pressures below 10^{-7} torr.

Consider for example the problem of maintaining a base pressure of 10^{-7} torr in a stainless-steel chamber. If the outgassing rate is 10^{-8} torr liter sec^{-1}cm^{-2}, a pumping speed of 0.1 liter sec^{-1} for every cm^2 of surface is required. For a container 30 cm in diameter and 30 cm high, a pumping speed of 400 liter sec^{-1} is needed. This speed is typical of a diffusion pump with a nominal 4-in. inlet port.



OIL DENSITIES (20° C)

di-n-butyl phthalate	1.044 gm cm^{-3}
Octoil	0.983
Octoil-S	0.912
DC-704	1.07
DC-705	1.09

Figure 3.3 Oil manometer. The trap is intended to contain the oil in the event that the low-pressure arm of the manometer is accidentally opened to the atmosphere.

3.3 PRESSURE MEASUREMENT

The pressure within a vacuum system may vary over ten or more orders of magnitude. No one gauge will operate over this great range, and thus most systems are equipped with several different gauges.

3.3.1 Mechanical Gauges

The simplest pressure gauges are hydrostatic gauges such as the oil or mercury manometer. A closed-end U-tube manometer filled with mercury is useful down to 1 torr. Diffusion-pump oils such as Octoil, di-*n*-butyl phthalate, or the Dow Corning silicone oils DC704 or DC705 may also be used as manometer fluids. Oil-filled manometers may be used down to 0.1 torr because the oil density is much less than that of mercury. As shown in Figure 3.3, an oil manometer should be made of glass tubing of at least 1-cm diameter in order to reduce the effects of capillary depression. To prevent oil adhesion to the walls, the manometer should be cleaned with dilute HF before filling. Because air is somewhat soluble in oil, a heater is usually provided to outgas the oil prior to use.[3] Several manometer designs have been developed to increase the surface area of oil exposed to the vacuum in order to hasten outgassing.[4] Both oil and mercury manometers will contaminate a vacuum system with vapor of the working fluid. If this is a problem, a cold trap is placed between the gauge and the system to condense the offending vapor. Mercury in a manometer can be isolated and protected from chemical attack by floating a few drops of silicone oil on top of the mercury column.

The McLeod gauge shown in Figure 3.4 is a sophisticated hydrostatic gauge and is sensitive to much lower pressures than a simple U-tube manometer. In a McLeod gauge a sample of gas is trapped and compressed by a known amount (typically 1000:1) by displacing the gas with mercury from the original volume into a much smaller volume. The pressure of the compressed gas is measured with a mercury manometer and the original pressure determined by use of the general gas law. Of course, it is not possible to use a McLeod gauge to measure the pressure of a gas that condenses upon compression. With care a high-quality McLeod gauge will provide accuracy of a few percent down to about 10^{-4} torr. Some gauges can be used at even lower pressures, but it is necessary to consider the error caused by the pumping action of mercury vapor streaming out of the gauge.[5] The useful pressure range of any one McLeod gauge is limited by geometrical constraints to about four orders of magnitude in pressure.

The McLeod gauge is constructed of glass and is easily broken. If the mercury is raised into the gas bulb too quickly, it can acquire sufficient momentum to shatter the glass. When a gauge is broken, the external air pressure frequently drives a quantity of mercury into the vacuum system. For this reason it is wise to place a ballast bulb of sufficient volume to contain the mercury charge of the gauge between the gauge and the system.

A number of different gauges depend upon the flexure of a metal tube or diaphragm as a measure of pressure. In the Bourdon gauge, a thin-wall, curved tube, closed at one end, is attached to the vacuum system. Pressure changes cause a change in curvature of the tube. A mechanical linkage to the tube drives a needle, which gives a pressure reading on a curved scale. Another type of mechanical gauge contains a chamber divided in two by a thin metal diaphragm. The volume on one side of the diaphragm is sealed, while the volume on the other side is attached to the system. A variation in pressure on one side relative to the other causes the diaphragm to flex, and this movement is sensed by a system of gears and levers, which drives a needle on the face of the gauge. The precision of these gauges is limited by hysteresis caused by friction in the linkage. To overcome this friction it is helpful to gently tap the gauge before making a reading. Unlike the liquid manometer, mechanical gauges are not absolute gauges. The pressure scale and zero location on these gauges must be calibrated against a McLeod gauge or U-tube manometer. Mechanical gauges are useful down to 1 torr. They offer the advantage of being insensitive to the chemical or physical nature of the gas. Excellent mechanical gauges are manufactured by Wallace & Tiernan and by Leybold-Heraeus.

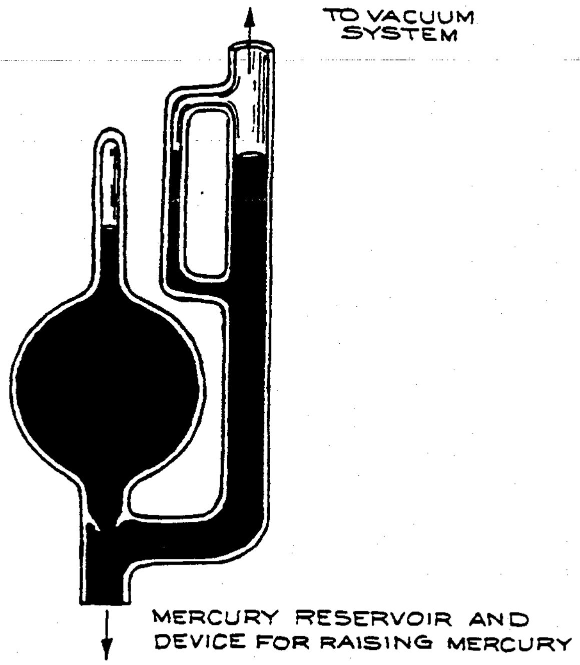


Figure 3.4 McLeod gauge.

A recent development of the diaphragm gauge is the capacitance manometer, wherein the diaphragm is one plate of an electrical capacitor. A change in the diaphragm position results in a change in capacitance, which is detected by a sensitive capacitance bridge. The MKS Baratron capacitance manometer is claimed to be accurate to a few percent at 10^{-4} torr.

3.3.2 Thermal-Conductivity Gauges

The thermal conductivity of a gas decreases from some constant value above 1 torr to essentially zero at about 10^{-3} torr. This change in thermal conductivity is used as an indication of pressure in the Pirani gauge and the thermocouple gauge. In both gauges a low-temperature filament is heated by a constant current. The temperature of the filament depends on the rate of heat loss to the surrounding gas. In the thermocouple gauge, the temperature is determined from the e.m.f. produced by a thermocouple in contact with the filament (Figure 3.5). In the Pirani gauge, a change in temperature of the filament results in a change in resistivity that is detected by a sensitive bridge.

The pressure indicated by a thermocouple gauge or a Pirani gauge depends upon the thermal conductivity of the gas. They are usually calibrated by the manufacturer for use with air. For other gases these gauges must be recalibrated point by point over their entire range, since thermal conductivity is a nonlinear function of pressure. In routine use a thermal-conductivity gauge can only be expected to be accurate to within a factor of two. This accuracy is adequate when the gauge is used to sense the foreline pressure of a diffusion pump or to determine whether the pressure in a system is sufficiently low to begin diffusion-pumping. The principal advantages of these gauges are their ease of use, ruggedness, and low cost.

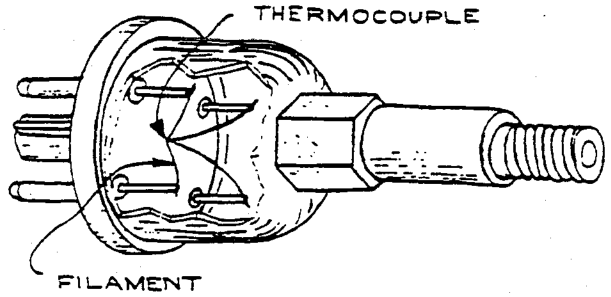


Figure 3.5 Thermocouple gauge.

3.3.3 Ionization Gauges

In the region of molecular flow, pressure is usually measured with an ion gauge. In this type of gauge, gas molecules are ionized by electron impact and the resulting positive ions are collected at a negatively biased electrode. The current to this electrode is a function of pressure. There are several types of ion gauges, differing primarily in the mechanism of electron production. The most common gauge is the thermionic or hot-cathode ionization gauge, shown schematically in Figure 3.6. Electrons from an electrically heated filament are accelerated through the gas toward a positively biased grid. Ions are collected at a central wire, and the positive ion current is measured by a sensitive electrometer. These gauges are useful in the 10^{-3}- to 10^{-9}-torr range. The indicated pressure depends upon the ionization cross section of the gas. When a gauge calibrated for air is used with hydrogen, it will read low by a factor of 3. With helium, it will be low by a factor of 8.

Ion gauges have the sometimes useful and sometimes detrimental property of functioning as pumps. This pumping action results from two mechanisms. Ions accelerated to and imbedded in the collector are effectively removed from the system. In addition, metal

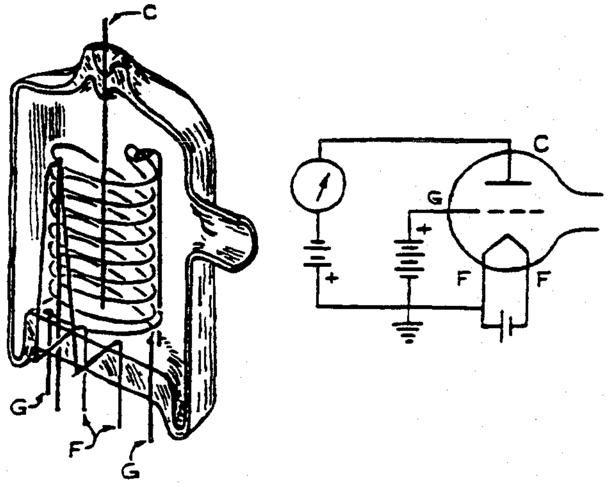


Figure 3.6 Bayard-Alpert type of thermionic ionization gauge.

evaporated from the filament deposits on the walls to produce a clean, chemically active surface that adsorbs nitrogen, oxygen, and water. At pressures below 10^{-7} torr this pumping action creates a pressure gradient, which causes the gauge to indicate a pressure different from the pressure in the system to which it is attached. On the other hand, it is possible to pump a small system solely with an ion gauge. Typically the pumping speed of an ion gauge is 0.2 liter sec^{-1} for N_2 when it is operated with an electron-emission current of 10 mA.

At low pressures the accuracy of an ion gauge is improved if the electrode surfaces and gauge walls are periodically outgassed by heating. With most commercial gauges this degassing is accomplished by heating the grid to incandescence by passing an a.c. current through it.

Other types of ionization gauges include the cold-cathode gauge (Penning gauge) and the Alphatron. In the first type a magnetically confined discharge is contained between two electrodes at room temperature. Molecules in the discharge are ionized and collected to produce a measurable current. These gauges are useful between 10^{-2} and 10^{-6} torr. In the Alphatron gauge the ionizing electrons are produced by a radium source. The Alphatron gauge operates between atmosphere and 10^{-3} torr.

3.3.4 Mass Spectrometers

Mass spectrometers may be used to determine the pa tial pressures of residual gases in a system and to detec leaks. Residual-gas analyzers (RGAs) may be of th magnetic-deflection, quadrupole, time-of-flight, r.f., o cycloidal type, but most commercial RGAs are of th first two types. Typically these devices cover the range of 1 to 200 amu and can detect partial pressures as low as 5×10^{-13} torr. Fixed-focus mass spectrometers set tc detect helium are widely used as sensitive leak detectors. Leaks are located by monitoring the helium concentration within a vacuum system while the exterior of the system is probed with a small jet of helium. The detection limit of these devices is of the order of 10^{-10} torr liter sec^{-1}.

The pressure ranges in which the various gauges are useful are shown in Figure 3.7.

3.4 VACUUM PUMPS

Vacuum pumps, like pressure gauges, operate in a limited pressure range. In general, a pump that operates in the viscous flow region will not operate in the molec-

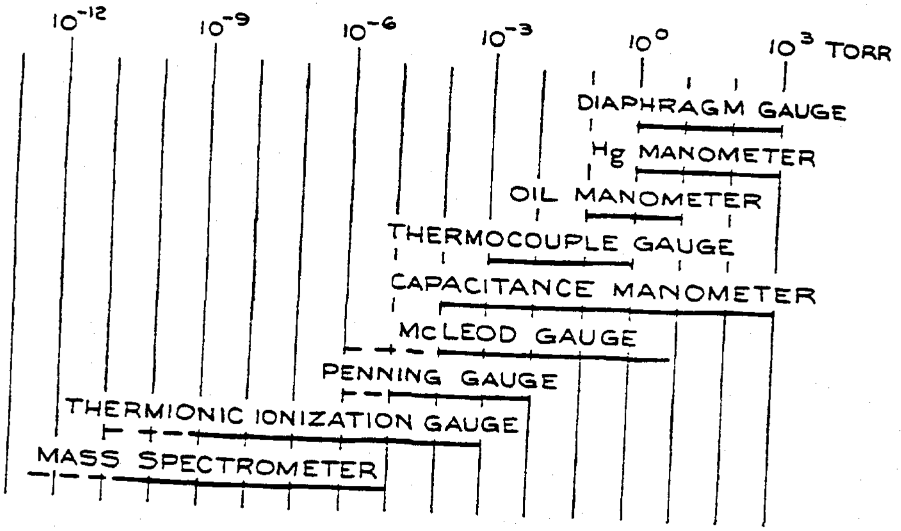


Figure 3.7 Operational pressure ranges of common gauges.

ular flow region and vice versa. The useful range of a pump is also limited by the vapor pressure of the materials of construction and the working fluids within the pump.

3.4.1 Mechanical Pumps

The pump most commonly used for attaining pressures down to a few millitorr is the oil-sealed rotary pump shown schematically in Figure 3.8. In this pump a rotor turns off-center within a cylindrical stator. The interior of the pump is divided into two volumes by spring-loaded vanes attached to the rotor. Gas from the pump inlet enters one of these volumes and is compressed and forced through a one-way valve to the exhaust. The seal between the vanes and the stator is maintained by a thin film of oil. The oil used in these pumps is a good-quality lubricating oil from which the high-vapor-pressure fraction has been removed. These pumps are also made in a two-stage version in which two pumps with rotors on a common shaft operate in series. Rotary pumps to be used for pumping condensable vapors are provided with a gas ballast. This is a valve that admits air to the compressed gas just prior to the exhaust cycle. This additional air causes the exhaust valve to open before the pressures of condensable vapors exceed their vapor pressure and thus prevents these vapors from condensing inside the pump.

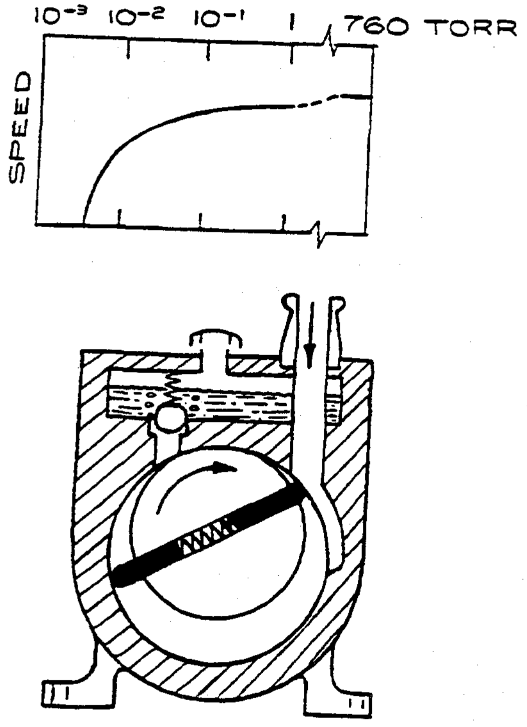


Figure 3.8 Two-vane, oil-sealed rotary pump.

Oil-sealed rotary pumps will operate well for years if the inner surfaces do not rust and the oil maintains its lubricating properties. It is wise to leave these pumps operating continuously so that the oil stays warm and dry. For storage a pump should be filled with new oil and the ports sealed. In use, an increase in the lowest attainable pressure (the base pressure) indicates that the oil has been contaminated with volatile materials. The dirty oil should be drained while the pump is warm. The pump should be filled with new oil, run for several minutes, drained, and refilled.

Rotary pumps are available with capacities of 1 to 500 litersec^{-1}. A single-stage pump is useful down to 50 mtorr, and a two-stage pump to 5 mtorr. Typical performance of a single-stage pump is indicated by the pumping-speed curve in Figure 3.8. With a two-stage pump, a base pressure of 10^{-4} torr can be achieved after a long pumping time if the back diffusion of oil vapor from the pump is suppressed by use of a sorption or liquid-air trap on the pump inlet. This is a simple scheme for evacuating small spectrometers and Dewar flasks or other vacuum-type thermal insulators.

The exhaust gases from an oil-sealed mechanical vacuum pump contain a mist of fine droplets of oil. This oil smoke is especially dense when the inlet pressure is in the 200- to 600-torr range. The oil droplets are extremely small, usually less than 5 microns. Over the course of time, this oil settles on the pump and its surroundings and collects dirt and grime. Furthermore, breathing the finely dispersed oil may injure the operator's lungs. Most pump manufacturers market filters to cope with this problem, and their use is recommended. An excellent system of coalescing-type filters is manufactured by Balston, Inc. These filters cause the droplets to accumulate into large drops, which run off into a sump at the bottom of the filter housing. Fittings are available to install these filters on most pumps. An alternative solution to the pump-exhaust problem, particularly when pumping toxic gases, is to vent the pump into a fume hood in the lab. Exhaust lines can be made of PVC drain pipe available from plumbing suppliers.

In order to achieve high pumping speeds in the 10- to 10^{-2}-torr region, a *Roots blower* can be used in series with an oil-sealed rotary pump. As illustrated in Figure 3.9, these pumps consist of a pair of counterrotating, two-lobed rotors on parallel shafts. Rotational speeds are about 3000 rpm. There is a clearance of a few thousandths of an inch between the rotors themselves and between the rotors and the housing. The roughing pump in series is required because there is no oil present in the Roots blower to attain a seal at high pressure. A Roots blower provides a compression ratio of the order of 10:1, and hence the speed required of the backing pump is correspondingly lower than that of the blower. Roots pumps are available with displacements of a few hundred to a few thousand liters per second.

There are now commercially available mechanical pumps, called *molecular-drag pumps*, which operate in the molecular flow regime. In these pumps one or more balanced rotors turn at 20,000 to 50,000 rpm within a slotted stator. The edge speed of the rotors approaches molecular velocities. When a molecule strikes a rotor, a significant component of momentum is transferred to the molecule in the direction of rotation. This transferred momentum causes molecules to move from the pump inlet toward the exhaust. The most common version of this pump is the *turbomolecular pump*, which, like a turbine, has a series of rotors with oblique radial slots turning between radially slotted stators. These pumps achieve a compression ratio of up to 10^6:1 provided the outlet pressure is kept below 100 mtorr. This requirement means that the turbomolecular pump must be run in a series with a conventional rotary pump, which is referred to as a *backing pump* or *forepump*. The forepump, pumping directly through the turbopump, can be used for initial roughing of a system from atmosphere to 10^{-1} torr. In this capacity the rotary pump is referred to as a *roughing pump*. Turbomolecular pumps are available with capacities of a few hundred to 10,000 liter sec^{-1}. They have the advantage over diffusion pumps of providing an oil-free and mercury-free vacuum. Their main disadvantage is cost. For comparable pumping speeds, turbomolecular pumps cost about ten times more than diffusion pumps.

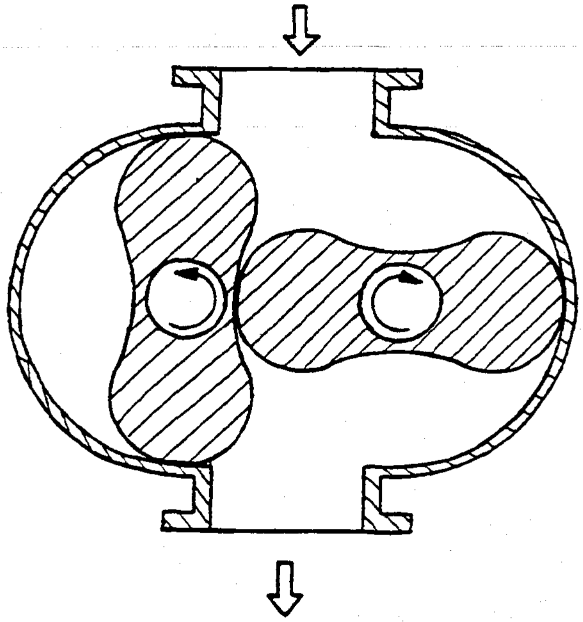


Figure 3.9 Roots blower.

3.4.2 Vapor Diffusion Pumps

In a diffusion pump, gas molecules are moved from inlet to outlet by momentum transfer from a directed stream of oil or mercury vapor. As shown in Figure 3.10, the working fluid is evaporated in an electrically heated boiler at the bottom of the pump. Vapor is conducted upward through a tower above the boiler to a nozzle or an array of nozzles from which the vapor is emitted in a jet directed downward and outward toward the pump walls. The walls of the pump are cooled so that molecules of the working fluid vapor condense before their motion is randomized by repeated collisions. The pump walls are usually water-cooled, although in some small pumps they are air-cooled. The condensate runs down the pump wall to return to the boiler.

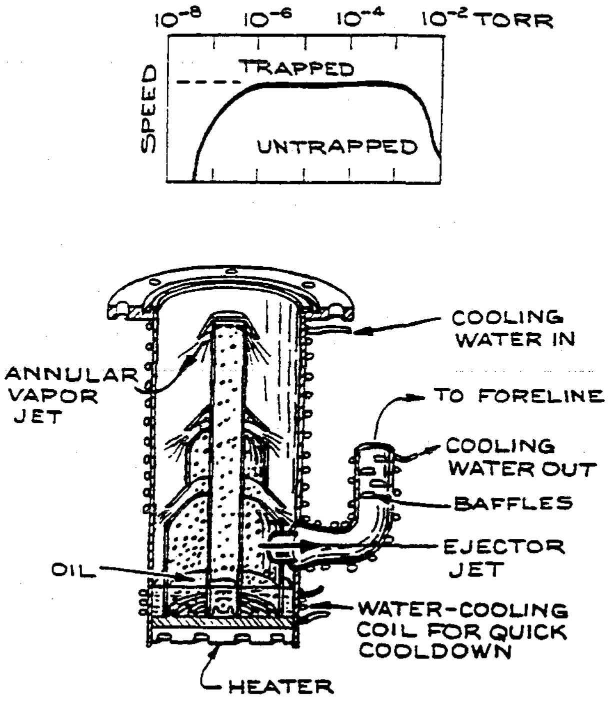


Figure 3.10 Diffusion pump.

As indicated by the pumping-speed curve in Figure 3.10, the pumping action of a diffusion pump begins to fail when the inlet pressure increases to the point where the mean free path is less than the distance from the vapor-jet nozzle to the wall. When this occurs the net downward momentum of vapor molecules is lost and the vapor begins to diffuse upward into the vacuum system. Diffusion pumping of a system may be initiated at 50 to 100 mtorr, but the system pressure should quickly fall below 1 mtorr or the system may become significantly contaminated with the vapor of the working fluid. Oil diffusion pumps can run against an outlet pressure of 300 to 500 mtorr, and mercury pumps can tolerate an outlet pressure of a few torr. Thus these pumps must be operated in series with a mechanical forepump. The pump speed is insensitive to foreline pressure up to some critical pressure. If this critical pressure is exceeded, the pump is said to stall. Stalling is a disaster, because hot pump-fluid vapor is flushed backwards through the pump into the system.

Oil diffusion pumps use low-vapor-pressure hydrocarbon or silicone oil as the working fluid. Hydrocarbon oils are subject to cracking and will oxidize if exposed to air when hot. Silicone oils are much less subject to chemical reaction and may be exposed to air when hot. Silicone oils are poor lubricants and thus should be used with a foreline trap that prevents their entering the backing pump. Many hydrocarbon oils have the advantage of being much less expensive than silicone oils. The absolute lowest pressure attainable with an untrapped diffusion pump is the room-temperature vapor pressure of the working fluid. The vapor pressures of some common pump oils are given in Table 3.2. For operation at pressures below 10^{-7} torr, and for a reasonably oil-free vacuum environment, a vapor trap must be placed immediately above a diffusion pump. Vapor traps tie up oil molecules by adsorption or condensation.

Mercury as a pump fluid has the advantage of being chemically inert and is used in systems where hydrocarbon contamination is unacceptable. Mercury diffusion pumps can tolerate inlet pressures ten times greater than the maximum inlet pressure tolerated by oil pumps. In addition the critical foreline pressure for a mercury pump is much higher. However, the room-temperature vapor pressure of mercury is about 10^{-3} torr, and thus an inlet cold trap is required to condense mercury vapor in order to achieve system pressures below 10^{-3} torr. Mercury amalgamates with many metals and alloys, particularly brass. Mercury vapor is toxic, and care must be taken to properly trap and vent the backing-pump exhaust.

To attain pressures in the 10^{-3}- to 10^{-9}-torr region, diffusion pumps are the simplest and least expensive route. Pumps with speeds of 50 to 50,000 liter sec^{-1} and with nominal inlet port diameters of 1 to 35 in. are available. Diffusion pumps are constructed of Pyrex glass, mild steel, or stainless steel. The jets and towers of pumps with steel barrels are aluminum. The choice of glass or steel usually depends upon the material that is used to construct the vacuum container, although it is also relatively simple to mate glass to steel with an elastomer gasket or through a metal-to-glass seal. For those systems that are intended to handle reactive gases, glass is the preferred material. However, glass pumps are available only in small sizes.

Table 3.2 PROPERTIES OF DIFFUSION-PUMP FLUIDS

Name (*chemical composition*)	*Boiling Point at 1 Torr* (°C)	*Approximate Room-Temperature Vapor Pressure* (torr)
(Mercury)	120	2×10^{-3}
(Di-*n*-butyl phthalate)	140	2×10^{-5}
Octoil (di-2-ethyl hexyl phthalate)	200	3×10^{-7}
Octoil-S (di-2-ethyl hexyl sebacate)	210	3×10^{-8}
Convoil-10 (saturated hydrocarbon)	150	10^{-4}
Convoil-20 (saturated hydrocarbon)	190	5×10^{-7}
Convalex-10 (polyphenyl ether)	280	3×10^{-9}
Neovac Sy (alkyldiphenyl ether)	240	10^{-8}
D.C. 702 (silicone)	180	5×10^{-7}
D.C. 704 (silicone)	210	6×10^{-8}
D.C. 705 (silicone)	250	10^{-9}

3.4.3 Sorption Pumps, Getter Pumps, Cryopumps, and Ion Pumps

A variety of vacuum pumps remove gas from a system by chemically or physically tying up molecules on a surface or by trapping them in the interior of a solid. Two of the principal advantages of these pumps are that they require no backing pump and they contain no fluids to contaminate the vacuum.

The simplest of this class of pumps is the *sorption pump*, illustrated in Figure 3.11. The sorbent material is activated charcoal or one of the synthetic zeolite materials known as molecular sieves. These materials are effective sorbents partly because of their huge surface area, which is of the order of thousands of square meters per gram. The most common molecular sieves are Linde 5A or 13X. The number in the sieve code specifies the pore size: 5A is preferred for air pumping, while 13X is used for trapping hydrocarbons. All of these materials will pump water and hydrocarbon vapors at room temperature, but they must be cooled to liquid-nitrogen temperature to absorb air. Sorbent materials do not trap hydrogen or helium at liquid-nitrogen temperature. In some cases, therefore, the pressure of hydrogen or helium in a system will establish the lowest attainable pressure. Sorption pumps must be provided with a poppet valve because the sorbent material releases all of its absorbed air as it warms to room temperature.

The sorbent is initially activated by baking to 300°C. After several pumping cycles the pores of the sorbent material will become clogged with water and the ef-

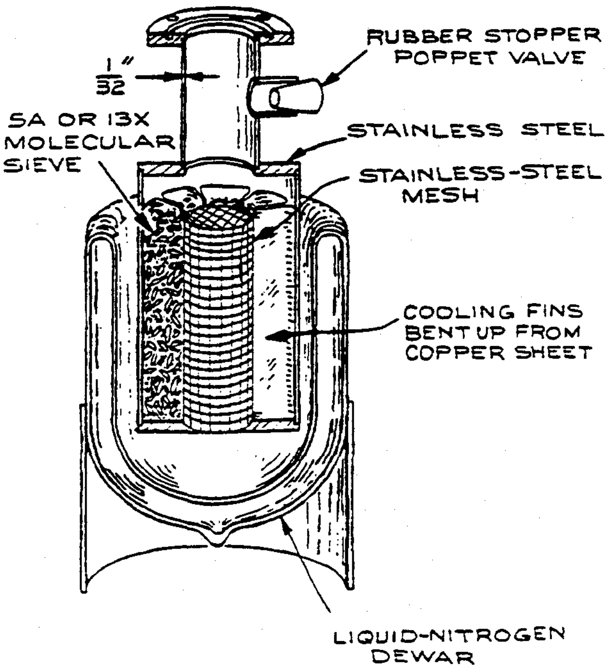

RUBBER STOPPER POPPET VALVE
1"/32
5A OR 13X MOLECULAR SIEVE
STAINLESS STEEL
STAINLESS-STEEL MESH
COOLING FINS BENT UP FROM COPPER SHEET
LIQUID-NITROGEN DEWAR

Figure 3.11 Sorption pump.

ficiency of the pump will deteriorate. Water is removed by baking the pump to 300°C with the poppet valve open and then cooling to room temperature with the valve closed so that moisture from the room is not reabsorbed. Baking can be accomplished by wrapping the pump with heating tape and covering with fiberglass insulation. Custom-made heating mantles can be obtained inexpensively from the Glas-Col Company.

In a well-designed pump, 50 g of dry Linde 5A will pump a 1-liter volume from atmosphere to less than 10^{-2} torr in about 20 minutes. The pump must be designed to provide good thermal contact between the sieve and the coolant, or the maximum pumping speed will not be achieved. In addition, the inlet tube should be as thin as possible to minimize heat conduction into the pump. Pressures below 10^{-6} may be attained in a system that has been roughed down. A simple way to achieve low pressures with sorption pumping is to use several pumps with appropriate valving so that one pump is used as a rough pump and successive pumps are used at lower pressures.

Clean surfaces of refractory metals such as titanium, molybdenum, tantalum, or zirconium will pump N_2, O_2, CO_2, H_2O, and CO by chemisorption. This process is called *gettering*. In a *getter pump*, the active metal surface is produced *in vacuo*. As in Figure 3.12, a simple pump can be made by wrapping titanium wire around a tungsten heater filament contained in a glass bulb. The filament is electrically heated to evaporate the titanium, which in turn condenses on the walls of the bulb. This pump operates effectively between 10^{-3} and 10^{-11} torr. The titanium surface must be renewed at a rate roughly equal to the rate at which a monolayer of gas is adsorbed. Down to 10^{-7} torr the filament is continuously heated. Below 10^{-7} torr the titanium need only be deposited periodically. The capacity of a titanium sublimation pump is about 30 torr·liter per gram of Ti. A forepump is not required, but a mechanical or sorption roughing pump is needed to reach a starting pressure of 10^{-2} to 10^{-3} torr.

Saturated hydrocarbons and rare gases are not adsorbed at room temperature. However, if a getter is cooled to the temperature of liquid nitrogen, hydrocarbons and argon are adsorbed. A titanium sublimation pump can be used to achieve an ultrahigh vacuum if it is used in conjunction with a small ion pump tl removes rare gases, or if, as will be described in a la section, it is used to pump a system that has be purged of inert gases prior to evacuation.

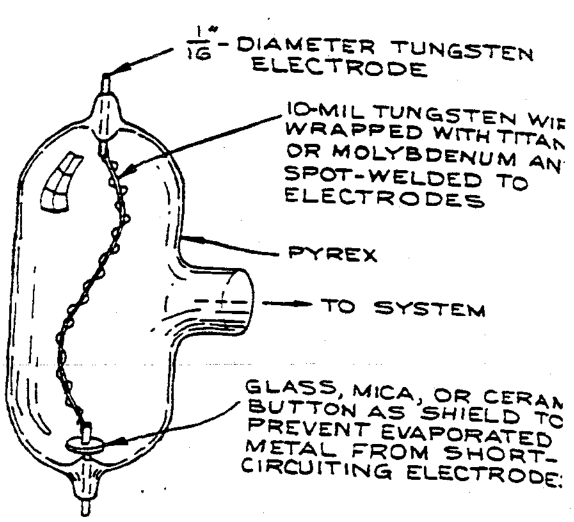

1" 16 - DIAMETER TUNGSTEN ELECTRODE
10-MIL TUNGSTEN WIR WRAPPED WITH TITAN OR MOLYBDENUM AN SPOT-WELDED TO ELECTRODES
PYREX
TO SYSTEM
GLASS, MICA, OR CERAM BUTTON AS SHIELD TO PREVENT EVAPORATED METAL FROM SHORT-CIRCUITING ELECTRODE

Figure 3.12 Simple titanium sublimation pump.

Any surface will act as a pump for a gas that co denses at the temperature of the surface. A pump th relies primarily upon condensation on a cold surface called a *cryopump*. Commercial versions of these pum incorporate a closed-circuit helium refrigerator to co the active surfaces. Working temperatures are typical below 20 K. These pumps usually employ two stages. I the first stage a metal surface is maintained at 30 to 5 K to trap water vapor, carbon dioxide, and the maj components of air. The second stage, maintained at 1 to 20 K, is coated with a cryosorbent material such a charcoal to provide pumping of neon, hydrogen, an helium. Cryopumps provide high pumping speed for th easily condensed gases. Their performance with heliun depends critically upon the quality and recent history o the cryosorbent surface.

An *ion pump* combines getter pumping with the pumping action exhibited by an ionization gauge. Within these pumps a magnetically confined discharge is maintained between a stainless-steel anode and a titanium cathode. The discharge is initiated by field emission when a potential of about 7 kV is placed across the electrodes. After the discharge is struck, a current-limited power supply maintains the discharge rate at 0.2 to 1.5 A, depending on the size of the pump. Inert-gas molecules, and other molecules as well, are ionized in the discharge and accelerated into the cathode with sufficient kinetic energy that they are permanently buried. Active gases are chemisorbed by titanium that has been sputtered off the cathode by ion bombardment and deposited on the anode. Ion pumps operate between 10^{-2} and 10^{-11} torr. Pumps with speeds from one to many thousands of liters per second are available.

Ion pumps require no cooling water or backing pump. They continue pumping by getter action even if the power fails. Ion pumps do not introduce hydrocarbon or mercury vapors into the vacuum. The positive-ion current to the cathode of an ion pump is a function of pressure; thus the pump serves as its own pressure gauge. Ion pumps cannot be used where stray electric and magnetic fields are unacceptable. The primary disadvantage of ion pumps is their cost. The price of an ion pump with its control unit is nearly an order of magnitude greater than for a comparable diffusion pump. However, on small ultrahigh-vacuum systems, an ion pump may compete economically with other pumps, since the pump obviates the need for a pressure gauge. The cost of a 1- to 5-liter sec^{-1} ion pump with a control unit that has a pressure readout is little more than the cost of an ion gauge and controller.

3.5 VACUUM HARDWARE

3.5.1 Materials

Borosilicate glasses such as Pyrex or Kimax glass are particularly well suited for the construction of small laboratory vacuum systems. These glasses are chemically inert and have a low coefficient of thermal expansion. Because of the plasticity of the material, complicated shapes are easily formed. Glass vacuum systems can be constructed and modified *in situ* by a moderately competent glassblower. The finished product does not have to be cleaned after working.

Hard glass tubing and glass vacuum accessories ar inexpensive. Stopcocks and Teflon-sealed vacuum valve: ball joints, taper joints, and O-ring-sealed joints; traps and diffusion pumps are all easily obtained at low cost In most cases only the glassblowing ability to mak straight butt joints and T-seals is required to make ; complete system from standard glass accessories.

Glass pipe and pipe fittings are manufactured for th chemical industry. A variety of standard shapes such a: elbows, tees, and crosses are available in pipe diameter: of $\frac{1}{2}$ to 6 in. Glass process pipe with Teflon seals can b used to make inexpensive vacuum equipment for us down to 10^{-8} torr. Couplings are also available fo joining glass pipe to metal pipe and to standard meta flanges and pipe fittings, and it is therefore easy t make a system that combines glass and metal vacuum accessories.

Glass and metal parts can be mated through a graded glass seal. Typically a graded seal appears to be simply a section of glass tubing butt-sealed to a section of Kovar metal tube. In fact, the glass tubing consists of a series of short pieces of glass whose coefficients of thermal expansion vary in small increments from that of hard glass to that of the metal. Graded seals are useful for joining glass accessories such as ion-gauge tubes to metal systems. Large, high-throughput glass stopcocks are not readily available. However, by using graded seals, large metal vacuum valves can be inserted into a glass vacuum line. Inexpensive graded seals in sizes up to 2 in. in diameter are available from commercial sources.

Brass and copper are useful vacuum materials. Brass has the advantage of being easily machined, and brass parts can be joined by either soft solder or silver solder. Unfortunately, brass contains a large percentage of zinc, whose volatility limits the use of brass to pressures above about 10^{-6} torr. Heating brass causes it to lose zinc quite rapidly. The vapor pressure of zinc is 10^{-5} torr at 200°C, and 2×10^{-3} torr at 300°C. Forelines for diffusion pumps are conveniently constructed of brass or copper tubing and standard plumbing elbows and tees. Ordinary copper water pipe is acceptable for forelines and other rough vacuum applications; however oxygen-free high-conductivity (OFHC) copper should be used for high temperature and high vacuum work. Copper can be heliarc-welded by a skilled technician.

The most desirable metal for the construction of high-vacuum apparatus is type-304 stainless steel. This material is strong, reasonably easy to machine, bakeable, and easy to clean after fabrication. Stainless-steel parts may be brazed or silver-soldered. Low-melting (230°C) silver-tin solder such as StainTin 157 PA (Eutectic Welding Alloys Co.) is useful for joining stainless-steel parts in the laboratory. However, it is best if stainless parts are fused together by arc welding using a nonconsumable tungsten electrode in an argon atmosphere. This process, known commonly as heliarc fusion welding or tungsten–inert-gas (TIG) welding, produces a very strong joint, and, because no flux or welding rod is used, such a joint is easily cleaned after welding. Most machine shops are prepared to do heliarc welding on a routine basis.

Type-304 stainless is widely used in the milk- and food-processing industry. As a result many stock shapes are commercially available at low cost. Elbows, tees, crosses, Y's, and many other fittings in sizes up to at least 6-in. diameter may be purchased from the Ladish Company or Alloy Products Company.

Many stainless steels are nearly nonmagnetic. The field produced by a piece of type 304 after machining is on the order of 10–30 milligauss at a distance of 1 cm. This magnetism can be reduced to 1 *milligauss* by quickly heating the material to 1100°C after machining and quenching in water.

Aluminum alloys, particularly the 6000 series, are used for vacuum apparatus. Aluminum has the advantage over stainless steel of being lighter and stiffer per unit weight, and much easier to machine. It is completely nonmagnetic. The chief disadvantages of aluminum stem from its porosity and the oxide layer that covers the surface. The rate of outgassing from aluminum is five to ten times greater than for stainless steel. Welds in aluminum are not as reliable as welds in stainless, and they tend to outgas volatile materials that are occluded in the weld. However, welded aluminum vacuum containers can be used down to about 10^{-7} torr. The hard oxide layer that forms instantly on a clean aluminum surface is an electrical insulator. This insulating surface tends to collect an electrical charge, which may be undesirable in some applications. Stray fields resulting from this charge may be eliminated by having critical aluminum parts copper- or gold-plated after fabrication.

Many plastics may be used at pressures down to 10^{-7} or 10^{-8} torr. The use of these materials is limited to varying degrees because they outgas air and plasticizers, and because they cannot be heated to high temperatures. Fluorocarbon polymers such as Teflon, Kel-F, and Viton-A have relatively low outgassing rates. Teflon can withstand temperatures up to 250°C, and its outgassing rate falls well below 10^{-8} torr liter sec^{-1}cm^{-2} after an initial pumpdown of a day at 100°C. Unfortunately, Teflon is relatively soft, and it cold-flows under mechanical pressure. Nylon, Delrin, and Vespel (polyimide) are harder than Teflon, and because they are self-lubricating they are useful as bearing surfaces. Delrin is preferred to nylon, since nylon is quite hygroscopic and outgasses water vapor after each exposure to air. Polyimide is somewhat hygroscopic as well, but may be baked to 250–300°C. Windows in vacuum systems may be made of acrylic plastic such as Plexiglas or Lucite. General Electric's Lexan is a very strong, tough, and machinable plastic well suited for use as a structural material or as an electrical insulator.

A variety of low-vapor-pressure sealers and adhesives have vacuum applications. Apiezon M grease and Dow-Corning silicone high-vacuum grease are used to seal stopcocks and ground-glass taper joints and, in some instances, as low-speed lubricants. Apiezon W black wax is useful for sealing windows to the ends of glass or metal tubes. This material melts at 60°C and has a room-temperature vapor pressure of about 10^{-6} torr. Glyptal is another low-vapor-pressure sealer and adhesive material.

Epoxy resins are particularly useful. Epoxy cements consist of a resin and a catalytic hardener, which are combined immediately before use. The proportions of resin and hardener and subsequent curing must be carefully controlled to prevent excessive outgassing from the hardened material. Small epoxy kits with resin and catalyst prepackaged in the correct proportions are available (e.g., from Tracon). Epoxy formulations with a range of flexibilities can be obtained, as well as ones with high electrical or thermal conductivity.

3.5.2 Demountable Vacuum Connections

Vacuum systems require detachable joints for convenience in assembling and servicing. There are a tremendous variety of vacuum connections, but demountable parts are most commonly joined by mating flanges or pipe threads that are sealed with some elastic material.

Pipe threads may be reliably sealed with Teflon thread dope that is sold in hardware stores for sealing threads in water pipes. This sealer consists of a thin Teflon tape that is stretched over the male thread before assembly. The tape must be replaced upon reassembly.

For pressures down to about 10^{-7} torr, vacuum connections are usually sealed with rubber O-rings. Several O-ring-sealed joints are illustrated in Figure 3.13. O-rings are circular gaskets with a round cross section. They are available in hundreds of sizes from 0.125-in. i.d. (inner diameter) with a cord diameter of 0.070 in. (nominally $\frac{1}{16}$ in.) to 2-ft i.d. with a cord diameter of 0.275 in. (nominally $\frac{1}{4}$ in.). Very large rings may be made from lengths of cord stock with the ends butted together and glued with Eastman 910 adhesive. O-rings are made of a variety of elastomers. The most common are Buna-N, a synthetic rubber, and Viton-A, a fluorocarbon polymer. Buna-N may be heated to 80°C and will not take a set after long periods of compression. Unfortunately this material outgasses badly, particularly after exposure to cleaning solvents. Viton-A has a low rate of outgassing and will withstand temperatures up to 250°C. Viton-A will take a set after baking. Generally, the advantages of Viton-A O-rings offset their higher cost.

O-ring-sealed flanges and "quick connects" are easily fabricated, and they are also available ready-made. Mating flanges have a groove that contains the ring after the flanges have been pulled into contact with one another. Usually, one flange is flat and the groove is cut into the mating flange, but flanges can be made sexless by cutting a groove of half the required depth in both flanges. The cross section of the groove should be about 10% greater than the cross section of the O-ring cord, since rubber is deformable but incompressible. The groove depth should be about 70% of the cord diameter for static seals and 80% of the cord diameter for dynamic seals. The i.d. of the groove should match the O-ring i.d. It is sometimes convenient to undercut the sides of the groove slightly to give a closed dovetail cross section rather than a rectangular cross section, so that the ring is retained in the groove during assembly.

For rings up to a foot in diameter, the cord diameter should be $\frac{1}{8}$ in. or less. Choose a ring to fit in a groove that is as close to the flange i.d. as possible, in order to minimize the amount of gas trapped in the narrow space between the mated flanges. The bolts or clamps that pull the flanges together should be as close to the groove as possible to prevent flange distortion.

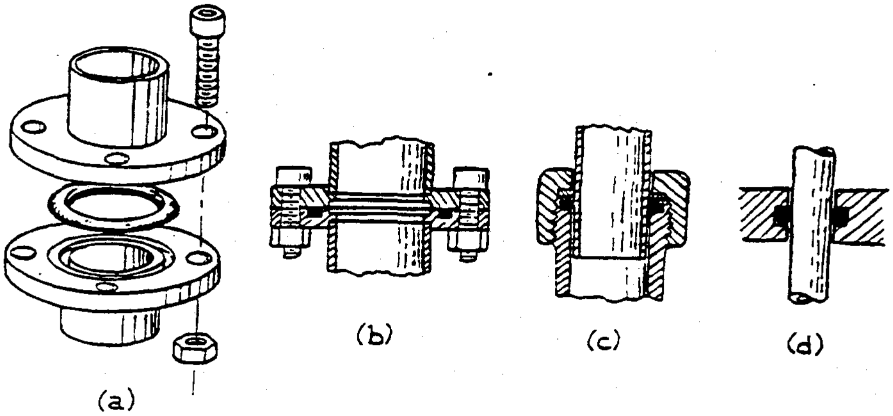


Figure 3.13 O-ring-sealed vacuum connections: (*a*) an exploded view of an O-ring-sealed flange joint; (*b*) a flange joint; (*c*) a quick connect; (*d*) a rotating-shaft seal.

For static seals, O-rings should be used dry. Before assembly, the groove should be cleaned and the ring wiped free of mold powder with a lintless cloth. O-rings should not be cleaned with solvents. For rotating seals a very light film of vacuum grease on the ring will prevent abrasion. Occasionally a film of grease may be required on a static O-ring to help make a seal on an irregular surface.

Metal sealing materials are required for ultrahigh-vacuum (UHV) work. Elastomers are unacceptable for this because of their high vapor pressure and because they cannot be baked to high temperatures. The most common UHV seal consists of a flat OFHC copper gasket trapped between knife edges on the faces of a pair of flanges as shown in Figure 3.14. These flanges and gaskets are available from a number of manufacturers of vacuum hardware. A reliable UHV seal can be made by using a gold wire O-ring in a groove in the same manner as elastomer O-rings are used. The O-ring is made by butt-welding the ends of an appropriate length of gold wire and then annealing the whole ring by heating it to a dull red with a cold flame and permitting it to cool in air. After fabrication the ring must be handled carefully because it is easily stretched. Metal seals can be only used once, since they are permanently deformed during installation. The metal is recoverable and may be recycled.

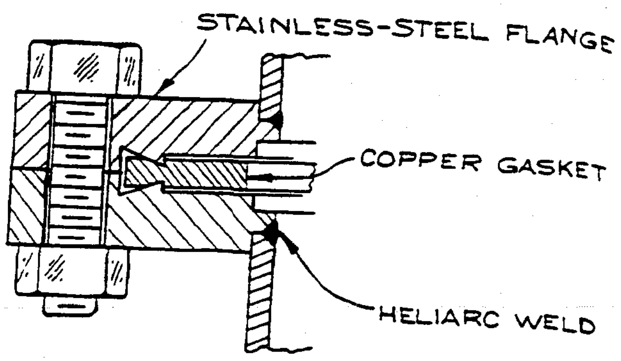


Figure 3.14 Detail of a bakeable, ultrahigh-vacuum flange seal.

3.5.3 Valves

A wide variety of valves are available commercially for use in glass systems and in high-vacuum and ultrahigh-vacuum metal systems. Generally, vacuum valves are sufficiently complex that it is uneconomical for the laboratory scientist to undertake their fabrication.

The two most common glass valves are illustrated in Figure 3.15. One (*a*) is a vacuum stopcock consisting of a tapered glass plug which is fitted into a glass body by lapping. The mating parts are sealed with a film of high-vacuum grease. As shown, these stopcocks are designed so that atmospheric pressure forces the plug into the body of the valve. The other, more modern type of glass valve (*b*) has a Teflon plug threaded into a glass body. The plug is sealed to the body with an O-ring. It is generally advisable to apply a light film of vacuum grease to the O-ring and the threads of these valves.

Several metal vacuum valves are illustrated in Figure 3.16. The bellows-sealed valve in Figure 3.16(*a*) is usually constructed of brass or stainless steel and is available in sizes suitable for use on tubing of $\frac{3}{8}$- (outer diameter), to $1\frac{5}{8}$-in. o.d. The drive screw that moves the sealing plate is contained within a bronze or stainless-steel bellows that *maintains* a vacuum seal while the plate moves from the open to the closed position. This type of valve is commonly used in the foreline of a diffusion pump, but it may also be used in small high-

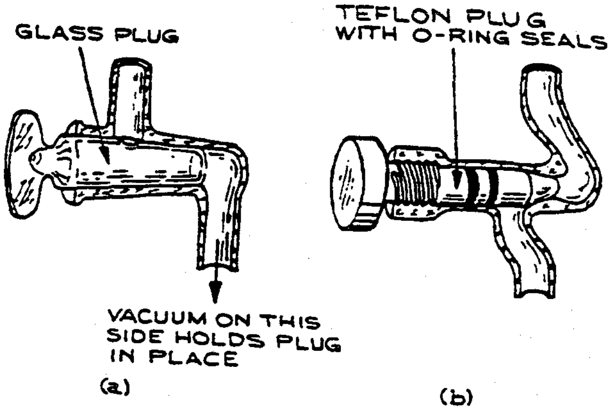

GLASS PLUG
TEFLON PLUG WITH O-RING SEALS
VACUUM ON THIS SIDE HOLDS PLUG IN PLACE
(a)
(b)

Figure 3.15 Glass vacuum valves: (*a*) valve with glass stopcock; (*b*) valve with Teflon plug.

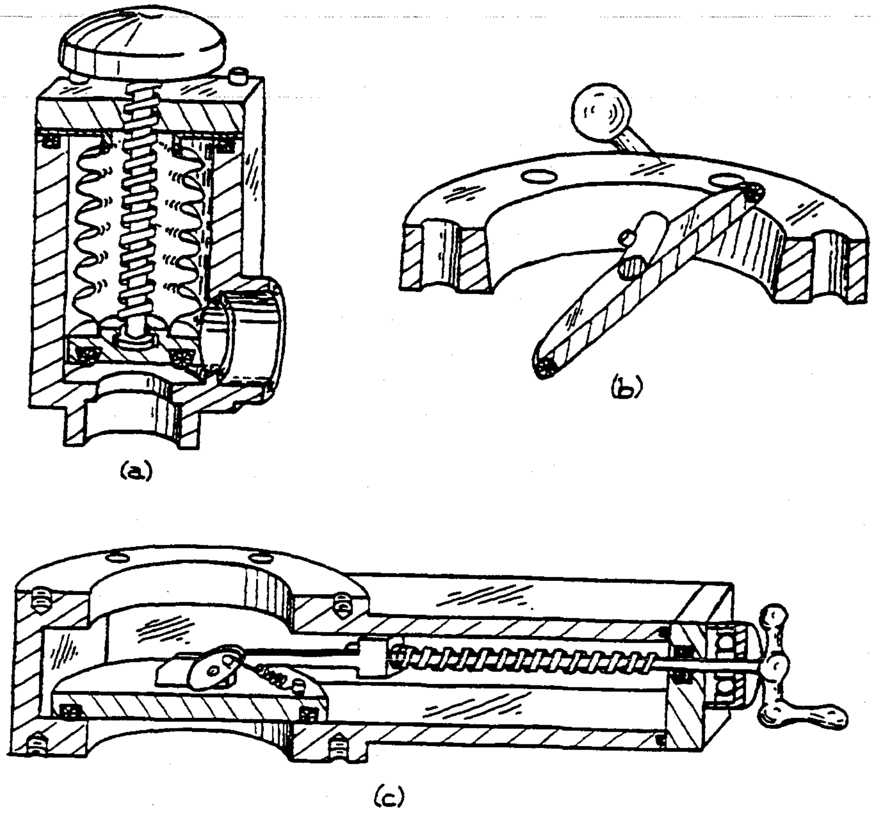


Figure 3.16 Metal high-vacuum valves: (*a*) bellows-sealed foreline valve; (*b*) quarter-swing gate valve; (*c*) sliding-gate valve.

vacuum systems. Stainless-steel, bellows-sealed valves with Viton O-rings can be heated to 200°C. Those with polyimide O-rings can be baked at 300°C.

The gate valves in Figure 3.16(*b*) and (*c*) are most often used to isolate a diffusion pump or ion pump from a high-vacuum chamber. These valves have very high conductance because of their low profile and because the sealing plate or gate does not significantly obstruct the valve aperture when the valve is open. Gate valves will seal against atmospheric pressure in either direction, but when used to isolate a pump they are usually installed so that atmospheric pressure in the chamber tends to hold them closed. The actuator may be either bellows-sealed or O-ring-sealed. The bellows seal is preferred. The body of a gate valve is cast from aluminum or stainless steel. They are available with nominal apertures of 2 to 10 in. to match the inlet apertures of most diffusion pumps.

3.5.4 Mechanical Motion in the Vacuum System

It is frequently necessary to transmit linear or rotary motion through the wall of a vacuum container. A simple and inexpensive linear-motion feedthrough can be made from the actuator mechanism of a small bellows-sealed vacuum valve. As shown in Figure 3.17, the valve body is truncated above the seat, a mounting flange is brazed to the body, and a fixture is brazed or screwed to the valve plate for attaching the mechanism to be driven inside the vacuum. Linear motion can be

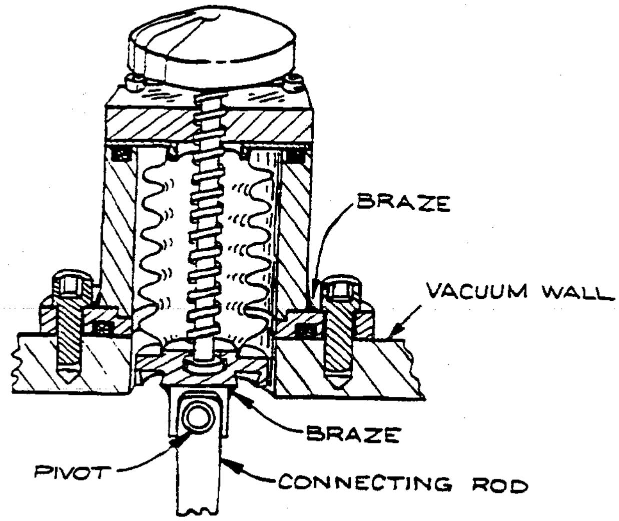


Figure 3.17 A bellows-sealed valve [Figure 3.16 (*a*)] converted to a linear-motion feedthrough.

converted to rotary motion inside the vacuum by applying the linear motion to a crank through a pivoting connecting rod. This scheme will provide up to 180° of rotary motion. Full 360° rotation may be achieved by means of a system of springs that carry the crank over center as shown in Figure 3.18.

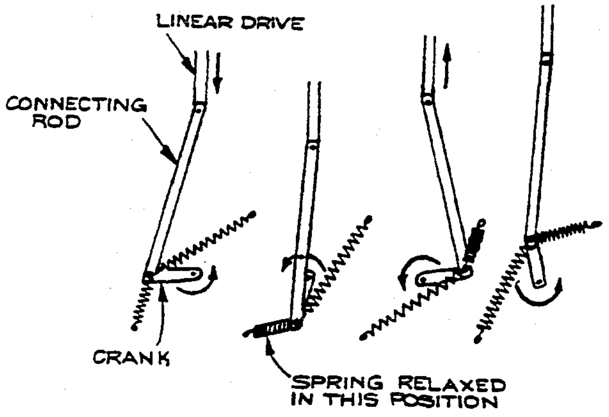


Figure 3.18 Spring arrangement to permit 360° rotation of a crank driven by a linear motion.

Rotary motion at speeds up to 100 rpm may be transmitted through an O-ring-sealed shaft as in Figure 3.13(*d*). These seals are inexpensive but may fail if the O-ring becomes abraded. In addition, lubricants and the elastomer O-ring material are exposed to the vacuum.

Rotary motion may be transmitted through a bellows-sealed wobble drive of the type shown in Figure 3.19. These drives are available commercially.

Moving parts can be magnetically coupled through a vacuum wall. All that is required is that a magnet be attached to the driving element and that the driven element be made of a magnetic material such as iron or nickel. Of course, the vacuum wall must be made of a nonmagnetic material such as Pyrex, brass, or type-304 stainless steel.

Metal surfaces become very clean *in vacuo*, particularly after baking, and metals in close contact tend to cold-weld. Because of this, unlubricated bearing surfaces within a vacuum system often become very rough after only a little use. There are a number of methods of improving bearing performance in a vacuum system without introducing high-vapor-pressure oils into the vacuum.

The tendency for a bearing to gall is reduced if the two mating bearing surfaces are made of different metals. For example, a steel shaft rotating without lubrication in a brass or bronze journal will hold up better than in a steel bushing. A solid lubricant may be applied to one of the bearing surfaces. Silver, lead-indium, and molybdenum disulfide have been used for this purpose. Graphite does not lubricate in a vacuum. MoS_2 is probably best. The lubricant should be burnished into the bearing surface. The part to be lubricated is placed in a lathe. As the part turns, the lubricant is applied and rubbed into the surface with the rounded end of a hardwood stick. By this means, the lubricant is forced into the pores. After burnishing, the surface should be wiped free of loose lubricant.

One component of a bearing may be fabricated of a self-lubricating material such as nylon, Delrin, Teflon, or polyimide. Teflon is good for this purpose, but its propensity to cold-flow will cause the bearing to become sloppy with time. Polyimide can be used in ultrahigh vacuum after baking to 250–300°C.

Brown, Sowinski, and Pertel[6] have overcome the

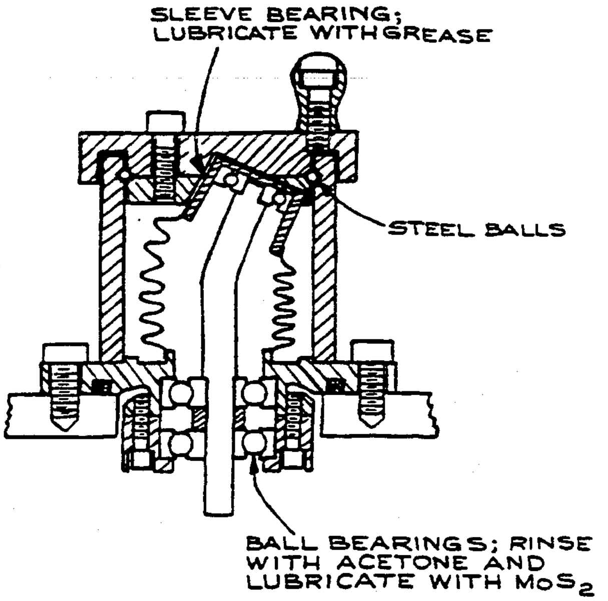

SLEEVE BEARING; LUBRICATE WITH GREASE
STEEL BALLS
BALL BEARINGS; RINSE WITH ACETONE AND LUBRICATE WITH MoS_2

Figure 3.19 A bellows-sealed, wobble-drive, rotary-motion feedthrough.

cold-flow problem in the design of a drive screw for use in a vacuum. Both the screw and its nut are steel, and lubrication is accomplished by placing a Teflon key in a slot cut in the side of the screw. Teflon is then continuously wiped onto the screw threads as the screw turns.

For very precise location of rotating parts and for high rotational speeds, ball bearings are required. Precision, very low-speed ball bearings that are bakeable can be made using sapphire balls running in a stainless-steel race. Inexpensive precision sapphire balls are available from Industrial Tectonics. A ball retainer is required to prevent the balls from rubbing against one another. As explained in Section 1.6.1, the race must be designed so that the balls rotate without slipping against the race surface. It is helpful to burnish the race with MoS_2.

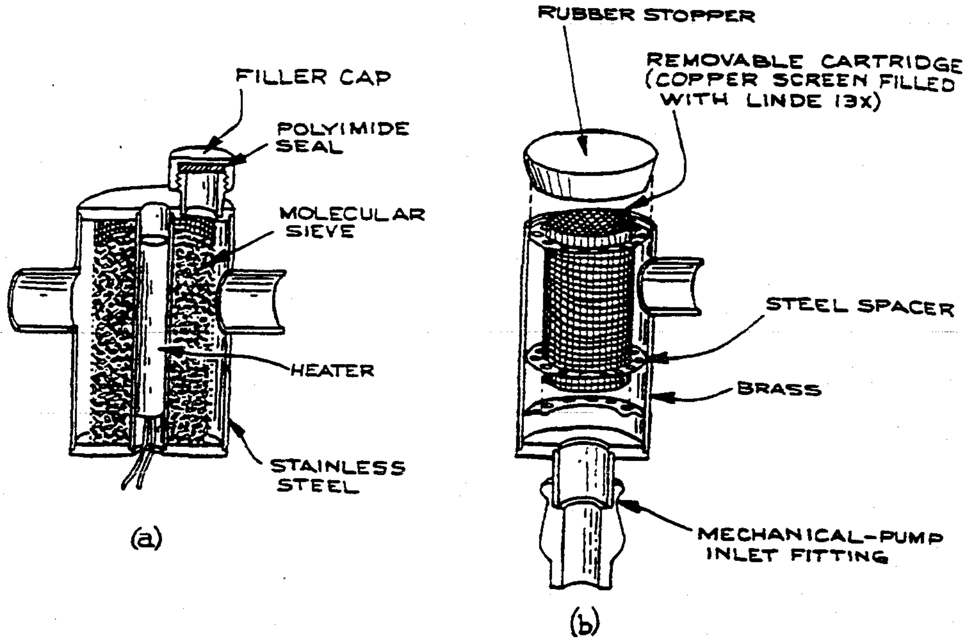
For high-speed applications a fluid lubricant is necessary. The following process has been developed at NASA-Goddard Space Flight Center.[7] Purchase high-quality stainless-steel ball bearings with side shields and phenolic ball retainers, such as the New Hampshire PPT series or Barden SST3 series. Remove the shields and leach clean the phenolic by boiling the bearings in chloroform-acetone, and then vacuum-dry at 100°C. The bearings are then lubricated by impregnating the phenolic with Du Pont Krytox 143AZ, a fluorinated hydrocarbon. Impregnation is accomplished by immersing the bearing in the lubricant and heating to 100°C at a pressure of 1 torr or less until air stops bubbling out of the phenolic. After this process the bearing must be wiped almost dry of lubricant. The Texwipe Company makes ultraclean foam cubes for this process. The amount of lubricant in the bearing is determined by weighing before and after impregnation. About 25 mg of Krytox is required for an R4 ($\frac{1}{4}$-in.) bearing, and about 50 mg is needed for an R8 ($\frac{1}{2}$-in.) bearing. This lubrication process should be carried out in a very clean environment, and the bearing should be inspected under a microscope for cleanliness before the side shields are replaced. Bearings treated in this manner have been run at speeds up to 60,000 rpm *in vacuo*.

3.5.5 Traps and Baffles

Traps are used in vacuum systems to intercept condensable vapors by means of chemisorption or physical condensation. Most high-vacuum systems have a trap in the foreline to prevent mechanical pump oil from backstreaming from the forepump to the diffusion pump. In addition, a trap is usually placed between a diffusion pump and a vacuum chamber to pump water vapor and to remove diffusion-pump fluid vapors that migrate backwards from the pump toward the chamber.

Foreline traps are similar in design to the molecular-sieve sorption pumps previously described, except, of course, that a trap must have both an inlet and an outlet. Two simple traps filled with 13X molecular sieve[8] are illustrated in Figure 3.20. The molecular sieve must initially be activated by baking to 300°C for several hours. In use the sieve material is regenerated at intervals of about a month by baking at 100°C to drive out absorbed oil and water. The baking may be done at atmospheric pressure or under a rough vacuum. If the baking is done in atmosphere, the sieve should be permitted to cool down from 100°C in vacuum to prevent reabsorption of water. If the baking is done *in*

Figure 3.20 Two designs for a molecular-sieve foreline trap.



situ, as would be the case for the trap in Figure 3.20(*a*), the trap should be isolated from the diffusion pump with a valve, and the line between the trap and the mechanical pump should be warmed to prevent desorbed vapors from condensing in the foreline. Also, the gas ballast of the pump should be opened while the sieve is baking to prevent water vapor from condensing in the pump.

The lowest pressure attainable with a two-stage mechanical pump is largely determined by the vapor pressure of the oil in the pump. By using a molecular-sieve trap in series with a mechanical pump to remove oil vapor it is possible to achieve pressures as low as 10^{-4} torr in a small system without using a diffusion pump.

A molecular sieve may also be used in a high-vacuum trap over the inlet of an oil diffusion pump. As shown in Figure 3.21, these traps are designed to be optically opaque so that a molecule cannot pass through the trap in a straight line. This precaution is necessary because this type of trap is intended for use at pressures where the mean free path of a molecule is very long. To ensure high conductance, the inlet and outlet ports should have the same cross-sectional area as the inlet of the attached pump. Also, the cross section perpendicular to the flow path through the trap (indicated by an arrow in Figure 3.21) should be at least as large as that of the inlet and outlet ports. The molecular sieve is activated by baking under vacuum to 300°C for at least six hours. The trap may be heated with heating tape covered with fiber-glass batting for insulation. Alternatively, a custom-made

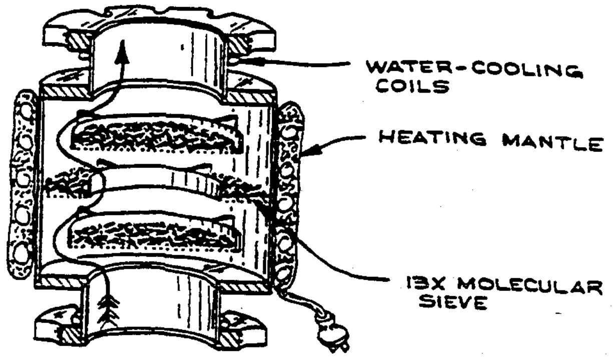


Figure 3.21 High-vacuum molecular-sieve trap.

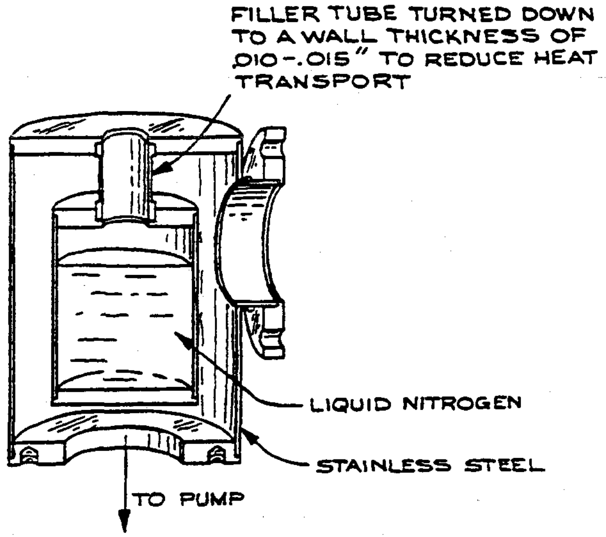

FILLER TUBE TURNED DOWN TO A WALL THICKNESS OF .010-.015" TO REDUCE HEAT TRANSPORT
LIQUID NITROGEN
STAINLESS STEEL
TO PUMP

Figure 3.22 Liquid-nitrogen-cooled trap.

mantle may be purchased from the Glas-Col Apparatus Company. The flanges of the trap are water-cooled to prevent overheating the seals.

The maintenance of a high-vacuum trap is different from that of a foreline trap. A high-vacuum molecular-sieve trap is placed above an oil diffusion pump, and a gate valve is located above the trap to permit isolation of the trap and pump stack from the vacuum chamber. After the initial bakeout the pump is run continuously so that the trap is always under vacuum. The isolation valve is closed whenever it is necessary to open the chamber to the atmosphere, and the chamber is rough-pumped before reopening the isolation valve. It is unwise to attempt to regenerate the molecular sieve by baking. The initial bake of the sieve results in the evolution of water vapor, but subsequent baking will drive out absorbed pump-fluid vapor. This oil vapor will condense on the bottom of the isolation-valve gate and will be exposed to the vacuum whenever the valve is open. After the initial bakeout, the molecular sieve will trap oil vapor effectively for a period of at least six months if it is not exposed to moist air during that period. When the sieve becomes clogged with absorbed vapor, the base pressure attainable with a typical pump-and-trap combination will begin to rise. At this time the molecular-sieve charge should be replaced. If it is absolutely necessary to stop the diffusion pump, the pump and trap should be filled with argon or dry nitrogen and isolated from the atmosphere in order to preserve the molecular sieve.

Liquid-nitrogen-cooled traps and baffles are frequently used with either oil or mercury diffusion pumps to condense backstreaming pump fluid. A metal cold trap is illustrated in Figure 3.22. These are very effective traps, but they have the disadvantage of requiring regular refilling with coolant. An alternative is to fill the trap with a low-melting liquid such as isopropyl alcohol and refrigerate the liquid with an immersion cooler. A temperature of −40°C is usually adequate. When a cold trap is permitted to warm up, it must be isolated from the vacuum chamber so that condensed material does not migrate into the chamber. Also the trap should be vented so that the evaporating condensate does not build up a dangerously high pressure.

An optically dense baffle of the type shown in Figure 3.23 is usually placed over the inlet of a diffusion pump. These baffles may be air-cooled, water-cooled, or cooled by a small refrigerator. When placed between a diffusion pump and a liquid-nitrogen trap or molecular-sieve trap, a baffle will significantly reduce the rate of contamination of the trap. When a baffled pump is used without a trap, the partial pressure of oil vapor in the vacuum system is reduced to the vapor pressure of oil at the temperature of the baffle.

3.5.6 Molecular Beams and Gas Jets

Very often it is necessary to introduce a gaseous sample into a vacuum system. The sample may be contained in a chamber with holes suitably located to admit probes such as light beams or electron beams. However, the walls of such a chamber may prove to be a hindrance to

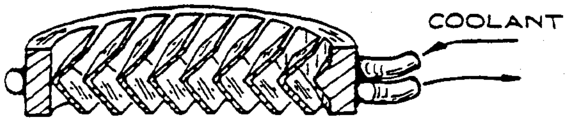


Figure 3.23 Cooled baffle.

the proposed experiment. In this case the sample may be introduced as an uncontained, but directed beam of atoms or molecules. The absence of walls is only one of several advantages to using a gas beam. Because of the directed velocities of the particles in the beam, it is possible to maintain the sample in a collisionless environment while still obtaining useful densities. The beam can be crossed with another beam or directed at a surface to obtain collisions of a specified orientation. In other cases, the expansion of a jet of gas results in extreme cooling to give a sample of gas with only a small number of quantum states populated. Lucas succinctly stated the case for atomic or molecular beams when he observed that "beams are employed in experiments where collisions . . . are either to be studied, or to be avoided."[9]

A gas beam is created by permitting the gas to flow into a vacuum through a tube. The result depends upon whether the flow exiting the tube is in the molecular or viscous flow regime. Of course the beam shape can be determined by appropriate apertures downstream of the channel through which the gas flows into the vacuum.

We begin with the molecular-flow case, where the mean free path λ of the gas at the outlet of the gas channel is greater than the diameter d of the channel. For the case of flow from a region of relatively high pressure into a vacuum through a round aperture (that is, a tube of length $l = 0$), the flux in a direction θ to the normal to the aperture surface is

$$I(\theta) = I(\theta = 0)\cos\theta \qquad (\text{atoms sec}^{-1}\text{sr}^{-1}),$$

and the beam width H = 120° (full width at half maximum). The beam can be narrowed by increasing the length l of the tube or channel through which the gas flows into the vacuum. A theoretical description of the gas beam issuing from such a channel depends upon the molecular diameter σ and the nature of the scattering of the molecules from the walls. A number of both theoretical and experimental investigations have been carried out in an effort to describe a gas beam in terms of easily measured parameters. In comparing experiment and theory it has turned out that the observed flux and the sharpness of the beam fall short of theoretical expectation by a factor of 2 to 5.

In general, both experimental and theoretical results imply that to obtain useful fluxes and reasonable collimation, a beam source consisting of a tube or channel must be at least ten times greater in length than its diameter, and the gas pressure behind this channel should be such that $d < \lambda < l$.

Lucas has devised a neat description of gas beams in the molecular-flow regime in terms of a set of reduced parameters.[9] The gas pressure P behind the channel (in torr), the beam width H (degrees), the gas flux I (atoms sec^{-1}sr^{-1}), and the throughput Q (atoms sec^{-1}) are related to the corresponding reduced parameters by

$$P = \frac{P_R}{l\sigma^2},$$

$$H = \frac{H_R d}{l},$$

$$I = \left(\frac{T}{295M}\right)^{1/2} \frac{d^2 I_R}{l\sigma^2},$$

$$Q = \left(\frac{T}{295M}\right)^{1/2} \frac{d^3 Q_R}{l^2\sigma^2},$$

where T is the absolute temperature, M is the molecular weight, d and l are in cm, and σ in Å (10^{-8} cm). For pressures roughly in the range where $d < \lambda < l$, the reduced half angle, flux, and throughput are related to the reduced pressure by

$$H_R = 2.48\times 10^2\sqrt{P_R},$$

$$I_R = 1.69\times 10^{20}\sqrt{P_R},$$

$$Q_R = 2.16\times 10^{21} P_R.$$

From his model calculations, Lucas has discovered that to obtain maximum intensity for a given half angle H, the reduced pressure P_R should not be less than unity.

For design work, a useful "optimum" equation can be derived by combining the above and setting $P_R = 1$ to give

$$I = 6.7\times 10^{17}\left(\frac{T}{295M}\right)^{1/2} \frac{dH}{\sigma^2} \qquad (\text{optimum}).$$

Then, for any gas and a given channel diameter and beam half angle, the axial intensity can be determined.

The tube length is fixed by the equations for H and $H_R(P_R = 1)$ above, and the input pressure by the equation for P. Note, however, that operating at somewhat higher pressures gives some control over the flux and throughput without significantly departing from the optimum condition.

The construction of low-pressure, single-channel gas-beam sources is fairly straightforward. An excellent source can be made of a hypodermic needle cut to the appropriate length. These needles are made of stainless steel, they are available in a wide range of lengths and diameters, and they come with a mounting fixture that is reasonably gastight.

The goal of high intensity in a sharp beam is incompatible with a single-channel source since for a fixed half angle, the gas load (throughput) increases more rapidly than the flux as the channel diameter is increased. The solution to this problem is to use an array of many tubes, each having a small aspect ratio (i.e., $d/l \ll 1$).[10] Tubes with diameters as small as 2×10^{-4} cm and lengths of 1×10^{-1} cm arranged in an array a centimeter or more across are commercially available in glass (e.g. from Galileo Electro-Optics Corp.).

When the gas pressure behind an aperture or nozzle leading to a vacuum is increased to the extent that the mean free path is much smaller than the dimensions of the aperture, a whole new situation ensues. Not only is the density of the resultant gas jet much greater than in the molecular-flow case, but the shape of the jet changes and the gas becomes remarkably cold. This is a direct result of collisions between molecules in the gas as it expands from the orifice into the vacuum. Because of collisions the velocities of individual molecules tend toward that of the bulk gas flow, just as an individual in a crowd tends to be dragged along with the crowd. The translational temperature of the gas, defined by the width of its velocity distribution, decreases, while the bulk-flow velocity increases. This conversion of random molecular motion into directed motion continues until the gas becomes too greatly rarefied by expansion, at which point the final temperature is frozen in. Because the mass-flow velocity increases while the local speed of sound, proportional to the square root of the translational temperature, decreases, the Mach number rises and the flow becomes supersonic. Inelastic molecular collisions in the expanding gas also cause internal molecular energy to flow into the kinetic energy of bulk flow, with the result that there may be substantial rotational and vibrational relaxation. Rotational temperatures less than 1 K and vibrational temperatures less than 50 K have been obtained. At these low temperatures only a small number of quantum states are occupied, a situation that is ideal for a variety of spectroscopic studies.

The low temperatures obtained in a supersonic jet can present some problems. Chief among these is that of condensation. Collisions in the high-density region of the jet cause dimer or even polymer formation. With high-boiling samples, bulk condensation may occur. Of course, if one wishes to study dimers or clusters this condensation is desirable. To avoid dimer formation, the sample gas is mixed at low concentration with helium. This *seeded gas* sample is then expanded in a jet. The degree of clustering of the seed-gas molecules is controlled by adjusting its concentration in the helium carrier.

The cooling effect as well as the directionality of the jet is lost if the expanding gas encounters a significant pressure of background gas in the vacuum chamber. Unfortunately, optimum operating conditions require a large throughput of gas into the chamber. The extent of cooling depends upon the probability of binary collisions, which is proportional to the product $P_0 d$ of the pressure behind the nozzle and the diameter of the nozzle. Furthermore, to minimize condensation the ratio d/P_0 should be as large as possible. The result, in practice, is that a large-capacity vacuum pumping system is needed. In typical continuously operating supersonic jet apparatus, source pressures of 10 to 100 atm have been used with nozzle diameters of 0.01 cm down to 0.0025 cm. The conductance of an aperture under these conditions is roughly

$$C = 15d^2 \text{ liter sec}^{-1},$$

when the diameter is in centimeters. This implies a throughput on the order of 10 torr liter sec^{-1}. To obtain a mean free path of several tens of centimeters, a base pressure of about 10^{-4} torr is required. Thus, a pump speed approaching 10,000 liter sec^{-1} may be necessary. Typically several large diffusion pumps are used. When possible the jet is aimed straight down the throat of a pump.

In many experiments the cold molecules in a supersonic jet are only probed periodically—as, for example, when doing laser spectroscopy with a pulsed laser. In this event the gas load imposed by the jet can be greatly reduced by operating the jet in a synchronously pulsed mode. A number of fast valves for this purpose have been devised.[11] The simplest are based upon inexpensive automobile fuel injector valves.[12] Jet sources providing gas pulses of duration less than a millisecond at a rate of 10 Hz are easily fabricated and are compatible with very modest vacuum systems.

Another efficient means of dealing with the background-gas problem has been demonstrated.[13] This relies upon the fact that interaction of the expanding gas with the background gas gives rise to a shock wave surrounding the gas jet. If the pressure P_0 behind the nozzle is increased, it is possible to achieve a mode of operation where the expanding gas behind the shock wave is unaffected by the background gas. In the region upstream of the shock front the gas behaves like a free jet expanding into a perfect vacuum. The distance from the nozzle to the shock front is

$$l = 0.67d\left(\frac{P_0}{P}\right)^{1/2},$$

where P_0 is the pressure in the nozzle, P the background pressure, and d the nozzle diameter.

When the nozzle pressure is sufficiently high to achieve a free jet length of usable dimensions, the background pressure will increase greatly. This state of affairs has a distinct advantage, since at a high background pressure, a large throughput can be achieved with a pump of moderate speed. For example, to obtain a free length $l = 1$ cm with a 0.01-cm nozzle and a source pressure of 10 atm, the background pressure should be about 400 mtorr. The throughput in this case is about 10 torr liter sec^{-1}, and the required pumping speed, 25 liter sec^{-1}, could be achieved by a large rotary mechanical pump or a Roots pump of modest capacity.

The fabrication of a nozzle is straightforward, although some difficulty may be encountered in making the necessary small hole. A skillful mechanical technician can drill a hole as small as 0.01-cm diameter. Smaller holes can be made by spark or electrolytic erosion, or by swaging a hole closed on a piece of hard wire and then withdrawing the wire. One of the first small, high-pressure nozzles[13] was made by drilling down the axis of a stainless-steel rod to within 0.1 mm of the end with a drill bit having a sharp conical point. A 0.0025-cm-diameter hole was then made through the remaining metal by spark erosion.

For many experiments, the gas flowing into a vacuum system through an aperture or a channel produces a beam that is too broad or too divergent for the intended application. In this case the beam shape can be defined by one or more apertures placed downstream from the source. Often it is advantageous to build these apertures into partitions that separate the vacuum housing into a succession of chambers. Each chamber can be evacuated with a separate pump. The pressure in the first will be highest, so that the large throughput obtained here will significantly reduce the gas load on the next pump. This scheme is called *differential pumping*.

An aperture downstream of a supersonic jet but close enough to be in the viscous flow region is called a *skimmer*. The design of these skimmers is critical, since they tend to produce turbulence, which will destroy the directionality of the flowing gas. A skimmer is usually a cone with its tip cut off and the truncated edge ground to knife sharpness. The details of the design, location, and fabrication of a skimmer are beyond the scope of this book, and the interested reader should refer to a specialized text on this subject.[14]

3.5.7 Electronics and Electricity *In Vacuo*

Electrical insulation inside a vacuum system is not generally a problem. At pressures below 10^{-4} torr a gap of 1 mm is adequate insulation up to at least 5000 volts. Initially, sparks may occur between closely spaced parts because of high field gradients around whiskers of metal. These whiskers quickly evaporate and sparks do not recur. High-voltage discharges can also occur along surfaces in a vacuum system. This is particularly a problem if the surfaces are dirty or hygroscopic. Inorganic salts on the surface of an insulator tend to absorb moisture and become conducting. After cleaning, insulators should be rinsed with distilled water followed by ethanol and then dried with hot air.

If the spacing between wires in a vacuum system cannot be reliably maintained, then insulation is required. Teflon-insulated hookup wire may be used down to 10^{-7} torr, although air bleeding out from under the insulation will slow pumpdown. At very low pressures and high temperatures, wires can be insulated by stringing ceramic beads or pieces of Pyrex tubing over them.

Solder should not be used for electrical connections in a vacuum system, because the lead in solder and soldering flux contaminate the vacuum. Mechanical connections are preferable. Connections to electrodes can be made by wrapping a wire under the head of a screw and tightening the screw. Wires may be jointed by slipping the ends into a piece of tubing and crimping the tubing onto the wire. Tungsten, molybdenum, nichrome, or stainless-steel wires may be spot-welded together. In welding refractory metals, a more secure weld is obtained if a piece of nickel foil is interposed between the wires.

Electronic devices tend to overheat in a vacuum, since the only cooling is by radiation. Electronic components such as power transistors and integrated circuits that must dissipate more than about $\frac{1}{2}$ watt are particularly unreliable. Such devices should be placed in vacuum-tight, air-filled boxes that are thermally connected to the vacuum wall.

A number of electrical feedthroughs are illustrated in Figure 3.24. The tungsten-wire feedthrough (*a*) is made by sealing the wire directly into Pyrex as described in Section 2.3.10. A simple electrical feedthrough for glass systems may be made by soldering or brazing an electrode into a Kovar-to-Pyrex graded seal as in (*b*). Ceramic-insulated terminal end bushings of the type shown in (*c*) are available commercially (e.g. from Ceramaseal). After an electrode wire is welded or brazed into the center hole, these bushings are easily joined to a metal vacuum wall by welding or brazing. There are, in addition to the feedthroughs shown in Figure 3.24, a number of different types sold by manufacturers of vacuum equipment.

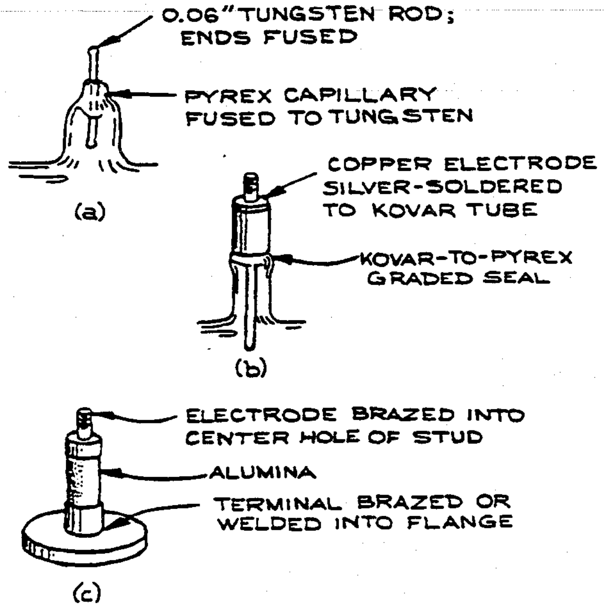

0.06" TUNGSTEN ROD; ENDS FUSED
PYREX CAPILLARY FUSED TO TUNGSTEN
(a)
COPPER ELECTRODE SILVER-SOLDERED TO KOVAR TUBE
KOVAR-TO-PYREX GRADED SEAL
(b)
ELECTRODE BRAZED INTO CENTER HOLE OF STUD
ALUMINA
TERMINAL BRAZED OR WELDED INTO FLANGE
(c)

Figure 3.24 Electrical feedthroughs employing (*a*) a tungsten-to-glass seal; (*b*) a Kovar-to-glass seal; (*c*) a ceramic-to-metal terminal bushing.

3.6 VACUUM-SYSTEM DESIGN AND CONSTRUCTION

Before beginning the design of a vacuum system, the size and shape of the vacuum chamber must be determined. The desired ultimate pressure and the composition of the residual gas in the chamber must be specified. In order to choose pumps for the system, a number of parameters must be determined in at least a semiquantitative manner. One must estimate the gas load on the pumps from outgassing as well as from gas introduced into the vacuum. The maximum tolerable pumpdown time should be specified.

The amount of money and time available are important considerations when designing a vacuum system. It is instructive to spend a few evenings leafing through vacuum-equipment manufacturers' catalogs in order to become familiar with the specifications and cost of commercial apparatus.

3.6.1 Some Typical Vacuum Systems

A diffusion-pumped vacuum system is shown schematically in Figure 3.25. For illustration, suppose the vacuum chamber has a volume of 10 liters and a surface area of 3000 cm^2, and is constructed of stainless steel with an initial outgassing rate of 10^{-7} torr liter sec^{-1} cm^{-2}. If the pump stack consists of a 2-in. oil diffusion pump with a speed of 150 liter sec^{-1} and a trap and baffle with a conductance of 300 liter sec^{-1}, then the net speed of the pump station will be

$$S = \left(\frac{1}{150 \text{ liter sec}^{-1}} + \frac{1}{300 \text{ liter sec}^{-1}}\right)^{-1} = 100 \text{ liter sec}^{-1}.$$

The ultimate pressure attainable against the outgassing load will be

$$P = \frac{Q}{S} \frac{(10^{-7} \text{ torr liter sec}^{-1}\text{cm}^{-2}) \times (3000 \text{ cm}^2)}{100 \text{ liter sec}^{-1}}$$
$$= 3 \times 10^{-6} \text{ torr}.$$

After a day of pumping and perhaps a light bake to 100°C, the outgassing rate should fall below 10^{-8} torr liter sec^{-1} cm^{-2} and the ultimate pressure will decrease to 3×10^{-7} torr or less.

To determine the speed required of the backing pump, first estimate the maximum throughput of the diffusion pump. If the diffusion pump is operated at pressures up to its stalling pressure of about 5×10^{-2} torr, the maximum throughput will be

$$Q_{\max} = P_{\max} S = (5 \times 10^{-2} \text{ torr}) \times (100 \text{ liter sec}^{-1})$$
$$= 5 \text{ torr liter sec}^{-1}.$$

To keep the foreline pressure below 300 mtorr at this throughput, the speed of the backing pump should be

$$S_p = \frac{Q_{\max}}{P} = \frac{5 \text{ torr liter sec}^{-1}}{3 \times 10^{-1} \text{ torr}}$$
$$= 16 \text{ liter sec}^{-1}.$$

This requirement can be met by the smallest single-stage or double-stage mechanical pump.

The conductance of the foreline should be at least twice the speed of the forepump, so that the forepump is not strangled by the foreline. To achieve a conduc-

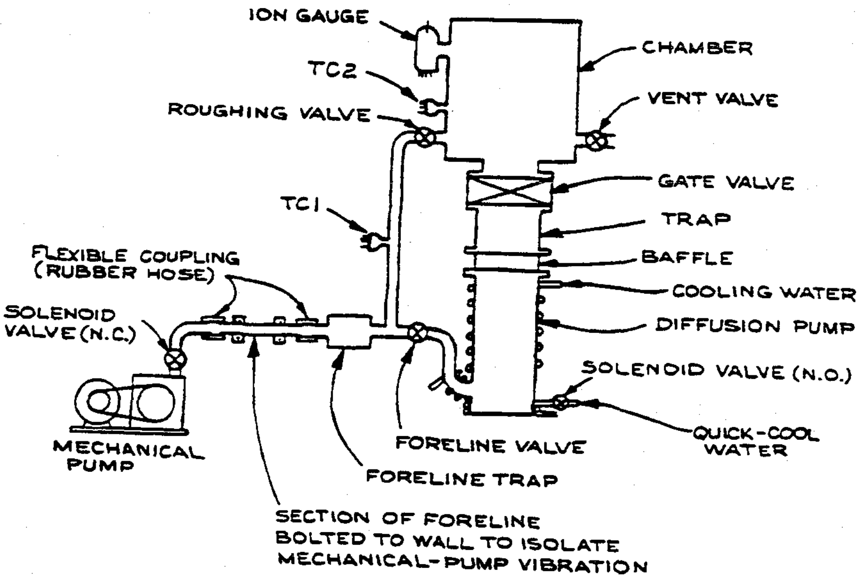


Figure 3.25 Schema of a diffusion-pumped vacuum system.

tance of 30 liter sec^{-1} at a pressure of 300 mtorr in a foreline 1 m long, the diameter (from Section 3.2.4) must be

$$D = \left(\frac{CL}{180P} \right)^{1/4} \text{cm} \qquad (L \text{ in cm})$$
$$= \left(\frac{30 \times 100}{180 \times 3 \times 10^{-1}} \right)^{1/4} = 2.7 \text{ cm}.$$

In this case a 1-in. copper water pipe would make an excellent foreline.

A number of inexpensive safeguards are incorporated in the design in Figure 3.25 so that the system will be failsafe in the event of a loss of electrical power. A normally open (N.O.) solenoid-operated water valve wired in parallel with the pump heater admits water to the diffusion pump quick-cooling coils if the power fails. Also, the diffusion pump is isolated from the mechanical pump by a normally closed (N.C.) solenoid valve to prevent air or oil from being sucked through the mechanical pump into the foreline. It is helpful to place a ballast volume in the foreline to maintain a low foreline pressure while the diffusion pump cools after a power outage. A useful failsafe mechanism, not shown in Figure 3.25, would be a water-pressure sensor in the water line that interrupts the electrical power if the water fails. Another useful, but expensive, accessory would be an electrically controlled, pneumatically activated gate valve that can isolate the vacuum chamber from the pump.

To activate a system of the type shown in Figure 3.25, starting with all valves closed and the pumps off, proceed as follows:

1. Turn on the mechanical pump.
2. When the pressure indicated by thermocouple gauge 1 (TC1) is below 200 mtorr, open the foreline valve.
3. When TC1 again indicates a pressure below 200 mtorr, turn on the cooling water and activate the diffusion pump. The pump will require about 20 minutes to reach operating temperature. When operating it makes a crackling sound.
4. Before pumping on the chamber with the diffusion pump, the pressure in the chamber must be reduced to a rough vacuum. Close the foreline valve and open the roughing valve. When the pressure indicated by TC2 is below 200 mtorr, close the roughing valve and open the foreline valve. The diffusion pump should not be operated for more than 2 minutes with the foreline valve closed. If necessary, interrupt the roughing procedure, close the roughing valve, and reopen the foreline valve for a moment to ensure that the diffusion-pump outlet pressure does not rise above about 200 mtorr.
5. Open the gate valve that isolates the diffusion pump from the chamber. Within about a minute the pressure indicated by TC2 should fall to a few millitorr and the ionization gauge can be turned on.

A diffusion-pumped system of the type shown in Figure 3.25 is one of the most convenient and least expensive systems for obtaining high or even ultrahigh vacuum. The system can be fabricated of metal or glass using either an oil or a mercury diffusion pump. Without a trap, a mercury pump will yield an ultimate pressure of about 10^{-3} torr. A baffled, but untrapped, oil diffusion pump on this system will result in an ultimate pressure slightly below 10^{-6} torr. Using a mercury diffusion pump with a liquid-nitrogen-cooled trap or an oil diffusion pump with either a molecular-sieve trap or a cold trap, an ultimate pressure of 10^{-8} torr can be achieved. An ultimate pressure of 10^{-10} torr is possible if the vacuum chamber is baked and only very low-vapor-pressure materials are exposed to the vacuum.

A small, inexpensive system, capable of producing an oil-free vacuum of about 10^{-10} torr in a 1-liter volume, is illustrated in Figure 3.26. The main residual gases in such a system are primarily water vapor and rare gases. The system must be heated to drive water vapor from the walls, and since none of the pumps are particularly efficient for rare gases, the system is purged with nitrogen before being evacuated in order to displace the rare gases.[15]

The pumpdown procedure for this vacuum system is as follows:

Figure 3.26 A small, clean, inexpensive ultrahigh-vacuum system.

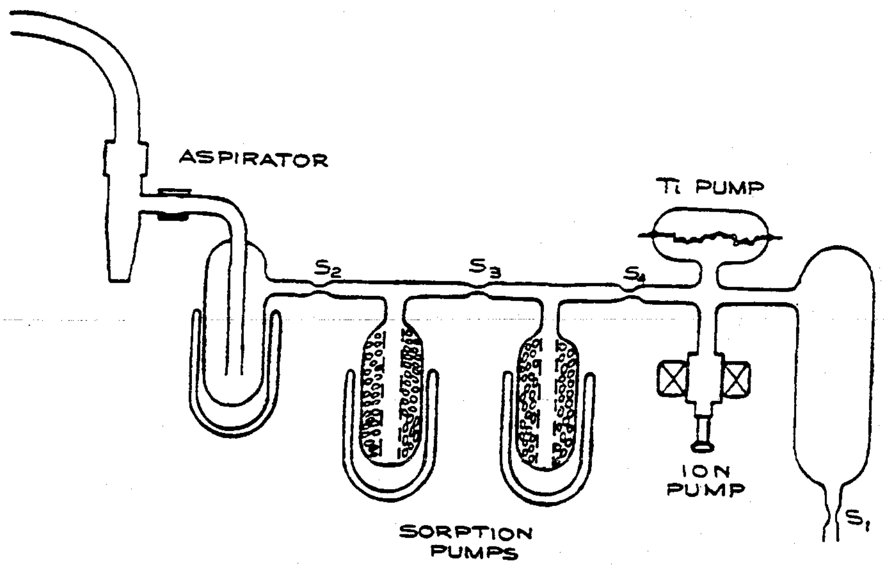
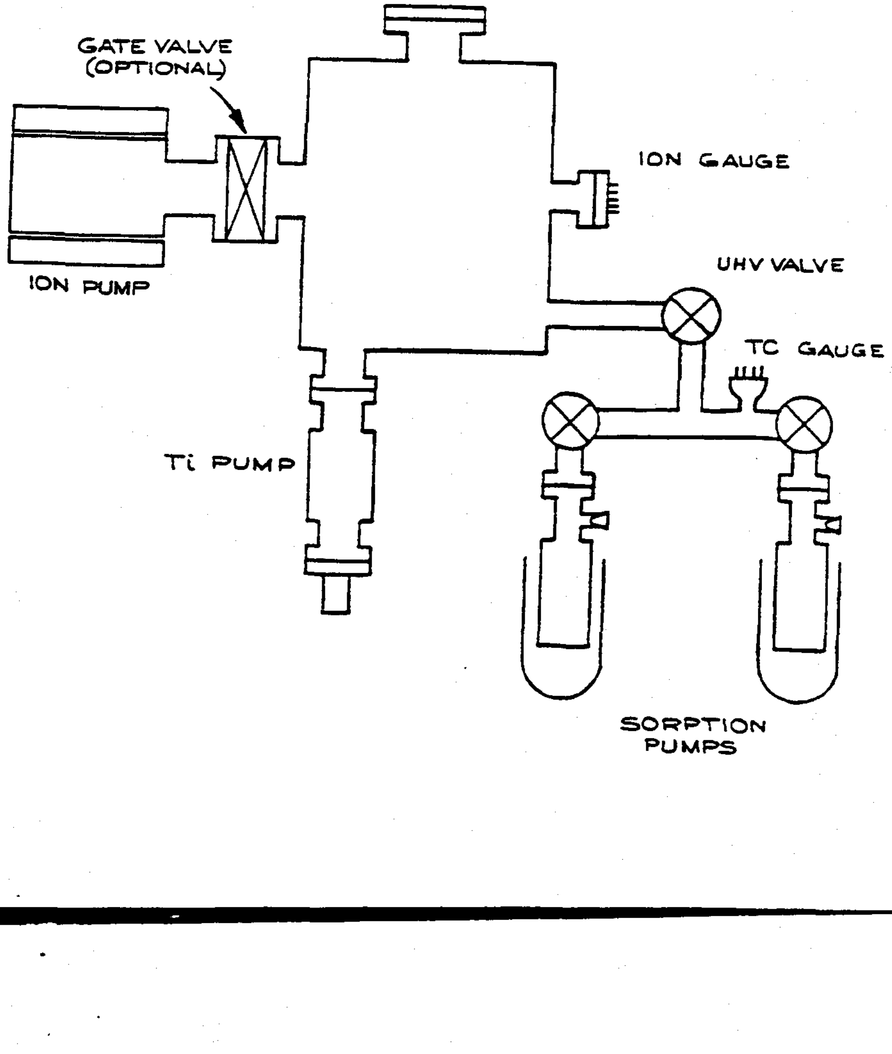


Figure 3.27 An oil-free ultrahigh-vacuum system.



1. Flow pure, dry nitrogen through the system from S_1 to the aspirator. Do not activate the aspirator. The nitrogen gas can be obtained from a Dewar filled with liquid nitrogen.
2. As nitrogen flows through the system, the molecular-sieve sorption pumps are heated to 300°C and the trap is cooled with liquid nitrogen. The remainder of the system should be heated to about 100°C with heating tape or by brushing the glass from S_1 toward V_1 with a cold flame or with hot air from a heat gun.
3. After about an hour, close pinchoff S_1 by heating the glass with a torch, and turn on the aspirator. The pressure will quickly fall to about 20 torr.
4. Seal off constriction S_2 and refrigerate the first molecular-sieve pump. The pressure will fall to about 10^{-4} torr. Continue to heat the system.
5. Seal off S_3 and cool the second sorption pump. The pressure will fall to about 10^{-7} torr. Continue heating the system for a short time to drive off the remaining adsorbed gases.
6. Pass a current through the filament of the titanium pump to evaporate a layer of titanium onto the glass envelope, and activate the ion pump.
7. Seal off S_4 and deposit a fresh layer of titanium. The pressure should fall well below 10^{-9} torr.

When sealing off the constriction it is wise to proceed slowly so that the pumps have time to take up any gas that is liberated.

A similar system can be made with the pinchoffs replaced by all-metal high-vacuum valves. In this case it would also be possible to construct the entire system of stainless steel.

A large, oil-free ultrahigh-vacuum system is illustrated in Figure 3.27. The system is roughed down with sorption pumps, and the ion pump is depended upon for the removal of rare gases. The system must be baked. A number of variations on this system are possible. The rough-pumping operation can be accelerated and the gas load on the sorption pumps reduced if the system is first roughed down with one of the compressed-air aspirators sold for this purpose. The pumping at high vacuum can be carried out with a cryopump rather than a getter pump. The choice of the high-vacuum pump or combination of pumps depends upon the composition of the residual gas as well as the composition of gases to be admitted to the system at high vacuum.

3.6.2 The Construction of Metal Vacuum Apparatus

Metal vacuum chambers are usually made up of one or more cylindrical sections, because a thin cylindrical shape has the strength to withstand external pressure and because metal tube stock is readily available. Even when tube stock is unavailable, a cylindrical section is easily fabricated by rolling sheet metal. The ends of a cylindrical chamber are most conveniently closed with flat plates as shown in Figure 3.28. A flat end plate must be quite thick to withstand atmospheric pressure, as discussed below.

The ability of a cylinder to resist collapsing under external pressure depends upon the length of the cylinder between supporting flanges, the diameter, the wall thickness, and the strength of the material. The A.S.M.E. has published standards for the wall thickness of cylindrical vessels under external pressure.[16] The data in

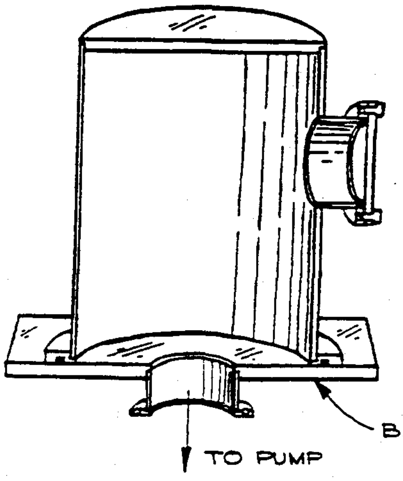


Figure 3.28 Typical metal vacuum chamber.

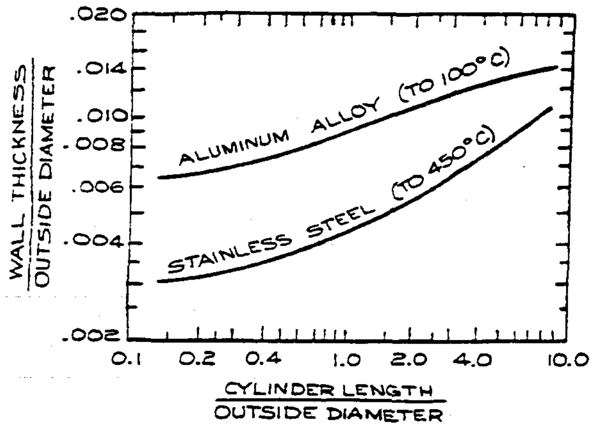


Figure 3.29 Minimum wall thickness for aluminum alloy and stainless-steel tubes under atmospheric pressure as a function of length between supporting flanges.

Figure 3.29 have been abstracted from the A.S.M.E. specifications for type 3003-T0 aluminum alloy and type 304 stainless steel respectively. These recommendations should be more than adequate for any of the 4000-, 6000-, or 7000-series tempered aluminum alloys (such as 2024-T4, 6061-T6 or 7075-T6) and adequate for most stainless steels. Soft aluminum such as the 1000-series alloys should not be used for vacuum chambers. Holes cut for ports in the side of a cylinder reduce the strength of the cylinder and should be avoided. If side ports are necessary the wall thickness should be increased over the recommendations of the code. Dents and out-of-roundness in a cylinder weaken it and should be avoided.

A flat end plate will bend a surprising amount under atmospheric pressure. The deflection at the center of a flat circular end plate clamped at its edges (e.g. by welding or otherwise firmly affixing it to a cylinder) is

$$\delta = \frac{3PR^4(1-\mu^2)}{16Ed^3},$$

and the maximum tensile stress, which occurs at the edge, is

$$s_{\max}(\text{edge}) = \frac{3}{4}\left(\frac{R}{d}\right)^2 P,$$

where R is the radius of the plate, d the thickness, P the external pressure, μ Poisson's ratio, and E the modulus of elasticity or Young's modulus of the material.[17] For metals Poisson's ratio is about 0.3. For steels with a Young's modulus of 3×10^7 psi, an acceptable value of the ratio of radius to thickness, R/d, is 30. In this case the relative deflection is

$$\frac{\delta}{R} = 0.002 \qquad (\text{steel}, R/d = 30).$$

A slightly thicker aluminum plate is required to achieve the same result:

$$\frac{\delta}{R} = 0.002 \qquad (\text{aluminum}, R/d = 20).$$

The deflection of an end plate is greater when the edges are not securely clamped, as is the case for almost all removable end plates. For a circular plate with unclamped edges,

$$\delta = \frac{3PR^4(5+\mu)(1-\mu)}{16Ed^3},$$

and the maximum tensile stress, which occurs at the center, is

$$s_{\max}(\text{center}) = \frac{3}{8}\left(\frac{R}{d}\right)^2(3+\mu)P.$$

A deflection of about $0.003R$ is usually acceptable for a demountable end plate:

$$\frac{\delta}{R} = 0.003 \qquad (\text{steel}, R/d = 20),$$

and

$$\frac{\delta}{R} = 0.003 \qquad (\text{aluminum}, R/d = 15).$$

Steel is about three times heavier than aluminum; so for a given maximum deflection, an aluminum plate is less than half as heavy as a steel plate. A vacuum-chamber design that incorporates a stainless-steel cylinder with a demountable aluminum end plate has much to say in its

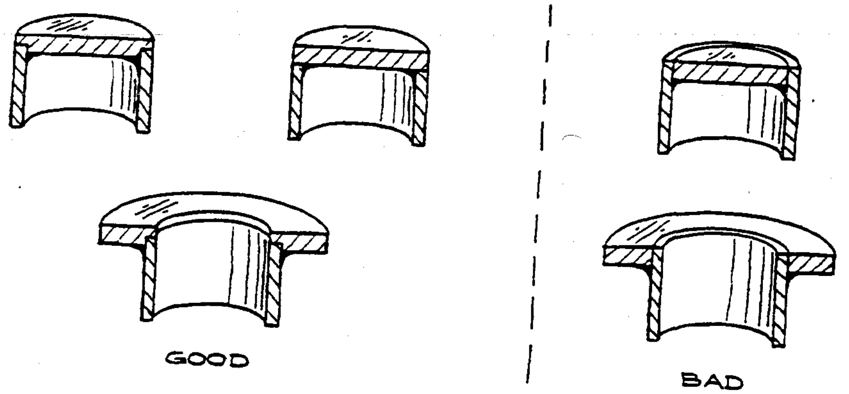


Figure 3.30 The design of joints to be soldered or brazed.

favor. Welds in the stainless cylinder are much stronger and more reliable than would be the case if the cylinder were made of aluminum. A removable aluminum end plate is lighter and much less expensive to machine than a steel one.

Metal parts may be joined by soldering, welding, or brazing. Heliarc welding produces the cleanest, strongest joint, but soldering and brazing result in less distortion of the joined parts and are more convenient operations in the laboratory. As shown in Figure 3.30, a joint that is to be soldered or brazed should be designed so that the metal parts take the thrust of the atmosphere and the soldering or brazing material serves only as a sealant. If possible, the joint should be designed to provide positive location of mating parts. The surfaces to be joined should be cleaned before assembly by sandblasting or by polishing with sandpaper so that the solder or brazing metal will flow into and completely fill the joint. This is necessary to achieve maximum strength and to eliminate the narrow gap, which is difficult to pump out and is liable to collect flux and other material that will contaminate the vacuum system. Lead-tin solder should only be used for rough vacuum applications such as a diffusion-pump foreline. Brazed joints are acceptable in high vacuum, but usually the joints should not be heated over about 100°C *in vacuo*, since most brazing alloys contain high-vapor-pressure constituents such as zinc or cadmium. Brazing is an excellent way of attaching thin metal parts, such as bellows, which can easily be burned through in a welding operation.

Parts to be joined by welding are designed so that the rate of heat dissipation while welding is the same to each part. This is necessary to prevent destructive thermal stresses from being frozen into the finished joint. When joining a heavy flange to a thin tube as shown in Figure 3.31, a notch or a groove is cut in the thick piece of metal near the joint to control the rate of dissipation of heat from the joint. Whenever possible, welds in the wall of a vacuum chamber should be on the inside. If an outside weld is necessary, it should be a full-penetration weld. If neither an inside nor a full-penetration outside weld is possible, the mating parts should fit together loosely so that the gap between is wide open.

Metal parts inside a vacuum system may be joined by welding or brazing, but in most cases nut-and-bolt assembly is more convenient. Special care must be taken to prevent air from being trapped under a screw in a

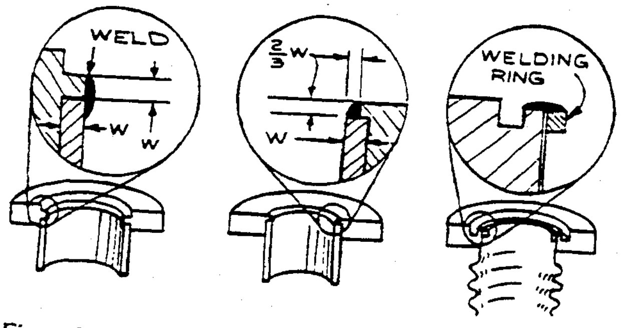


Figure 3.31 The design of joints to be heliarc-welded.

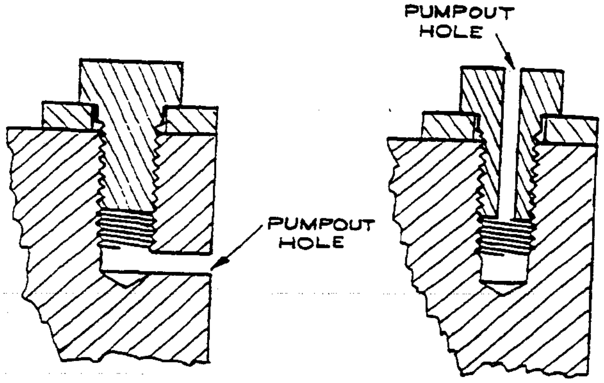


Figure 3.32 Pumpout holes for blind screw holes.

blind hole. The placement of pumpout holes for blind screw holes is illustrated in Figure 3.32.

3.6.3 Surface Preparation

Surfaces that are to be exposed to a vacuum should be free of substances that have a significant vapor pressure, and they should be as smooth as possible to minimize the microscopic surface area and thus minimize the amount of adsorbed gas. The substances that must be removed from the surfaces of vacuum apparatus are mainly hydrocarbon oils and greases, and inorganic salts that are hygroscopic and outgas water vapor.

The two preferred treatments for metal vacuum-system surfaces are electropolishing and bead blasting. Conventional wheel polishing and buffing is unsatisfactory, since these processes tend to flatten surface burrs and trap gas underneath. Electropolishing is conveniently carried out in the laboratory, although most machine shops are prepared to do it routinely. An electropolishing solution for stainless steel recommended by Armco Steel consists of 50 parts by volume of citric acid and 15 parts by volume of sulfuric acid plus enough water to make 100 parts of solution. The solution should be used at a temperature of 90°C. A current density of about 0.1 $\mathrm{A\,cm^{-2}}$ at 6 to 12 volts is required. A copper cathode is used with the piece to be polished serving as the anode.

Blasting with 20–30-μm glass beads effectively reduces the adsorbing area of a metal surface and thus reduces the rate of outgassing from the surface. Glass beading can be carried out in a conventional sandblasting apparatus. The machine should be carefully cleaned of coarse grits before use.

Glass may be effectively cleaned with a 10% solution of HF. The routine cleaning procedure for metal parts is as follows:

1. Scrub with a strong solution of detergent (Alconox or liquid dishwashing detergent is fine).
2. Rinse with very hot water.
3. Rinse with distilled water.
4. Rinse with pure methanol.

Do not touch clean parts with bare hands. Disposable plastic gloves (free of talc) are convenient for handling vacuum apparatus after cleaning.

An ultrasonic cleaner is particularly effective for preparing vacuum parts. The cleaner should be filled with a detergent solution. If two cleaners are available, fill one with detergent solution and the other with distilled water, and use them in sequence with a hot-water rinse between.

A vapor-phase degreaser is one of the most effective tools for cleaning metal parts after fabrication. The part to be cleaned is suspended in a cloud of hot solvent vapor. Vapor condenses on the part, heating it and rinsing it with pure hot solvent. As illustrated in Figure 3.33, a vapor degreaser is easily and inexpensively constructed. Any steel container, from a coffee can to a 55-gallon drum, can be used, depending on the size of the objects to be cleaned. The cooling coil around the top of the can is necessary to condense solvent vapors before they escape the container. Trichloroethylene is the preferred solvent. A vapor degreaser should always be operated under a fume hood to prevent contamination of the laboratory with toxic solvent vapors. Objects to be cleaned in the degreaser should first be washed with detergent and rinsed so as not to load the degreaser with volatile oils.

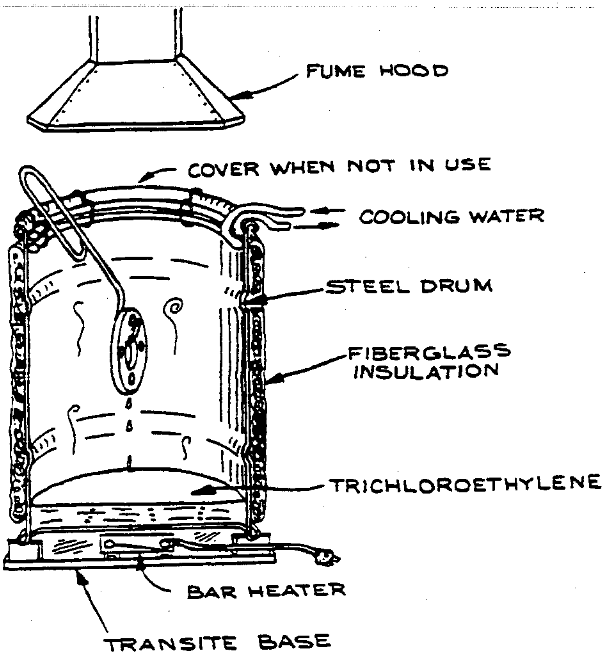


Figure 3.33 Vapor-phase degreaser.

3.6.4 Leak Detection

A leak that raises the base pressure of a system above about 10^{-6} torr can usually be found by probing suspected locations on the outside of the vacuum chamber with a liquid or vapor for which the gauge sensitivity or pump speed is very different from that for air. A squeeze bottle of acetone or a spray can of liquid Freon cleaner is a useful tool. These liquids will usually cause a very abrupt increase in indicated pressure as they flow through a leak, but sometimes rapid evaporation of a liquid through a leak will cause the liquid to freeze and temporarily plug the leak, causing the pressure to fall. A disadvantage of this method is that the solvent may contaminate O-rings. A small jet of helium is also a useful leak probe, since an ionization gauge is very insensitive to helium. The indicated pressure will fall when helium is introduced into a leak.

In a glass system a leak that raises the pressure into the range from 10 mtorr to several torr can be located with a Tesla coil. The surface of the glass is brushed with the discharge from the Tesla coil. The discharge is preferentially directed toward the leak, and a bright white spot will reveal the location as the discharge passes through. Avoid very thin glass walls and glass-to-metal seals, as the Tesla discharge can hole the glass in these fragile areas.

For very small leaks in high-vacuum systems a mass-spectrometer leak detector is needed.

3.6.5 Ultrahigh Vacuum

To achieve pressures much below 10^{-7} torr, baking is required in order to remove water and hydrocarbons from vacuum-system walls. Heating to 50–100°C for several hours will improve the ultimate pressure of most systems by an order of magnitude. A true ultrahigh-vacuum system must be baked to at least 250°C for several hours, and a system contaminated with hydrocarbons should be baked at 400°C. After a rigorous preliminary bake *in vacuo* to remove deeply adsorbed material, the system may be baked more gently after subsequent exposures to air.

A small system can be wrapped with heating tape and then covered with fiberglass or asbestos insulation. For larger systems an oven constructed of Transite or sheet metal insulated with fiberglass can be erected around the vacuum chamber and heated with bar heaters. Batts of fiberglass insulation intended for use in automobile engine compartments are readily available. Do not use home-insulating fiberglass materials. Metal seals must be used in a metal vacuum system that is to be baked. It is, however, often necessary to join the vacuum chamber to its pumping station with an O-ring-sealed flange. In this case, a cooling coil should be installed around the flange (as in Figure 3.21) to prevent melting the O-ring while the chamber is being baked.

CITED REFERENCES

1. N. Milleron, Res. Dev., 21, No. 9, 40, 1970.
2. The derivation of the conductance equations is given by

C.M. VanAtta, *Vacuum Science and Engineering*, McGraw-Hill, New York, 1965, pp. 44–62.

3. M.A. Biondi, Rev. Sci. Instr., 24, 989, 1953.
4. A.T.J. Hayward, J. Sci. Instr., 40, 173, 1963; C. Veillon, Rev. Sci. Instr., 41, 489, 1970.
5. H. Ishii and K. Nakayama, *Transactions of the 2nd International Congress on Vacuum Science and Technology*, Vol. 1, Pergamon Press, Elmsford, N.Y., 1961, p. 519; R.J. Tunnicliffe and J.A. Rees, Vacuum, 17, 457, 1967; T. Edmonds and J.P. Hobson, J. Vac. Sci. Tech., 2, 182, 1965.
6. G.R. Brown, P.J. Sowinski, and R. Pertel, Rev. Sci. Instr., 43, 334, 1972.
7. An application of this process is described by J.H. Moore and C. B. Opal, Space Sci. Instr., 1, 377, 1975.
8. The trap shown in Figure 3.20(*b*) is described by W.W. Roepke and K.G. Pung, Vacuum, 18, 457, 1968.
9. C.B. Lucas, Vacuum, 23, 395, 1973.
10. R.H. Jones, D.R. Olander, and V.R. Kruger, J. Appl. Phys., 40, 4641, 1969; D.R. Olander, J. Appl. Phys., 40, 4650, 1969; D.R. Olander, J. Appl. Phys., 41, 2769, 1970; D.R. Olander, R.H. Jones, and W.J. Siekhaus, J. Appl. Phys., 41, 4388, 1970; W.J. Siekhaus, R.H. Jones, J. Appl. Phys., 41, 4392, 1970.
11. M.G. Liverman, S.M. Beck, D.L. Monts, and R.E. Smalley, J. Chem. Phys., 70, 192, 1979; W.R. Gentry and C.F. Giese, Rev. Sci. Instr., 49, 595, 1978.
12. D. Bassi, S. Iannotta, and S. Niccolini, Rev. Sci. Instr., 52, 8, 1981; F.M. Behlen, Chem. Phys. Lett., 60, 364, 1979.
13. R.E. Smalley, D.H. Levy, and L. Wharton, J. Chem. Phys., 64, 3266, 1976.
14. J.B. Anderson, R.P. Andres, and J.B. Fenn, in *Advances in Chemical Physics*, Vol. 10, *Molecular Beams*, J. Ross, Ed., Wiley, New York, 1966, Chapter 8; H. Pauly and J.P. Toennies, in *Methods of Experimental Physics*, Vol. 7, Part A, B. Bederson and W.L. Fite, Eds., Academic Press, New York, 1968, Chapter 3.1.
15. This system is an adaptation of a design described by N.W. Robinson, *Ultra-high Vacuum*, Chapman and Hall, London, 1968, pp. 72–73.
16. *A.S.M.E. Boiler and Pressure Vessel Code*, Section VIII, Division 1, Appendix V, 1974.
17. J.F. Harvey, *Pressure Vessel Design*, Van Nostrand, Princeton, N.J., 1963, pp. 89–91.

GENERAL REFERENCES

Comprehensive Texts on Vacuum Technology

S. Dushman, *Scientific Foundations of Vacuum Technique*, 2nd edition, J. M. Lafferty, Ed., Wiley, New York, 1961.

G. Lewin, *Fundamentals of Vacuum Science and Technology*, McGraw-Hill, New York, 1965.

C.M. Van Atta, *Vacuum Science and Engineering*, McGraw-Hill, New York, 1965.

Criteria for Selection and Sizing Vacuum Pumps

J.F. O'Hanlon, *A User's Guide to Vacuum Technology*, Wiley, New York, 1980.

Design of Vacuum Systems

N. T. M. Dennis and T. A. Heppell, *Vacuum System Design*, Chapman and Hall, London, 1968.

G. W. Green, *The Design and Construction of Small Vacuum Systems*, Chapman and Hall, London, 1968.

R. P. LaPelle, *Practical Vacuum Systems*, McGraw-Hill, New York, 1972.

Outgassing Data

W. A. Campbell, Jr., R. S. Marriott, and J. J. Park, *A Compilation of Outgassing Data for Spacecraft Materials*, NASA Technical Note TND-7362, NASA, Washington, D.C., 1973.

Properties of Materials Used in Vacuum Systems

W. Espe, *Materials of High Vacuum Technology*: Vol. 1, *Metals and Metaloids*; Vol. 2, *Silicates*; Vol. 3, *Auxiliary Materials*, Pergamon Press, Oxford, 1968 (a translation of the original German published in 1960).

Sealing Ceramics and Glass to Metal, Heat-Treating, Cleaning, Building Joints, and Feedthroughs

F. Rosebury, *Handbook of Electron Tube and Vacuum Techniques*, Addison-Wesley, Reading, Mass., 1969.

Ultrahigh Vacuum

P.A. Redhead, J.P. Hobson, and E.V. Kornelsen, *The Physical Basis of Ultrahigh Vacuum*, Chapman and Hall, London, 1968.

R.W. Roberts and T.A. Vanderslice, *Ultrahigh Vacuum and Its Applications*, Prentice-Hall, Englewood Cliffs, N.J., 1963.

W. Robinson, *The Physical Principles of Ultrahigh Vacuum Systems and Equipment*, Chapman and Hall, London, 1968.

MANUFACTURERS AND SUPPLIERS

Ace Glass, Inc.
P.O. Box 688
1430 Northwest Blvd.
Vineland, NJ 08360
(Glass vacuum accessories and glass process pipe)

Alloy Products Company
1045 Perkins Ave.
Waukesha, WI 53186

Balston, Inc.
703 Massachusetts Ave.
Lexington, Mass. 02173

Ceramaseal, Inc.
New Lebanon Center
New York, NY 12126

Eutectic Welding Alloys Co.
40-42 172nd St.
Flushing, NY 11358

Galileo Electro-Optics Corp.
Galileo Park
Sturbridge, MA 01518

Glas-Col Apparatus Co.
709 Hulman St.
Terre Haute, IN 47802

Industrial Tectonics, Inc.
P.O. Box 1128
Ann Arbor, MI 48106

Ladish Co.
Cudahy, WI 53110

Lebold-Heraeus,
200 Seco Rd.,
Monroeville, PA 15146

MKS Instruments
22 3rd Ave.
Burlington, MA 01803

Neslab Instruments, Inc.
871 Islington St.,
Portsmouth, NH 03801
(Refrigerators and immersion coolers)

Texwipe Company
Hillsdale, NJ 07642

Tra-Con, Inc.
55-T North St.
Medford, MA 02155

Wallace & Tiernan
25 Main St.
Belleville, NJ 07109

Weiss, R. (1972) Quarterly Progress Report of the Research Laboratory of Electronics of the Massachusetts Institute of Technology, Vol. 105, P54

REF. XX

B. ELECTROMAGNETICALLY COUPLED BROADBAND GRAVITATIONAL ANTENNA

1. Introduction

The prediction of gravitational radiation that travels at the speed of light has been an essential part of every gravitational theory since the discovery of special relativity. In 1918, Einstein,[1] using a weak-field approximation in his very successful geometrical theory of gravity (the general theory of relativity), indicated the form that gravitational waves would take in this theory and demonstrated that systems with time-variant mass quadrupole moments would lose energy by gravitational radiation. It was evident to Einstein that since gravitational radiation is extremely weak, the most likely measurable radiation would come from astronomical sources. For many years the subject of gravitational radiation remained the province of a few dedicated theorists; however, the recent discovery of the pulsars and the pioneering and controversial experiments of Weber[2,3] at the University of Maryland have engendered a new interest in the field.

Weber has reported coincident excitations in two gravitational antennas separated 1000 km. These antennas are high-Q resonant bars tuned to 1.6 kHz. He attributes these excitations to pulses of gravitational radiation emitted by broadband sources concentrated near the center of our galaxy. If Weber's interpretation of these events is correct, there is an enormous flux of gravitational radiation incident on the Earth.

Several research groups throughout the world are attempting to confirm these results with resonant structure gravitational antennas similar to those of Weber. A broadband antenna of the type proposed in this report would give independent confirmation of the existence of these events, as well as furnish new information about the pulse shapes.

The discovery of the pulsars may have uncovered sources of gravitational radiation which have extremely well-known frequencies and angular positions. The fastest known pulsar is NP 0532, in the Crab Nebula, which rotates at 30.2 Hz. The gravitational flux incident on the Earth from NP 0532 at multiples of 30.2 Hz can be 10^{-6} erg/cm^2/s at most. This is much smaller than the intensity of the events measured by Weber. The detection of pulsar signals, however, can be benefited by use of correlation techniques and long integration times.

The proposed antenna design can serve as a pulsar antenna and offers some distinct advantages over high-Q acoustically coupled structures.

2. Description of a Gravitational Wave in the General Theory of Relativity

In his paper on gravitational waves (1918), Einstein showed by a perturbation argument that a weak gravitational plane wave has an irreducible metric tensor in an almost Euclidean space. The total metric tensor is $g_{ij} = \eta_{ij} + h_{ij}$, where

$$\eta_{ij} = \begin{pmatrix} 1 & & & \bigcirc \\ & -1 & & \\ & & -1 & \\ \bigcirc & & & -1 \end{pmatrix}$$

is the Minkowski background metric tensor, h_{ij} is the perturbation metric tensor resulting from the gravitational wave, and it is assumed that all components of this tensor are much smaller than 1. If the plane wave propagates in the x_1 direction, it is always possible to find a coordinate system in which h_{ij} takes the irreducible form

$$h_{ij} = \left(\begin{array}{cc:cc} \bigcirc & & \bigcirc & \\ \hdashline & & h_{22} & h_{23} \\ \bigcirc & & h_{32} & h_{33} \end{array}\right)$$

with $h_{22} = -h_{33}$, and $h_{23} = h_{32}$. The tensor components have the usual functional dependence $f(x_1-ct)$.

To gain some insight into the meaning of a plane gravitational wave, assume that the wave is in the single polarization state $h_{23} = h_{32} = 0$, and furthermore let $h_{22} = -h_{33} = h \sin (kx_1-\omega t)$. The interval between two neighboring events is then given by

$$ds^2 = g_{ij}dx^i dx^j = c^2dt^2 - \left[dx_1^2 + (1+h\sin(kx_1-\omega t))\,dx_2^2 + (1-h\sin(kx_1-\omega t))\,dx_3^2\right].$$

The metric relates coordinate distances to proper lengths. In this metric coordinate time is proper time; however, the spatial coordinates are not proper lengths. Some reality can be given to the coordinates by placing free noninteracting masses at various points in space which then label the coordinates. The proper distance between two coordinate points may then be defined by the travel time of light between the masses. Assume a light source at $x_2 = -a/2$ and a receiver at $x_2 = a/2$. For light, the total interval is always zero so that

$$ds^2 = 0 = c^2dt^2 - (1+h\sin(kx_1-\omega t))\,dx_2^2.$$

Since $h \ll 1$,

$$cdt = \left[1 + \frac{h}{2}\sin(kx_1-\omega t)\right] dx_2.$$

If the travel time of light, Δt, is much less than the period of the wave, the integral for

Δt becomes simple and we get

$$\Delta t = \left(1 - \frac{h}{2} \sin \omega t\right) \frac{a}{c}.$$

In the absence of the gravitational wave $\Delta t = \ell_o/c = a/c$, the coordinate distance becomes the proper length. The variation in Δt because of the gravitational wave is given by

$$\delta \Delta t = \left(\frac{h}{2} \sin \omega t\right) \frac{\ell_o}{c}.$$

This can be interpreted as though the gravitational wave produces a strain in space in the x_2 direction of

$$\frac{\Delta \ell}{\ell_o} = \frac{h}{2} \sin \omega T = \frac{h_{22}}{2}.$$

There is a comparable strain in the x_3 direction, however, inverted in phase.

This geometric description of the effects of a gravitational wave is useful for showing the interaction of the wave with free stationary particles. It becomes cumbersome when the particles have coordinate velocities or interact with each other. Weber[4] has developed a dynamic description of the effect of a gravitational wave on interacting matter which has negligible velocity. For the case of two masses m separated by a proper distance ℓ along the x_2 direction that are coupled by a lossy spring, the equation for the differential motion of the masses in the gravitational wave of the previous example becomes

$$\frac{d^2 x_{2R}}{dt^2} + \frac{\omega_o}{Q} \frac{dx_{2R}}{dt} + \omega_o^2 x_{2R} = c^2 R_{2020} \ell,$$

where x_{2R} is the proper relative displacement of the two masses, and R_{2020} is that component of the Riemannian curvature tensor which interacts with the masses to give relative displacements in the x_2 direction; it can be interpreted as a gravitational gradient force.

For the plane wave,

$$R_{2020} = \frac{1}{2c^2} \frac{d^2 h_{22}}{dt^2}$$

If the masses are free, the equation of differential motion becomes

$$\frac{d^2 x_{2R}}{dr^2} = \frac{1}{2} \frac{d^2 h_{22}}{dt^2} \ell$$

and, for zero-velocity initial conditions, the strain becomes $\frac{x_{2R}}{\ell} = \frac{1}{2} h_{22}$, which is the same result as that arrived at by the geometric approach.

The intensity of the gravitational wave in terms of the plane-wave metric tensor is given by Landau and Lifshitz[5] as

$$I_g = \frac{c^3}{16\pi G}\left[\left(\frac{dh_{23}}{dt}\right)^2 + \frac{1}{4}\left(\frac{dh_{22}}{dt} - \frac{dh_{33}}{dt}\right)^2\right]. \tag{1}$$

3. Gravitational Radiation Sources — Weber Events and Limits on Pulsar Radiation

The strain that Weber observes in his bars is of the order of $\Delta\ell/\ell \sim 10^{-16}$. If the strain is caused by impulsive events that can excite a 1.6 kHz oscillation in the bars, the events must have a rise time of 10^{-3} second or less — the fact that the bars have a high Q does not enter into these considerations. The peak incident gravitational flux of these events is truly staggering. Using Eq. 1, we calculate $I_g \geq 5 \times 10^9$ erg/s/cm^2.

If the sources of this radiation, which are alleged to be at the center of the galaxy, radiate isotropically, each pulse carries at least 5×10^{52} ergs out of the galaxy, the equivalent of the complete conversion to gravitational energy of 1/40 of the sun's rest mass. Weber observes on the average one of these events per day. At this rate the entire known rest mass of the galaxy would be converted into gravitational radiation in 10^{10} years. Gravitational radiation would then become the dominant energy loss mechanism for the galaxy.

Gravitational radiation by pulsar NP 0532, even at best, is not expected to be as spectacular as the Weber pulses. Gold[6] and Pacini[7] have proposed that pulsars are rotating neutron stars with off-axis magnetic fields. In a neutron star the surface magnetic field can be so large ($\sim 10^{12}$-10^{13} G) that the magnetic stresses perceptibly distort the star into an ellipsoid with a principal axis along the magnetic moment of the star. The star, as viewed in an inertial coordinate system, has a time-dependent mass quadrupole moment that could be a source of gravitational radiation at twice the rotation frequency of the star. Gunn and Ostriker[8] have made a study of this pulsar model and conclude from the known lifetime and present decay of the rotation frequency of NP 0532 that no more than 1/6 of the rotational energy loss of the pulsar could be attributed to gravitational radiation. The measured and assumed parameters for NP 0532 are listed below.

Rotation Frequency	$\nu = 30.2155 \ldots (\pm 3.4 \times 10^{-9})$ Hz
Slowdown Rate	$d\nu/dt = -3.859294 \pm .000053 \times 10^{-10}$ Hz/s
Distance	d = 1.8 kpc
Mass	$m = 1.4\ m_\odot$
Radius	r = 10 km.

The gravitational radiation intensity at 60.4 Hz incident on the Earth must be less than $I_g \leq 1 \times 10^{-6}$ erg/cm^2/s. The strain amplitude corresponding to this intensity is $\Delta\ell/\ell \leq 10^{-24}$.

4. Proposed Antenna Design

The principal idea of the antenna is to place free masses at several locations and measure their separations interferometrically. The notion is not new; it has appeared as a gedanken experiment in F. A. E. Pirani's[9] studies of the measurable properties of the Riemann tensor. However, the realization that with the advent of lasers it is feasible to detect gravitational waves by using this technique grew out of an undergraduate seminar that I ran at M.I.T. several years ago, and has been independently discovered by Dr. Philip Chapman of the National Aeronautics and Space Administration, Houston.

A schematic diagram of an electromagnetically coupled gravitational antenna is shown in Fig. V-20. It is fundamentally a Michelson interferometer operating in vacuum with the mirrors and beam splitter mounted on horizontal seismometer suspensions. The suspensions must have resonant frequencies far below the frequencies in the gravitational wave, a high Q, and negligible mechanical mode cross coupling. The laser beam makes multiple passes in each arm of the interferometer. After passing through the beam splitter, the laser beam enters either interferometer arm through a hole in the reflective coating of the spherical mirror nearest the beam splitter. The beam is reflected and refocused by the far mirror, which is made slightly astigmatic. The beam continues to bounce back and forth, hitting different parts of the mirrors, until eventually it emerges through another hole in the reflective coating of the near mirror. The beams from both arms are recombined at the beam splitter and illuminate a photodetector. Optical delay lines of the type used in the interferometer arms have been described by Herriott.[10] An experimental study of the rotational and transverse translational stability of this kind of optical delay line has been made by M. Wagner.[11]

The interferometer is held on a fixed fringe by a servo system which controls the optical delay in one of the interferometer arms. In such a mode of operation, the servo output signal is proportional to the differential strain induced in the arms. The servo signal is derived by modulating the optical phase in one arm with a Pockel-effect phase shifter driven at a frequency at which the laser output power fluctuations are small, typically frequencies greater than 10 kHz. The photo signal at the modulation frequency is synchronously detected, filtered, and applied to two controllers: a fast controller which is another Pockel cell optical phase shifter that holds the fringe at high frequencies, and a slow large-amplitude controller that drives one of the suspended masses to compensate for thermal drifts and large-amplitude low-frequency ground noise.

The antenna arms can be made as large as is consistent with the condition that the travel time of light in the arm is less than one-half the period of the gravitational wave

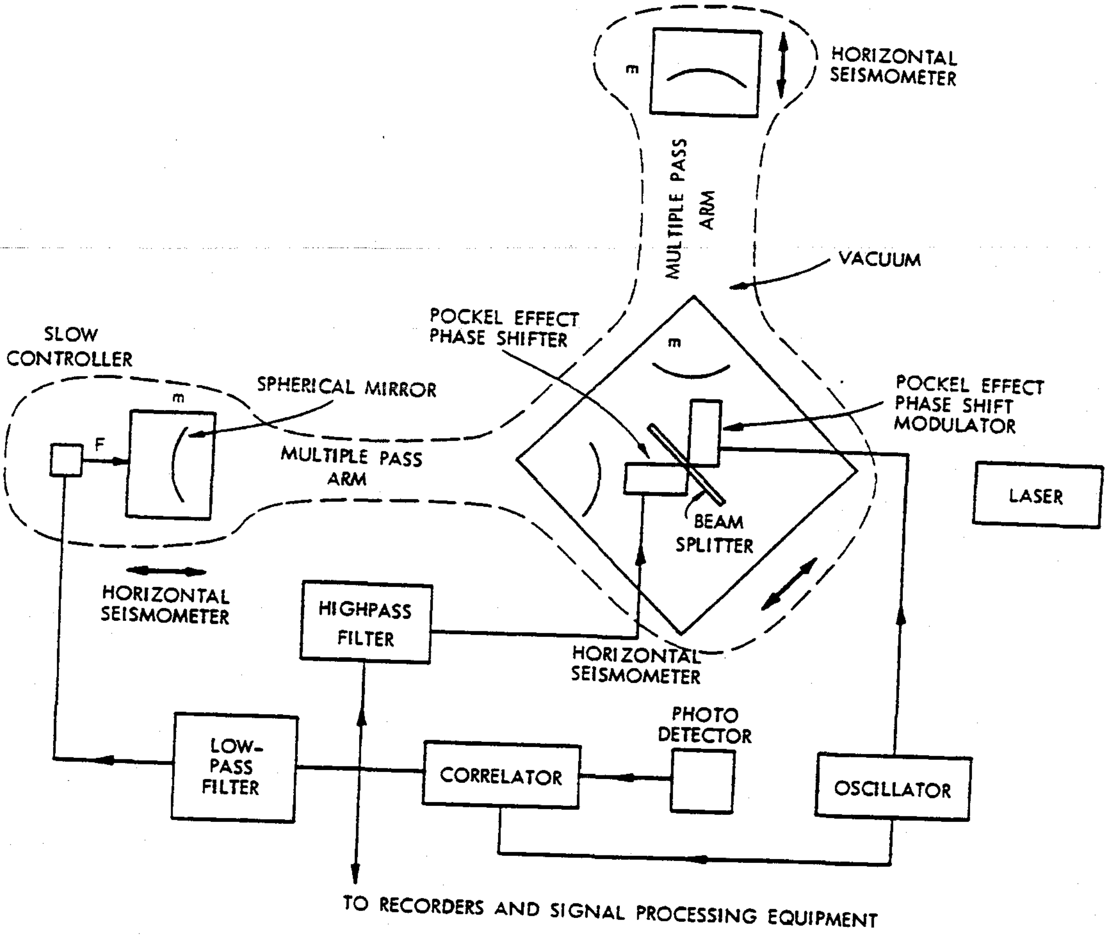


Fig. V-20. Proposed antenna.

that is to be detected. This points out the principal feature of electromagnetically coupled antennas relative to acoustically coupled ones such as bars; that an electromagnetic antenna can be longer than its acoustic counterpart in the ratio of the speed of light to the speed of sound in materials, a factor of 10^5. Since it is not the strain but rather the differential displacement that is measured in these gravitational antennas, the proposed antenna can offer a distinct advantage in sensivity relative to bars in detecting both broadband and single-frequency gravitational radiation. A significant improvement in thermal noise can also be realized.

5. Noise Sources in the Antenna

The power spectrum of noise from various sources in an antenna of the design shown in Fig. V-20 is estimated below. The power spectra are given in equivalent displacements squared per unit frequency interval.

a. Amplitude Noise in the Laser Output Power

The ability to measure the motion of an interferometer fringe is limited by the fluctuations in amplitude of the photo current. A fundamental limit to the amplitude noise in a laser output is the shot noise in the arrival rate of the photons, as well as the noise generated in the stochastic process of detection. At best, a laser can exhibit Poisson amplitude noise. This limit has been approached in single-mode gas lasers that are free of plasma oscillations and in which the gain in the amplifying medium at the frequency of the oscillating optical line is saturated.[12, 13]

The equivalent spectral-noise displacement squared per unit frequency interval in an interferometer of the design illustrated by Fig. V-20, illuminated by a Poisson noise-limited laser and using optimal signal processing, is given by

$$\frac{\Delta x^2(f)}{\Delta f} \geqslant \frac{hc\lambda}{8\pi^2 \epsilon P b^2 \; e^{-b(1-R)}},$$

where h is Planck's constant, c the velocity of light, λ the wavelength of the laser light, ϵ the quantum efficiency of the photodetector, P the total laser output power, b the number of passes in each interferometer arm, and R the reflectivity of the spherical mirrors. The expression has a minimum value for $b = 2/(1-R)$.

As an example, for a 0.5 W laser at 5000 Å and a mirror reflectivity of 99.5% using a photodetector with 50% quantum efficiency, the minimum value of the spectral noise power is

$$\frac{\Delta x^2(f)}{\Delta f} \geqslant 10^{-33} \text{ cm}^2/\text{Hz}.$$

b. Laser Phase Noise or Frequency Instability

Phase instability of the laser is transformed into displacement noise in an interferometer with unequal path lengths. In an ideal laser the phase noise is produced by spontaneous emission which adds photons of random phase to the coherent laser radiation field. The laser phase performs a random walk in angle around the noise-free phase angle given by $\phi_o = \omega_o t$. The variance in the phase grows as $(\overline{\Delta\phi})^2 = t/st_c$, where s is the number of photons in the laser mode, t_c the laser cavity storage time, and t the observation time. This phase fluctuation translates into an oscillating frequency width of the laser given by $\delta = 1/4\pi t_c s$.

Armstrong[14] has made an analysis of the spectral power distribution in the output of a two-beam interferometer illuminated by a light source in which the phase noise has a Gaussian distribution in time. By use of his results, the equivalent power spectrum of displacement squared per unit frequency in the interferometer is given by

$$\frac{\Delta x^2(f)}{\Delta f} = \frac{4}{3}\lambda^2\delta^2\tau^3$$

for the case $f\tau \ll 1$ and $\delta\tau \ll 1$, where τ is the difference in travel time of light between the two paths in the interferometer.

The main reason for using a Michelson interferometer in the gravity antenna is that τ can be made small (equal to zero, if necessary), so that excessive demands need not be made on the laser frequency stability. In most lasers δ is much larger than that because of spontaneous emission, especially for long-term measurements (large τ). For small τ, however, δ does approach the theoretical limit. In a typical case δ might be of the order of 10 Hz and τ approximately 10^{-9} second, which gives

$$\frac{\Delta x^2(f)}{\Delta f} \leqslant 10^{-34}\ \text{cm}^2/\text{Hz}.$$

c. Mechanical Thermal Noise in the Antenna

Mechanical thermal noise enters the antenna in two ways. First, there is thermal motion of the center of mass of the masses on the horizontal suspensions and second, there is thermal excitation of the internal normal modes of the masses about the center of mass. Both types of thermal excitation can be handled by means of the same technique. The thermal noise is modeled by assuming that the mechanical system is driven by a stochastic driving force with a spectral power density given by

$$\frac{\Delta F^2(f)}{\Delta f} = 4kT\alpha \qquad \text{dyn}^2/\text{Hz},$$

where k is Boltzmann's constant, T the absolute temperature of the damping medium, and α the damping coefficient. We can express α in terms of Q, the resonant frequency, ω_o, of the mechanical system, and the mass. Thus $\alpha = m\omega_o/Q$. The spectral power density of the displacement squared, because of the stochastic driving force on a harmonic oscillator, is

$$\frac{\Delta x^2(f)}{\Delta f} = \frac{1}{m^2\omega_o^4}\,\frac{1}{(1-z^2)^2 + z^2/Q^2}\,\frac{4kT\omega_o m}{Q},$$

where $z = \omega/\omega_o$. The seismometer suspension should have a resonant frequency much lower than the frequency of the gravitational wave that is to be detected; in this case $z \gg 1$ and $Q \gg 1$, to give

$$\frac{\Delta x^2(f)}{\Delta f} = 4\,\frac{\omega_o}{\omega^4}\,\frac{kT}{mQ}.$$

On the other hand, the lowest normal-mode frequencies of the internal motions of the masses, including the mirrors and the other suspended optical components, should be higher than the gravitational wave frequency. Some care must be taken to make the entire suspended optical system on each seismometer mount as rigid as possible. For the internal motions $z \ll 1$ and $Q \gg 1$, so that

$$\frac{\Delta x^2(f)}{\Delta f} = \frac{4kT}{\omega_o^3 mQ}.$$

It is clear that, aside from reducing the temperature, the thermal noise can be minimized by using high-Q materials and a high-Q suspension, as long as the gravitational wave frequency does not fall near one of the mechanical resonances. The range of Q for internal motions is limited by available materials: quartz has an internal Q of approximately 10^6, while for aluminum it is of the order of 10^5. The Q of the suspension can be considerably higher than the intrinsic Q of materials. The relevant quantity is the ratio of the potential energy stored in the materials to that stored in the Earth's gravitational field in the restoring mechanism.

The suspensions are critical components in the antenna, and there is no obvious optimal design. The specific geometry of the optics in the interferometer can make the interferometer output insensitive to motions along some of the degrees of freedom of the suspension. For example, the interferometer shown in Fig. V-20 is first-order insensitive to motions of the suspended masses transverse to the direction of propagation of light in the arms. It is also first-order insensitive to rotations of the mirrors. Motions of the beam splitter assembly along the 45° bisecting line of the interferometer produce common phase shifts in both arms and therefore do not appear in the interferometer output. Nevertheless, the success of the antenna rests heavily on the mechanical design of the suspensions because the thermal noise couples in through them, and they also have to provide isolation from ground noise.

The general problem with suspensions is that in the real world they do not have only one degree of freedom but many, and these modes of motion tend to cross-couple nonlinearly with each other, so that, by parametric conversion, noise from one mode appears in another. A rule of thumb, to minimize this problem in suspensions, is to have as few modes as possible, and to make the resonance frequencies of the unwanted modes high relative to the operating mode.[15]

It is still worthwhile to look at an example of the theoretical thermal noise limit of a single-degree-of-freedom suspension. If the internal Q is 10^5, the mass 10 kG, and the lowest frequency resonance in the mass 10 kHz, the thermal noise from internal motions at room temperature for frequencies less than 10 kHz is

$$\frac{\Delta x^2(f)}{\Delta f} \sim 10^{-35}\ \mathrm{cm}^2/\mathrm{Hz}.$$

The thermal noise from center-of-mass motion on the suspension for a $Q \sim 10^4$ and a resonant frequency of 5×10^{-2} Hz becomes

$$\frac{\Delta x^2(f)}{\Delta f} \sim \frac{10^{-24}}{f^4}\ \mathrm{cm}^2/\mathrm{Hz}$$

for frequencies greater than the resonant frequency of the suspension. With the chosen sample parameters, the Poisson noise in the laser amplitude is larger than the thermal noise at frequencies greater than 200 Hz. An antenna that might be used in the pulsar radiation search would require, at room temperature, an mQ product 10^2 larger than the example given, to match the Poisson noise of the laser.

d. Radiation-Pressure Noise from the Laser Light

Fluctuations in the output power of the laser can drive the suspended masses through the radiation pressure of light. In principle, if the two arms of the interferometer are completely symmetric, both mechanically and optically, the interferometer output is insensitive to these fluctuations. Since complete symmetry is hard to achieve, this noise source must still be considered. An interesting point is that although one might find a high modulation frequency for the servo system where the laser displays Poisson noise, it is the spectral power density of the fluctuations in the laser output at the lower frequency of the gravitational wave which excites the antenna. In other words, if this is a serious noise source, the laser has to have amplitude stability over a wide range of frequencies.

Radiation-pressure noise can be treated in the same manner as thermal noise. If the laser displays Poisson noise, the spectral power density of a stochastic radiation-pressure force on one mirror is

$$\frac{\Delta F_{\mathrm{rad}}^2(f)}{\Delta f} = \frac{4b^2hP}{\lambda c} \qquad \mathrm{dyn}^2/\mathrm{Hz},$$

where b is the number of times the light beam hits the mirror, and P is the average total laser power. Using the same sample parameters for the suspension as we used in calculating the thermal noise, and those for the laser in the discussion of the amplitude noise, the ratio of stochastic radiation pressure forces relative to stochastic thermal forces is

$$\frac{\Delta F^2_{\text{rad}}(f)}{\Delta F^2_{\text{thermal}}(f)} \sim 10^{-6}.$$

e. Seismic Noise

If the antenna masses were firmly attached to the ground, the seismic noise, both through horizontal and tilt motions of the ground, would be larger than any of the other noise sources considered thus far. The seismic noise on the earth at frequencies higher than 5 Hz has been studied by several investigators[16-18] at various locations both on the surface and at different depths. In areas far from human industrial activity and traffic, the high-frequency noise can be characterized by a stationary random process. The noise at the surface appears higher than at depths of 1 km or more, but an unambiguous determination of whether the high-frequency noise is due to Rayleigh or to body waves has not been carried out. Measurements made in a zinc mine at Ogdensburg, New Jersey,[16] at a depth of approximately 0.5 km have yielded the smallest published values of seismic noise. In the region 10-100 Hz, the power spectrum is approximated by

$$\frac{\Delta x_\ell^2(f)}{\Delta f} \sim \frac{3 \times 10^{-14}}{f^4} \text{ cm}^2/\text{Hz}.$$

Although the spectrum has not been measured at frequencies higher than 100 Hz, it is not expected to decrease more slowly with frequency at higher frequencies. Surface measurements are typically larger by an order of magnitude.

By mounting the antenna masses on horizontal seismometer suspensions, we can substantially reduce the seismic noise entering the interferometer. The isolation provided by a single-degree-of-freedom suspension is given by

$$\left|\frac{\Delta x_m(f)}{\Delta x_\ell(f)}\right|^2 = \frac{[(1-z^2)+(2/Q)^2]^2+(z^3/Q)^2}{[(1-z^2)^2+(z/Q)^2]^2},$$

where $z = f/f_o$, and f_o is the resonant frequency of the suspension. $\Delta x_m(f)$ is the displacement of an antenna mass at frequency f relative to an inertial frame, and $\Delta x_\ell(f)$ is the motion of the Earth measured in the same reference frame.

At frequencies for which $z \gg 1$, the isolation ratio is

$$\left|\frac{\Delta x_m(f)}{\Delta x_\ell(f)}\right|^2 \sim \left(\frac{f_o}{f}\right)^4 + \left(\frac{f_o}{f}\right)^2 \frac{1}{Q^2}.$$

For the sample suspension parameters given, the estimated seismic noise entering the antenna is

$$\frac{\Delta x^2(f)}{\Delta f} > \frac{2 \times 10^{-18}}{f^8} \text{ cm}^2/\text{Hz}; \quad 10 < f < 10 \text{ kHz}$$

with the average seismic driving noise at the Earth's surface assumed. For frequencies higher than 100 Hz, the effect of seismic noise is smaller than the noise from the laser amplitude fluctuations.

Although the isolation is adequate for detecting Weber-type events, an antenna to detect pulsar radiation would require better rejection of the ground noise. Several approaches are possible. Clearly, the suspension period can be increased to be longer than 20 s, but suspensions of very long periods are difficult to work with. Several shorter period suspensions may be used in series, since their isolation factors multiply. The disadvantage of this is that by increasing the number of moving members, the mode cross-coupling problem is bound to be aggravated.

An interesting possibility of reducing the seismic noise is to use a long-baseline antenna for which the period of the gravitational wave is much shorter than the acoustic travel time through the ground between the antenna end points. In this situation, the sections of ground at the end points are uncoupled from each other and the gravitational wave moves the suspended mass in the same way as the ground around it. In other words, there is little differential motion between the suspended mass and the neighboring ground because of the gravitational wave. Differential motion would result primarily from seismic noise. The differential motion can be measured by using the suspended mass as an inertial reference in a conventional seismometer. This information can be applied to the interferometer output to remove the seismic-noise component.

f. Thermal-Gradient Noise

Thermal gradients in the chamber housing the suspension produce differential pressures on the suspended mass through the residual gas molecules. The largest unbalanced heat input into the system occurs at the interferometer mirror where, after multiple reflections, approximately 1/10 of the laser power will be absorbed.

The excess pressure on the mirror surface is approximately $p \sim nk\Delta T$, where n is the number of gas molecules/cm^3, k is Boltzmann's constant, and ΔT is the difference in temperature between the mirror surface and the rest of the chamber. The fluctuations in ΔT can be calculated adequately by solving the one-dimensional problem of thermal diffusion from the surface into the interior of the mirror and the associated antenna mass, which are assumed to be at a constant temperature.

The mirror surface temperature fluctuations, $\Delta T(f)$, driven by incident intensity fluctuations $\Delta I(f)$, is given by

$$\Delta T(f) = \frac{\Delta I(f)}{4\epsilon\sigma T_o^3 + (\pi c_v \rho k_t)^{1/2} f^{1/2}} .$$

The first term in the denominator is the radiation from the surface, with ϵ the emissivity, σ the Stefan-Boltzmann constant, and T_o the ambient temperature. The second term is due to thermal diffusion from the surface into the interior, with c_v the specific heat, ρ the density, and k_t the thermal conductivity of the mirror.

If the laser exhibits Poisson noise, the spectral force density on the antenna mass becomes

$$\frac{\Delta F^2(f)}{\Delta f} = \frac{2(nk)^2}{f(\pi c_v \rho k_t)} \frac{hc}{\lambda} \overline{P} \qquad \text{dyn}^2/\text{Hz}.$$

Radiation is neglected because it is much smaller than the thermal diffusion. Using the following parameters for glass, $c_v \sim 10^6$ erg/gm °K, $\rho \sim 4$, $k_t \sim 10^3$ erg/s cm °K, an average laser power of 0.5 W and a vacuum of 1×10^{-8} mm Hg, the ratio of the thermal-gradient noise to the thermal noise forces in the sample suspension is

$$\frac{\Delta F^2_{T,G}(f)}{\Delta F^2_{th}(f)} \sim \frac{10^{-15}}{f} .$$

g. Cosmic-Ray Noise

The principal component of the high-energy particle background both below and on the Earth's surface is muons with kinetic energies[19] greater than 0.1 BeV. A muon that passes through or stops in one of the antenna masses imparts momentum to the mass, thereby causing a displacement that is given by

$$\Delta x = \frac{\Delta E \cos \theta}{m\omega_o c} ,$$

where ΔE is the energy loss of the muon in the antenna mass, θ the angle between the displacement and the incident muon momentum, m the antenna mass, and ω_o the suspension resonant frequency.

The energy loss of muons in matter is almost entirely through electromagnetic interactions so that the energy loss per column density, k(E), is virtually constant with energy for relativistic muons. A 10^{-1} BeV muon loses 3 MeV/gm/cm^2, while a 10^4 BeV muon loses ~30 MeV/gm/cm^2.

The vertical flux of muons at sea level with an energy greater than 10^{-1} BeV is approximately 10^{-2} particles/cm^2 sec sr. For energies larger than 10 BeV, the integrated flux varies as $\sim 10^{-1}/E^2$(BeV).

Since the flux falls off steeply with energy and the energy loss is almost independent of energy, the bulk of the muon events will impart the same momentum to the suspension. If we use the following sample suspension parameters, $m \sim 10^4$ g, $f_o \sim 5 \times 10^{-2}$ Hz, $\rho \sim 3$, and typical linear dimensions ~10 cm, the average energy loss per muon is $\sim 10^{-1}$ BeV. At sea level the antenna mass might experience impulsive displacements of $\sim 10^{-18}$ cm occurring at an average rate of once a second. An event arising from the passage of a 10^4 BeV muon results in a displacement of 10^{-17} cm at a rate of once a year.

Although the shape of the antenna mass can be designed to reduce somewhat the effect and frequency of muon interactions especially if we take advantage of the anisotropy of the muon flux, the best way of reducing the noise is to place the antenna masses underground. The pulse rate at depths of 20 m, 200 m, and 2 km is approximately 3×10^{-2}, 10^{-4}, 10^{-9} pulses/second.

If the antenna output is measured over times that include many muon pulses, as it would be in a search for pulsar radiation, the noise can be treated as a stationary distribution. Under the assumption that the muon events are random and, for ease of calculation, that the magnitude of the momentum impacts is the same for all muons, the spectral power density of displacement squared of the antenna mass is

$$\frac{\Delta x^2(f)}{\Delta f} = \frac{4N(\Delta E/c)^2}{(2\pi)^4 m^2 f^4} \qquad \text{cm}^2/\text{Hz}$$

for $f \gg f_o$, where N is the average number of pulses per second, $\Delta E/c$ the momentum imparted to the mass per pulse, and m the antenna mass. For the sample suspension parameters at sea level

$$\frac{\Delta x^2(f)}{\Delta f} \sim 10^{-40}/f^4 \qquad \text{cm}^2/\text{Hz}.$$

h. Gravitational-Gradient Noise

The antenna is sensitive to gravitational field gradients, that is, differential gravitational forces exerted on the masses defining the ends of the interferometer arms. No data are available concerning high-frequency gravitational gradients that occur naturally on or near the surface of the earth. Two effects can bring about gravitational-gradient noise: first, time-dependent density variations in both the atmosphere and the ground, and second, motions of existing inhomogeneities in the mass distribution around the antenna.

An estimate of these two effects can be made with a crude model. Assume that one of the antenna masses is at the boundary of a volume that has a fluctuating density. The amount of mass that can partake in a coherent density fluctuation at frequency f and exert a force on the mass is roughly that included in a sphere with a radius equal to half the acoustic wavelength, λ, in the ground. The fluctuating gravitational force on the mass is

$$\frac{F_g(f)}{m} \sim \frac{2}{3} \pi\lambda\Delta\rho(f)\, G,$$

where $\Delta\rho(f)$ is the density fluctuation at frequency f, and G the Newtonian gravitational constant. The density fluctuations driven by ground noise in the sphere are

$$\Delta\rho(f) = 3\langle \rho \rangle \frac{\Delta x_e(f)}{\lambda},$$

where $\langle \rho \rangle$ is the average density of the ground, and $\Delta x_e(f)$ is the ground noise displacement. If f is larger than the resonant frequency of the suspension, the ratio of the displacement squared of the mass to that of the ground motion is given by

$$\frac{\Delta x_m^2(f)}{\Delta x_e^2(f)} = \left[\frac{\langle \rho \rangle G}{2\pi f^2}\right]^2 .$$

For the earth, this isolation factor is

$$\frac{\Delta x_m^2(f)}{\Delta x_e^2(f)} \sim \frac{10^{-14}}{f^4},$$

which is much smaller than the isolation factor for the attenuation of direct ground motion by the sample suspension.

A comparable approach can be used in estimating the effect of motions of inhomogeneities in the distribution of matter around the antenna which are driven by ground noise. If we assume an extreme case of a complete inhomogeneity, for example, an atmosphere-ground interface, the mass that partakes in a coherent motion, $\Delta x(f)$, could be $m \sim \lambda^3\langle \rho \rangle$. The fluctuating force on the nearest antenna mass is

$$\frac{F_g(f)}{m} = \frac{2}{3} \pi G\langle \rho \rangle\, \Delta x(f).$$

The isolation factor is

$$\frac{\Delta x_m^2(f)}{\Delta x_e^2(f)} \sim \left[\frac{G\langle \rho \rangle}{6\pi f^2}\right]^2,$$

which is comparable to the isolation factor attributable to density fluctuations. These factors become smaller if the distance between the masses is less than λ.

i. Electric Field and Magnetic Field Noise

Electric fields in dielectric-free conducting vacuum chambers are typically 10^{-3} V/cm. These fields result from variations in the work function of surfaces and occur even when all surfaces in a system are constructed of the same material, since the work function of one crystal face is different from that of another. Temporal fluctuations in these fields are caused by impurity migrations and variations in adsorbed gas layers. Little is known about the correlation time of these fluctuations, except that at room temperature it seems to be longer than a few seconds and at cryogenic temperatures it is possible to keep the fields constant to better than 10^{-12} V/cm for several hours.[20]

The electric force on a suspended antenna mass is

$$F_e \sim \frac{1}{4\pi} \mathcal{E}^2 A,$$

where A is the exposed antenna surface, and $\mathcal{E}$ is the fluctuating electric field at the surface. Under the assumption that the power spectrum of the field fluctuations is similar to that of the flicker effect in vacuum tubes or to the surface effects in semiconductors, both of which come from large-scale, but slow, changes in the surface properties of materials, the electric force power spectrum might be represented by

$$\frac{\Delta F_e^2(f)}{\Delta f} \sim \frac{\frac{2}{\pi}\langle F_e^2\rangle\, 1/\tau_o}{(1/\tau_o)^2 + (2\pi f)^2} \qquad \text{dyn}^2/\text{Hz},$$

where τ_o is the correlation time of the fluctuations, and $\langle F_e^2\rangle$ is the average electric force squared.

If the gravitational wave frequency is much greater than $1/\tau_o$ and also higher than the resonant frequency of the suspension, the power spectrum of the displacements squared becomes

$$\frac{\Delta x^2(f)}{\Delta f} = \frac{\langle \mathcal{E}^4\rangle A^2}{32\pi^6 m^2 \tau_o f^4} \qquad \text{cm}^2/\text{Hz}.$$

For $m \sim 10^4$ gm, $A \sim 10^2$ cm^2, $\mathcal{E} \sim 10^{-5}$ stat V/cm and $\tau_o \sim 1$ s,

$$\frac{\Delta x^2(f)}{\Delta f} \sim 10^{-38}/f^4 \qquad \text{cm}^2/\text{Hz}.$$

This noise is considerably less than that from the Poisson noise of the laser. Nevertheless, it is necessary to take care to shield, electrostatically, the deflection mirror surfaces.

Geomagnetic storms caused by ionospheric currents driven by the solar wind and cosmic rays create fluctuating magnetic fields at the surface of the Earth. The smoothed power spectrum of the magnetic field fluctuations in mid-latitude regions at frequencies greater than 10^{-3} Hz is approximately[21]

$$B^2(f) \sim B_o^2/f^2 \qquad G^2/\text{Hz},$$

with $B_o \sim 3 \times 10^{-8}$ G. Large pulses with amplitudes $\sim 5 \times 10^{-3}$ G are observed occasionally; the rise time of these pulses is of the order of minutes.[22]

Fluctuating magnetic fields interact with the antenna mass primarily through eddy currents induced in it or, if it is constructed of insulating material, in the conducting coating around the antenna that is required to prevent charge buildup. The interaction, especially at low frequencies, can also take place through ferromagnetic impurities in nonmagnetic materials. Magnetic field gradients cause center-of-mass motions of the suspended mass. Internal motions are excited by magnetic pressures if the skin depth is smaller than the dimensions of the antenna mass.

In an extreme model it would be assumed that the fluctuating magnetic fields are completely excluded by the antenna mass and that the field changes over the dimensions of the mass are equal to the fields. The magnetic forces are $F_m = \frac{1}{4\pi} B^2 A$.

The power spectrum for center-of-mass motions, with $f \gg f_o$, becomes

$$\frac{\Delta x^2(f)}{\Delta f} = \frac{A^2 B_o^4}{16\pi^3 m^2 f^4} \qquad \text{cm}^2/\text{Hz}.$$

For the sample suspension, using the smoothed power spectrum of magnetic field fluctuations, we have

$$\Delta x^2(f) \sim 10^{-36}/f^4 \qquad \text{cm}^2/\text{Hz}.$$

The displacements arising from internal motions driven by magnetic pressures at frequencies lower than the internal resonant frequency, $f_{o_{int}}$, are given by

$$\frac{\Delta x^2(f)}{\Delta f} = \frac{A^2 B_o^4}{16\pi^3 m^2 f_{o_{int}}^2 f^2} \qquad \text{cm}^2/\text{Hz}.$$

Although the disturbances caused by the smoothed power spectrum do not appear troublesome in comparison with the other noise sources, the occasional large magnetic pulses will necessitate placing both conducting and high-μ magnetic shields around the antenna masses. (It is not inconceivable that Weber's coincident events may be caused by pulses in geomagnetic storms, if his conducting shielding is inadequate. It would require a pulse of 10^{-2} G with a rise time $\sim 10^{-3}$ s to distort his bars by $\Delta\ell/\ell \sim 10^{-16}$.)

6. Detection of Gravitational Waves in the Antenna Output Signal

The interferometer (servo) output signal is filtered after detection. The gravitational wave displacements in the filtered output signal are given by

$$\Delta x_g^2 = \frac{1}{4} \int_0^\infty |F(f)|^2\, h^2(f)\, \ell^2\, df,$$

where F(f) is the filter spectral response, $h^2(f)$ is the spectral power density of the gravitational wave metric components, and ℓ is the arm length of the antenna interferometer. The noise displacements in the filtered output signal are given by

$$\Delta x_n^2 = \int_0^\infty |F(f)|^2 \frac{\Delta x_n^2(f)}{\Delta f}\, df,$$

where $\Delta x_n^2(f)/\Delta f$ is the spectral power density of the displacement noise. In order to observe a gravitational wave, the signal-to-noise ratio has to be greater than 1. That is, $\Delta x_g^2/\Delta x_n^2 > 1$.

The dominant noise source for the antenna appears to be the amplitude fluctuations in the laser output power. When translated into equivalent displacement of the masses, the noise has been shown to have a flat spectrum given by $\Delta x_n^2(f)/\Delta f \sim 10^{-33}$ cm^2/Hz.

If we assume this noise and an idealized unity gain bandpass filter with cutoff frequencies f_2 and f_1, then the signal-to-noise ratio becomes

$$\frac{\Delta x_g^2}{\Delta x_n^2} = \frac{1/4 \int_{f_1}^{f_2} h_2(f)\, \ell^2\, df}{\frac{\Delta x_n^2(f)}{\Delta f}\,(f_2 - f_1)}.$$

For continuous gravitational waves, the minimum detectable gravitational wave metric spectral density is then

$$h^2(f) > \frac{4}{\ell^2} \frac{\Delta x_n^2(f)}{\Delta f} \approx \frac{4 \times 10^{-33}}{\ell^2(\text{cm})} \qquad \text{Hz}^{-1}.$$

Detectability criteria for pulses cannot be so well defined; a reasonable assumption is that the pulse "energy" be equal to the noise "energy." The optimum filter should have a bandwidth comparable to the pulse bandwidth. The spectral density of a pulse of duration τ is roughly distributed throughout a $1/\tau$ bandwidth. A possible signal-to-noise criterion for pulses is then

$$\Delta x_g^2 \tau > \frac{\Delta x_n^2(f)}{\Delta f},$$

or in terms of h,

$$h^2 \tau > \frac{4}{\ell^2} \frac{\Delta x_n^2(f)}{\Delta f}.$$

As an example, the Weber pulses induce impulsive strains of $h \sim 2 \times 10^{-16}$ for a duration of approximately 10^{-3} s, so that $h^2\tau \sim 4 \times 10^{-35}$. A 1-m interferometer arm antenna of the proposed design would have a noise "energy" of 4×10^{-37}, so that the signal-to-noise ratio for Weber events would approach 100/1.

A meaningful search for the pulsar radiation requires a more elaborate and considerably more expensive installation. The spectral density of the pulsar gravitational wave metric is

$$h^2(f) = h_o^2 \delta(f-f_p),$$

where f_p is a multiple of the pulsar rotation frequency. The signal-to-noise ratio is

$$\frac{\Delta x_g^2}{\Delta x_n^2} = \frac{1/4\, h_o^2 \ell^2}{\frac{\Delta x_n^2}{\Delta f}(f_2-f_1)}.$$

By coherent amplitude detection, using a reference signal at multiples of the pulsar rotation frequency, we can reduce the filter bandwidth by increasing the postdetection integration time. The integration time, t_{int}, required to observe the pulsar radiation with a signal-to-noise ratio greater than 1 is given by

$$t_{int} > \frac{4\, \frac{\Delta x_n^2(f_p)}{\Delta f}}{h_o^2 \ell^2}.$$

Assuming the Gunn-Ostriker upper limit for the gravitational radiation of the Crab Nebula pulsar, $h_o \sim 2 \times 10^{-24}$, and an antenna with a 1-km interferometer arm, we find that the integration time is around one day.

An interesting point, suggested by D. J. Muehlner, is that the Weber events, if they are gravitational radiation pulses, could constitute the dominant noise in a pulsar radiation search. Under the assumption that the Weber pulses cause steplike strains, h_o, at an average rate of n per second, and that the integration time includes many pulses, the power spectrum of displacement squared is given roughly by

$$\frac{\Delta x^2(f)}{\Delta f} \sim \frac{N\ell^2 h_o^2}{16\pi^2 f^2} \qquad \text{cm}^2/\text{Hz}.$$

With $f \sim 60$ Hz, $h_o \sim 10^{-16}$, $\ell \sim 10^5$ cm, and $N \sim 10^{-5}$/s, the noise is $\sim 10^{-32}$ cm^2/Hz, which is greater than the Poisson noise of the laser. Large pulses can be observed directly in the broadband output of the antenna and can therefore be removed in the data analysis of the pulsar signal. If the energy spectrum of gravitational radiation pulses is, however, such that there is a higher rate for lower energy pulses, in particular, if Nh_o^2 is constant as h_o gets smaller, gravitational radiation may prove to be the dominant noise source in the pulsar radiation measurements.

Appendix

Comparison of Interferometric Broadband and Resonant Bar Antennas for Detection of Gravitational Wave Pulses

Aside from their greater possible length, interferometric broadband antennas have a further advantage over bars, in that the thermal noise in the detection bandwidth for the gravitational wave pulse is smaller than for the bar. In the following calculation it is assumed that the thermal noise is the dominant noise in both types of antennas.

Let the gravitational radiation signal be a pulse given by

$$h(t) = \begin{cases} 0 & t < 0 \\ h & 0 \leqslant t \leqslant t_o \\ 0 & t > t_o \end{cases}$$

The spectral energy density of the pulse is

$$h^2(\omega) = \frac{2h^2t_o^2}{(2\pi)^2}\,\frac{\sin^2 \omega t_o/2}{(\omega t_o/2)^2} \qquad 0 \leqslant \omega < \infty$$

$$\cong \begin{cases} \dfrac{2h^2t_o^2}{(2\pi)^2} & 0 < \omega \leqslant \pi/t_o \\ 0 & \omega > \pi/t_o \end{cases} \qquad \text{the equivalent energy box spectrum}$$

The gravitational force spectral density is

$$F_g^2(\omega) = \frac{1}{4}\omega^4 h^2(\omega)\,\ell^2 m^2.$$

Using the dynamic interpretation for the interaction of the bar with the gravitational-wave pulse, the "energy" in the bar after the pulse excitation is given by

$$\int_0^\infty x_g^2(t)\,dt = E_g = \frac{h^2 t_o^2 \ell^2 \omega_o Q}{4\pi} \qquad Q \gg 1, \quad \frac{\pi}{t_o} > \omega_o,$$

where ω_o is the resonant frequency of the bar.

The pulse "energy" is distributed throughout the ringing time of the bar so that

$$E_g \sim x_g^2(t)\,2Q/\omega_o,$$

and the average displacement of the ends of the bar becomes

$$x_g^2(t) \sim \frac{h^2 t_o^2 \omega_o^2 \ell^2}{8\pi}.$$

The average thermal-noise displacement is

$$\langle x_{TH}^2\rangle \sim \frac{4kT}{m\omega_o^2}.$$

The thermal noise also rings on the average for a period $\tau \sim 2Q/\omega_o$.

The signal-to-noise ratio for the bar is given by

$$\frac{x_g^2}{\langle x_{TH}^2\rangle} \sim \frac{h^2\ell^2(t_o\omega_o)^2\,m\omega_o^2}{32\pi kT}.$$

Now make the same calculation for the broadband antenna with a filter matched to the pulse spectrum. The displacement spectrum is

$$x^2(\omega) = h^2(\omega)\,\ell^2 = \frac{2h^2 t_o^2 \ell^2}{(2\pi)^2}.$$

The pulse "energy" in a filter with matched bandwidth and a low-frequency cutoff ω_L is

$$E_g = 2\pi \int_{\omega_L}^{\pi/t_o} x^2(\omega)\, d\omega \cong \ell^2 h^2 t_o \qquad \omega_L \ll \pi/t_o.$$

If the resonant frequency, ω_o, of the suspension is smaller than ω_L, and the suspension has a high Q, the thermal "energy" in the same bandwidth is given by

$$E_{TH} = \langle x_{TH}^2 \rangle\, t_o = t_o \int_{\omega_L}^{\pi/t_o} \frac{1}{m^2\omega^4} \frac{4kT\omega_o m}{Q}\, d\omega.$$

Generally $\omega_L \ll \frac{\pi}{t_o}$, the thermal "energy" becomes

$$E_{TH} \sim \frac{4kT\omega_o t_o}{3Qm\omega_L^3}.$$

The signal-to-noise ratio for the broadband antenna is

$$\frac{E_g}{E_{TH}} = \frac{x_g^2}{\langle x_{TH}^2 \rangle} = \frac{3h^2\ell^2 Qm\omega_L^3}{4kT\omega_o}.$$

The signal-to-noise ratio for the broadband antenna relative to the equivalent-length resonant bar antenna at the same temperature is

$$R = \frac{(S/N)_{BB}}{(S/N)_B} = \frac{24\pi Q_{BB} m_{BB} \omega_L^3/\omega_{oBB}}{(t_o\omega_{oB})^2\, m_B \omega_{oB}^2}.$$

The best case for the bar is a pulse with $t_o \sim \frac{\pi}{\omega_o}$. If we assume Weber bar parameters $m_B \sim 10^6$ g, $\omega_{oB} \sim 10^4$ and the sample suspension parameters previously given, $Q_{BB} \sim 10^4$, $m_{BB} \sim 10^4$ g, $\omega_L \sim 10^3$, $\omega_{oBB} \sim 3 \times 10^{-1}$, the signal-to-noise ratio approaches $\sim 10^4$. This entire factor cannot be realized because the laser amplitude noise dominates in the interferometric antenna.

R. Weiss

References

1. A. Einstein, Sitzber. deut. Akad. Wiss. Berlin, Kl. Math. Physik u. Tech. (1916), p. 688; (1918), p. 154.
2. J. Weber, Phys. Rev. Letters 22, 1320 (1969).
3. J. Weber, Phys. Rev. Letters 25, 180 (1970).

4. J. Weber, Phys. Rev. 117, 306 (1960).

5. L. D. Landau and E. M. Lifshitz, The Classical Theory of Fields (Pergamon Press, London and New York, 1962).

6. T. Gold, Nature 218, 731 (1968).

7. F. Pacini, Nature 219, 145 (1968).

8. J. P. Ostriker and J. E. Gunn, Astrophys. J. 157, 1395 (1969).

9. F. A. E. Pirani, Acta. Phys. Polon. 15, 389 (1956).

10. D. R. Herriott and H. J. Schulte, Appl. Opt. 4, 883 (1965).

11. M. S. Wagner, S.B. Thesis, Department of Physics, M.I.T., June 1971 (unpublished).

12. G. Blum and R. Weiss, Phys. Rev. 155, 1412 (1967).

13. G. F. Moss, L. R. Miller, and R. L. Forward, Appl. Opt. 10, 2495 (1971).

14. J. A. Armstrong, J. Opt. Soc. Am. 56, 1024 (1966).

15. R. Weiss and B. Block, J. Geophys. Res. 70, 5615 (1965).

16. B. Isacks and J. Oliver, Bull. Seismol. Soc. Am. 54, 1941 (1964).

17. G. E. Frantti, Geophys. 28, 547 (1963).

18. E. J. Douze, Bull. Seismol. Soc. Am. 57, 55 (1967).

19. M. G. K. Menon and P. V. Ramana Murthy, in Progress in Elementary Particle and Cosmic Ray Physics, Vol. 9 (North-Holland Publishing Co., Amsterdam, 1967).

20. F. C. Witteborn and W. M. Fairbank, Phys. Rev. Letters 19, 1049 (1967).

21. W. H. Campbell, Ann. Geophys. 22, 492 (1966).

22. T. Sato, Rep. Ionosphere Space Res. Japan 16, 295 (1962).

REF. YY

PHYSICAL REVIEW D VOLUME 17, NUMBER 2 15 JANUARY 1978

Wideband laser-interferometer graviational-radiation experiment

Robert L. Forward

Hughes Research Laboratories, Malibu, California 90265

(Received 12 September 1977)

A wideband laser-interferometer gravitational-radiation antenna was constructed and used to search for gravitational radiation in the frequency band from 1 to 20 kHz. The antenna consisted of a Michelson interferometer with the beamsplitter and retroreflectors attached to masses on soft suspensions that allowed essentially free motion above the suspension frequencies. The strains in the gravitational radiation produce a differential path length change in the two arms of the interferometer which is detected by a pair of balanced photodetectors. The interferometer used a folded-path configuration with an effective length of 8.5 m. The sensitivity of the interferometer was calibrated with signals from a piezoelectric displacement transducer. The strain noise in a 1-Hz bandwidth was less than 0.3 fm/m from 1 to 3 kHz, and less than 0.1 fm/m above 3 kHz, where it was essentially photon-noise limited. (For comparison, the kT strain noise in a room-temperature, 2-m long, 1000-kg, elastic solid bar antenna is 0.14 fm/m.) The laser interferometer was operated as a detector for gravitational radiation for 150 h during the nights and weekends from the period 4 October through 3 December 1972. During the same period, bar antennas were operated by the Maryland, Glasgow, and Frascati groups, with 18 events reported by the Frascati group in their single bar, 22 single-bar events and no coincidences reported by the Glasgow group in their two bars, and 28 coincidences reported by the Maryland group between the Argonne bar and the Maryland bar and/or disk antennas. The various bar antenna systems were quite different but in general were sensitive to gravitational-radiation strain spectral components with an amplitude of the order of 0.1 fm/m in a narrow band of frequencies about the resonant frequency of the bar. The wideband interferometer data was analyzed by ear, with the detection sensitivity estimated to be of the order of 1–10 fm/m (depending upon the signature of the signal) for the total of the gravitational-radiation strain spectral components in the band from 1–20 kHz. No significant correlations between the Malibu interferometer output and any of tne bar events or coincidences were observed.

I. WIDEBAND INTERFEROMETER ANTENNA

The wideband laser-interferometer gravitational-radiation antenna consists of a laser-excited Michelson interferometer with the beamsplitter and reflectors attached to kilogram-sized masses on soft suspensions that allow essentially free motion above the suspension frequencies. As shown in Fig. 1, when the direction of gravitational radiation is along one of the interferometer arms, that arm does not experience any differential motion between the beamsplitter and retroreflector, and thus acts as a reference arm for the interferometer. The strains produced by the gravitational radiation (of proper polarization) then act on the other arm of the interferometer, causing a differential ac motion of the beamsplitter and retroreflector at the frequency of the gravitational radiation.[1,2] When the direction of the gravitational radiation is at right angles to the plane of the interferometer, one arm will decrease in length and the other will increase, resulting in a doubling of the signal.

A. Coupling of radiation to antenna

Gravitational radiation couples to the laser-interferometer antenna by causing relative motion between the beamsplitter and mirrors in the interferometer. The strains and relative motions induoed by the gravitational radiation are readily derived from the weak-field approximation to the Einstein field equations,

$$R_{\alpha\beta} - \tfrac{1}{2} g_{\alpha\beta} R = \frac{8\pi G}{c^4} T_{\alpha\beta} \ . \tag{1}$$

The gravitational-radiation field is assumed to be a weak perturbation on the metric $g_{\alpha\beta} = \eta_{\alpha\beta} + h_{\alpha\beta}$, that propagates as a tensor wave:

$$\frac{\partial^2 h_{\alpha\beta}}{\partial x_\alpha{}^2} - \frac{1}{c^2}\frac{\partial h_{\alpha\beta}}{\partial t^2} = 0 \ . \tag{2}$$

The general form of a gravitational wave propagating in the z direction will have the character[3]

$$h_{\alpha\beta} = (h_t t_{\alpha\beta} + h_s s_{\alpha\beta}) e^{-i(\omega t - kz)} \ , \tag{3}$$

where h_t and h_s are the scalar amplitudes of the two states of polarization of the wave. For linearly polarized radiation one state of polarization is the tension-compression polarization represented by

$$t_{\alpha\beta} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{4}$$

Copyright 1978 by The American Physical Society

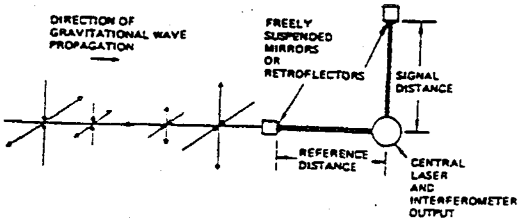

DIRECTION OF GRAVITATIONAL WAVE PROPAGATION
FREELY SUSPENDED MIRRORS OR RETROFLECTORS
SIGNAL DISTANCE
REFERENCE DISTANCE
CENTRAL LASER AND INTERFEROMETER OUTPUT

FIG. 1. Wideband laser-interferometer antenna.

or $h_{11} = -h_{22}$, which consists of a tension along the x axis and a compression along the y axis [see Fig. 2(a)].

The other state of polarization is the shear polarization represented by

$$s_{\alpha\beta} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \qquad (5)$$

or $h_{12} = h_{21}$, which consists of shear forces along the bisectors of the x and y axes [see Fig. 2(b)]. From inspection it is easily seen that the two polarizations are orthogonal since $s_{\alpha\beta}t^{\alpha\beta} = 0$ and that the shear polarization $s_{\alpha\beta}$ is just the tension polarization rotated through 45° (demonstrating the spin-2 characteristics of the radiation).

If desired, the flux of this gravitational radiation in power per unit area can be obtained from the relation[4]

$$I = \frac{c^3}{16\pi G}\left[\dot{h}_{12}{}^2 + \tfrac{1}{4}(\dot{h}_{11} - \dot{h}_{22})^2\right] . \qquad (6)$$

Each arm of the interferometer consists of a pair of freely suspended masses holding a beam-splitter or a retroreflector and separated by a distance $\zeta^\beta = l^\beta + \xi^\beta(t)$, where l^β is the nominal separation distance of the masses without excitation, and $\xi^\beta(t)$ is the time-varying portion of the displacement caused by the radiation.

The gravitational radiation interacts with the pair of masses through the equation of geodesic deviation for small, nonrelativistic motions as follows:

$$\frac{d^2\zeta^\beta}{dt^2} = -c^2R^\beta{}_{0\alpha0}\zeta^\alpha \approx -c^2R^\beta{}_{0\alpha0}l^\alpha . \qquad (7)$$

If the detecting masses are entirely free, then the equations simplify to give the relative displacement ξ^β between the masses in terms of the gravitational-field strength,

$$\ddot{\xi}^\beta = -c^2R^\beta{}_{0\alpha0}l^\alpha = \tfrac{1}{2}\ddot{h}^\beta{}_\alpha l^\alpha . \qquad (8)$$

The equivalent strain $\epsilon^\beta{}_\alpha = \xi^\beta / l^\alpha$ over the distance l^α is then given by

$$\ddot{\epsilon}^\beta{}_\alpha = -c^2R^\beta{}_{0\alpha0} = \tfrac{1}{2}\ddot{h}^\beta{}_\alpha , \qquad (9)$$

which shows that the strain is a direct measure of the gravitational-field strength.

The Michelson interferometer used as the antenna consists of two orthogonal arms of the same nominal length l. The interferometer produces a change in output when there is a difference in the two path lengths. For an antenna with one arm along the x axis and the other along the y axis, the output is

$$\xi^1 - \xi^2 = \tfrac{1}{2}(h^1{}_1 l^1 - h^2{}_2 l^2) \qquad (10)$$

or

$$\Delta\xi = \tfrac{1}{2}(h_{11} - h_{22})l = lh , \qquad (11)$$

where $h_{\alpha\beta}$ is measured in the coordinate system of the antenna. This scalar output can be obtained by assuming a tensor format for the combined response of the two arms of the antenna

$$A^{\alpha\beta} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} l , \qquad (12)$$

and carrying out the operation

$$\Delta\xi = \tfrac{1}{2}h_{\alpha\beta}A^{\alpha\beta} . \qquad (13)$$

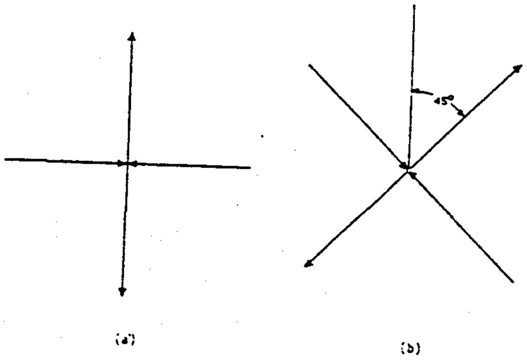


FIG. 2. The two states of polarization of tensor gravitational radiation.

B. Detection-sensitivity pattern

The detection-sensitivity pattern of the laser-interferometer antenna for the gravitational radiation of different polarizations coming from different directions is a complex one.

In Fig. 3 we assume that the laser-interferometer antenna lies in the x-y plane with one arm along the x axis and the other along the y

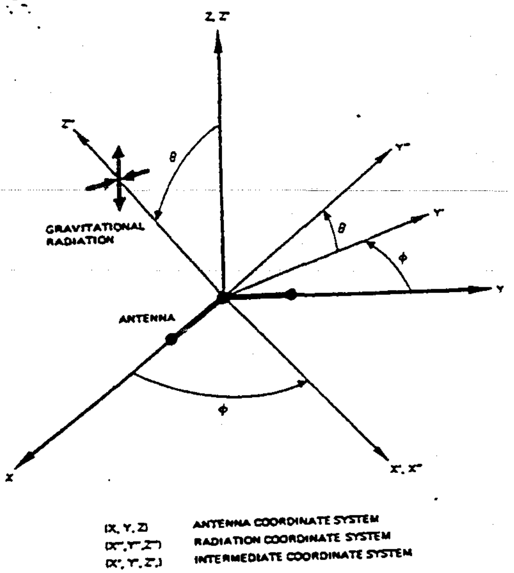

Z, Z'
Z''
θ
GRAVITATIONAL RADIATION
Y''
θ
Y'
φ
Y
ANTENNA
φ
X
X', X''
(X, Y, Z) ANTENNA COORDINATE SYSTEM
(X'', Y'', Z'') RADIATION COORDINATE SYSTEM
(X', Y', Z') INTERMEDIATE COORDINATE SYSTEM

FIG. 3. Coordinates for calculation of interaction of antenna with radiation.

axis, while the gravitational radiation is arriving from an arbitrary direction (θ, ϕ) with an arbitrary polarization. In the coordinate system (x'', y'', z'') of the gravitational radiation, we can separate the radiation into its two linear polarizations $t_{\alpha\beta}$, with the tension lying in the x-y plane of the antenna, and $s_{\alpha\beta}$ at 45° to it.

To convert the two tensor polarizations from the radiation-coordinate system to the antenna-coordinate system, we use the general form of the rotation matrix[5] with $\psi = 0$:

$$R^\alpha_\beta = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\phi & \sin\phi & 0 \\ 0 & -\cos\theta\sin\phi & \cos\theta\cos\phi & \sin\theta \\ 0 & \sin\theta\sin\phi & -\sin\theta\cos\phi & \cos\theta \end{bmatrix} . \tag{14}$$

Plane-polarized radiation propagating along the z'' direction and polarized along the x''-y'' direction

$$h''_{\alpha\beta} = h_t t''_{\alpha\beta} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} h_t , \tag{15}$$

is then converted by the rotation matrix into the antenna-coordinate system by the operations

$$t_{\alpha\beta} = R^{-1}{}_\alpha{}^\gamma t''_{\gamma\delta} R^\delta{}_\beta , \tag{16}$$

or

$$t_{\alpha\beta} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \cos^2\phi - \cos^2\theta\sin^2\phi & (1+\cos^2\theta)\sin\phi\cos\phi & \sin\theta\cos\theta\sin\phi \\ 0 & (1+\cos^2\theta)\sin\phi\cos\phi & \sin^2\phi - \cos^2\theta\cos^2\phi & -\sin\theta\cos\theta\cos\phi \\ 0 & \sin\theta\cos\theta\sin\phi & -\sin\theta\cos\theta\cos\phi & -\sin^2\theta \end{bmatrix} . \tag{17}$$

The antenna response to this polarization is

$$\Delta\xi = \tfrac{1}{2} h_{\alpha\beta} A^{\alpha\beta} = \tfrac{1}{2} h_t t_{\alpha\beta} A^{\alpha\beta}$$
$$= \tfrac{1}{2} h_t l (1 + \cos^2\theta) \cos 2\phi . \tag{18}$$

The shape of the antenna-response pattern in the azimuthal and polar directions are given in Fig. 4.

Plane-polarized radiation propagating along the z'' direction and polarized in the shear direction to the x''-y'' axes,

$$h''_{\alpha\beta} = h_s s''_{\alpha\beta} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} h_s , \tag{19}$$

is converted by the rotation matrix into the antenna-coordinate system by

$$s_{\alpha\beta} = R^{-1}{}_\alpha{}^\gamma s''_{\gamma\delta} R^\delta{}_\beta \tag{20}$$

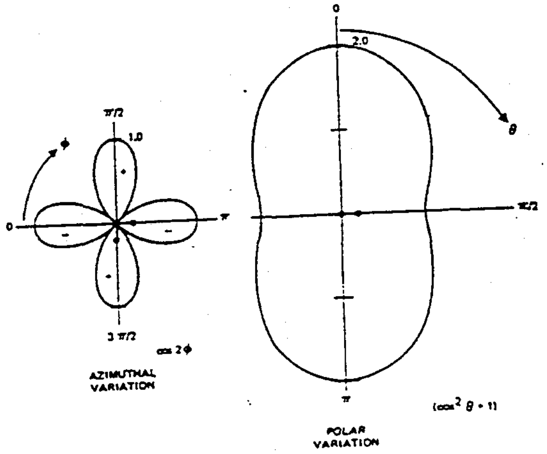


FIG. 4. Detection sensitivity pattern for linearly polarized gravitational radiation with one direction of polarization in plane of the antenna.

or

$$s_{\alpha\beta} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & -\sin 2\phi \cos\theta & \cos 2\phi \cos\theta & \cos\phi \sin\theta \\ 0 & \cos 2\phi \cos\theta & \sin 2\phi \cos\theta & \sin\phi \sin\theta \\ 0 & \cos\phi \sin\theta & \sin\phi \sin\theta & 0 \end{bmatrix}, \tag{21}$$

and the antenna response to the polarization is

$$\Delta\xi = \tfrac{1}{2} h_{\alpha\beta} A^{\alpha\beta} = \tfrac{1}{2} h_s s_{\alpha\beta} A^{\alpha\beta}$$
$$= -h_s l \cos\theta \sin 2\phi \ . \tag{22}$$

The antenna response patterns in the polar and azimuthal directions are shown in Fig. 5.

When we analyze the response of each arm individually we see that each has a $\cos^2\theta$ radiation pattern, similar to that of a simple mass quadrupole, and is the same as that of an elastic solid bar antenna (neglecting Poisson ratio effects). The combined antenna pattern is thus seen to be the coherent addition of two orthogonal $\cos^2\theta$ type patterns.

The total combined response of the antenna to the radiation of arbitrary polarization is a complex shape—an artist's conception of the pattern is shown in Fig. 6.

When the antenna was operated as a detector for gravitational radiation, it was situated in the basement laboratory of the Hughes Research Laboratories, Malibu, California. The Research Laboratories are located at 119° west longitude and 34° north latitude. The arms of the antenna were tangent to the earth's surface and were pointing at 53° and 143° from north.

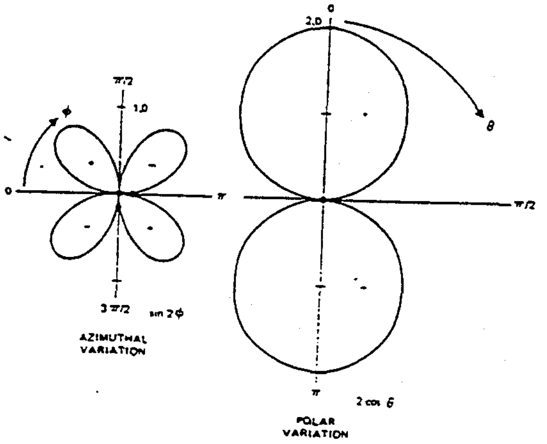


FIG. 5. Detection sensitivity pattern for linearly polarized gravitational radiation with a bisector of the polarization directions in the plane of the antenna.

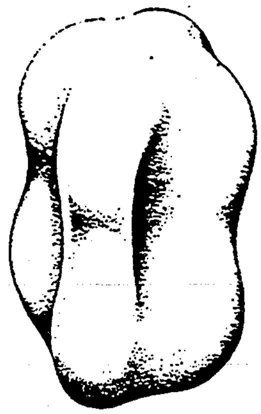


FIG. 6. Radiation pattern of Malibu wideband-interferometric gravitational antenna.

From Fig. 6 we see that the major response of the antenna was in the directions normal to the plane of the antenna. The secondary-response characteristics of the antenna were along the four tangential lobes aligned with the antenna arms. If we translate the antenna pattern to the center of the earth, then the main lobes were at

119° west longitude, 34° north latitude—Malibu,
61° east longitude, 34° south latitude—Southwest Pacific,

while the four tangent lobes lie on a great circle with Malibu as the center and were located at

7° west longitude, 29° north latitude—Morocco,
173° east longitude, 29° south lattitude—New Zealand,
65° west longitude, 43° south latitude—Argentina,
115° east longitude, 43° north latitude—Gobi Desert.

(One pair of antenna nulls was oriented close to the polar directions.)

The New Zealand lobe strongly overlaps the antenna patterns of the Glasgow bar antennas, while the Argentina and Gobi Desert lobes overlap the Maryland bar-antenna pattern.

For the night hours of the fall of 1972, the center of the galaxy passed under the earth so the antenna sensitivity was a maximum for gravitational radiation from that direction. With the combination of the multilobed pattern and the earth rotation, most of the sky (except for the poles) was scanned during the period data was collected.

II. ANTENNA DESIGN

A. Interferometer

The interferometer used was a folded Michelson interferometer. The light source for the interferometer was a large laser on a 3-m granite slab

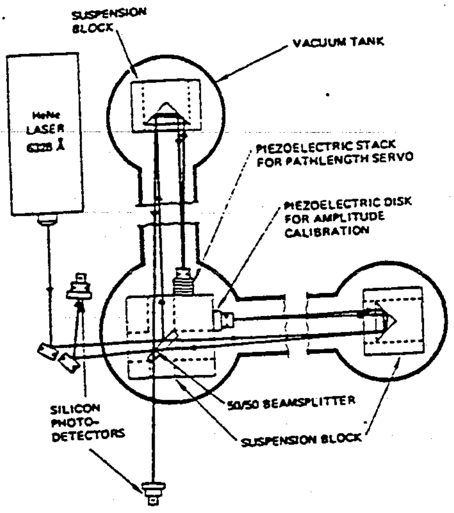


FIG. 7. Schematic of folded optical path.

on low-frequency air mounts. As is shown in Fig 7, the beam from the laser enters through a window in a 0.5-m-diameter cylindrical vacuum tank and is separated into two beams by a 50/50 beamsplitter inside. One beam travels down a 2 m length of evacuated aluminum irrigation pipe to a second cylindrical vacuum tank containing a high-quality optical-corner-cube retroreflector. The beamsplitter and retroreflectors were mounted in support blocks designed to have no internal mechanical resonances below 25 kHz. The reflected beam returns down the pipe to the center vacuum tank where it strikes a mirror mounted on the same support block as the beamsplitter. The mirror is also coupled to a stack of piezoelectric length transducers that are used in a low-frequency servo system to maintain the interferometer pathlength constant at the lower frequencies. The beam returns to the retroreflector where it is again reflected (in the process canceling some of the optical errors in the retroreflector) and sent back down to the beamsplitter.

The second beam from the beamsplitter travels down another evacuated section of pipe to another cylindrical vacuum tank mounted on a separate vibration isolation table. The laser beam in this arm is reflected from the retroreflector and returns to a mirror also mounted on the same block as the beamsplitter. This mirror is coupled to a piezoelectric disk that can be driven with calibrated voltages to induce a known high-frequency mirror motion that can be used to calibrate the interferometer. This beam also goes back through its retroreflector and returns to the beamsplitter where it is combined with the first beam. In this folded configuration, the effective length of each arm is 4.25 m. For gravitational radiation orthogonal to the plane of the interferometer, the effective length of the interferometer is 8.5 m.

The combined beams from the two arms were then brought out through windows where they were detected by a differential pair of silicon photodetectors.

B. Isolation system

The primary isolation table for the interferometer was a 0.3 × 1 × 3 m 2300-kg granite slab on four Firestone 1 × 84D-1 air mounts. The resulting combination had a resonant frequency of approximately 1.5 Hz.

A second isolation table of 0.3 × 1 × 1 m granite slab on similar air mounts was used for support of the other arm of the interferometer. The interferometer was enclosed in a vacuum housing that was kept evacuated to nominal forepump vacuum levels (<100 μ Hg).

The vacuum system and isolation tables were designed so that after an initial checkout and operation with 2-m sections of aluminum irrigation pipe (8.5 m total interferometer pathlength), those sections could be replaced with longer sections (up to 1 km) with a substantial increase in interferometer-gravitational radiation-strain sensitivity for the same photon-noise-limited displacement sensitivity.

C. Retroreflectors

The retroreflectors used at the ends of the arms in the interferometer were 5-cm-diameter fused-silica corner cubes. The return beam from a corner retroreflector is elliptically polarized. If, however, the beam is reflected back on itself, then the ellipticity of the polarization is corrected. This characteristic of the retroreflector and the desire to have all the active components in the system at one location (near the beamsplitter), to keep the remote portions of the antenna as simple as possible, led to the folded-beam configuration of Fig. 7. A flat mirror was used for the intermediate reflection at the end of the beam (back at the beamsplitter) to prevent beam translation with corner rotation.

D. Suspensions

The beamsplitter and the retroreflectors were mounted in holes bored into aluminum cubes approximately 10 cm on a side. With these dimensions, the first longitudinal vibrational mode is about 25 kHz, above the 1 to 20 kHz search band. No additional modes in the 1 to 20 kHz band should have been generated by the holes. These optical support blocks were placed on top of stacks of alternating 6-mm neoprene rubber pads and 2.5-

cm brass plates stacked to the desired height. These suspensions had a typical frequency of 10 Hz. Although they would not be adequate in a search for radiation at pulsar frequencies 0.1 −100 Hz, they were more than adequate for a search in the 1 to 20 kHz band.

E. Laser

The laser used in the interferometer was a modified Spectra-Physics Model 125, which nominally emits a 65-mW multimode. The laser was first modified by a new etalon design[6] to produce between 35 to 55 mW single mode. The new etalon design had high stability (11 min. between mode hops) and low noise, except when the etalon was exactly on its center mode, when it was noticeably noiser.

The noise in the 1 to 20 kHz region was further reduced by replacing the Spectra-Physics power supply with a Fluke precision high-voltage power supply, and then deliberately inducing a laser-plasma oscillation at 100 kHz. With these modifications, the laser noise was drastically reduced from the normal noise level of the Model 125. Figure 8 shows the laser-noise spectrum for 9 mA of detector current as measured through a 10-Hz bandwidth filter. The actual laser noise above 2 kHz is about 5 times the photon-noise limit of the detected photons.

The noise measurement was then repeated using a beamsplitter and optics to simulate the balanced detector operation of the actual interferometer. Figure 9 shows the laser-noise spectrum as measured by a pair of balanced photodetectors, each carrying about 2.2 mA of photoelectrons. The measured noise is now very close to the photon-noise limit of the detected photons.

F. Photodetection system

The photodetectors used in the interferometer were United Detector Technology Type PIN-25 Schottky barrier photodiodes. They had an active area of 6 cm² (2.8 cm diameter) and a linear range of 10 mW/cm² (better than 60 mW per detector). This was more than adequate for the 10 to 20 mW expected at each detector. The linearity of the detectors used in the interferometer was checked up to 6 mA of detected current (about 24 mW incident power).

The optical transmission through the interferometer beamsplitter mirrors and retroreflectors was measured as 0.7. The reflection loss at each photodetector was 0.6 and the quantum detection efficiency was 0.7, for a combined optical efficiency of 0.3. Thus, for a nominal 35 mW of single-mode laser power, 24 mW made it through the interferometer, 14 mW reached the photodetector, and 10 mW was detected. All of these efficiencies could have been improved somewhat with further effort. The most obvious improvement would have been to have the photodetectors fabricated with a front surface-contact layer that would be compatible with an antireflection coating.

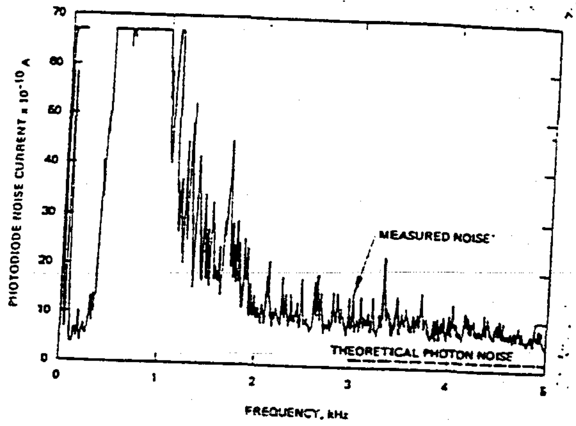


FIG. 8. Laser noise measured in 10-Hz bandwidth.

The photodiodes were operated at a nominal bias voltage of 10 V, which gave a capacitance of 1700 pF and a dark current of 5 μA. The load resistor used with the photodiodes was nominally 5 kΩ which gave a 3 dB frequency rolloff at 20 kHz.

The photodetectors were operated as a balanced pair in a circuit configuration (see Fig. 10) with a special filter that allowed the signals in the 0.5 to 25 kHz band to pass on to the preamplifier while blocking the high-frequency plasma oscillation at 100 kHz (induced for laser-noise reduction) and the low-frequency line harmonics that could have saturated the preamplifier.

The circuit also allowed for the extraction of the

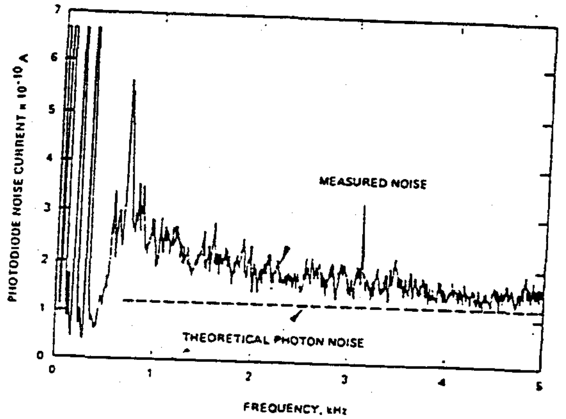


FIG. 9. Laser noise in 10-Hz bandwidth with balanced detectors.

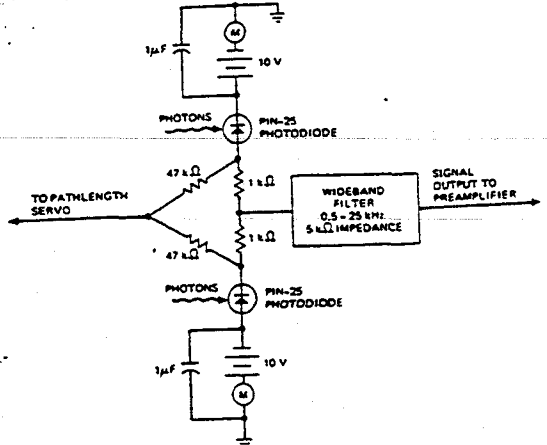


FIG. 10. Photodiode bias and filter circuit.

160-Hz pathlength servo signal that was used in a narrow-band feedback circuit to control a piezoelectric stack in one arm of the interferometer to maintain equal illumination of the two photodetectors.

The bandpass filter used in the circuit was a special design consisting of a 5 kΩ impedance, 5-pole, high-pass Butterworth filter with a low-frequency rolloff of 550 Hz and a slope of 26 dB/octave. This was followed by a 5 kΩ impedance, 5-pole, low-pass, Tschebyscheff filter with a high-frequency rolloff of 25 kHz and a slope of 30 dB/octave (see Fig. 11) The insertion loss of the filter was less than 1 dB from 1 to 20 kHz (see Fig. 12).

III. INTERFEROMETER S/N ANALYSIS

In the Michelson interferometer shown in Fig. 13, the single-mode laser power P entering the interferometer produces a photon flux at the entrance to the beamsplitter of

$$\phi_0 = \frac{P}{h\nu} , \tag{23}$$

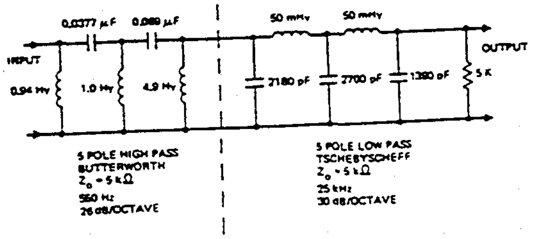


FIG. 11. Low-insertion loss wideband filter.

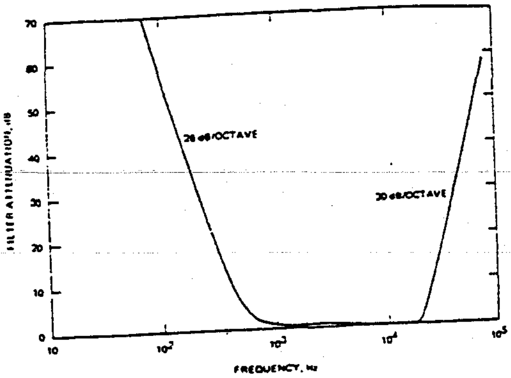


FIG. 12. Filter bandpass attenuation.

where

$h = 6.626 \times 10^{-34}$ J sec ,

$\nu = 4.74 \times 10^{14}$ Hz ,

$h\nu = 3.14 \times 10^{-19}$ J .

(A single-mode laser power of 50 mW is equivalent to a photon flux of 1.6×10^{17} photons/sec.)

The laser beam is divided at the 50-50 beamsplitter. One half travels down arm one and is delayed by the pathlength $\zeta_1 = l_1 + \xi_1(t)$ of arm one, while the other half travels down arm two and is delayed by that pathlength $\zeta_2 = l_2 + \xi_2(t)$. On their return to the beamsplitter, the two beams are again split and one half of each beam is combined with one half of the other at the two photodetectors.

The photon flux reaching the two photodetectors is then given by

$$\phi_1 = \eta \frac{\phi_0}{2} \left[1 - \cos \frac{4\pi}{\lambda} (\zeta_1 - \zeta_2) \right] , \tag{24}$$

$$\phi_2 = \eta \frac{\phi_0}{2} \left[1 + \cos \frac{4\pi}{\lambda} (\zeta_1 - \zeta_2) \right] , \tag{25}$$

where η is the transmission efficiency of the optics

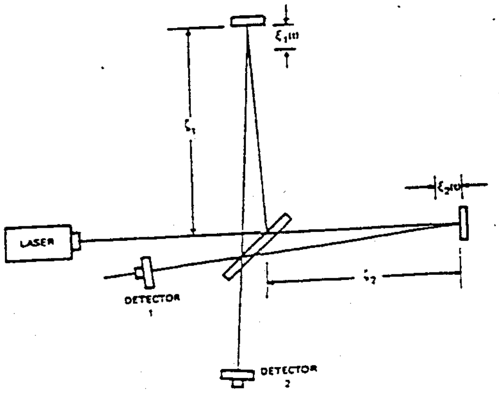


FIG. 13. Michelson interferometer schematic.

in the interferometer, and the difference in the pathlengths of the two arms of the interferometer is

$$\zeta_1 - \zeta_2 = l_1 - l_2 + \xi_1 - \xi_2$$

or

$$\Delta\zeta = \Delta l + \Delta\xi \ . \tag{26}$$

The detectors were operated in a differential signal mode. For maximum sensitivity the nominal pathlength l_1 of one arm was adjusted by the low-frequency pathlength servo so that there was nominally half the total power on each detector. This occurs when

$$\frac{4\pi\Delta l}{\lambda} = \frac{\pi}{2} \pm \pi \tag{27}$$

or

$$\Delta l = \frac{\lambda}{8} \pm \frac{\lambda}{4} \ . \tag{28}$$

At this interferometer setting,

$$\cos\frac{4\pi}{\lambda}(\Delta l + \Delta\xi) = -\sin\frac{4\pi}{\lambda}\Delta\xi \approx -\frac{4\pi}{\lambda}\Delta\xi \tag{29}$$

[where we have assumed $\Delta\xi(t)$ is small compared to the wavelength].

The flux incident at the two detectors is then

$$\phi_1 = \eta\frac{\phi_0}{2}\left(1 + \frac{4\pi}{\lambda}\Delta\xi\right) , \tag{30}$$

$$\phi_2 = \eta\frac{\phi_0}{2}\left(1 - \frac{4\pi}{\lambda}\Delta\xi\right) . \tag{31}$$

As the pathlength difference $\Delta\xi(t)$ due to the gravitational-radiation strains changes with time, the flux at one detector will increase, while the flux in the other will decrease.

In a measurement time interval τ, the number of photoelectrons produced in each detector is

$$N_1 = \frac{\eta\phi_0\tau}{2}\left(1 + \frac{4\pi}{\lambda}\Delta\xi\right) , \tag{32}$$

$$N_2 = \frac{\eta\phi_0\tau}{2}\left(1 - \frac{4\pi}{\lambda}\Delta\xi\right) , \tag{33}$$

where the quantum efficiency of the photodetector is now included in the efficiency coefficient η.

The number of photoelectrons in each photodetector has an average value[7] of

$$\langle N_1\rangle = \langle N_2\rangle = = \tfrac{1}{2}(\eta\tau\langle\phi_0\rangle) , \tag{34}$$

and a variance of

$$\mathrm{var}N_2 = \mathrm{var}N_1 = \langle(N_1 - \langle N_1\rangle)^2\rangle = \langle N_1\rangle \tag{35}$$

that is proportional to the average number of detected photons.

Each measurement interval has a measured number of photoelectrons that is different (usually) from the average number; this difference,

$$\Delta N_1 = N_1 - \langle N_1\rangle , \tag{36}$$

is greater or less depending upon the length of the measurement time (number of measured photoelectrons). The time average of this difference is

$$\Delta n_1 = \frac{\Delta N_1}{\tau} . \tag{37}$$

The spectral intensity in photoelectrons²/sec² Hz of the noise is then given by[7]

$$\begin{aligned} S(f) &= \lim_{\tau\to\infty} 2\tau\langle\Delta n_1^2\rangle \\ &= \lim_{\tau\to\infty} 2\frac{\langle\Delta N_1^2\rangle}{\tau} \\ &= \lim_{\tau\to\infty} \frac{2\langle N_1\rangle}{\tau} \\ &= \eta\langle\phi_0\rangle \\ &= \frac{\eta\langle P\rangle}{h\nu} . \end{aligned} \tag{38}$$

This spectral intensity of photoelectron noise S produces a squared noise current in a bandwidth B given by

$$\begin{aligned} I_{n_1}^2 = I_{n_2}^2 &= e^2 SB \\ &= e^2\eta\langle\phi_0\rangle B = \frac{\eta\langle P\rangle e^2 B}{h\nu} \\ &= 2e\langle I\rangle B , \end{aligned} \tag{39}$$

where $\langle I\rangle = \frac{1}{2}(\eta\langle\phi_0\rangle)e$ is the average photocurrent in each photodetector. The noise current is the same in both photodetectors since their average detected flux levels are kept equal.

The signal currents and the noise currents from the photodetectors are converted into voltages by means of the load resistor of nominal resistance $R \approx 5$ kΩ that terminates the wideband filter (see Fig. 11). The load resistor also contributes the following Johnson noise voltage:

$$V_R^2 = 4kTBR ,$$

where

$$\begin{aligned} k &= 1.38\times 10^{-23}\ \mathrm{J/K} , \\ T &= 290\ \mathrm{K} . \end{aligned} \tag{40}$$

The Johnson noise of a 5-kΩ resistor in a 10-Hz bandwidth is

$$\begin{aligned} V_R &= (4kTBR)^{1/2} \\ &= 27\ \mathrm{nV} . \end{aligned} \tag{41}$$

During the operation of the interferometer, the load resistor was carrying a photocurrent that was never less than 2 mA. This level of photocurrent would generate in a bandwidth of 10 Hz, a noise voltage due to shot noise of the photoelectrons of

$$V_P = (2eIB)^{1/2}R \\ = 370\ \text{nV} . \qquad (42)$$

Thus, in all cases, the photocurrent shot noise was very much larger than the Johnson noise in the load resistor, so we can ignore the noise contribution of the Johnson noise in our analyses.

The low-noise differential amplifier used in the experiment was a PAR CR-4, with selected front-end FETs. The amplifier noise figure for the 5-kΩ load was 0.2 dB. This had a negligible effect on the overall system performance.

The time-varying signal voltages are coherent, while the photoelectron noise currents are incoherent. Thus the differential amplifier measures a squared signal voltage as follows:

$$\begin{aligned} V_S{}^2 &= (V_1 - V_2)^2 \\ &= (I_1 - I_2)^2 R^2 \\ &= (\phi_1 - \phi_2)^2 e^2 R^2 = \left(\frac{\eta \phi_0 4\pi \Delta\xi}{\lambda} \right)^2 e^2 R^2 , \qquad (43) \end{aligned}$$

while in a bandwidth B it measures a squared shot-noise voltage given by

$$\begin{aligned} V_n{}^2 &= V_{n_1}{}^2 + V_{n_2}{}^2 = (I_{n_1}{}^2 + I_{n_2}{}^2)R^2 \\ &= 2e^2 \eta \phi_0 B R^2 . \qquad (44) \end{aligned}$$

Thus the signal-to-noise (S/N) power of the measured differential signal is

$$\frac{S}{N} = \frac{2\eta\phi_0}{B} \left(\frac{\Delta\xi}{\bar{\lambda}} \right)^2 . \qquad (45)$$

For unity S/N, the detection sensitivity is then

$$\frac{\Delta\xi^2}{B} = \frac{\bar{\lambda}^2}{2\eta\phi_0} = \frac{\bar{\lambda}^2}{2\eta P/h\nu} = \frac{\bar{\lambda}^2}{2I/e} , \qquad (46)$$

where ϕ_0 and P are the initial flux and power of the laser, η is the light incident to photoelectron-conversion efficiency, and I is the combined photocurrent of the two detectors.

For $\eta P = 10$ mW of detected laser power, the theoretical photon-noise limited displacement sensitivity is

$$\frac{\Delta\xi}{B^{1/2}} = \frac{\bar{\lambda}}{(2\eta P/h\nu)^{1/2}} = 0.4\ \text{fm/Hz}^{1/2} . \qquad (47)$$

If we assume an interferometer pathlength of 8.5 m, this converts to a strain sensitivity of 0.5 $\times 10^{-16}$ m/m Hz$^{1/2}$.

IV. DISPLACEMENT TRANSDUCER CALIBRATION

To calibrate the interferometer for displacement sensitivity we used a piezoelectric disk attached to the back of one of the interferometer mirrors (see Fig. 7). The mirror was first driven by a high dc voltage on the piezoelectric until the illumination on the photodetectors went from minimum to maximum. The interferometer pathlength thus had changed by $\lambda/2$ and the mirror had moved $\lambda/4$. The displacement sensitivity obtained was

$$\sigma = 1.6 \times 10^{-10}\ \text{m/V} . \qquad (48)$$

The drive on the piezoelectric crystal was then reduced from a very high level down to a drive level where the signal was hidden by the laser noise. During this decrease in the drive level the interferometer output remained a linear function of the drive level (to within the noise level of the measurements).

The piezoelectric displacement calibration technique was also cross-checked at the low drive levels by the use of a potassium dihydrogen phosphate (KDP) optical modulator. The KDP modulator was calibrated for a half-wavelength path difference and found to have a displacement sensitivity of

$$\sigma_k = 2.0 \times 10^{-11}\ \text{m/V} . \qquad (49)$$

The voltage on the KDP pathlength modulator was then reduced to where it gave the same interferometer response,as the piezoelectric displacement transducer. The calculated pathlength difference from the KDP modulator was 1.0 pm, within 10% of the displacement calculated for the piezoelectric modulator for the same interferometer output.

V. INTERFEROMETER SENSITIVITY

The sensitivity of the interferometer as a function of frequency was determined by a frequency scan of the interferometer output using a spectrum analyzer with a 10-Hz bandwidth filter (see Fig. 14). During the scan a calibration signal of 100 μV rms was placed on the calibration piezoelectric displacement transducer to insert a 16-fm amplitude signal. The amplitude of the calibration signal was about 5.7 times the noise level in the region between 4 and 5 kHz. The amplitude of the noise measured through a 10-Hz bandwidth filter in that region is therefore about 2.8 fm or 0.9 fm for a 1-Hz bandwidth.

If we assume that the effective length of the interferometer is 8.5 m then the equivalent strain sensitivity is 0.1 fm/m per root Hz at the higher frequencies, rising slightly to 0.3 fm/m per root Hz at 1 kHz. For comparison, the kT strain noise

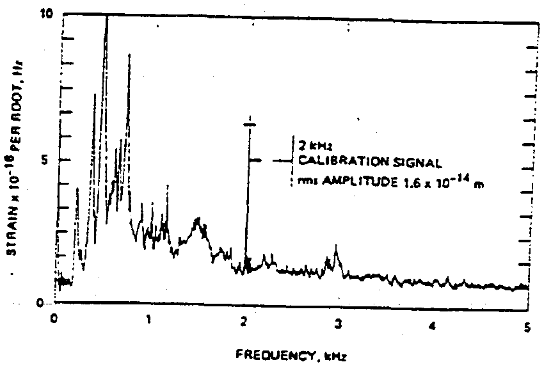


FIG. 14. Strain sensitivity of interferometer antenna.

in a room-temperature, 2-m long, 1000-kg elastic solid bar antenna is 0.14 fm/m.

VI. OPERATION OF THE ANTENNA

Our ultimate plan for the antenna system was to move the interferometer to a remote site and replace the 2-m evacuated pipes with much longer sections of evacuated irrigation pipe. Similar laser interferometer systems up to a kilometer in arm length had already been demonstrated by many others for geophysical studies.[8] Since the sensitivity of the interferometer was fixed at a certain level of displacement sensitivity by the photon noise, an increase in interferometer arm length should give a proportional increase in strain sensitivity for the same laser-power level. The funding for this next move proved to be unavailable so we concluded the program by operating the system as it was, despite the high level of acoustic, electromagnetic and vibrational noise from the other activities in the building.

Since the output of the laser interferometer was a wideband analog signal in the audio region, the signal was recorded directly onto magnetic tape through one channel of a high quality stereo tape recorder. The other channel of the recorder was used as the monitor channel. To monitor the system and environmental noises we combined the outputs of a photodetector to detect the audio-frequency noises in a sample of the laser beam, a microphone to detect acoustic noises in the room, a wideband seismometer to detect floor motion, and an amplifier to detect audio voltages on the power lines. The combined signal was then placed on the monitor channel.

To provide a constant time and amplitude calibration, an accurate 2-kHz signal of about 10-fm amplitude was maintained on the calibration transducer. At 15-min intervals, the time from WWV would be acoustically introduced into the system and recorded on both channels.

Later analysis of the magnetic tapes revealed that about once every two minutes there would be a short coherent signal audible to the ear over the white-noise hiss of the photon noise. Most of these were chirps from a spurious laser mode mixing with the main laser mode, clicks as the laser switched modes, and tones from thermal contractions exciting mechanical vibrations in structures. Nearly all of these were also found in the monitor channel. About once every ten minutes there would be an audible chirp or tone that was not on the monitor channel. Some of these were digitally analyzed and compared with the calibration signal. A typical audible signal was about 1 to 5 times the power of the calibration signal or about 20 to 100 times the photon-noise limit in a 1-Hz bandwidth.

VII. CALIBRATION OF EAR

When the interferometer was working well, we were able to hear single-frequency 3- to 10-kHz tones of 10-fm rms amplitude introduced into the interferometer by the piezoelectric displacement transducer.

Since the noise level of the interferometer in that band is about 0.9 fm/Hz$^{1/2}$, this means that the audio system, including our ear-brain combination, had an effective detection bandwidth of about 120 Hz.

Although the detection capability for single tones, chirps, and impulsive events will be different, we feel that if we were listening carefully for a signal that exceeded an rms amplitude of about 10 fm or a strain amplitude of 1 fm/m, that we would have detected it by ear. We could then pin down the exact position on the tape and carry out a detailed digital analysis of that section of tape that would produce data with time resolution of 40 μsec, amplitude resolution of 0.9 fm/Hz,$^{1/2}$ and strain resolution of 0.1 fm/m Hz$^{1/2}$.

However, because of the uncertainties in the signatures of gravitational-radiation signals and the capability of our ear-brain combination to recognize an unknown signature as something "unusual" in a background of white noise, we will be conservative and say only that a gravitational wave with a total strain level of 10 fm/m over the audio band (1 to 20 kHz), would have been easily detected by ear on our tapes.

To give an example of the detailed structure that can be extracted from a short signal with a sophisticated digital signal processor replacing the analog ear-brain signal processing system, we analyzed one short (~10 msec) tone that occurred at 09 h 46 min 21 sec GMT Sunday 3 December

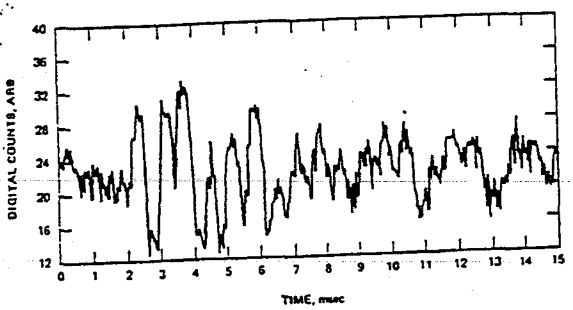


FIG. 15. Time history of typical signal.

1972. The detailed signature can be seen in the digitized time history of the signal (see Fig. 15), and the frequency spectrum can be seen in the power spectral density plot of Fig. 16. (The peak at 3000 Hz is the calibration signal of about 10 fm.) Although this signal occurred within 3 sec. of a coincidence between the two 1660 Hz antennas in the Maryland system, it is not a good candidate. The energy at 1660 Hz is negligible and the time delay exceeds the estimated errors in the measurement of absolute time (0.6 sec) in the two systems.

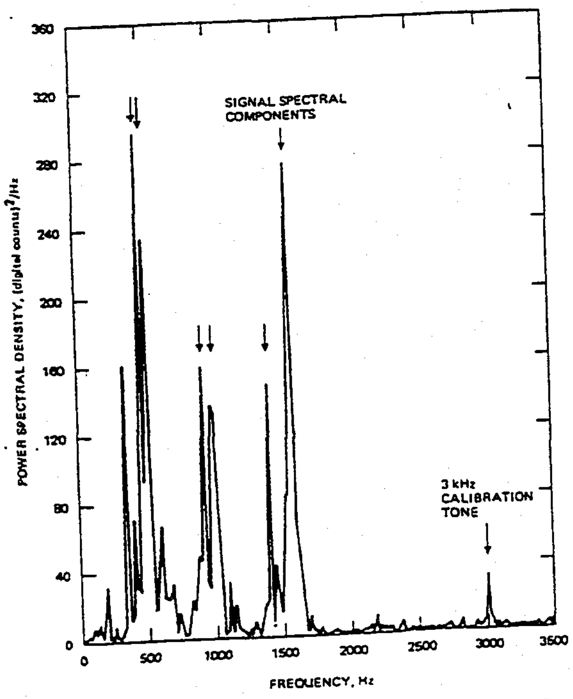


FIG. 16. Power spectral density plot of typical signal.

VIII. COMPARISON OF DATA WITH OTHER OBSERVERS

During October 1972 the Frascati group were operating a single elastic-solid type gravitational-radiation antenna.[9] During the hours when the Malibu antenna was in operation between the dates from 15 to 25 October 1972, the Frascati group recorded 18 instances when the antenna output produced a significant level of response. None of these coincided with a signal in the Malibu antenna.

During the fall of 1972, the Glasgow group were operating a pair of wideband (BW ≈ 800 Hz) elastic-solid type antennas. During the hours when the Malibu antenna was in operation, their system responded to 22 events where the signal from one or the other of their two antennas exceeded a previously set threshold. None of these coincide with a signal in the Malibu antenna. The one "distinctive signal" reported by the Glasgow group[10] occurred at 13 h 07 min 29 sec GMT 5 September 1972, which was prior to the start of the Malibu data collection period.

Gravitational-radiation antennas of the elastic-solid type have been under development at the University of Maryland since 1959.[11] Statistically significant numbers of coincidences between antennas at the University of Maryland and the Argonne National Laboratory have been reported since 1969.[12]

During the fall of 1972, the Maryland group was recording coincidences between a 66-cm diameter, 1.5-m long resonant aluminum cylinder at College Park, Maryland and a similar cylinder at Argonne, as well as coincidences between a resonant aluminum disk at College Park and the cylinder at Argonne. The Argonne and College Park cylinders had a nominal resonant frequency of 1660 Hz. The disk was originally designed to search for 1660-Hz scalar radiation and was constructed to have a radially symmetric mode at 1660 Hz that would be excited by scalar radiation. It was also instrumented to detect a mode at 1100 Hz that would only be excited by tensor gravitational radiation.

During the time that the Malibu antenna was operational, the Maryland group recorded 28 coincidences between either the bar at Argonne and the bar at Maryland, or the bar at Argonne and the disk at Maryland. Because of triple coincidences and close-spaced coincidences, the 28 Maryland coincidences fell into 20 two-minute time blocks. Of the 20 time blocks, 7 blocks (containing 17 coincidences) had audible signals in the Malibu interferometer that were within 10 sec of the Maryland coincidences and were not audible in the monitoring channel.

Both raw power and derivative power-squared digitized data plots digitized to 0.1-sec accuracy were obtained from the Maryland group and compared with the 0.2-sec accuracy Malibu data. None of the audible Malibu signals fell within 0.6 sec of a Maryland-Argonne coincidence.

It is difficult to compare the relative detection capabilities of the various antennas since their amplitude sensitivities, bandwidths, and signal processing techniques differ widely.

In general we can say that the bar-antenna detection systems responded to gravitational-radiation strains with spectral components near the resonant frequency of the bar that had an amplitude of the order of 0.1 fm/m, while the interferometer antenna responded to gravitational-radiation strains with spectral components in the band from 1–20 kHz. The sensitivity of the interferometer is highly dependent upon the signature of the signal and the processing technique and varies from 0.1 fm/m for known narrow-band signals to 10 fm/m for noiselike signals.

However, the lack of a significant correlation between the interferometer output and the bar events and coincidences can be used to put an upper limit on the gravitational-radiation strain amplitude during the bar coincidences and events. Thus, at the time one of the bar-antenna systems produced an event or coincidence corresponding to a gravitational-radiation signal with an amplitude of 0.1 fm/m due to spectral components in a narrow band around the bar resonance, the amplitude of the gravitational-radiation spectral components in the entire band from 1–20 kHz was definitely less than 10 fm/m and was probably less than 1 fm/m.

ACKNOWLEDGMENTS

The work on the design, construction, operation, and data analysis of the wideband laser-interferometer antenna was a multiyear effort involving many people.

The initial design and studies were carried out by Gaylord Moss, Larry Miller, and Jay Melosh, with advice from Philip Chapman and Ranier Weiss. The construction of the steadily more complex and capable designs were predominantly the brilliant work of Gaylord Moss with some inspired early feasibility demonstrations by Larry Miller. Mr. Moss and Mr. Miller were ably assisted by Larry Matheney, Don Boswell, Ken Craig, Cesar DeAnda, Ismael Charnabroda, Craig Spencer, and Dale Sipma at various times during the program. The operation of the laser interferometer as a detector for gravitational radiation required a dedicated effort which involved sitting nearly motionless for hours at a time, monitoring the interferometer and data collection system performance. I was ably assisted on alternate shifts by Gaylord Moss, who often collected data all night and spent the next day fixing a balky laser or piece of electronics.

The data analysis was another long and arduous effort which was made much easier by the able assistance of Ray Lahr in the double-blind audio analysis and the Hughes digitalization and display facility operated by Brad Tuttle and Al Rieser.

I would also like to acknowledge the encouragement, cooperation, and exchange of data of the other groups who were operating gravitational-radiation detectors at the same time—Professor R. W. P. Drever and Roger Bland of the University of Glasgow, Dr. K. Maischberger of the European Space Research Organization, Frascati, Italy and especially Professor J. Weber and Bruce Webster of the University of Maryland.

I finally would like to thank Dr. George Smith and the Hughes Aircraft Company for the moral and financial support during the three years it took to get the system operational.

[1] G. E. Moss, L. R. Miller, and R. L. Forward, Appl. Opt. 10, 2495 (1971).

[2] R. L. Forward and G. E. Moss, American Physical Society Winter Meeting, Los Angeles, CA, 1972 (unpublished).

[3] R. L. Forward, Gen. Relativ. Gravit. 2, 149 (1971).

[4] L. Landau and E. Lifshitz, *The Classical Theory of Fields* (Addison-Wesley, Reading, Mass., 1951), p. 326.

[5] H. Goldstein, *Classical Mechanics* (Addison-Wesley, Cambridge, Mass., 1956), pp. 107–109.

[6] G. E. Moss, Appl. Opt. 10, 2565 (1971).

[7] A. van der Ziel, *Noise* (Prentice Hall, Englewood Cliffs, New Jersey, 1971), pp. 14–16.

[8] V. Vali and R. C. Bostrom, Earth Planet. Sci. Lett. 4, 436 (1968).

[9] D. Bramanti and K. Maischberger, Lett. Nuovo Cimento 4, 1007 (1972).

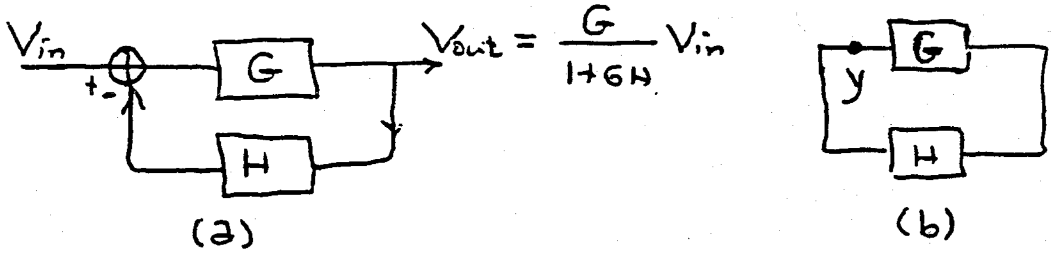
[10] R. W. P. Drever *et al.*, Nature 246, 340 (1973).

[11] J. Weber, Phys. Rev. 117, 306 (1960).

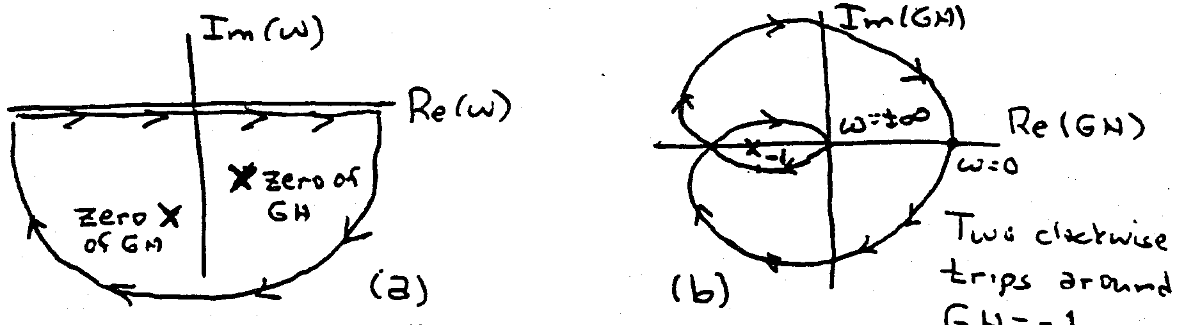
[12] J. Weber, Phys. Rev. Lett. 22, 1320 (1969).

181635

c. The Nyquist criterion for the stability of a control loop [e.g., read *Dorf*, pp. 309–333]. [The Nyquist criterion, in a nutshell, is this: Consider a simple feedback loop of form shown in (a) below. If the input and output ports are shut, the resulting closed loop shown in (b) can oscillate at certain complex eigenfrequencies without any stimulus. Those frequencies are easily deduced from the requirement that the amplitude y at the indicated point must satisfy $y = G(\omega)H(\omega)y$, and therefore $y(1+GH) = 0$, and therefore *the loop's frequencies of self oscillation are the zeroes of* $1 + G(\omega)H(\omega)$.



Since the time dependence of these oscillations is $e^{+j\omega t}$, if there are any zeroes of $1 + GH$ in the *lower-half* complex frequency plane (any eigenfrequencies ω with negative imaginary parts), then the amplitude of the closed loop's oscillations will grow in time; i.e., the closed loop will be unstable. The number of zeroes in the lower-half frequency plane can be inferred from the Cauchy theorem of complex variable theory: Construct the curve $G(\omega)H(\omega)$ in the complex plane, with ω running along the real axis from $-\infty$ to $+\infty$, and then swinging down around the lower half frequency plane and back to $-\infty$; see drawing (a) below. The number of times that this curve, $G(\omega)H(\omega)$ encircles clockwise the point $GH = -1$ (on the real axis) is the number of zeroes of $1+GH$ minus the number of poles of $1 + GH$; see drawing (b) below. For feedback loops there usually are no poles of $1 + GH$ [such a pole would give precisely zero output/input in the feedback loop of (a) above], so usually the number of clockwise trips around $GH = -1$ is the number of zeroes in the complex frequency plane. Thus, if there are no clockwise trips, the closed loop is stable; if there are some, it is unstable. This is the Nyquist criterion for stability.]

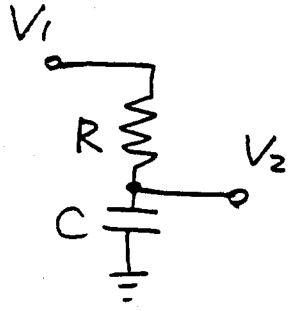


Suggested Supplementary Reading:

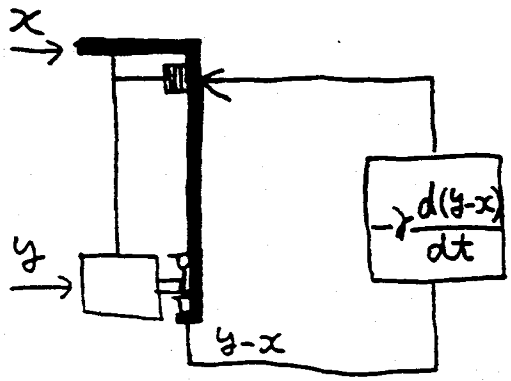
5. Read whichever of items 3. and 4. you did not do as "assigned reading".

A Few Suggested Problems

1. Use the Nyquist criterion for the stability of a feedback loop to show that, when the Bode diagram has the qualitative form shown on transparency 23 of Kawamura's lecture (where $f = \omega/2\pi$), then the loop is stable if the phase of GH at the unity gain point is $\phi > -180^o$, and unstable if $\phi < -180^o$. [Hint: show that, because in the time domain the equations describing most any servo loop are real, when ω is real then $G(-\omega)H(-\omega)$ is the complex conjugate of $G(+\omega)H(+\omega)$. This permits you to construct the Nyquist curve in the frequency-response plot for both positive and negative ω from Kawamura's positive-frequency Bode diagram.] For what shapes of Bode diagrams will this $\phi > -180^o$ stability citericn remain true?(Consider, for example, the issue of how many unity gain points there are).
2. Construct a complex frequency-response curve and also a Bode diagram for the following pass $R-C$ circuit. From the Bode diagram infer that this circuit is a low-pass filter.



3. In his lecture [transparencies numbered 15–17], Kawamura described the damping of the swing of a pendulum via a feedblack loop that produces a displacement $\delta x = -\gamma dy/dt$ of the pendulum's support point, where γ is the damping constant and y is the horizontal position of the pendulum's mass. Of course, in order to implement this, one needs some fixed object with respect to which y is measured. In transparency 15 that object is the shadow sensor, but nothing is said about what that sensor is attached to. A practical approach is to attach the sensor to the pendulum's support point, as shown below. Then the feedback displacement is $\delta x = -\gamma d(y-x)/dt$, where x is the instantaneous horizontal position of the support point. Repeat Kawamura's analysis [transparencies 15–17]] for this feedback system.



4. Suppose that one were to try to damp the (low-frequency, 1 Hz) swing of the pendulum in problem 3 not with a feedback displacement $\delta x = -\gamma d(y-x)/dt$, but instead with a feedback displacement that is $-ay$ (for some constant $a > 1$) at low frequencies (near 1 Hz) but that shuts off at higher frequencies (above 10 Hz), where the gravity waves are to be measured. Suppose one implements this feedback displacement by simply passing a voltage, proportional to y, through a low-pass $R-C$ filter of the sort discussed in problem 2. Show that the resulting damping system will be unstable.
5. Derive the relation $\delta l = d_1 \delta\theta_1 + d_2 \delta\theta_2$ on transparency 28 of Kawamura's lecture.

LECTURE 11.

Optical Topology for Locking and Control of an Interferometer, and Signal Extraction

Lecture by Martin Regehr

Assigned Reading:

EE. P. W. Milonni and J. H. Eberly, *Lasers* (Wiley, New York, 1988): sections 12.9 "AM Locking" and 12.10 "FM Locking," pp. 385–390. [Here you are asked to focus on the description of AM modulation and FM modulation as putting side bands onto a carrier frequency. Of particular interest is the fact that a sinusoidal FM modulation produces a whole series of side bands, whose strengths are described by Bessel functions. When the modulation amplitude is small compared to a radian, only the first side bands dominate.]

FF. C. N. Man, D. Shoemaker, M. Pham Tu and D. Dewey, "External modulation technique for sensitive interferometric detection of displacements," *Physics Letters A*, **148**, 8–16. [This paper describes in detail a technique used in LIGO to circumvent laser noise that is seriously in excess of standard photon shot noise in the gravitational wave's kHz band. The trick is to upconvert the gravitational-wave signal to ~ 10 MHz frequency, where the laser's noise is near the shot-noise level. This is achieved by modulating the laser light at ~ 10 MHz (i.e. put 10 MHz side bands on the light's $\sim 10^{15}$ Hz carrier frequency), and then arranging that the gravitational-wave signal becomes a ~ 1 kHz side band of the 10 MHz side band.]

Suggested Supplementary Reading:

Read in one or more electronics or laser textbooks about places elsewhere in experimental physics and engineering where noise is circumvented by upconverting a signal to higher frequency via modulation, and then recovering the signal by synchronous demodulation. For example, read about "lock-in amplifiers," which do this. Two references dealing with this were passed out in class:

GG. John H. Moore, Christopher C. Davis, and Michael A. Coplan, *Building Scientific Apparatus: A Practical Guide to Design and Construction* (Addison-Wesley, 1983), Sec. 6.8.3 "The lock-in amplifier and gated integrator or boxcar," (pp. 435–437).

HH. Paul Horowitz and Winfield Hill, *The Art of Electronics* (Cambridge University Press, Cambridge, 1980), Sec. 14.15 "Lock-in detection" (pp. 628–631) and an earlier section to which it refers, Sec. 9.29 "PLL components, Phase detector" (pp. 429–430).

A Few Suggested Problems: See the next page.

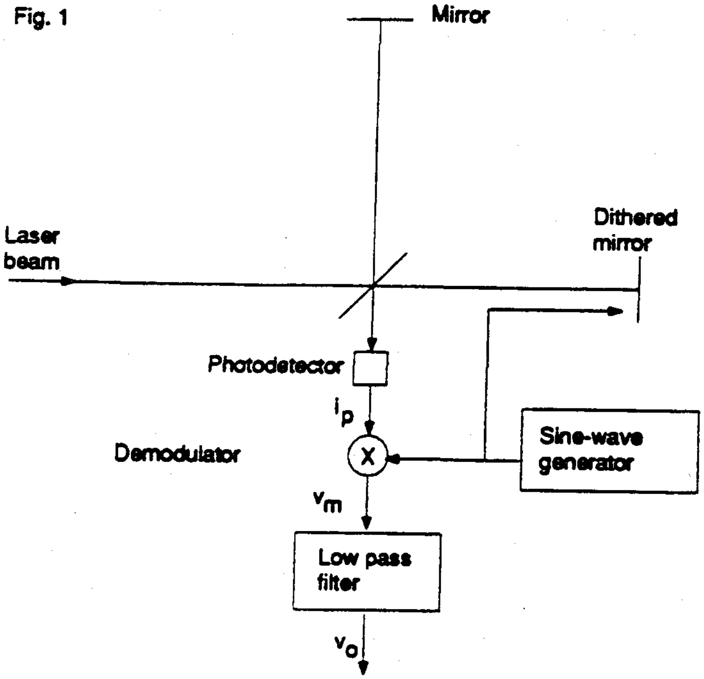
Problems for Ph 103 5/4/94

1. In this problem we will calculate the shot noise limited sensitivity of an ideal Michelson interferometer which is modulated around the dark fringe by dithering one of the mirrors with $\delta \sin \omega t$. We model the demodulator as a device which multiplies its input by $\sin \omega t$ and the low-pass filter as a device which averages over an interval T:

$$v_o(t) = \frac{1}{T} \int_{t-T}^{t} v_m(t')dt'$$

and for convenience we choose T to be an integral number of modulation periods $T = \frac{2\pi n}{\omega}$. Assume that the interferometer is small enough that we can neglect the light travel time from the dithered mirror to the photodetector, and that δ is very small $\delta \ll \lambda$.

Fig. 1



a. Find the derivative of the low-pass filter output with respect to displacement of the mirror which is not being dithered. It should be a function of the amplitude δ of the dithering.

b. Find the shot noise in $v_m(t)$ at the demodulator output (assuming that the ouput is at a dark fringe except for the dither):

   i. Use the time averaged photocurrent to calculate the shot noise $S_{i_p}(f)$.

   ii. Assume that the shot noise at the mixer output is related to $S_{i_p}(f)$ by the time average of the square of the mixer gain, i.e.:

$$\begin{aligned} S_{v_m}(f) &= \overline{S_{i_p}(f)(\sin^2 \omega t)} \\ &= \frac{1}{2} S_{i_p}(f) \end{aligned}$$

It should also be a function of the dithering amplitude. Finally $S_{v_o}(f) = S_{v_m}(f)$ since the low-pass filter passes noise in the signal band virtually unattenuated.

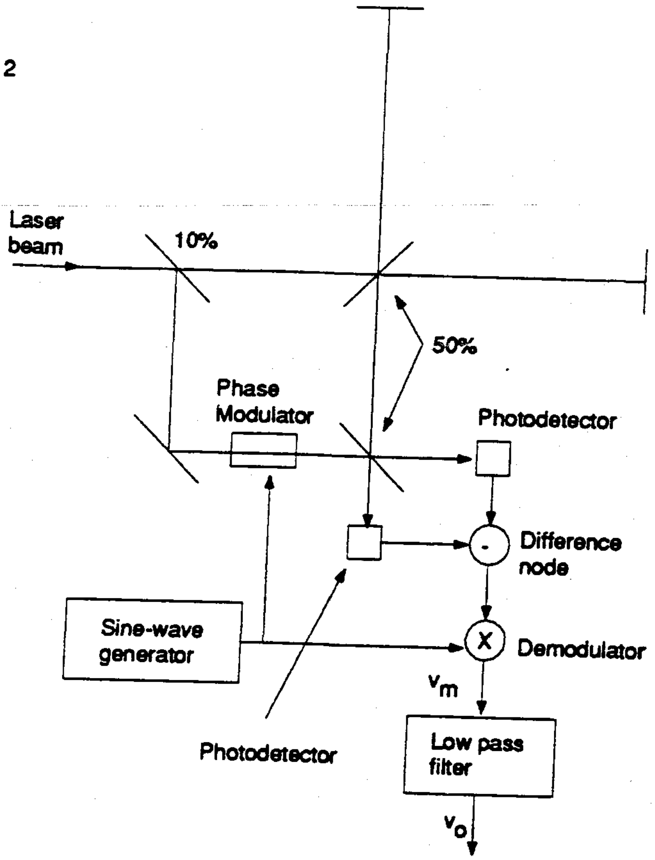
c. Take the ratio of the above two quantities to find the shot noise limited displacement sensitivity

$$S_x^{\frac{1}{2}}(f) \equiv \frac{\sqrt{S_{v_m}(f)}}{\frac{dv_o}{dx}}$$

It should be a function of the optical power and wavelength, and independent of the dithering amplitude.

2. Consider the externally modulated Michelson interferometer shown in Figure 2. Find the derivative of the low-pass filter output with respect to displacement of one of the Michelson end mirrors, assuming a pick-off which diverts 10% of the power from the main beam, a 50/50 beam splitter, a 50/50 beam combiner, a demodulator modeled as in question 1, the input to which is the difference in the photocurrents, and a low-pass filter modeled as above. Write your answer in terms of the optical power and Bessel functions of the modulation index.

Fig. 2


Laser beam
10%
50%
Phase Modulator
Photodetector
Difference node
Sine-wave generator
Demodulator
v_m
Photodetector
Low pass filter
v_o

Applied Optics, in press

Mirror Orientation Noise in a Fabry-Perot Interferometer Gravitational Wave Detector

Seiji Kawamura and Michael E. Zucker

LIGO Project
California Institute of Technology
Pasadena, California 91125 U.S.A.

ABSTRACT

The influence of angular mirror orientation errors on the length of a Fabry-Perot resonator is analyzed geometrically. Under conditions where dominant errors are static or vary slowly over time, the analysis allows a simple prediction of the spectrum of short-term cavity length fluctuations resulting from mirror orientation noise. The resulting model is applicable to the design of mirror control systems for LIGO (the Laser Interferometer Gravitational–wave Observatory), which will monitor separations between mirrored surfaces of suspended inertial test bodies to measure astrophysical gravitational radiation. The analysis was verified by measuring the response of the LIGO 40 meter interferometer testbed to rotation of its mirrors.

Key words: Fabry-Perot, cavity, gravitational wave, interferometer, alignment.

1. Introduction

LIGO (the Laser Interferometer Gravitational–wave Observatory) will employ Fabry-Perot optical cavities to detect and measure small changes in separation, having characteristic durations of a few milliseconds and magnitudes of order 10^{-18} meter, between inertial gravitational test bodies separated by orthogonal 4 kilometer baselines[1, 2]. The four test bodies, made of fused quartz, will be polished and coated to form two resonant Fabry-Perot optical cavities. Laser light will be split by a beamsplitter and made to resonate in each of the two orthogonal cavities; the resonant fields are extracted and interfered to measure the difference between their phases, which depend sensitively on the mirror separations. To isolate the test bodies from external forces, they will be suspended as pendula whose natural periods, of order one second, are significantly longer than characteristic signal timescales. Band-limited control systems will damp their rotational rigid-body normal modes, which also have periods of order one second, to maintain optical alignment of the resonators. The required displacement sensitivity places a direct limit on the allowable linear momentum imparted to test bodies by seismic noise, thermal fluctuations, and side effects of the control systems.

Excess noise from angular fluctuations of each mirror, arising from external torques, also influences optical cavity length. A simple geometric model predicting the apparent cavity length change due to test body rotation was developed and evaluated on the LIGO 40 m interferometer testbed. The model accurately predicts the interferometer's response to experimental probe torques applied to its test bodies. Applying the model to predict the effect of residual torques from the interferometer's mirror angle control systems indicated such noise limited the interferometer's sensitivity at frequencies below 700 Hz. New lower-noise control systems were developed and implemented, resulting in a substantial improvement in displacement sensitivity.

2. Optical Cavity Length

The round trip optical phase for the resonator's TEM_{00} mode, which is measured and interpreted as apparent mirror displacement, depends on the *optical length* l, the length of the line segment which lies perpendicular to both mirror surfaces (i.e. the optic axis). Our objective is to quantify the relationship between the measured quantity l and the desired *inertial length* L, the separation between the test bodies' centers of mass, as each body rotates in response to external torques. We will employ the orthogonality of small rotations about the two relevant mutually perpendicular axes ("altitude" and "azimuth") to treat their influences on l independently; thus, in what follows, the normals to both mirrors and the line joining the test bodies can be assumed to lie in a common plane. We will also assume that no net forces are applied (i.e., L is constant).

The optical length of a Fabry–Perot cavity depends on the *deviation angles* θ_1 and θ_2, defined as shown in Figure 1. These angles are both equal to zero when the optical axis coincides with the line joining the centers of mass of the test bodies. To second order in θ_1 and θ_2, which are presumed to be small, the optical length of the cavity is given by[3]

$$l = l_0 + \alpha\theta_1^2 + \beta\theta_2^2 + \gamma\theta_1\theta_2 \qquad (1)$$

where, for the half-symmetric cavity geometry illustrated[4],

$$\alpha = \frac{1}{2}(R + a_2 - L) \quad \text{and}$$
$$\beta = -\frac{\gamma}{2} = \frac{1}{2}(R + a_2) \ .$$

Here R is the radius of curvature of the concave mirror M_2 and $l_0 \equiv L - a_1 - a_2$ is the optical length when $\theta_1 = \theta_2 = 0$. The time evolution $l(t)$ can be computed from (1) for arbitrary $\theta_1(t)$ and $\theta_2(t)$.

For typical cases of experimental interest, θ_1 and θ_2 will be random processes, whose properties are summarized by measured power spectral density functions. We are interested in estimating the power spectrum of the resulting random process l for comparison with expected signals and other noise. Since the power spectrum is the expectation value of the squared modulus of a random process' Fourier transform[5], we begin by transforming (1) into the frequency domain. By the convolution theorem, the transform of (1) can be written

$$\tilde{l}(f) = l_0\delta(f) + \alpha\,\tilde{\theta}_1\otimes\tilde{\theta}_1\,(f) + \beta\,\tilde{\theta}_2\otimes\tilde{\theta}_2\,(f) + \gamma\,\tilde{\theta}_1\otimes\tilde{\theta}_2\,(f) \qquad (3)$$

where $\delta(f)$ is the Dirac delta function and the operator $\otimes$ denotes convolution;

$$\tilde{a}\otimes\tilde{b}\ (f) \equiv \int_{-\infty}^{\infty} a(f')\, b(f-f')\, df' \ . \tag{4}$$

We will subsequently replace each angle fluctuation's Fourier transform with the square root of the spectral density of the corresponding random process, and take the result to be the square root of the spectral density of l. For this substitution to be justified the phases of the Fourier components must truly be random, that is, components at different frequencies must not on average be correlated. True random noise processes like those considered here will satisfy this criterion [6].

While (3) applies to any Fabry–Perot cavity, the application to gravitational wave detection permits an intuitively appealing simplification. Terrestrial gravitational wave detectors will monitor $l(t)$ only at frequencies above a few tens of Hertz, while the alignment errors θ_1 and θ_2 are dominated by static ("D.C.") or slowly varying components (at frequencies below 10 Hz). Faster angle fluctuations, at frequencies within the observation band, will be considerably smaller. This spectral character arises from long–term thermal drift and the increase of seismic vibration with decreasing frequency, combined with the low–pass filtering action of the suspension. In such a situation we may write

$$\theta_\nu(t) = \overline{\theta_\nu} + \epsilon_\nu(t) \tag{5}$$

where $|\epsilon_\nu(t)| \ll |\overline{\theta}_\nu|$,

$$\overline{\theta_\nu} \equiv \frac{1}{T} \int_{-T/2}^{T/2} \theta_\nu(t)\, dt$$

is the average taken over some long interval T, and $\nu = 1$ or 2 for mirror M_1 or M_2, respectively. To leading order in $\epsilon_\nu/\overline{\theta}_\nu$ we find

$$l(t) \simeq l_0 + \frac{1}{2}\overline{d_1}\ \left[\overline{\theta_1} + 2\epsilon_1(t)\right] + \frac{1}{2}\overline{d_2}\ \left[\overline{\theta_2} + 2\epsilon_2(t)\right]\ , \tag{7}$$

where the quantities

$$\begin{aligned} d_1 &\equiv 2\alpha\theta_1 - 2\beta\theta_2 \text{ and} \\ d_2 &\equiv 2\beta(\theta_2 - \theta_1) \end{aligned} \tag{8}$$

are the moment arms separating the perturbed optic axis (i.e. the position of the resonating beam on each mirror) from each test body's center of rotation (Figure 1). The Fourier transform of (7) is then simply

$$\tilde{l}(f) \simeq \overline{d_1}\ \tilde{\epsilon}_1(f) + \overline{d_2}\ \tilde{\epsilon}_2(f) \quad (f \neq 0)\ , \tag{9}$$

agreeing with the intuitive notion that the apparent displacement of the portion of the mirror at the beam location is given by the angle fluctuation multiplied by the moment arm.

Equation 9 can be generalized if we let $\overline{\theta}_\nu$ be a large (in the mean-square sense) slowly varying, rather than static, misalignment. If the dominant misalignment is confined to a small region of the spectrum $|f| < |w|$, we can write

$$\tilde{\theta}_\nu(f) = \tilde{\theta}^w_\nu(f < w) + \tilde{\epsilon}_\nu(f > w) \tag{10}$$

where

$$\int_{-w}^{w} \left|\tilde{\theta}_\nu^w(f)\right|^2 df \ \gg \int_{-\infty}^{\infty} |\tilde{\epsilon}_\nu(f)|^2 \, df \tag{11}$$

(i.e. the low-frequency portion carries substantially more spectral power than the high-frequency portion). We then find the analogous expression to (9), again to leading order, is

$$\tilde{l}(f) \simeq \int_{-w}^{w} \tilde{d}_1(f') \ \tilde{\epsilon}_1(f-f') \, df' + \int_{-w}^{w} \tilde{d}_2(f') \ \tilde{\epsilon}_2(f-f') \, df' \quad (f > 2w). \tag{12}$$

This result can also be derived by restricting the range of the convolution integrals in (3) to frequencies where at least one parent spectrum carries significant power.

For a static misalignment, $\tilde{d}_\nu(f) = \overline{d_\nu} \, \delta(f)$, we recover Equation 9 and a mirror angle fluctuation at frequency f_0 linearly induces a Fourier component of optical length at the same frequency. This can be shown to give an adequate prediction of the optical length spectrum, without resorting to the more general form (12), if $\overline{d_\nu} \gg d_\nu^{rms}$, where

$$(d_\nu^{rms})^2 \equiv \frac{1}{T} \int_0^T \left[d_\nu(t) - \overline{d}_\nu\right]^2 dt \quad = \int_{-w}^{w} \left|\tilde{d}_\nu(f)\right|^2 [1 - \delta(f)] \, df \tag{13}$$

is the mean-square deviation of the optic axis (excluding the static offset).

On the other hand, if $\overline{d_\nu} \lesssim 2 \ d_\nu^{rms}$ the optical path length spectrum will not be linearly related to the angle spectra and will contain frequency-shifted sideband products. Nevertheless, in situations of interest the high-frequency mirror angle spectra $\tilde{\epsilon}_\nu(f)$ are often smooth and nearly constant over a band of frequencies

$\Delta f \gtrsim 2w$. Such spectra arise from electronic noise in the mirror control system or thermal fluctuations in the suspension, for example. We can then derive the approximation

$$\left|\tilde{l}(f)\right|^2 \approx \left[\overline{d_1}^2 + (2d_1^{rms})^2\right] |\tilde{\epsilon}_1(f)|^2 + \left[\overline{d_2}^2 + (2d_2^{rms})^2\right] |\tilde{\epsilon}_2(f)|^2 \qquad (f > 2w). \tag{14}$$

Under the above conditions, we may replace the squared Fourier transforms with the power spectral densities $S_{\epsilon_1}(f)$ and $S_{\epsilon_2}(f)$ of random processes ϵ_1 and ϵ_2 to obtain

$$S_l(f) \approx \left[\overline{d_1}^2 + (2d_1^{rms})^2\right] S_{\epsilon_1}(f) + \left[\overline{d_2}^2 + (2d_2^{rms})^2\right] S_{\epsilon_2}(f) \qquad (f > 2w), \tag{15}$$

where $S_l(f) \equiv \lim_{T\to\infty} 2\left|\tilde{l}(f)\right|^2/T$ is the power spectral density of induced cavity length fluctuation.

In the above we treat static and RMS errors as separate coupling parameters for purely practical reasons, related to our intended application. The specialization of d_ν^{rms} to specifically exclude "D.C." is motivated by the qualitative difference in the origins and methods for dealing with fluctuations which occur on timescales much longer than gravitational wave measurements, as opposed to those which occur on comparable or shorter timescales. Long–term misalignments in gravitational wave interferometers result from thermal, tidal or relaxation effects, and can be reduced to an acceptable level by periodic readjustment. Shorter–term misalignments, generally driven by seismic noise, are only reduced to a finite extent by action of the mirror control system; the residual RMS seismic excitation thus constitutes a practical lower limit to the coupling between angle and length.

3. Experimental Tests

This model was tested on the LIGO 40 meter interferometer[7]. Briefly, this instrument comprises two orthogonal 40 m cavities whose 1.5 kilogram test bodies are suspended by wires so they are essentially free to translate in a horizontal plane and to rotate about axes perpendicular to the laser beam (Figure 2). Referring to Figure 1, the geometrical parameters of each cavity are $R = 62$ m, $L = 40$ m, and $a_1 = a_2 = 6$ cm.

The pendulum mode and both rotational normal modes of the suspended bodies have eigenfrequencies near 1 Hz. Each is controlled in azimuth (ϕ) and elevation (θ) by a control system which derives appropriate feedback signals from an optical lever sensor and applies magnetic corrective torques to the suspension. Manual offsets can be introduced and trimmed to adjust and optimize (or degrade) the alignment of the two cavities.

The control electronics are provided with summing nodes through which probe signals are introduced to measure the transfer characteristics for various degrees of freedom. The induced rotations are calibrated by measuring the motion of the optical lever reflections. The interferometer's differential displacement output signal, which represents the difference between the cavity optical lengths, is monitored and itself periodically calibrated against a known test force applied electromagnetically to one test body. The differential displacement induced by the probe signal reflects the change in the length $l(t)$ of the probed cavity. This length signal is Fourier analyzed with a digital spectrum analyzer to pick out the components resulting from the probes (while discriminating against other components) and to measure their power spectra.

In the presence of a large static D.C. misalignment $\overline{d_\nu}$ between the cavity axis and one mirror's center of mass, Equation 9 implies that the coupling of that mirror's angle to apparent cavity length will be $\tilde{l}(f)/\tilde{\epsilon}_\nu(f) = \overline{d_\nu}$. To test this prediction, a small sinusoidal probe torque resulting in a vertical oscillation $\epsilon_2(t) = a \cos(2\pi f_0 t)$ with $a = 3.5 \times 10^{-9}$ radian was applied to one mirror (M_2 in Figure 1) at $f_0 = 250$ Hz. The probe amplitude was calibrated by observing the deflection of an optical lever beam reflected from that mirror. The interferometer cavity length signal was recorded as the vertical distance $\overline{d_2}$ between the cavity axis and the center of that mirror was varied by adjusting the alignment controls for both mirrors and the direction of the incident laser beam. The true position of the cavity axis was monitored by photographing scattered light from the resonating cavity mode against the outline of the mirror. Figure 3 shows the Fourier component of the measured cavity length at f_0 divided by the induced mirror angle probe amplitude, $\tilde{l}(f_0)/\tilde{\epsilon}_2(f_0)$, vs. $\overline{d_2}$. Since the exact position of the test body's center of mass was not accurately known, we have chosen the reference point $\overline{d_2} = 0$; the measured slope agrees, within experimental errors, with that predicted by Equation 9.

In more realistic situations, the mirror angle spectra will consist of many uncorrelated components, so the cross products of many different frequency pairs will generally be superimposed at each frequency in the convolved result. To investigate this case, a band-limited random noise test signal was generated to induce random angle fluctuations of M_2 principally between 200 and 315 Hz ,

$$|\tilde{\epsilon}_2(f)| \approx \begin{cases} 0, & f \lesssim 200 \text{ Hz} \\ K, & 200 \text{ Hz } \lesssim f \lesssim 315 \text{ Hz} \\ 0, & f \gtrsim 315 \text{ Hz} \end{cases} \qquad (16)$$

where K is a constant. The cavity length's spectral density was monitored in this frequency band as a function of $\overline{d_2}$. As shown in Figure 4, the magnitude of the effective transfer function for this random noise probe signal is approximately proportional to the magnitude of $\overline{d_2}$ at large offsets, as in Figure 3, but remains essentially constant for very small $\overline{d_2}$. To investigate this behavior, the light transmitted through M_2 was analyzed with a position–sensing quadrant photodetector to measure low–frequency fluctuation of the cavity axis position $\tilde{d}_2(f \ll f_0)$. Integrating this position spectrum over frequency (and excluding D.C. as in Equation 13) gave $d_2^{rms} \approx (0.2 \pm 0.1)$ mm. From Equation 15 we expect the linear proportionality approximation to fail for $|\overline{d_2}| \lesssim 2d_2^{rms} \approx (0.4 \pm 0.2)$ mm, essentially the region in which the data shown in Figure 4 deviate from the linear prediction. While the method used to obtain the data in Figures 3 and 4 can be used for each mirror to empirically adjust all $\overline{d_\nu}$ precisely to zero, the average coupling coefficient will generally reach a nonzero minimum of order $2\,d_\nu^{rms}$.

To quantitatively test (12) a two–frequency probe signal

$$\tilde{\theta}_2(f) = \theta_2^a\,\delta(f_a) + \epsilon_2\,\delta(f_b) \qquad (17)$$

was imposed on M_2 with a large low–frequency component ($\theta_2^a = 1.6 \times 10^{-6}\,\mathrm{rad}_{rms}$, $f_a = 10$ Hz) and a small high–frequency component ($\epsilon_2 = 5.0 \times 10^{-9}\,\mathrm{rad}_{rms}$, $f_b = 250$ Hz). The natural $\tilde{\theta}_2(f)$ spectrum was small enough to be completely dominated by these probe components in their respective frequency regimes. The cavity length signal displayed the expected pair of equal–amplitude Fourier components at 260 Hz and 240 Hz, with measured amplitudes of $(3.9 \pm 1.0) \times 10^{-13}\,\mathrm{m}_{rms}$; direct substitution of the probe spectrum (17) into Equation 12 would predict

$$\tilde{l}(f) = 3.4 \times 10^{-13}\,\mathrm{m}_{rms} \times \;\; [\delta(f_b + f_a) + \;\delta(f_b - f_a)]\,, \qquad (18)$$

agreeing within measurement errors.

The $f_a = 10$ Hz component of the probe rotation was then removed and the 250 Hz test signal was reduced in amplitude to $\epsilon_2 = 1.5\times10^{-9}\mathrm{rad}_{rms}$. Analysis of the cavity length spectrum then revealed natural sidebands, symmetric around the 250 Hz probe. The vertical position of the cavity axis at M_2 was simultaneously measured by the transmitted–beam quadrant photodetector. The detailed spectral shape found for the cavity length signal matches the spectrum of cavity axis position, upshifted by 250 Hz and with an amplitude close to that predicted by Equation 12 (Figure 5). In effect, the monochromatic 250 Hz probe has acted as a virtual delta function in the convolution integral, "picking out" of the convolution an upshifted replica of the residual low-frequency beam offset spectrum.

4. Application examples

The following examples illustrate the use of Equation 15 to understand the effect of orientation fluctuations, in the first instance to improve an existing interferometer and in the second to develop specifications for a future instrument.

40 meter interferometer: Our analysis was applied to the 40 meter interferometer in 1991 to analyze its noise spectrum and achieve a significant improvement in its sensitivity. The transfer functions from mirror torque control signals to mirror angles and the residual noise in these control signals were measured, and combined with estimates of the displacement between each cavity mirror's center of mass and the optical axis of that cavity, to form an estimated noise contribution from each angular degree of freedom to the total displacement noise. These estimates indicated that mirror orientation noise was a significant component of the observed displacement spectrum[8]. The conclusion was confirmed by demonstrating that improved coincidence between optical and inertial axes (achieved by iterative realignment) or reduced high–frequency angle noise (achieved by reducing control system gain) could reduce interferometer noise over a broad spectral band, principally between 50 and 700 Hz.

Although high–frequency angle fluctuations can arise from many phenomena, the dominant source in this case was found to be excess noise in the mirror orientation control systems, resulting from a combination of direction and intensity fluctuation in the optical lever laser beams, electronic noise in control electronics, and inadequate filtering outside the control band. Improved feedback electronics were developed with lower unity gain frequency, increased filter attenuation and lower intrinsic electronic noise. At frequencies near 100 Hz, the improved controllers would contribute a factor of 10^5 less noise than the systems they replaced for a given offset between cavity optical and inertial axes. This change reduced the interferometer displacement noise by a significant factor at frequencies between 50 and 700 Hz, leaving the sensitivity limited by other noise mechanisms except at narrow mechanical resonances (Figure 6).

LIGO interferometer: Specifications for LIGO test body orientation controls can also be derived by application of the model. Initial LIGO interferometers are expected to achieve total displacement noise spectral densities below 10^{-19} $\mathrm{m}/\sqrt{\mathrm{Hz}}$ at frequencies near 100 Hz[1]. To meet this goal Equation 15 permits angular fluctuations no larger than

$$S_{\theta_\nu}^{1/2}(f) \lesssim 10^{-17} \ \frac{\mathrm{rad}}{\sqrt{\mathrm{Hz}}} \tag{19}$$

for each angular degree of freedom of the four cavity mirrors at frequencies $f \sim 100$ Hz, assuming the projected $d_\nu^{rms} \sim 0.5\,\mathrm{mm}$ is achieved at remote LIGO sites[9]. Here we have omitted the additional linear contribution due to nonzero $\overline{d_\nu}$, since it can be empirically trimmed to zero by monitoring the response of each mirror to probe torques; see Figures 3 and 4 and discussion.

The improved 40 m interferometer mirror control systems described above induce approximately $5 \times 10^{-16}\ \mathrm{rad}/\sqrt{\mathrm{Hz}}$ angle fluctuations near 100 Hz, principally due to electronic noise in their active filters. Substitution of passive or lower–noise active filtering would give a reduction in noise by a factor of order 30 or more at 100 Hz. In addition, the remote LIGO sites are seismically quieter than the campus laboratory by approximately an order of magnitude at relevant frequencies. The dynamic reserve of the controllers can be correspondingly reduced, reducing the torque induced by a given amount of electronic noise in the circuitry.

5. Conclusions

Simple geometric considerations can accurately model the sensitivity of Fabry–Perot cavities to high–frequency angular rotations of their mirrors, allowing straightforward analysis of torque–induced noise. In addition to guiding the design of improved mirror control systems for the LIGO 40 m interferometer, which greatly enhanced the performance of that instrument, application of the model to planned full–scale observatory interferometers indicates that with available control technology, angular rotations need not compromise LIGO performance at target sensitivity levels.

Acknowledgments: We wish to thank all who contributed to the construction and improvement of the interferometer, and are grateful for the continuing support of LIGO science, engineering and administrative staff. We are especially grateful to J. Chapsky for designing the improved control electronics and to A. Gillespie, F. Raab, R. Spero and S. Whitcomb for their assistance in measuring and analyzing the data and preparing the manuscript. Our work is supported by National Science Foundation grant PHY-9210038.

References

1. A. Abramovici, W. E. Althouse, R. W. P. Drever, Y. Gürsel, S. Kawamura, F. J. Raab, D. Shoemaker, L. Sievers, R. E. Spero, K. S. Thorne, R. E. Vogt, R. Weiss, S. E. Whitcomb, and M. E. Zucker, "LIGO: the laser interferometer gravitational wave observatory," Science **256**, 325-333 (1992).

2. R. E. Vogt, "The U. S. LIGO project," in *Proceedings of the Sixth Marcel Grossmann Meeting on General Relativity,* H. Sato and T. Nakamura, eds. (World Scientific, Singapore, 1991), p. 244.

3. S. L. Smith, *A Search for Gravitational Waves from Coalescing Binary Stars Using the Caltech 40 Meter Gravity Wave Detector,* Ph.D. thesis (California Institute of Technology, Pasadena, Ca., 1988), p. 37.

4. Equation 1 and the rest of the discussion can be readily applied to any cavity geometry by suitably redefining α, β and γ. In general, β and γ are not linearly dependent.

5. R. G. Brown, *Introduction to Random Signal Analysis and Kalman Filtering* (Wiley, New York, 1983), Chap. 2, p. 85.

6. S. O. Rice, "Mathematical analysis of random noise," in *Selected Papers on Noise and Stochastic Processes,* N. Wax, ed. (Dover, New York, 1954), pp. 133-294.

7. M. E. Zucker, "The LIGO 40 m prototype laser interferometer gravitational wave detector," in *Proceedings of the Sixth Marcel Grossmann Meeting on General Relativity,* H. Sato and T. Nakamura, eds. (World Scientific, Singapore, 1991), p. 224.

8. S. Kawamura, "Test mass orientation noise in the LIGO 40 m prototype," in *Proceedings of the Sixth Marcel Grossmann Meeting on General Relativity,* H. Sato and T. Nakamura, eds. (World Scientific, Singapore, 1991), p. 1486.

9. Remote LIGO sites exhibit approximately one tenth the seismic amplitude found in the campus laboratory at frequencies which determine the RMS mirror angle fluctuation. In addition, current seismic isolation stack and suspension designs show one fifth the resonant amplification effect found to enhance RMS mirror motion in the 40 m interferometer. While the hundredfold increase in length would proportionately increase the d_ν^{rms} arising from a given RMS angle fluctuation, these factors should reduce that angle fluctuation by a factor of order fifty, leading us to expect only a doubling in net d_ν^{rms} in full–scale interferometers.

Figure Captions

Figure 1. Geometry of a half-symmetric Fabry-Perot cavity consisting of flat (M_1) and concave (M_2) test body/mirrors.

Figure 2. Suspended test body and orientation control system used in the LIGO 40 meter interferometer. The quadrant photodetector and electronic processor determine angular error signals from the position of an auxiliary laser beam reflected from the mirrored surface of the test body. These signals are filtered and amplified and applied to pairs of electromagnetic coils near the suspension point. The coils interact with permanent magnets (poled oppositely to induce torque) on an intermediate platform which is suspended by a single wire so that it is free to rotate. The torque developed on this platform is transmitted to the test body below by the suspension wires, inducing rotation of the mirror and closing the feedback loop. Probe torques are introduced by adding currents to the feedback signals driving the coils.

Figure 3. Linear mirror angle—cavity length coupling coefficient $\tilde{l}(f_0)/\tilde{\varepsilon}_2(f_0)$ as a function of beam axis position $\overline{d_2}$. The dashed line is the curve $\tilde{l}(f_0)/\tilde{\varepsilon}_2(f_0) = \overline{d_2}$ predicted by Equation 9. The zero for the $\overline{d_2}$ axis has been chosen for best fit.

Figure 4. Interferometer displacement due to bandlimited random mirror angle noise as a function of $\overline{d_2}$. The coupling magnitude appears proportional to $|\overline{d_2}|$ down to $|\overline{d_2}| \approx 0.4$ mm, below which it is approximately constant. By integrating the spectrum of position fluctuations in the transmitted cavity mode, it was found that $d_2^{rms} \approx 0.2 \text{ mm} \pm 0.1 \text{ mm}$ during this experiment, in agreement with the transition from linear behavior predicted by Equation 15.

Figure 5. Spectrum of low-frequency beam axis position fluctuations (heavy line, upper frequency scale and right-hand magnitude scale) is replicated as upper and lower sidebands of an artificially induced 250 Hz probe angle fluctuation in the interferometer displacement spectrum (thinner line, lower and left scales). The relative scales are chosen in accord with Equation 12 using the known 1.5×10^{-9} rad_{rms} amplitude of the 250 Hz probe.

Figure 6. Spectral density of apparent cavity mirror displacement in the 40m interferometer with the original orientation control system (upper curve, heavy line) and after installing the new orientation feedback electronics (middle curve, heavy line). The lower, thin curve depicts the estimated displacement noise induced by residual electronic noise in the new controller for one axis of one test body, using Equation 15 and an assumed $\overline{d_2} \approx 1$ mm. The peak at 212 Hz is a resonance of the wire suspension system, where the coupling of external torque to the mirror is locally enhanced.

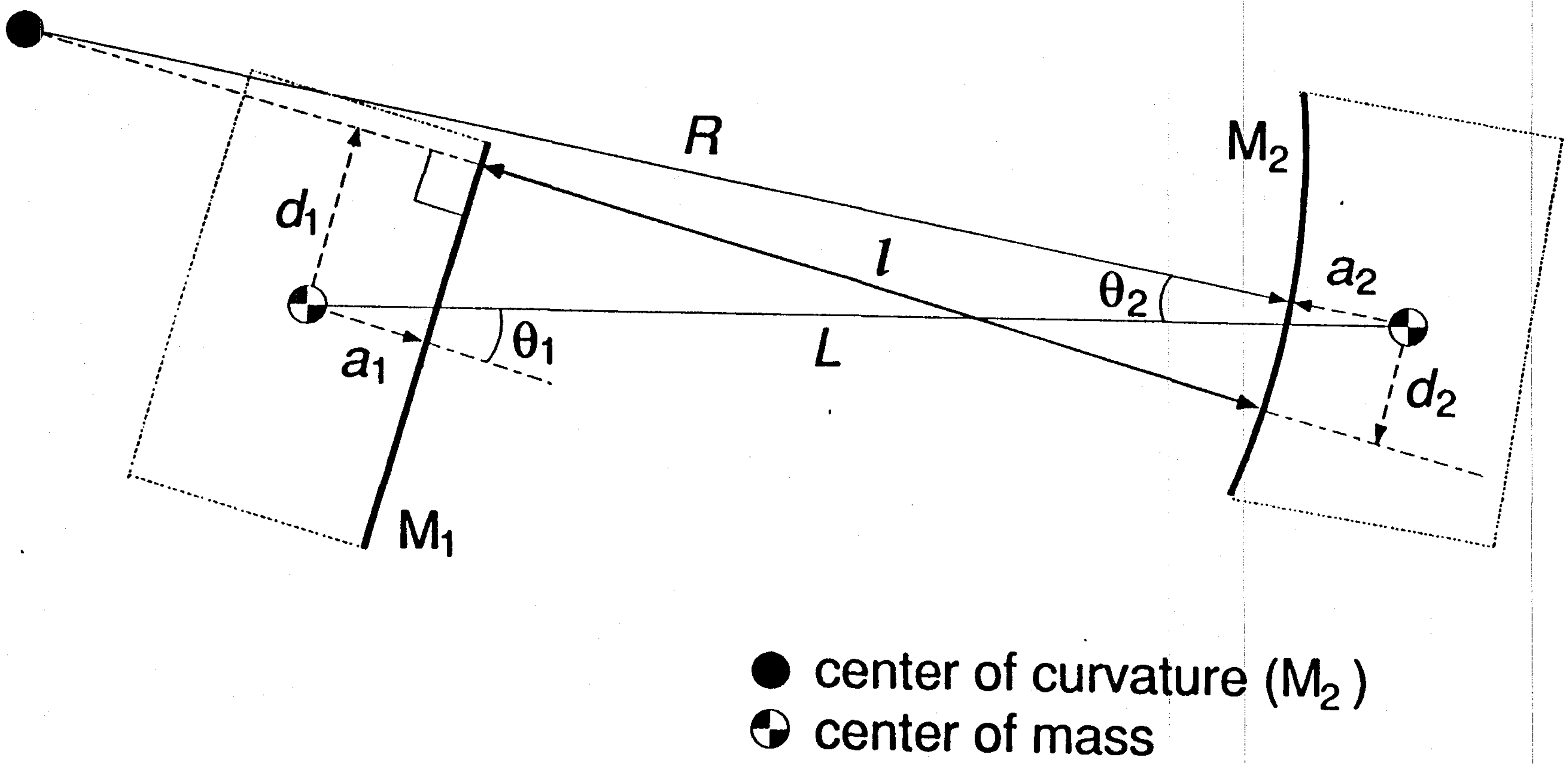

R
M2
d1
l
θ2
a2
θ1
a1
L
d2
M1
● center of curvature (M2)
⊕ center of mass

Figure 1

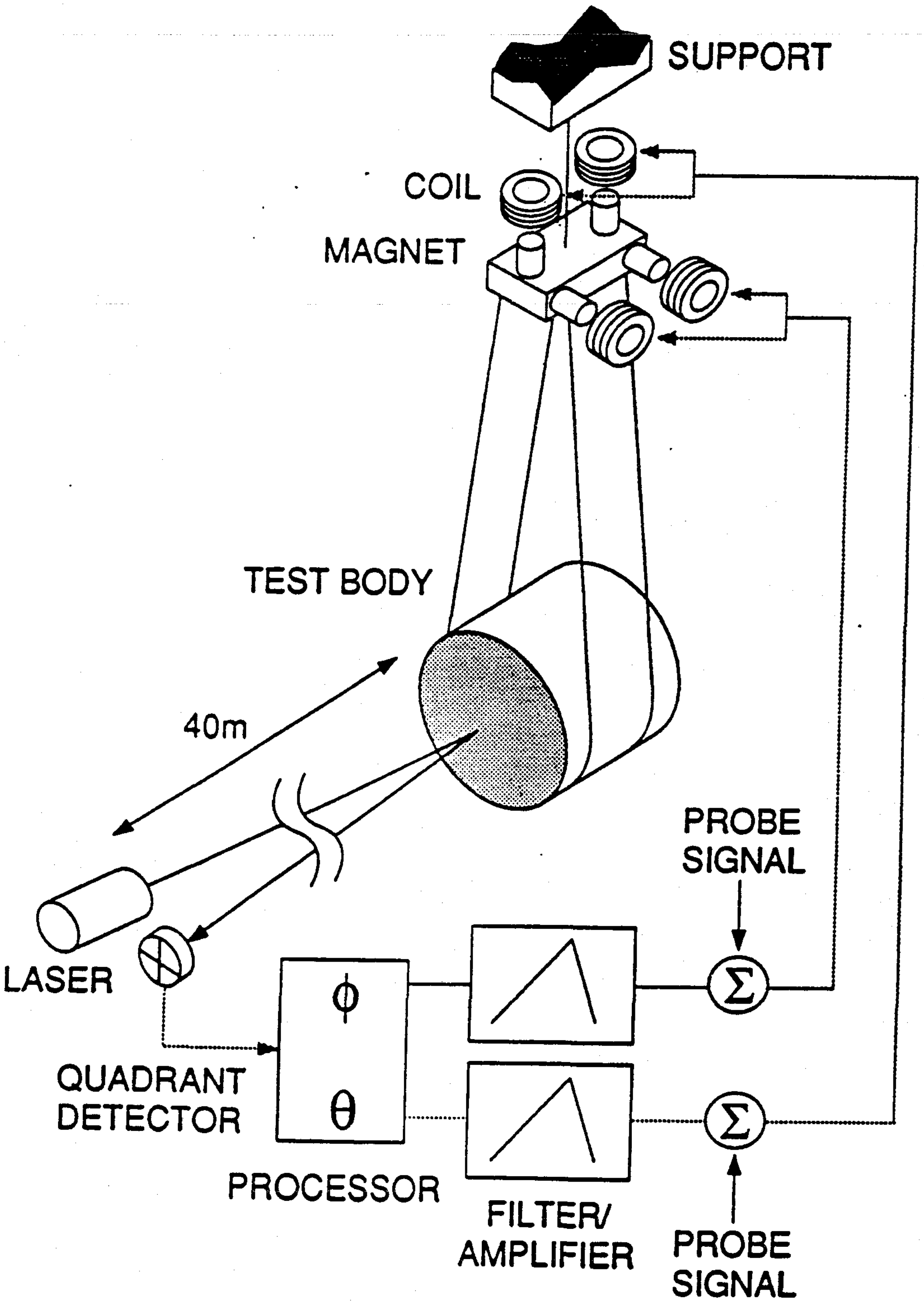

SUPPORT
COIL
MAGNET
TEST BODY
40m
PROBE
SIGNAL
LASER
ϕ
Σ
QUADRANT
DETECTOR
θ
Σ
PROCESSOR
FILTER/
AMPLIFIER
PROBE
SIGNAL

Figure 2

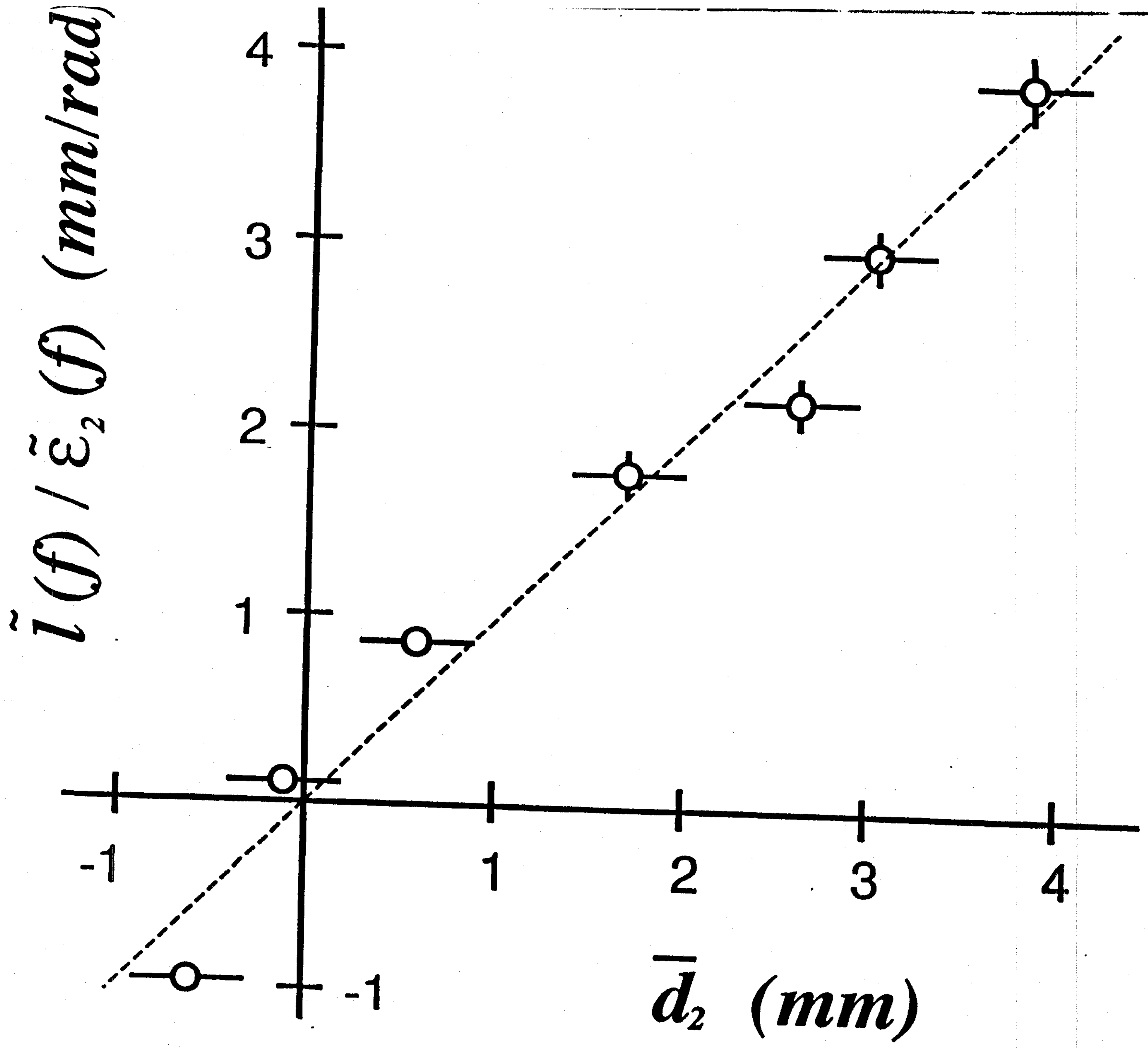
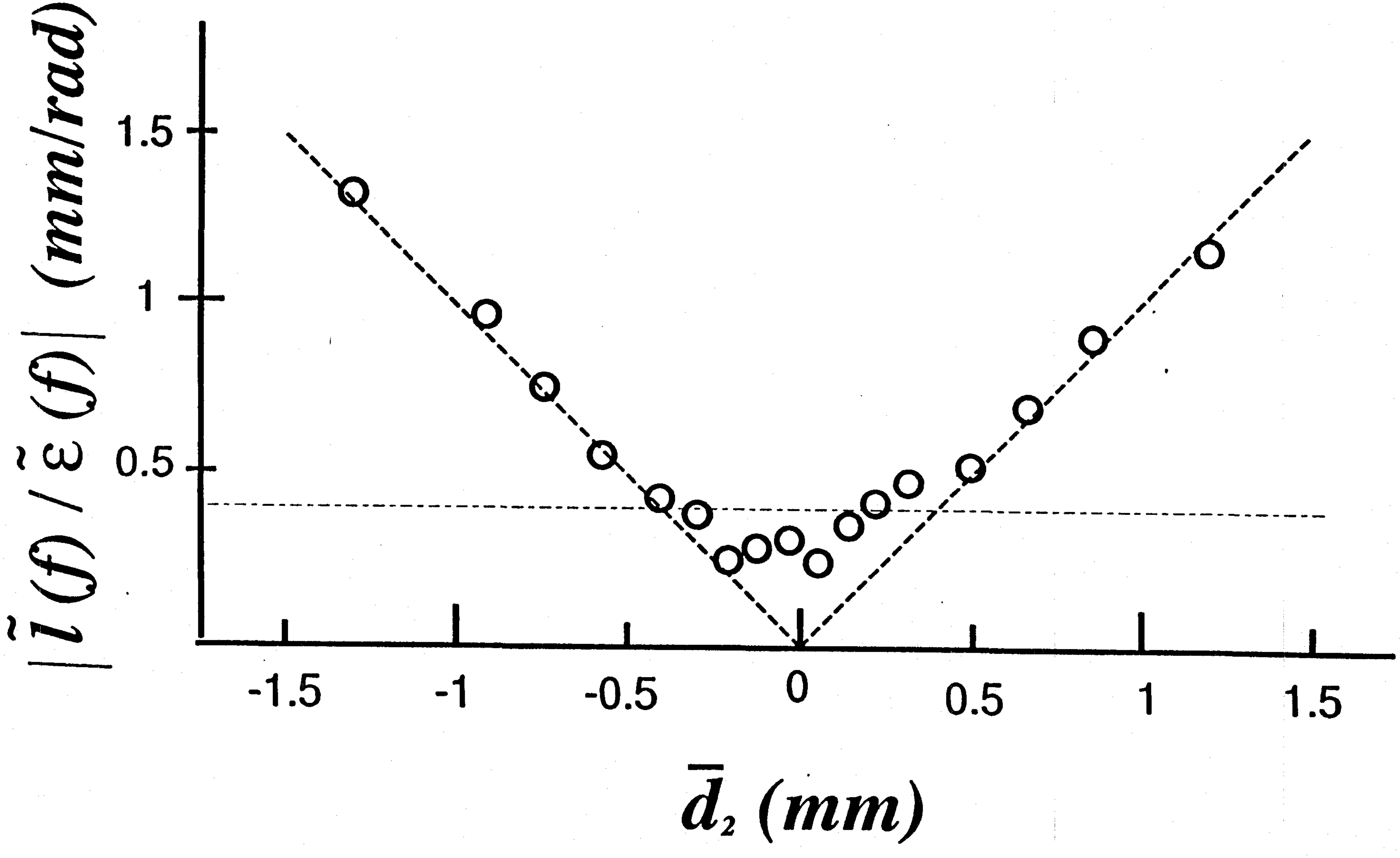


Figure 3


$|\tilde{l}(f) / \tilde{\varepsilon}(f)|$ (mm/rad)
1.5
1
0.5
-1.5
-1
-0.5
0
0.5
1
1.5
$\bar{d}_2$ (mm)

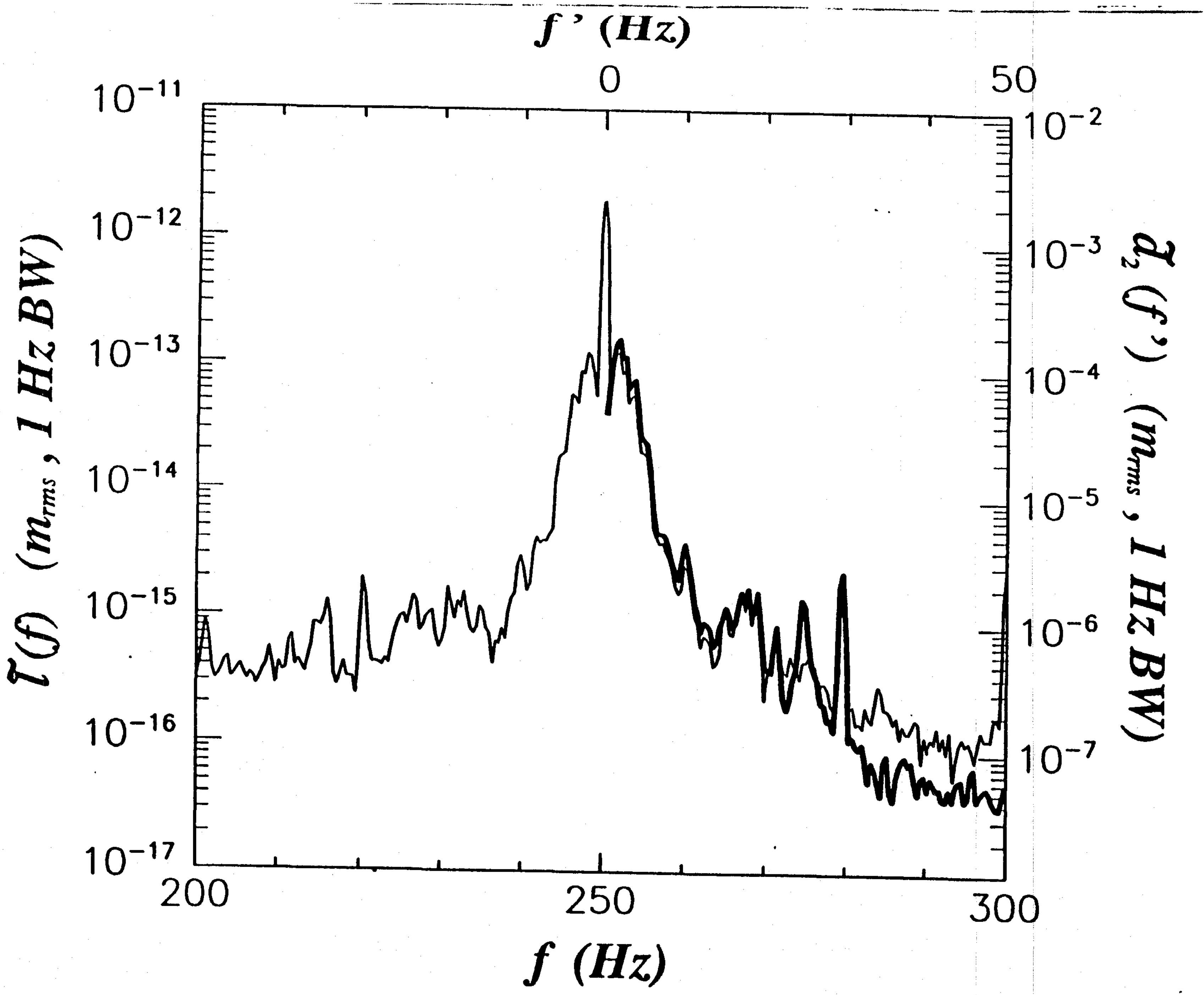
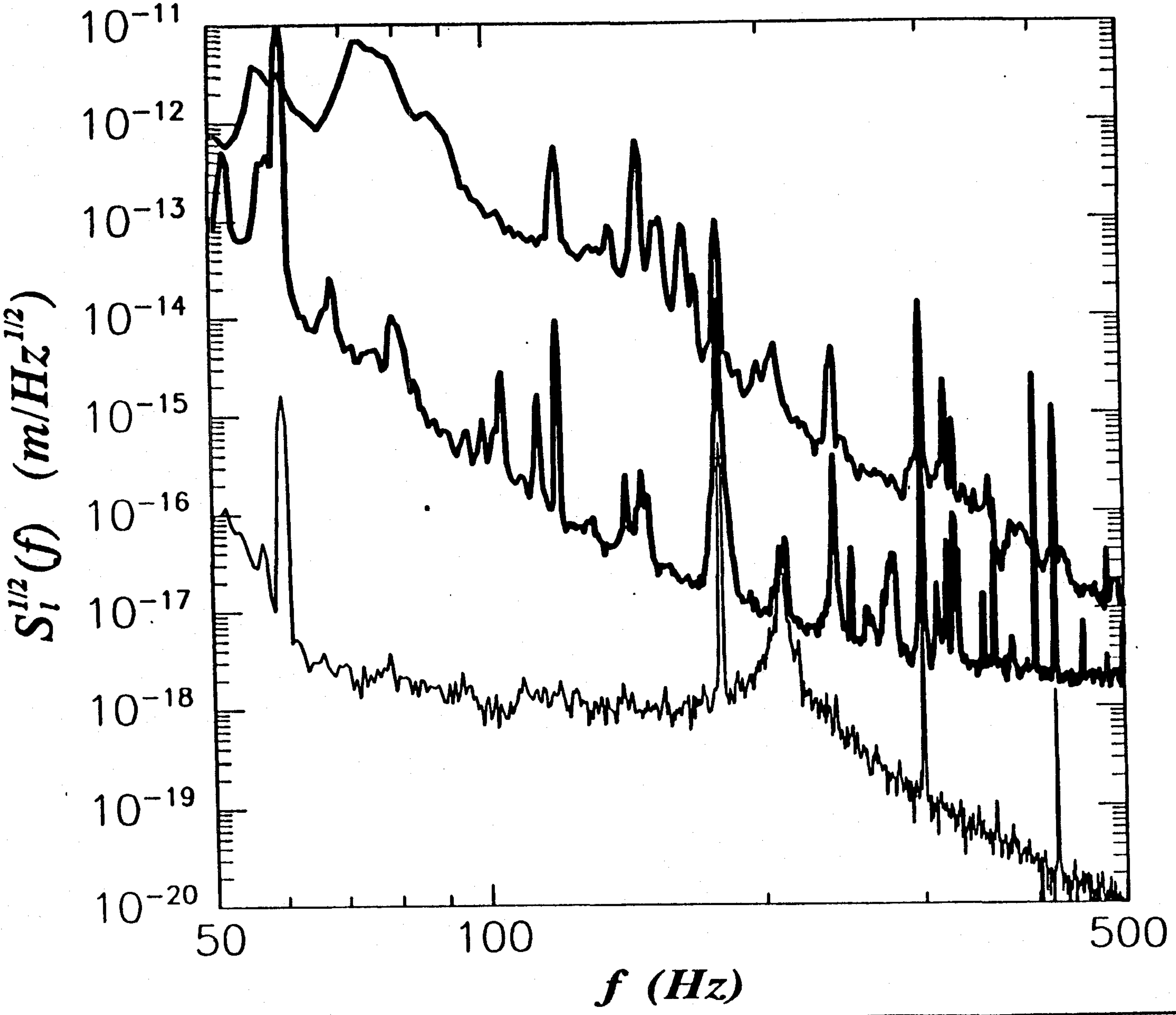

f ' (Hz)
0
50
10^{-11}
10^{-12}
10^{-13}
10^{-14}
10^{-15}
10^{-16}
10^{-17}
$\tilde{l}(f)$ (m_{rms} , 1 Hz BW)
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}
$\tilde{d}_z(f')$ (m_{rms} , 1 Hz BW)
200
250
300
f (Hz)

Figure 5



10^{-11}
10^{-12}
10^{-13}
10^{-14}
10^{-15}
10^{-16}
10^{-17}
10^{-18}
10^{-19}
10^{-20}
$S_l^{1/2}(f)$ $(m/Hz^{1/2})$
50
100
500
f (Hz)

REF. CC

Figure 6

A double pendulum vibration isolation system for a laser interferometric gravitational wave antenna

M. Stephens, P. Saulson,[a)] and J. Kovalik
Massachusetts Institute of Technology. 20F-001. 77 Massachusetts Ave.. Cambridge, Massachusetts 02139

(Received 14 August 1990; accepted for publication 3 December 1990)

We have developed a nested double pendulum suspension system for the test masses of a laser interferometric gravitational wave antenna. The system consists of a mass hung as a pendulum inside a shell mass that is also hung as a pendulum. A set of two-degree-of-freedom reflective "shadow detectors" senses motion of the shell relative to ground. Identical sensors measure motion of the mass relative to the shell. The equations of motion were solved to find the resonances and mode shapes for all of the rigid body degrees of freedom. The predicted resonant frequencies agree well with the measured frequencies. A damping system has been implemented that damps the resonances by applying forces to the shell mass alone. The vibration transfer function along the optic axis was measured. It shows the steep f^{-4} decline expected of a double pendulum. We have also measured the vertical vibration transfer function and the cross coupling due to misalignment. A set of plates on the inner surface of the shell allows the application of low noise electrostatic forces directly to the test mass for high-bandwidth control such as interferometer fringe lock. We have measured the response of the system to this input, and compared it to that predicted by our model equations of motion. We have determined that there exist stable feedback loops that can maintain fringe lock. The possibilities of active isolation are discussed.

I. INTRODUCTION

Interferometric gravitational wave detectors measure extremely small relative displacements between the mirrored faces of several "test masses" arranged in the L-shaped configuration of a Michelson interferometer.[1] The design of the test mass suspensions must satisfy several demands simultaneously. First, the test masses must be approximately free in a frequency range of interest so that they can respond to the weak gravitational forces in a gravitational wave. Second, they must be free from mechanical noise that might mask the tiny motions expected from a gravitational wave. Finally, they must be controlled so the interferometric read-out system can be aligned and held locked to a fringe.

The first requirement, dynamical response approximately that of a free mass, can be satisfied by supporting the test masses with some form of compliant connection to the outside world. At low frequencies, the response of the test mass is dominated by the spring constant of its mounting. But above the resonant frequency of the oscillator, the response approaches that of a free mass.

The compliant mounting of the test mass is a start toward satisfying the second requirement as well. One of the largest sources of mechanical noise is the ubiquitous background of seismic noise communicated to the test masses through their supports. The compliant suspension functions as a vibration isolator, attenuating the external vibrations at frequencies high compared to the resonance. If the isolation provided by one oscillator [proportional to $(f/f_0)^2$, where f_0 is the resonant frequency of the oscillator] is insufficient, it can be supplemented by additional mass-spring stages.

Noise can also be generated within the structure itself. The most fundamental mechanism of this type is thermal noise, the analog of Brownian motion in the macroscopic suspension. The amplitude of this noise is minimized by arranging for the lowest possible level of mechanical losses in the final stage of the suspension. A pendulum is almost always used for the suspension of the mass itself, because of its low losses.

There are several different control functions to be performed. The low mechanical losses used to reduce thermal noise have as their side effect high-Q resonances. Servo damping is used to reduce the mechanical excitation of these modes. In addition, control forces must be applied to the compliantly mounted test mass to keep it aligned with the rest of the optical system, and to adjust the optical path length to fix the relative phase of the interfering light beams in the interferometer. All of these functions must be performed without short circuiting the vibration isolation of the system, and without introducing significant amounts of mechanical noise.

A number of approaches to meeting these demands have been taken by workers in the field.[2-6] In this article, we describe a suspension system that we designed with the goal of meeting the stringent specifications necessary to detect and study gravitational waves from expected astrophysical sources with a detector such as the long-baseline

[a)] Also at Joint Institute for Laboratory Astrophysics, University of Colorado at Boulder, Campus Box 440, Boulder, CO 80309-0440. Present address: Department of Physics, Syracuse University, Syracuse, NY 13244-1130.

interferometers being developed by the Caltech/MIT LIGO project and several groups around the world.

In the next section, we explain the specific design requirements we set for the suspension, showing how this led to a system with the features that we adopted. The succeeding sections describe an extensive series of tests that we performed to try to determine how well the suspension performed. Finally, we describe what aspects of the performance still need verification.

II. DESIGN PRINCIPLES

A central theme of the design was to keep the structure of the test mass itself as simple as possible. This was done in an attempt to maintain the high mechanical Q of the test mass and thus minimize the thermal excitation of the internal resonances of the test mass. We strove to find alternatives to control mechanisms that involve attachment of magnets for actuators or shadow tabs for position sensors. This is the reason for the choice of electrostatic actuators for the forces that need to be applied to the test mass itself. The pendulum is hung by a single loop of wire around its middle, as done by the Max Planck group.[5] For applications that require direct sensing of the position of the test mass, we made provision for reflective optical sensors.

Another theme of the design was to attempt to achieve sufficient isolation so that gravitational wave interferometers could be limited by the fundamental measurement noise, photon shot noise, down to a frequency of 100 Hz. We wanted to do this without the use of elastomer-based isolation stacks[7,8] or air-spring systems,[9] believing that a simpler system would suffice. This does not preclude the use of other isolators with this system.

The solution we adopted to meet these requirements is the nested double pendulum. The vibration transfer function should be that of two low-frequency oscillators in series, $f_{0_1}^2 f_{0_2}^2/f^4$ in the limit of high frequencies where f_{0_1} and f_{0_2} are the resonant frequencies of the two-oscillator system. By nesting the test mass inside the closely fitted shell mass, we provided a platform from which to apply the control forces. This platform is itself vibration-isolated by virtue of its placement on the upper mass of the double pendulum, although it is not as quiet as the test mass itself. (An alternative design based on similar ideas has been pursued by Cantley and co-workers[4]).

This design offers the additional benefit that measurement of the relative displacement of the test mass with respect to the shell mass gives an error signal that can be nulled by a servosystem that applies appropriate forces to the shell. This is the classic configuration for active vibration isolation, which effectively reduces the frequency of one of the normal modes of the isolation system (see the reference section). Although we did not test the system in this style of operation, we studied some of the relevant transfer functions to learn its possibilities and limitations.

As a final design principle, we attempted to minimize the use of materials whose outgassing could interfere with the need to achieve a good vacuum in the gravitational wave interferometer.

III. SYSTEM DESCRIPTION

The nested double pendulum consists of a shell mass surrounding a 10 kg mirror mass. The mirror mass is a 20.32-cm-diam, 10.16-cm-long cylinder. This aspect ratio was chosen so that the frequencies of the lowest orders of the vibrational modes of the cylinder are as high as possible.[10,11] In this prototype, the mirror mass was made of aluminum. For an actual test mass in a full scale gravitational wave detector it most likely would be a quartz mirror.

The mirror mass is held about the middle by a single 0.2-mm-diam tungsten wire. The wire is close to breaking; thus, the violin mode of the wire is at the highest possible frequency. Two 0.32-cm-diam cylinders at ten degrees above the midline increase the normal force of the wire on the mirror mass at the point where the wire leaves the mirror mass to prevent slippage. The cylinders are held in place by friction. .

The mirror mass has four polished flats on it that are used in conjunction with the reflective position sensors to measure motion of the mirror mass relative to the shell.

The 10 kg aluminum shell has a 25.4 cm o.d. and is 18 cm long. The mirror mass hangs centered within the shell with a 1 mm gap on all sides between the inner edge of the shell and the mirror mass. Clamps on the top of the shell hold the two ends of the tungsten wire that suspends the mirror mass.

To allow for the electrostatic actuators mentioned in the previous discussion, annuli at the front and back of the shell overlap the front and back of the mirror mass by 1.27 cm. The shell is cut into sections that are insulated from each other by thin teflon sheets. Thus a voltage applied to the front and back plates can produce a force on the electrically grounded mirror mass. Both the front and the back overlapping sections of the shell are cut into four quadrants, allowing an axial force and two torques to be applied to the mirror mass from the shell.

The shell also has four holes drilled in it that allow mounting of reflective position sensors. These sensors are attached to the shell and, by using the reflective surfaces on the mirror mass, they measure motion of the mirror mass relative to the shell. Four position sensors attached to the ground use four reflective surfaces on the shell to measure motion of the shell relative to the ground.

The sensors consist of shadow detectors, which use a small infrared LED shining through a hole in the center of a quadrant photodiode. The light shines through the hole onto a polished flat on the shell mass that has a small dark patch in the center. The difference in the photocurrents from two opposite quadrants indicates the relative displacement of the sensor and the shell mass (Fig. 1). The sensitivity is limited by shot noise down to frequencies below 5 Hz, at a level of 4×10^{-10} m/$\sqrt{\mathrm{Hz}}$. This signal is linear over approximately 2 mm.

Attached to the shell are six magnets that are used by magnetic actuators. A 750 turn coil wound on a teflon core is attached to the ground near each magnet, allowing adjustment of the position of the shell relative to the ground (a variant on a design developed by the Max Planck

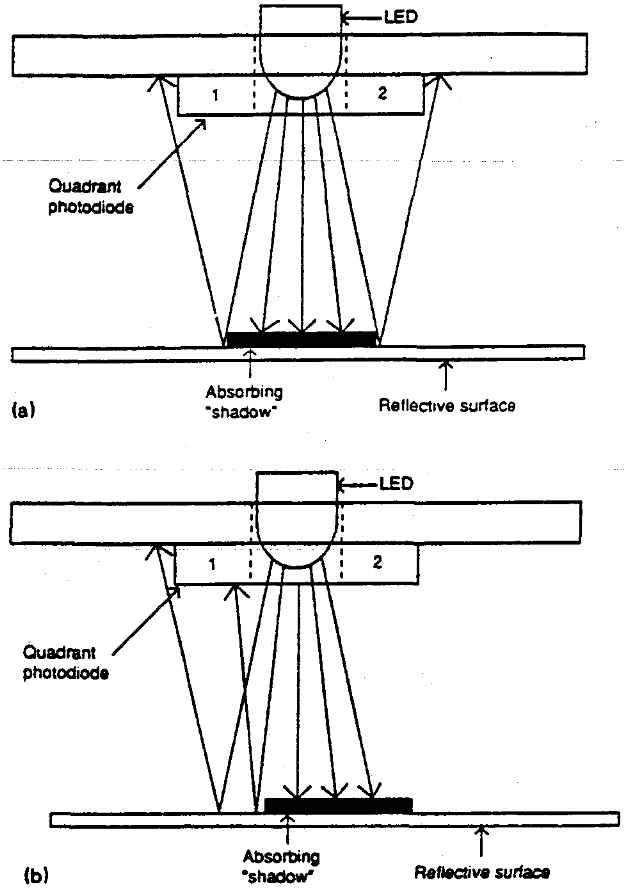


FIG. 1. (a) An aligned shadow detector. The size of the absorbing patch is chosen so that no light falls on quadrant 1 or quadrant 2. (b) The reflective surface has moved. Now more light falls on quadrant 1 than quadrant 2 so that the difference between the signals from the two quadrants is nonzero.

group[5]). The magnets are placed so that all six degrees of freedom of the shell can be controlled. The actuators produce a maximum force of about 0.03 N over a range of 1 mm.

The shell itself is hung as a pendulum. Four tungsten wires (also 0.2 mm in diameter), each in series with a coil spring, hold the shell and mirror mass. The wires are clamped to the shell at ten degrees above the midline.

Figure 2 shows a diagram of the suspension and defines the coordinate system that will be used throughout this article.

The linearized equations of motion for the double pendulum suspension can be derived by finding the Lagrangian for the coupled system and then writing down the Euler-Lagrange equations in the small angle approximation. This is simplified by recognizing that for a system with no misalignment, the motion can be described by four combinations of coupled motion. These are translation along the optic axis and rotation about the y axis, translation along the y axis and rotation about the optic axis, translation along the vertical axis, and rotation about the vertical axis.

Once the Euler–Lagrange equations have been derived, the dynamics of the system are completely described. The

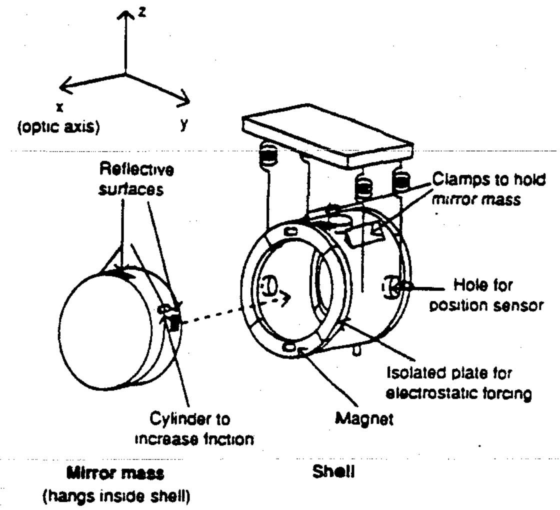


FIG. 2. Nested double pendulum suspension system. The mirror mass has been drawn out of the shell to allow a better view of the hanging method used.

equations of motion can be solved numerically, and the resonant frequencies and the eigenvectors of the rigid body normal modes are easily computed. A computer model was used to predict resonant frequencies and transfer functions between pairs of drive and sensor signals. Table I presents a comparison of the rigid body resonant frequencies predicted by the model and the corresponding measured frequencies. The predicted frequencies agree closely with the measured frequencies. The largest error is in the prediction of the two highest frequencies, which are primarily due to the stretching of the tungsten wire that holds the mirror mass and is probably due to an error in the estimate of the spring constant of the tungsten wire.

TABLE I. Predicted and measured rigid body resonances. The third column briefly describes the coupled motion for each resonance.

Predicted (H2)	Measured (H2)	Coupled motions
0.711	0.690	Translation along the y axis and rotation about the optic axis
0.800	0.812	Translation along the optic axis and rotation about the y axis
0.918	0.920	Rotation about the vertical axis
1.07	1.05	Translation along the optic axis and rotation about the y axis
1.22	1.22	Translation along the optic axis and rotation about the y axis
1.66	1.7	Translation along the y axis and rotation about the optic axis
2.05	1.9	Rotation about the vertical axis
3.23	2.88	Translation along the optic axis and rotation about the y axis
3.57	3.63	Translation along the vertical axis
4.75	4.84	Translation along the y axis and rotation about the optic axis
31.0	27.0	Translation along the vertical axis
31.2	37.0	Translation along the y axis and rotation about the optic axis

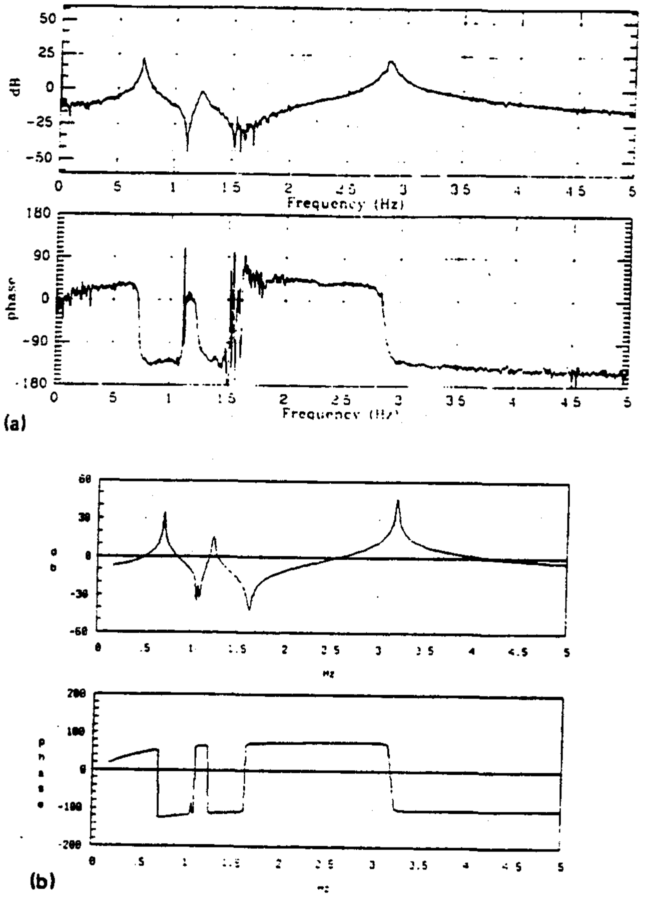


FIG. 3. (a) A typical open loop transfer function for damping. This is the ratio of the voltage applied to the driver for the magnetic actuators to the voltage received from the appropriate position sensor. (b) The same transfer function predicted by the computer model.

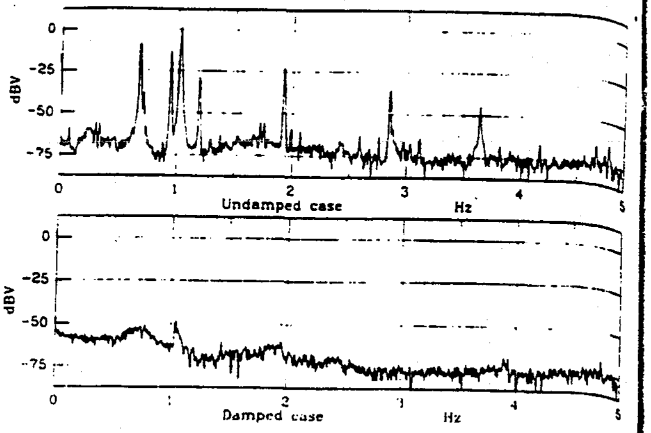


FIG. 4. Power spectra of the output of a position sensor not used in any damping loop. The first spectrum was taken with all loops open, the second with them closed. The sensitivity of this position sensor was 40 V/cm. The suspension is driven by the ambient seismic vibration in the laboratory.

IV. DAMPING

Because the suspension has low losses, the seismically driven motion of the mirror mass at the resonant frequencies is too large to allow a suspended optical cavity to remain in resonance. Therefore the rigid body motion at the resonant frequencies must be artificially damped. A damping system has been implemented in which a sensor detects motion of the shell mass at a position on the shell very close to the point at which the damping force is applied. Because the phase of the transfer function never goes through a phase shift greater than 180°, independent of the number of resonances in the signal, the sensor signal is easily used as an error signal in a feedback loop. Figure 3 shows a typical open loop transfer function. The phase characteristics are similar for all of the loops.

For the double pendulum suspension, damping of the ten lowest resonant frequencies in vacuum has been achieved with four independent damping loops. Two loops are needed to damp the four resonances that couple translation along the optic axis with rotation about the y axis. Two more loops are needed to damp the six other resonances. Figure 4 compares the power spectrum measured with the damping loops open and closed as measured by a position sensor that was not used in any of the damping loops.

By damping the motion of the suspension using forces applied only to the first pendulum stage, one can take advantage of the filtering of electronic noise by the second pendulum stage. The position sensors sense motion of the outer shell relative to the ground, and the magnetic pushers push on the shell from the ground. Pushing on the shell alone damps the motion of both the shell and the mirror mass because the motions of both masses at the resonant frequencies are strongly coupled. With noise from the coil driver electronics at 1.2×10^{-10} N/$\sqrt{\text{Hz}}$, a low closed-loop gain results in a controller induced position noise at the inner mass of 3×10^{-19} m/$\sqrt{\text{Hz}}$ at 100 Hz. The noise due to damping the mirror mass directly would be (f/f_0^2) larger, or 2×10^{-15} m/$\sqrt{\text{Hz}}$ at 100 Hz. Any noise on the magnetic actuators due to fluctuating magnetic fields in the area is also filtered by (f_0/f^2).

V. ISOLATION AND CROSS COUPLING

Isolation transfer functions along the optic axis and in the vertical direction were measured. Cross coupling between vertical motion and optic axis motion and between motion along the y axis and motion along the optic axis were also measured. All measurements were made in vacuum to prevent acoustic excitation of the mirror mass.

To measure the isolation transfer function along the optic axis in vacuum an electromagnetic shaker was attached to the suspension point of the double pendulum via a bellows feedthrough. Accelerometers were placed along the optic axis of the mirror mass, the shell, and the suspension point. The shaker was used to shake the suspension point, and transfer functions between the shell and the suspension point and between the mirror mass and the suspension point were measured. To measure the cross-coupling between motion along the y axis and the optic

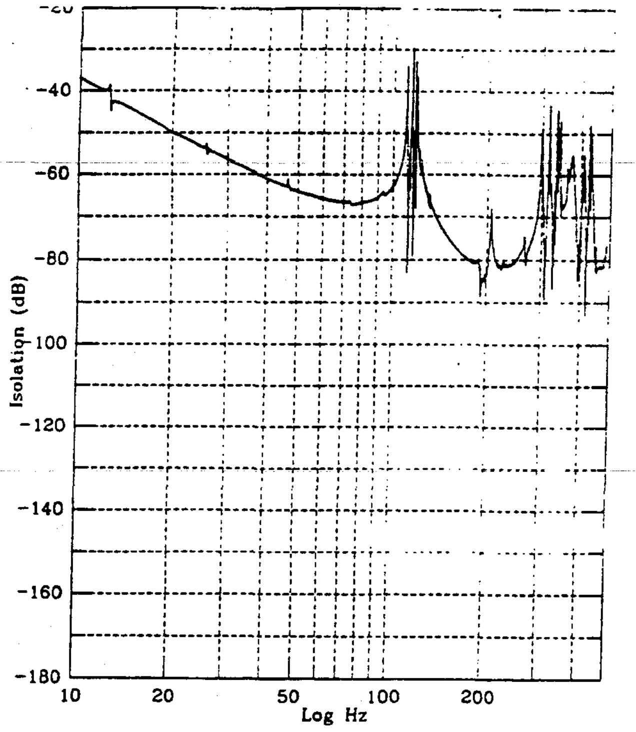


FIG. 5. Optic axis isolation of shell relative to the suspension point.

axis, the same shaker configuration was used but the suspension was turned 90° in the vacuum tank.

The shell transfer function along the optic axis follows the $1/f^2$ behavior expected for a single pendulum up to 113 Hz, where the internal resonances of the four coil springs dominate the transfer function (one could reduce the Q of the internal resonances with a passive magnetic damping scheme). The shell continues to isolate beyond these resonances, but the harmonics of the spring resonances at 339 Hz and the violin modes of the four wires holding the shell mass (at 430 Hz) begin to compromise the isolation (Fig. 5). In addition, the inner mass transfer function follows the $1/f^4$ behavior expected for a double pendulum up to 113 Hz. It continues to isolate as $1/f^2$ with respect to the shell despite the internal spring resonances that compromise the shell's isolation. At 250 Hz an isolation of 160 dB was measured (Fig. 6). The violin mode of the two inner wires is expected at 1200 Hz; however, measurements of the isolation of the inner mass could not be made beyond 250 Hz because the suspension isolation was so effective that the measurements were dominated by electronic noise in the accelerometers.

The vertical isolation of the suspension and the vertical-to-optic-axis cross coupling were also measured. There are two vertical resonances, one at 3.6 Hz due to the coil springs, and one at 26 Hz due to the stretching of the wire holding the mirror mass. The expected $1/f^2$ isolation of the mirror mass before the 26 Hz vertical resonance, and the $1/f^4$ isolation after this resonance was confirmed (Fig. 7).

Figures 8 and 9 show the cross-coupling of y-axis motion to optic axis motion and the cross-coupling of vertical

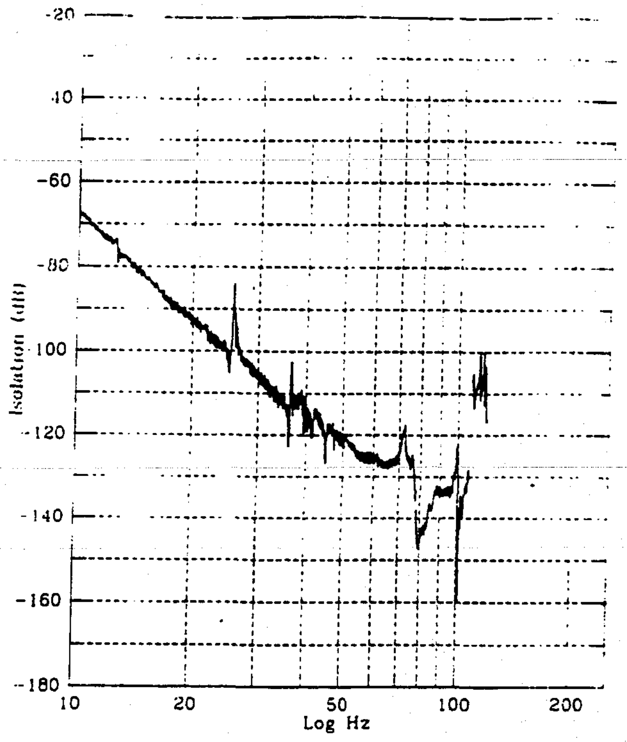


FIG. 6. Optic axis isolation of mirror mass relative to the suspension point.

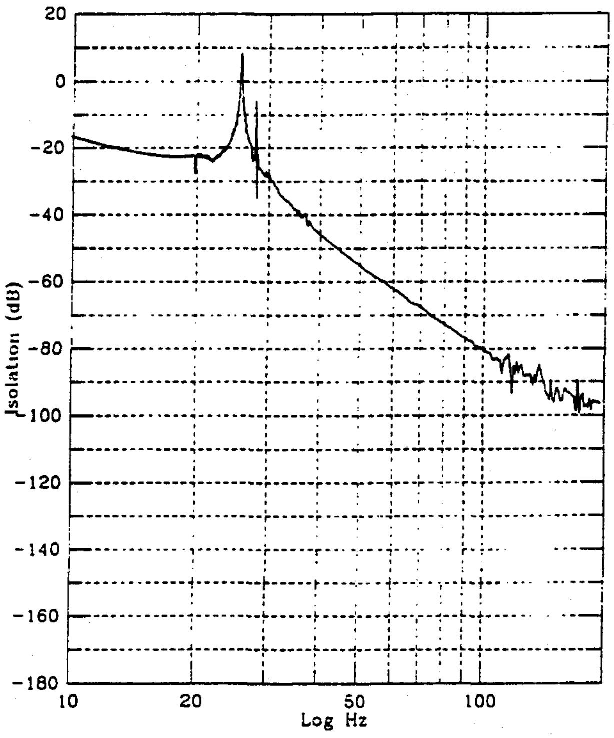


FIG. 7. Vertical isolation of the mirror mass relative to the suspension point.

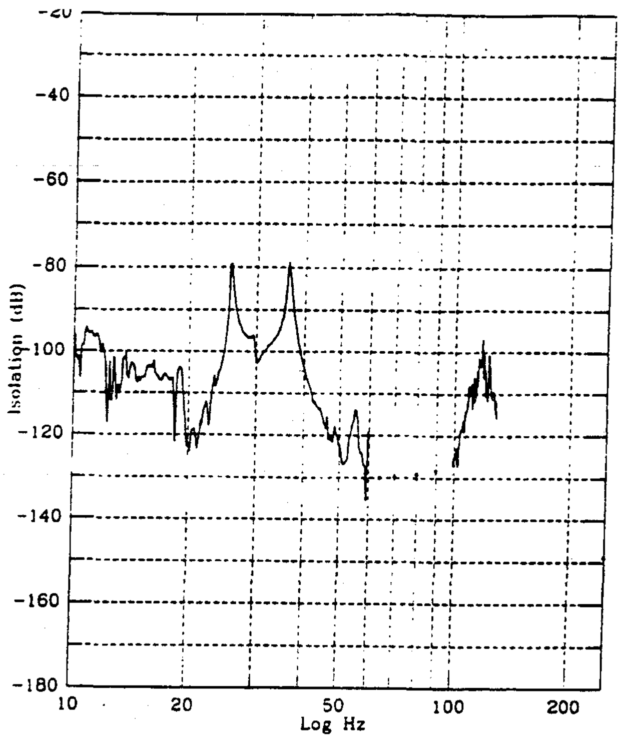


FIG. 8. *Y* axis to optic axis cross coupling.

motion to optic axis motion. Figure 10 shows a composite of Figures 6, 7, 8, and 9 to demonstrate the actual optic axis isolation of the suspension system. It also includes the vertical isolation reduced by 65 dB, showing the contribution of vertical motion to the interferometer noise if the arms are $\frac{2}{3}$ mrad away from level. (Arms that are 4 km long, if level at their midpoints, will depart from level at their ends by about $\frac{2}{3}$ mrad because of the curvature of the earth.) It is apparent that for these tests the isolation of the suspension below 80 Hz is dominated by the vertical to horizontal cross coupling. Above 80 Hz the cross coupling is comparable to the optic axis isolation and does not seriously compromise the isolation.

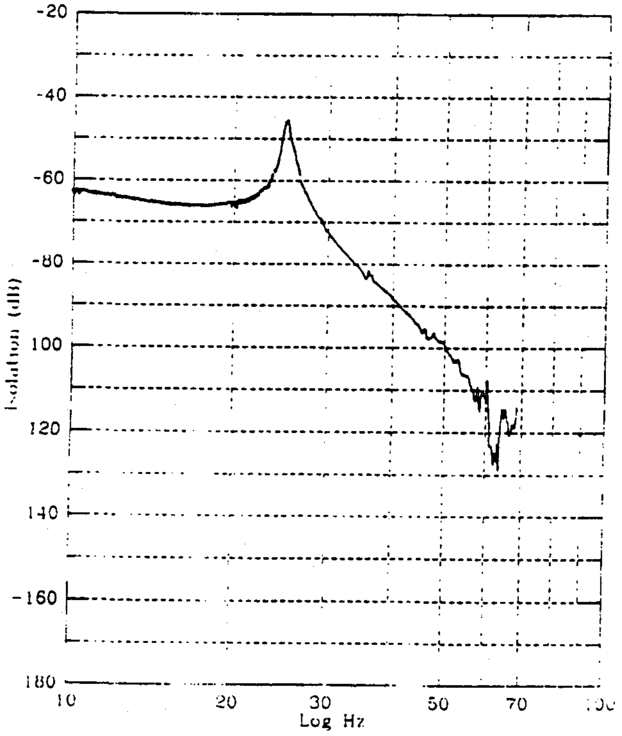


FIG. 9. Vertical axis to optic axis cross coupling.

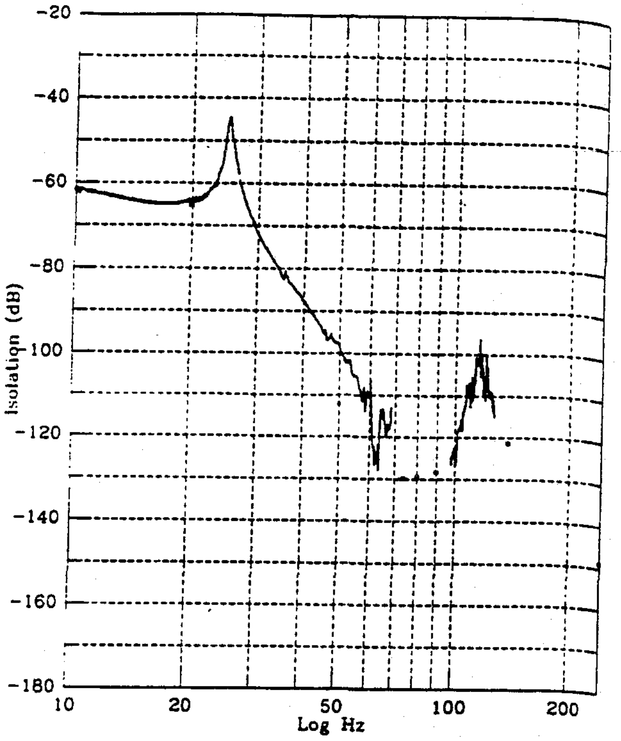


FIG. 10. Composite optic axis isolation. The ground noise is assumed to be isotropic.

An investigation into the cause of the cross coupling indicates that the main contribution to the cross-coupling is misalignment at the takeoff point of the single wire about the mirror mass. For example, if the loop of wire leaves the mirror mass at different points relative to the center of mass on each side the mirror mass will hang at a slight angle. If in addition the wire leaves the mirror mass at a slightly higher point on one side relative to the other, a vertical drive will cause a nonzero torque about the center of mass. This in turn will cause a rotation about the *y* axis. Because the rotation point is by design above the center of mass, this rotation will cause a first-order translation of the center of mass. The actual motions of the mirror mass caused by cross coupling were more complicated than this, involving rotations about the vertical axis and the *y* axis and translations along the optic axis and the *y* axis.

These cross-coupling measurements should be taken to be "worst case" measurements since no special alignment tools were made to minimize misalignment during hanging.

VI. FINE CONTROL

Fine control of the mirror mass is necessary to keep the interferometer fringe locked. This double pendulum suspension was designed so that the fine control could be done with high voltage actuators. By pushing on the mirror mass from the shell one can take advantage of the f^2 isolation of the shell to push from a quiet platform, thus reducing the amount of seismic noise transmitted to the mirror mass through the actuators. By having high voltage plates on both the front and back of the electrically grounded mirror mass, an effective restoring force that is much smaller than any of the pendulum restoring forces can be obtained. This provides a forcer that does not interfere with the suspension isolation in addition to providing a force that is linear with respect to applied voltage.

The effective restoring force for small displacements of the mirror mass is calculated as:

$$\left.\frac{dF}{dx}\right|_{x_0} = \frac{A_{\rm plate}V_{\rm bias}^2}{\pi x_0^3}. \qquad (1)$$

In this prototype suspension, the gap size $x_0 = 0.1$ cm, the bias voltage $V_{\rm bias} = 1.33$ statvolts, is applied to both the front and the back plate, each plate section has an area $A_{\rm plate} = 19$ cm^2 and the mass $m = 10^4$ gm. The effective resonant frequency due to this restoring is $f_0 = 0.165$ Hz. This effective frequency is much lower than any of the pendulum resonant frequencies, thus it has a negligible effect on the dynamics and isolation of the suspension.

With an input noise to the high voltage amplifiers of 2 nV/$\sqrt{\rm Hz}$ and a gain of 50, the position noise at the mirror mass from the amplifier noise is 2×10^{-19} m/$\sqrt{\rm Hz}$ at 100 Hz. This noise will be reduced by the loop gain in the interferometer locking servos.

Implementation of the fine control requires three servoloops, one to control translation of the mirror mass along the optic axis, one to control rotation about the y axis, and one to control rotation about the vertical axis. There must therefore be three stable loops. Not only must each individual loop be stable when closed, but all three loops must be closed at once without losing stability.

The first two of these loops interact because of the way the mirror mass is hung, with a single wire about the center. A force that translates the mirror mass along the optic axis couples to a rotation about the y axis, and a force that rotates the mirror about the y axis couples to a translation along the optic axis. By using normal coordinate transformations[12] the drives for these two coupled loops were optimized so that at frequencies much higher than the resonant frequencies this coupling was minimized. The Bode plots for each particular drive were calculated with the computer model. The two drives were then implemented on the prototype suspension. Each drive consisted of a combination of forces on the shell from the magnetic pushers and forces on the mirror mass from the high voltage pushers.

The rotation about the y axis was measured with an optical lever, and the translation along the optic axis was measured with a capacitance transducer. Small differences in spacing between the front and back plates were compensated for by adjusting $V_{\rm bias}$ on one of the plates.

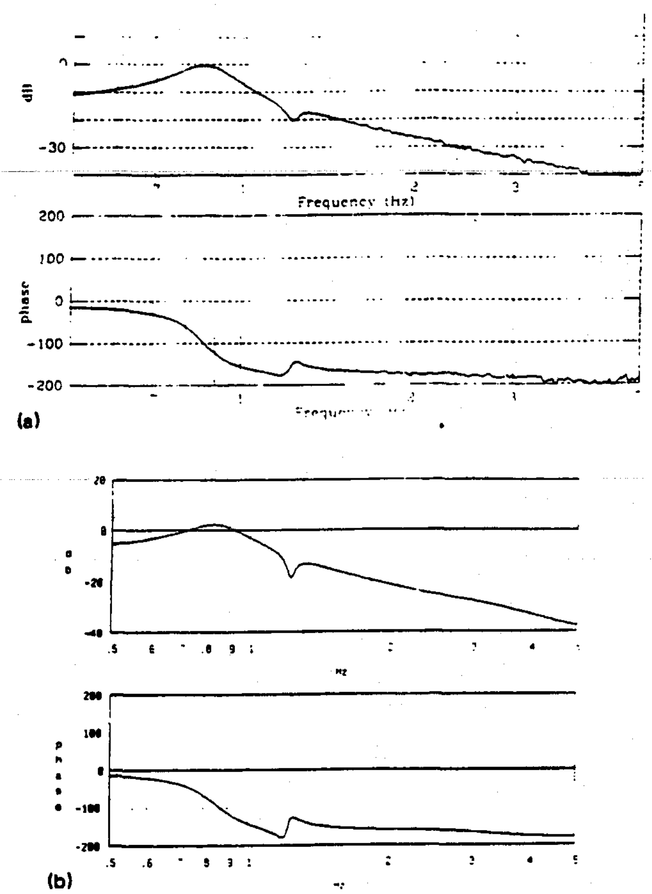


FIG. 11. (a) Measured open loop transfer function for maximum translation drive. This is the ratio of the voltage received from the capacitance transducer to the voltage applied to the drive circuitry. (b) Corresponding transfer function predicted by the computer model.

Figure 11 compares the measured and predicted transfer functions for the drive designed to optimize translation of the center of mass of the mirror mass while minimizing the coupling to rotation about the y axis. Figure 12 compares the measured and predicted transfer function for the drive designed to optimize rotation about the y axis while minimizing the coupling to the translation of the center of mass of the mirror mass. These transfer functions have the proper phase characteristics needed for use in a stable feedback loop.

A servo system with both the translation loop and the y-axis rotation loop closed simultaneously was simulated on the computer. Stable operation is allowed. The point of this simulation was to demonstrate that both loops could be closed simultaneously while retaining stability; all of the detection circuitry was assumed to have an infinite bandwidth and the closed loop gain in this simulation could be chosen arbitrarily high.

VII. ACTIVE ISOLATION

As mentioned in the Introduction, we explored the possibilities of active isolation with this suspension. (For

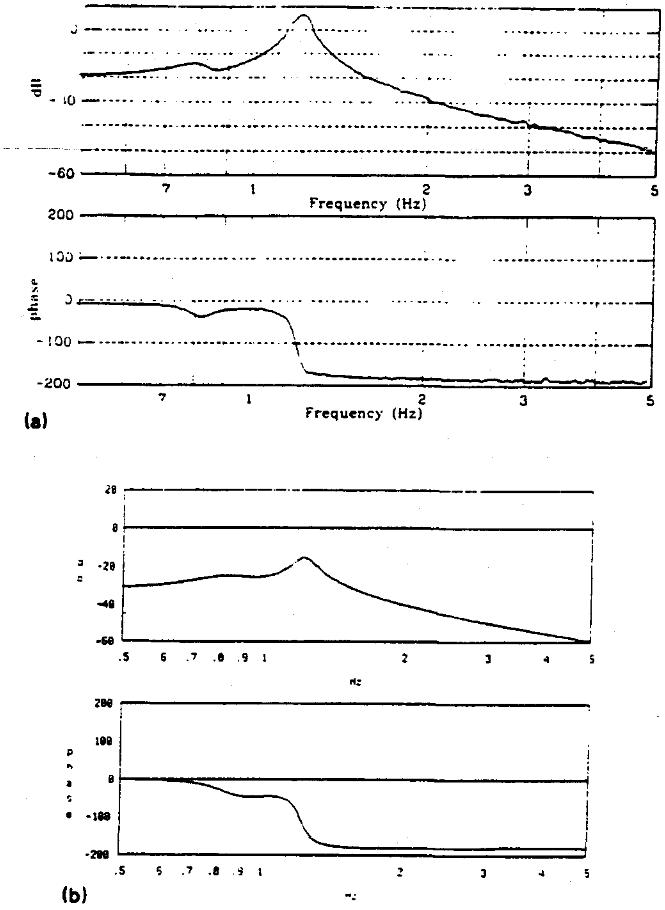


FIG. 12. (a) Measured open loop transfer function for maximum tilt drive. This is the ratio of the voltage received from the optical lever to the voltage applied to the drive circuitry. (b) Corresponding transfer function predicted by the computer model.

previous discussions of the topic, see Refs. 2 and 13.) The basic idea is to apply forces to the shell mass to null an error signal representing the relative displacement of the shell and the mirror mass. When the shell is accurately following the mirror mass (serving as an inertial proof mass), then it is transmitting less vibration than it would in the open-loop state. In the process, the servo loop has reduced the frequency of the gravest pendulation mode of the system.

Several aspects of the design will make it difficult to achieve much improvement in the seismic isolation from this suspension from active techniques. One problem is the coupling between rotation of the shell and translation of the mirror mass. This is of a fundamentally different character than the coupling of the displacements of the centers of mass of the shell and the mirror. The relative displacement due to shell mass displacement vanishes in the limit of zero frequency, but a rotation of the shell causes a relative displacement of the centers of mass at zero frequency. This means that, unless the forces from the upper and lower actuators on the shell were precisely balanced, the low-frequency performance of an active isolation loop will always be limited by the rotations caused by the mismatch. This requirement could be alleviated if the suspension were

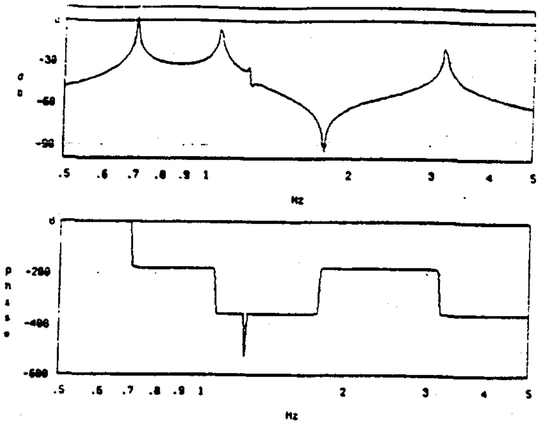


FIG. 13. Predicted transfer function indicating a stable active isolation loop.

redesigned so that the wire for the mirror mass left the shell at the point about which the shell rotates. Figure 13 demonstrates that there is a conditionally stable active isolation loop if one can perfectly balance the driving forces.

A second problem is perhaps more serious. As we have shown, the asymmetry in the isolation of the suspension is great enough that vertical vibration of the mirror is close to dominating the interferometer's response to seismic noise. This means that if we want to use active isolation to reduce the amount of seismic noise, we will need to improve the vertical isolation almost as much as the horizontal isolation. But the fact that the suspension is much stiffer in the vertical direction will substantially complicate implementation of a vertical active isolation loop. In particular, the signal-to-noise ratio for sensing vertical motion is greatly reduced because the relative motion between the shell and the mirror mass is small. The replacement of the relatively stiff tungsten wire holding the mirror mass with a softer vertical isolator would improve the vertical isolation of the system and make it easier to actively isolate in the vertical direction if necessary.

For these reasons, we are not hopeful that this suspension's vibration isolation can be much improved by active isolation techniques although a few design changes could once again make this a candidate for active isolation. Another possibility for active isolation can be found in specially designed pre-filters from which the high-Q pendulum suspensions are hung.[14]

VIII. DISCUSSION

The results of our series of tests show this nested double pendulum to be a credible design for a test mass suspension in a high performance gravitational wave interferometer. The vibration isolation is consistent with the model. The damping servos perform well. The transfer functions for the fine control loops appear to have the required form. Active isolation remains a possibility, but would require some changes in the design.

Some aspects of the performance can only be fully tested when the suspension has been incorporated into a working interferometer. For example. tests with a sensitive interferometer would verify that the isolation we measured at large noise levels is obtained at ambient noise levels, and that there is no unexpected additive noise component.

ACKNOWLEDGMENTS

We would like to thank Rai Weiss for extensive contributions. We would like to thank Nelson Christensen for work on the high voltage amplifiers. We would also like to thank Ron Drever, Bob Spero, and Mike Zucker for useful discussions.

This work was supported by NSF Grant PHY-8803557.

[1] K. S. Thorne, in *300 Years of Gravitation*, edited by S. W. Hawking and W. Israel (Cambridge University Press, Cambridge, 1987), Chap. 9.

[2] N. A. Robertson, R. Drever, I. Kerr, and J. Hough, Phys. E **15**, 1101 (1982).

[3] R. Del Fabbro, A. Di Virgilio, A. Giazotto, H. Kautzky, V. Montelatici, and D. Passuello, Phys Lett. A **132**, 237 (1988)

[4] C. A. Cantley, N. A. Robertson, and J. A. Hough, presented at the 12th International Conference on General Relativity and Gravitation, Boulder, CO, 1989.

[5] D. Shoemaker, R. Schilling, L. Schnupp, W. Winkler, K. Maischberger, and A. Rudiger, Phys. Rev. D **38**, 423 (1988).

[6] P. Linsay, and D Shoemaker, Rev. Sci. Instrum. **53**, 1014 (1982).

[7] R. Drever, in *Gravitational Radiation*, edited by N. Deruelle and T. Piran (North-Holland, Amsterdam, 1983), p. 330.

[8] R. Wolf, Undergraduate Research Thesis, California Institute of Technology, 1987.

[9] R. Del Fabbro, A. Di Virgilio, A. Giazotto, H. Kautzky, V. Montelatici, and D. Passuello, Rev. Sci. Instrum. **59**, 292 (1988).

[10] G. W. McMahon, J. Acoust. Soc. Am. **36**, 85 (1964).

[11] J. R. Hutchinson, J. Appl. Math. **47**, 901 (1980).

[12] H. Goldstein, *Classical Mechanics*, 2nd ed. (Addison-Wesley, Reading, 1980).

[13] P. R. Saulson, Rev Sci Instrum. **55**, 1315 (1984).

[14] R. T. Stebbins, P. L. Bender, J. E. Faller, D. B. Newell, C. C. Speake, presented at 12th International Conference on General Relativity and Gravitation, Boulder, CO (1989).

REF. DD

Fifth Edition

Modern Control Systems

Richard C. Dorf
University of California, Davis

ADDISON-WESLEY PUBLISHING COMPANY

Reading, Massachusetts • Menlo Park, California • New York
Don Mills, Ontario • Wokingham, England • Amsterdam • Bonn
Sydney • Singapore • Tokyo • Madrid • San Juan

This book is in the Addison-Wesley Series in Electrical and Computer Engineering: Control Engineering

Adaptive Control, 09720	Karl J. Åström and Björn Wittenmark
Introduction to Robotics, Second Edition, 09528	John J. Craig
Modern Control Systems, Fifth Edition, 14278	Richard C. Dorf
Digital Control of Dynamic Systems, Second Edition, 11938	Gene F. Franklin, J. David Powell, and Michael L. Workman
Computer Control of Machines and Processes, 10645	John G. Bollinger and Neil A. Duffie
Feedback Control of Dynamic Systems, 11540	Gene F. Franklin, J. David Powell, and Abbas Emami-Naeini
Adaptive Control of Mechanical Manipulators, 10490	John J. Craig
Modern Control System Theory and Application, Second Edition, 07494	Stanley M. Shinners

Sponsoring Editor, Tom Robbins
Production Administrator, Helen Wythe
Copy Editor, Herb Merritt
Text Designer, Patricia O. Williams
Art Consultant, Joseph Vetere
Illustrators, Illustrated Arts
Robert Forget
Cover Designer, Marshall Henrichs

LIBRARY OF CONGRESS
Library of Congress Cataloging-in-Publication Data

Dorf, Richard C.
Modern control systems/Richard C. Dorf.—5th ed.
p. cm.—(Addison-Wesley series in electrical engineering)
Includes index.
ISBN 0-201-14278-3:
1. Feedback control systems. I. Title. II. Series.
TJ216.D67 1989
629.8'3—dc19 88-17656
CIP

Reprinted with corrections November, 1990.

Copyright © 1989, 1986, 1980, 1974, 1967, by Addison-Wesley Publishing Company, Inc.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher. Printed in the United States of America.

FGHIJ-DO-93210

CHAPTER 7

Frequency Response Methods

Preview

We have examined the use of test input signals such as a step and a ramp signal. In this chapter, we will use a steady-state sinusoidal input signal and consider the response of the system as the frequency of the sinusoid is varied. Thus we will look at the response of the system to a changing frequency, ω.

We will examine the response of $G(s)$ when $s = j\omega$ and develop several forms of plotting the complex number for $G(j\omega)$ when ω is varied. These plots provide insight regarding the performance of a system. We are able to develop several performance measures for the frequency response of a system. The measures can be used as system specifications and we can adjust parameters in order to meet the specifications.

We will consider the graphical development of one or more forms for the frequency response plot. We can then proceed to use computer-generated data to readily obtain these plots.

7.1 Introduction

In the preceding chapters the response and performance of a system has been described in terms of the complex frequency variable s, and the location of the poles and zeros on the s-plane. A very practical and important alternative approach to the analysis and design of a system is the *frequency response* method. *The frequency response of a system is defined as the steady-state response of the system to a sinusoidal input signal.* The sinusoid is a unique input signal, and the resulting output signal for a linear system, as well as signals throughout the system, is sinusoidal in the steady state; it differs from the input waveform only in amplitude and phase angle.

One advantage of the frequency response method is the ready availability of sinusoid test signals for various ranges of frequencies and amplitudes. Thus the experimental determination of the frequency response of a system is easily accomplished and is the most reliable and uncomplicated method for the experimental analysis of a system. Often, as we shall find in Section 7.4, the unknown transfer function of a system can be deduced from the experimentally determined frequency response of a system [2,7]. Furthermore, the design of a system in the frequency domain provides the designer with control of the bandwidth of a system and some measure of the response of the system to undesired noise and disturbances.

A second advantage of the frequency response method is that the transfer function describing the sinusoidal steady-state behavior of a system can be obtained by replacing s with $j\omega$ in the system transfer function $T(s)$. The transfer function representing the sinusoidal steady-state behavior of a system is then a function of the complex variable $j\omega$ and is itself a complex function $T(j\omega)$ which possesses a magnitude and phase angle. The magnitude and phase angle of $T(j\omega)$ are readily represented by graphical plots that provide a significant insight for the analysis and design of control systems.

The basic disadvantage of the frequency response method for analysis and design is the indirect link between the frequency and the time domain. Direct correlations between the frequency response and the corresponding transient response characteristics are somewhat tenuous, and in practice the frequency response characteristic is adjusted by using various design criteria which will normally result in a satisfactory transient response.

The Laplace transform pair was given in Section 2.4 and is written as:

$$F(s) = \mathcal{L}\{f(t)\} = \int_0^\infty f(t)e^{-st}\,dt \qquad (7.1)$$

and

$$f(t) = \mathcal{L}^{-1}\{F(s)\} = \frac{1}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty} F(s)e^{st}\,ds, \tag{7.2}$$

where the complex variable $s = \sigma + j\omega$. Similarly, the *Fourier transform* pair is written as

$$F(j\omega) = \mathcal{F}\{f(t)\} = \int_{-\infty}^{\infty} f(t)e^{-j\omega t}\,dt \tag{7.3}$$

and

$$f(t) = \mathcal{F}^{-1}\{\mathcal{F}(j\omega)\} = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(j\omega)e^{j\omega t}\,d\omega. \tag{7.4}$$

The Fourier transform exists for $f(t)$ when

$$\int_{-\infty}^{\infty} |f(t)|\,dt < \infty.$$

The Fourier and Laplace transforms are closely related, as we can see by examining Eqs. (7.1) and (7.3). When the function $f(t)$ is defined only for $t \geq 0$, as is often the case, the lower limits on the integrals are the same. Then, we note that the two equations differ only in the complex variable. Thus, if the Laplace transform of a function $f_1(t)$ is known to be $F_1(s)$, one can obtain the Fourier transform of this same time function $F_1(j\omega)$ by setting $s = j\omega$ in $F_1(s)$.

Again, one might ask, because the Fourier and Laplace transforms are so closely related, why not always use the Laplace transform? Why use the Fourier transform at all? The Laplace transform permits us to investigate the s-plane location of the poles and zeros of a transfer $T(s)$ as in Chapter 6. However, the frequency response method allows us to consider the transfer function $T(j\omega)$ and concern ourselves with the amplitude and phase characteristics of the system. This ability to investigate and represent the character of a system by amplitude and phase equations and curves is an advantage for the analysis and design of control systems.

If we consider the frequency response of the closed-loop system, we might have an input $r(t)$ that has a Fourier transform, in the frequency domain, as follows:

$$R(j\omega) = \int_{-\infty}^{\infty} r(t)e^{-j\omega t}\,dt. \tag{7.5}$$

Then the output frequency response of a single-loop control system can be obtained by substituting $s = j\omega$ in the closed-loop system relationship, $C(s) = T(s)R(s)$, so that we have

$$C(j\omega) = T(j\omega)R(j\omega) = \frac{G(j\omega)}{1 + G(j\omega)H(j\omega)}R(j\omega). \tag{7.6}$$

Utilizing the inverse Fourier transform, the output transient response would be

$$c(t) = \mathcal{F}^{-1}\{C(j\omega)\} = \frac{1}{2\pi}\int_{-\infty}^{\infty} C(j\omega)e^{j\omega t}\,d\omega. \tag{7.7}$$

However, it is usually quite difficult to evaluate this inverse transform integral for any but the simplest systems, and a graphical integration may be used. Alternatively, as we will note in succeeding sections, several measures of the transient response can be related to the frequency characteristics and utilized for design purposes.

7.2 Frequency Response Plots

The transfer function of a system $G(s)$ can be described in the frequency domain by the relation

$$G(j\omega) = G(s)|_{s=j\omega} = R(j\omega) + jX(j\omega), \tag{7.8}$$

where

$$R(j\omega) = Re(G(j\omega))$$

and

$$X(j\omega) = \mathrm{Im}(G(j\omega)).$$

Alternatively, the transfer function can be represented by a magnitude $|G\,(j\omega)|$ and a phase $\phi(j\omega)$ as

$$G(j\omega) = |G(j\omega)|e^{j\phi(j\omega)} = |G(\omega)|\underline{/\phi(\omega)}, \tag{7.9}$$

where

$$\phi(\omega) = \tan^{-1} X(\omega)/R(\omega)$$

and

$$|G(\omega)|^2 = (R(\omega))^2 + (X(\omega))^2.$$

The graphical representation of the frequency response of the system $G(j\omega)$ can utilize either Eq. (7.8) or Eq. (7.9). The *polar plot* representation of the frequency response is obtained by using Eq. (7.8). The coordinates of the polar plot are the real and imaginary parts of $G(j\omega)$ as shown in Fig. 7.1. An example of a polar plot will illustrate this approach.

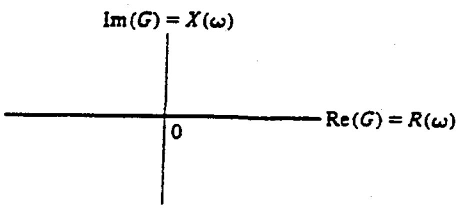


Figure 7.1. The polar plane.

■ Example 7.1

A simple *RC* filter is shown in Fig. 7.2. The transfer function of this filter is

$$G(s) = \frac{V_2(s)}{V_1(s)} = \frac{1}{RCs + 1}, \tag{7.10}$$

and the sinuoidal steady-state transfer function is

$$G(j\omega) = \frac{1}{j\omega(RC) + 1} = \frac{1}{j(\omega/\omega_1) + 1}, \tag{7.11}$$

where

$$\omega_1 = 1/RC.$$

Then the polar plot is obtained from the relation

$$\begin{aligned} G(j\omega) &= R(\omega) + jX(\omega) \\ &= \frac{1 - j(\omega/\omega_1)}{(\omega/\omega_1)^2 + 1} \\ &= \frac{1}{1 + (\omega/\omega_1)^2} - \frac{j(\omega/\omega_1)}{1 + (\omega/\omega_1)^2}. \end{aligned} \tag{7.12}$$

The locus of the real and imaginary parts is shown in Fig. 7.3 and is easily shown to be a circle with the center at (½, 0). When $\omega = \omega_1$, the real and imaginary parts are equal, and the angle $\phi(\omega) = 45°$. The polar plot can also be readily obtained from Eq. (7.9) as

$$G(j\omega) = |G(\omega)| \underline{/\phi(\omega)}, \tag{7.13}$$

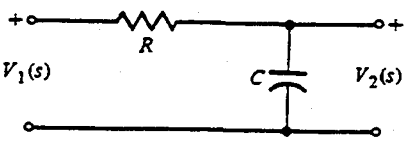


Figure 7.2. An *RC* filter.

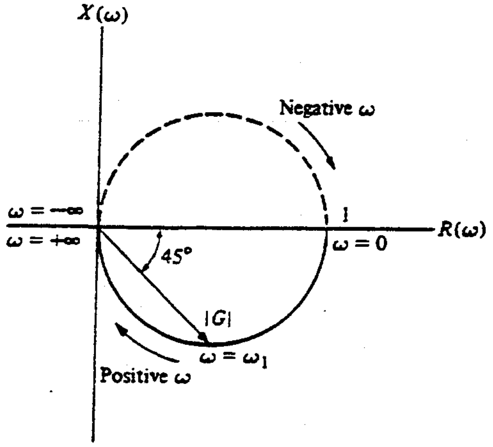


Figure 7.3. Polar plot for *RC* filter.

where

$$|G(\omega)| = \frac{1}{(1 + (\omega/\omega_1)^2)^{1/2}} \quad \text{and} \quad \phi(\omega) = -\tan^{-1}(\omega/\omega_1).$$

Clearly, when $\omega = \omega_1$, magnitude is $|G(\omega_1)| = 1/\sqrt{2}$ and phase $\phi(\omega_1) = -45°$. Also, when ω approaches $+\infty$, we have $|G(\omega)| \to 0$ and $\phi(\omega) = -90°$. Similarly, when $\omega = 0$, we have $|G(\omega)| = 1$ and $\phi(\omega) = 0$.

■ Example 7.2

The polar plot of a transfer function will be useful for investigating system stability and will be utilized in Chapter 8. Therefore, it is worthwhile to complete another example at this point. Consider a transfer function

$$|G(s)|_{s=j\omega} = G(j\omega) = \frac{K}{j\omega(j\omega\tau + 1)} = \frac{K}{j\omega - \omega^2\tau}. \tag{7.14}$$

Then the magnitude and phase angle are written as

$$|G(\omega)| = \frac{K}{(\omega^2 + \omega^4\tau^2)^{1/2}} \tag{7.15}$$

and

$$\phi(\omega) = -\tan^{-1}\left(\frac{1}{-\omega\tau}\right).$$

The phase angle and the magnitude are readily calculated at the frequencies $\omega = 0$, $\omega = 1/\tau$, and $\omega = +\infty$. The values of $|G(\omega)|$ and $\phi(\omega)$ are given in Table 7.1, and the polar plot of $G(j\omega)$ is shown in Fig. 7.4.

There are several possibilities for coordinates of a graph portraying the fre-

Table 7.1

ω	0	$\frac{1}{2}\tau$	$1/\tau$	∞
$\lvert G(\omega)\rvert$	∞	$4K\tau/\sqrt{5}$	$K\tau/\sqrt{2}$	0
$\phi(\omega)$	$-90°$	$-117°$	$-135°$	$-180°$

quency response of a system. As we have seen, one may choose to utilize a polar plot to represent the frequency response (Eq. 7.8) of a system. However, the limitations of polar plots are readily apparent. The addition of poles or zeros to an existing system requires the recalculation of the frequency response as outlined in Examples 7.1 and 7.2. (See Table 7.1.) Furthermore, the calculation of the frequency response in this manner is tedious and does not indicate the effect of the individual poles or zeros.

Therefore, the introduction of *logarithmic plots*, often called *Bode plots*, simplifies the determination of the graphical portrayal of the frequency response. The logarithmic plots are called Bode plots in honor of H. W. Bode, who used them extensively in his studies of feedback amplifiers [4,5]. The transfer function in the frequency domain is

$$G(j\omega) = |G(\omega)|e^{j\phi(\omega)}. \tag{7.16}$$

The natural logarithm of Eq. (7.16) is

$$\ln G(j\omega) = \ln |G(\omega)| + j\phi(\omega), \tag{7.17}$$

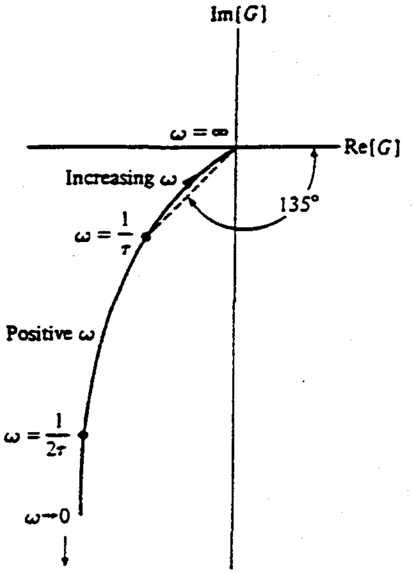


Figure 7.4. Polar plot for $G(j\omega) = K/j\omega(j\omega\tau + 1)$.

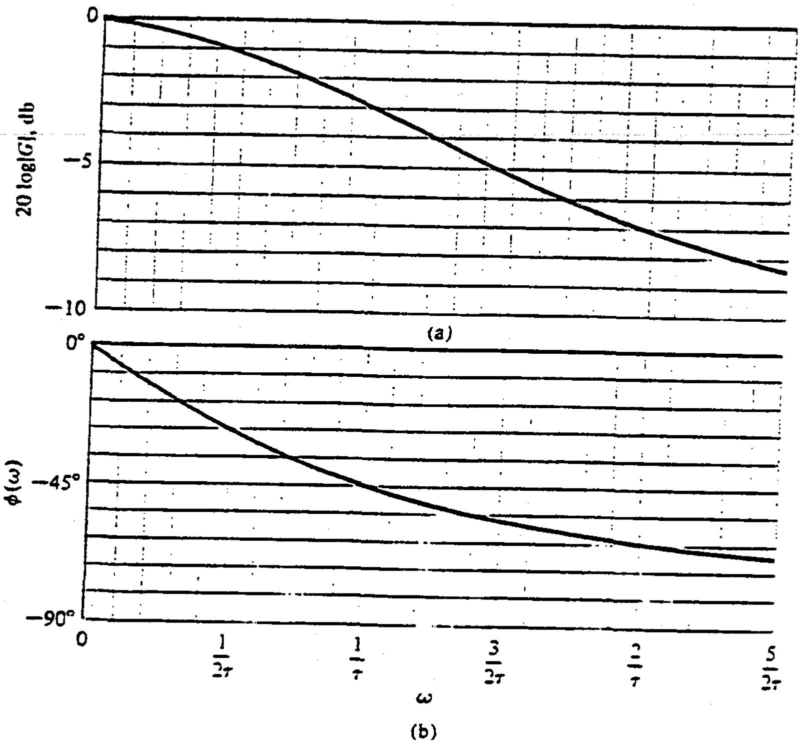


Figure 7.5. Bode diagram of $G(j\omega) = 1/(j\omega\tau + 1)$.

where $\ln |G|$ is the magnitude in nepers. The logarithm of the magnitude is normally expressed in terms of the logarithm to the base 10, so that we use

$$\text{Logarithmic gain} = 20 \log_{10} |G(\omega)|,$$

where the units are decibels (db). A decibel conversion table is given in Appendix E. The logarithmic gain in db and the angle $\phi(\omega)$ can be plotted versus the frequency ω by utilizing several different arrangements. For a Bode diagram, the plot of logarithmic gain in db versus ω is normally plotted on one set of axes and the phase $\phi(\omega)$ versus ω on another set of axes as shown on Fig. 7.5. For example, the Bode diagram of the transfer function of Example 7.1 can be readily obtained, as we will find in the following example.

■ Example 7.3

The transfer function of Example 7.1 is

$$G(j\omega) = \frac{1}{j\omega(RC) + 1} = \frac{1}{j\omega\tau + 1}, \tag{7.18}$$

where

$$\tau = RC,$$

the time constant of the network. The logarithmic gain is

$$20 \log |G| = 20 \log \left(\frac{1}{1 + (\omega\tau)^2}\right)^{1/2} = -10 \log (1 + (\omega\tau)^2). \tag{7.19}$$

For small frequencies, that is $\omega \ll 1/\tau$, the logarithmic gain is

$$20 \log |G| = -10 \log (1) = 0 \text{ db}, \qquad \omega \ll 1/\tau. \tag{7.20}$$

For large frequencies, that is $\omega \gg 1/\tau$, the logarithmic gain is

$$20 \log |G| = -20 \log \omega\tau \qquad \omega \gg 1/\tau, \tag{7.21}$$

and at $\omega = 1/\tau$, we have

$$20 \log |G| = -10 \log 2 = -3.01 \text{ db}.$$

The magnitude plot for this network is shown in Fig. 7.5(a). The phase angle of this network is

$$\phi(j\omega) = -\tan^{-1} \omega\tau. \tag{7.22}$$

The phase plot is shown in Fig. 7.5(b). The frequency $\omega = 1/\tau$ is often called the *break frequency* or *corner frequency.*

Examining the Bode diagram of Fig. 7.5, we find that a linear scale of frequency is not the most convenient or judicious choice and we should consider the use of a logarithmic scale of frequency. The convenience of a logarithmic scale of frequency can be seen by considering Eq. (7.21) for large frequencies $\omega \gg 1/\tau$, as follows:

$$20 \log |G| = -20 \log \omega\tau = -20 \log \tau - 20 \log \omega. \tag{7.23}$$

Then, on a set of axes where the horizontal axis is log ω, the asymptotic curve for $\omega \gg 1/\tau$ is a straight line, as shown in Fig. 7.6. The slope of the straight line can be ascertained from Eq. (7.21). An interval of two frequencies with a ratio equal

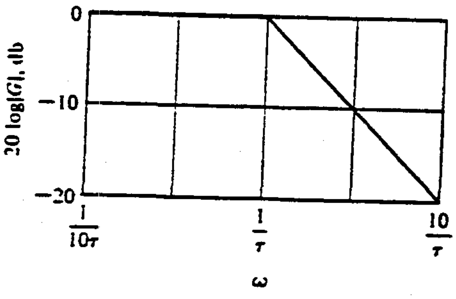


Figure 7.6. Asymptotic curve for $(j\omega\tau + 1)^{-1}$.

to ten is called a decade, so that the range of frequencies from ω_1 to ω_2, where $\omega_2 = 10\omega_1$, is called a decade. Then, the difference between the logarithmic gains, for $\omega \gg 1/\tau$, over a decade of frequency is

$$\begin{aligned}
20 \log |G(\omega_1)| - 20 \log |G(\omega_2)| &= -20 \log \omega_1\tau - (-20 \log \omega_2\tau) \\
&= -20 \log \frac{\omega_1\tau}{\omega_2\tau} \qquad (7.24) \\
&= -20 \log (1/10) = +20 \text{ db}.
\end{aligned}$$

That is, the slope of the asymptotic line for this first-order transfer function is -20 db/decade, and the slope is shown for this transfer function in Fig. 7.6. Instead of using a horizontal axis of $\log \omega$ and linear rectangular coordinates, it is simpler to use semilog paper with a linear rectangular coordinate for db and a logarithmic coordinate for ω. Alternatively, one could use a logarithmic coordinate for the magnitude as well as for frequency and avoid the necessity of calculating the logarithm of the magnitude.

The frequency interval $\omega_2 = 2\omega_1$ is often used and is called an octave of frequencies. The difference between the logarithmic gains for $\omega \gg 1/\tau$, for an octave, is

$$\begin{aligned}
20 \log |G(\omega_1)| - 20 \log |G(\omega_2)| &= -20 \log \frac{\omega_1\tau}{\omega_2\tau} \\
&= -20 \log (1/2) = 6.02 \text{ db}.
\end{aligned} \qquad (7.25)$$

Therefore the slope of the asymptotic line is -6 db/octave or -20 db/decade.

The primary advantage of the logarithmic plot is the conversion of multiplicative factors such as $(j\omega\tau + 1)$ into additive factors $20 \log (j\omega\tau + 1)$ by virtue of the definition of logarithmic gain. This can be readily ascertained by considering a generalized transfer function as

$$G(j\omega) = \frac{K_b \prod_{i=1}^{Q} (1 + j\omega\tau_i)}{(j\omega)^N \prod_{m=1}^{M} (1 + j\omega\tau_m) \prod_{k=1}^{R} (1 + (2\zeta_k/\omega_{n_k})j\omega + (j\omega/\omega_{n_k})^2)}. \qquad (7.26)$$

This transfer function includes Q zeros, N poles at the origin, M poles on the real axis, and R pairs of complex conjugate poles. Clearly, obtaining the polar plot of such a function would be a formidable task indeed. However, the logarithmic magnitude of $G(j\omega)$ is

$$\begin{aligned}
20 \log |G(\omega)| = 20 \log K_b &+ 20 \sum_{i=1}^{Q} \log |1 + j\omega\tau_i| \\
&- 20 \log |(j\omega)^N| - 20 \sum_{m=1}^{M} \log |1 + j\omega\tau_m| \qquad (7.27) \\
&- 20 \sum_{k=1}^{R} \log \left| 1 + \left(\frac{2\zeta_k}{\omega_{n_k}}\right) j\omega + \left(\frac{j\omega}{\omega_{n_k}}\right)^2 \right|,
\end{aligned}$$

and the Bode diagram can be obtained by adding the plot due to each individual factor. Furthermore, the separate phase angle plot is obtained as

$$\phi(\omega) = +\sum_{i=1}^{Q} \tan^{-1} \omega\tau_i - N(90°) - \sum_{m=1}^{M} \tan^{-1} \omega\tau_m - \sum_{k=1}^{R} \tan^{-1}\left(\frac{2\zeta_k\omega_{n_k}\omega}{\omega_{n_k}^2 - \omega^2}\right), \tag{7.28}$$

which is simply the summation of the phase angles due to each individual factor of the transfer function.

Therefore, the four different kinds of factors that may occur in a transfer function are

1. constant gain K_b,
2. poles (or zeros) at the origin $(j\omega)$,
3. poles or zeros on the real axis $(j\omega\tau + 1)$,
4. complex conjugate poles (or zeros) $[1 + (2\zeta/\omega_n)j\omega + (j\omega/\omega_n)^2]$.

We can determine the logarithmic magnitude plot and phase angle for these four factors and then utilize them to obtain a Bode diagram for any general form of a transfer function. Typically, the curves for each factor are obtained and then added together graphically to obtain the curves for the complete transfer function. Furthermore, this procedure can be simplified by using the asymptotic approximations to these curves and obtaining the actual curves only at specific important frequencies.

Constant Gain K_b. The logarithmic gain is

$$20 \log K_b = \text{constant in db},$$

and the phase angle is zero. The gain curve is simply a horizontal line on the Bode diagram.

Poles (or Zeros) at the Origin $(j\omega)$. A pole at the origin has a logarithmic magnitude

$$20 \log \left|\frac{1}{j\omega}\right| = -20 \log \omega \text{ db} \tag{7.29}$$

and a phase angle $\phi(\omega) = -90°$. The slope of the magnitude curve is -20 db/decade for a pole. Similarly for a multiple pole at the origin, we have

$$20 \log \left|\frac{1}{(j\omega)^N}\right| = -20\, N \log \omega, \tag{7.30}$$

and the phase is $\phi(\omega) = -90°N$. In this case the slope due to the multiple pole is

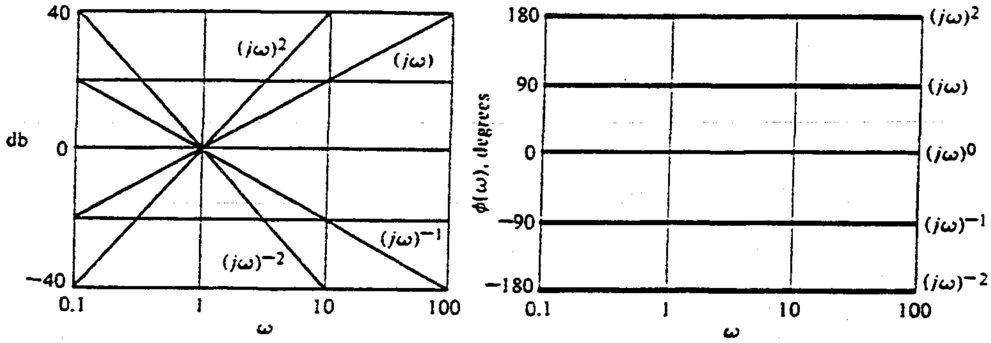


Figure 7.7. Bode diagram for $(j\omega)^{\pm N}$.

$-20N$ db/decade. For a zero at the origin, we have a logarithmic magnitude

$$20 \log |j\omega| = +20 \log \omega, \tag{7.31}$$

where the slope is +20 db/decade and the phase angle is +90°. The Bode diagram of the magnitude and phase angle of $(j\omega)^{\pm N}$ is shown in Fig. 7.7 for $N = 1$ and $N = 2$.

Poles or Zeros on the Real Axis. The pole factor $(1 + j\omega\tau)^{-1}$ has been considered previously and we found that

$$20 \log \left|\frac{1}{1 + j\omega\tau}\right| = -10 \log (1 + \omega^2\tau^2). \tag{7.32}$$

The asymptotic curve for $\omega \ll 1/\tau$ is $20 \log 1 = 0$ db, and the asymptotic curve for $\omega \gg 1/\tau$ is $-20 \log \omega\tau$ which has a slope of −20 db/decade. The intersection of the two asymptotes occurs when

$$20 \log 1 = 0 \text{ db} = -20 \log \omega\tau$$

or when $\omega = 1/\tau$, the *break frequency.* The actual logarithmic gain when $\omega = 1/\tau$ is −3 db for this factor. The phase angle is $\phi(\omega) = -\tan^{-1} \omega\tau$ for the denominator factor. The Bode diagram of a pole factor $(1 + j\omega\tau)^{-1}$ is shown in Fig. 7.8.

The Bode diagram of a zero factor $(1 + j\omega\tau)$ is obtained in the same manner as that of the pole. However, the slope is positive at +20 db/decade, and the phase angle is $\phi(\omega) = +\tan^{-1} \omega\tau$.

A linear approximation to the phase angle curve can be obtained as shown in Fig. 7.8. This linear approximation, which passes through the correct phase at the break frequency, is within 6° of the actual phase curve for all frequencies. This approximation will provide a useful means for readily determining the form of the phase angle curves of a transfer function $G(s)$. However, often the accurate phase angle curves are required and the actual phase curve for the first-order factor must be drawn. Therefore it is often worthwhile to prepare a cardboard (or plastic) template which can be utilized repeatedly to draw the phase curves for

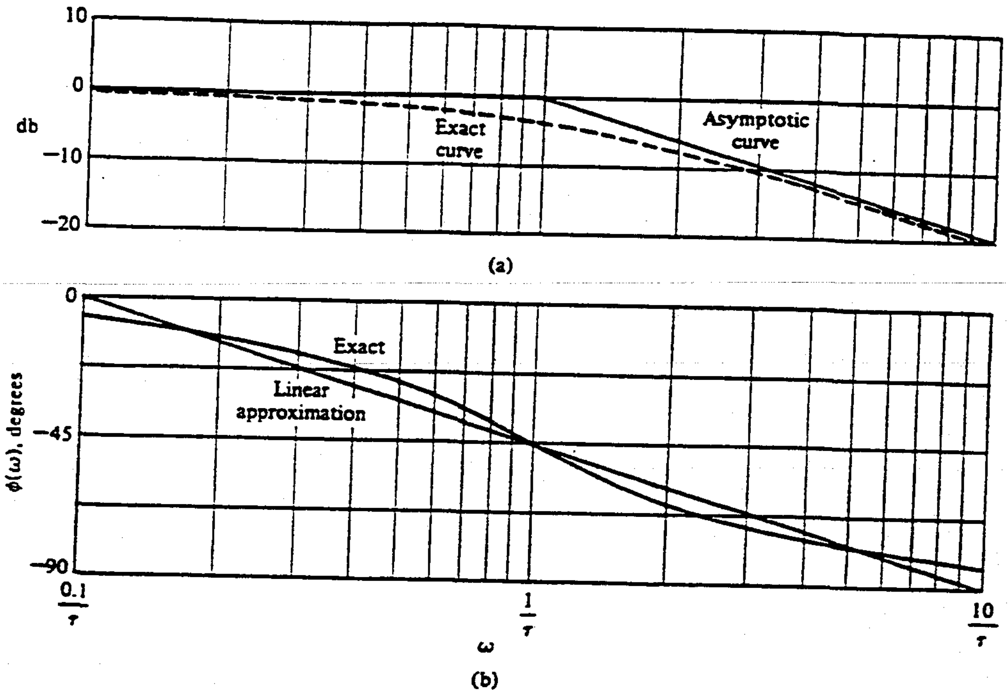


Figure 7.8. Bode diagram of $(1 + j\omega\tau)^{-1}$.

the individual factors. The exact values of the frequency response for the pole $(1 + j\omega\tau)^{-1}$ as well as the values obtained by using the approximation for comparison are given in Table 7.2.

Complex Conjugate Poles or Zeros $[1 + (2\zeta/\omega_n)j\omega + (j\omega/\omega_n)^2]$. The quadratic factor for a pair of complex conjugate poles can be written in normalized form as

$$[1 + j2\zeta u - u^2]^{-1}, \tag{7.33}$$

Table 7.2

$\omega\tau$	0.10	0.50	0.76	1	1.31	2	5	10
20 log $\|(1 + j\omega\tau)^{-1}\|$, db	−0.04	−1.0	−2.0	−3.0	−4.3	−7.0	−14.2	−20.04
Asymptotic approximation, db	0	0	0	0	−2.3	−6.0	−14.0	−20.0
$\phi(\omega)$, degrees	−5.7	−26.6	−37.4	−45.0	−52.7	−63.4	−78.7	−84.3
Linear approximation, degrees	0	−31.5	−39.5	−45.0	−50.3	−58.5	−76.5	−90.0

where $u = \omega/\omega_n$. Then, the logarithmic magnitude is

$$20 \log |G(\omega)| = -10 \log ((1 - u^2)^2 + 4\zeta^2 u^2), \tag{7.34}$$

and the phase angle is

$$\phi(\omega) = -\tan^{-1}\left(\frac{2\zeta u}{1 - u^2}\right). \tag{7.35}$$

When $u \ll 1$, the magnitude is

$$\text{db} = -10 \log 1 = 0 \text{ db},$$

and the phase angle approaches 0°. When $u \gg 1$, the logarithmic magnitude approaches

$$\text{db} = -10 \log u^4 = -40 \log u,$$

which results in a curve with a slope of -40 db/decade. The phase angle, when $u \gg 1$, approaches $-180°$. The magnitude asymptotes meet at the 0-db line when $u = \omega/\omega_n = 1$. However, the difference between the actual magnitude curve and the asymptotic approximation is a function of the damping ratio and must be accounted for when $\zeta < 0.707$. The Bode diagram of a quadratic factor due to a pair of complex conjugate poles is shown in Fig. 7.9. The maximum value of the frequency response, M_{p_ω}, occurs at the *resonant frequency* ω_r. When the damping ratio approaches zero, then ω_r approaches ω_n, the natural frequency. The resonant frequency is determined by taking the derivative of the magnitude of Eq. (7.33) with respect to the normalized frequency, u, and setting it equal to zero. The resonant frequency is represented by the relation

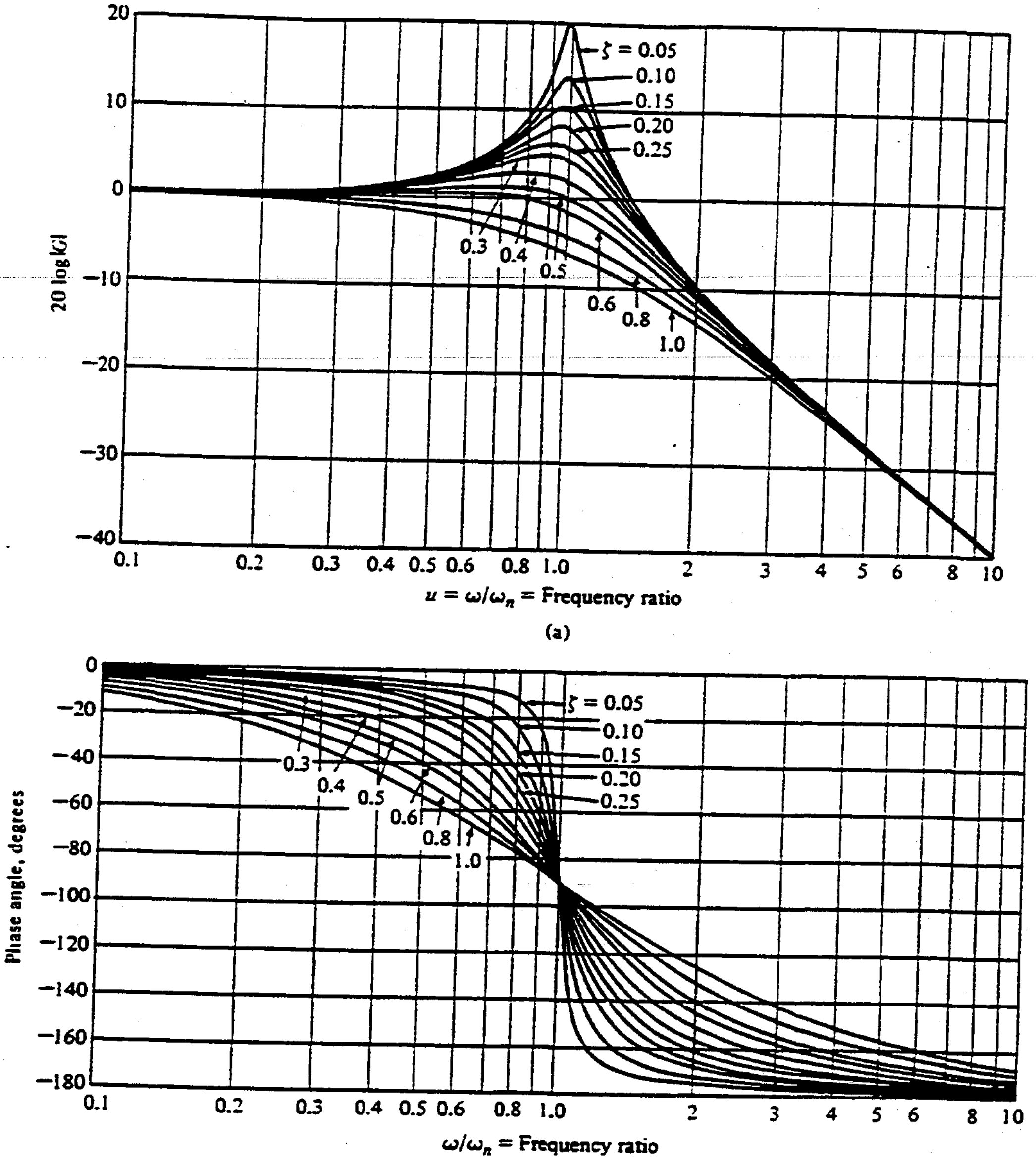
$$\omega_r = \omega_n\sqrt{1 - 2\zeta^2}, \qquad \zeta < 0.707, \tag{7.36}$$

and the maximum value of the magnitude $|G(\omega)|$ is

$$M_{p_\omega} = |G(\omega_r)| = (2\zeta\sqrt{1 - \zeta^2})^{-1}, \qquad \zeta < 0.707, \tag{7.37}$$

for a pair of complex poles. The maximum value of the frequency response M_{p_ω}, and the resonant frequency ω_r are shown as a function of the damping ratio ζ for a pair of complex poles in Fig. 7.10. Assuming the dominance of a pair of complex conjugate closed-loop poles, we find that these curves are useful for estimating the damping ratio of a system from an experimentally determined frequency response.

The frequency response curves can be evaluated graphically on the s-plane by determining the vector lengths and angles at various frequencies ω along the



(a)

(b)

Figure 7.9. Bode diagram of $G(j\omega) = [1 + (2\zeta/\omega_n)j\omega + (j\omega/\omega_n)^2]^{-1}$.

($s = +j\omega$)-axis. For example, considering the second-order factor with complex conjugate poles, we have

$$G(s) = \frac{1}{(s/\omega_n)^2 + 2\zeta s/\omega_n + 1} = \frac{\omega_n^2}{s^2 + 2\zeta\omega_n s + \omega_n^2}. \qquad (7.38)$$

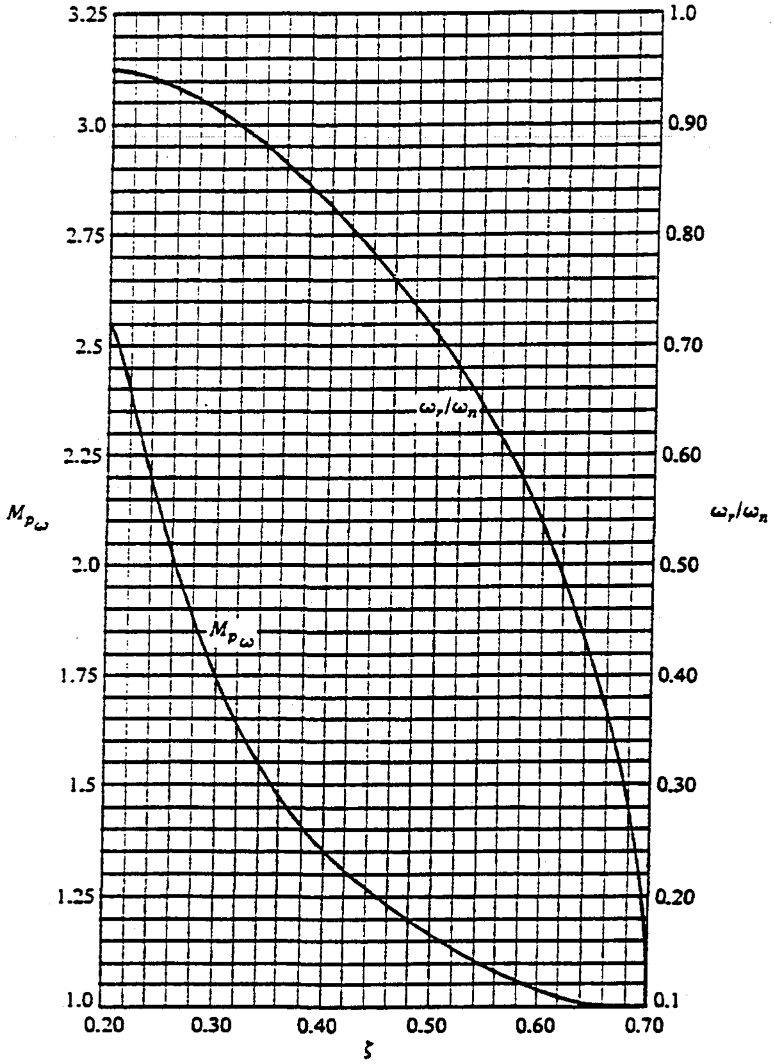
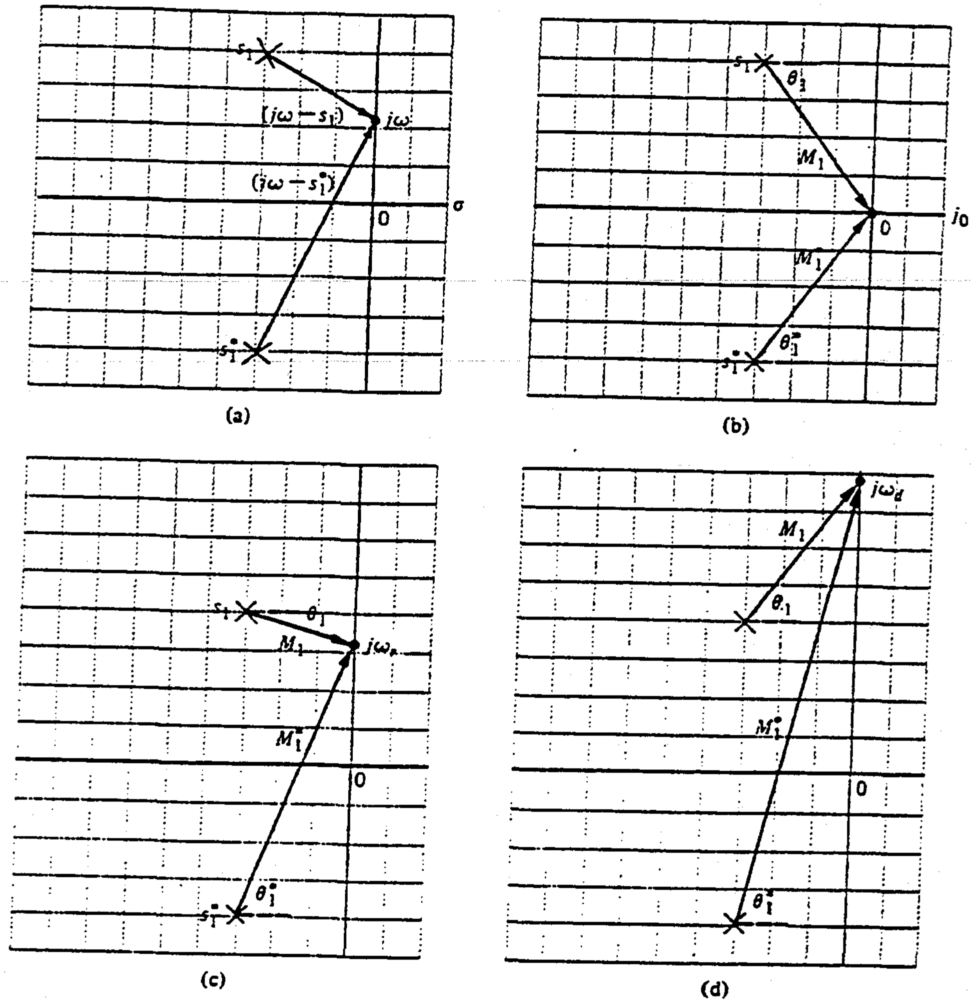


Figure 7.10. The maximum of the frequency response, M_{p_ω}, and the resonant frequency, ω_r, versus ζ for a pair of complex conjugate poles.

The poles lie on a circle of radius ω_n and are shown for a particular ζ in Fig. 7.11(a). The transfer function evaluated for real frequency $s = j\omega$ is written as

$$G(j\omega) = \left.\frac{\omega_n^2}{(s - s_1)(s - s_1^*)}\right|_{s=j\omega} = \frac{\omega_n^2}{(j\omega - s_1)(j\omega - s_1^*)}, \tag{7.39}$$

where s_1 and s_1^* are the complex conjugate poles. The vectors $(j\omega - s_1)$ and $(j\omega - s_1^*)$ are the vectors from the poles to the frequency $j\omega$, as shown in Fig. 7.11(a).



(a)

(b)

(c)

(d)

Figure 7.11. Vector evaluation of the frequency response.

Then the magnitude and phase may be evaluated for various specific frequencies. The magnitude is

$$|G(\omega)| = \frac{\omega_n^2}{|j\omega - s_1|\,|j\omega - s_1^*|}, \tag{7.40}$$

and the phase is

$$\phi(\omega) = -\underline{/(j\omega - s_1)} - \underline{/(j\omega - s_1^*)}.$$

The magnitude and phase may be evaluated for three specific frequencies:

$$\omega = 0, \quad \omega = \omega_r, \quad \omega = \omega_d,$$

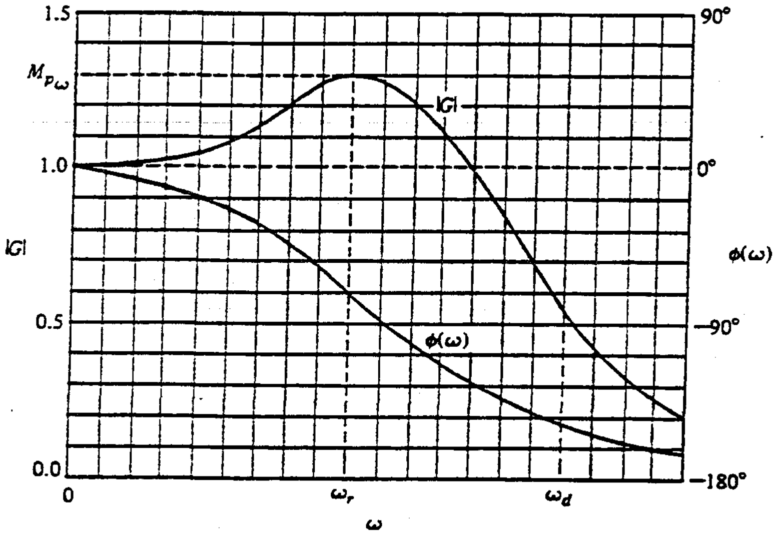


Figure 7.12. Bode diagram for complex conjugate poles.

as shown in Figs. 7.11(b), 7.11(c), and 7.11(d), respectively. The magnitude and phase corresponding to these frequencies are shown in Fig. 7.12.

■ Example 7.4

As an example of the determination of the frequency response using the pole-zero diagram and the vectors to $j\omega$, consider the twin-T network shown in Fig. 7.13 [13]. The transfer function of this network is

$$G(s) = \frac{E_{\text{out}}(s)}{E_{\text{in}}(s)} = \frac{(s\tau)^2 + 1}{(s\tau)^2 + 4s\tau + 1}, \tag{7.41}$$

where $\tau = RC$. The zeros are at $\pm j1$ and the poles are at $-2 \pm \sqrt{3}$ in the $s\tau$-plane as shown in Fig. 7.14(a). Clearly, at $\omega = 0$, we have $|G| = 1$ and $\phi(\omega) = 0°$. At $\omega = 1/\tau$, $|G| = 0$ and the phase angle of the vector from the zero at $s\tau = j1$ passes through a transition of 180°. When ω approaches ∞, $|G| = 1$ and

Figure 7.13. Twin-T network.

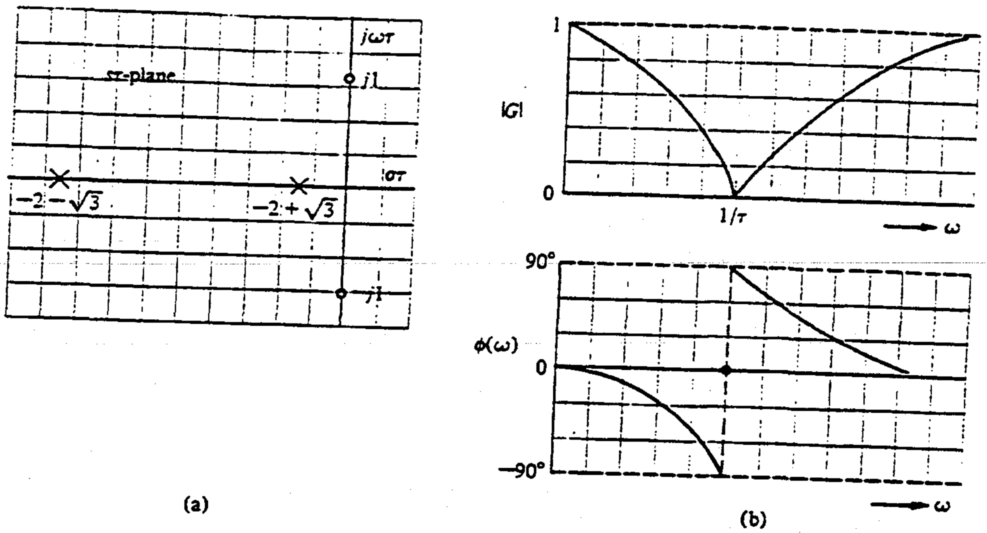


Figure 7.14. Twin-T network. (a) Pole–zero pattern. (b) Frequency response.

$\phi(\omega) = 0$ again. Evaluating several intermediate frequencies, one can readily obtain the frequency response as shown in Fig. 7.14(b).

In the previous examples the poles and zeros of $G(s)$ have been restricted to the left-hand plane. However, a system may have zeros located in the right-hand s-plane and may still be stable. Transfer functions with zeros in the right-hand s-plane are classified as *nonminimum phase-shift* transfer functions. If the zeros of a transfer function are all reflected about the $j\omega$-axis, there is no change in the magnitude of the transfer function, and the only difference is in the phase-shift characteristics. If the phase characteristics of the two system functions are compared, it can be readily shown that the net phase shift over the frequency range from zero to infinity is less for the system with all its zeros in the left-hand s-plane. Thus, the transfer function $G_1(s)$, with all its zeros in the left-hand s-plane is called a *minimum phase* transfer function. The transfer function $G_2(s)$, with $|G_2(j\omega)| = |G_1(j\omega)|$ and all the zeros of $G_1(s)$ reflected about the $j\omega$-axis into the right-hand s-plane, is called a *nonminimum phase* transfer function. Reflection of any zero or pair of zeros into the right-half plane results in a nonminimum phase transfer function.

The two pole–zero patterns shown in Fig. 7.15(a) and 7.15(b) have the same amplitude characteristics as can be deduced from the vector lengths. However, the phase characteristics are different for Fig. 7.15(a) and Fig. 7.15(b). The minimum phase characteristic of Fig. 7.15(a) and the nonminimum phase characteristic of Fig. 7.15(b) are shown in Fig. 7.16. Clearly, the phase shift of

$$G_1(s) = \frac{s + z}{s + p}$$

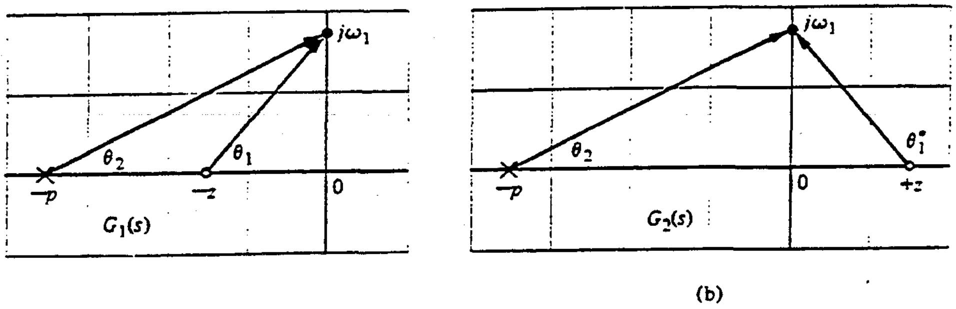


Figure 7.15. Pole–zero patterns giving the same amplitude response and different phase characteristics.

ranges over less than 80°, whereas the phase shift of

$$G_2(s) = \frac{s - z}{s + p}$$

ranges over 180°. The meaning of the term minimum phase is illustrated by Fig. 7.16. The range of phase shift of a minimum phase transfer function is the least possible or minimum corresponding to a given amplitude curve, whereas the range of the nonminimum phase curve is greater than the minimum possible for the given amplitude curve.

A particularly interesting nonminimum phase network is the *all-pass* network, which can be realized with a symmetrical lattice network [7]. A symmetrical pattern of poles and zeros is obtained as shown in Fig. 7.17(a). Again, the magnitude $|G|$ remains constant; in this case, it is equal to unity. However, the angle varies from 0° to −360°. Because $\theta_2 = 180° - \theta_1$ and $\theta_2^* = 180° - \theta_1^*$, the phase is given by $\phi(\omega) = -2(\theta_1 + \theta_1^*)$. The magnitude and phase characteristic of the all-pass network is shown in Fig. 7.17(b). A nonminimum phase lattice network is shown in Fig. 7.17(c).

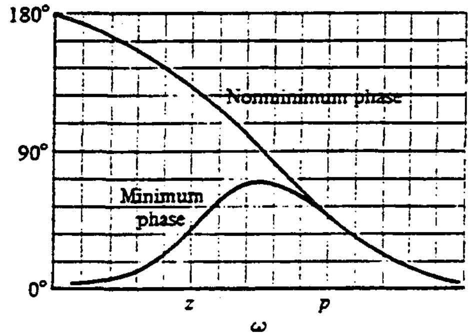


Figure 7.16. The phase characteristics for the minimum phase and nonminimum phase transfer function.

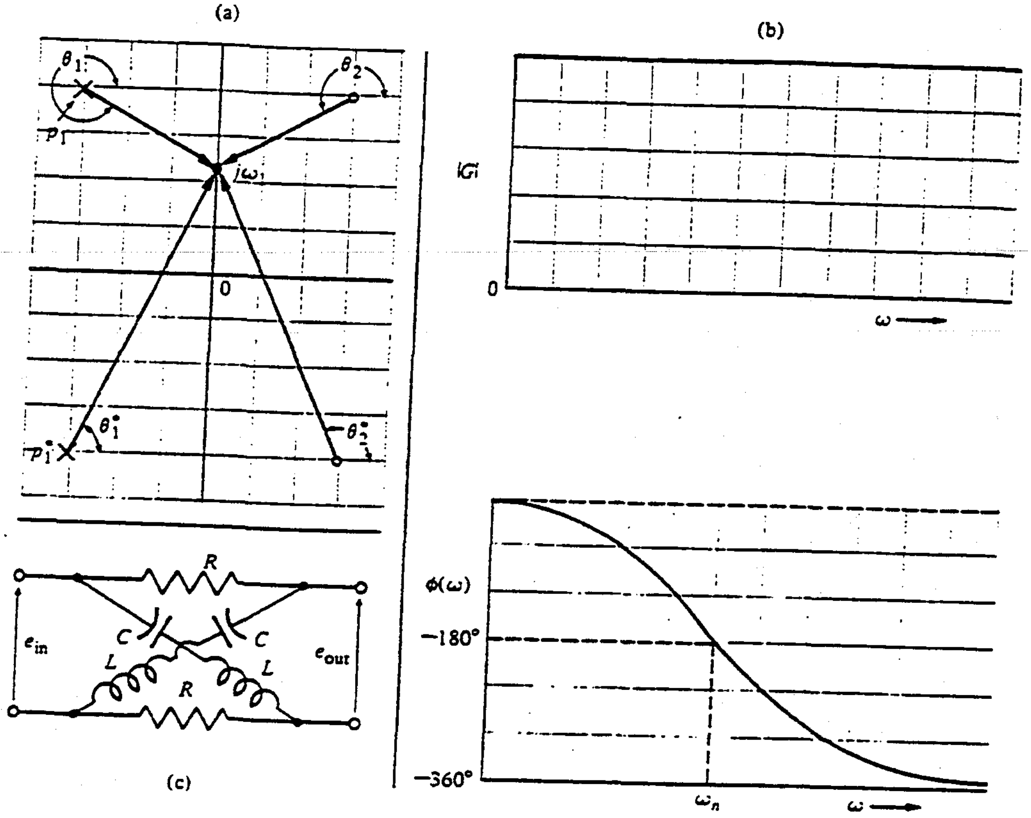


Figure 7.17. The all-pass network pole-zero pattern and frequency response.

7.3 Example of Drawing the Bode Diagram

The Bode diagram of a transfer function $G(s)$, which contains several zeros and poles, is obtained by adding the plot due to each individual pole and zero. The simplicity of this method will be illustrated by considering a transfer function that possesses all the factors considered in the preceding section. The transfer function of interest is

$$G(j\omega) = \frac{5(1 + j0.1\omega)}{j\omega(1 + j0.5\omega)(1 + j0.6(\omega/50) + (j\omega/50)^2)} \cdot \qquad (7.42)$$

The factors, in order of their occurrence as frequency increases, are

1. a constant gain $K = 5$,
2. a pole at the origin,

CHAPTER 8

Stability in the Frequency Domain

Preview

As we noted in earlier chapters, it is important to determine whether a system is stable. If it is stable, then the degree of stability is important to determine. We may use the frequency response of a transfer function around a feedback loop $GH(j\omega)$ to provide answers to our inquiry about the system's relative stability.

We will use some concepts developed in the theory of complex variables to obtain a stability criterion in the frequency domain. Then this criterion can be extended to indicate relative stability by indicating how close we come to operating at the edge of instability.

We will then demonstrate how one can examine the frequency response of the closed-loop transfer function, $T(j\omega)$, as well as the loop transfer function $GH(j\omega)$.

Finally, we will use these methods to analyze the response and performance of a system with a pure time delay, without attenuation, located within the feedback loop of a closed-loop control system.

8.1 Introduction

For a control system, it is necessary to determine whether the system is stable. Furthermore, if the system is stable, it is often necessary to investigate the relative stability. In Chapter 5, we discussed the concept of stability and several methods of determining the absolute and relative stability of a system. The Routh–Hurwitz method discussed in Chapter 5 is a useful method for investigating the characteristic equation expressed in terms of the complex variable $s = \sigma + j\omega$. Then in Chapter 6, we investigated the relative stability of a system utilizing the root locus method, which is also in terms of the complex variable s. In this chapter, we are concerned with investigating the stability of a system in the real frequency domain, that is, in terms of the frequency response discussed in Chapter 7.

The frequency response of a system represents the sinusoidal steady-state response of a system and provides sufficient information for the determination of the relative stability of the system. The frequency response of a system can readily be obtained experimentally by exciting the system with sinusoidal input signals; therefore it can be utilized to investigate the relative stability of a system when the system parameter values have not been determined. Furthermore, a frequency-domain stability criterion would be useful for determining suitable approaches to altering a system in order to increase its relative stability.

A frequency domain stability criterion was developed by H. Nyquist in 1932 and remains a fundamental approach to the investigation of the stability of linear control systems [1, 2]. The *Nyquist stability criterion* is based upon a theorem in the theory of the function of a complex variable due to Cauchy. Cauchy's theorem is concerned with the *mapping* of *contours* in the complex s-plane, and fortunately the theorem can be understood without a formal proof, which uses complex variable theory.

In order to determine the relative stability of a closed-loop system, we must investigate the characteristic equation of the system:

$$F(s) = 1 + P(s) = 0. \tag{8.1}$$

For a multiloop system, we found in Section 2.7 that, in terms of signal-flow graphs, the characteristic equation is

$$F(s) = \Delta(s) = 1 - \Sigma L_n + \Sigma L_m L_q \cdots,$$

where $\Delta(s)$ is the graph determinant. Therefore we can represent the characteristic equation of single-loop or multiple-loop systems by Eq. (8.1), where $P(s)$ is a rational function of s. In order to assure stability, we must ascertain that all the zeros of $F(s)$ lie in the left-hand s-plane. Nyquist proposed, in order to investigate this, a mapping of the right-hand s-plane into the $F(s)$-plane. Therefore, to utilize and understand Nyquist's criterion, we shall first consider briefly the mapping of contours in the complex plane.

8.2 Mapping of Contours in the s-Plane

We are concerned with the mapping of contours in the s-plane by a function $F(s)$. A *contour map* is a contour or trajectory in one plane mapped or translated into another plane by a relation $F(s)$. Since s is a complex variable, $s = \sigma + j\omega$, the function $F(s)$ is itself complex and can be defined as $F(s) = u + jv$ and can be represented on a complex $F(s)$-plane with coordinates u and v. As an example, let us consider a function $F(s) = 2s + 1$ and a contour in the s-plane as shown in Fig. 8.1(a). The mapping of the s-plane unit square contour to the $F(s)$ plane is accomplished through the relation $F(s)$, and so

$$u + jv = F(s) = 2s + 1 = 2(\sigma + j\omega) + 1. \tag{8.2}$$

Therefore, in this case, we have

$$u = 2\sigma + 1 \tag{8.3}$$

and

$$v = 2\omega. \tag{8.4}$$

Thus the contour has been mapped by $F(s)$ into a contour of an identical form, a

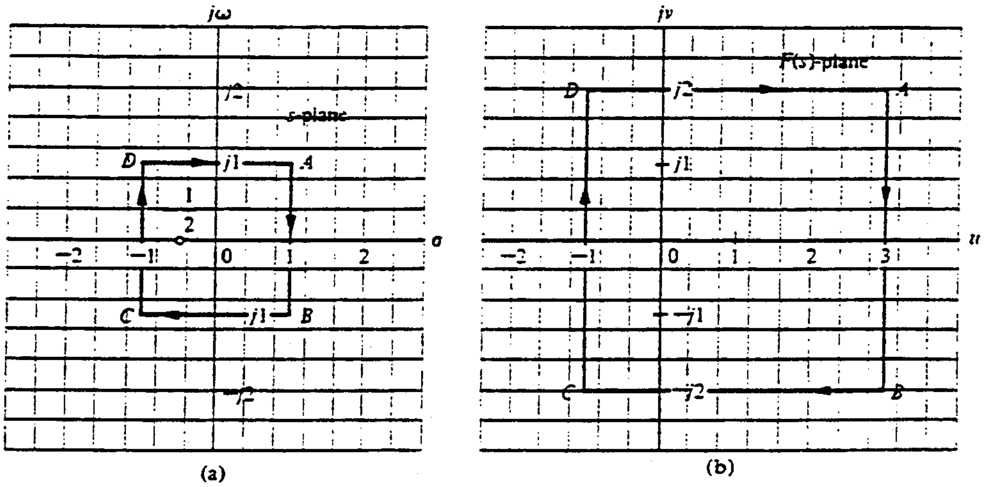


Figure 8.1. Mapping of a square contour by $F(s) = 2s + 1 = 2(s + \frac{1}{2})$.

square, with the center shifted by one unit and the magnitude of a side multiplied by 2. This type of mapping, which retains the angles of the s-plane contour on the $F(s)$-plane, is called a *conformal mapping.* We also note that a closed contour in the s-plane results in a closed contour in the $F(s)$-plane.

The points A, B, C, and D, as shown in the s-plane contour, map into the points A, B, C, and D shown in the $F(s)$-plane. Furthermore, a direction of traversal of the s-plane contour can be indicated by the direction $ABCD$ and the arrows shown on the contour. Then a similar traversal occurs on the $F(s)$-plane contour as we pass $ABCD$ in order, as shown by the arrows. By convention, the area within a contour to the right of the traversal of the contour is considered to be the *area enclosed* by the contour. Therefore we will assume *clockwise traversal* of a contour to be positive and the area enclosed within the contour to be on the right. This convention is opposite to that usually employed in complex variable theory but is equally applicable and is generally used in control system theory. The reader might consider the area on the right as he walks along the contour in a clockwise direction and call this rule "clockwise and eyes right."

Typically, we are concerned with an $F(s)$ that is a rational function of s. Therefore it will be worthwhile to consider another example of a mapping of a contour. Let us again consider the unit square contour for the function

$$F(s) = \frac{s}{s+2}. \tag{8.5}$$

Several values of $F(s)$ as s traverses the square contour are given in Table 8.1, and the resulting contour in the $F(s)$-plane is shown in Fig. 8.2(b). The contour in the $F(s)$-plane encloses the origin of the $F(s)$-plane because the origin lies within the enclosed area of the contour in the $F(s)$-plane.

Cauchy's theorem is concerned with the mapping of a function $F(s)$, which has a finite number of poles and zeros within the contour so that we may express $F(s)$ as

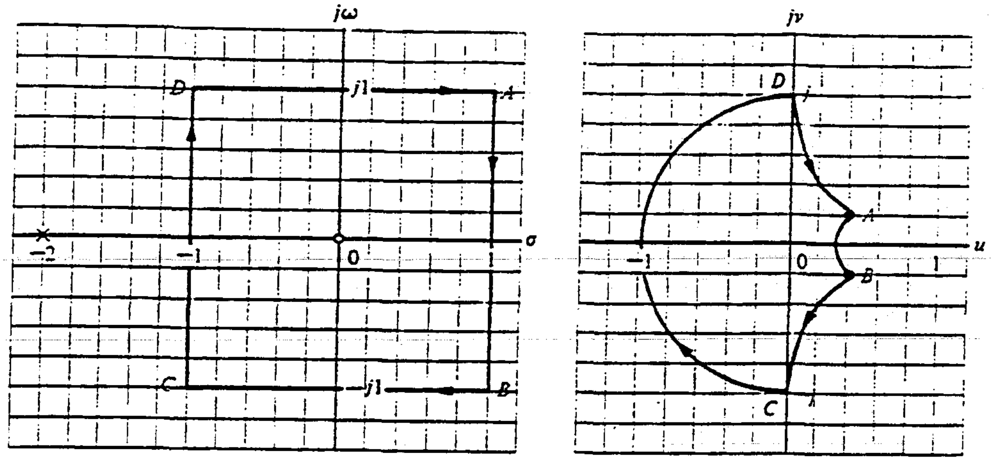
$$F(s) = \frac{K\prod_{i=1}^{n}(s+s_i)}{\prod_{k=1}^{M}(s+s_k)}, \tag{8.6}$$

where s_i are the zeros of the function $F(s)$ and s_k are the poles of $F(s)$. The function $F(s)$ is the characteristic equation, and so

$$F(s) = 1 + P(s), \tag{8.7}$$

Table 8.1

$s = \sigma + j\omega$	Point A $1 + j1$	1	Point B $1 - j1$	$-j1$	Point C $-1 - j1$	-1	Point D $-1 + j1$	$j1$
$F(s) = u + jv$	$\frac{4+2j}{10}$	$\frac{1}{3}$	$\frac{4-2j}{10}$	$\frac{1-2j}{5}$	$-j$	-1	$+j$	$\frac{1+2j}{5}$



(a)

(b)

Figure 8.2. Mapping for $F(s) = s/(s + 2)$.

where

$$P(s) = \frac{N(s)}{D(s)}.$$

Therefore, we have

$$F(s) = 1 + \frac{N(s)}{D(s)} = \frac{D(s) + N(s)}{D(s)} = \frac{K\prod_{i=1}^{n}(s + s_i)}{\prod_{k=1}^{M}(s + s_k)}, \tag{8.8}$$

and the poles of $P(s)$ are the poles of $F(s)$. However, it is the zeros of $F(s)$ that are the characteristic roots of the system and that indicate the response of the system. This is clear if we recall that the output of the system is

$$C(s) = T(s)R(s) = \frac{\Sigma P_k\Delta_k}{\Delta(s)} R(s) = \frac{\Sigma P_k\Delta_k}{F(s)} R(s), \tag{8.9}$$

where P_k and Δ_k are the path factors and cofactors as defined in Section 2.7.

Now, reexamining the example when $F(s) = 2(s + \frac{1}{2})$, we have one zero of $F(s)$ at $s = -\frac{1}{2}$, as shown in Fig. 8.1. The contour that we chose (that is, the unit square) enclosed and encircled once the zero within the area of the contour. Similarly, for the function $F(s) = s/(s + 2)$, the unit square encircled the zero at the origin but did not encircle the pole at $s = -2$. The encirclement of the poles and zeros of $F(s)$ can be related to the encirclement of the origin in the $F(s)$-plane by a *theorem* of *Cauchy*, commonly known as the *principle of the argument*, which states [3, 7]:

> If a contour Γ_s in the s-plane encircles Z zeros and P poles of $F(s)$ and does not pass through any poles or zeros of $F(s)$ as the traversal is in the clockwise direction

along the contour, the corresponding contour Γ_F in the $F(s)$ - plane encircles the origin of the $F(s)$-plane $N = Z - P$ times in the clockwise direction.

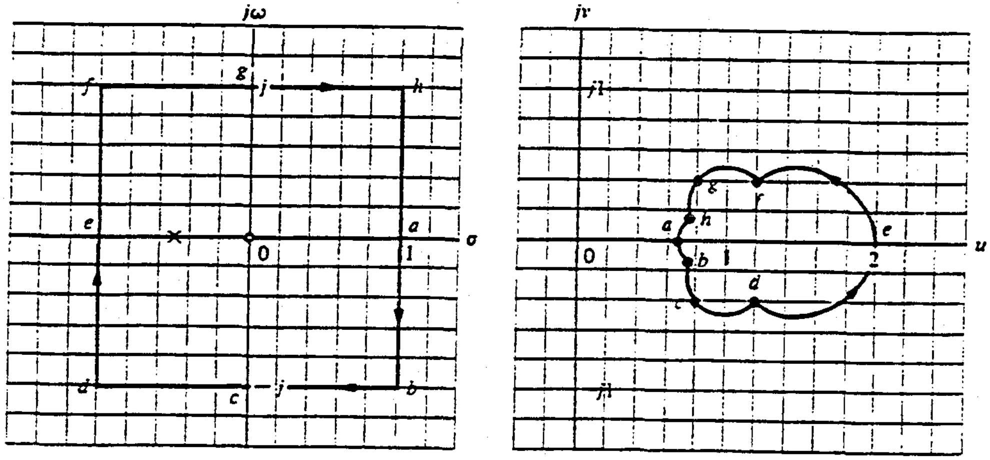
Thus for the examples shown in Figs. 8.1 and 8.2, the contour in the $F(s)$-plane encircles the origin once, because $N = Z - P = 1$, as we expect. As another example, consider the function $F(s) = s/(s + ½)$. For the unit square contour shown in Fig. 8.3(a), the resulting contour in the $F(s)$ plane is shown in Fig. 8.3(b). In this case, $N = Z - P = 0$ as is the case in Fig. 8.3(b), since the contour Γ_F does not encircle the origin.

Cauchy's theorem can be best comprehended by considering $F(s)$ in terms of the angle due to each pole and zero as the contour Γ_s is traversed in a clockwise direction. Thus let us consider the function

$$F(s) = \frac{(s + z_1)(s + z_2)}{(s + p_1)(s + p_2)}, \tag{8.10}$$

where z_i is a zero of $F(s)$ and p_k is a pole of $F(s)$. Equation (8.10) can be written as

$$\begin{aligned} F(s) &= |F(s)| \underline{/F(s)} \\ &= \frac{|s + z_1|\,|s + z_2|}{|s + p_1|\,|s + p_2|} (\underline{/s + z_1} + \underline{/s + z_2} - \underline{/s + p_1} - \underline{/s + p_2}) \qquad (8.11) \\ &= |F(s)|(\phi_{z_1} + \phi_{z_2} - \phi_{p_1} - \phi_{p_2}). \end{aligned}$$



(a) (b)

Figure 8.3. Mapping for $F(s) = s/(s + ½)$.

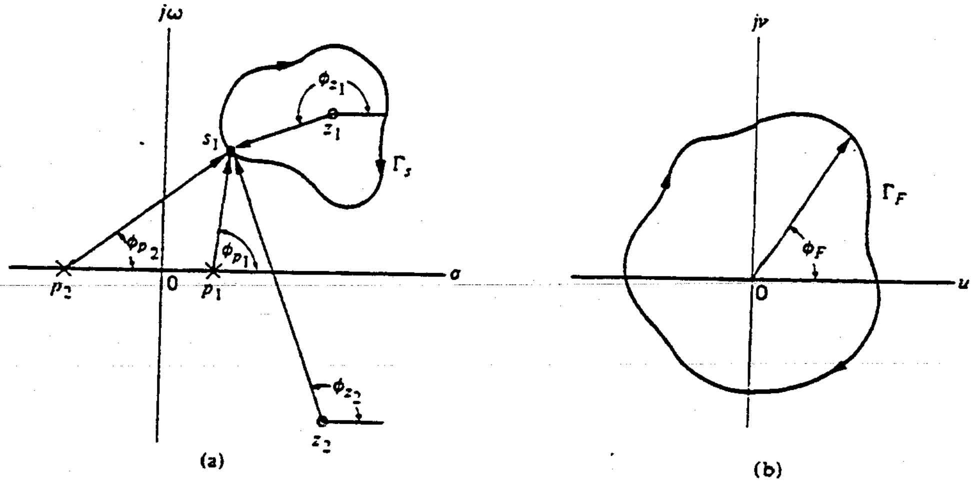


Figure 8.4. Evaluation of the net angle of Γ_F.

Now, considering the vectors as shown for a specific contour Γ_s (Fig. 8.4a), we can determine the angles as s traverses the contour. Clearly, the net angle change as s traverses along Γ_s a full rotation of 360° for ϕ_{p1}, ϕ_{p2} and ϕ_{z2} is zero degrees. However, for ϕ_{z1} as s traverses 360° around Γ_s, the angle ϕ_{z1} traverses a full 360° clockwise. Thus, as Γ_s is completely traversed, the net angle of $F(s)$ is equal to 360° since only one zero is enclosed. If Z zeros were enclosed within Γ_s, then the net angle would be equal to $\phi_z = 2\pi(Z)$ rad. Following this reasoning, if Z zeros and P poles are encircled as Γ_s is traversed, then $2\pi(Z) - 2\pi(P)$ is the net resultant angle of $F(s)$. Thus the net angle of Γ_F of the contour in the $F(s)$-plane, ϕ_F, is simply

$$\phi_F = \phi_Z - \phi_P$$

or

$$2\pi N = 2\pi Z - 2\pi p, \tag{8.12}$$

and the net number of encirclements of the origin of the $F(s)$-plane is $N = Z - P$. Thus for the contour shown in Fig. 8.4(a), which encircles one zero, the contour Γ_F shown in Fig. 8.4(b) encircles the origin once in the clockwise direction.

As an example of the use of Cauchy's theorem, consider the pole–zero pattern shown in Fig. 8.5(a) with the contour Γ_s to be considered. The contour encloses and encircles three zeros and one pole. Therefore we obtain

$$N = 3 - 1 = +2,$$

and Γ_F completes two clockwise encirclements of the origin in the $F(s)$-plane as shown in Fig. 8.5(b).

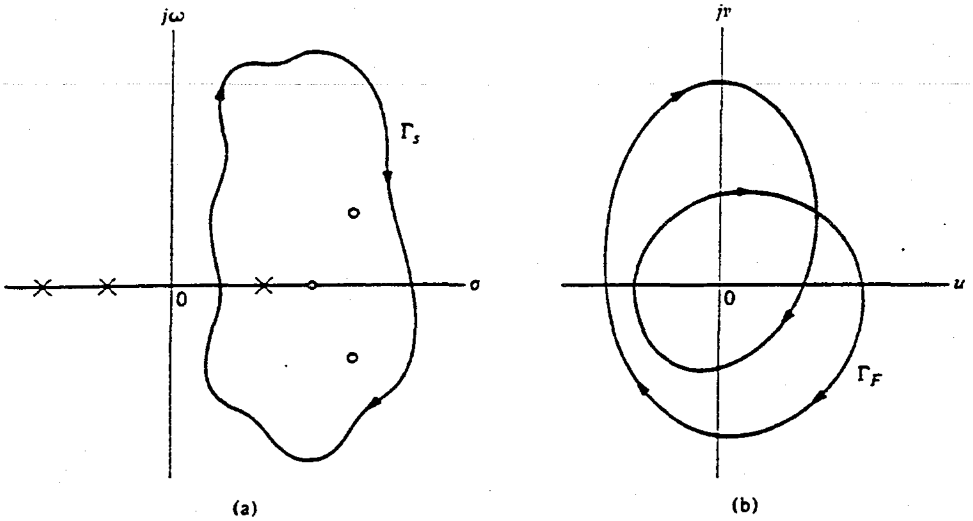


Figure 8.5. Example of Cauchy's theorem.

For the pole and zero pattern shown and the contour Γ_s as shown in Fig. 8.6(a), one pole is encircled and no zeros are encircled. Therefore we have

$$N = Z - P = -1,$$

and we expect one encirclement of the origin by the contour Γ_F in the $F(s)$-plane. However, since the sign of N is negative, we find that the encirclement moves in the counterclockwise direction as shown in Fig. 8.6(b).

Now that we have developed and illustrated the concept of mapping of contours through a function $F(s)$, we are ready to consider the stability criterion proposed by Nyquist.

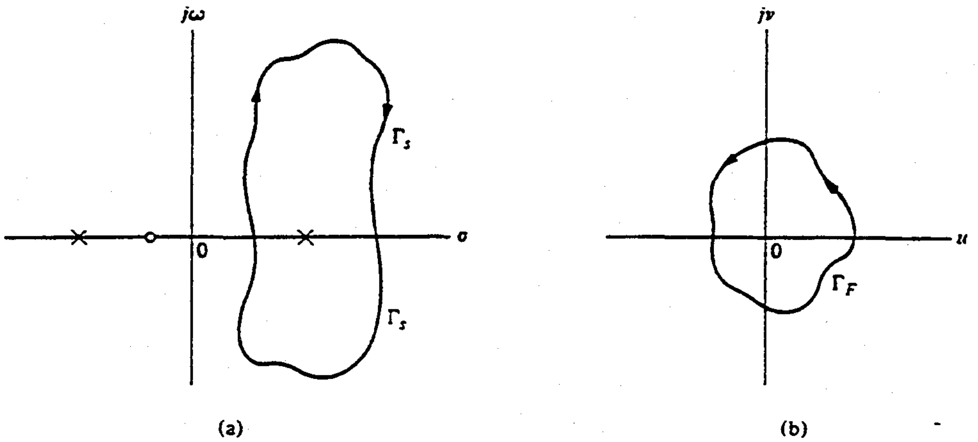


Figure 8.6. Example of Cauchy's theorem.

8.3 The Nyquist Criterion

In order to investigate the stability of a control system, we consider the characteristic equation, which is $F(s) = 0$, so that

$$F(s) = 1 + P(s) = \frac{K\prod_{i=1}^{n}(s + s_i)}{\prod_{k=1}^{M}(s + s_k)} = 0. \tag{8.13}$$

For a system to be stable, all the zeros of $F(s)$ must lie in the left-hand s-plane. Thus we find that the roots of a stable system [the zeros of $F(s)$] must lie to the left of the $j\omega$-axis in the s-plane. Therefore we chose a contour Γ_s in the s-plane which encloses the entire right-hand s-plane, and we determine whether any zeros of $F(s)$ lie within Γ_s by utilizing Cauchy's theorem. That is, we plot Γ_F in the $F(s)$-plane and determine the number of encirclements of the origin N. Then the number of zeros of $F(s)$ within the Γ_s contour [and therefore unstable zeros of $F(s)$] is

$$Z = N + P. \tag{8.14}$$

Thus if $P = 0$, as is usually the case, we find that the number of unstable roots of the system is equal to N, the number of encirclements of the origin of the $F(s)$ plane.

The Nyquist contour that encloses the entire right-hand s-plane is shown in Fig. 8.7. The contour Γ_s passes along the $j\omega$-axis from $-j\infty$ to $+j\infty$, and this part of the contour provides the familiar $F(j\omega)$. The contour is completed by a semicircular path of radius r where r approaches infinity.

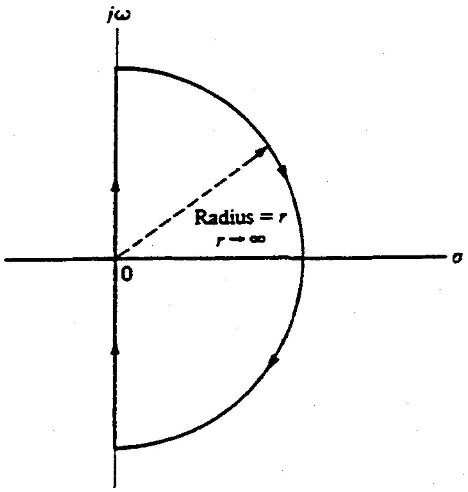


Figure 8.7. The Nyquist contour.

Now, the Nyquist criterion is concerned with the mapping of the characteristic equation

$$F(s) = 1 + P(s) \tag{8.15}$$

and the number of encirclements of the origin of the $F(s)$-plane. Alternatively, we may define the function $F'(s)$ so that

$$F'(s) = F(s) - 1 = P(s). \tag{8.16}$$

The change of functions represented by Eq. (8.16) is very convenient because $P(s)$ is typically available in factored form, while $1 + P(s)$ is not. Then the mapping of Γ_s in the s-plane will be through the function $F'(s) = P(s)$ into the $P(s)$-plane. In this case the number of clockwise encirclements of the origin of the $F(s)$-plane becomes the number of clockwise encirclements of the -1 point in the $F'(s) = P(s)$ plane because $F'(s) = F(s) - 1$. Therefore, the *Nyquist stability criterion* can be stated as follows:

> A feedback system is stable if and only if the contour Γ_p in the $P(s)$-plane does not encircle the $(-1, 0)$ point when the number of poles of $P(s)$ in the right-hand s-plane is zero ($P = 0$).

When the number of poles of $P(s)$ in the right-hand s-plane is other than zero, the Nyquist criterion is:

> A feedback control system is stable if and only if, for the contour Γ_p, the number of counterclockwise encirclements of the $(-1, 0)$ point is equal to the number of poles of $P(s)$ with positive real parts.

The basis for the two statements is the fact that for the $F'(s) = P(s)$ mapping, the number of roots (or zeros) of $1 + P(s)$ in the right-hand s-plane is represented by the expression

$$Z = N + P.$$

Clearly, if the number of poles of $P(s)$ in the right-hand s-plane is zero ($P = 0$), we require for a stable system that $N = 0$ and the contour Γ_p must not encircle the -1 point. Also, if P is other than zero and we require for a stable system that $Z = 0$, then we must have $N = -P$, or P counterclockwise encirclements.

It is best to illustrate the use of the Nyquist criterion by completing several examples.

■ Example 8.1

A single-loop control system is shown in Fig. 8.8, where

$$GH(s) = \frac{K}{s(\tau s + 1)}. \tag{8.17}$$

In this single-loop case, $P(s) = GH(s)$ and we determine the contour $\Gamma_p = \Gamma_{GH}$ in

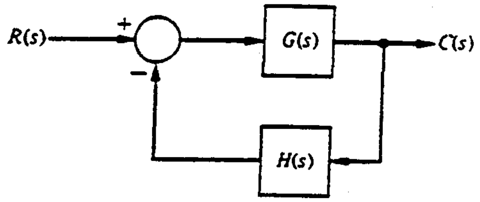


Figure 8.8. Single-loop feedback control system.

the $GH(s)$-plane. The contour Γ_s in the s-plane is shown in Fig. 8.9(a), where an infinitesimal detour around the pole at the origin is effected by a small semicircle of radius ϵ, where $\epsilon \rightarrow 0$. This detour is a consequence of the condition of Cauchy's theorem which requires that the contour cannot pass through the pole of the origin. A sketch of the contour Γ_{GH} is shown in Fig. 8.9(b). Clearly, the portion of the contour Γ_{GH} from $\omega = 0^+$ to $\omega = +\infty$ is simply $GH(j\omega)$, the real frequency polar plot. Let us consider each portion of the Nyquist contour Γ_s in detail and determine the corresponding portions of the $GH(s)$-plane contour Γ_{GH}.

(a) The Origin of the s-Plane. The small semicircular detour around the pole at the origin can be represented by setting $s = \epsilon e^{j\phi}$ and allowing ϕ to vary from $-90°$ at $\omega = 0^-$ to $+90°$ at $\omega = 0^+$. Because ϵ approaches zero, the mapping $GH(s)$ is

$$\lim_{\epsilon \to 0} GH(s) = \lim_{\epsilon \to 0} \left(\frac{K}{\epsilon e^{j\phi}}\right) = \lim_{\epsilon \to 0} \left(\frac{K}{\epsilon}\right) e^{-j\phi}. \tag{8.18}$$

Therefore, the angle of the contour in the $GH(s)$-plane changes from 90° at $\omega = 0^-$ to $-90°$ at $\omega = 0^+$, passing through 0° at $\omega = 0$. The radius of the contour in

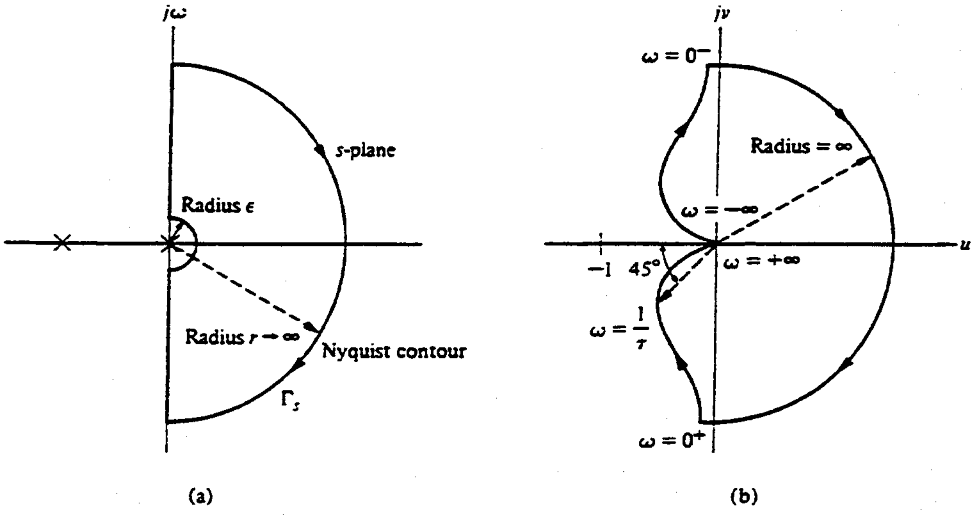


Figure 8.9. Nyquist contour and mapping for $GH(s) = K/s(\tau s + 1)$.

the $GH(s)$-plane for this portion of the contour is infinite, and this portion of the contour is shown in Fig. 8.9(b).

(b) The Portion from $\omega = 0^+$ to $\omega = +\infty$. The portion of the contour Γ_s from $\omega = 0^+$ to $\omega = +\infty$ is mapped by the function $GH(s)$ as the real frequency polar plot because $s = j\omega$ and

$$GH(s)|_{s=j\omega} = GH(j\omega) \tag{8.19}$$

for this part of the contour. This results in the real frequency polar plot shown in Fig. 8.9(b). When ω approaches $+\infty$, we have

$$\begin{aligned}\lim_{\omega\to+\infty} GH(j\omega) &= \lim_{\omega\to+\infty} \frac{K}{+j\omega(j\omega\tau + 1)} \\ &= \lim_{\omega\to\infty} \left|\frac{K}{\tau\omega^2}\right| \underline{/-(\pi/2) - \tan^{-1}\omega\tau}.\end{aligned} \tag{8.20}$$

Therefore the magnitude approaches zero at an angle of $-180°$.

(c) The Portion from $\omega = +\infty$ to $\omega = -\infty$. The portion of Γ_s from $\omega = +\infty$ to $\omega = -\infty$ is mapped into the point zero at the origin of the $GH(s)$-plane by the function $GH(s)$. The mapping is represented by

$$\lim_{r\to\infty} GH(s)|_{s=re^{j\phi}} = \lim_{r\to\infty} \left|\frac{K}{r^2}\right| e^{-2j\phi} \tag{8.21}$$

as ϕ changes from $\phi = +90°$ at $\omega = +\infty$ to $\phi = -90°$ at $\omega = -\infty$. Thus the contour moves from an angle of $-180°$ at $\omega = +\infty$ to an angle of $+180°$ at $\omega = -\infty$. The magnitude of the $GH(s)$ contour when r is infinite is always zero or a constant.

(d) The Portion from $\omega = -\infty$ to $\omega = 0^-$. The portion of the contour Γ_s from $\omega = -\infty$ to $\omega = 0^-$ is mapped by the function $GH(s)$ as

$$GH(s)|_{s=-j\omega} = GH(-j\omega). \tag{8.22}$$

Thus we obtain the complex conjugate of $GH(j\omega)$, and the plot for the portion of the polar plot from $\omega = -\infty$ to $\omega = 0^-$ is symmetrical to the polar plot from $\omega = +\infty$ to $\omega = 0^+$. This symmetrical polar plot is shown on the $GH(s)$-plane in Fig. 8.9(b).

Now, in order to investigate the stability of this second-order system, we first note that the number of poles P within the right-hand s-plane is zero. Therefore, for this system to be stable, we require $N = Z = 0$, and the contour Γ_{GH} must not encircle the -1 point in the GH-plane. Examining Fig. 8.9(b), we find that irrespective of the value of the gain K and the time constant τ, the contour does not encircle the -1 point, and the system is always stable. As in Chapter 6, we are considering positive values of gain K. If negative values of gain are to be considered, one should use $-K$, where $K \geq 0$.

We may draw two general conclusions from this example as follows:

1. The plot of the contour Γ_{GH} for the range $-\infty < \omega < 0^-$ will be the complex conjugate of the plot for the range $0^+ < \omega < +\infty$ and the polar plot of $GH(s)$ will be symmetrical in the $GH(s)$-plane about the u-axis. Therefore *it is sufficient to construct the contour Γ_{GH} for the frequency range $0^+ < \omega < +\infty$ in order to investigate the stability.*
2. The magnitude of $GH(s)$ as $s = re^{j\phi}$ and $r \to \infty$ will normally approach zero or a constant.

■ Example 8.2

Let us again consider the single-loop system shown in Fig. 8.8 when

$$GH(s) = \frac{K}{s(\tau_1 s + 1)(\tau_2 s + 1)}. \tag{8.23}$$

The Nyquist contour Γ_s is shown in Fig. 8.9(a). Again this mapping is symmetrical for $GH(j\omega)$ and $GH(-j\omega)$ so that it is sufficient to investigate the $GH(j\omega)$-locus. The origin of the s-plane maps into a semicircle of infinite radius as in the last example. Also, the semicircle $re^{j\phi}$ in the s-plane maps into the point $GH(s) = 0$ as we expect. Therefore, in order to investigate the stability of the system, it is sufficient to plot the portion of the contour Γ_{GH} which is the real frequency polar plot $GH(j\omega)$ for $0^+ < \omega < 0^+ +\infty$. Therefore, when $s = +j\omega$, we have

$$\begin{aligned} GH(j\omega) &= \frac{K}{j\omega(j\omega\tau_1 + 1)(j\omega\tau_2 + 1)} \\ &= \frac{-K(\tau_1 + \tau_2) - jK(1/\omega)(1 - \omega^2\tau_1\tau_2)}{1 + \omega^2(\tau_1^2 + \tau_2^2) + \omega^4\tau_1^2\tau_2^2} \\ &= \frac{K}{[\omega^4(\tau_1 + \tau_2)^2 + \omega^2(1 - \omega^2\tau_1\tau_2)^2]^{1/2}} \\ &\quad \times \underline{/-\tan^{-1}\omega\tau_1 - \tan^{-1}\omega\tau_2 - (\pi/2)}. \end{aligned} \tag{8.24}$$

When $\omega = 0^+$, the magnitude of the locus is infinite at an angle of $-90°$ in the $GH(s)$-plane. When ω approaches $+\infty$, we have

$$\begin{aligned} \lim_{\omega\to\infty} GH(j\omega) &= \lim_{\omega\to\infty} \left|\frac{1}{\omega^3}\right| \underline{/-(\pi/2) - \tan^{-1}\omega\tau_1 - \tan^{-1}\omega\tau_2} \\ &= \left(\lim_{\omega\to\infty} \left|\frac{1}{\omega^3}\right|\right) \underline{/-(3\pi/2)}. \end{aligned} \tag{8.25}$$

Therefore $GH(j\omega)$ approaches a magnitude of zero at an angle of $-270°$. In order for the locus to approach at an angle of $-270°$, the locus must cross the u-axis in the $GH(s)$-plane as shown in Fig. 8.10. Thus it is possible to encircle the -1 point as is shown in Fig. 8.10. The number of encirclements, when the -1 point lies within the locus as shown in Fig. 8.10, is equal to two and the system is unstable

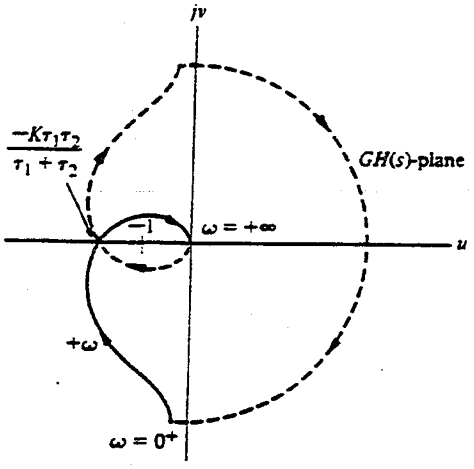


Figure 8.10. Nyquist diagram for $GH(s) = K/s(\tau_1 s + 1)(\tau_2 s + 1)$.

with two roots in the right-hand s-plane. The point where the $GH(s)$-locus intersects the real axis can be found by setting the imaginary part of $GH(j\omega) = u + jv$ equal to zero. Then we have from Eq. (8.24)

$$v = \frac{-K(1/\omega)(1-\omega^2\tau_1\tau_2)}{1+\omega^2(\tau_1^2+\tau_2^2)+\omega^4\tau_1^2\tau_2^2} = 0. \tag{8.26}$$

Thus $v = 0$ when $1 - \omega^2\tau_1\tau_2 = 0$ or $\omega = 1/\sqrt{\tau_1\tau_2}$. The magnitude of the real part, u, of $GH(j\omega)$ at this frequency is

$$\begin{aligned} u &= \left.\frac{-K(\tau_1+\tau_2)}{1+\omega^2(\tau_1^2+\tau_2^2)+\omega^4\tau_1^2\tau_2^2}\right|_{\omega^2=1/\tau_1\tau_2} \\ &= \frac{-K(\tau_1+\tau_2)\tau_1\tau_2}{\tau_1\tau_2+(\tau_1^2+\tau_2^2)+\tau_1\tau_2} = \frac{-K\tau_1\tau_2}{\tau_1+\tau_2}. \end{aligned} \tag{8.27}$$

Therefore, the system is stable when

$$\frac{-K\tau_1\tau_2}{\tau_1+\tau_2} \geq -1$$

or

$$K \leq \frac{\tau_1+\tau_2}{\tau_1\tau_2}. \tag{8.28}$$

■ Example 8.3

Again let us determine the stability of the single-loop system shown in Fig. 8.8 when

$$GH(s) = \frac{K}{s^2(\tau s + 1)}. \tag{8.29}$$

The real frequency polar plot is obtained when $s = j\omega$, and we have

$$\begin{aligned} GH(j\omega) &= \frac{K}{-\omega^2(j\omega\tau + 1)} \\ &= \frac{K}{[\omega^4 + \tau^2\omega^6]^{1/2}} \underline{/-\pi - \tan^{-1}\omega\tau}. \end{aligned} \qquad (8.30)$$

We note that the angle of $GH(j\omega)$ is always $-180°$ or greater, and the locus of $GH(j\omega)$ is above the u-axis for all values of ω. As ω approaches 0^+, we have

$$\lim_{\omega\to 0+} GH(j\omega) = \left(\lim_{\omega\to 0+} \left|\frac{K}{\omega^2}\right|\right) \underline{/-\pi}. \qquad (8.31)$$

As ω approaches $+\infty$, we have

$$\lim_{\omega\to +\infty} GH(j\omega) = \left(\lim_{\omega\to +\infty} \frac{K}{\omega^3}\right) \underline{/-3\pi/2.} \qquad (8.32)$$

At the small semicircular detour at the origin of the s-plane where $s = \epsilon e^{j\phi}$, we have

$$\lim_{\epsilon\to 0} GH(s) = \lim_{\epsilon\to 0} \frac{K}{\epsilon^2} e^{-2j\phi}, \qquad (8.33)$$

where $-\pi/2 \leq \phi \leq \pi/2$. Thus the contour Γ_{GH} ranges from an angle of $+\pi$ at $\omega = 0^+$ to $-\pi$ at $\omega = 0^-$ and passes through a full circle of 2π rad as ω changes from $\omega = 0^-$ to $\omega = 0^+$. The complete contour plot of Γ_{GH} is shown in Fig. 8.11. Because the contour encircles the -1 point twice, there are two roots of the closed-loop system in the right-hand plane and the system, irrespective of the gain K, is unstable.

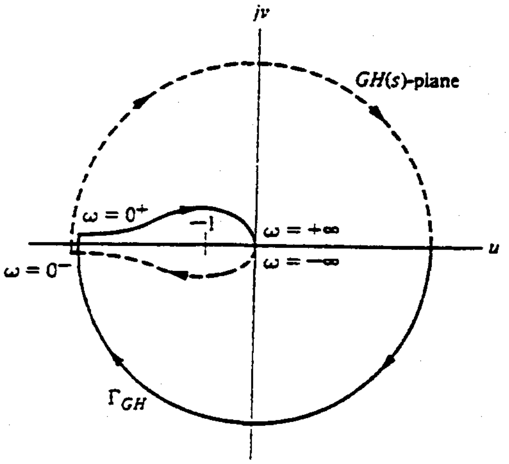


Figure 8.11. Nyquist contour plot for $GH(s) = K/s^2(\tau s + 1)$.

■ Example 8.4

Let us consider the control system shown in Fig. 8.12 and determine the stability of the system. First, let us consider the system without derivative feedback so that $K_2 = 0$. Then we have the open-loop transfer function

$$GH(s) = \frac{K_1}{s(s-1)}. \tag{8.34}$$

Thus the open-loop transfer function has one pole in the right-hand plane, and therefore $P = 1$. In order for this system to be stable, we require $N = -P = -1$, one counterclockwise encirclement of the -1 point. At the semicircular detour at the origin of the s-plane, we let $s = \epsilon e^{j\phi}$ when $-\pi/2 \leq \phi \leq \pi/2$. Then we have, when $s = \epsilon e^{j\phi}$,

$$\lim_{\epsilon\to 0} GH(s) = \lim_{\epsilon\to 0} \frac{K_1}{-\epsilon e^{j\phi}} = \left(\lim_{\epsilon\to 0}\left|\frac{K_1}{\epsilon}\right|\right) \angle -180^\circ - \phi. \tag{8.35}$$

Therefore this portion of the contour Γ_{GH} is a semicircle of infinite magnitude in the left-hand GH-plane, as shown in Fig. 8.13. When $s = j\omega$, we have

$$\begin{aligned} GH(j\omega) &= \frac{K_1}{j\omega(j\omega - 1)} = \frac{K_1}{(\omega^2 + \omega^4)^{1/2}} \angle (-\pi/2) - \tan^{-1}(-\omega) \\ &= \frac{K_1}{(\omega^2 + \omega^4)^{1/2}} \angle (+\pi/2) + \tan^{-1}\omega. \end{aligned} \tag{8.36}$$

Finally, for the semicircle of radius r as r approaches infinity, we have

$$\lim_{r\to\infty} GH(s)\big|_{s=re^{j\phi}} = \left(\lim_{r\to\infty}\left|\frac{K_1}{r^2}\right|\right) e^{-2j\phi}, \tag{8.37}$$

where ϕ varies from $\pi/2$ to $-\pi/2$ in a clockwise direction. Therefore the contour Γ_{GH}, at the origin of the GH-plane, varies 2π rad in a counterclockwise direction, as shown in Fig. 8.13. Several important values of the $GH(s)$-locus are given in Table 8.2. The contour Γ_{GH} in the $GH(s)$-plane encircles the -1 point once in the clockwise direction and $N = +1$. Therefore

$$Z = N + P = 2. \tag{8.38}$$

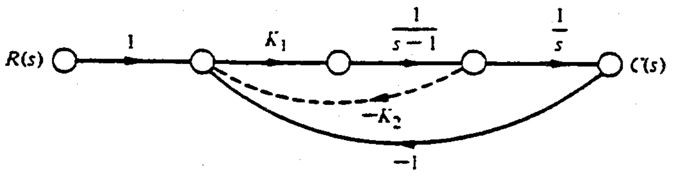


Figure 8.12. Second-order feedback control system.

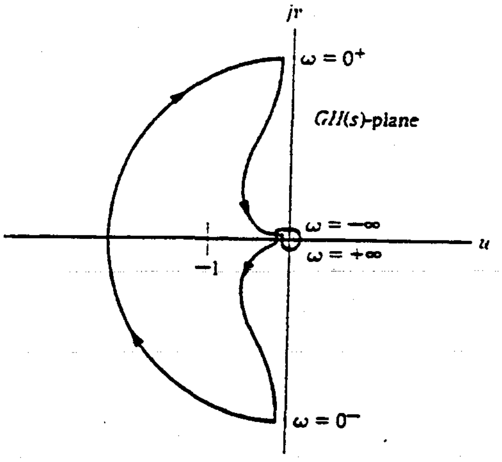


Figure 8.13. Nyquist diagram for $GH(s) = K_1/s(s - 1)$.

and the system is unstable because two zeros of the characteristic equation, irrespective of the value of the gain K, lie in the right half of the s-plane.

Let us now reconsider the system when the derivative feedback is included in the system shown in Fig. 8.12. Then the open-loop transfer function is

$$GH(s) = \frac{K_1(1 + K_2 s)}{s(s - 1)}. \tag{8.39}$$

The portion of the contour Γ_{GH} when $s = \epsilon e^{j\phi}$ is the same as the system without derivative feedback, as is shown in Fig. 8.14. However, when $s = re^{j\phi}$ as r approaches infinity, we have

$$\lim_{r\to\infty} GH(s)|_{s=re^{j\phi}} = \lim_{r\to\infty} \left|\frac{K_1K_2}{r}\right| e^{-j\phi}, \tag{8.40}$$

and the Γ_{GH}-contour at the origin of the GH-plane varies π rad in a counterclockwise direction, as shown in Fig. 8.14. The frequency locus $GH(j\omega)$ crosses the u-axis and is determined by considering the real frequency transfer function

$$\begin{aligned} GH(j\omega) &= \frac{K_1(1 + K_2 j\omega)}{-\omega^2 - j\omega} \\ &= \frac{-K_1(\omega^2 + \omega^2 K_2) + j(\omega - K_2\omega^3)K_1}{\omega^2 + \omega^4}. \end{aligned} \tag{8.41}$$

Table 8.2

s	$j0^-$	$j0^+$	$j1$	$+j\infty$	$-j\infty$
$\lvert GH\rvert/K_1$	∞	∞	$1/\sqrt{2}$	0	0
$\underline{/GH}$	$-90°$	$+90°$	$+135°$	$+180°$	$-180°$

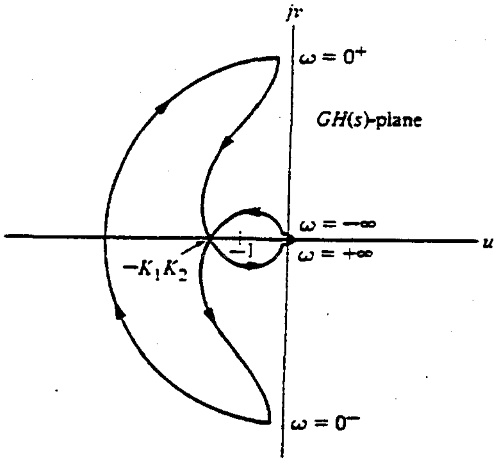


Figure 8.14. Nyquist diagram for $GH(s) = K_1(1 + K_2s)/s(s - 1)$.

The $GH(j\omega)$-locus intersects the u-axis at a point where the imaginary part of $GH(j\omega)$ is zero. Therefore,

$$\omega - K_2\omega^3 = 0$$

at this point, or $\omega^2 = 1/K_2$. The value of the real part of $GH(j\omega)$ at the intersection is then

$$u|_{\omega^2=1/K_2} = \left.\frac{-\omega^2K_1(1 + K_2)}{\omega^2 + \omega^4}\right|_{\omega^2=1/K_2} = -K_1K_2. \tag{8.42}$$

Therefore, when $-K_1K_2 < -1$ or $K_1K_2 > 1$, the contour Γ_{GH} encircles the -1 point once in a counterclockwise direction, and therefore $N = -1$. Then Z, the number of zeros of the system in the right-hand plane, is

$$Z = N + P = -1 + 1 = 0. \tag{8.43}$$

Thus the system is stable when $K_1K_2 > 1$. Often, it may be useful to utilize a computer or calculator program to calculate the Nyquist diagram [5].

8.4 Relative Stability and the Nyquist Criterion

We discussed the relative stability of a system in terms of the s-plane in Section 5.3. For the s-plane, we defined the relative stability of a system as the property measured by the relative settling time of each root or pair of roots. We would like to determine a similar measure of relative stability useful for the frequency-response method. The Nyquist criterion provides us with suitable information concerning the absolute stability and, furthermore, can be utilized to define and ascertain the relative stability of a system.

The Nyquist stability criterion is defined in terms of the $(-1, 0)$ point on the polar plot or the 0 db, 180° point on the Bode diagram or log-magnitude–phase diagram. Clearly, the proximity of the $GH(j\omega)$-locus to this stability point is a measure of the relative stability of a system. The polar plot for $GH(j\omega)$ for several values of K and

$$GH(j\omega) = \frac{K}{j\omega(j\omega\tau_1 + 1)(j\omega\tau_2 + 1)} \tag{8.44}$$

is shown in Fig. 8.15. As K increases, the polar plot approaches the -1 point and eventually encircles the -1 point for a gain $K = K_3$. We determined in Section 8.3 that the locus intersects the u-axis at a point

$$u = \frac{-K\tau_1\tau_2}{\tau_1 + \tau_2}. \tag{8.45}$$

Therefore the system has roots on the $j\omega$-axis when

$$u = -1 \quad \text{or} \quad K = \left(\frac{\tau_1 + \tau_2}{\tau_1\tau_2}\right).$$

As K is decreased below this marginal value, the stability is increased and the margin between the gain $K = (\tau_1 + \tau_2)/\tau_1\tau_2$ and a gain $K = K_2$ is a measure of the relative stability. This measure of relative stability is called the *gain margin* and is defined as *the reciprocal of the gain* $|GH(j\omega)|$ *at the frequency at which the phase angle reaches 180°* (that is, $v = 0$). The gain margin is a measure of the

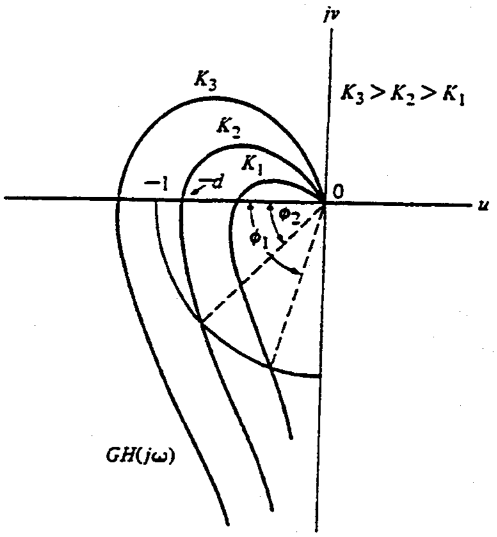


Figure 8.15. The polar plot for $GH(j\omega)$ for three values of gain.

factor by which the system gain would have to be increased for the $GH(j\omega)$ locus to pass through the $u = -1$ point. Thus, for a gain $K = K_2$ in Fig. 8.15, the gain margin is equal to the reciprocal of $GH(j\omega)$ when $v = 0$. Because $\omega = 1/\sqrt{\tau_1\tau_2}$ when the phase shift is 180°, we have a gain margin equal to

$$\frac{1}{|GH(j\omega)|} = \left[\frac{K_2\tau_1\tau_2}{\tau_1 + \tau_2}\right]^{-1} = \frac{1}{d}. \tag{8.46}$$

The gain margin can be defined in terms of a logarithmic (decibel) measure as

$$20 \log \left(\frac{1}{d}\right) = -20 \log d \text{ db}. \tag{8.47}$$

Therefore, for example, when $\tau_1 = \tau_2 = 1$, the system is stable when $K \leq 2$. Thus when $K = K_2 = 0.5$, the gain margin is equal to

$$\frac{1}{d} = \left[\frac{K_2\tau_1\tau_2}{\tau_1 + \tau_2}\right]^{-1} = 4, \tag{8.48}$$

or, in logarithmic measure,

$$20 \log 4 = 12 \text{ db}. \tag{8.49}$$

Therefore the gain margin indicates that the system gain can be increased by a factor of four (12 db) before the stability boundary is reached.

An alternative measure of relative stability can be defined in terms of the phase angle margin between a specific system and a system that is marginally stable. Several roots of the characteristic equation lie on the $j\omega$-axis when the $GH(j\omega)$-locus intersects the $u = -1$, $v = 0$ point in the GH-plane. Therefore, a measure of relative stability, the *phase margin,* is defined as *the phase angle through which the $GH(j\omega)$ locus must be rotated in order that the unity magnitude $|GH(j\omega)| = 1$ point passes through the* $(-1, 0)$ *point in the $GH(j\omega)$ plane.* This measure of relative stability is called the phase margin and is equal to the additional phase lag required before the system becomes unstable. This information can be determined from the Nyquist diagram as shown in Fig. 8.15. For a gain $K = K_2$, an additional phase angle, ϕ_2, may be added to the system before the system becomes unstable. Furthermore, for the gain K_1, the phase margin is equal to ϕ_1, as shown in Fig. 8.15.

The gain and phase margins are easily evaluated from the Bode diagram, and because it is preferable to draw the Bode diagram in contrast to the polar plot, it is worthwhile to illustrate the relative stability measures for the Bode diagram. The critical point for stability is $u = -1$, $v = 0$ in the $GH(j\omega)$ plane which is equivalent to a logarithmic magnitude of 0 db and a phase angle of 180° in the Bode diagram.

The gain margin and phase margin can be readily calculated by utilizing a computer program [6]. A computer program for accomplishing this calculation is given in Table 8.3 in the computer language BASIC [5]. This program can readily be converted to other languages. The symbols used are: W = ω; G2 = $|G(s)|^2$;

Table 8.3. A Computer Program in BASIC Computer Language for Calculating the Gain Margin and Phase Margin for the Third-Order System $GH(j\omega) = 1/j\omega(j\omega + 1)(0.2j\omega + 1)$

```
 10 LET W = 0.1
 20 GOSUB 100
 30 IF G2 < =1 THEN 60
 35 IF G2 < =100 THEN 50
 40 LET W = 2*W
 45 GO TO 20
 50 LET W = 1.02*W
 55 GO TO 20
 60 IF P> =180 THEN 140
 65 PRINT "UNITY GAIN", "W =" W, "P=" P
 70 LET W=1.02*W
 75 GOSUB 100
 80 IF P> =180 THEN 90
 85 GO TO 70
 90 PRINT "W=" W, "GAIN MARGIN =" 4.343*LOG(1/G2)
 95 GO TO 200
100 LET P = 57.3*(ATN(W) + ATN(0.2*W) + 1.571)
110 LET X = W*W
120 LET G2 = 1/((1 + X)*(1 + 0.04*X)*X)
130 RETURN
140 PRINT "W =" W, "SYSTEM UNSTABLE"
200 END
```

P = phase of $G(s)$; PM = phase margin. The program is shown for the case where $GH(j\omega)$ is as given in Eq. (8.50). The calculations commence at $\omega = 0.1$ and increase by 2% at each iteration at line 50.

The Bode diagram of

$$GH(j\omega) = \frac{1}{j\omega(j\omega + 1)(0.2j\omega + 1)} \tag{8.50}$$

is shown in Fig. 8.16. The phase angle when the logarithmic magnitude is 0 db is equal to 137°. Thus the phase margin is $180° - 137° = 43°$, as shown in Fig. 8.16. The logarithmic magnitude when the phase angle is $-180°$ is -15 db, and therefore the gain margin is equal to 15 db, as shown in Fig. 8.16.

The frequency response of a system can be graphically portrayed on the logarithmic-magnitude–phase-angle diagram. For the log-magnitude–phase diagram, the critical stability point is the 0 db, $-180°$ point, and the gain margin and phase margin can be easily determined and indicated on the diagram. The log-magnitude–phase locus of

$$GH_1(j\omega) = \frac{1}{j\omega(j\omega + 1)(0.2j\omega + 1)} \tag{8.51}$$

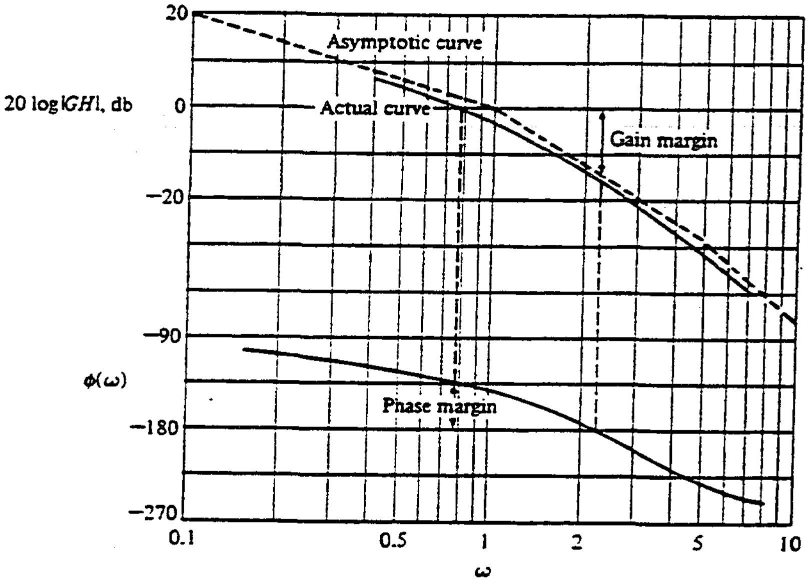


Figure 8.16. Bode diagram for $GH_1(j\omega) = 1/j\omega(j\omega + 1)(0.2j\omega + 1)$.

is shown in Fig. 8.17. The indicated phase margin is 43° and the gain margin is 15 db. For comparison, the locus for

$$GH_2(j\omega) = \frac{1}{j\omega(j\omega + 1)^2} \tag{8.52}$$

is also shown in Fig. 8.17. The gain margin for GH_2 is equal to 5.7 db, and the phase margin for GH_2 is equal to 20°. Clearly, the feedback system $GH_2(j\omega)$ is relatively less stable than the system $GH_1(j\omega)$. However, the question still remains: How much less stable is the system $GH_2(j\omega)$ in comparison to the system $GH_1(j\omega)$? In the following paragraph we shall answer this question for a second-order system, and the usefulness of the relation that we develop will depend upon the presence of dominant roots.

Let us now determine the phase margin of a second-order system and relate the phase margin to the damping ratio ζ of an underdamped system. Consider the loop-transfer function

$$GH(j\omega) = \frac{\omega_n^2}{j\omega(j\omega + 2\zeta\omega_n)}. \tag{8.53}$$

The characteristic equation for this second-order system is

$$s^2 + 2\zeta\omega_n s + \omega_n^2 = 0.$$

Therefore the closed-loop roots are

$$s = -\zeta\omega_n \pm j\omega_n\sqrt{1-\zeta^2}.$$

The magnitude of the frequency response is equal to 1 at a frequency ω_c, and thus

$$\frac{\omega_n^2}{\omega_c(\omega_c^2 + 4\zeta^2\omega_n^2)^{1/2}} = 1. \tag{8.54}$$

Rearranging Eq. (8.54), we obtain

$$(\omega_c^2)^2 + 4\zeta^2\omega_n^2(\omega_c^2) - \omega_n^4 = 0. \tag{8.55}$$

Solving for ω_c, we find that

$$\frac{\omega_c^2}{\omega_n^2} = (4\zeta^4 + 1)^{1/2} - 2\zeta^2. \tag{8.56}$$

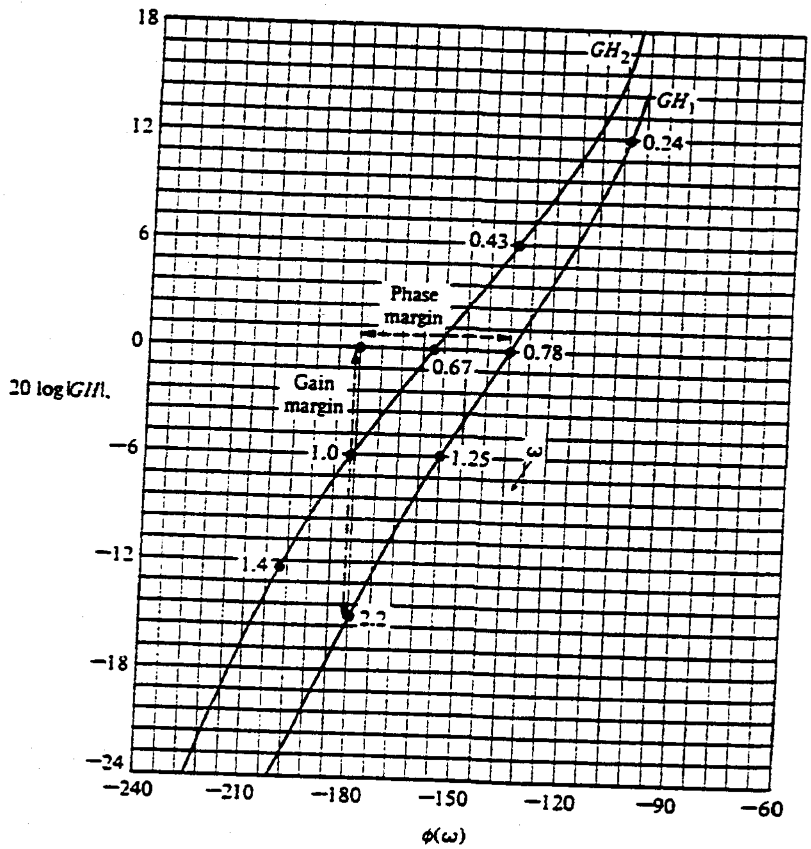


Figure 8.17. Log-magnitude–phase curve for GH_1 and GH_2.

The phase margin for this system is

$$\begin{aligned}\phi_{pm} &= 180^\circ - 90^\circ - \tan^{-1}\left(\frac{\omega_c}{2\zeta\omega_n}\right)\\ &= 90^\circ - \tan^{-1}\left(\frac{1}{2\zeta}[4\zeta^4 + 1)^{1/2} - 2\zeta^2]^{1/2}\right) \qquad (8.57)\\ &= \tan^{-1}\left(2\zeta\left[\frac{1}{(4\zeta^4 + 1)^{1/2} - 2\zeta^2}\right]^{1/2}\right).\end{aligned}$$

Equation (8.57) is the relationship between the damping ratio ζ and the phase margin ϕ_{pm} that provides a correlation between the frequency response and the time response. A plot of ζ versus ϕ_{pm} is shown in Fig. 8.18. The actual curve of ζ versus ϕ_{pm} can be approximated by the dashed line shown in Fig. 8.18. The slope of the linear approximation is equal to 0.01, and therefore an approximate linear relationship between the damping ratio and the phase margin is

$$\zeta = 0.01\phi_{pm}, \qquad (8.58)$$

where the phase margin is measured in degrees. This approximation is reasonably accurate for $\zeta \leq 0.7$, and is a useful index for correlating the frequency response with the transient performance of a system. Equation (8.58) is a suitable approximation for a second-order system and may be used for higher-order systems if one can assume that the transient response of the system is primarily due to a

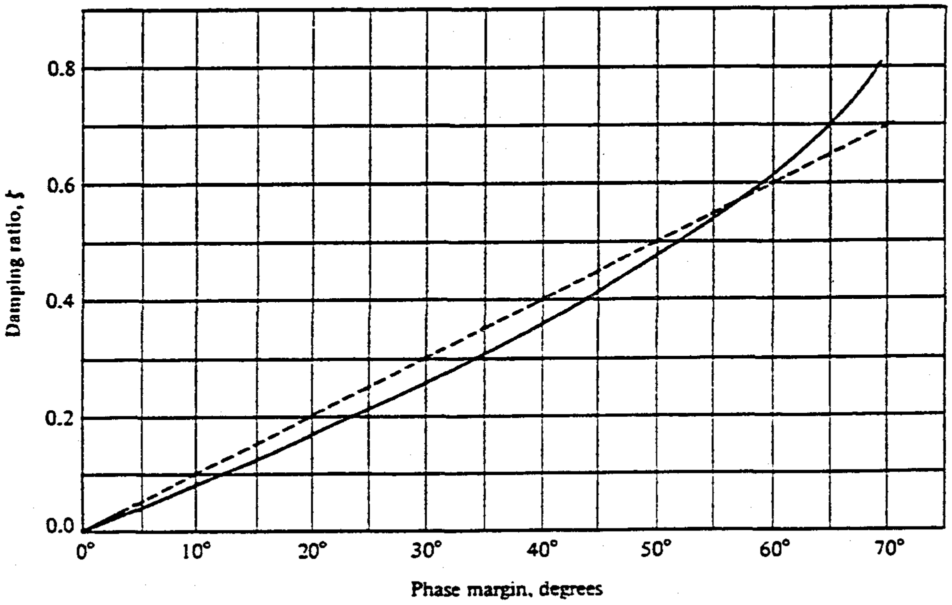


Figure 8.18. Damping ratio vs. phase margin for a second-order system.

pair of dominant underdamped roots. The approximation of a higher-order system by a dominant second-order system is a useful approximation indeed! Although it must be used with care, control engineers find this approach to be a simple, yet fairly accurate, technique of setting the specifications of a control system.

Therefore, for the system with a loop-transfer function

$$GH(j\omega) = \frac{1}{j\omega(j\omega + 1)(0.2j\omega + 1)}, \tag{8.59}$$

we found that the phase margin was 43°, as shown in Fig. 8.16. Thus the damping ratio is approximately

$$\zeta \simeq 0.01\phi_{pm} = 0.43. \tag{8.60}$$

Then the peak response to a step input for this system is approximately

$$M_{pt} = 1.22 \tag{8.61}$$

as obtained from Fig. 4.8 for $\zeta = 0.43$.

The phase margin of a system is a quite suitable frequency response measure for indicating the expected transient performance of a system. Another useful index of performance in the frequency domain is M_{p_ω}, the maximum magnitude of the closed-loop frequency response, and we shall now consider this practical index.

8.5 The Closed-Loop Frequency Response

The transient performance of a feedback system can be estimated from the closed-loop frequency response. The *closed-loop frequency response* is the frequency response of the closed-loop transfer function $T(j\omega)$. The open- and closed-loop frequency responses for a single-loop system are related as follows:

$$\frac{C(j\omega)}{R(j\omega)} = T(j\omega) = \frac{G(j\omega)}{1 + GH(j\omega)}. \tag{8.62}$$

The Nyquist criterion and the phase margin index are defined for the open-loop transfer function $GH(j\omega)$. However, as we found in Section 7.2, the maximum magnitude of the closed-loop frequency response can be related to the damping ratio of a second-order system of

$$M_{p_\omega} = |G(\omega_r)| = (2\zeta\sqrt{1 - \zeta^2})^{-1}, \qquad \zeta < 0.707. \tag{8.63}$$

This relation is graphically portrayed in Fig. 7.10. Because this relationship between the closed-loop frequency response and the transient response is a useful relationship, we would like to be able to determine M_{p_ω} from the plots completed

REF. EE

LASERS

PETER W. MILONNI
Los Alamos National Laboratory
Los Alamos, New Mexico

JOSEPH H. EBERLY
University of Rochester
Rochester, New York

WILEY

A WILEY-INTERSCIENCE PUBLICATION

JOHN WILEY & SONS

New York / Chichester / Brisbane / Toronto / Singapore

Copyright © 1988 by John Wiley & Sons, Inc.

All rights reserved. Published simultaneously in Canada.

Reproduction or translation of any part of this work beyond that permitted by Section 107 or 108 of the 1976 United States Copyright Act without the permission of the copyright owner is unlawful. Requests for permission or further information should be addressed to the Permissions Department, John Wiley & Sons, Inc.

Library of Congress Cataloging in Publication Data:

Milonni, Peter W.
Lasers / Peter W. Milonni and Joseph H. Eberly.
p. cm.

"A Wiley-Interscience publication."
Bibliography: p.
ISBN 0-471-62731-3
1. Lasers. I. Eberly, Joseph H. II. Title.

QC688.M55 1988
621.36'6—dc19
87-26347
CIP

Printed in the United States of America

10 9 8 7 6 5 4

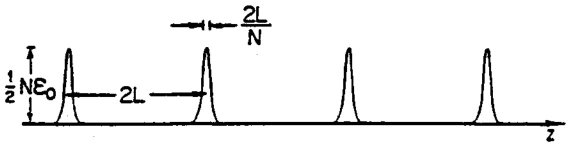


Figure 12.10 A mode-locked pulse train as a function of coordinate z, observed at a fixed instant of time.

The fact that the pulses of a mode-locked train are separated in time by the round-trip cavity transit time $2L/c$ suggests a "bouncing-ball" picture of a mode-locked laser: we can regard the mode locking as generating a pulse of duration $2L/cN$, and this pulse keeps bouncing back and forth between the cavity mirrors. Focusing our attention on a particular plane of constant z in the resonator, we observe a train of identical pulses moving in either direction.

In most lasers the phases ϕ_n of the different modes will undergo random and uncorrelated variations in time. In this case the total intensity is the sum of the individual mode intensities. In mode-locked lasers, however, the mode phases are correlated and the total intensity is not simply the sum of the individual mode intensities. In fact the individual pulses in the mode-locked train have an intensity N times larger than the sum of the individual mode intensities. The average power, however, is essentially unaltered by mode-locking the laser (Problem 12.6).

• Before discussing how mode locking can be accomplished, it is worth noting that "phase locking" or "synchronization" phenomena occur in many *nonlinear* oscillatory systems besides lasers, and indeed these phenomena have been known for a very long time. C. Huygens (1629–1695), for instance, observed that two pendulum clocks hung a few feet apart on a thin wall tend to have their periods synchronized as a result of their small coupling via the vibrations of the wall. Near the end of the nineteenth century Lord Rayleigh found that two organ pipes of slightly different resonance frequencies will vibrate at the same frequency when they are sufficiently close together.

The contractive pulsations of the heart's muscle cells become phase-locked during the development of the fetus. Fibrillation of the heart occurs when they get out of phase for some reason, and results in death unless the heart can be shocked back into the normal condition of cell synchronization. There are other biological examples of phase locking, but detailed theoretical analyses are obviously extremely difficult or impossible for such complex systems. Modern applications of synchronization principles are made in high-precision motors and control systems. •

12.9 AM MODE LOCKING

The process by which phase or mode locking is forced upon a laser is fundamentally a nonlinear one, and a rigorous analysis of it is complicated. We will therefore rely largely on semiquantitative explanations.

Consider again the scalar electric field

$$E_m(z, t) = \mathcal{E}_m \sin k_m z \sin (\omega_m t + \phi_m) \tag{12.9.1}$$

associated with a longitudinal mode. Suppose that the amplitude $\mathcal{E}_m$ is not constant but rather is modulated periodically in time according to the formula

$$\mathcal{E}_m = \mathcal{E}_0(1 + \epsilon \cos \Omega t) \tag{12.9.2}$$

where Ω is the modulation frequency and $\mathcal{E}_0$ and ϵ are constants. Thus we have an amplitude-modulated field

$$E_m(z, t) = \mathcal{E}_0(1 + \epsilon \cos \Omega t) \sin (\omega_m t + \phi_m) \sin k_m z \tag{12.9.3}$$

Since

$$\begin{aligned}\cos \Omega t \sin (\omega_m t + \phi_m) = \tfrac{1}{2} \sin (\omega_m t + \phi_m + \Omega t)&\\ + \tfrac{1}{2} \sin (\omega_m t + \phi_m - \Omega t)&\end{aligned} \tag{12.9.4}$$

we can write the field (12.9.3) as a sum of harmonically varying parts:

$$\begin{aligned}E_m(z, t) = \mathcal{E}_0\Big\{\sin (\omega_m t + \phi_m) + \frac{\epsilon}{2} \sin \big[(\omega_m + \Omega)t + \phi_m\big]&\\ + \frac{\epsilon}{2} \sin \big[(\omega_m - \Omega)t + \phi_m\big]\Big\} \sin k_m z&\end{aligned} \tag{12.9.5}$$

The frequency spectrum of the field (12.9.5) is shown in Figure 12.11. The amplitude modulation of the field (12.9.1) of frequency ω_m has generated *sidebands* of frequency $\omega_m \pm \Omega$. These sidebands are displaced from the *carrier frequency* ω_m by precisely the modulation frequency Ω. Sideband generation is a well-known consequence of amplitude modulation.

In a laser the mode amplitudes $\mathcal{E}_m$ are determined by the condition that the gain equals the loss. If the loss (or gain) is periodically modulated at a frequency Ω, we expect the fields $E_m(z, t)$ associated with the various modes to be amplitude-modulated (AM) with this frequency. In other words, we expect sidebands to be generated about each mode frequency ω_m, as in (12.9.5). In particular, if the modulation frequency Ω is equal to the mode frequency spacing

$$\Delta = \omega_{m+1} - \omega_m = \pi c/L \tag{12.9.6}$$

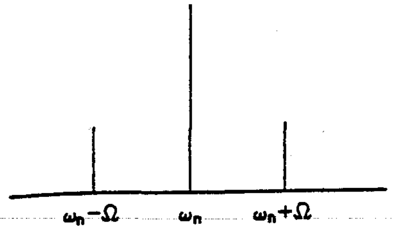


Figure 12.11 Frequency spectrum of the amplitude modulated field (12.9.5). The sidebands at $\omega_n \pm \Omega$ have amplitudes $\epsilon/2$ times as large as the carrier amplitude at ω_n. In this case $\epsilon/2 < 1$.

the sidebands associated with each mode match exactly the frequencies of the two adjacent modes (Figure 12.12). In this case each mode becomes strongly coupled to its nearest-neighbor modes, and it turns out that there is a tendency for the modes to lock together in phase. Loss or gain modulation at the mode separation frequency Δ is therefore one way of mode locking. Borrowing terminology from radio engineering, we call this *AM mode locking*.

The dimensionless factor ϵ appearing in (12.9.2) is called the modulation index. It is usually small, but it must be large enough to couple the different modes sufficiently strongly. This is analogous to the synchronization phenomenon observed in the 17th century by Huygens with pendulum clocks. Their frequencies were locked together when the clocks were mounted just a meter or so apart, but larger separations weakened their coupling and destroyed the locking effect. If ϵ is too large, on the other hand, the locking effect is also weakened. This is analogous to the distortion arising in AM radio electronic systems when the carrier wave is "overmodulated," i.e., when $\epsilon > 1$. (See also Problem 12.7.)

A heuristic way to understand why AM mode locking occurs in lasers is first to suppose that lasing can occur only in brief intervals when the periodically modulated loss is at a minimum. These minima occur in time intervals of $T = 2\pi/\Delta = 2L/c$ if the modulation frequency $\Omega = \Delta$. Between these times of minimum loss the loss is too large for laser oscillation. Thus we can have laser oscillation only if it is possible to generate a train of short pulses separated in time by T. This is possible if the modes lock together and act in unison, for then we generate a mode-locked train of pulses separated by time T. Thus mode locking has been described as a kind of "survival of the fittest" phenomenon.

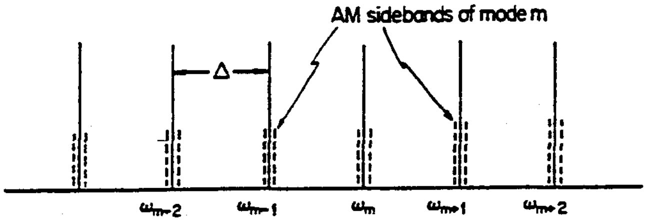


Figure 12.12 Longitudinal modes amplitude-modulated at the frequency Δ equal to their spacing. For clarity the AM sidebands are indicated as dashed lines slightly dispaced from the mode frequencies ω_m.

12.10 FM MODE LOCKING

We will now consider the case where the *phase* of the field (12.9.1) is periodically modulated rather than the amplitude:

$$E_m(z, t) = \mathcal{E}_m \sin k_m z \sin (\omega_m t + \phi_m + \delta \cos \Omega t) \qquad (12.10.1)$$

The dimensionless constant δ gives the amplitude of the modulation of frequency Ω. As in the case of amplitude modulation, this phase modulation gives rise to sideband frequencies about the carrier frequency ω_m. As we will now see, however, the phase modulation produces a whole series of sidebands.

The time-dependent part of (12.10.1) may be written as

$$\sin (\omega_m t + \phi_m + \delta \cos \Omega t) = \sin (\omega_m t + \phi_m) \cos (\delta \cos \Omega t) + \cos (\omega_m t + \phi_m) \sin (\delta \cos \Omega t) \qquad (12.10.2)$$

Now we make use of two mathematical identities:[2]

$$\cos (x \cos \theta) = J_0(x) + 2 \sum_{k=1}^{\infty} (-1)^k J_{2k}(x) \cos (2k\theta) \qquad (12.10.3a)$$

and

$$\sin (x \cos \theta) = 2 \sum_{k=0}^{\infty} (-1)^k J_{2k+1}(x) \cos \left[(2k + 1)\theta\right] \qquad (12.10.3b)$$

where $J_n(x)$ is the Bessel function of the first kind of order n. The first few lowest-order Bessel functions are plotted in Figure 12.13. These plots are all we will need

[2] See, for example M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions* (Dover, New York, 1971), formulas 9.1.44 and 9.1.45.

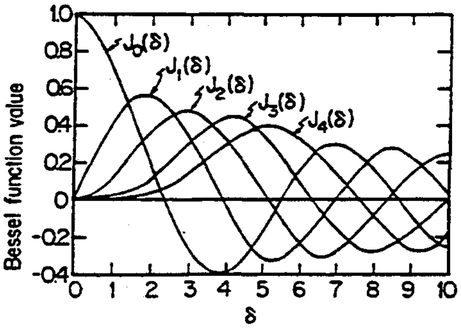


Figure 12.13 The first few lowest-order Bessel functions of the first kind, $J_n(\delta)$.

to know about them. The functions (12.10.3) appear in (12.10.2) with $x = \delta$ and $\theta = \Omega t$. Thus

$$\begin{aligned}
&\sin(\omega_m t + \phi_m + \delta\cos\Omega t) \\
&\quad = \sin(\omega_m t + \phi_m)\left(J_0(\delta) + 2\sum_{k=1}^{\infty}(-1)^k J_{2k}(\delta)\cos(2k\Omega t)\right) \\
&\qquad + 2\cos(\omega_m t + \phi_m)\sum_{k=0}^{\infty}(-1)^k J_{2k+1}(\delta)\cos[(2k+1)\Omega t] \\
&\quad = \sin(\omega_m t + \phi_m)[J_0(\delta) - 2J_2(\delta)\cos 2\Omega t \\
&\qquad + J_4(\delta)\cos 4\Omega t - 2J_6(\delta)\cos 6\Omega t + \cdots] \\
&\qquad + 2\cos(\omega_m t + \phi_m)[J_1(\delta)\cos\Omega t - J_3(\delta)\cos 3\Omega t \\
&\qquad + J_5(\delta)\cos 5\Omega t - \ldots]
\end{aligned} \tag{12.10.4}$$

Using the identities

$$\sin x\cos y = \tfrac{1}{2}[\sin(x+y) + \sin(x-y)]$$

$$\cos x\cos y = \tfrac{1}{2}[\cos(x+y) + \cos(x-y)]$$

therefore, we have

$$\begin{aligned}
&\sin(\omega_m t + \phi_m + \delta\cos\Omega t) \\
&\quad = J_0(\delta)\sin(\omega_m t + \phi_m) \\
&\qquad + J_1(\delta)\left\{\cos[(\omega_m + \Omega)t + \phi_m] + \cos[(\omega_m - \Omega)t + \phi_m]\right\} \\
&\qquad - J_2(\delta)\left\{\sin[(\omega_m + 2\Omega)t + \phi_m] + \sin[(\omega_m - 2\Omega)t + \phi_m]\right\} \\
&\qquad - J_3(\delta)\left\{\cos[(\omega_m + 3\Omega)t + \phi_m] + \cos[(\omega_m - 3\Omega)t + \phi_m]\right\} \\
&\qquad + J_4(\delta)\left\{\sin[(\omega_m + 4\Omega)t + \phi_m] + \sin[(\omega_m - 4\Omega)t + \phi_m]\right\} \\
&\qquad + J_5(\delta)\left\{\cos[(\omega_m + 5\Omega)t + \phi_m] + \cos[(\omega_m - 5\Omega)t + \phi_m]\right\} \\
&\qquad - \ldots
\end{aligned} \tag{12.10.5}$$

after a simple rearrangement of terms in (12.10.4).

Whereas amplitude modulation produces one sideband on either side of the carrier frequency ω_m, phase modulation in general produces a whole series of pairs of sidebands. If the "modulation index" δ is somewhat less than unity, however, we observe from (12.10.5) and Figure 12.13 that the first pair of sidebands

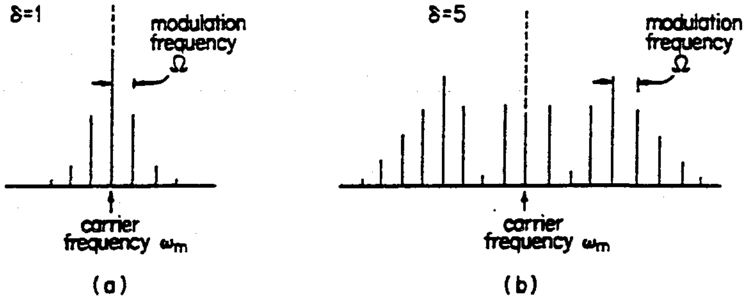


Figure 12.14 Frequency spectrum of the function (12.10.5) for the modulation index (*a*) $\delta = 1$ and (*b*) $\delta = 5$.

at $\omega_m \pm \Omega$ is strongest. As the strength of the modulation increases, i.e., as δ increases, more sideband pairs become important. Figure 12.14 shows the frequency spectrum of the function (12.10.5) for $\delta = 1$ and $\delta = 5$.

Again borrowing the terminology of radio engineering, we refer to this type of modulation as *frequency modulation* (FM). As in the AM case, frequency modulation at the mode separation frequency $\Omega = \Delta = \pi c/L$ causes the sidebands associated with each mode to be in resonance with the carrier frequencies of other modes. This results in a strong coupling of these modes and a tendency for them to lock together and produce a mode-locked train of pulses. This is called *FM mode locking*.

• Information cannot be transmitted with a purely monochromatic wave. The basic idea of radio communication is to modulate a monochromatic (carrier) wave in some way (AM or FM), transmit it, then demodulate it at a receiver to recover the information contained in the original modulation. In the AM case the sidebands imposed on the carrier wave are displaced from the carrier by an amount equal to the modulation frequency, independently of the modulation index ϵ. In the FM case, on the other hand, the "width" of the modulation about the carrier is directly proportional to the corresponding index δ, approximately independently of the modulation frequency Ω. This makes FM transmission less susceptible to interference from extraneous sources (lightning, electric power generators, etc.) than AM if its modulation index is large. At the same time, there is a disadvantage to FM in that the amplifiers in the transmitter and receiver must have large bandwidths in order to pick up a good portion of the sideband spectrum. A large bandwidth is most easily obtained at higher carrier frequencies; this is analogous to the fact that the bandwidth of a laser cavity increases with frequency if the cavity Q is held constant [Eq. (11.9.21)], and explains why FM radio stations broadcast at higher frequencies than AM stations (Problem 2.9). •

12.11 METHODS OF MODE LOCKING

Lasers can be mode-locked in a variety of ways. We will focus our attention on three common and illustrative techniques.

External modulation technique for sensitive interferometric detection of displacements

C.N. Man, D. Shoemaker [1], M. Pham Tu and D. Dewey [2]
Centre National de la Recherche Scientifique, Groupe de Recherches sur les Ondes de Gravitation,
Bât. 104, Université Paris-Sud, 91405 Orsay, France

Received 3 May 1990; accepted for publication 28 May 1990
Communicated by J.P. Vigier

We present and demonstrate experimentally a novel modulation technique for the shot-noise limited measurement of the differential displacement between arms of a Michelson interferometer. In this scheme electro-optic modulation is applied to a reference beam *external* to the Michelson system, avoiding modulator-induced losses and wave front distortions, and allowing efficient power increase through recycled-light operation.

1. Introduction

Michelson interferometers allows measurement of differential displacements to $\lambda/10^9$ and better; this high sensitivity is being exploited in gravitational-wave (GW) detection systems [1]. Typically, electro-optic modulators *internal* to the Michelson arms are used to move the measurement band to high frequencies, hold the interferometer output on a dark fringe, and yield a shot-noise limited measurement. Next generation long-baseline GW detectors will incorporate schemes to achieve greater sensitivity through increased power. A promising method is *recycling* in which the bright-fringe output port is recycled back into the interferometer input [2,3]; the resulting Fabry–Perot cavity can produce stored power gains in excess of 30 for realistic conditions.

In the context of these recycled GW detectors the use of internal modulators is undesirable for several reasons. They introduce wave front distortions which reduce the interferometer contrast. In addition the (Pockels cell) modulators may have losses of several percent which can limit the achievable recycling gain. Finally, large scale systems will have large diameter and/or high power-density beams; fabricating suitable modulators may be difficult.

With the above motivation we present the external modulation scheme [#1] and verify its operation in a (simple prototype) recycled interferometric system.

2. Simplified theory

We describe the modulation scheme in an intuitive framework giving straightforward equations which will be applied to our experimental results. We present a simplified description in which all beams are considered to be in the TEM_{00} mode, and we make no attempt at a global optimization, e.g., the effects of reference beam intensity and recycling parameters on overall performance.

2.1. Internal modulation

Fig. 1 shows the arrangement of the "classical" internal modulation interferometric system [6,7]. The input laser light is split by a (nominally 50–50) beam splitter and is interferometrically recombined after

[1] Present address: Room 20F-001, MIT, 77 Massachusetts Avenue, Cambridge, MA 02139, USA.

[2] Present address: Room 37-662D, MIT, 77 Massachusetts Avenue, Cambridge, MA 02139, USA.

[#1] In the context of microwave spectroscopy, see e.g., ref. [4]; in the present context, see ref. [5].

0375-9601/90/$ 03.50 © 1990 – Elsevier Science Publishers B.V. (North-Holland)

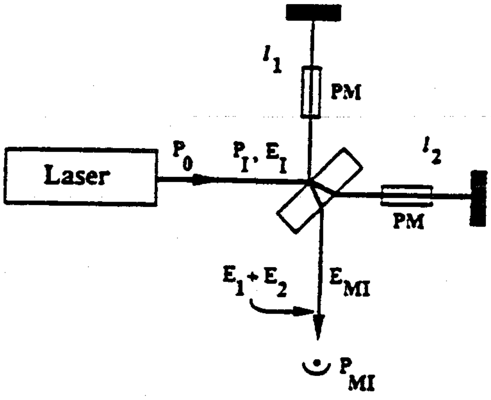


Fig. 1. Internal modulation scheme. Phase modulators PM internal to the interferometer arms provide fringe interrogation.

passage through the interferometer arms; here l_1 and l_2 are the total equivalent path lengths of the two arms which may be many times the physical separation between the beam-splitter and mirrors due to use of delay lines [6] or Fabry–Perot cavities [8]. The goal of the detection system is to measure changes in the path length difference, l_1-l_2. Because of low-frequency (<5 MHz) laser amplitude fluctuations above those due to shot noise, a modulation scheme [6] is typically employed to move the measurement to a higher frequency (10 MHz); the (electro-optic) phase-modulating crystals PM shown are used for this purpose and are typically sinusoidally modulated in anti-phase.

The operation of the internal modulation system can be understood through the use of simple phasor diagrams in which a beam given by $EF(x, y)e^{ikz-i\omega t}$ is represented by a vector of length $|E|$ and angle kz, with $k=2\pi/\lambda$. The assumption is that in a given plane (of the photodetector say) all beams have the same (complex) spatial form, $F(x, y)$, and time dependence, $e^{-i\omega t}$; thus the phasor parameters can uniquely specify the beam. In fig. 2 we show such a phasor diagram representing the electric fields at the Michelson output. E_1 and E_2 are phasors representing the output fields from the two arms of the interferometer, $E_1e^{ikl_1}$, $E_2e^{ikl_2}$; their vector sum is represented by E_{MI}, making an angle of Φ_D with respect to the average phase $\Phi_C=k(l_1+l_2)/2$. The dotted line is the locus of E_{MI} as a function of the path length difference $\phi_0=k(l_1-l_2)$. Note that the minimum amplitude of E_{MI} (the interferometer dark fringe) occurs at $\phi_0=\pi$ and at this point the phase Φ_D changes rapidly with the path length difference; here Φ_{MI}, the interferometer output phase, is clearly the sum of Φ_C (the average phase) and Φ_D.

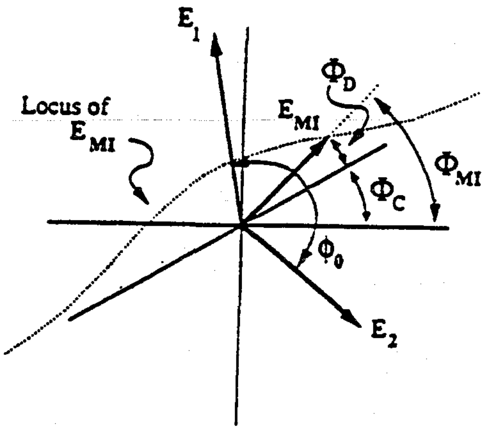


Fig. 2. Phasor diagram for Michelson interferometer output field E_{MI}. The locus of points swept out by E_{MI} is shown as a function of path length difference $\phi_0=k(l_1-l_2)$; the output phase Φ_{MI} is the sum of a common mode phase $\Phi_C=k(l_1+l_2)/2$ and the difference phase Φ_D.

Calculating the output field gives

$$E_{MI}=E_{MI}e^{i\Phi_{MI}}, \qquad (1)$$

where

$$E_{MI}^2=(E_1^2+E_2^2)\{1+C_{00}\cos[k(l_1-l_2)]\} \qquad (2)$$

and

$$\Phi_{MI}=\Phi_C+\Phi_D=k(l_1+l_2)/2 + \arctan\left(\frac{E_1-E_2}{E_1+E_2}\tan[k(l_1-l_2)/2]\right). \qquad (3)$$

The contrast $C_{00}=2E_1E_2/(E_1^2+E_2^2)$ implies a contrast model in which non-unity contrast is due to beam intensity mismatch. In reality other defects contribute to non-ideal contrast, perhaps the largest among them being wavefront distortion, the effect of which is to add a non-interfering constant component to the light intensity. This term is easily added to our equations and is used when comparing experimental results.

For near equal intensities, $E_1 \approx E_2$, and operating around a dark fringe, $\phi_0 \equiv k(l_1-l_2)=\pi+k\Delta l$, the amplitude E_{MI} is roughly constant:

$$E_{\mathrm{MI}}^2=(E_1^2+E_2^2)[1-C_{00}\cos(k\Delta l)] \tag{4}$$

and shows a quadratic variation with changes in Δl. The phase

$$\Phi_{\mathrm{MI}}=k(l_1+l_2)/2+\frac{E_1+E_2}{E_1-E_2}k\Delta l/2 \tag{5}$$

changes rapidly with Δl due to the amplifying factor $(E_1+E_2)/(E_1-E_2)\approx\sqrt{2/(1-C_{00})}$ and less rapidly with the common mode phase $k(l_1+l_2)$.

With the addition of anti-phase modulation between the two arms,

$$\phi_{\mathrm{mod}}=\phi_0+m\cos(\Omega t)\,, \tag{6}$$

the amplitude E_{MI} is modulated (quadratically); when ϕ_0 is equal to π, symmetry results in an output signal (detected power) which has no component at Ω. Changes in the path-length difference Δl result in an asymmetry and thus produce a signal component at Ω. Note that because the photo-detection process is phase-insensitive, common mode arm length changes (which rotate the output field in this picture) do not affect or appear in the internal modulation scheme output.

With the addition of the relative phase modulation of eq. (6) the output power can be expressed as a sum of terms at DC and the harmonics of Ω; keeping only the leading terms, we have

$$E_{\mathrm{MI}}^2=S_{\mathrm{DC}}+S_\Omega\sin(\Omega t)\,, \tag{7}$$

with

$$S_{\mathrm{DC}}=E_1^2+E_2^2+2E_1E_2J_0(m)\cos(\phi_0)\,, \tag{8}$$

$$S_\Omega=4E_1E_2J_1(m)\sin(\phi_0)\,. \tag{9}$$

For the case when $E_1=E_2=E_0/\ \ 2$ $(C_{00}=1)$, and operating around the dark fringe $(\phi_0=k(l_1-l_2)=\pi+k\Delta l)$, the detected photocurrent is

$$I_{\mathrm{DC}}^{\mathrm{int}}=\tfrac{1}{2}I_0[1-J_0(m)]\,, \tag{10}$$

$$I_\Omega^{\mathrm{int}}=\frac{2\pi}{\lambda}I_0J_1(m)\Delta l\,, \tag{11}$$

where I_0 is the photocurrent due to the power $E_0^2=(E_1+E_2)^2$ incident on the photodetector. The signal at Ω is demodulated and measured in the presence of the shot-noise due to I_{DC} (expressed as a linear spectral density i_{noise} in units of $\mathrm{A}/\sqrt{\mathrm{Hz}}$) which is given by

$$i_{\mathrm{noise}}=\sqrt{2}\sqrt{2eI_{\mathrm{DC}}^{\mathrm{int}}}\,, \tag{12}$$

where the pre-factor of $\sqrt{2}$ is a result of the demodulation process. This current noise can be expressed as an equivalent displacement noise (in units of $\mathrm{m}/\sqrt{\mathrm{Hz}}$) using the sensitivity $\mathrm{d}I_\Omega^{\mathrm{int}}/\mathrm{d}\Delta l$ from eq. (11), giving

$$\Delta l_{\mathrm{int}}=\frac{\sqrt{2}\sqrt{2eI_{\mathrm{DC}}^{\mathrm{int}}}}{\mathrm{d}I_\Omega^{\mathrm{int}}/\mathrm{d}\Delta l}=\frac{\sqrt{2eI_0[1-J_0(m)]}}{(2\pi/\lambda)I_0J_1(m)}\,. \tag{13}$$

Thus, the signal-to-noise ratio is a function of m; and in the case of perfect contrast $(C_{00}\rightarrow 1)$, the optimum value of the modulation tends toward 0 $(m\rightarrow 0)$, and we have a limiting displacement sensitivity of

$$\Delta l_{\mathrm{MI}}=\frac{\lambda}{2\pi}\sqrt{\frac{2e}{I_0}}\,. \tag{14}$$

Note that I_0 is the photocurrent equivalent of the input power E_0^2 to the interferometer. This shot-noise-limited sensitivity is the standard for comparison with other modulation schemes.

2.2. *External modulation*

In external modulation operation, the Michelson output field is brought to interference with a laser-derived reference beam through the Mach–Zehnder optical arrangement shown in fig. 3 resulting in the phasor diagram of fig. 4. The Michelson output field, E_{MI}, is near zero when on a dark fringe. For deviations in the path length difference, the amplitude and phase of E_{MI} change. Interference between this non-zero Michelson output field and the reference beam produces a signal sensitivite to changes in E_{MI} and hence path length difference. Note that variations in the common-mode length of the Michelson arms rotate E_{MI} in the phasor diagram and thus must be controlled to maintain a constant operating phase ϕ_0^{MZ}, which is the difference in phase between the Michelson output and the reference beam at the Mach–Zehnder beam splitter. As in the internal modulation case , the relative phase, $\phi_{\mathrm{mod}}^{\mathrm{MZ}}=\phi_0^{\mathrm{MZ}}+m\cos(\Omega t)$, is modulated at a high fre-

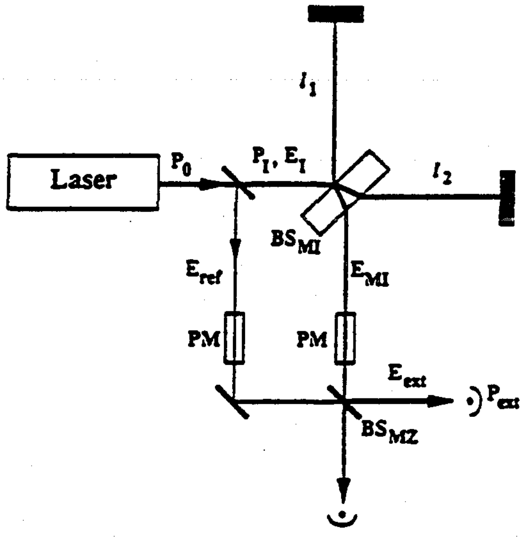


Fig. 3. External modulation scheme. The phase modulators have been removed from the arms of the Michelson interferometer; instead, a reference beam is phase modulated and is brought into interference with the Michelson output beam.

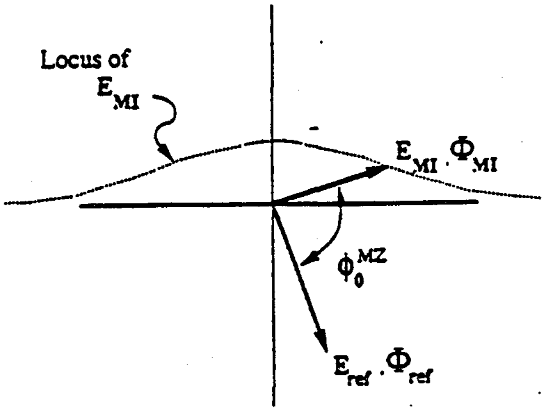


Fig. 4. Phasor diagram for external modulation showing the phase relationship between the Michelson output and the reference beam.

quency, Ω, and the photocurrent component at Ω is the desired measure of E_{MI} and thus path length changes.

Quantitatively, the case of external modulation can be easily calculated using the modulation eqs. (7)–(9) with the following identifications:

$$E_1^{MZ} = |E_{ref}|, \quad E_2^{MZ} = |E_{MI}|,$$

$$\phi_0^{MZ} \equiv \Phi_{ref} - \Phi_{MI}$$

$$= \Phi_{ref} - k(l_1 + l_2)/2 - \frac{E_1 + E_2}{E_1 - E_2} k\Delta l/2 , \qquad (15)$$

where we have assumed that the Michelson operates around a dark fringe and that all of the Michelson output light appears at the Mach–Zehnder output. The resulting output at DC and Ω are then

$$S_{DC}^{ext} = E_{ref}^2 + E_{MI}^2 + 2E_{ref}E_{MI}J_0(m)\cos(\phi_0^{MZ}) ,$$

$$S_{\Omega}^{ext} = 4E_{ref}E_{MI}k\Delta l\, J_1(m)\sin(\phi_0^{MZ}) . \qquad (16)$$

Operating about $\phi_0^{MZ} = \Phi_{ref} - k(l_1 + l_1)/2 = \pi$ gives maximum sensitivity of the signal to the displacement $k\Delta l$. In analogy to eqs. (10),(11) the photodetector currents can be written as

$$I_{DC}^{ext} = I_{ref} + I_{MI} - 2\sqrt{I_{ref}I_{MI}}\, J_0(m) ,$$

$$I_{\Omega}^{ext} = \frac{2\pi}{\lambda} 4\sqrt{I_{ref}I_{MI}}\, J_1(m) \frac{E_1 + E_2}{E_1 - E_2} \frac{\Delta l}{2}$$

$$= \frac{2\pi}{\lambda} 2\sqrt{I_{ref}I_0}\, J_1(m)\Delta l , \qquad (17)$$

where it is interesting to note that we have used $\sqrt{I_{MI}}\,(E_1 + E_2)/(E_1 - E_2) = \sqrt{I_0}$ to eliminate the apparent contrast dependence of (17).

The resulting shot-noise limited displacement sensitivity of the external modulation scheme is thus

$$\Delta l_{ext} = \frac{\sqrt{2}\sqrt{2eI_{DC}^{ext}}}{(2\pi/\lambda)2\sqrt{I_{ref}I_0}\, J_1(m)}$$

$$= \frac{\lambda}{2\pi} \frac{\sqrt{2I_{DC}^{ext}}\ \sqrt{e}}{\sqrt{I_{ref}I_0}\, J_1(m)} . \qquad (18)$$

Note that I_{ref} and I_0 above correspond to the values measured *after* the Mach–Zehnder beam splitter. Also I_0 above is equal to that of eq. (14) for similar Michelson interferometers with the same input power in each scheme (assuming that I_{ref} is negligible compared to I_0 and that all of the light from the Michelson output is detected). Thus, the sensitivities for the two techniques can be directly compared. Two cases are of particular interest. In the first case, when I_{DC}^{ext} is dominated by the terms of eq. (17), the optimum intensity of the reference beam is the Michelson output intensity ($I_{ref} = I_{MI}$) and, as in the case of internal modulation, the optimum modulation tends towards zero ($m \to 0$). This leads to a sensitivity equal to that of internal modulation; to realize this experimentally an asymmetric Mach–Zehnder beam splitter could be used (i.e., 99/1) with the output taken at the high-transmission port for the Michelson beam. In the second case (relevant to the experiments carried out here) $I^{\text{ext}}_{\text{DC}}$ is dominated by I_{ref}, which is much larger than I_{MI} and other terms, e.g., non-TEM$_{00}$ light in the beams. In this case the maximum value of $J_1(m)$ (≈ 0.582 for $m \approx 1.80$ radians) gives the external scheme an increased noise level by a factor of 1.2 over internal modulation. (This is provided that all light out of the Michelson is utilized; in this case both output ports of a 50/50 Mach–Zehnder beam splitter can be used.)

3. Experimental verification

In this section we describe the experimental setup used to verify external modulation operation, the initial tests performed to calibrate and verify operation of the locking and demodulation electronics systems, the measurement procedure for taking data, and our results with a recycled externally modulated system.

3.1. Experimental setup

The experiment is built on two optical tables as shown in fig. 5. A granite table holds the laser and associated optics; the interferometer is built up of standard mirror mounts on a mechanically isolated, suspended aluminum platform in a vacuum tank (which is closed but not evacuated for our measurements).

The argon-ion laser (Spectra-Physics 171) is stabilized in frequency by the Pound–Drever technique [9] to a Fabry–Perot reference cavity FP. The control elements are fast and slow piezoelectric transducers (PZTs) on the two laser cavity mirrors [10], giving a unity-gain frequency of 60 kHz, and a suppression of frequency fluctuations of 40 dB at 10 kHz. The laser is isolated from the rest of the experiment by a Faraday isolator FI1 and an acousto-optic modulator AO. The light is carried via a single-mode fiber (to suppress beam pointing fluctuations) to the interferometer optical table.

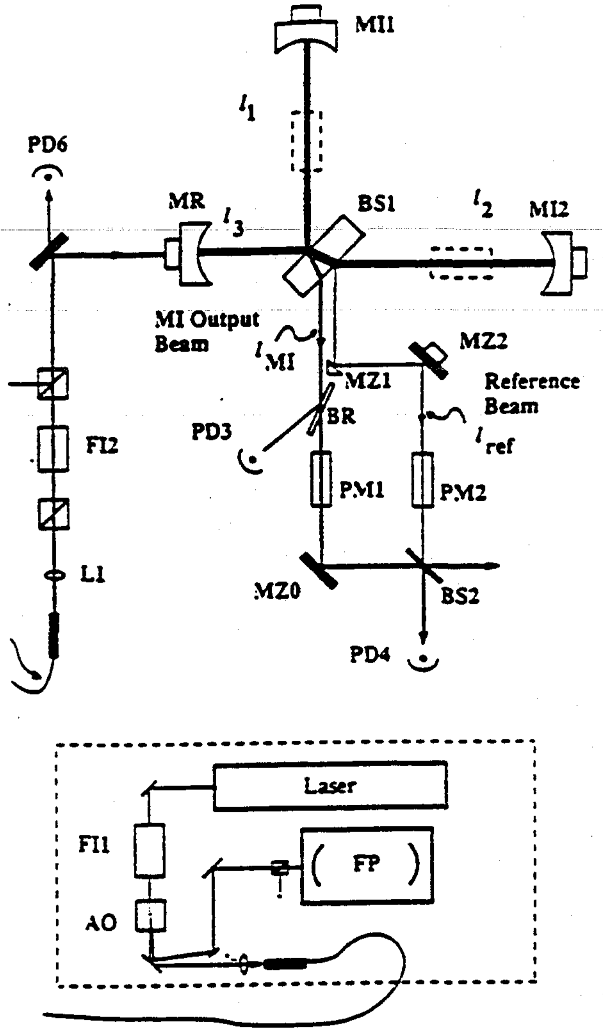


Fig. 5. Experimental arrangement to verify external modulation.

The light from the fiber is matched into the optical system with lens L1 (focal length $f = 16$ mm), and once again isolated from the laser and fiber by FI2. The Michelson interferometer MI consists of a 50–50 beam splitter BS1 and the 78 cm concave mirrors MI1 and MI2, placed $l_1 = 18$ cm from BS1. For recycling, mirror MR (101 cm radius concave) is placed at a distance 12 cm from BS1. The output of the Michelson interferometer MI passes through Pockels cell PM1, is reflected by mirror MZ0, and is brought to the Mach–Zehnder beam splitter BS2. The reference beam for the Mach–Zehnder interferometer is taken from the AR-coated face of BS1, guided by mirrors MZ1 and MZ2 through Pockels cell PM2 to interference with the output beam of the Michelson interferometer on BS2. Note that this differs from the reference source indicated in fig. 4, which is taken from the laser directly. A potential disadvantage of the present arrangements is that the reference is now taken asymmetrically from one arm of the interferometer, but a calculation of this effect shows that it makes at most a small correction to the sensitivity (<5%). However the advantage of this arrangement is the fact that the light for the reference arm has travelled almost the same path length as the main Michelson output, giving an insensitivity to frequency noise and assuring matched optical wave fronts. To preserve this advantage, the path length from BS1 to BS2 for the Michelson beam, l_{M1}, and the reference beam, l_{ref}, are adjusted for near equality.

The Michelson output is held on a dark fringe by a 13 kHz synchronous detection system. The error signal for this system is derived from photodiode PD3 which monitors the Michelson output intensity picked off by the near-Brewster plate BR; feedback is applied to PZTs mounted on mirrors MI1 and MI2. The depth of modulation is sufficiently small so that the intensity on the dark fringe is not significantly increased over that due to the finite contrast. In addition to providing the servo error signal, the known amplitude of the mirror motion at 13 kHz also serves to produce a calibration peak to determine the sensitivity of the detection scheme.

The laser intensity is actively stabilized (using PD6 and AO) in a band around 13 kHz to improve the signal-to-noise ratio for the Michelson servo. Another servo system using a PZT on MR, photodiode PD2, and a 19 kHz synchronous detection holds the recycling cavity on resonance.

Pockels cell PM1 impresses a 10 MHz phase modulation on the Michelson output beam for the Mach–Zehnder synchronous detection; PM2 ensures that the optical paths l_{M1} and l_{ref} are matched. One of the Mach–Zehnder output beams falls on photodiode PD4; this photocurrent is demodulated at 10 MHz, and this output signal is sent to a spectrum analyzer to determine the signal-to-noise ratio. This same output signal also contains the 13 kHz calibration signal which we wish to maximize to obtain the correct operating phase, i.e., the best signal to noise ratio. Demodulating this signal at 13 kHz, we obtain a DC level which can be maximized, through changes in the bias voltage applied to the PZT on MZ2, to set the correct operating phase ϕ_0^{MZ}.

3.2. Internal modulation tests

As a preliminary to our external modulation measurements, a simple internal modulation Michelson was configured allowing us to optimize the 13 kHz locking system, establish noise levels in the 10 MHz synchronous detection system, and calibrate the phase-modulating Pockels cells. The Michelson interferometer was configured with the Pockels cells in the arms of the interferometer (shown dashed in fig. 5) and photodiode PD4 was placed at the output of the interferometer. With the interferometer held on a dark fringe by the 13 kHz synchronous detection system through PZT feedback, a computer controlled measuring sequence measured the signal (the 13 kHz calibration peak) and the broadband noise (around 60 kHz) as a function of the modulation depth m for various power levels. These automated measurements could be taken quickly and with great repeatability; some hitherto unrecognized problems (e.g., rapid changes in the interferometer contrast through Pockels cell heating) were discovered and corrected. Also, a calibration of the modulation depth m as a function of our attenuator-controlled RF (10 MHz) drive level was obtained and verified.

The overall agreement between the calculated displacement sensitivities Δl_{int} and the measured values of Δl_{int} for this internal modulation system were quite good: the error was of the order of $\Delta l^{meas}_{int}/\Delta l^{calc}_{int} = 1.2$. Plots of calculated versus measured signal, and calculated versus measured noise, show the range of signal levels for which the system remained linear, and comparison with the shot noise due to an incandescent bulb or unmodulated (10 MHz) laser light shows that the increased experimental Δl_{int} is due to an excess of noise in the interferometer output of 1.2 (rather than a lack of (13 kHz) signal). This excess seems to be electrical (e.g., pre-amplifier, mixer) in nature, and, thus, should also be relevant for the external modulation technique results.

3.3. External-modulation measurement procedure

The Michelson interferometer beam splitter BS1 is adjusted for the best contrast (typically 99.5%); this is done while continuously locked on a dark fringe to avoid false minima while searching in adjustment space. If the measurement is to be made with recycling, the recycling mirror MR is put in place, and it and the position of lens L1 are adjusted to match and align into the recycling cavity; about 90% of the light can be coupled into the fundamental mode.

Correct alignment of the Mach–Zehnder interferometer requires that the TEM$_{00}$ component of the Michelson output beam be colinear with the (TEM$_{00}$) reference beam on the Mach–Zehnder beam splitter BS2. However, on the dark fringe of the Michelson, the residual output beam is primarily due to wave front distortion from the two arms of the interferometer, and the power in the TEM$_{00}$ mode is negligibly small and thus difficult to identify. For alignment purpose, the Michelson output is deliberately held away from the dark fringe so that the TEM$_{00}$ mode dominates. This allows for a good alignment criterion; typically 90% of the possible contrast (given the two interfering beam intensities) is achieved. Once the system is aligned, this Michelson fringe offset is removed.

With the phase difference of the Mach–Zehnder held at the correct value for sensing differential changes a series of measurements at various modulation depths m are made under computer control; for each measurement several experimental parameters are measured. Photodiode PD6, which is calibrated via a wattmeter (Photon Control), gives a measure of the incident laser power P_{inc}. The reference beam intensity I_{ref} is measured at PD4 (by blocking the Michelson output beam I_{MI}). For nonrecycled measurements, the Michelson bright fringe intensity I_0 is measured directly at PD4 by blocking I_{ref} and holding the Michelson on a bright fringe; this allows us to measure the (constant) ratio between I_{ref} and I_0 which can be used to deduce I_0 in the case of recycled operation. The photocurrent in PD4 during the spectral (signal and noise) measurement, I_{data}, also is recorded. Finally, for each measurement the level of the 13 kHz calibration signal V_{calib} and the average noise spectral density V_{noise} in a band from 56 to 65 kHz are measured and recorded.

3.4. External modulation results

Table 1 shows a summary of measurements made using the external modulation technique, with and without recycling, and at various modulation depths and powers. With the known displacement amplitude of the 13 kHz calibration signal ($x_{calib}=1.2\times10^{-10}\,\mathrm{m_{rms}}$), the equivalent displacement noise level Δl_{ext} can be calculated from

$$\Delta l_{ext}^{meas}=V_{noise}\frac{x_{calib}}{V_{calib}}\,. \tag{19}$$

To obtain the expected signal-to-noise ratio for our experiment, we use eq. (18) modified to take into account the experimental situation:

$$\Delta l_{ext}^{calc}=\frac{\lambda}{2\pi}\frac{\sqrt{e(I_{data}+I_{det})}}{J_1(m)\sqrt{I_{ref}I_0}}\,. \tag{20}$$

I_0 and I_{ef} are here measured *after* the Michelson beamsplitter. The current due to light falling on the photodetector I_{data} is the sum of the Mach–Zehnder interference term I_{DC}^{ext} plus the current due to any non-TEM$_{00}$ light from contrast defects in the Michelson. Our Michelson output beam consists primarily of power which is not in the TEM$_{00}$ mode and does not interfere with the exclusively TEM$_{00}$-mode reference beam. The electronic noise of the photodiode–amplifier combination, which adds in quadrature to the shot noise due to I_{data}, can be characterized as an effective current I_{det}. Note that the noise was often dominated by the PD4 photodetector–amplifier system noise of 4.4 pA/$\sqrt{\mathrm{Hz}}$, which is equivalent to the shot noise due to a photocurrent of $I_{det}=0.062$ mA.

The maximum value of $J_1(m)$ used here, 0.46 (for $m=1.06$ radians) is 0.8 of the maximum possible, 0.582 ($m=1.8$); we have demonstrated an m dependence that agrees with our theory over a wide range of modulation depths.

The recycling factor (increase in the circulating light power) for the cavity can be determined by looking at the change in the ratio of I_{ref} to P_{inc} without and with recycling; here we have a gain of 27.5. Wit the measured transmission of the recycling mirror of $T=3\%$ this allows us to calculate the total losses

Table 1
Results of external modulation measurements. Using the measured parameters and values of known constants ($I_{\rm det}=0.062$ mA, $\lambda=514.5$ nm) the expected displacement sensitivity has been calculated. This is compared to the measured sensitivity derived from the calibration and noise levels at the detection system output and the known calibration peak level ($x_{\rm calib}=1.2\times10^{-10}$ m rms). The average agreement of $\Delta l^{\rm meas}/\Delta l^{\rm calc}\sim1.9$ results from a slight excess of noise (of order 1.2) and a smaller than expected signal (by 1.6). The recycling gain is determined by noting that recycled values of $I_0/P_{\rm inc}$ are 27.5 times larger than in the non-recycled cases.

	Measured parameters					Measurements		$\Delta l_{\rm external}$ ($\times10^{-15}$ m/$\sqrt{\rm Hz}$)	
	$P_{\rm inc}$ (mW)	$I_{\rm ref}$ (mA)	I_0 (mA)	$J_1(m)$	$I_{\rm dark}$ (mA)	$V_{\rm calib}$ (mV rms)	$V_{\rm noise}$ (μV/$\sqrt{\rm Hz}$)	calculated	measured
without recycling	12.6	0.005	0.87	0.26	0.005	15	3.5	16.13	28.00
	12.6	0.005	0.87	0.46	0.005	27	3.5	9.11	15.56
	30.8	0.012	2.13	0.46	0.015	59	3.5	3.96	7.12
with recycling	3.8	0.039	6.93	0.46	0.042	200	5.3	1.40	3.18
	3.8	0.039	6.93	0.46	0.038	210	4.4	1.37	2.51
	8.5	0.094	16.90	0.13	0.091	130	5.6	2.47	5.17

in the interferometer to be of the order of 3.4%.

The average ratio of the calculated signal to noise to the measured signal to noise is 1.9. With the additional information gathered in the internal modulation tests, the external modulation results for the noise and the signal can be interpreted separately; for the noise, we measure the same excess that was seen for the internal modulation scheme. However, the measured signal was less than the calculated signal for the external modulation detection by a factor of 1.6. A partial explanation may be given by the fine contrast of the MZ interferometer. The signal is expected to be reduced by the contrast defect of 0.9, and with these two corrections the ratio of the measured to the expected signal-to-noise ratios for the external modulation scheme is 1.4, a reasonable agreement for these initial proof-of-concept experiments.

4. Conclusions

External modulation is a technique for the fringe interrogation in a Michelson interferometer that avoids some of the practical difficulties associated with the customary internal modulation. The losses, beam distortion, and power-dependent blooming due to electro-optic modulators in the Michelson arms are eliminated. These advantages are particularly important in recycled interferometers. We have demonstrated a reasonable agreement between theory and experiment in the case where the reference beam intensity dominates. Future experiments should explore the global optimization of the reference beam intensity, mode cleaning to eliminate non-TEM$_{00}$ components in the Michelson output, methods to hold the reference phase to the optimum value, and the use of two photodetectors. In addition, a more detailed theory incorporating these refinements and including a more complete description of the contrast is desirable.

Acknowledgement

We thank A. Brillet, R. Schilling and L. Schnupp for many valuable discussions. D.S. and D.D. thank CNRS for grant support.

References

[1] D.G. Blair and M.J. Buckingham, eds., Proc. Fifth Marcel Grossmann Meeting on General relativity, Perth, 1988 (World Scientific, Singapore, 1989), and references therein.

[2] R.W.P. Drever, in: Gravitational radiation, Les Houches 1982, eds. N. Deruelle and T. Piran (North-Holland Amsterdam, 1983) pp. 321–338.

[3] J.Y. Vinet, B. Meers, C.N. Man and A. Brillet, Phys. Rev. D 38 (1988) 433.

[4] C. Townes, Microwave spectroscopy (McGraw-Hill, New York, 1955) p. 426.

[5] R.W.P. Drever: Ideas for interferometric detectors for gravitational radiation, at 10th Int. Conf. on General relativity and gravitation, Padua 4–9 July 1983.

[6] R. Weiss, Q. Prog. Rep. Res. Lab. Electron. MIT 105 (1972) 54.

[7] D. Shoemaker, R. Schilling, L. Schnupp, W. Winkler, K. Maischberger and A. Rudiger, Phys. Rev. D 38 (1988) 423.

[8] R.W.P. Drever, G.M. Ford, J. Hough, I.M. Kerr, A.J. Munley, J.R. Pugh, N.A. Robertson and H. Ward, in: Proc. 9th Int. Conf. on General relativity and gravitation, Jena, 1980, ed. E. Schmutzer (VEB Deutscher Verlag der Wissenschaften, Berlin, 1983).

[9] R.W.P. Drever, J.L. Hall, F.V. Kowalski, J. Hough, G.M. Ford, A.J. Munley and H. Ward, Appl. Phys. B 31 (1983) 97.

[10] C.N. Man and A. Brillet, J. Appl. Phys. 56 (1984) 71.

REF GG

Building Scientific Apparatus

A Practical Guide to Design and Construction

John H. Moore ▪ Christopher C. Davis ▪ Michael A. Coplan

University of Maryland

1983

ADDISON-WESLEY PUBLISHING COMPANY
Advanced Book Program / World Science Division
London ▪ Amsterdam ▪ Don Mills, Ontario ▪ Sydney ▪ Tokyo

Figures 4.17, 4.18, 4.19(*a*), 4.45, and 4.119 are reprinted with permission from *Handbook of Lasers*, R. J. Pressley, Ed., CRC Press, Cleveland, 1971. Copyright The Chemical Rubber Co., CRC Press, Inc.

Illustrations by James S. Kempton

Library of Congress Cataloging in Publication Data

Moore, John H., 1941–
Building scientific apparatus.

Includes index.
1. Scientific apparatus and instruments—Handbooks, manuals, etc. 2. Instrument manufacture—Handbooks, manuals, etc.
I. Davis, Christopher C., 1944– . II. Coplan, Michael A., 1938– . III. Title.
Q185.M66 1983 681'.75 82-8726

ISBN 0-201-05532-5 AACR2

Copyright © 1983 by Addison-Wesley Publishing Company, Inc.
Published simultaneously in Canada

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher, Addison-Wesley Publishing Company, Inc., Advanced Book Program/World Science Division, Reading, Massachusetts 01867, U.S.A.

Printed in the United States of America

ABCDEFGHIJ-HA-898765432

From Moore, Davis & Coplan Building Scientific Apparatus, 1983

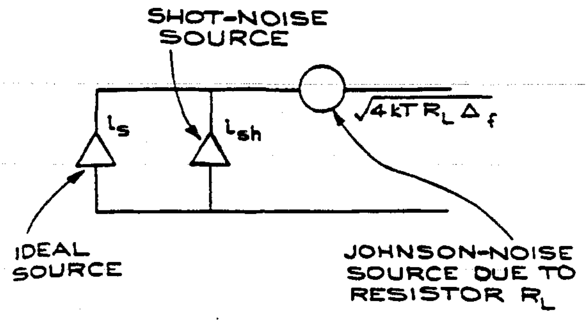


Figure 6.104 Explicit representation of noise sources in a detector.

this case the SNR is optimized by using as large a value of R_L as possible and an amplifier with high input impedance and low v_n and i_n. Amplifiers with FET input stages are most appropriate in this case. Maximum practical values of R_L are limited by the capacitance C of the input circuit (including cables) and the input impedance of the amplifier. There is no benefit in having R_L greater than the input impedance of the amplifier or so large that the time constant R_LC requires one to work at low frequencies where $1/f$ noise dominates. The justification for using large values of R_L comes from the fact that the voltage signal from a current source is proportional to R_L, while the Johnson noise is proportional to $\sqrt{R_L}$.

An alternative way of specifying amplifier performance is with the *equivalent noise temperature*, *Te*, defined as the increase in temperature of the source resistor necessary to produce the observed noise at the amplifier output, the amplifier being considered noiseless for this purpose. We have

$$T_e = T(10^{\mathrm{NF}/10} - 1),$$

where T is the absolute temperature of the source resistor. Using the example of the amplifier with an NF of 3 dB and a source resistor at 300 K,

$$T_e = (300\ \mathrm{K})(10^{0.3} - 1) = 300\ \mathrm{K}.$$

In this case the amplifier introduces an amount of noise equal to the noise from the source resistor.

6.8.3 The Lock-in Amplifier and Gated Integrator or Boxcar

The proper matching of a signal source to an amplifier is the first step to be taken in the recovery of a signal accompanied by noise. Once this has been done correctly, a number of signal-enhancing techniques can be used to extract the signal.

Because of the difficulty in making d.c. measurements due to zero drifts, amplifier instabilities, and flicker noise, the signal of interest should, if at all possible, be at a frequency sufficiently high that the dominant noise is white noise. Often choppers are used to convert a d.c. or low-frequency signal to a higher-frequency one. Clearly it is wise to choose a frequency far from the power-line frequency and its harmonics. Also to be avoided are frequencies near known noise frequencies.

Signal-enhancing techniques are mainly based on bandwidth reduction. Since the white-noise power per unit bandwidth is constant, reducing the bandwidth will reduce the noise power proportionally. Of course, in the limit as the bandwidth goes to zero, the noise power goes to zero and the signal power as well. Common bandwidth-narrowing circuits are the resonant filter and low-pass filter. For resonant filters, the sharpness of the resonance is given in terms of Q, which is approximately equal to $f_0/\Delta f$, where f_0 is the frequency to which the filter is tuned and Δf is the bandwidth. Practical values of Q for electronic resonant filters vary from 10 to 100. A simple detection system might take the form shown in Figure 6.105. The rectifier converts the a.c. signal back to d.c., so it can be recorded on a chart recorder or read from a meter. With this arrangement the noise passed by the filter is rectified along with the signal and adds to the recorded d.c. level.

The effective bandwidth can be substantially narrowed if in the above arrangement the ordinary rectifier is replaced by a synchronous rectifier. This kind of rectifier acts like a switch that is opened and closed in synchronization with the chopper. Since the phase relations for each of the noise components are random (a requirement of white noise), they will tend to cancel each other when averaged, and a low-pass filter at the output of the rectifier can be used to do the averaging. The RC time constant of the filter is related to the

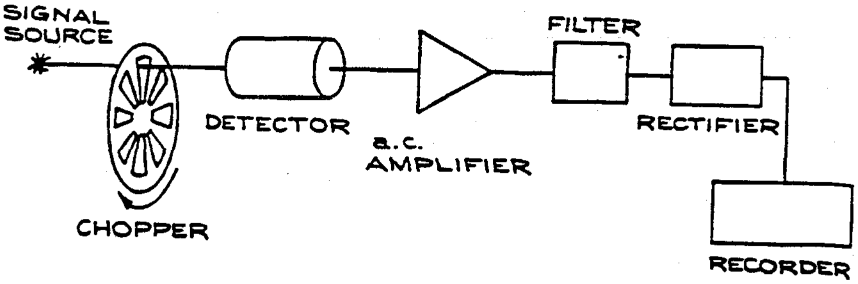


Figure 6.105 Simple detection and signal-enhancement system.

effective passband of the system, Δf, by

$$\Delta f = \frac{1}{4RC}$$

for a single-section low-pass filter, and

$$\Delta f = \frac{1}{4nRC}$$

for n concatenated low-pass filters, each with time constant RC. Quite small values of Δf are possible with such a system, the only limitation being the length of time necessary for the measurement. If Δf is the passband, the measurement time is approximately $1/\Delta f$, so that it is necessary for the source signal to remain stable over times comparable to $1/\Delta f$. Normally the chopping frequency is chosen to be ten times the highest component frequency to be recovered in the source signal.

The system described above is often called a *lock-in amplifier* or *phase-sensitive detector*. The details of the operation of such devices are well documented.[9] However, as far as signal enhancement is concerned, the formulas given above are sufficient for estimating the benefits of such devices. It is well to remember that the SNR is proportional to $1/\sqrt{\Delta f}$, so that narrowing the passband by a factor of 4 increases SNR by a factor of 2, but increases the time of measurement by a factor of 4.

When the signal to be detected has the form of a repetitive low-duty-cycle train of pulses, the lock-in amplifier may not be the best method for signal enhancement. Here the *duty cycle* is defined as the fraction of the time during which the signal of interest is present. With low-duty cycles, signal information is available for only a fraction of the total time, while noise is always on the line. With timing and gating circuits it is possible to connect the signal line to an RC integrating circuit only during those times when the signal is present. The time constant of the integrator is then chosen to be very much larger than the period of the pulse train. The time required for the capacitor to charge to 99% of the final voltage level is $4.6RC$, so that 5 time constants after the first gate opening the capacitor should be charged to within 1% of its final steady-state value, provided the signal is continuously present. If the signal is present for a fraction γ of the time, the time constant of the integrator must be increased to $5RC/\gamma$. With this system, known as a *gated* or *boxcar integrator*, the SNR is increased only by increasing the time of the measurement. The effective bandwidth of the instrument is $\gamma/4RC$. Table 6.35 compares the important parameters associated with lock-in amplifiers and gated integrators.

6.8.4 Signal Averaging

With a repetitive signal, improvement in the SNR can be effected by merely averaging the signal over many cycles. Let SNR_0 represent the SNR in one cycle, and N the number of cycles over which one averages. The SNR improvement is proportional to $\sqrt{N}$. If each cycle lasts for a time τ, the time T necessary to arrive at a specified SNR is given by

$$T = \left(\frac{\text{SNR}_0}{\text{SNR}}\right)^2 \tau.$$

Lock-in amplifiers and gated integrators are inherently superior to simple signal averaging because of the

Table 6.35 COMPARISON OF LOCK-IN AMPLIFIER AND GATED INTEGRATOR

	Lock-in Amplifier	*Gated Integrator*
Duty cycle δ	> 0.05	< 0.50
Bandwidth	$1/8RC$	$\gamma/4RC$
Minimum measurement time	$5RC$	$5RC/\gamma$
Highest recoverable signal frequency	$1/20\pi RC$ for chopping frequency $\geqslant 1/2\pi RC$	$\gamma/20\pi RC$ for repetition frequency $\geqslant 1/2\pi RC$
Design notes	1. Determine $f_{\max}$, the highest frequency component of the signal to be recovered. 2. Choose f_s, the chopping frequency, where $f_s = 10 f_{\max}$. 3. Choose low-pass filter constants $1/RC = 2\pi f_s$ so as to pass frequency $f_{\max}$. 4. The bandwidth Δf is $1/8RC$, and a tuned amplifier with a Q of 10 is sufficient to pass all frequency components of the signal to $f_{\max}$.	1. Required measurement time is $5\ RC/\gamma$. 2. Bandwidth is $\gamma/4RC$. 3. Preferred to a lock-in amplifier for signal repetition rates $\leqslant 10$ Hz and low duty cycle.

band-narrowing functions they perform; however, the bandwidth improvement is only an advantage when the highest component frequency in the signal to be recovered permits a long integrating time.

6.8.5 Waveform Recovery

The gated integrator can be converted to an instrument for waveform recovery by the inclusion of variable delay and variable gate-width functions. A separate RC network is required for each delay time in such a scheme. A schematic representation of such a system is shown in Figure 6.106. Each separate delay corresponds to a different part of the waveform. The duty cycle for each gate opening is $\tau/T = \gamma$, so that the effective bandwidth is $\gamma/4RC$. An integration time of $5RC/\gamma$ is needed for the charge on each capacitor to reach a steady-state value. If the waveform has been divided into N parts, the total time of measurement is $5NRC/\gamma$.

Digital schemes for recovering waveforms use ADCs to convert the analog signal level to a digital number, which is then recorded in the appropriate time channel with the aid of delay and gating signals. With such instruments, called *digital signal analyzers*, SNRs are only limited by the length of time of the measurement; however, the long-term stability of the various components limits the use of very long measuring times. If the absolute value of the waveform is needed, one must record the number of cycles treated by the instrument, and data rates are limited by the conversion speed of the ADC that encodes the amplitude information. An advantage of digital signal analysis is the increased versatility in data manipulation and the production of permanent records on paper or magnetic media. Such data are then available for numerical analysis.

Table 6.45 SEMICONDUCTOR INTEGRATED-CIRCUIT CODE PREFIXES

Company	*Prefix*ª
Analog Devices	AD
Advanced Micro Devices	Am
General Instrument	AY, GIC, GP
Intel	C, I
RCA	CA, CD, CDP
TRW	CA, TDC, MPY, CMP, DAC, MAT, OP
Precision Monolithics	PM, REF, SSS
National Semiconductor	DM, LF, LFT, LH, LM, NH
Fairchild	F, μA, μL, U*nx*
Ferranti	FSS, ZLD
GE	GEL
Harris	HA
Motorola	HEP, MC, MCC, MCM, MFC, MM, MWM
Intersil	ICH, ICL, ICM, IM
ITT	ITT, MIC
Siliconix	L, LD
Fugitsu	MB
Mostek	MK
Plessey	MN, SL, SP
Signetics	N, NE, S, SE, SP
Raytheon	R, RAY, RC, RM
Texas Instruments	SN, TMS
Sprague	ULN, ULS
Westinghouse	WC, WM
Hewlett-Packard	5082-*nnnn*

ª*x* = number; *n* = letter.

Sometimes specially selected or matched components are used. Replacement with off-the-shelf units may not work in this case.

CITED REFERENCES

1. H.W. Bode, *Network Analysis and Feedback Amplifier Design*, Van Nostrand, Princeton, N.J., 1945.
2. Electronic Design, 24, 63, 1976.
3. J. Millman and C.C. Halkias, *Integrated Electronics: Analog and Digital Circuits and Systems*, McGraw-Hill, New York, 1972, pp. 244–245.
4. J.G. Graeme, *Operational Amplifiers, Design and Application*, G.E. Tobey and L.P. Huelsman, Eds., McGraw-Hill, New York, 1971; *Designing with Operational Amplifiers*, McGraw-Hill, New York, 1977.
5. E. Fairstein and J. Hahn, "Nuclear Pulse Amplifiers—Fundamentals and Design Practice," Nucleonics, 23, No. 7, 56, 1965; ibid., No. 9, 81, 1965; ibid., No. 11, 50, 1965; ibid., 24, No. 1, 54, 1966; ibid., No. 3, 68, 1966.
6. *Voltage Regulator Handbook*, National Semiconductor Corporation, Santa Clara, Calif.
7. *Standard Nuclear Instrument Modules*, adopted by AEC Committee on Nuclear Instrument Modules, U.S. Government Publication TID-20893 (Rev. 3).
8. S. Letzter and N. Webster, "Noise in Amplifiers," IEEE Spectrum, August 1970, pp. 67–75.
9. John C. Fisher, "Lock in the Devil, Educe Him or Take Him for the Last Ride in a Boxcar?" Tek Talk, Princeton Applied Research, 6, No. 1.
10. R. Morrison, *Grounding and Shielding Techniques in Instrumentation*, Wiley, New York, 1967.
11. *Floating Measurements and Guarding*, Application Note 123, Hewlett-Packard, 1970.
12. *Circuits for Electronics Engineers*, S. Weber, Ed., Electronics Book Series, McGraw-Hill, New York, 1977; *Circuit Design Idea Handbook*, W. Furlow, Ed., Cahner's Books, Boston, 1974; *Electronics Circuit Designer's Casebook*, Electronics, New York; *Signetics Analog Manual, Applications, Specifications*, Signetics Corporation, Sunnyvale, Calif.; *Linear Applications Handbook*, Vols. 1 and 2, National Semiconductor Corporation, Santa Clara, Calif.
13. W.R. Blood, Jr., *MECL Applications Handbook*, 2nd edition, Motorola Semiconductor Products, 1972.

GENERAL REFERENCES

CAMAC and IEEE-488 (GPIB or HP-IB)

CAMAC: A Modular Instrumentation System for Data Handling, ESONE Committee, Report EUR 4100, 1972, Chapters 4–6.

CAMAC Tutorial Issue, IEEE Trans. Nucl. Sci., NS-20, No. 2, April 1973.

D. Horelick and R.S. Larsen, CAMAC: "A Modular Standard," IEEE Spectrum, April 1976, p. 50.

KEF. HH

THE ART OF ELECTRONICS

Paul Horowitz HARVARD UNIVERSITY

Winfield Hill SEA DATA CORPORATION, NEWTON, MASSACHUSETTS

CAMBRIDGE UNIVERSITY PRESS
Cambridge
London New York
New Rochelle Melbourne Sydney

Published by the Press Syndicate of the University of Cambridge
The Pitt Building, Trumpington Street, Cambridge CB2 1RP
32 East 57th Street, New York, NY 10022, USA
296 Beaconsfield Parade, Middle Park, Melbourne 3206, Australia

© Cambridge University Press 1980

First published 1980
Reprinted 1981 (thrice), 1982

Printed in the United States of America
Typeset by Science Press, Inc., Ephrata, Pennsylvania
Printed and bound by Hamilton Printing Company, Rensselaer, New York

Library of Congress Cataloging in Publication Data

Horowitz, Paul, 1942–

The art of electronics.

1. Electronics. 2. Electronic circuit design.
I. Hill, Winfield, joint author. II. Title.
TK7815.H67 1980 621.381 79-27170
ISBN 0 521 23151 5 hard covers
ISBN 0 521 29837 7 paperback

from Horowitz & Hill, The Art of Electronics, 1980

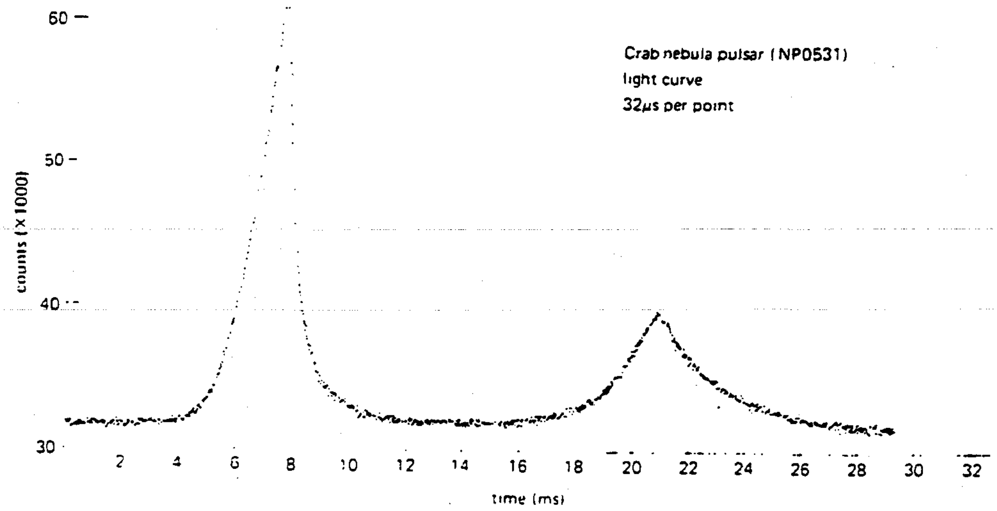


Figure 14.34
Crab nebula pulsar brightness versus time (light curve).

own well-defined periodicity may in fact be the most difficult to work with, since you have to know the periodicity precisely. The graph of the "light curve" (brightness versus time) in Figure 14.34 is an example. We made this curve by using an MCS on the output of a photomultiplier stationed at the focus of a 60-inch telescope, run exactly in synchronism with the pulsar's rotation. Even with that size telescope it required an average of approximately 5 million sweeps to generate such a clean curve, since the average number of detected photons for each entire pulsar pulse was about 1. With such a short period, that puts enormous accuracy requirements on the MCS channel-advance circuitry, in this case requiring clocks of part-per-billion stability and frequent adjustment of the clock rate to compensate for the earth's motion.

It is worth saying again that the essence of signal averaging is a reduction in bandwidth, gained by running an experiment for a long period of time. The bottom line here is the total length of the experiment; the particular rate of scanning, or modulation, is usually not important, as long as it takes you far enough from the $1/f$ noise present near dc. You can think of the modulation as simply shifting the signal you wish to measure from dc up to the modulating frequency. The effect of the long data accumulation is then to center an effective bandwidth $\Delta f = 1/T$ at f_{mod}, rather than at dc.

14.15 Lock-in detection

This is a method of considerable subtlety. In order to understand the method, it is necessary to take a short detour into the phase detector, a subject we first took up in Section 9.29. (Attached)

Phase detectors

In Section 9.29 we described phase detectors that produce an output voltage proportional to the phase difference between two digital (logic-level) signals. For purposes of lock-in detection, you need to know about linear phase detectors, since you are nearly always dealing with analog voltage levels.

The basic circuit is shown in Figure 14.35. An analog signal passes through a linear amplifier whose gain is reversed by a square-wave "reference" signal controlling a FET switch. The output signal passes through a low-pass filter, RC. That's all there is to it. Let's see what you can do with it.

Phase-detector output. To analyze the

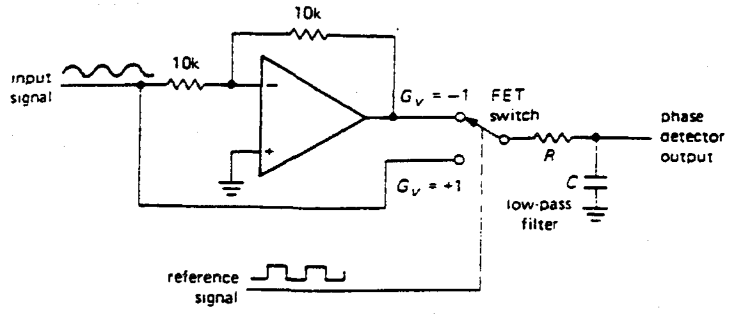


Figure 14.35
Phase detector for linear input signals.

phase detector operation, let's assume we apply a signal

$E_s \cos(\omega t + \phi)$

to such a phase detector, whose reference signal is a square wave with transitions at the zeros of sin ωt, i.e., at $t = 0, \pi/\omega, 2\pi/\omega$, etc. Let us further assume that we average the output, V_{out}, by passing it through a low-pass filter whose time constant is longer than one period:

$\tau = RC >> T = 2\pi/\omega$

Then the low-pass-filtered output is

$$\langle E_s \cos(\omega t + \phi)\rangle \Big|_0^{\pi/\omega} - \langle E_s \cos(\omega t + \phi)\rangle \Big|_{\pi/\omega}^{2\pi/\omega}$$

where the brackets represent averages, and the minus sign comes from the gain reversal over alternate half cycles of V_{ref}. As an exercise, you can show that

$\langle V_{out}\rangle = -(2E_s/\pi)\sin\phi$

EXERCISE 14.2
Perform the indicated averages by explicit integration to obtain the preceding result for unity gain.

Our result shows that the averaged output, *for an input signal of the same frequency as the reference signal,* is proportional to the amplitude of V_s and sinusoidal in the relative phase.

We need one more result before going on: What is the output voltage for an input signal whose frequency is close to (but not equal to) the reference signal? This is easy, since in the preceding equations the quantity ϕ now varies slowly, at the difference frequency:

$$\cos(\omega + \Delta\omega)t = \cos(\omega t + \phi) \quad \text{with} \quad \phi = t\Delta\omega$$

giving an output signal that is a slow sinusoid:

$V_{out} \doteq (2E_s/\pi)\sin(\Delta\omega)t$

which will pass through the low-pass filter relatively unscathed if $\Delta\omega < 1/\tau = 1/RC$ and will be heavily attenuated if $\Delta\omega > 1/\tau$.

The lock-in method

Now the so-called lock-in (or phase-sensitive) amplifier should make sense. First you make a weak signal periodic, as we've discussed, typically at a frequency in the neighborhood of 100Hz. The weak signal, contaminated by noise, is amplified and phase-detected relative to the modulating signal. Look at Figure 14.36. You need an experiment with two "knobs" on it, one for fast modulation in order to do phase detection and one for a slow sweep through the interesting features of the signal (in NMR, for example, the fast modulation might bε a small 100Hz modulation of the magne ic field, and the slow modulation might be a frequency sweep of 10 minutes' duration through the resonance). The phase shifter is adjusted to give maximum output signal, and the low-pass filter is set for a time constant long enough to give good signal/noise ratio. The low-pass-filter rolloff

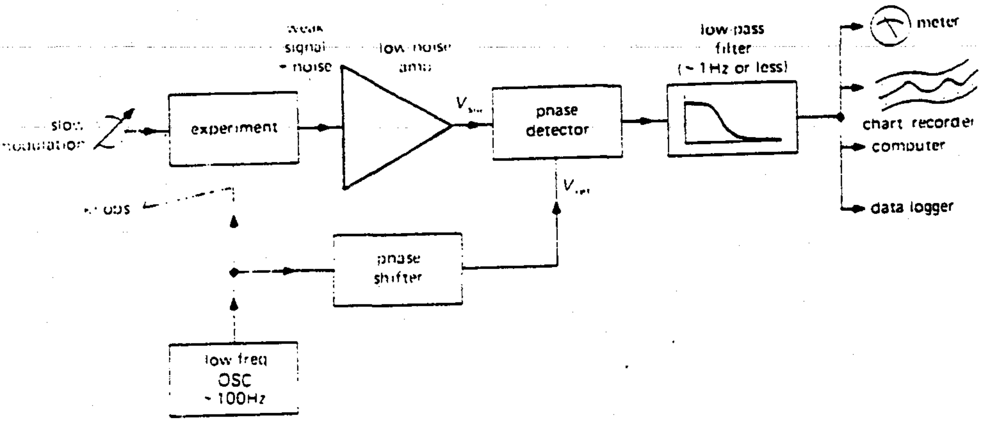


Figure 14.36
Lock-in detection.

sets the bandwidth, so a 1Hz rolloff, for example, gives you sensitivity to spurious signals and noise only within 1 Hz of the desired signal. The bandwidth also determines how fast you can adjust the "slow modulation," since now you must not sweep through any features of the signal faster than the filter can respond. People use time constants of fractions of a second up to tens of seconds and often do the slow modulation with a geared-down clock motor turning an actual knob on something!

Note that lock-in detection amounts to bandwidth narrowing again, with the bandwidth set by the postdetection low-pass filter. As with signal averaging, the effect of the modulation is to center the signal at the fast modulation frequency, rather than at dc, in order to get away from 1/*f* noise (flicker noise, drifts, and the like).

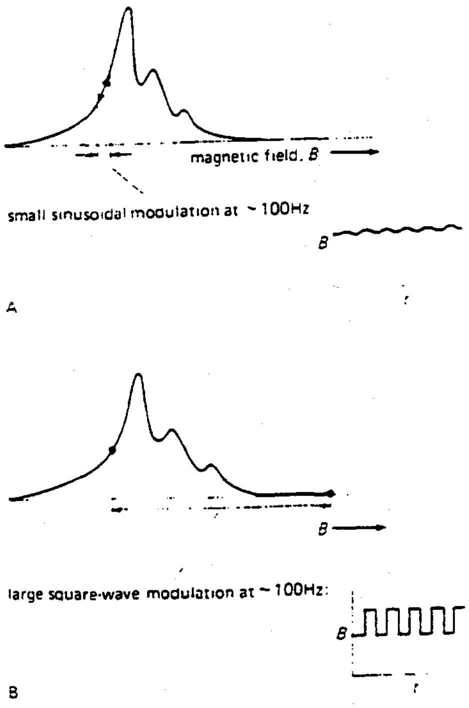


Figure 14.37
Lock-in modulation methods. A: Small sinusoid. B: Large square wave.

Two methods of "fast modulation"

There are two ways to do the fast modulation: The modulation waveform can be either a very small sine wave or a very large square wave compared with the features of the sought-after signal (line shape versus magnetic field, for example, in NMR), as sketched in Figure 14.37. In the first case the output signal from the phase-sensitive detector is proportional to the *slope* of the line shape (i.e., its derivative), whereas in the second case it is proportional to the line shape itself (providing there aren't any other lines out at the other endpoint of the modulation waveform). This is the reason all those simple NMR resonance lines come out looking like dispersion curves (Fig. 14.38).

For large-shift square-wave modulation

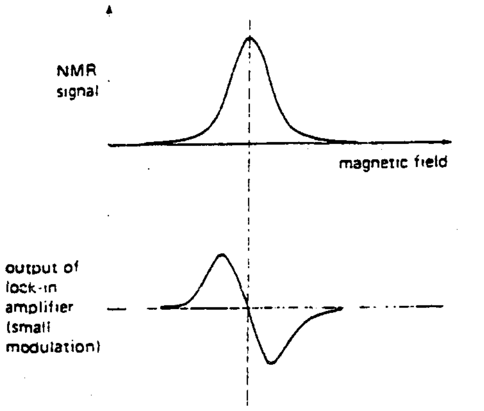


Figure 14.38
Line shape differentiation resulting from lock-in detection.

there's a clever method for suppressing modulation feedthrough, in cases where that is a problem. Figure 14.39 shows the modulation waveform. The offsets above and below the central value kill the signal, causing an on/off modulation of the signal at *twice* the fundamental of the modulating waveform. This is a method for use in special cases only; don't get carried away by the beauty of it all!

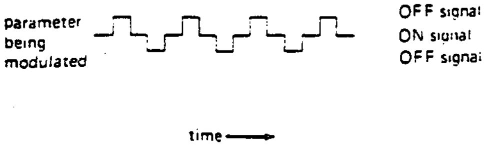


Figure 14.39
Modulation scheme for suppressing modulation feedthrough.

Large-amplitude square-wave modulation is a favorite with those dealing in infrared astronomy, where the telescope secondary mirrors are rocked to switch the image back and forth on an infrared source. It is also popular in radioastronomy, where it's called a Dicke switch.

Commercial lock-in amplifiers have a variable-frequency modulating source and tracking filter, a switchable time-constant post-detection filter, a good low-noise wide-dynamic-range amplifier (you wouldn't be using lock-in detection if you weren't having noise problems), and a nice linear phase detector. They also let you use an external source of modulation. There's a knob that adjusts the phase shift, so you can maximize the detected signal. The whole item comes packed in a handsome cabinet, with a meter to read output signal. Typically these things cost a few thousand dollars.

In order to illustrate the power of lock-in detection, we usually set up a small demonstration for our students. We use a lock-in to modulate a small LED of the kind used for panel indicators, with a modulation rate of a kilohertz or so. The current is very low, and you can hardly see the LED glowing in normal room light. Six feet away a phototransistor looks in the general direction of the LED, with its output fed to the lock-in. With the room lights out, there's a tiny signal from the phototransistor at the modulating frequency (mixed with plenty of noise), and the lock-in easily detects it, using a time constant of a few seconds. Then we turn the room lights on (fluorescent), at which point the signal from the phototransistor becomes just a huge messy 120Hz waveform, jumping in amplitude by 50dB or more. The situation looks hopeless on the oscilloscope, but the lock-in just sits there, unperturbed, calmly detecting the same LED signal at the same level. You can check that it's really working by sticking your hand in between the LED and detector. It's darned impressive.

14.16 Pulse-height analysis

A pulse-height analyzer (PHA) is a simple extension of the multichannel scaler principle, and it is a very important instrument in nuclear and radiation physics. The idea is simplicity itself: Pulses with a range of amplitudes are input to a peak-detector/ADC circuit that converts the relative pulse height to a channel address. A multichannel scaler then increments the selected address. The result is a graph that is a histogram of pulse heights. That's all there is to it.

The enormous utility of pulse-height

dly vanishing. And with proper design conservative application, the PLL is as ble a circuit element as an op-amp or flop.

Figure 9.51 shows the classic PLL configtion. The phase detector is a device that mpares two input frequencies, generating output that is a measure of their phase ference (if, for example, they differ in quency, it gives a periodic output at the ference frequency). If f_{IN} doesn't equal , the phase-error signal, after being ered and amplified, causes the VCO quency to deviate in the direction of f_{IN}. If nditions are right (lots more on that soon), VCO will quickly "lock" to f_{IN}, maintaing a fixed phase relationship with the input gnal.

At that point the filtered output of the hase detector is a dc signal, and the control nput to the VCO is a measure of the input equency, with obvious applications to tone ecoding (used in digital transmission over elephone lines) and FM detection. The VCO utput is a locally generated frequency equal to f_{IN}, thus providing a clean replica of f_{IN}, which may itself be noisy. Since the VCO output can be a triangle wave, sine wave, or whatever, this provides a nice method of generating a sine wave, say, locked to a train of input pulses.

In one of the most common applications of PLLs, a modulo-*n* counter is hooked between the VCO output and the phase detector, thus generating a multiple of the input reference frequency f_{IN}. This is an ideal method for generating clocking pulses at a multiple of the power-line frequency for integrating A/D converters (dual-slope, chargebalancing), in order to have infinite rejection of interference at the power-line frequency and its harmonics. It also provides the basic technique of frequency synthesizers.

9.29 PLL components

Phase detector

Let's begin with a look at the phase detector. There are actually two basic types, sometimes referred to as type I and type II. The type I phase detector is designed to be driven by analog signals or digital squarewave signals, whereas the type II phase detector is driven by digital transitions (edges). They are typified by the type I 565 (linear) and the type II 4044 (TTL):the CMOS 4046 contains both.

The simplest phase detector is the type I (digital), which is simply an exclusive-OR gate (Fig. 9.52). With low-pass filtering, the graph of the output voltage versus phase difference is as shown, for input square waves of 50% duty cycle. The type I (linear) phase detector has similar output-voltage-versus-phase characteristics, although its internal circuitry is actually a "four-quadrant multiplier," also known as a "balanced mixer." Highly linear phase detectors of this type are essential for *lock-in detection*, a lovely technique we will discuss in Section 14.15.

The type II phase detector is sensitive only to the relative timing of *edges* between the signal and VCO input, as shown in Figure 9.53. The phase comparator circuitry generates either *lead* or *lag* output pulses, depending on whether the transitions of the VCO output occur before or after the transitions of the reference signal, respectively. The width of these pulses is equal to the time between the respective edges, as

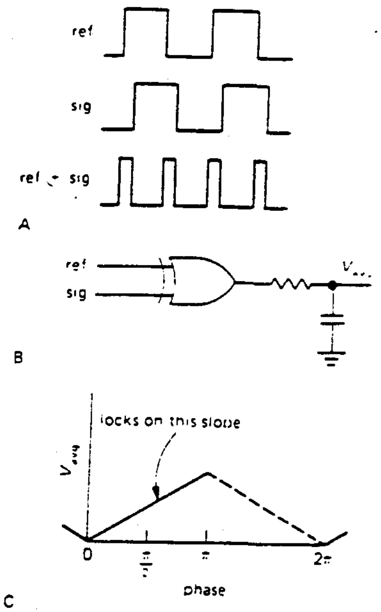


Figure 9.52
Exclusive-OR-gate phase detector (type I).

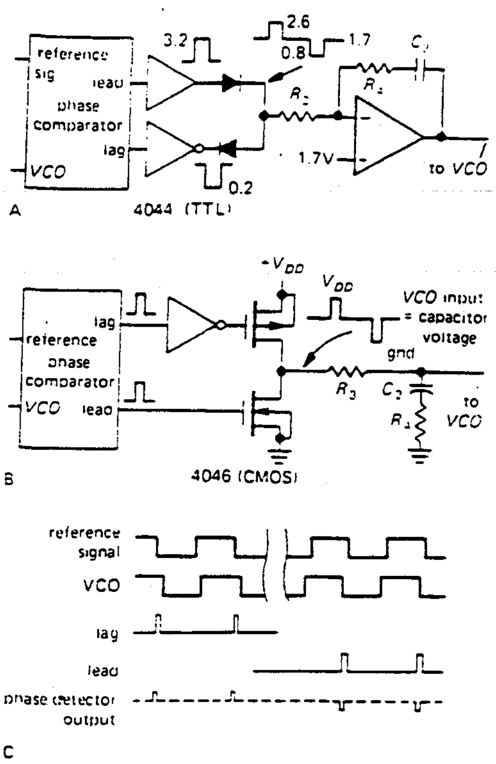


Figure 9.53
Edge-sensitive lead-lag phase detector (type II).

shown. The output circuitry then either sinks or sources current (respectively) during those pulses and is otherwise open-circuited, generating an average output voltage versus phase difference like that in Figure 9.54. This is completely independent of the duty cycle of the input signals, unlike the situation with the type I phase comparator discussed earlier. Another nice feature of this phase detector is the fact that the output pulses disappear entirely when the two signals are in lock. This means that there is no "ripple" present at the output to generate periodic phase modulation in the loop, as there is with the type I phase detector.

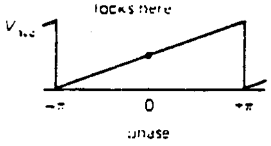


Figure 9.54

Here is a comparison of the properties of the two basic types of phase detector:

	Type I Exclusive-OR	*Type II edge-triggered ("charge pump")*
Input duty cycle	50% optimum	irrelevant
Lock on harmonic?	yes	no
Rejection of noise	good	poor
Residual ripple at $2f_{IN}$	high	low
Lock range (L)	full VCO range	full VCO range
Capture range	fL ($f < 1$)	L
Output frequency when out of lock	f_{center}	f_{min}

There is one additional point of difference between the two kinds of phase detectors. The type I detector is always generating an output wave, which must then be filtered by the loop filter (much more on this later). Thus in a PLL with type I phase detector the loop filter acts as a low-pass filter, smoothing this full-swing logic-output signal. There will always be residual ripple, and consequent periodic phase variations, in such a loop. In circuits where phase-locked loops are used for frequency multiplication or synthesis, this adds "phase-modulation sidebands" to the output signal (see Section 13.17).

By contrast, the type II phase detector generates output pulses only when there is a phase error between the reference and VCO signal. Since the phase detector output otherwise looks like an open circuit, the loop filter capacitor then acts as a voltage-storage device, holding the voltage that gives the right VCO frequency. If the reference signal moves away in frequency, the phase detector generates a train of short pulses, charging (or discharging) the capacitor to the new voltage needed to put the VCO back into lock.

□ *VCOs*

An essential component of a phase-locked loop is an oscillator whose frequency can be controlled by the phase detector output. Some PLL ICs include a VCO (e.g., the linear

BATCH
START

Lecture 10

Control Systems for Test-mass Position & Orientation

STAPLE
OR
DIVIDER

LECTURE 12.

Seismic Isolation

Lecture by Lisa Sievers

Assigned Reading:

II. Leonard Meirovitch, *Elements of Vibration Analysis* (McGraw-Hill, 1986), pp. 48–58. [This reference develops the basic concepts of using mass-spring-damper systems for vibration isolation; and it discusses the measurement of vibrations, and two types of damping that can occur in mechanical systems: viscous damping and structural damping. These two types of damping will play an important role in the lectures on thermal noise next week.]

JJ. R. del Fabbro, A. di Virgilio, A. Giazotto, H. Kautzky, V. Montelatici, and D. Passuello, "Three-dimensional seismic super-attenuator for low frequency gravitational wave detection," *Physics Letters A*, **124**, 253–257 (1987). [This reference describes and analyzes an early version of the ambitious mass-spring-damper vibration-isolation stack that is being developed by the Pisa, Italy group as their prime contribution to the VIRGO Project. The analysis of the LIGO isolation stacks is similar, though their initial design is less ambitious.]

Suggested Supplementary Reading:

II. Leonard Meirovitch, *Elements of Vibration Analysis* (McGraw-Hill, 1986), pp. 39–48. [This is largely foundational material underlying the assigned reading (item 1. above); you may find it helpful.

KK. C. A. Cantley, J. Hough, and N. A. Robertson, "Vibration isolation stacks for gravitational wave detectors—Finite element analysis," *Rev. Sci. Instrum.*, **63**, 2210–2219 (1992). [This paper, by the Glasgow gravity-wave group, illustrates an isolation-stack analysis that is more sophisticated than the simple models used in class and in reference 2, and that reveals pitfalls in the design of a stack.]

LL. M. Stephens, P. Saulson, and J. Kovalik, "A double pendulum vibration isolation system for a laser interferometric gravitational wave antenna," *Rev. Sci. Instrum.*, **62**, 924–932 (1991). [This paper, passed out for other reasons in Lecture 10, analyses the use of compound pendula for vibration isolation.]

MM. L. Ju, D. G. Blair, H. Peng, and F. van Kann, "High dynamic range measurements of an all metal isolator using a sapphire transducer," *Mass. Sci. Technol.*, **3**, 463–470 (1992). [This paper, by the Perth resonant-bar gravitational-wave-detector group, describes a type of all-metal isolator which might be a precursor to an isolation stack for advanced LIGO detectors; see transparencies 20, 21, and 22 of Sievers' lecture.]

A Few Suggested Problems: See the next page.

1. You have the 2 stage spring/mass stack shown in Figure 1 and want to decide the best mass ratio m_1/m_2 in the 2 stages so that you achieve maximum isolation at frequencies well above $w_o^2=k_1/m_1$. A good designer would assume that the springs are compressed to their maximum limit in order to get the most bang for their buck, therefore the strain energy in each spring should be assumed equal ($k_1/m_1 = k_2/(m_1+m_2)$). Show that the transmissibility $X_1(f)/X_g(f)$ is maximized as the mass ratio m_1/m_2 goes to zero but that a point of diminishing returns is reached when the ratio is about 1.

2. Work out the equations of motion for the 1 and 2 stage pendula shown in Figure 2. Compare the amount of isolation achieved at two different frequencies: $\omega = 2\sqrt{\frac{g}{l}}$ and $\omega = 10\sqrt{\frac{g}{l}}$

3. A method for mechanically damping a high Q mechanical resonance is to use a "proof mass damper" as shown in Figure 3. The proof mass damper is a damped oscillator whose mass is much smaller than the mass to be damped and whose resonant frequency and damping coefficient is tuned specifically to damp the system in the most effective way. Assume m_1=20m_2, f_1=4Hz, and $f_2=(\frac{1}{1+m_2/m_1})f_1$. Plot the transmissibility function $X_1(f)/X_g(f)$ for 3 different damping coefficients, c. [Definition of damping coefficient, c: If a particle of mass m moves under the combined influence of a linear restoring force -kx and a resisting force $-c\dot{x}$, the differential equation which describes the motion is $m\ddot{x} + c\dot{x} + kx = 0$ .... c is inversely proportional to Q]

   1. c=0
   2. c=infinity
   3. $c=2m_2(2\pi f_1)\sqrt{\frac{3m_2/m_1}{8(1+m_2/m_1)^3}}$

   The third case is the case where you get the maximum attenuation possible (i.e. $X_1(f_1)/X_g(f_1)=\sqrt{1+2m_1/m_2}$)

   [A proof mass damper has been experimentally implemented in Mark I. One of the stacks (i.e. optics plate mounted on rubber), had a high Q horizontal resonance at f_1=4Hz. In a compact vacuum sealed vessel, we built a pendulum whose bob was 1/20 the mass of the offending optics plate. The pendulum was partially submerged in motor oil whose damping coefficient was given in (3). The length of the pendulum bob was tuned to the resonant frequency of f_2. The stack resonance was damped without compromising the isolation at higher frequencies.]

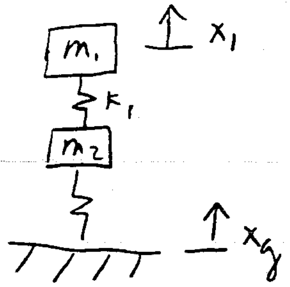


Figure 1

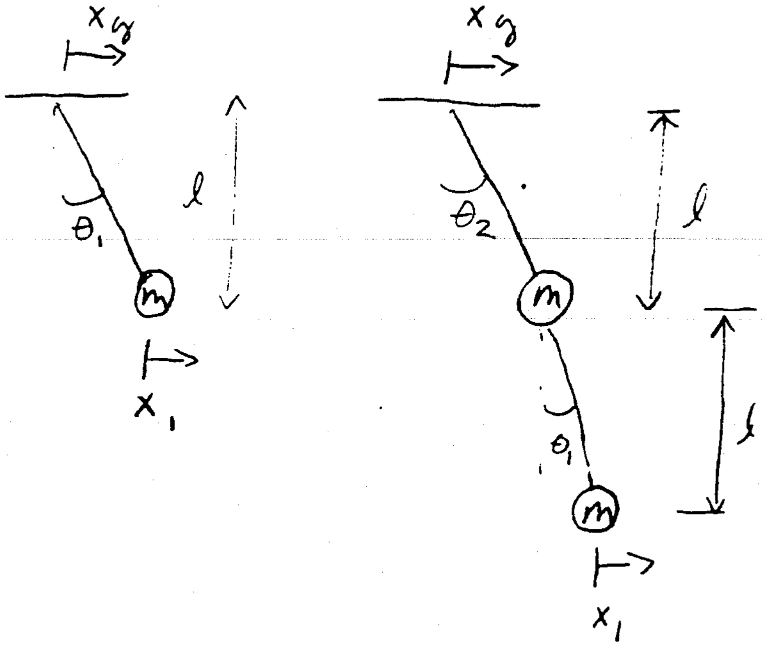
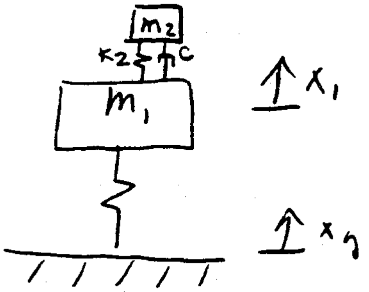


Figure 2



$$2\pi f_1 = \sqrt{\frac{k_1}{m_1}}$$

$$2\pi f_2 = \sqrt{\frac{k_2}{m_2}}$$

Figure 3

LECTURES 13 & 14

Thermal Noise

Lectures by Aaron Gillespie

Assigned Reading:

NN. Herbert B. Callen and Theodore A. Welton, "Irreversibility and generalized noise," *Phys. Rev.* **83**, 34–40 (1951). [This paper derives a generalized version of the fluctuation-dissipation theorem (Nyquist's theorem), cf. Lecture 2. The terminology and notation are quite different from what is now standard. Equations (4.8) in the quantum regime and (4.11) in the classical limit describe the mean square value $\langle V^2 \rangle$ of the fluctuating "generalized force" V in terms an integral over the real part $R(\omega)$ of a complex generalized impedance $Z(\omega)$. In modern language, one switches from ω to $f = \omega/2\pi$ and thereby rewrites (4.11) as $\langle V^2 \rangle = 4kT \int R(f) df$, and one then identifies the contribution at frequency f as the "Spectral density" of V:

$$G_V(f) \equiv S_V(f) \equiv \tilde{V}^2(f) = 4kTR(f); \tag{1}$$

and similarly for the quantum-regime formula (4.8).]

OO. Peter R. Saulson, "Thermal noise in mechanical experiments," *Phys. Rev. D* **42**, 2437–2445 (1990). [This paper applies the generalized fluctuation-dissipation theorem of paper 1. to thermal noise in a mechanical oscillator. The fluctuation-dissipation theorem is Eq. (5) of this paper; and it implies that the spectral density of the oscillator's displacement $x(t)$ has the form (16),

$$G_x(f) = \tilde{x}^2(f) = \frac{4k_BTk\phi(\omega)}{\omega[(k - m\omega^2)^2 + k^2\phi^2]}. \tag{2}$$

Here $k = m\omega_o^2$ is the oscillator's spring constant with ω_o its angular eigenfrequency; $\omega \equiv 2\pi f$ is angular frequency; and $k\phi(\omega)$ is $R(\omega)$, the real part of the generalized impedance. The key issue raised in this paper is "What is the frequency dependence of $\phi(\omega)$?" For viscous damping, $\phi \propto \omega$; for structural damping, ϕ is independent of ω.]

PP. Aaron Gillespie and Frederick Raab, "Thermal noise in the test mass suspensions of a laser interferometer gravitational-wave detector prototype," *Phys. Lett. A*, **178**, 357–363 (1993). [In this paper strong evidence is given that for the flexural motion of the wire from which a test mass hangs, the damping is structural, i.e. ϕ is independent of ω. Note that in this paper the phrase "lineshape" is sometimes used for the spectral density $\tilde{x}^2(f)$.]

Suggested Supplementary Reading:

e. More on the fluctuation-dissipation theorem:

Herbert B. Callen and Richard F. Greene, "On a theorem of irreversible thermodynamics," *Phys. Rev.*, **86**, 702 (1952).

f. Elasticity theory:

L. D. Landau and E. M. Lifshitz, *Theory of Elasticity* (Pergamon Press, New York, 1959).

g. Losses in materials:

g1. A.S. Nowick and B.S. Berry, *Anelastic Relaxation in Crystalline Solids*, (Academic Press, New York, 1972). [A good, general book.]

g2. A.L. Kimball and D.E. Lovell, "Internal friction in solids," *Phys. Rev.* **30** 948 (1927). [An early reference for losses independent of frequency, i.e. structural damping, in solid materials.]

g3. Clarence Zener, "Internal friction in solids: I. Theory of internal friction in reeds," *Phys. Rev.* **52** 230 (1937); "Internal friction in solids: II. General theory of thermoelastic internal friction," *Phys. Rev.* **53**, 90 (1938). [The original references for thermoelastic damping.]

h. Vibrations of Cylinders:

QQ. Aaron Gillespie and Frederick Raab, "Thermally excited vibrations of the mirrors of a laser interferometer gravitational-wave detector," unpublished (1994).

h2. James R. Hutchinson, "Vibrations of solid cylinders," *J. Appl. Mech.*, **47**, 901 (1980).

h3. James R. Hutchinson, "Axisymmetric vibrations of free finite-length rod," *J. Acoust. Soc. Am.* **51**, 233 (1972).

h4. G. W. McMahon, "Experimental study of vibrations of solid, isotropic elastic cylinders," J. Acoust. Soc. Am. **36**, 85 (1964).

i. "Exact" solution for a pendulum with a finite size, lossy wire:

Gabriela I. Gonzalez and Peter R. Saulson, "Brownian motion of a mass suspended by an anelastic wire," *J. Acoust. Soc. Am.*, in press (1994). [See Aaron Gillespie (x2128) for a copy.]

j. Ultra-High Q pendula:

j1. V.B. Braginsky, V.P. Mitrofanov, and O.A. Okhrimenko, "Pendulum fused silica oscillators with small dissipation," *Phys. Lett. A*, **175**, 82 (1993).

j2. D.G. Blair, L. Ju, and M. Notcutt, "Ultra high Q pendulum suspensions for gravitational wave detectors," *Rev. Sci. Instrum.*, **64**, 1899 (1993).

k. Some current experimental work:

k1. J.E. Logan, N.A. Robertson, J. Hough, and P.J. Veitch, "An investigation of coupled resonances in materials suitable for test masses in graviational wave detectors," *Phys. Lett. A* **161**, 101 (1991).

k2. J.E. Logan, N.A. Robertson, and J. Hough, "An investigation of limitations to quality factor measurements of suspended masses due to resonances in the suspension wires," *Phys. Lett. A*, **170**, (1992).

RR. A. Gillespie and F. Raab, "Suspension losses in the pendula of laser interferometer gravitational wave detectors," *Phys Lett A*, in press (1994).

A Few Suggested Problems:

1. *Relationship between the equipartition theorem and the fluctuation-dissipation theorem.* Consider a pendulum with mass $m = 1kg$ and swing frequency $f_o = 2\pi\omega_o = 1$ Hz, and with damping such that, when it is driven at angular frequency ω, its equation of motion is

$$m\ddot{x} = -k(1 + i\phi(\omega))x + Fe^{i\omega t}; \qquad (3)$$

where $k = m\omega_o^2$ and x is the transverse position of the pendulum mass. Recall from Gillespie's lecture or the above assigned reading that the pendulum's position will exhibit fluctuations with spectral density given by Eq. (2) above.

   a. If the pendulum is set swinging freely, what is its damping time, i.e. the time τ_* for its energy of swing to be damped by $1/e$?

   b. Take $\phi(\omega) = 10^{-7}\omega/\omega_o$, corresponding to weak frictional damping. Integrate the thermal noise spectrum to find the pendulum's rms velocity $v_{\rm rms} = \langle\dot{x}^2\rangle^{1/2}$. Compare your result with the rms velocity predicted by the equipartition theorem.

   c. What is the full width at half maximum (FWHM) of the thermal noise spectral density $\tilde{x}^2(f)$ in terms of f_o and $\phi(\omega)$?

   d. What fraction of the pendulum's total rms energy lies in the frequency band defined by the FWHM around the resonant frequency? What fraction lies within 10 FWHM?

   e. Take $\phi(\omega) = 10^{-7}$ independent of ω, corresponding to structural damping. Notice that at $\omega \ll \omega_o$, the pendulum's motions are characterized by flicker noise, $\tilde{x}^2(f) \propto 1/f$, and that as a result the integral of the spectral density diverges. Can you explain physically how this comes about? Take as a lower cutoff frequency the inverse of the lifetime of the universe (10^{10} years). What then is the pendulum's rms velocity?

2. *Johnson noise and thermal noise due to eddy current damping.* In Gillespie's lecture the eddy current damping of a pendulum due to a simple current loop and a resistor was found; see his transparency number 20.

   a. What is the spectral density of the pendulum's displacement, $\tilde{x}^2(f)$?

   b. An electrical resistor has Johnson noise, $\tilde{V}^2(f) = 4k_BTR$ where R is its resistance. Compute the pendulum's spectral density $\tilde{x}^2(f)$ due to the Johnson noise associated with the flow of current in the resistor.

   c. Is there a difference, physically, between the damping processes in parts a. and b.?

3. *Pendulum losses due to flexing of the wire material.* This problem examines how the losses in the pendulum due to flexing of the wire material and the associated thermal noise scale with the parameters of the pendulum.

   a. Show that, for a pendulum of fixed mass, the thinner one makes the support wire, the weaker will be the damping of the pendulum's swing.

   For parts b–d, assume that the wire is stressed to its maximum safe value, i.e. that its tension per unit area is held at the maximum (a fixed constant independent of the pendulum's other parameters).

b. How do the pendulum's losses, i.e. $\phi(\omega)$, scale with its mass? How does the off-resonance thermal noise scale with the mass?

c. One way of getting high Q pendulums is to use ribbons with rectangular cross sections rather than circular wires. How do the losses in the pendulum scale with the ribbon thickness (its short dimension)? How does the off-resonance thermal noise scale?

d. How do the losses and thermal noise in the vertical mode scale with mass? with ribbon thickness?

4. *Effective mass coefficients in a simple model of a test mass.* The concept of "effective mass coefficients" was introduced in Gillespie's lecture (see his transparency numbers 40-41 and also see reference 7.a of the supplementary reading). Explicit calculation of these effective mass coefficients is a recent development in the gravitational-wave field. Previously, experimenters used a model which assumed that the laser was an ideal one-dimensional beam and the mass was one dimensional. In this model, the modes can be found by solving the one-dimensional acoustic wave equation:

$$\frac{\partial^2 u}{\partial z^2} = \frac{1}{c^2}\frac{\partial^2 u}{\partial t^2}, \tag{4}$$

where c is the sound velocity and $u(z,t)$ is the longitudinal displacement of the mass's material.

a. What are the eigenfunctions of the modes? (The boundary condition is no stress at the end faces: $\partial u/\partial z = 0$ at $z = \pm h/2$ where h is the length of the mass.)

b. What are the effective mass coefficients as a function of ω_n the resonant frequency of the $n'th$ mode? How do these compare to the actual effective mass coefficients of the 40m prototype's test masses for the lowest-frequency mode, ω_0? for higher-frequency modes?

c. What is the mode density $\rho(\omega)$?

d. What is the low-frequency ($\omega \ll \omega_0$) thermal noise as a function of the highest frequency mode included in the analysis?

e. Experimenters knew that this model was flawed in that the axisymmetric modes comprised a two-dimensional system ($\rho(\omega) \propto \omega$), so they used the effective mass coefficients of the one-dimensional model but changed the mode density to $\rho(\omega) \propto \omega$. What is the low-frequency thermal noise in this case as a function of the highest frequency mode included?

KEF 11

McGRAW-HILL
BOOK COMPANY
New York
St. Louis
San Francisco
Auckland
Düsseldorf
Johannesburg
Kuala Lumpur
London
Mexico
Montreal
New Delhi
Panama
Paris
São Paulo
Singapore
Sydney
Tokyo
Toronto

LEONARD MEIROVITCH
College of Engineering
Virginia Polytechnic Institute and State University

Elements of Vibration Analysis

To my wife and
to the memory of my parents

This book was set in Times New Roman.
The editors were B. J. Clark and M. E. Margolies;
the cover was designed by S. Bernstein;
the production supervisor was Sam Ratkewitch.
The drawings were done by ANCO Technical Services.
Kingsport Press, Inc., was printer and binder.

Library of Congress Cataloging in Publication Data

Meirovitch, Leonard.
Elements of vibration analysis.
Bibliography: p.
1. Vibration. I. Title.
QA935.M53 531'.32 74-13980
ISBN 0-07-041340-1

ELEMENTS
OF
VIBRATION
ANALYSIS

Copyright © 1975 by McGraw-Hill, Inc. All rights reserved.
Printed in the United States of America. No part of this publication may be reproduced,
stored in a retrieval system, or transmitted, in any form or by any means,
electronic, mechanical, photocopying, recording, or otherwise,
without the prior written permission of the publisher.

101112 KPKP 8987654

McGRAW-HILL
BOOK COMPANY
New York
St. Louis
San Francisco
Auckland
Düsseldorf
Johannesburg
Kuala Lumpur
London
Mexico
Montreal
New Delhi
Panama
Paris
São Paulo
Singapore
Sydney
Tokyo
Toronto

LEONARD MEIROVITCH
College of Engineering
Virginia Polytechnic Institute and State University

Elements of Vibration Analysis

2

FORCED VIBRATION OF SINGLE-DEGREE-OF-FREEDOM LINEAR SYSTEMS

2.1 GENERAL CONSIDERATIONS

A very important subject in vibrations is the response of systems to external excitations. The excitations, for example, can be in the form of initial displacements, initial velocities, or both. Following initial excitation alone, the system vibrates freely, for which reason such motion is referred to as *free vibration*, a subject discussed in Chap. 1. However, the excitation can also be in the form of forces which persist for an extended period of time. Such vibration is called *forced vibration* and is the subject of this chapter. For linear systems it is possible to obtain the response to initial conditions and external forces separately, and then combine them to obtain the total response of the system. This is based on the so-called principle of superposition.

The procedure for obtaining the response of a system to external forces depends to a large extent on the type of excitation. In this chapter we follow a pattern of increasing complexity, beginning the discussion with the simple harmonic excitation, extending it to periodic excitation, and culminating with nonperiodic excitation. Because of its fundamental nature and because it has a multitude of practical applications, the case of harmonic excitation is discussed in great detail. The principle of superposition receives special attention, as it forms the basis for the analysis of linear systems. A rigorous discussion of the principle is provided. The case of periodic excitation can be reduced to that of harmonic excitation by regarding the periodic forcing function as a superposition of harmonic functions through the use of standard Fourier series. To discuss the response to nonperiodic excitation, the impulsive response and convolution integral are introduced. Finally, the system response by the Fourier transformation and the Laplace transformation is introduced, and the many advantages of the latter method are pointed out.

2.2 RESPONSE TO HARMONIC EXCITATION. FREQUENCY RESPONSE

Let us consider the damped second-order linear system depicted in Fig. 1.6. The differential equation of motion of the system, Eq. (1.8), is

$$m\ddot{x}(t) + c\dot{x}(t) + kx(t) = F(t) \tag{2.1}$$

where all the quantities are as defined in Sec. 1.3. The homogeneous solution of Eq. (2.1), obtained by letting $F(t) = 0$, was discussed in Chap. 1, and will not be repeated here. We shall concentrate our attention on the particular solution instead. First we wish to consider the simplest case, namely, the response of the system to harmonic excitation. To this end, we let the force have the form

$$F(t) = kf(t) = kA \cos \omega t \tag{2.2}$$

where the excitation frequency ω is sometimes referred to as the *driving frequency*. Note that $f(t)$ and A have units of displacement. The introduction of the function $f(t)$ may appear artificial at this point, but it permits the construction of a nondimensional ratio of response to excitation, as we shall see later. Nondimensional ratios often enhance the usefulness of a particular analysis by extending its applicability to a larger variety of cases. Inserting Eq. (2.2) into (2.1), and dividing through by m, we obtain

$$\ddot{x}(t) + 2\zeta\omega_n\dot{x}(t) + \omega_n{}^2x(t) = \frac{k}{m}f(t) = \omega_n{}^2A \cos \omega t \tag{2.3}$$

The solution of the homogeneous differential equation (obtained by letting $A = 0$) dies out with time for $\zeta > 0$, for which reason it is called the *transient solution*. On the other hand, the particular solution does not vanish for large t and is known as the *steady-state solution* to the harmonic excitation in question. By virtue of the fact that the system is linear, the principle of superposition holds (see Sec. 2.7), so that the transient and steady-state solutions can be obtained separately and then combined to obtain the complete solution.

Because the excitation force is harmonic, it can be verified easily that the steady-state response $x(t)$ is also harmonic and has the same frequency ω. Noting

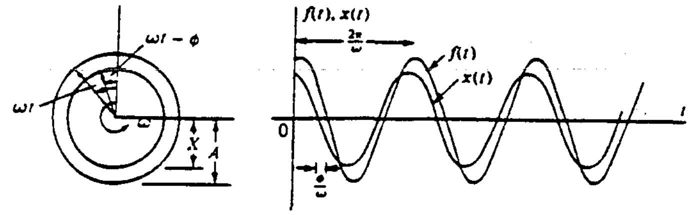


FIGURE 2.1

that the left side of Eq. (2.3) contains both even- and odd-order time derivatives of $x(t)$, we assume a solution in the form

$$x(t) = X \cos(\omega t - \phi). \tag{2.4}$$

where X and ϕ are the amplitude and phase angle of the response, respectively, quantities that must yet be determined. Inserting solution (2.4) into Eq. (2.3), we arrive at

$$X[(\omega_n^2 - \omega^2)\cos(\omega t - \phi) - 2\zeta\omega_n\omega \sin(\omega t - \phi)] = \omega_n^2 A \cos \omega t \tag{2.5}$$

and, recalling the trigonometric relations

$$\cos(\omega t - \phi) = \cos \omega t \cos \phi + \sin \omega t \sin \phi$$
$$\sin(\omega t - \phi) = \sin \omega t \cos \phi - \cos \omega t \sin \phi$$

we can equate the coefficients of $\cos \omega t$ and $\sin \omega t$ on both sides of Eq. (2.5), with the result

$$\begin{aligned} X[(\omega_n^2 - \omega^2)\cos\phi + 2\zeta\omega_n\omega \sin\phi] &= \omega_n^2 A \\ X[(\omega_n^2 - \omega^2)\sin\phi - 2\zeta\omega_n\omega \cos\phi] &= 0 \end{aligned} \tag{2.6}$$

Solving Eqs. (2.6), we obtain

$$X = \frac{A}{\{[1 - (\omega/\omega_n)^2]^2 + (2\zeta\omega/\omega_n)^2\}^{1/2}} \tag{2.7}$$

and

$$\phi = \tan^{-1} \frac{2\zeta\omega/\omega_n}{1 - (\omega/\omega_n)^2} \tag{2.8}$$

which, when inserted into Eq. (2.4), completes the particular solution of Eq. (2.3). Typical plots of the excitation and response are shown in Fig. 2.1.

At this point, it appears desirable to represent the harmonic excitation by a complex vector, because such a representation possesses many advantages in the derivation of the response. To show this, let us recall from Sec. 1.5 that

$$e^{i\omega t} = \cos \omega t + i \sin \omega t \tag{2.9}$$

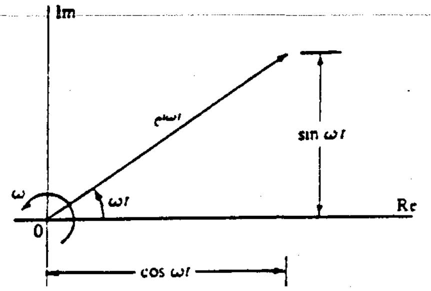


FIGURE 2.2

Equation (2.9) is represented graphically in the complex plane of Fig. 2.2, from which we conclude that $e^{i\omega t}$ can be regarded as a complex vector of unit magnitude rotating in the complex plane with angular velocity ω. Hence, $\cos \omega t$ is simply the projection of the complex vector $e^{i\omega t}$ on the real axis, i.e., the real part of $e^{i\omega t}$.

Let us consider again the damped second-order linear system described by Eq. (2.1), and represent the excitation by the complex form

$$f(t) = Ae^{i\omega t} \tag{2.10}$$

with the tacit understanding that *the excitation is given only by the real part of $f(t)$*. Then *the response will also be given only by the real part of $x(t)$*, where $x(t)$ is the solution of Eq. (2.1) subject to the excitation in the form (2.10). The quantity A is generally a complex number.

In view of the above discussion, the response will be regarded as the real part of $x(t)$, where $x(t)$ is a complex quantity satisfying the differential equation

$$\ddot{x}(t) + 2\zeta\omega_n\dot{x}(t) + \omega_n^2 x(t) = \omega_n^2 f(t) = \omega_n^2 Ae^{i\omega t} \tag{2.11}$$

Concentrating on the steady-state solution, we can easily show that it has the form

$$x(t) = \text{Re}\left(\frac{\omega_n^2 Ae^{i\omega t}}{\omega_n^2 - \omega^2 + i2\zeta\omega\omega_n}\right) = \text{Re}\left[\frac{Ae^{i\omega t}}{1 - (\omega/\omega_n)^2 + i2\zeta\omega/\omega_n}\right] \tag{2.12}$$

where Re designates the real part of the expression inside brackets. We notice from Eq. (2.12) that the response $x(t)$ is proportional to the force $f(t) = F(t)/k$, the proportionality factor being

$$H(\omega) = \frac{1}{1 - (\omega/\omega_n)^2 + i2\zeta\omega/\omega_n} \tag{2.13}$$

which is known as the *complex frequency response*. Recalling that the force in the spring is $F_s(t) = kx(t)$, we conclude from Eqs. (2.12) and (2.13) that

$$H(\omega) = \frac{x(t)}{f(t)} = \frac{kx(t)}{F(t)} = \frac{F_s(t)}{F(t)} \tag{2.14}$$

or the complex frequency response can be identified as the nondimensional ratio between the force in the spring and the actual excitation force $F(t)$. In defining

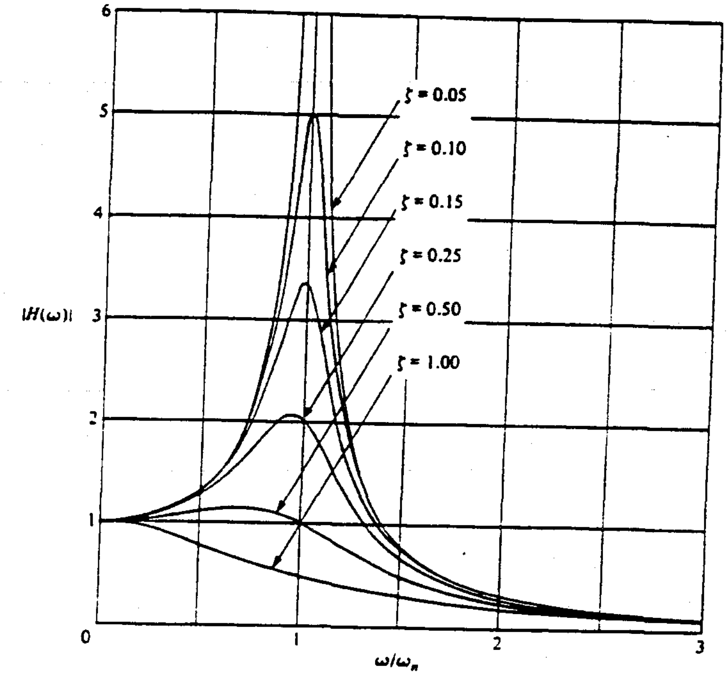


FIGURE 2.3

the force in the spring, it is perhaps worth reminding ourselves that $x(t)$ is measured from the static equilibrium position.

From complex algebra, we conclude that the absolute value of $H(\omega)$, called the *magnification factor*, is equal to the nondimensional ratio between the amplitude X of the response $x(t)$ and the amplitude A of the excitation $f(t)$ [see Eq. (2.7)], namely,

$$|H(\omega)| = \frac{1}{\{[1 - (\omega/\omega_n)^2]^2 + [2\zeta(\omega/\omega_n)]^2\}^{1/2}} \tag{2.15}$$

Figure 2.3 shows plots of $|H(\omega)|$ versus ω/ω_n for various values of ζ, which permit the observation that damping tends to diminish amplitudes and to shift the peaks to the left of the vertical through $\omega/\omega_n = 1$. To find the values at which the peaks of the curves occur, we use the standard technique of calculus for finding stationary values of a function, namely, we differentiate Eq. (2.15) with respect to ω and set the result equal to zero. This leads us to the conclusion that the peaks occur at

$$\omega = \omega_n(1 - 2\zeta^2)^{1/2} \tag{2.16}$$

indicating that the maxima do not occur at the undamped natural frequency ω_n but for values $\omega/\omega_n < 1$, depending on the amount of damping. Clearly, for

$\zeta > 1/\sqrt{2}$ the response has no peaks, and for $\zeta = 0$ there is a discontinuity at $\omega/\omega_n = 1$. The case $\zeta = 0$ corresponds to the undamped case in which the homogeneous differential equation reduces to that of a harmonic oscillator, leading us to the conclusion that when the driving frequency ω approaches the natural frequency ω_n the response of the harmonic oscillator tends to increase indefinitely. In such a case the harmonic oscillator is said to approach a *resonance condition* characterized by violent vibration. However, solution (2.12) is no longer valid at resonance; a new solution of Eq. (2.3) corresponding to $\omega = \omega_n$ is obtained later in this section.

We notice that for light damping, such as when $\zeta < 0.05$, the maximum of $|H(\omega)|$ occurs in the immediate neighborhood of $\omega/\omega_n = 1$. Introducing the notation $|H(\omega)|_{\max} = Q$, we obtain for small values of ζ

$$Q \cong \frac{1}{2\zeta} \tag{2.17}$$

and the curves $|H(\omega)|$ versus ω/ω_n are nearly symmetric with respect to the vertical through $\omega/\omega_n = 1$ in that neighborhood. The symbol Q is known as the *quality factor* because in many electrical engineering applications, such as the tuning circuit of a radio, the interest lies in an amplitude at resonance that is as large as possible. The symbol is often referred to as the "Q" of the circuit. The points P_1 and P_2, where the amplitude of $|H(\omega)|$ falls to $Q/\sqrt{2}$, are called *half power points* because the power absorbed by the resistor in an electric circuit or by the damper in a mechanical system responding harmonically at a given frequency is proportional to the square of the amplitude (see Sec. 2.6). The increment of frequency associated with the half power points P_1 and P_2 is referred to as the *bandwidth* of the system. For light damping, it is not difficult to show that the bandwidth has the value

$$\Delta\omega = \omega_2 - \omega_1 \cong 2\zeta\omega_n \tag{2.18}$$

Moreover, comparing Eqs. (2.17) and (2.18), we conclude that

$$Q \cong \frac{1}{2\zeta} \cong \frac{\omega_n}{\omega_2 - \omega_1} \tag{2.19}$$

which can be used as a quick way of estimating Q or ζ.

At this point let us turn our attention to the phase angle. If we recall that $e^{i\phi} = \cos\phi + i\sin\phi$, we can use expressions (2.13) and (2.15) and show without much difficulty that

$$H(\omega) = |H(\omega)|e^{-i\phi} \tag{2.20}$$

where

$$\phi = \tan^{-1}\frac{2\zeta\omega/\omega_n}{1 - (\omega/\omega_n)^2} \tag{2.21}$$

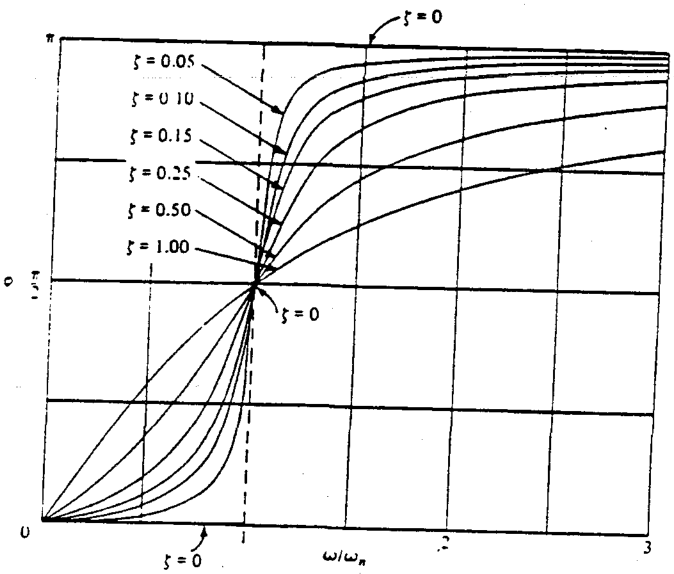


FIGURE 2.4

which is the same as Eq. (2.8) obtained previously. In view of Eqs. (2.20) and (2.21), solution (2.12) can be written in the form

$$x(t) = \mathrm{Re}\,[AH(\omega)e^{i\omega t}] = \mathrm{Re}\,[A|H(\omega)|e^{i(\omega t-\phi)}] \tag{2.22}$$

from which we conclude that ϕ is the phase angle, namely, the angle between the excitation and the response (see Sec. 2.3). Figure 2.4 plots ϕ versus ω/ω_n for selected values of ζ. We notice that all curves pass through the point $\phi = \pi/2$, $\omega/\omega_n = 1$. Moreover, for $\omega/\omega_n < 1$ the phase angle tends to zero, whereas for $\omega/\omega_n > 1$ it tends to π.

For $\zeta = 0$ the plot ϕ versus ω/ω_n has a discontinuity at $\omega/\omega_n = 1$, jumping from $\phi = 0$ for $\omega/\omega_n < 1$ to $\phi = \pi$ for $\omega/\omega_n > 1$. This can be easily explained because for $\zeta = 0$ solution (2.22) reduces to

$$x(t) = \mathrm{Re}\left[\frac{1}{1-(\omega/\omega_n)^2}Ae^{i\omega t}\right] \tag{2.23}$$

so that the response is in phase for $\omega/\omega_n < 1$ and 180° out of phase for $\omega/\omega_n > 1$. Equation (2.23) also shows clearly that the response of a harmonic oscillator increases without bounds as the driving frequency ω approaches the natural frequency ω_n.

Finally, let us consider the case of the harmonic oscillator at resonance. In this case the differential equation of motion, Eq. (2.3), reduces to

$$\ddot{x}(t) + \omega_n^2 x(t) = \omega_n^2 A \cos \omega_n t \tag{2.24}$$

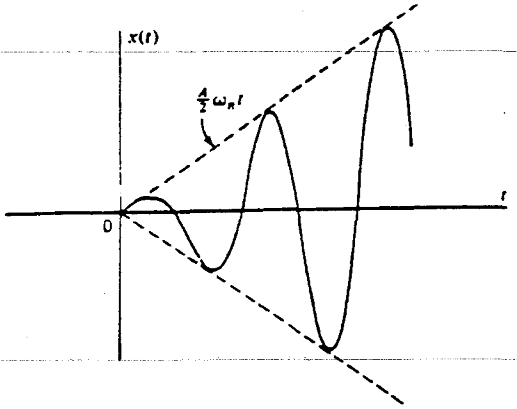


FIGURE 2.5

Note that, because the velocity term is zero, there is no need to use the complex vector form for the excitation and response. It is not difficult to verify by substitution that the particular solution of Eq. (2.24) is

$$x(t) = \frac{A}{2}\,\omega_n t \sin \omega_n t \tag{2.25}$$

which represents oscillatory response with an amplitude increasing linearly with time. This implies that the response undergoes wild fluctuations as t becomes large. Physically, however, the response cannot grow indefinitely, as at a certain time the small-motions assumption implicit in linear systems is violated. Because the excitation is a cosine function and the response is a sine function, there is a 90° phase angle between them, as can also be concluded from Fig. 2.4. The response $x(t)$, as given by Eq. (2.25), is plotted in Fig. 2.5 as a function of time.

The complete solution of Eq. (2.3) is obtained by adding the homogeneous solution, Eq. (1.28), to the particular solution, Eq. (2.22). To avoid confusion, we shall change the notation of the amplitude in Eq. (1.28) from A to C and of the phase angle from ϕ to ψ. Moreover, we shall assume without loss of generality that A in Eq. (2.22) is a real quantity, so that the complete solution becomes

$$x(t) = A|H(\omega)| \cos (\omega t - \phi) + Ce^{-\zeta\omega_n t} \cos (\omega_d t - \psi) \tag{2.26}$$

It is clear that the first term on the right of Eq. (2.26) represents the steady-state solution, whereas the second term represents the transient solution.

Example 2.1 Rotating Unbalanced Masses

As an illustration of a system subjected to harmonic excitation, we consider the case of Fig. 2.6, in which two eccentric masses $m/2$ rotate in opposite directions with constant angular velocity ω. The reason for having two equal masses rotating

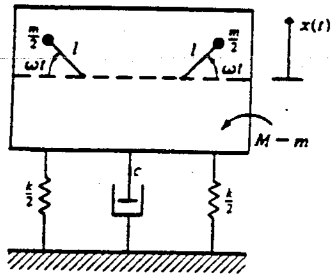


FIGURE 2.6

in opposite directions is that the horizontal components of excitation of the two masses cancel each other. On the other hand, the vertical components of excitation add. The vertical displacement of the eccentric masses is $x + \ell \sin \omega t$, where x is measured from the equilibrium position. In view of this, it is not difficult to show that the differential equation of the system is

$$(M - m)\frac{d^2x}{dt^2} + m\frac{d^2}{dt^2}(x + \ell \sin \omega t) + c\frac{dx}{dt} + kx = 0 \tag{a}$$

which can be rewritten in the form

$$M\ddot{x}(t) + c\dot{x}(t) + kx(t) = m\ell\omega^2 \sin \omega t = \mathrm{Im}(m\ell\omega^2 e^{i\omega t}) \tag{b}$$

where Im denotes the imaginary part of the expression within parentheses. From Eq. (2.22) we conclude that the response is

$$\begin{aligned} x(t) &= \mathrm{Im}\left[\frac{m}{M}\ell\left(\frac{\omega}{\omega_n}\right)^2 |H(\omega)|e^{i(\omega t - \phi)}\right] \\ &= \frac{m}{M}\ell\left(\frac{\omega}{\omega_n}\right)^2 |H(\omega)| \sin(\omega t - \phi), \qquad \omega_n^2 = \frac{k}{M} \end{aligned} \tag{c}$$

in which the phase angle ϕ is given by Eq. (2.21). Writing the response in the form

$$x(t) = X \sin(\omega t - \phi) \tag{d}$$

we conclude that

$$X = \frac{m\ell}{M}\left(\frac{\omega}{\omega_n}\right)^2 |H(\omega)| \tag{e}$$

Hence, in this particular case the indicated nondimensional ratio is

$$\frac{MX}{m\ell} = \left(\frac{\omega}{\omega_n}\right)^2 |H(\omega)| \tag{f}$$

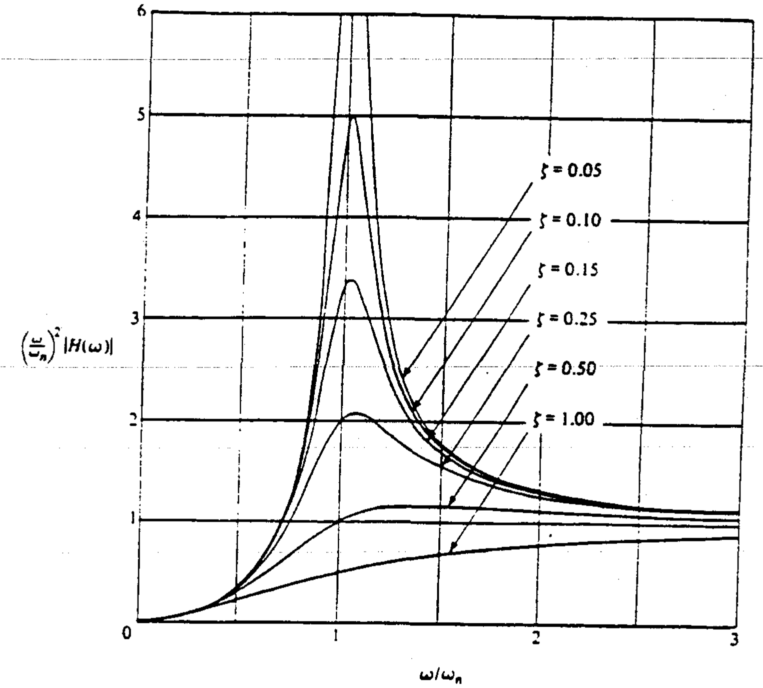


FIGURE 2.7

instead of $|H(\omega)|$ alone, so that Fig. 2.3 is not applicable. Plots of $(\omega/\omega_n)^2|H(\omega)|$ versus ω/ω_n with ζ as a parameter are shown in Fig. 2.7. On the other hand, the plot ϕ versus ω/ω_n remains as in Fig. 2.4.

We note that for $\omega \to 0$, $(\omega/\omega_n)^2|H(\omega)| \to 0$, whereas for $\omega \to \infty$, $(\omega/\omega_n)^2|H(\omega)| \to 1$. At the same time, from Eq. (2.21), we conclude that as $\omega \to \infty$, $\phi \to \pi$. Since the mass $M - m$ undergoes the displacement Im x, whereas the mass m undergoes the displacement $\text{Im}(x + \ell e^{i\omega t})$, it follows that for large driving frequencies ω the masses $M - m$ and m move in such a way that the mass center of the system tends to remain stationary. This is true regardless of the amount of damping. Note that $\text{Im}\, x = -X \sin \omega t$ for large ω.

Example 2.2 Harmonic Motion of the Support

Another illustration of a system subjected to harmonic excitation is that in which the support undergoes harmonic motion. Considering Fig. 2.8, the differential equation of motion can be shown to have the form

$$m\ddot{x} + c(\dot{x} - \dot{y}) + k(x - y) = 0 \tag{a}$$

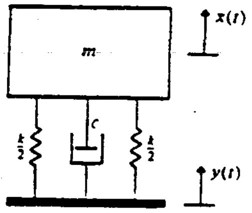


FIGURE 2.8

leading to

$$\ddot{x} + 2\zeta\omega_n\dot{x} + \omega_n^2 x = 2\zeta\omega_n\dot{y} + \omega_n^2 y \tag{b}$$

Letting the harmonic displacement of the support be given by

$$y(t) = \mathrm{Re}\,(Ae^{i\omega t}) \tag{c}$$

the response can be written as

$$x(t) = \mathrm{Re}\left[\frac{1 + i2\zeta\omega/\omega_n}{1 - (\omega/\omega_n)^2 + i2\zeta\omega/\omega_n} Ae^{i\omega t}\right] \tag{d}$$

Following a procedure similar to that used previously, the response can be written in the form

$$x(t) = X\cos(\omega t - \phi_1) \tag{e}$$

where

$$X = A\left\{\frac{1 + (2\zeta\omega/\omega_n)^2}{[1 - (\omega/\omega_n)^2]^2 + (2\zeta\omega/\omega_n)^2}\right\}^{1/2} = A\left[1 + \left(\frac{2\zeta\omega}{\omega_n}\right)^2\right]^{1/2} |H(\omega)| \tag{f}$$

and

$$\phi_1 = \tan^{-1}\frac{2\zeta(\omega/\omega_n)^3}{1 - (\omega/\omega_n)^2 + (2\zeta\omega/\omega_n)^2} \tag{g}$$

Hence, in this case the indicated nondimensional ratio is

$$\frac{X}{A} = \left[1 + \left(\frac{2\zeta\omega}{\omega_n}\right)^2\right]^{1/2} |H(\omega)| \tag{h}$$

where the ratio X/A is known as *transmissibility*. Curves X/A versus ω/ω_n with ζ as a parameter are plotted in Fig. 2.9. Moreover, curves ϕ_1 versus ω/ω_n for various values of ζ are shown in Fig. 2.10. Again for $\zeta = 0$ the response is either in phase with the excitation for $\omega/\omega_n < 1$ or completely out of phase with the excitation for $\omega/\omega_n > 1$.

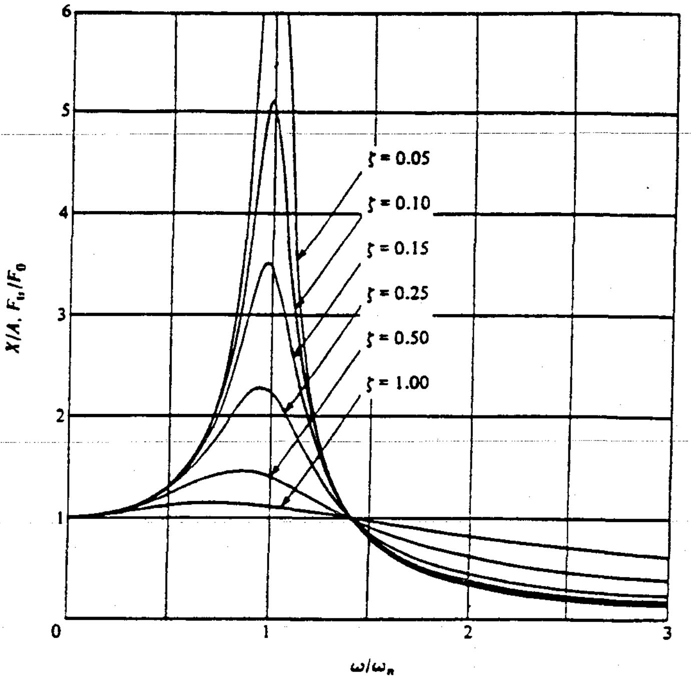


FIGURE 2.9

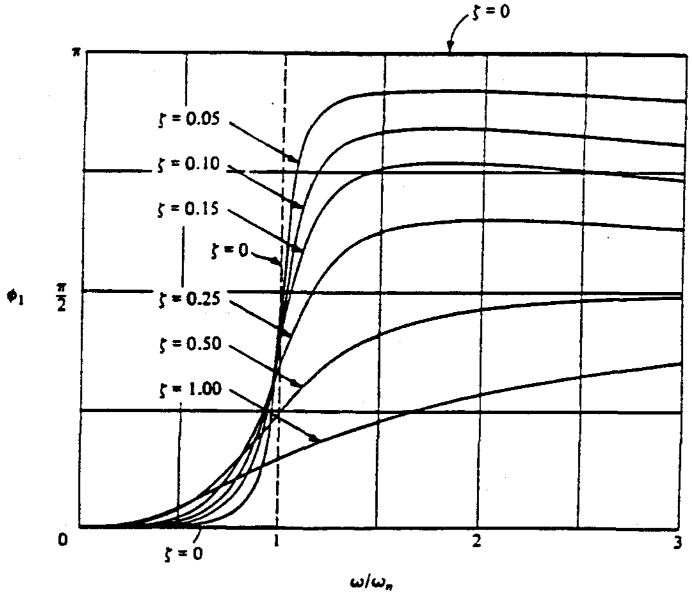


FIGURE 2.10

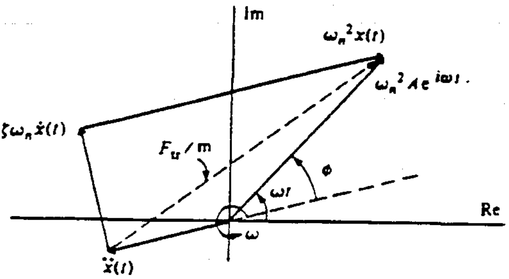


FIGURE 2.11

2.3 COMPLEX VECTOR REPRESENTATION OF HARMONIC MOTION

The representation by complex vectors of the harmonic excitation and the response of a damped system to that excitation can be given an interesting geometric interpretation by means of a diagram in the complex plane. Indeed, differentiating Eq. (2.22) with respect to time, we obtain

$$\dot{x}(t) = i\omega A|H(\omega)|e^{i(\omega t-\phi)} = i\omega x(t) \tag{2.27a}$$

$$\ddot{x}(t) = (i\omega)^2 A|H(\omega)|e^{i(\omega t-\phi)} = -\omega^2 x(t) \tag{2.27b}$$

Because i can be written as $i = \cos \pi/2 + i \sin \pi/2 = e^{i\pi/2}$, we conclude that the velocity leads the displacement by the phase angle $\pi/2$ and that it is multiplied by the factor ω. Moreover, because -1 can be expressed as $-1 = \cos \pi + i \sin \pi = e^{i\pi}$, it follows that the acceleration leads the displacement by the phase angle π and that it is multiplied by the factor ω^2.

In view of the above, we can represent Eq. (2.11) in the complex plane shown in Fig. 2.11. There is no loss of generality in regarding the amplitude A as a real number, which is the assumption implied in Fig. 2.11. The interpretation of Fig. 2.11 is that the sum of the complex vectors $\ddot{x}(t)$, $2\zeta\omega_n\dot{x}(t)$, and $\omega_n^2 x(t)$ balances $\omega_n^2 A e^{i\omega t}$, which is precisely the requirement that Eq. (2.11) be satisfied. Note that the entire diagram rotates in the complex plane with angular velocity ω. It is clear that considering only the real part of the response is the equivalent of projecting the diagram on the real axis. We can just as easily retain the projections on the imaginary axis, or any other axis, without affecting the nature of the response; retaining the real part happens to be very convenient. In view of this, it is also clear that the assumption that A is real is immaterial. Choosing A as a complex quantity, or considering projections on an axis other than the real axis, would merely imply the addition of a phase angle ψ to all the vectors in Fig. 2.11, without changing their relative position. This is equivalent to multiplying both sides of Eq. (2.11) by the constant factor $e^{i\psi}$.

2.4 VIBRATION ISOLATION

In many systems of the type shown in Fig. 1.7, we are interested in transmitting as little vibration as possible to the base. This problem can become critical when the excitation is harmonic. Clearly, the force transmitted to the base is through springs and dampers. From Fig. 2.11, we conclude that the amplitude of that force is

$$F_{tr} = m[(2\zeta\omega_n\dot{x})^2 + (\omega_n{}^2x)^2]^{1/2} \tag{2.28}$$

where the amplitude of the velocity is simply ωx. Hence, we have

$$F_{tr} = kx\left[1 + \left(\frac{2\zeta\omega}{\omega_n}\right)^2\right]^{1/2} \tag{2.29}$$

But from Eq. (2.22), if we recall that the phase angle is of no consequence, we conclude that

$$F_{tr} = Ak\left[1 + \left(\frac{2\zeta\omega}{\omega_n}\right)^2\right]^{1/2} |H(\omega)| \tag{2.30}$$

Because $Ak = F_0$ is the amplitude of the actual excitation force, the nondimensional ratio F_{tr}/F_0 is a measure of the force transmitted to the base. The ratio can be written as

$$\frac{F_{tr}}{F_0} = \left[1 + \left(\frac{2\zeta\omega}{\omega_n}\right)^2\right]^{1/2} |H(\omega)| \tag{2.31}$$

and is recognized as the *transmissibility* given by Eq. (*h*) of Example 2.2. Hence, the plots F_{tr}/F_0 versus ω/ω_n are the same as the plots X/A versus ω/ω_n shown in Fig. 2.9. It is not difficult to show that when $\omega/\omega_n = \sqrt{2}$ the full force is transmitted to the base, $F_{tr}/F_0 = 1$. For values $\omega/\omega_n > \sqrt{2}$ the force transmitted tends to decrease with increasing driving frequency ω, regardless of ζ. Interestingly, damping does not alleviate the situation, and, in fact, for $\omega/\omega_n > \sqrt{2}$, the larger the damping, the larger the transmitted force. Recalling, however, that in increasing the driving frequency we would have to go through a resonance condition for zero damping, we conclude that a small amount of damping is desirable. Moreover, the case of zero damping represents only an idealization which does not really exist, and in practice a small amount of damping is always present.

2.5 VIBRATION MEASURING INSTRUMENTS

There are basically three types of vibration measuring instruments, namely, those measuring accelerations, velocity, and displacements. We shall discuss the first and the third only. Many instruments consist of a case containing a spring-damper-

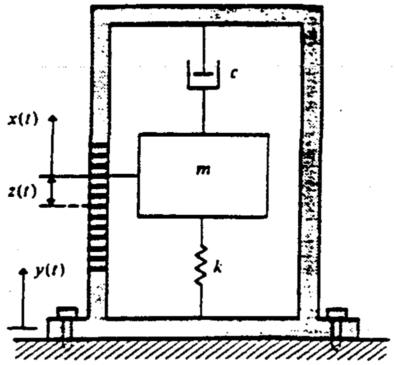


FIGURE 2.12

mass system of the type shown in Fig. 2.12, and a device measuring the displacement of the mass relative to the case. The mass is constrained to move along a given axis. The displacement of the mass relative to the case is generally measured electrically. Damping may be provided by a viscous fluid inside the case.

The displacement of the case, the displacement of the mass relative to the case, and the absolute displacement of the mass are denoted by $y(t)$, $z(t)$, and $x(t)$, respectively, so that $x(t) = y(t) + z(t)$. The relative displacement $z(t)$ is the one measured, and from it we must infer the motion $y(t)$ of the case. Although we wish ultimately to determine $y(t)$, it is the response $z(t)$ which is the variable of interest. Using Newton's second law, we can write the equation of motion

$$m\ddot{x}(t) + c[\dot{x}(t) - \dot{y}(t)] + k[x(t) - y(t)] = 0 \tag{2.32}$$

which, upon elimination of $x(t)$, can be rewritten as

$$m\ddot{z}(t) + c\dot{z}(t) + kz(t) = -m\ddot{y}(t) \tag{2.33}$$

Assuming harmonic excitation, $y(t) = Ae^{i\omega t}$, Eq. (2.33) leads to

$$m\ddot{z} + c\dot{z} + kz = Am\omega^2 e^{i\omega t} \tag{2.34}$$

which is similar in structure to Eq. (*b*) of Example 2.1. By analogy, the response is

$$z(t) = A\left(\frac{\omega}{\omega_n}\right)^2 |H(\omega)|e^{i(\omega t - \phi)} \tag{2.35}$$

where the phase angle ϕ is given by Eq. (2.21). Introducing the notation $z(t) = z_0 e^{i(\omega t - \phi)}$, we conclude that the plot z_0/A versus ω/ω_n is identical to that given in Fig. 2.7. The plot is shown again in Fig. 2.13 on a scale more suitable for our purposes here.

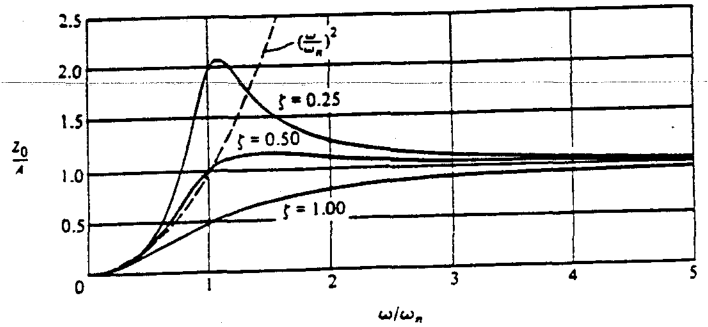


FIGURE 2.13

For small values of the ratio ω/ω_n the value of the magnification factor $|H(\omega)|$ is nearly unity and the amplitude z_0 can be approximated by

$$z_0 \cong A\left(\frac{\omega}{\omega_n}\right)^2 \tag{2.36}$$

so that z_0 is proportional to the acceleration of the case. Hence, if the frequency ω of the harmonic motion of the case is sufficiently low relative to the natural frequency of the system that the amplitude ratio z_0/A can be approximated by the parabola $(\omega/\omega_n)^2$ (see Fig. 2.13), the instrument can be used as an *accelerometer*. Because the range of ω/ω_n in which the amplitude ratio can be approximated by $(\omega/\omega_n)^2$ is the same as the range in which $|H(\omega)|$ is approximately unity, it will prove advantageous to refer to the plot $|H(\omega)|$ versus ω/ω_n instead of the plot z_0/A versus ω/ω_n. Figure 2.14 shows plots $|H(\omega)|$ versus ω/ω_n in the range $0 \leq \omega/\omega_n \leq 1$, with ζ acting as a parameter. From Fig. 2.14, we conclude that the range in which $|H(\omega)|$ is approximately unity is very small for light damping, which implies that the natural frequency of lightly damped accelerometers must be appreciably larger than the frequency of the harmonic motion to be measured. To increase the range of utility of the instrument, larger damping is necessary. It is clear from that figure that the approximation is valid for a larger range of ω/ω_n if $0.65 < \zeta < 0.70$. Indeed, for $\zeta = 0.7$ the accelerometer can be used in the range $0 \leq \omega/\omega_n \leq 0.4$ with less than 1 percent error, and the range can be extended to $\omega/\omega_n \leq 0.7$ if proper corrections, based on the instrument calibration, are made. Accelerometers with a piezoelectric crystal serving both as the spring and the sensor and with a frequency range to 5,000 cps (Hz) are commercially available.

Also from Fig. 2.13, we notice that for very large values of ω/ω_n, the ratio $z_0/A = (\omega/\omega_n)^2|H(\omega)|$ approaches unity, regardless of the amount of damping. Hence, if the object is to measure displacements, then we should make the natural frequency of the systems very low relative to the excitation frequency, in which case the instrument is called a *vibrometer*. For a *seismograph*, which is an instru-

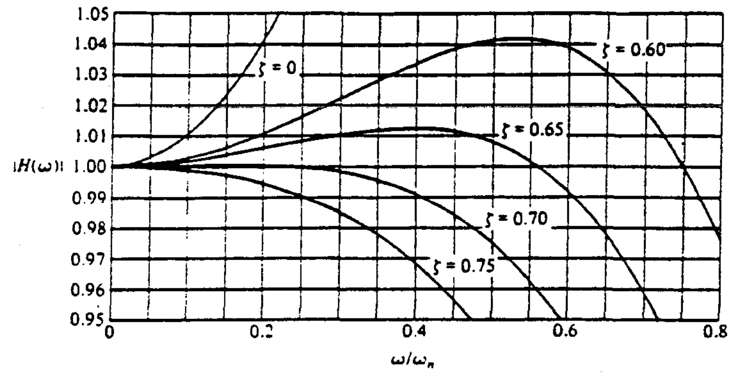


FIGURE 2.14

ment designed to measure earth displacements such as those caused by earthquakes or underground nuclear explosions, the requirement for a low natural frequency dictates that the spring be very soft and the mass relatively heavy, so that, in essence, *the mass remains nearly stationary in inertial space* while the case, being attached to the ground, moves relative to the mass. Displacement-measuring instruments are generally undamped.

Because seismographs require a much larger mass than accelerometers and the relative motion of the mass in a seismograph is nearly equal in magnitude to the motion to be measured, seismographs are considerably larger in size than accelerometers. In view of this, if the interest lies in displacements, it may prove more desirable to use an accelerometer to measure the acceleration of the case, and then integrate twice with respect to time to obtain the displacement.

2.6 ENERGY DISSIPATION. STRUCTURAL DAMPING

In Sec. 2.2 we have shown that the response of a spring-damper-mass system subjected to a harmonic excitation given by the real part of

$$F(t) = Ake^{i\omega t} \tag{2.37}$$

is given by the real part of

$$x(t) = A|H(\omega)|e^{i(\omega t-\phi)} = Xe^{i(\omega t-\phi)} \tag{2.38}$$

where

$$X = A|H(\omega)| \tag{2.39}$$

can be interpreted as the maximum displacement amplitude. Moreover, we have shown in Sec. 2.3 that there is no loss of generality by regarding A as a real number. Clearly, because of damping, the system is not conservative, and indeed energy is dissipated. Since this energy dissipation must be equal to the work done by the external force, we can write the expression for the energy dissipated per cycle of vibration in the form

$$\Delta E_{cyc} = \int_{cyc} F\,dx = \int_0^{2\pi/\omega} F\dot{x}\,dt \tag{2.40}$$

where we recall that only the real parts of F and $\dot{x}$ must be considered. Inserting Eqs. (2.27*a*) and (2.37) into (2.40), we obtain

$$\begin{aligned}\Delta E_{cyc} &= -kA^2|H(\omega)|\omega \int_0^{2\pi/\omega} \cos\omega t \sin(\omega t - \phi)\,dt \\ &= m\omega_n{}^2A^2|H(\omega)|\pi \sin\phi \end{aligned}\tag{2.41}$$

From Eqs. (2.8) and (2.15), it is not difficult to show that

$$\sin\phi = 2\zeta\frac{\omega}{\omega_n}|H(\omega)| = \frac{c\omega}{m\omega_n{}^2}|H(\omega)| \tag{2.42}$$

where it is recalled that $\zeta = c/2m\omega_n$. Inserting Eqs. (2.39) and (2.42) into (2.41), we obtain the simple expression

$$\Delta E_{cyc} = c\pi\omega X^2 \tag{2.43}$$

from which it follows that the energy dissipated per cycle is directly proportional to the damping coefficient c, the driving frequency ω, and the square of the response amplitude.

Experience shows that energy is dissipated in all real systems, including those systems for which the mathematical model makes no specific provision for damping, because energy is dissipated in real springs as a result of internal friction. In contrast to viscous damping, damping due to internal friction does not depend on velocity. Experiments performed on a large variety of materials show that energy loss per cycle due to internal friction is roughly proportional to the square of the displacement amplitude,*

$$\Delta E_{cyc} = \alpha X^2 \tag{2.44}$$

where α is a constant independent of the frequency of the harmonic oscillation. This type of damping, called *structural damping*, is attributed to the *hysteresis phenomenon* associated with cyclic stress in elastic materials. The energy loss per cycle of stress is equal to the area inside the hysteresis loop shown in Fig. 2.15. Hence, comparing Eqs. (2.43) and (2.44), we conclude that systems possessing

* See L. Meirovitch, "Analytical Methods in Vibrations," p. 402, The Macmillan Co., New York, 1967.

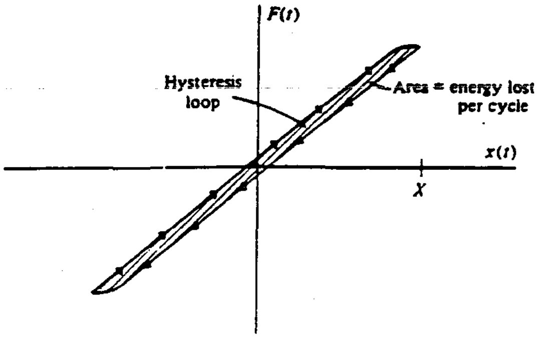


FIGURE 2.15

structural damping and subjected to harmonic excitation may be treated as if they were subjected to an equivalent viscous damping

$$c_{eq} = \frac{\alpha}{\pi\omega} \tag{2.45}$$

This enables us to write Eq. (1.8) in the form

$$m\ddot{x}(t) + \frac{\alpha}{\pi\omega}\dot{x}(t) + kx(t) = Ake^{i\omega t} \tag{2.46}$$

where consideration has been given to Eqs. (2.37) and (2.45). Because $\dot{x} = i\omega x$, we can rewrite Eq. (2.46) in the form

$$m\ddot{x}(t) + k(1 + i\gamma)x(t) = Ake^{i\omega t} \tag{2.47}$$

where

$$\gamma = \frac{\alpha}{\pi k} \tag{2.48}$$

is called the *structural damping factor*. The quantity $k(1 + i\gamma)$ is called *complex stiffness*, or occasionally also *complex damping*.

The steady-state solution of Eq. (2.47) is the real part of

$$x(t) = \frac{Ae^{i\omega t}}{1 - (\omega/\omega_n)^2 + i\gamma} \tag{2.49}$$

and, in contrast to viscous damping, for structural damping the maximum amplitude is obtained exactly for $\omega = \omega_n$.

One word of caution is in order. *The analogy between structural and viscous damping is valid only for harmonic excitation*, because implied in the foregoing development is the assumption of system harmonic response at the driving frequency ω.

THREE-DIMENSIONAL SEISMIC SUPER-ATTENUATOR FOR LOW FREQUENCY GRAVITATIONAL WAVE DETECTION

R. DEL FABBRO, A. DI VIRGILIO, A. GIAZOTTO, H. KAUTZKY [1], V. MONTELATICI [2] and D. PASSUELLO

INFN Sezione di Pisa and Dipartimento di Fisica, Università di Pisa, Pisa, Italy

Received 26 June 1987; accepted for publication 27 July 1987

Communicated by J.P. Vigier

We present the study of a passive n-fold pendulum to be used as a three-dimensional seismic noise attenuator. A 7-fold pendulum, under construction at the INFN laboratory in Pisa, is expected to provide a horizontal and vertical attenuation factor of 10^{-11} and 10^{-9} respectively at 10 Hz and is capable to sustain a 400 kg test mass used in a large base interferometric gravitational wave antenna.

1. The seismic noise reduction is a crucial problem in the task of setting up a gravitational wave antenna sensitive in the low frequency range, where one hopes to detect waves emitted by rotating massive stellar objects like pulsars as well as signals due to coalescing neutron stars and collapsing bodies.

In this paper we present the results we have obtained in calculating the attenuation functions of a passive n-fold pendulum. Such a pendulum is capable of seismic noise reduction both in horizontal and vertical direction. These seismic attenuators are designed in such a way as to be capable to sustain heavy test masses for a large base interferometric gravitational wave antenna and to reach a horizontal and vertical attenuation factor of the order of 10^{-11} and 10^{-9} at 10 Hz respectively.

One-dimensional active seismic attenuation systems have been realized [1–3] reaching a horizontal attenuation value of about 10^{-6} at 10 Hz. These systems are rather difficult to operate at a high amplification level, due to the presence of feedback instabilities, and furthermore it is rather inconceivable to design a multi-stage three-dimensional (3D) system capable of sustaining such heavy test masses. For these reasons we were led to study a passive three-dimensional attenuator. In particular the attenuation in the vertical direction is exploited by gas springs, which have the advantage of not having hysteresis and low frequency normal modes as mechanical springs have. Furthermore a gas spring can lift very heavy loads and still have a very low stiffness.

2. First we present a study devoted to solving the problem of obtaining a relevant suppression of the horizontal component of the seismic noise in a frequency range as low as possible. Let us consider a system composed by a cascade of n masses connected by wires and suspended at a fixed frame. We call this system an n-fold passive pendulum.

In the small angle approximation the equations of the horizontal motion can be decoupled, therefore we limit ourselves to study the system in one dimension. In our scheme we suppose the wires to be unstretchable and the pendulum masses concentrated in their own CMS. In the study of the n-fold pendulum we have assumed the following conditions:

(a) the total length L of the pendulum is a fixed input parameter;

(b) the pendulum masses have an equal value of m, with the exception of the test mass, which has mass fm, where f is an input parameter;

(c) the distance between the contiguous masses is L/n;

[1] Fermilab, Batavia, USA.

[2] Laboratorio ENEA, Frascati, Italy.

(d) the relaxation time coefficient τ is equal for all pendulum stages, τ is a fixed input parameter.

For each pendulum mass one can write the equation of the motion, therefore for an n-fold pendulum one obtains

$$[Z/\Omega_0^2+2(n+f-k)-3]X_{k+1}=(n+f-k-1)X_k$$

$$+(n+f-k-2)X_{k+2} \quad (k=0, n-2),$$

$$(Z/\Omega_0^2+1)X_n=X_{n-1}, \tag{1}$$

where the last equation cannot be represented in the sequential form, since it refers to the last mass, and the other $n-1$ are labelled by the index k, where the value $k=0$ refers to the suspension point. The quantity X_k represents the horizontal displacement of the kth mass. The complex quantity Z and the real one Ω_0^2 are defined as follows:

$$Z=-\Omega^2+\mathrm{i}\Omega/\tau, \quad \Omega_0^2=gn/L,$$

where Ω is the circular frequency and g is the gravity acceleration constant. If we define a set of n quantities A_k:

$$A_{k+1}=A_0+2(n+f-k)-3 \quad (k=0, n-2),$$

$$A_n=A_0+1, \tag{2}$$

where A_0 is defined as Z/Ω_0^2, then the system of equations (1) can be written in the form

$$A_{k+1}X_{k+1}=(n+f-k-1)X_k$$

$$+(n+f-k-2)X_{k+2} \quad (k=0, n-2),$$

$$A_nX_n=X_{n-1}. \tag{3}$$

The system of equations (3) can be put in the following way:

$$X_k=B_{k+1}X_n \quad (k=0, n-2),$$

$$X_{n-1}=B_nX_n, \tag{4}$$

where the $n+1$ quantities B_k are defined as follows:

$$B_{n+1}=1, \quad B_n=A_n=A_0+1,$$

$$B_{k+1}=\frac{A_{k+1}B_{k+2}-(n+f-k-2)B_{k+3}}{(n+f-k-1)}$$

$$(k=0, n-2). \tag{5}$$

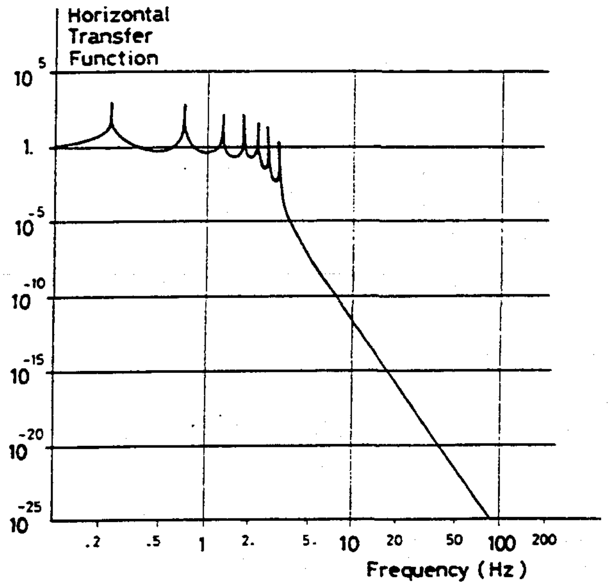


Fig. 1. Horizontal attenuation function ($n=7$) versus frequency.

From the set (4) we can take the equation for $k=0$:

$$X_n/X_0=1/B_1, \tag{6}$$

which gives the ratio of the amplitudes of the last mass over the amplitude of the pendulum suspension point. Hence the complex function $1/B_1$ expresses the horizontal transfer function of the n-fold pendulum. The study of the absolute value of the transfer function (6) shows three well separated regions in the frequency parameter:

(a) a flat region at the lowest frequencies, where the attenuation is 1;

(b) a region where n resonance peaks rise over an almost flat base, the peak widths depend on the parameter τ;

(c) a region where the attenuation function decreases with the law of ν^{-2n}.

The third region is suitable in suppressing seismic noise. Pushing this region toward the low frequencies implies keeping the pendulum resonances as low as possible. In fig. 1 a typical horizontal transfer function in the case of $n=7$ and $f=4$ is shown. The flat region ranges from 0 to about 0.1 Hz; from 0.1 Hz to about 4 Hz there is the peaks region and above 4 Hz the attenuation function begins to decrease. The last region from 4 Hz to infinity shows an exponential falling with a power -14. The noise suppression

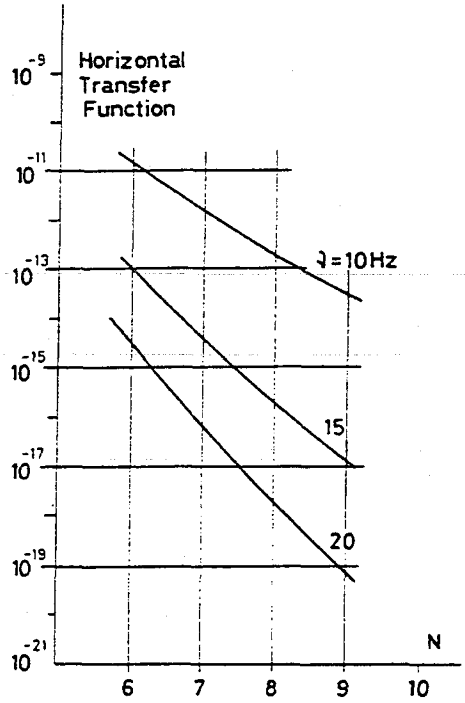


Fig. 2. Horizontal attenuation versus *n*.

factor is rather impressive: at 10 Hz the transfer function reaches the value of 1.45×10^{-12}. In fig. 2 is shown the attenuation function of a multiple pendulum versus the number of stages ($n=6$, 7, 8, 9) at three fixed frequencies: 10, 15 and 20 Hz.

3. Let us present now the study of an *n*-fold pendulum having a gas spring at each stage in the vertical direction only. It is easy to verify that a vertical cascade of *n* springs and *n* masses is capable to suppress the vertical component of the seismic noise. The choice of gas springs instead of mechanical ones is due to some merits of the former:

(a) the capability of bearing heavy masses with a rather low stiffness;

(b) the lack of hysteresis effects;

(c) the ease of operation in adjusting the spring length.

This last point is rather important if one has to arrange a long chain of gas springs. The performance of a gas spring is described elsewhere [4], in this section we discuss the attenuation factor of a vertical *n*-fold harmonic oscillator. The stiffness K_i of each spring is given by (see ref. [3])

$$K_i=\gamma(P_0+Mg/S_0)S_0^2/V+K_m, \tag{7}$$

where V is the cylinder volume, S_0 the piston area, P_0 the external pressure, γ the adiabatic constant and Mg the total weight supported by the jth stage and K_m the total stiffness of mechanical origin (bellows etc.).

The *n* equations of the *n*-fold harmonic oscillators are

$$ZY_j=\Omega_j^2(Y_{j-1}-Y_j)+\Omega_{j-1}^2(Y_{j+1}-Y_j)$$

$$(j=1,n-1),$$

$$ZY_n=\Omega_n^2(Y_{n-1}-Y_n)m/M_t, \tag{8}$$

where $\Omega_i^2=K_i/m$ and $Z=-\Omega^2+i\Omega/\tau$ and Y_j is the vertical displacement of the jth mass. Ω is the circular frequency and τ the relaxation time, supposed to be equal for all springs. M_t is the lowest mass i.e. the test mass of the antenna. In expression (8) the vertical displacement Y_{j-1} for $j=1$ represents the displacement of the suspension point. Let us define the $n+1$ quantities A_j,

$$A_n=1,\quad A_{n-1}=(M_t/m)(Z/\Omega_n^2)+1,$$

$$A_{n-1-j}=[(Z+\Omega_{n-j}^2+\Omega_{n+1-j}^2)A_{n-j}$$

$$-\Omega_{n+1-j}^2A_{n+1-j}]/\Omega_{n-j}^2\quad (j=1,n-1). \tag{9}$$

Substituting eqs. (9) into eqs. (8) we obtain

$$Y_{j-1}=A_{j-1}Y_n\quad (j=1,n). \tag{10}$$

From the first equation of the system (10) ($j=1$), we get the relation between the displacement of the suspension point and the test mass:

$$Y_n/Y_0=1/A_0. \tag{11}$$

The complex function $1/A_0$ is called the vertical transfer function and its absolute value is plotted in fig. 3 versus the frequency for $n=7$. The stiffness values K_i have been deduced by fitting the experimental data (see fig. 6 of ref. [3]) and extrapolating to higher masses. The vertical attenuation function looks very similar to the horizontal one, and in complete analogy three different regions can be distinguished with the same characteristics. The vertical attenuation function is 2.38×10^{-9} at 10 Hz. In fig. 4 the attenuation function is shown at 10, 15 and 20

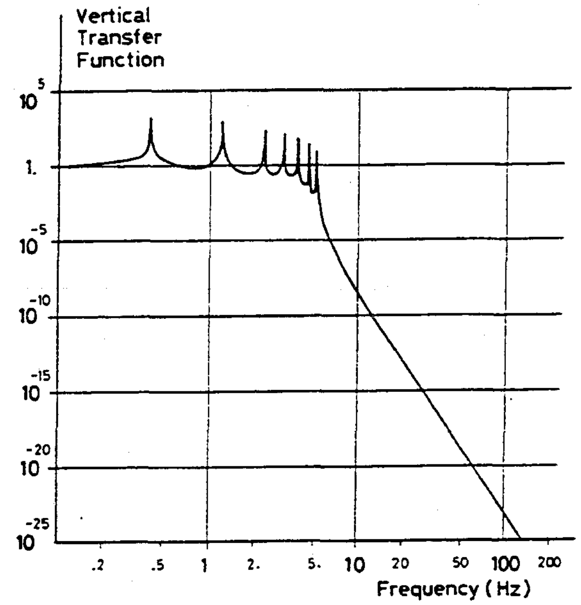


Fig. 3. Vertical attenuation function ($n=7$) versus frequency.

Hz as a function of n ($n=6$, 7, 8, 9). A criterium to equalize the horizontal and vertical attenuation, leads to the choice of $n=7$, assuming the vertical to horizontal coupling of the motion to be of the order of 10^{-2}.

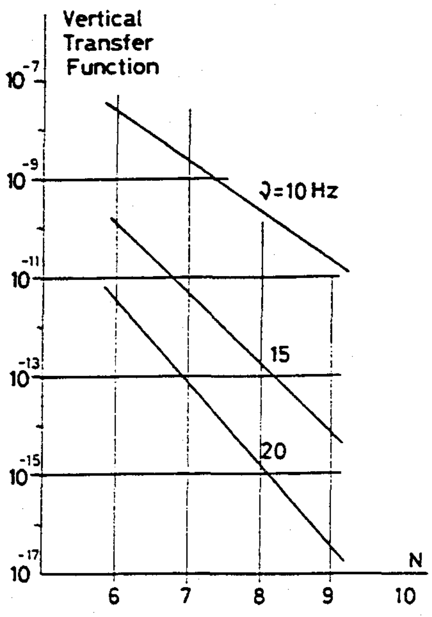


Fig. 4. Vertical attenuation versus n.

4. Let us evaluate the oscillator resonances coming from the described attenuator. Both the horizontal and vertical transfer functions (eqs. (6) and (11)), can be represented as a complex polynomial expression of the order $2n$ in the following way:

$$\mathrm{TF}=1/F,$$

$$F=\sum a_j(\mathrm{i}\Omega)^j \quad (j=0,2n). \tag{12}$$

Since the a_j are real it follows that the roots of F are conjugate. If we indicate with $\mathrm{i}\Omega=\mathrm{i}\omega_j-1/2\tau_j$ the roots of F, the TF can be written

$$\mathrm{TF}=\prod\frac{\omega_j^2}{\omega_j^2-\Omega^2+\mathrm{i}\Omega/\tau_j} \quad (j=1,n). \tag{13}$$

The TF absolute value which results is:

$$|\mathrm{TF}|=\prod\frac{\omega_j^2}{[(\omega_j^2-\Omega^2)^2+(\Omega/\tau_j)^2]^{1/2}} \quad (j=1,n). \tag{14}$$

From expression (14) it appeares that the TF behaviour in the different frequency regions is:

$$\mathrm{Re}\,\Omega\ll\omega_j, \quad \mathrm{Re}\,\Omega\sim\omega_j, \quad \mathrm{Re}\,\Omega\gg\omega_j,$$

as discussed above and grafically shown in figs. 1 and 3.

The resonance frequency $\nu_j=\omega_j/2\pi$ has been calculated both in the horizontal and vertical case for $L=5$ m, $\tau_j=10^3$ s and K_j taken by the experimental data of ref. [3] and extrapolated up to $M=10^3$ kg. The values obtained are listed in table 1.

Table 1

j	ν_j (Hz)	
	horizontal	vertical
1	0.24401	0.39565
2	0.75831	1.22524
3	1.35458	2.26042
4	1.90512	3.12087
5	2.37280	3.84846
6	2.78994	4.61551
7	3.25575	5.17409

5. The 3D 7-fold pendulum discussed in this paper seems to meet the attenuation factors as required by a low frequency interferometric gravitational wave antenna [5] working down to 10 Hz. The construction of such a seismic attenuator is under way at the INFN Laboratory in Pisa.

The authors are grateful to Professor A. Stefanini for valuable suggestions and useful discussions.

References

[1] A. Giazotto, D. Passuello and A. Stefanini, Rev. Sci. Instrum. 57 (1986) 1145.

[2] N.A. Robertson, R.W.P. Drever, I. Kerr and J. Hough, J. Phys. E 15 (1982) 1101.

[3] P.R. Saulson, Rev. Sci. Instrum. 55 (1984) 1315.

[4] R. Del Fabbro, A. Di Virgilio, A. Giazotto, H. Kautzky, V. Montelatici and D. Passuello, Performance of a gas spring harmonic oscillator, to be published.

[5] A. Giazotto et al., Proposal for "Antenna interferometrica a grande base per la ricerca di onde gravitazionali", May 1987, INFN PI/AE 87/1.

REF.KK

Vibration isolation stacks for gravitational wave detectors—Finite element analysis

C. A. Cantley, J. Hough, and N. A. Robertson
Department of Physics & Astronomy, University of Glasgow, Glasgow G12 8QQ, Scotland

R. J. S. Greenhalgh
Rutherford Appleton Laboratory, Chilton, Didcot, Oxon OX11 0QX, England

(Received 11 October 1991; accepted for publication 2 January 1992)

Seismic isolation is a necessity for many delicate experiments such as those designed to search for gravitational radiation. One method of providing a significant amount of isolation has been the use of multiple stage stacks of alternating layers of dense material (metal) and elastic material (rubber). In this work finite element analysis is used to model stack systems in a relatively realistic way, allowing the importance of cross coupling of degrees of freedom to be evaluated. It becomes clear that care has to be taken with the geometrical construction in order to achieve the isolation predicted by simple dynamical analysis.

I. INTRODUCTION

Gravitational wave detectors currently being developed rely on sensing the extremely small motions which are expected to be produced by a gravitational wave passing through a mechanical system. The system may be a bar of low loss metal cooled to liquid-helium temperatures[1] or a series of test masses which are hung as pendulums at the ends of two perpendicular arms of a laser interferometer.[2] In both of these cases it is essential to isolate the mechanical system from the surroundings to reduce spurious noise introduced by local seismic or mechanical disturbances. In general the detector systems are arranged so that the motions to be monitored lie in the horizontal plane. Thus it is particularly important that the vibration isolation system is designed such that it can provide a high degree of horizontal isolation. However, since geometrical effects can cross-couple motion in one direction into motion in another it is sensible to aim for isolation factors in the remaining degrees of freedom (vertical, tilt, and rotation) which are suitably large. In most designs of gravitational wave detector, pendulum suspensions are used to provide a significant degree of isolation but in general further isolation is required. This may be provided by placing vibration isolation stacks, consisting of alternating layers of metal and elastic material, between the pendulum suspension points and the ground.

Using simple dynamical theory the transmissibility of unidirectional motion at the base of a vibration isolation stack to the corresponding displacement at the center of mass of the top plate of the stack can be readily estimated as a function of frequency. In such an analysis the stack is treated as a system of point masses connected with springs and dampers. In practice however, cross coupling is liable to be an important factor in determining the effectiveness of the isolation stack. It is not straightforward to develop an analytical model which includes the effects of distributed masses—and it is precisely these effects which can lead to cross coupling. One of our principal objectives in this work has been to quantify the degree of cross coupling occurring within a particular design of isolation stack. For this purpose a finite element model representing an aluminum plate supported by four rubber pieces was generated to represent one stage of a stack and the finite element program MSC/NASTRAN,[3,4] was used to carry out various analyses. The model dimensions are shown in Fig. 1 and the resulting resonant frequencies are typical of what could be applicable to a prototype laser interferometric gravitational wave detector. Note that to achieve the best overall isolation with such a system the ratio of the rubber stiffness to the mass of the supported plate is chosen to give the system a reasonably low resonant frequency in each dimension.

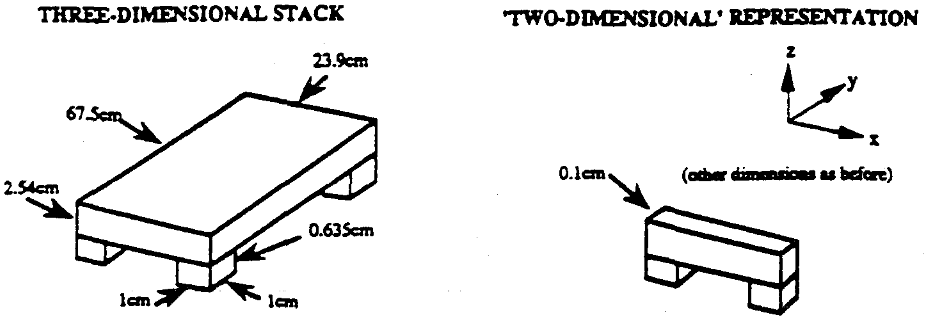
Investigations were conducted on a "single stage" stack model and also on a "double stage" which consisted of two single stages in series. The effects of a stiffness imbalance of the rubber at either end of a given stage of stack were also investigated. The results of these tests were generalized to allow an understanding of the performance of a stack containing four stages. Up to five stages have been considered in the planning of the seismic isolation systems for long baseline detectors.[2]

The work discussed in this article is most relevant to gravitational wave detectors which use laser interferometry to sense relative motions of separated masses suspended as pendulums, but is also of relevance in any field where stacks may be used as part or all of an isolation system.

II. PERFORMANCE REQUIRED FROM VIBRATION ISOLATION STACKS

The performance required from a stack system depends on the particular application, but we assume that the performance needed is such that an interferometric gravitational wave detector of arm length 3 km may approach a sensitivity for measuring strains in space of $\sim 10^{-24}/\sqrt{\mathrm{Hz}}$ at 100 Hz.[2] We also assume that the test masses are suspended in the simplest way, by wires as simple pendulums with horizontal and vertical resonant frequencies of 1 and 20 Hz, respectively. In this case, at 100 Hz, move-

0034-6748/92/042210-10$02.00 © 1992 American Institute of Physics


THREE-DIMENSIONAL STACK

'TWO-DIMENSIONAL' REPRESENTATION

FIG. 1. Dimensions of the two-dimensional finite element model compared to the three-dimensional stack.

ments at the top of the stack of less than 4×10^{-18} m/$\sqrt{\text{Hz}}$ in the vertical direction and 1.5×10^{-17} m/$\sqrt{\text{Hz}}$ in the horizontal direction are required. A figure of 1% coupling of vertical into horizontal movement in the pendulum suspension has been used to calculate the vertical limit as such a figure has been suggested by the experimental measurements of Del Fabbro *et al.*[5]

Obviously the isolation required from the stack depends on the magnitude of ground movement in the appropriate sense. A figure of $\sim10^{-11}$ m/$\sqrt{\text{Hz}}$ at 100 Hz, with a spectrum falling at 12 dB per octave, was assumed in the horizontal and vertical directions for the work in this paper;[2] and ground tilt of $\sim2\times10^{-11}$ rad/$\sqrt{\text{Hz}}$ at 100 Hz, as recently measured at a prototype detector site, with a spectrum falling somewhere between 12 dB per octave and 18 dB per octave, was used.[6]

III. GENERATION OF THE FINITE ELEMENT MODEL

It was decided to represent the three-dimensional system with what was essentially a two-dimensional cross-sectional slice through it in the x-z plane (see Fig. 1). This was done so that the complexity involved in the generation of the model and in the interpretation of the results obtained was kept to a minimum. Note that in modeling the stack in this way the investigations involving rotational motion were restricted to those where the axes of rotation were parallel to the y axis (tilting motions).

The structure was modeled by generating a mesh of nodes and connecting them with plate elements having the desired physical properties (see Fig. 2). Note that some of the physical properties of the materials used in the model had to be scaled so that the two-dimensional finite element model exhibited behavior appropriate to its three-dimensional counterpart.

Synthetic rubber is extensively used in isolation stacks. The horizontal and vertical resonant frequencies (f_h and f_v) of the plate (mass $m = 11.1$ Kg) on four pieces of Neoprene rubber of the size shown in Fig. 1 were determined experimentally. The value of f_h was used to calculate the horizontal stiffness (k_h) of the loaded rubber using the equation

$$k_h = m(2\pi f_h)^2. \tag{1}$$

Since the model used here was a thin slice through the real system, a fictitious density had to be assigned to the modeled aluminum plate to give the correct loading conditions per unit area of rubber support. The shear modulus (G) and the Young's modulus (E) of the aluminum were also scaled up by the same factor as the density in order to maintain the stiffness properties of the plate. The damping of the internal modes of the plate was modeled using a "structural element damping coefficient" facility in the NASTRAN package.[4] For initial models the damping was chosen so that the Q's of the internal modes of the aluminum plate were ~20, while for later ones the modes were critically damped ($Q\sim0.5$).

Further, since one slice-like rubber support in the model represented two block-like supports in the real system, the moduli of the rubber were scaled to give the stack realistic stiffness properties. The shear modulus of the rubber was determined using the equation

$$G = k_h h/A \tag{2}$$

where h is the height of the rubber supports, A is the total top surface area of the rubber, and then scaled to compensate for the reduction in area A.

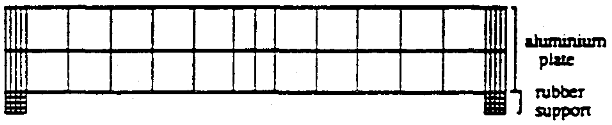


FIG. 2. Finite element model of the one-stage vibration isolation stack. The system is divided into a number of elements as shown.

There is approximately no volume change when rubber undergoes tension or compression,[7] hence Poisson's ratio $\upsilon\sim0.5$. The value of Young's modulus for the rubber was calculated by the finite element program according to the relationship:

$$E = 2G(1+\upsilon) \tag{3}$$

giving

$$E\sim3G. \tag{4}$$

A structural element damping coefficient was assigned to the rubber so that the internal modes of the rubber supports were damped to give an internal Q of ~5, this being a typical observed value.[8] Viscous damping of the fundamental stack resonances was incorporated by connecting the nodes at the bases and tops of the rubber supports with orthogonally connected damping elements in

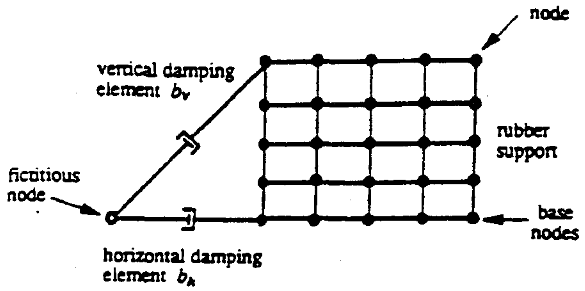

node
vertical damping element b_v
rubber support
fictitious node
base nodes
horizontal damping element b_h

FIG. 3. Damping elements are connected across each rubber support to provide viscous damping of the supported plate to the ground in both the horizontal and vertical directions. To show the damping elements clearly, they are not drawn at right angles to each other. Each pair of horizontal and vertical elements are connected orthogonally using a massless "fictitious node." Using "multipoint constraints" (Ref. 3) this node can be instructed to follow the vertical motion of the horizontal element and the horizontal motion of the vertical element. In this way the levels of damping in the two directions can be independently defined. Note that each pair of nodes along the base and top of a given support are connected in this way.

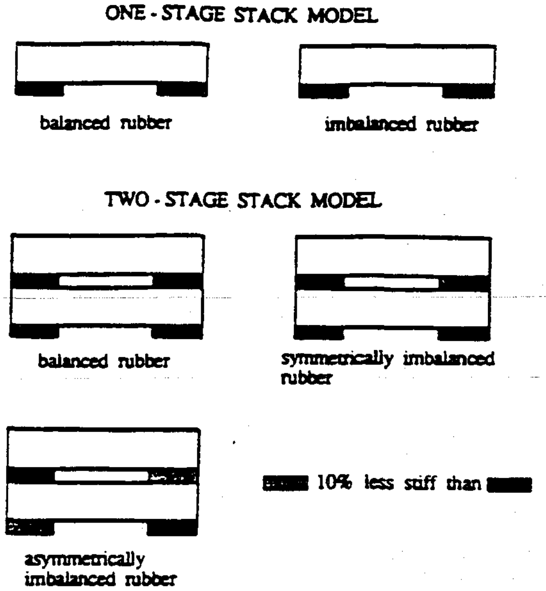


FIG. 4. The various stack systems investigated.

:he horizontal and vertical directions (see Fig. 3). The size ɔf the horizontal damping factor b_h was established using :he equation

$$b_h = 2\pi m f_h/Q_h, \tag{5}$$

where Q_h, the corresponding quality factor, was chosen to be ~5 as suggested by some elementary experimental tests. These experiments suggested that the Q of the vertical mode of each stage would be somewhat higher than of the horizontal mode. Thus the size of the damping factor for vertical motion, b_v, was calculated to give $Q_v \sim 15$. Note hat the vertical stiffness and viscous damping of the rubber supports also determines the tilting resonant frequency and corresponding quality factor. Further details of this inite element model are given in Ref. 6.

Despite the approximate nature of this model it incorporates the essential features of each stage of the stack and hould lead to a reasonable understanding of the performance and an indication of the cross-coupling mechanisms which take place in multistage vibration isolation stacks.

V. METHODS OF ANALYSIS

Eigenvalue analyses were carried out on the single and wo-stage models illustrated in Fig. 4. The horizontal to orizontal and vertical to vertical frequency response analses for each model were carried out by driving the base odes of the rubber supports (see Fig. 3) sinusoidally in ıe relevant direction with unit amplitude displacement at ll frequencies in the range considered. In each case the :sulting amplitudes of motion of the centers of mass of the ıpported plates in the appropriate directions were observed to give a value for the transmissibility at each freuency. To investigate the isolation for tilting motion the ase nodes were driven with a unit vertical input at the end odes of the base of the rubber supports and with progresvely decreasing displacement towards the center of the late as illustrated in Fig. 5. The amplitudes of tilting motion of the plates about their centers of mass were evaluated in each case (at frequencies below the internal resonances of the plates) by noting the relative vertical displacements of the end and center nodes of the plates.

Investigations into the cross coupling of tilting ground motion to horizontal and vertical motion of the centers of mass of the plates were carried out using the same progressive driving mechanism described above. The amplitudes of the horizontal or vertical motion of the centers of mass of the plates were compared to the angle of tilt at the base of the stack to give a value for the transmissibility in units of m/rad.

In the analysis of the remaining cross-coupling effects, unit vertical or horizontal input at the driven base nodes was used and the orthogonal directional components at the center of mass nodes of the plates were noted. The amplitude of tilting motion of the plates about their centers of mass were also evaluated in each case to give the transmissibility in units of rad/m.

V. EIGENVALUE ANALYSES

The one-stage stack model was constrained to move in the x-z plane. Therefore only two of the three possible translational degrees of freedom remained. Similarly only

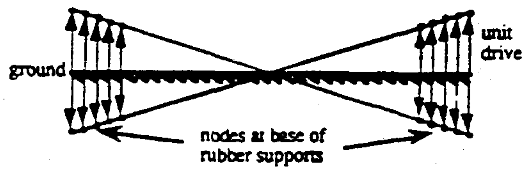


FIG. 5. Progressive driving mechanism for investigations of couplings of tilting motions.

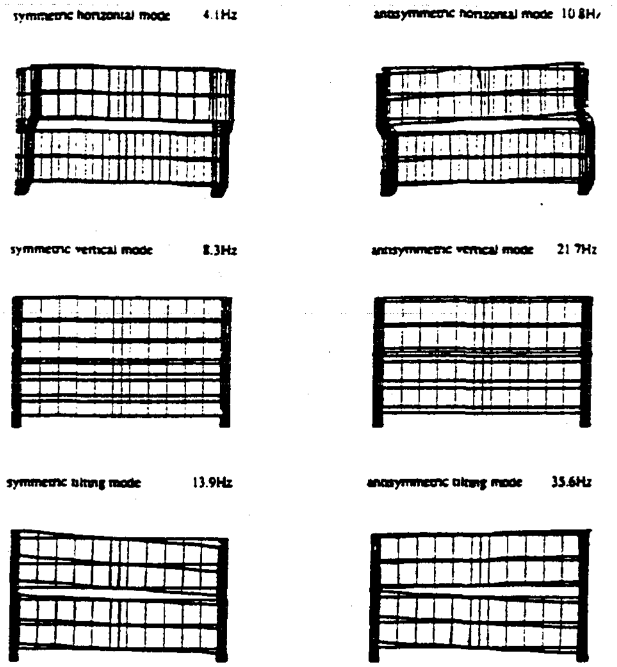

symmetric horizontal mode 4.1Hz
antisymmetric horizontal mode 10.8Hz
symmetric vertical mode 8.3Hz
antisymmetric vertical mode 21.7Hz
symmetric tilting mode 13.9Hz
antisymmetric tilting mode 35.6Hz

FIG. 6. Modes present in the two-stage stack with balanced rubber properties. The modes occur in pairs corresponding to the oscillations in the two stages being in or out of phase with each other. Note that the horizontal and tilting modes are not pure but are in fact weakly coupled together, as can be seen in the figure. The step-like nature of the displacements is an artifact of the plotting method used.

one rotational degree of freedom remained unconstrained, this being tilting motion about the y axis (refer to Fig. 1). As a consequence of these constraints the modeled stack had only three fundamental modes of oscillation, namely, the horizontal, vertical, and tilting modes. With rubber of identical stiffness properties on each side of the stack the fundamental frequencies were at $f_h \sim 7$ Hz, $f_v \sim 13$ Hz, and $f_t \sim 22$ Hz. (These values agreed reasonably well with those predicted using simple theory.) The lowest internal mode of the metal plate was at ~ 2 kHz.

The first six normal modes associated with the two-stage stack having balanced rubber properties are shown in Fig. 6. The dotted lines represent the undisturbed position of the mesh for reference. These diagrams were generated using the graphics package FEMVIEW.[9]

With a 10% reduction in the stiffness of the rubber on one side of the stack, the computed fundamental stack frequencies were slightly lower in value, as one would expect, and the mode shapes were asymmetrical about the y-z plane.

A complete summary of the fundamental frequencies obtained for the various stack models analyzed is given in Fig. 7.

VI. TRANSMISSIBILITY OF ONE- AND TWO-STAGE STACKS

The transmissibility of the one-stage stack was analyzed for a number of possible couplings.

(a) *Vertical drive to vertical response; Horizontal drive to horizontal response; Tilt drive to tilt response.* These are due to direct coupling of the displacements through the mass/spring (plate/rubber) arrangement.

one-stage stack			
MODEL	f_h	f_v	f_t
balanced	6.7	13.4	22.1
imbalanced	6.6	13.1	21.5

two-stage stack						
MODEL	f_h^s	f_h^a	f_v^s	f_v^a	f_t^s	f_t^a
balanced	4.1	10.8	8.3	21.7	13.9	35.6
symmetrically imbalanced	4.0	10.5	8.1	21.1	13.6	34.7
asymmetrically imbalanced	4.0	10.5	8.1	21.2	13.5	34.7

FIG. 7. Summary of the fundamental resonant frequencies for the various stack systems investigated. All frequencies are given in hertz. Superscripts denote symmetric (s) or antisymmetric (a) mode. Subscripts denote horizontal (h), vertical (v), or tilt (t). Imbalanced figures are for a 10% reduction in the rubber stiffness on one side of each stage.

(b) *Horizontal drive to tilt response; Tilt drive to horizontal response.* The first type of coupling is due to the effective torques on the supported plate (Fig. 8). One of the torques results from the line of action of the horizontal force exerted by the top of the rubber on the base of the metal plate being offset vertically from the center of mass of the plate. The other torque, acting across the top of each rubber support results from the distortion of the rubber.

The second type of coupling is due to the geometrical effect of the vertical offset of the center of mass of the plate from the axis of tilt (Fig. 9).

(c) *Tilt drive to vertical response; Vertical drive to tilt response; Vertical drive to horizontal response; Horizontal drive to vertical response.* For the linear analysis used in this work these occur only when the rubber stiffness is imbalanced and effectively arise due to torques exerted on the supported plates by the rubber pads. There are of course tilt to vertical and horizontal to vertical cross couplings when the rubber stiffness is balanced; these are second-order effects producing responses at frequencies other than the driving frequency, and since seismic noise has a relatively small amplitude these couplings are not generally

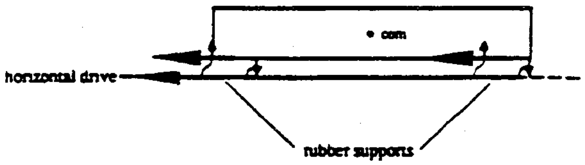


FIG. 8. Mechanism for the cross coupling of horizontal to tilting motion in a one-stage stack. The plate is subject to torques which force it to rotate about its center of mass (com). One of the torques results from the shear force of the rubber acting on the metal plate being vertically offset from the center of mass of the plate. The other torque, acting across the top of each rubber support, results from the distortion of the rubber.

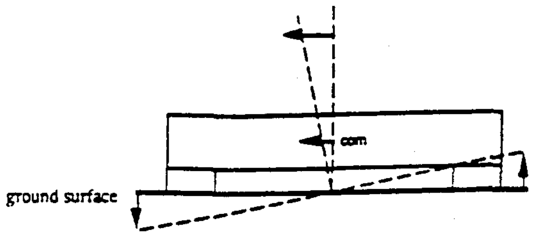


FIG. 9. Mechanism for the cross coupling of tilt to horizontal motion in a one-stage stack. At low frequencies the system is essentially rigid and the center of mass (com) of the plate is driven linearly in the horizontal direction. The larger the vertical offset the larger this effect.

significant. The results of these various analyses are shown in Figs. 10–17. In Fig. 10, the transmissibility curve for vertical drive to vertical response, the aluminum plate was modeled to have a Q of 20 and the effect of the lowest plate resonance can easily be seen. It should be noted that for the three-dimensional stack shown in Fig. 1 the first plate resonance will be at a lower frequency than that of the two-dimensional model presented here. As such plate resonances compromise the isolation significantly and are liable to appear at frequencies of interest for gravitational wave detection it is important that they are well damped. In the remaining two-dimensional analyses, in order to simplify the situation it was assumed that the plates were critically damped ($Q=0.5$). The effect of this damping is also shown in Fig. 10.

The form of each transmissibility curve in Figs. 10–17 can be understood relatively easily from consideration of the dynamics of the system. In situations where a single resonance of the mass/rubber stage is dominant the transmissibility falls at approximately 12 dB per octave above the resonance at f_0 up to a frequency Qf_0, where Q is the quality factor of the resonance, and at 6 dB per octave above this frequency. Where there are two resonances involved and where these resonances are effectively in series, e.g., in going from tilt drive to horizontal response (Fig. 13), the transmissibility falls at twice this rate. In some cases, at high frequency, the transmissibility curve flattens out as a result of computational rounding effects discussed more fully below.

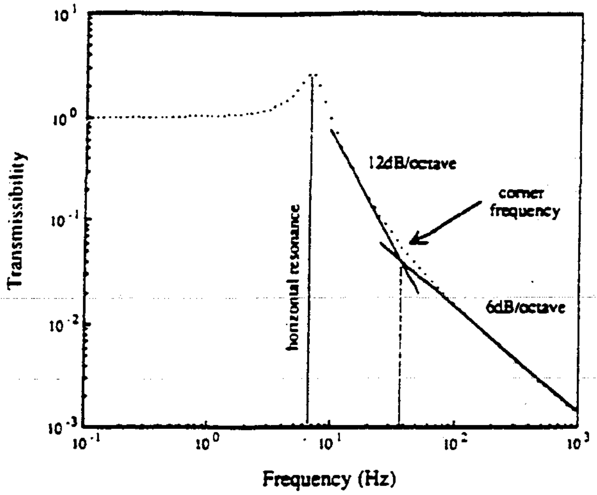


FIG. 11. Horizontal drive to horizontal response in the one-stage stack.

It should be noted that certain of the couplings between directions disappear if the geometry can be arranged to remove the vertical offsets mentioned above. However such arrangements are potentially difficult to realize. The magnitude of coupling in set (c) described above depends on the magnitude of the imbalance of the stiffness of the rubber pads. For convenience a summary table of the transmissibility values obtained for the one-stage stack at a frequency of 100 Hz is given in Fig. 18.

The situation with a two-stage stack is somewhat more complicated since there are a larger number of degrees of

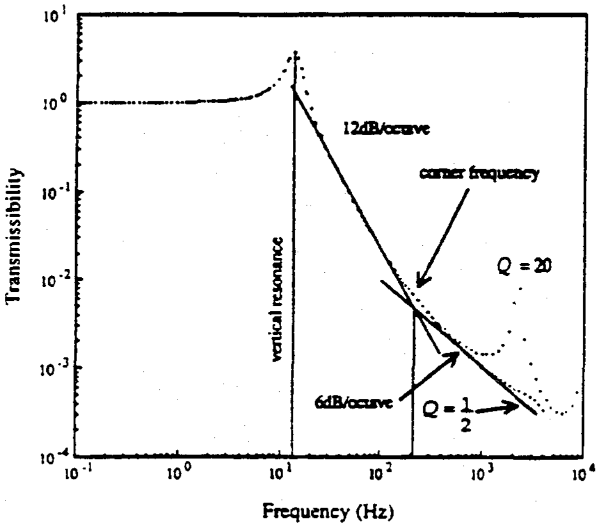


FIG. 10. Vertical drive to vertical response in the one-stage stack.

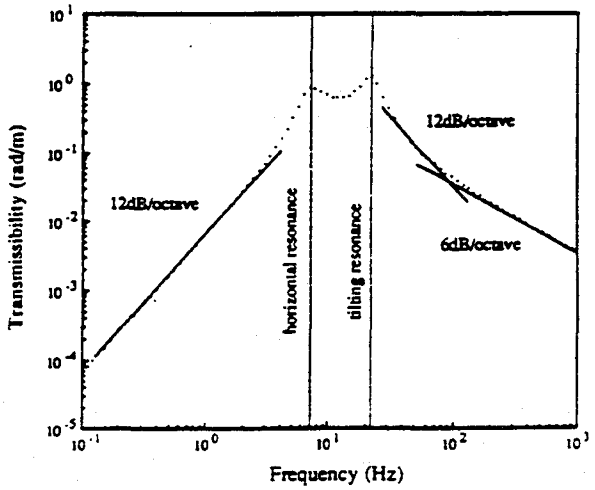


FIG. 12. Horizontal drive to tilt response in the one-stage stack.

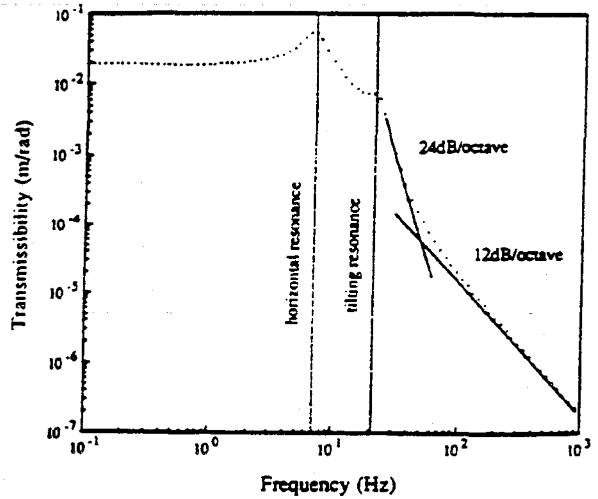


FIG. 13. Tilt drive to horizontal response in the one-stage stack.

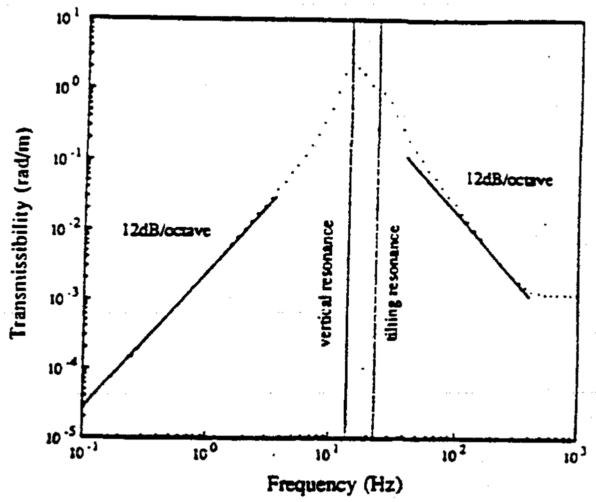


FIG. 15. Vertical drive to tilt response in the one-stage stack with a 10% stiffness imbalance in the rubber springs.

freedom and various competing routes for the transmission of motions exist. Consider, for example, the case of horizontal response of the second plate driven by horizontal motion of the base. The largest effect, as one would expect, is due to the horizontal motion of the base exciting horizontal motion in the first plate, which in turn excites horizontal motion of the second plate. There are other effects, however, such as horizontal base motion exciting tilt of the first plate which leads to horizontal motion of the second plate.

Some of the results for the arrangements of two-stage stacks given in Fig. 4 are shown in Figs. 19–25, and again these can be understood from the dynamics of the system. In most cases the response at the first plate is similar to that for the single stage stack and the second stage provides extra isolation as would be intuitively expected. However there are three exceptions to this, and these result from the dynamics involved in the transmission route which dominates:

(1) *Tilt to horizontal transmissibility* (Fig. 21): Tilt drive (base) to horizontal response (center of mass 2) comes about from tilting motion (base) to tilting motion (center of mass 1) followed by conversion to horizontal motion at the top of mass 1 (intraplate conversion) and then to horizontal motion (center of mass 2).

(2) *Vertical to horizontal transmissibility* (Fig. 24): Vertical drive (base) to horizontal response (center of mass 2) comes about from vertical motion (base) to tilting motion (center of mass 1) *to horizontal motion* (top of mass 1) to horizontal motion (center of mass 2).

(3) *Horizontal to vertical transmissibility* (Fig. 25): Horizontal drive (base) to vertical response (center of

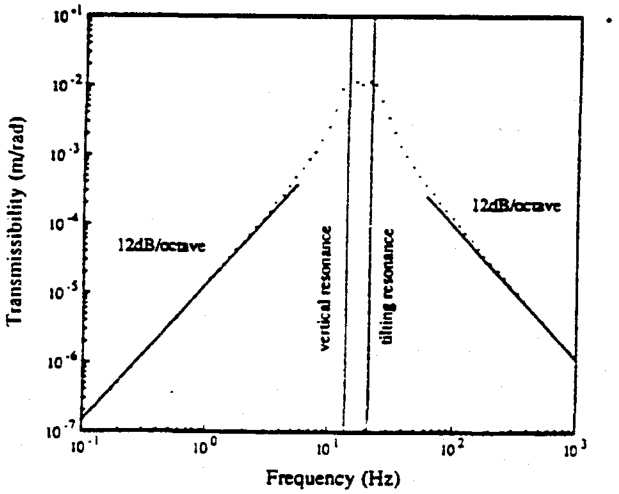


FIG. 14. Tilt drive to vertical response in the one-stage stack with a 10% stiffness imbalance in the rubber springs.

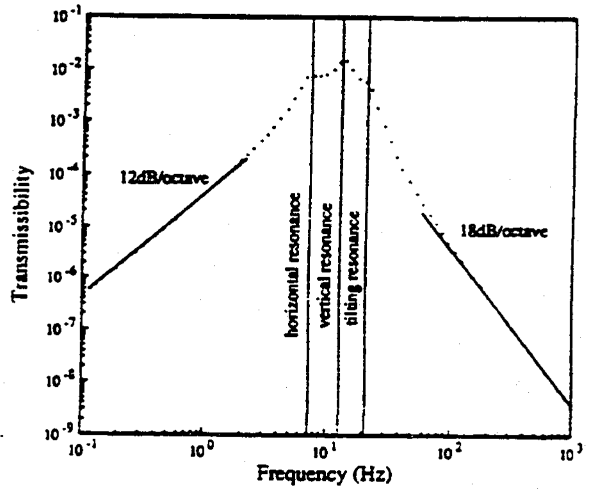


FIG. 16. Vertical drive to horizontal response in the one-stage stack with a 10% stiffness imbalance in the rubber springs.

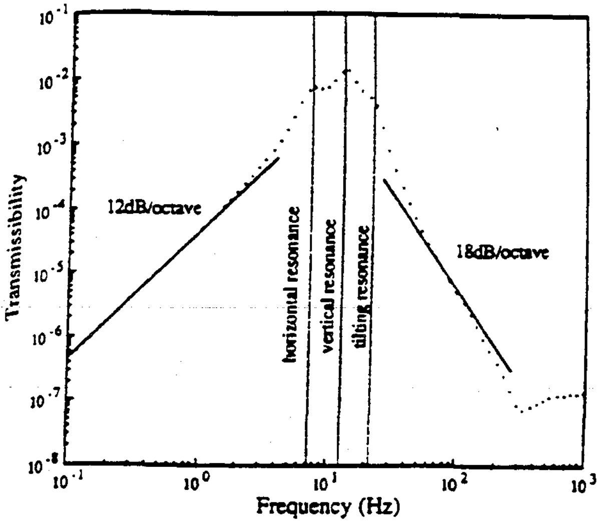


FIG. 17. Horizontal drive to vertical response in the one-stage stack with a 10% stiffness imbalance in the rubber springs.

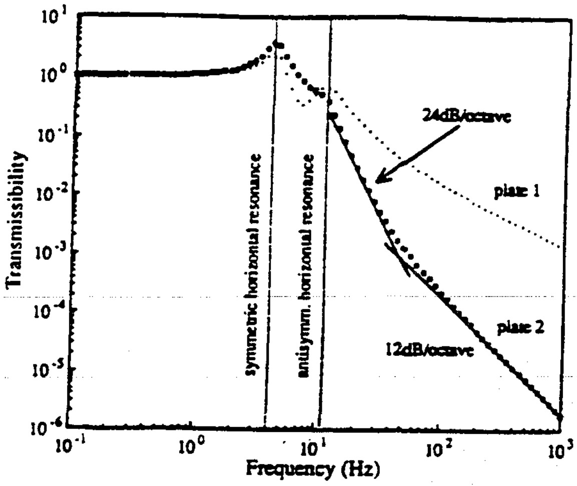


FIG. 19. Horizontal drive to horizontal response in the two-stage stack.

mass 2) arises from horizontal motion (base) to tilting motion (center of mass 1) to vertical motion (center of mass 2).

Other interesting factors emerge from these analyses. For example in the case of vertical drive (base) to horizontal response (center of mass 2) the horizontal motion at stage 1 is smaller than that at stage 2 (Fig. 24). This arises from a balancing of the restoring forces acting at plate 1 due to the presence of stage 2 above it. In most cases for the two-stage system it does not matter whether the 10% imbalance in rubber properties is symmetrical or asymmetrical for the two stages. An exception to this is the response of stage 1 in the case of horizontal drive (base) to vertical response (center of mass 2). Figure 25 shows the situation for symmetrically imbalanced rubber where there is little dynamical balancing for stage 1. Even with the presence of the overlying stage there remains a difference in the vertical stiffness at each end of plate 1. Its induced tilting motion will therefore not be about its center of mass and will lead to vertical movement of the center of mass. For asymmetrically imbalanced rubber there is a dynamical balancing effect on stage 1 which reduces this vertical movement by a large factor.

VII. EXTENSION TO THE UNDERSTANDING OF A FOUR-STAGE STACK

It is clearly possible to carry out the same types of analyses for multiple stage stacks. However, in the work described above we encountered two numerical problems. One was due to the fact that the stiffnesses of the aluminum and rubber are very disparate. The other problem was seen at high frequencies where the degree of isolation is such that the output is very small compared to the input. The first problem manifests itself in error messages from NASTRAN and leads to potential inaccuracies in the results. At the suggestion of the NASTRAN vendors we did

OUT IN	*HORIZONTAL*	*VERTICAL*	*TILT*
HORIZONTAL	-1.6×10^{-2}	**-5.8×10^{-6}**	-4.4×10^{-2} rad/m
VERTICAL	**-5.1×10^{-6}**	-2.3×10^{-2}	**-2.3×10^{-2} rad/m**
TILT	-2.1×10^{-5} m/rad	**-1.2×10^{-4} m/rad**	-6.4×10^{-2}

FIG. 18. Summary of the various transmissibility values at 100 Hz for the one-stage stack. Figures in bold are given for a 10% stiffness imbalance in the rubber springs.

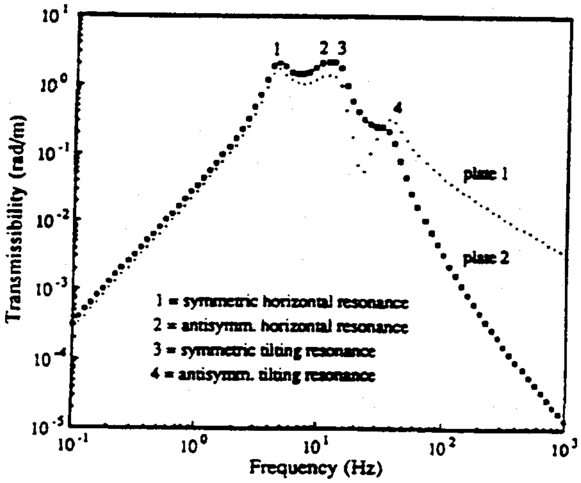


FIG. 20. Horizontal drive to tilt response in the two-stage stack.

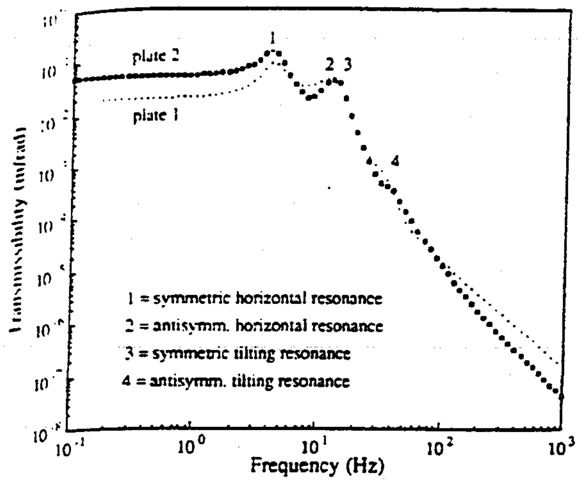


FIG. 21. Tilt drive to horizontal response in the two-stage stack.

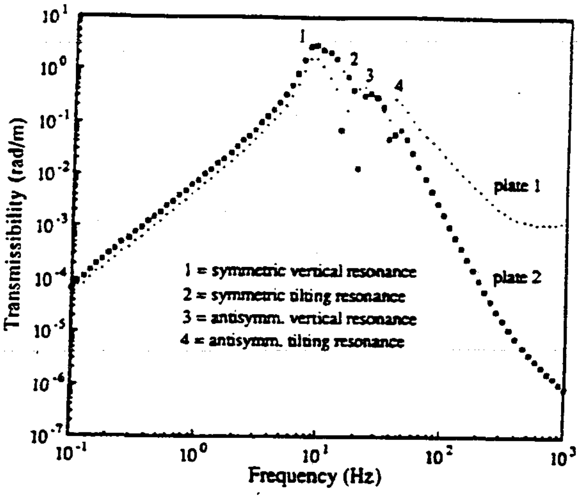


FIG. 23. Vertical drive to tilt response in the two-stage stack with a 10% stiffness imbalance in the rubber springs.

not discard results when error messages occurred, but took care to examine the results carefully for the kind of inaccuracies characteristic of such problems. The second problem manifested itself as a flattening out of transmissibility at a value of around 10^{-7} or lower. This effect was reduced to this acceptable level by using the "direct" rather than "modal" method of calculating transmissibilities.[4] However the direct method becomes very demanding in computing time for models of more than two stages.

Fortunately, from considerations of the mechanisms involved in the transmissibility of one- and two-stage stacks, it is possible to predict relatively easily the main features of the isolation performance of a stack with more stages. For example, the principal transmission routes for both the vertical and horizontal responses of a four-stage stack are shown in Figs. 26 and 27, and the performance at 100 Hz in each case has been calculated. It is clear from these results that by using four stages of the particular design considered here, sufficient vertical isolation can be achieved and that the horizontal isolation is much better than required for our particular purpose. (If there were much less cross coupling of vertical to horizontal motion occurring in the subsequent pendulum suspension, less vertical isolation would be required and the number of stages required in the stack could be reduced.)

We assume that the stages have the dimensions used earlier but in order to generalize the analyses we allow the possibility of embedding the bases of the rubber supports into the underlying aluminum plates to control the degree of conversion of tilting motion of the center of mass of a plate to horizontal motion at the base of the rubber (intraplate conversion). The relevant variable here is *a*, the vertical distance between the center of mass of a metal plate and the level above it at which the rubber contacts the plate. We also allow for the fact that the suspension point of the pendulum system mounted from the stack may

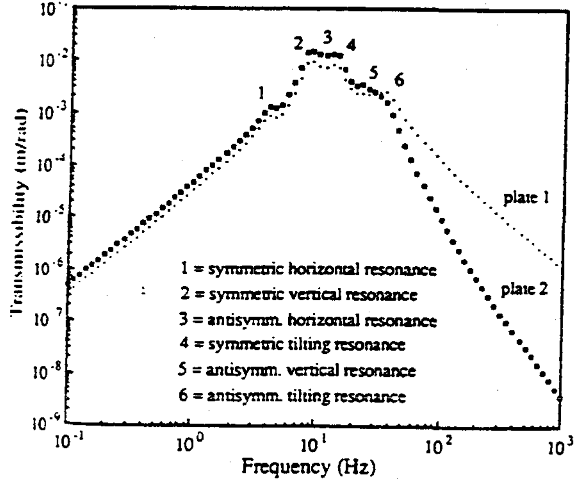


FIG. 22. Tilt drive to vertical response in the two-stage stack with a 10% stiffness imbalance in the rubber springs.

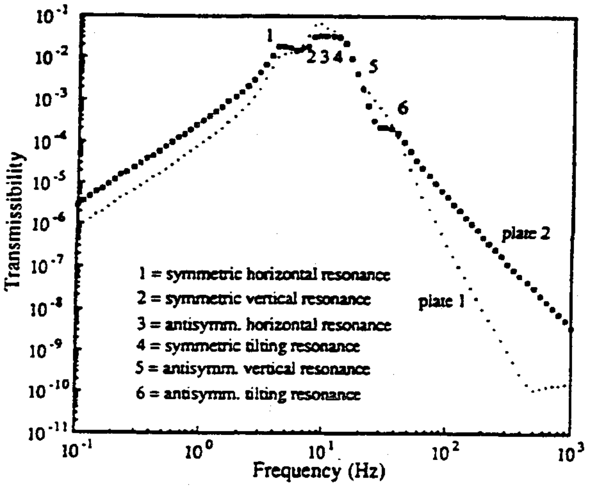


FIG. 24. Vertical drive to horizontal response in the two-stage stack with a 10% stiffness imbalance in the rubber springs.

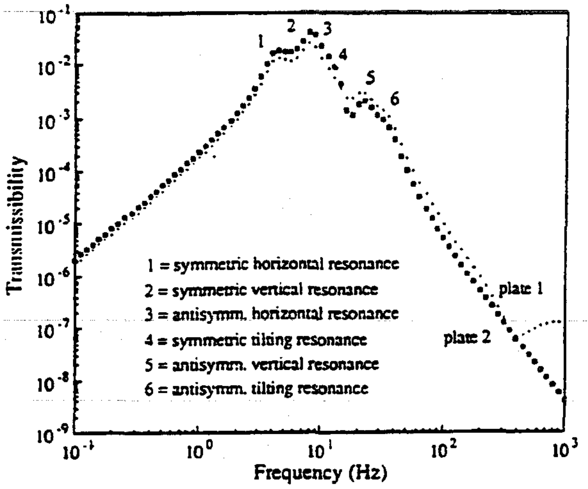


FIG. 25. Horizontal drive to vertical response in the two-stage stack with a 10% stiffness imbalance in the rubber springs.

be displaced a vertical distance z and a horizontal distance x from the center of mass of the top plate. It is interesting to see that for the dimensions chosen the most important route for the vertical motion is the simplest route, i.e., by direct vertical coupling from stage to stage. However in the horizontal case the simplest type of coupling up the stack is not the most important mechanism. For the four-stage stack studied the most important mechanism (route 2) involves intraplate conversion, but the effect of this may be greatly reduced by allowing the rubber springs to be embedded into the underlying plates, especially near the top of the stacks. Assuming this is done, there remains another important transmission route for introducing horizontal motion to the pendulum suspension point. This is direct transmission of tilts up the stack and linear conversion to horizontal movement by the vertical offset of the suspension point. Reduction of this coupling to a level such that the noise introduced will be less than that introduced by the simplest horizontal transmission route will require very careful adjustment of the suspension point.

It is worth recalling from Fig. 8 that in a single stage the conversion of horizontal movement into tilt comes about from two apparently separate torques exerted on the plate. Simple dynamical analysis shows that these are related to each other and cancel each other out to give no tilt if the center of mass of the plate is arranged to lie on a line passing through the midpoints of the rubber supports. Unfortunately simplicity of design is lost if we wish to extend this to a multiple stage stack arrangement and cut out horizontally driven tilts, as the center lines of the rubber supports for each stage would have to be coincident. Fortunately it would seem that it is not very important to do this for our particular application.

VIII. DISCUSSION

From the finite element analysis of one- and two-stage stack systems and the extension of these findings to a four-stage stack it is clear that the simple dynamical models often may not be adequate as they do not account for some potentially important routes for noise transmission. However the analyses do suggest that performance of the order required for gravitational wave detectors should be attainable. Particularly high performance is achievable if care is taken in the mounting of the rubber at stages close to the top of the stack to reduce intraplate conversion, and if care is taken in reducing the separation in the vertical and horizontal directions of the suspension point of the pendulum from the center of mass of the top plate.

It should be pointed out that gravity was not included in this work. Its inclusion may affect overall static stability and would probably have some dynamic effects, such as the increased loading on the lower stages of the stack causing changes of stiffness of the rubber and hence changes in the resonant frequencies of the lower stages. However provided attention is paid to maintaining static stability, and provided any changes in rubber stiffness are allowed for at

	transmission route	transmissibility (T) at 100Hz	vertical motion (dz) of suspension point at 100Hz
1	v_{base} - v_{com1} - v_{com2} - v_{com3} - v_{com4} (v_{sp})	$\sim 2.8 \times 10^{-7}$	$\sim 2.8 \times 10^{-18}$ m/$\sqrt{\text{Hz}}$
2	t_{base} - t_{com1} - t_{com2} - t_{com3} - $[t_{com4}$ - $v_{sp}]$	$\sim 3.4 \times 10^{-8} \left[\frac{x}{2.0 \times 10^{-3}}\right]$ m/rad	$\sim 6.8 \times 10^{-19} \left[\frac{x}{2.0 \times 10^{-3}}\right]$ m/$\sqrt{\text{Hz}}$
3	† t_{base} - t_{com1} - t_{com2} - t_{com3} - v_{com4} (v_{sp})	$\sim 3.1 \times 10^{-8}$ m/rad	$\sim 6.2 \times 10^{-19}$ m/$\sqrt{\text{Hz}}$
4	h_{base} - t_{com1} - t_{com2} - t_{com3} - $[t_{com4}$ - $v_{sp}]$	$\sim 2.3 \times 10^{-8} \left[\frac{x}{2.0 \times 10^{-3}}\right]$	$\sim 2.3 \times 10^{-19} \left[\frac{x}{2.0 \times 10^{-3}}\right]$ m/$\sqrt{\text{Hz}}$

FIG. 26. The main transmission routes predicted to give rise to vertical motion at the top of the four-stage stack. Horizontal, vertical, and tilting motions are denoted by h, v, and t, respectively. Round brackets indicate that the motion of the pendulum suspension point (sp) is equivalent to that of the center of mass (com) of the top plate of the stack. Square brackets are used where cross coupling of motions occurs within a metal plate. Here x is the horizontal offset between the pendulum suspension point and the center of mass of the top plate of the stack. We have used an offset x of 2 mm as the reference offset since this would be a reasonable estimate of the accuracy to which the top plate center of mass and suspension point might be made coincident. The † symbol in route 3 indicates that a 10% stiffness imbalance exists in the rubber springs of stage 4.

transmission route	transmissibility (T) at 100Hz	horizontal motion (dx) of suspension point at 100Hz
1 h_{base} - h_{com1} - h_{com2} - h_{com3} - h_{com4} (h_{sp})	-6.6×10^{-8}	-6.6×10^{-19} m/$\sqrt{\mathrm{Hz}}$
2 t_{base} - t_{com1} - t_{com2} - [t_{com3} - h_{top3}] - h_{com4} (h_{sp})	$-5.5 \times 10^{-8} \left[\frac{a}{1.3 \times 10^{-2}}\right]$ m/rad	$-1.1 \times 10^{-18} \left[\frac{a}{1.3 \times 10^{-2}}\right]$ m/$\sqrt{\mathrm{Hz}}$
3 t_{base} - t_{com1} - [t_{com2} - h_{top2}] - h_{com3} - h_{com4} (h_{sp})	$-1.4 \times 10^{-8} \left[\frac{a}{1.3 \times 10^{-2}}\right]$ m/rad	$-2.8 \times 10^{-19} \left[\frac{a}{1.3 \times 10^{-2}}\right]$ m/$\sqrt{\mathrm{Hz}}$
4 h_{base} - t_{com1} - t_{com2} - [t_{com3} - h_{top3}] - h_{com4} (h_{sp})	$-3.7 \times 10^{-8} \left[\frac{a}{1.3 \times 10^{-2}}\right]$	$-3.7 \times 10^{-19} \left[\frac{a}{1.3 \times 10^{-2}}\right]$ m/$\sqrt{\mathrm{Hz}}$
5 † v_{base} - t_{com1} - t_{com2} - [t_{com3} - h_{top3}]- h_{com4} (h_{sp})	$-2.0 \times 10^{-8} \left[\frac{a}{1.3 \times 10^{-2}}\right]$	$-2.0 \times 10^{-19} \left[\frac{a}{1.3 \times 10^{-2}}\right]$ m/$\sqrt{\mathrm{Hz}}$
6 t_{base} - t_{com1} - t_{com2} - t_{com3} - [t_{com4} - h_{sp}]	$-3.4 \times 10^{-8} \left[\frac{z}{2.0 \times 10^{-3}}\right]$ m/rad	$-6.8 \times 10^{-19} \left[\frac{z}{2.0 \times 10^{-3}}\right]$ m/$\sqrt{\mathrm{Hz}}$
7 h_{base} - t_{com1} - t_{com2} - t_{com3} - [t_{com4} - h_{sp}]	$-2.3 \times 10^{-8} \left[\frac{z}{2.0 \times 10^{-3}}\right]$	$-2.3 \times 10^{-19} \left[\frac{z}{2.0 \times 10^{-3}}\right]$ m/$\sqrt{\mathrm{Hz}}$

FIG. 27. The main transmission routes predicted to give rise to horizontal motion at the top of the four-stage stack. Horizontal, vertical, and tilting motions are denoted by h, v, and t, respectively. Round brackets indicate that the motion of the suspension point (sp) is equivalent to that of the center of mass (com) of the top plate of the stack. Square brackets are used where cross coupling of motions occurs within a metal plate. Here, a is the vertical offset between the center of mass of a plate and the base level of the rubber supports for the stage above, and z is the vertical offset between the pendulum suspension point and the center of mass of the top plate of the stack. We have used an offset z of 2 mm as the reference offset since this would be a reasonable estimate of the accuracy to which the top plate center of mass and suspension point might be made coincident. The † symbol in route 5 indicates that a 10% stiffness imbalance exists in the rubber springs of stage 1.

the design stage, such effects are not expected to significantly alter the findings presented above. It should also be noted that this was not a full three-dimensional an..:ysis. Rotations of the stack about a vertical axis were not accounted for in the finite element model: yet these could be important for the coupling of rotational ground noise to horizontal noise particularly if the suspension point of the pendulum was offset in the y direction from the axis of symmetry of the stack (Fig. 1). Also stiffness imbalance in the rubber springs at either side of the stack in this situation could lead to more complicated transmission routes and would justify further analysis.

ACKNOWLEDGMENTS

We would like to thank J. E. Hall of Rutherford Appleton Laboratory and our colleagues in the Gravitational Waves Group at Glasgow University, in particular P. J. Veitch, for many useful discussions and suggestions. This work was supported by SERC and the University of Glasgow and one of us (C.A.C.) wishes to thank SERC for the award of a studentship.

[1] D. G. Blair, in *The Detection of Gravitational Radiation*, edited by D. G. Blair (Cambridge University, Cambridge, England, 1991), p. 73.

[2] J. Hough, B. J. Meers, G. P. Newton, N. A. Robertson, H. Ward, G. Leuchs, T. M. Niebauer, A. Rudiger, R. Schilling, L. Schnupp, H. Walther, W. Winkler, B. F. Schutz, J. Ehlers, P. Kafka, G. Schafer, M. W. Hamilton, I. Schutz, H. Welling, J. R. J. Bennett, I. F. Corbett, B. H. W. Edwards, R. J. S. Greenhalgh, and V. Kose, Proposal for a joint German-British interferometric gravitational wave detector [Technical Report MPQ 147 (Max-Planck Institute of Technology and GWD/137/JH(89) (Rutherford Appleton Laboratory), September, 1989].

[3] The MacNeal-Schwendler Corporation, Handbook for Linear Analysis, MSC/Nastran Version 64 (California, 1985).

[4] The MacNeal-Schwendler Corporation, Handbook for Dynamic Analysis, MSC/Nastran Version 64 (California, 1983).

[5] R. Del Fabbro, A. Di Virgilio, A. Giazotto, H. Kautzky, V. Montelatici, and D. Passuello, Phys. Lett. A 133, 9 (1988).

[6] C. A. Cantley, PhD thesis, Glasgow University, Scotland, G12 8QQ, 1991.

[7] J. C. Snowdon, NBS Handbook 128, US Department of Commerce/National Bureau of Standards (Washington, DC, 1979).

[8] R. J. S. Greenhalgh, Rutherford Appleton Laboratory, Chilton, Oxfordshire (private communication, 1989).

[9] FEGS Ltd. FAMresult Reference Manual (Cambridge, 1987).

[10] J. E. Hall, Rutherford Appleton Laboratory, Chilton, Oxfordshire (private communication, 1991).

REF. LL

A Double Pendulum Vibration Isolation System for a Laser Interferometer Gravitational Wave Antenna

M. Stephens, P. Saulson, and J. Kovalik

Rev. Sci. Instrum. **62**, 924–932 (1991)

This reference turns out to be the same as Reference CC, which appears earlier in this volume. Therefore, we do not reproduce it here.

REF. MM

Meas. Sci. Technol. 3 (1992) 463–470. Printed in the UK

High dynamic range measurements of an all metal isolator using a sapphire transducer

L Ju, D G Blair, H Peng and F van Kann

Department of Physics, University of Western Australia, Nedlands, 6009 Australia

Received 30 October 1991, accepted for publication 15 January 1992

Abstract. We report the performance and mechanical properties of an all metal mass–spring isolator with corner frequency about 44 Hz. Noise floor measurements using a prototype sapphire transducer and results of active damping of the low-frequency normal modes are given. Our measurements cover a total dynamic range of 160 dB, and show that no excess noise is generated above the thermal noise limit, even in the presence of high-amplitude excitation.

1. Introduction

It is of great importance that terrestrial gravitational wave detectors of both resonant bar and laser interferometer types are isolated from external vibrations such as the seismic noise background. Mechanical suspensions are required to filter out the noise.

Traditional isolators use alternating layers of lead and rubber [1]. They provide good isolation under normal conditions. The disadvantage of these isolators is that they are not suitable in high vacuum and have problems of degassing and thermal stability. Another type of isolator, the gas–spring isolator [2], has high load bearing capacity and low resonant frequency. However, gas–springs have strong temperature coefficients, so such isolators experience problems of thermal stability as well as complexity. A variety of suspensions have been developed for gravitational wave detectors [3–5]. All are designed to have normal mode resonant frequencies well below the antenna frequency and internal modes of the isolator elements above the frequency range of interest. In general the normal modes define a set of low-frequency resonances. Internal modes of the mass or spring elements are generally at high audio frequencies. Thus, such isolators have good isolation above a low-frequency corner (the highest of the low-frequency modes), and below the high-frequency internal resonances. For resonant bar antennae, the isolation band required is from a few hundred Hz to a few kHz while in interferometer antennae, the corner frequency needs to be pushed to a few tens of Hz or lower. The isolators must also be strong enough to support the weight of up to several tons.

The low-frequency normal modes are a major problem, however. In the case of the University of Western Australia (UWA) parametric transducer on the Nb bar antenna, low-frequency modes modulate the cavity resonant frequency, which itself modulates both the antenna Q factor and the antenna resonant frequency [6, 7]. The operation of an interferometer is identical to that of a parametric transducer. The parametric interaction is weaker due to the higher photon frequency, but the normal modes make it difficult to keep the interferometer cavity stable [8]. High damping is required to make the amplitude of the low-frequency normal modes as low as possible.

Since the isolator must be designed to achieve very low vibration levels, it is difficult to test the properties of the isolator by conventional accelerometers. Electrical pick-up from vibrators is also a problem using conventional accelerometers.

We present here an all metal multistage mass–spring isolator having a corner frequency about 44 Hz. Active damping is used to damp the normal mode amplitudes at low frequencies. A newly developed sapphire transducer [9, 10] which is not influenced by electrical pick-up is used to measure the noise performance of the isolation stack. This is achieved using radiation coupling to the transducer, which avoids all vibration coupling along cables. This allows vibration performances to be tested down to about 10^{-14} m Hz$^{-1/2}$.

The isolator described here is suitable for use both at low temperatures and high vacuum. The overall goal of this work is to investigate isolators at the highest possible sensitivity, and to search for non-ideal behaviour, nonlinear noise generation mechanisms, etc. Work to date has used local measurements in a single isolator. We extend this to relative measurements between isolators, using the most sensitive optical readout possible, and under high levels of acoustic excitation.

0957-0233/92/050463 + 08 $04.50 © 1992 IOP Publishing Ltd

2. Modelling of multistage isolators

A one-dimensional numerical computer model is used to study the vertical vibration behaviour of the isolator. The displacement of the *i*th element x_i is given by

$$[-\omega_i^2 m_i + k_i + k_{i+1} + i\omega_i(d_i + d_{i+1})]x_i - (k_i + i\omega_{i-1}d_i)x_{i-1}$$
$$-(k_{i+1} + i\omega_{i+1}d_{i-1})x_{i+1} = F_i + \sqrt{4\kappa T_i d_i}$$

where k_i is the spring constant, d_i is the damping, F_i the applied force for the *i*th mass–spring element and $\sqrt{4\kappa T_i d_i}$ is introduced to the *i*th element as a Nyquist force. Here κ is Boltzmann's constant, m_i the *i*th mass and T_i the temperature of the *i*th element. The frequency $f_i = \omega_i/2\pi = 1/2\pi\ (k_i/m_i)^{1/2}$ and the Q factors $Q_i = \omega_i m_i/d_i$ are inserted as adjustable parameters. Seismic noise with an amplitude (in metres) and frequency dependence close to that observed in a laboratory environment, given by $x_0 = 10^{-6}/f^2$, is also introduced to the first element.

It is well known that at frequencies well above the natural frequency of the elements of the isolator there will be an amplitude attenuation about $(f/f_0)^{2N}$ where N is the number of elements used in the isolator, and f_0 is the natural frequency. But at the normal mode frequencies the amplitude is enhanced by the Q factor of the mode. However, by actively damping the motion of the elements, one can attenuate the normal mode amplitude at low frequencies. Such damping should normally have the characteristics of cold damping, contributing noise far less than that expected for a passively damped system.

Using the above model, we have experimented with a range of Q factors in different elements. We find that significantly damping one element in the stack can give good attenuation of the low-frequency peaks without seriously degrading the high-frequency performance, even in the case of passive damping. Figure 1 shows response curves from the computer model with damping applied to different elements in the presence of f^{-2} seismic noise. From this we can see that there is a trade-off between damping the low-frequency normal modes and degrading the high-frequency performance of the isolator. To achieve a low amplitude in the first mode it is essential to have high damping in the first spring. By adding damping to the fourth spring all the normal modes are effectively suppressed without increasing the high-frequency noise, which is dominated by Nyquist noise.

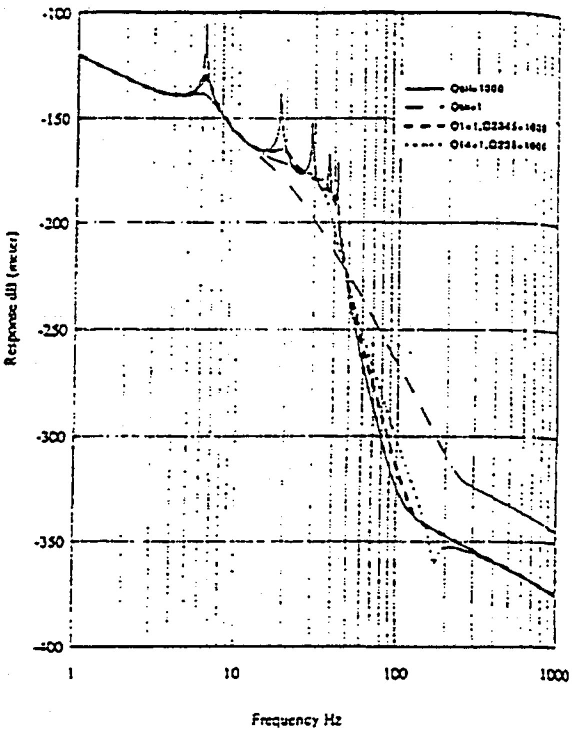


Figure 1. Computer modelling of damping on different elements.

Active damping should always contribute less high-frequency noise than passive damping, although this source of noise does not appear serious for the Q factors considered here.

3. Construction

3.1. Design of the isolator

The main idea of the isolator is that it is composed of metal suitable for both high vacuum and cryogenic

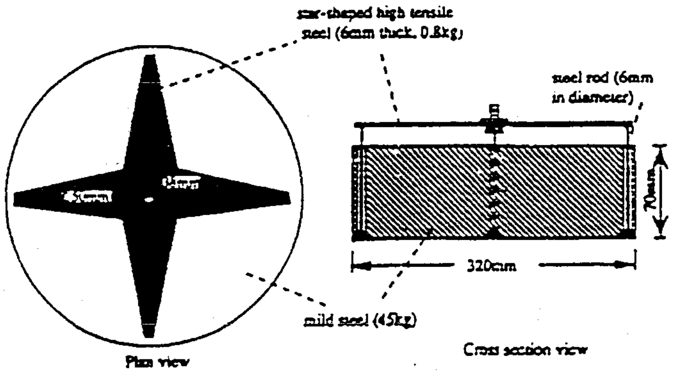

star-shaped high tensile steel (6mm thick, 0.8kg)
steel rod (6mm in diameter)
45mm
12mm
70mm
320mm
mild steel (45kg)
Plan view
Cross section view

Figure 2. Isolator element.

conditions, is as simple as possible, and at the same time strong enough to support tons of weight. The spring–mass element of the isolator is shown in figure 2. The mass of the element is a 45 kg mild steel disc. The spring is a 6 mm thick, star-shaped, high-tensile steel (ASTMA514) plate which has tensile strength of 1260 MPa (1.26×10^9 N m^{-2}). The shape of the plate is designed to have approximately uniform stress over the plate and to sustain a maximum load of about 3000 kg. Strength tests of the spring give a spring constant of $k = 1.1 \times 10^6$ N m^{-1} which agrees reasonably with the theoretical value of 1.2×10^6 N m^{-1}. The disc mass is connected to the spring with four 110 mm long, 6 mm diameter high-tensile bolts. These act as pendulums to provide flexibility in orthogonal horizontal directions. The natural frequency of the element is 22 Hz in the longitudinal mode, and about 10 Hz horizontally.

The isolator is suspended in a steel frame in a vacuum tank as shown in figure 3. Two mass loaded high-power loudspeakers are mounted vertically and horizontally to provide excitation in orthogonal directions. The driving force is provided by the reaction force from the acceleration of the loudspeaker diaphragm. At frequencies below about 15 Hz the displacement of the loudspeaker is dominated by the restoring force of its spring loaded diaphragm. Adding extra weight to the diaphragm improves the low-frequency response for reaction driving and increases the dynamic range. However, reaction force driving is intrinsically limited at low frequencies, because of the finite linear travel that can be achieved: in practice this is limited to about 10 mm of travel.

3.2. Active damping arrangement

Since there are large phase shifts between the different elements, we have used a single-stage active feedback loop as shown in figure 4. The signal from one element is detected by a geophone (a velocity sensor with sensitivity 28 V m^{-1} s), amplified, filtered and fed back through a servo-controller to a loudspeaker attached to the same element. The forces are generated entirely by inertial reaction as described above. To increase the low-

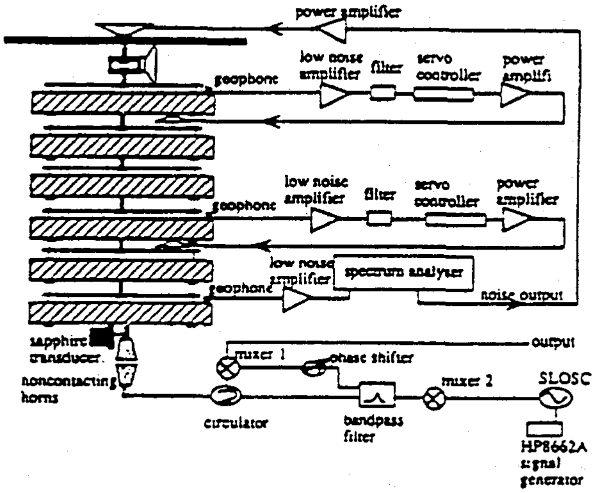


Figure 4. Single-stage active feedback arrangement and the microwave circuit of the sapphire transducer.

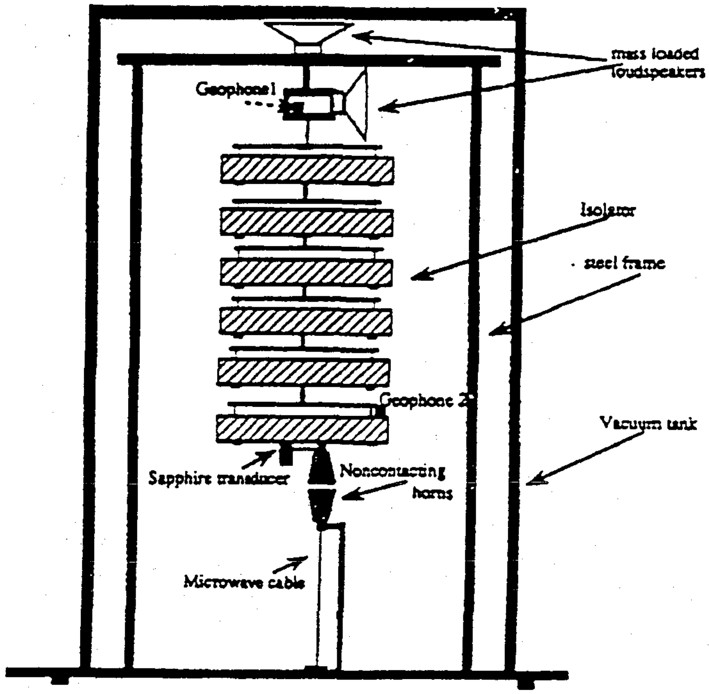


Figure 3. Isolator configuration and its testing arrangement.

frequency response the loudspeakers here are also mass loaded. Reaction force driving entirely avoids external vibrations entering through the driver. This cannot be achieved by any force driver attached to the outside world. This method does, however, lead to some stability difficulties.

The filter in the feedback loop is either set up as a low-pass filter, to achieve broad band velocity feedback, or as a high-Q band-pass filter, with a typical bandwidth of less than 1 Hz. Generally the narrow band feedback is applied to the first or second normal mode, while broad band feedback (typical bandwidth 0–100 Hz) is used to damp the higher modes. Active damping loops were tested on various stages of the isolator.

4. Results

4.1. The response of the isolator

The response of the isolator is in good agreement with the theoretical model. Figure 5(*a*) shows the experimental response curve of a five-stage isolator under vertical excitation driven by a white noise source. Figure 5(*b*) is the theoretical response curve. It can be seen that they are in good agreement. It can also be seen that there is a sharp roll-off at about 44 Hz. Because of the dynamic range of the amplifier and spectrum analyser on which the experimental results are shown, the amplitude above 60 Hz is buried in the instrument noise floor. Separate measurements of the amplitude at higher frequencies

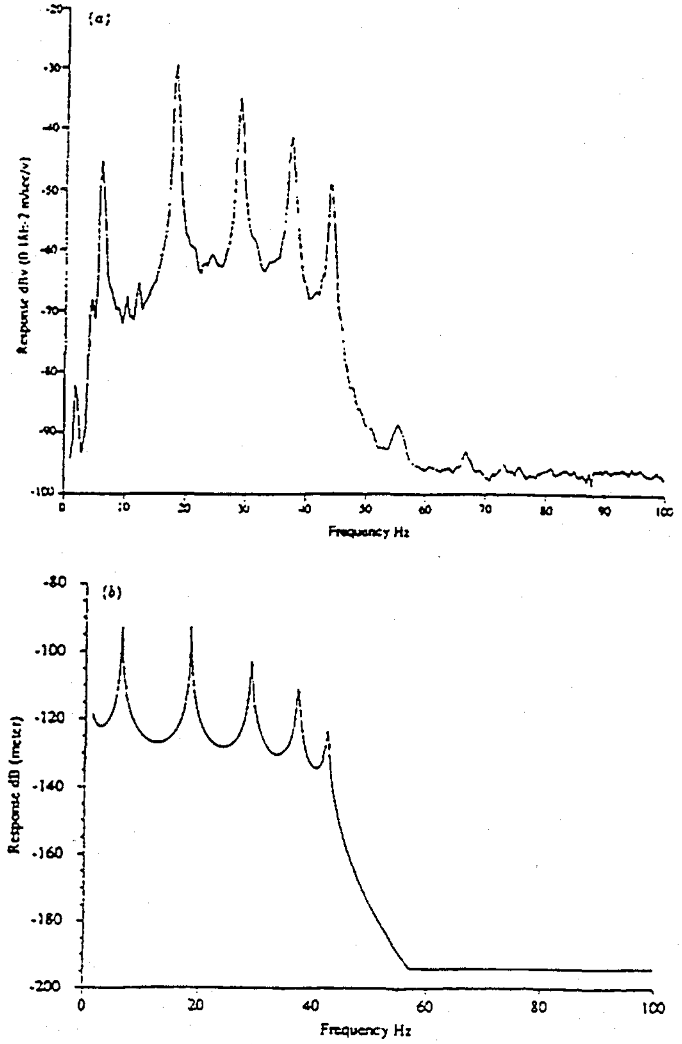


Figure 5. Response curve of a five-stage isolator and a noise floor set by the geophone sensitivity. (*a*) Velocity signal from geophone: (*b*) theoretical amplitude response curve.

and low frequencies must be used to achieve greater dynamic range. These are described below.

The horizontal response is more complicated, but the corner frequency is as expected—a typical result is shown in figure 6. The isolator clearly gives good attenuation in horizontal directions, although there is some mixing of the horizontal response with the vertical response, causing some additional low-level normal modes.

4.2. Sapphire dielectric resonator transducer measurements

As mentioned above electrical pick-up is a serious problem for geophone transducers, causing a strong reduction in measurement dynamic range when the isolator is driven at high powers. This problem has been overcome by using the newly developed microwave sapphire dielectric resonator transducer [8, 9], which is more sensitive and is not affected by electrical pick-up. The transducer can also be used in a non-contacting mode which eliminates vibration coupling through the microwave cables. This is achieved by coupling the microwave input and output signals through a pair of non-contacting horns as shown in figure 3. The sapphire transducer utilizes the strong tuning achieved when the spacing between two sapphire dielectric resonators is altered. The tuning coefficient is typically 1 MHz μm^{-1} and the resonator Q factor is about 10^5. The transducer is pumped by an ultra-low-noise sapphire loaded superconducting cavity (SLOSC) oscillator. The transducer is attached to the bottom of the isolator. The microwave signal from the SLOSC source is modulated by the sapphire transducer and demodulated in the mixer, as shown in figure 4. The final signal is amplified and displayed on a spectrum analyser. Under ambient seismic drive on a five-stage isolator, the noise floor is −143.6 dB V $Hz^{-1/2}$ which corresponds to the amplitude of 1.8×10^{-14} m $Hz^{-1/2}$, calibrated by using the position bandwidth [9] of the sapphire transducer. The performance of the sapphire transducer is demonstrated by a comparison of the response of a geophone and the sapphire transducer in a four-stage isolator experiment shown in figure 7. It can be seen clearly that the signal detected by the sapphire transducer at 200–300 Hz is completely hidden in the electrical pick-up floor of the geophone. (A narrow peak at 90 Hz is due to a resonance in the microwave horn.)

The sapphire transducer is designed to have a mechanical resonant frequency of 60 Hz. This was chosen to reduce its sensitivity to the low-frequency normal modes. Still, due to the enormous dynamic range expected over the frequency range considered here, we are faced with the problem that the instrumental dynamic range (of amplifiers and spectrum analysers) is insufficient. To solve this, we apply only high-frequency noise output to the driver to prevent extra excitation of the low-frequency modes. In addition, it is necessary to use low-noise amplifiers with high-pass filters to prevent overloads of the low-frequency normal modes. This method allowed both high gain and high excitation to be achieved without overloading the instruments.

When a five-stage isolator is excited vertically to the maximum available vibration power, there is no detectable signal from the sapphire transducer above 300 Hz. The noise floor corresponds to an amplitude of 1.8×10^{-14} m $Hz^{-1/2}$. The isolation can be measured by comparing curves (*a*), (*b*) and (*c*) in figure 8. Curve (*a*) shows the measured applied excitation at geophone 1 in figure 3. Curve (*b*) shows the normal level of seismic noise compared with the theoretical curve $10^{-6}f^{-2}$ (broken line). Curve (*c*) shows the observed noise floor of the sapphire transducer. The measured limit to the isolator is about 100 dB at 300 Hz, falling to 80 dB at

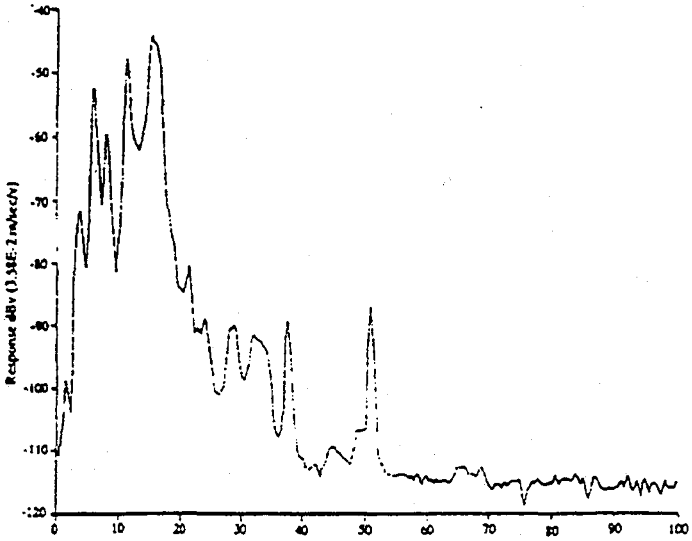


Figure 6. Horizontal velocity response curve of a five-stage isolator.