

# CONTEMPORARY PHYSICS                    VOL.I TRIESTE SYMPOSIUM 1968

PROCEEDINGS OF THE INTERNATIONAL SYMPOSIUM  
TRIESTE, 7-28 JUNE 1968, ORGANIZED BY THE  
INTERNATIONAL CENTRE FOR THEORETICAL PHYSICS, TRIESTE

Theory of condensed matter and related problems

Plasma physics, turbulence, quantum optics and statistical mechanics

Astrophysics, quasars and pulsars

Gravitational theory and cosmology

Scientific Secretary: L. FONDA   Symposium Director: ARDUS SALAM



INTERNATIONAL ATOMIC ENERGY AGENCY, VIENNA, 1969



**CONTEMPORARY PHYSICS:  
TRIESTE SYMPOSIUM 1968**

**VOL.I**

The following States are Members of the International Atomic Energy Agency:

AFGHANISTAN	GHANA	PAKISTAN
ALBANIA	GREECE	PANAMA
ALGERIA	GUATEMALA	PARAGUAY
ARGENTINA	HAITI	PERU
AUSTRALIA	HOLY SEE	PHILIPPINES
AUSTRIA	HUNGARY	POLAND
BELGIUM	ICELAND	PORTUGAL
BOLIVIA	INDIA	ROMANIA
BRAZIL	INDONESIA	SAUDI ARABIA
BULGARIA	IRAN	SENEGAL
BURMA	IRAQ	SIERRA LEONE
BYELORUSSIAN SOVIET SOCIALIST REPUBLIC	ISRAEL	SINGAPORE
CAMBODIA	ITALY	SOUTH AFRICA
CAMEROON	IVORY COAST	SPAIN
CANADA	JAMAICA	SUDAN
CEYLON	JAPAN	SWEDEN
CHILE	JORDAN	SWITZERLAND
CHINA	KENYA	SYRIAN ARAB REPUBLIC
COLOMBIA	KOREA, REPUBLIC OF	THAILAND
CONGO, DEMOCRATIC REPUBLIC OF	KUWAIT	TUNISIA
COSTA RICA	LEBANON	TURKEY
CUBA	LIBERIA	UGANDA
CYPRUS	LIBYA	UKRAINIAN SOVIET SOCIALIST REPUBLIC
CZECHOSLOVAK SOCIALIST REPUBLIC	LIECHTENSTEIN	UNION OF SOVIET SOCIALIST REPUBLICS
DENMARK	LUXEMBOURG	UNITED ARAB REPUBLIC
DOMINICAN REPUBLIC	MADAGASCAR	UNITED KINGDOM OF GREAT BRITAIN AND NORTHERN IRELAND
ECUADOR	MALAYSIA	UNITED STATES OF AMERICA
EL SALVADOR	MALI	URUGUAY
ETHIOPIA	MEXICO	VENEZUELA
FINLAND	MONACO	VIET-NAM
FRANCE	MOROCCO	YUGOSLAVIA
GABON	NETHERLANDS	ZAMBIA
GERMANY, FEDERAL REPUBLIC OF	NEW ZEALAND	
	NICARAGUA	
	NIGER	
	NIGERIA	
	NORWAY	

The Agency's Statute was approved on 23 October 1956 by the Conference on the Statute of the IAEA held at United Nations Headquarters, New York; it entered into force on 29 July 1957. The Headquarters of the Agency are situated in Vienna. Its principal objective is "to accelerate and enlarge the contribution of atomic energy to peace, health and prosperity throughout the world".

INTERNATIONAL CENTRE FOR THEORETICAL PHYSICS, TRIESTE

# CONTEMPORARY PHYSICS: TRIESTE SYMPOSIUM 1968

PROCEEDINGS OF THE INTERNATIONAL SYMPOSIUM  
ON CONTEMPORARY PHYSICS  
ORGANIZED BY AND HELD AT THE  
INTERNATIONAL CENTRE FOR THEORETICAL PHYSICS, TRIESTE  
FROM 7 TO 28 JUNE 1968

SCIENTIFIC SECRETARY: L. FONDA  
SYMPOSIUM DIRECTOR: ABDUS SALAM

*In two volumes*

## VOL.I

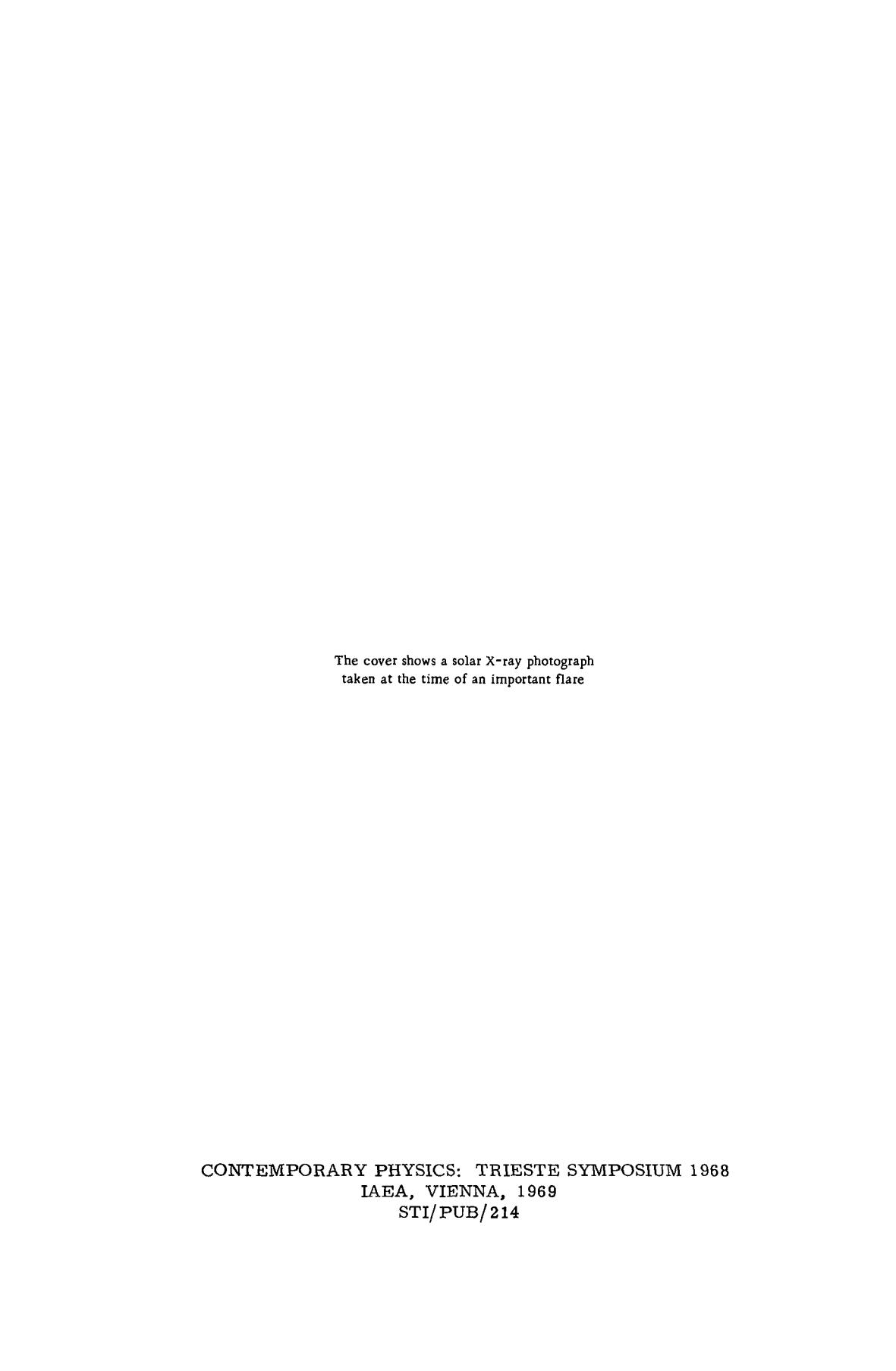
THEORY OF CONDENSED MATTER  
AND RELATED PROBLEMS

PLASMA PHYSICS, TURBULENCE, QUANTUM OPTICS  
AND STATISTICAL MECHANICS

ASTROPHYSICS, QUASARS AND PULSARS

GRAVITATIONAL THEORY AND COSMOLOGY

INTERNATIONAL ATOMIC ENERGY AGENCY  
VIENNA, 1969



The cover shows a solar X-ray photograph  
taken at the time of an important flare

CONTEMPORARY PHYSICS: TRIESTE SYMPOSIUM 1968  
IAEA, VIENNA, 1969  
STI/PUB/214

## **Contributors to Vol. I:**

A. Abrikosov, P. W. Anderson, J. G. Bolton,  
E. M. Burbidge, B. F. Burke, A. G. W. Cameron,  
B. Coppi, P. G. de Gennes, S. Deser, R. H. Dicke,  
P. A. M. Dirac, S. Doniach, T. Dupree, R. A. Ferrell,  
Michael E. Fisher, V. A. Fock, W. A. Fowler, T. Gold,  
J. B. Keller, I. M. Khalatnikov, R. P. Kraft, E. H. Lieb,  
E. M. Lifshitz, P. C. Martin, E. Montroll,  
P. Morrison, R. Penrose, C. Pethick, D. Pines,  
I. Prigogine, M. N. Rosenbluth, B. Rossi,  
E. E. Salpeter, M. Schmidt, J. R. Schrieffer,  
D. W. Sciama, F. G. Smith, H. Suhl, W. Thirring,  
W. B. Thompson, C. H. Townes, J. Weber, S. Weinberg



## FOREWORD

In June 1968, the International Centre for Theoretical Physics was privileged to hold an extended symposium on contemporary physics – an event probably unique in that its aim was to range over and review not just one specialized aspect of modern physics but its entire spectrum. The symposium brought together a group of 300 leading specialists in particle physics, theory of condensed matter, astrophysics, relativity, plasma physics, cosmology, nuclear physics, quantum electronics and biophysics; they lived together for three weeks, to deliver, to discuss and to listen to lectures in depth on their various specialities. The intention was to acquire, if possible, a deeper sense of the scope and unified nature of the general subject – physics – by sharing the insights of its fascinatingly diverse disciplines.

The present two volumes are a record of the proceedings of the symposium. The first groups papers presented on condensed and plasma matter, on earth and in the universe; the second deals with theory of matter at very small distances and fundamentals of quantum theory.

How far did we succeed in achieving the objectives of the symposium? Did we succeed in communicating to those in other disciplines the present situation in our own – the problems, the prospects? Did we emerge, three weeks after, as universal in our outlook and competence as the great physicists of even forty years ago, who could, with felicity, pass from one discipline to another enriching each? How much cross-fertilization was achieved? To what extent were the hierarchy of disciplines and the barriers between generations broken? In retrospect – and particularly in retrospect – the answers to these questions appear positive: we succeeded perhaps partially; but far more than anyone anticipated. There is no doubt that communication is a difficult art. There was so much ground to be covered before the treasures of one's own discipline could be unfolded. We could not easily forget those in the audience belonging to our own specialities. We were all guilty of exaggerating the importance of the problems of just this day – this will be found reflected in the proceedings, where survey lectures designed expressly for those in other disciplines are juxtaposed with contributions unabashedly specialized. We debated long on whether to publish the latter. In the end we decided to do so, for these contributions do sample in depth the present day's preoccupations with a vital, living subject in all its kaleidoscopic aspects, and they give a perspective to the general surveys.

Notwithstanding these specialist contributions, however, there did shine through the symposium the recognition that physics is still very much a unity – that even though they may appear under different names and guises in differing specialities, the conceptual constructs are the same; that results and techniques, both experimental and theoretical, developed for one discipline are relevant in others. If anything, we grew fonder of, more enthusiastic about our own discipline, after recognizing the problems which the others face.

But, even more than the scientific, for most of us the symposium was a human experience of a lifetime. I do not believe there ever has assembled, at one time and place and for so long, such a galaxy of first-rate physicists from so diverse fields; most lecturers were the men who had been responsible for creating their subjects. Then there was the civilizing influence of the presence among us, over three weeks, of some of the grand old men of physics — the founders of modern quantum theory — who so warmly responded to the invitation to participate. We were privileged to hear evening lectures from some of them (to be published as a Supplement to the IAEA Bulletin) on their life of physics. These were magical evenings and perhaps the most memorable part of the symposium. Finally, there was the physical beauty of the surroundings — we had just moved into the Centre's new premises, tastefully completed on the warm Adriatic shores just before the symposium began and graciously presented to the Centre by the Community of Trieste.

The symposium was made financially possible by generous grants from: Ford Foundation; Italian Ministry of Education; Region of Friuli-Venezia Giulia; Consorzio per l'Incremento degli studi e delle ricerche degli Istituti di Fisica dell'Università di Trieste; UNESCO; Academy of Sciences of the USSR; Royal Society (London); US Atomic Energy Commission; National Science Foundation (US); and Gulf General Atomic Inc.; as well as a personal donation from Mr. Cecil H. Green of Texas Instruments Inc. We received warm hospitality from His Highness Prince Raimondo Torre e Tasso and from the City of Trieste — this and an infinity of other courtesies we owe to the tireless organization of Professor Paolo Budini.

An international Organizing Committee, chaired till his tragic death by Professor J.R. Oppenheimer, was responsible for drawing up the programme — it was shaped finally through the dedicated efforts of the monitors, Professors G. Burbidge, D. Pines and M. Rosenbluth, of the Scientific Secretary of the symposium, Professor L. Fonda, and, particularly, of Professor R.E. Marshak. The symposium would never have been held if, at all stages, we had not had the assured warm support of Dr. S. Eklund.

The members of the Organizing Committee were as follows: Professors H. Alfvén, V. Ambartsumyan, N.N. Bogolyubov, Aage Bohr, H. Bondi, S. Brenner, A. de-Shalit, J.F. Dyson, M. Gell-Mann, M. Goldberger, I.M. Khalatnikov, R.E. Marshak, A. Matveyev, D. Pines, I. Prigogine, R.Z. Sagdeev, M. Sandoval Vallarte, V. Soloviev, J. Tiomno, L. Van Hove, and V.F. Weisskopf.

Abdus Salam

## CONTENTS OF VOL. I

### THEORY OF CONDENSED MATTER AND RELATED PROBLEMS

An introduction to quantum liquids: Fermi liquids and collective modes .....	3
D. Pines	
Phase transitions and critical phenomena .....	19
Michael E. Fisher	
Macroscopic coherence and superfluidity .....	47
P. W. Anderson	
Microscopic theory of superconductivity .....	55
J. R. Schrieffer	
Theory of Bose-Fermi quantum liquids .....	71
I. M. Khalatnikov	
Paramagnetism in Fermi liquids .....	87
S. Doniach	
Transport coefficients of a normal Fermi liquid .....	93
C. Pethick	
Magnetic impurities in non-magnetic metals .....	97
A. Abrikosov	
Solid-state problems for particle physicists .....	123
P. C. Martin	
Field theory of phase transitions .....	129
R. A. Ferrell	
Localized magnetic moments in metals .....	157
H. Suhl	
Survey of the one-dimensional many-body problem and two-dimensional ferro-electric models .....	163
E. H. Lieb	
Three examples of one-dimensional systems .....	177
E. Montroll	
Some applications of path integrals and diagrammatic methods to chemical physics .....	195
P. G. de Gennes	

### PLASMA PHYSICS, TURBULENCE, QUANTUM OPTICS AND STATISTICAL MECHANICS

Plasma physics: general survey .....	205
M. N. Rosenbluth	
Non-linear plasma physics .....	221
T. Dupree	
Controlled thermonuclear research .....	237
W. B. Thompson	

The problem of plasma confinement .....	249
B. Coppi	
Survey of the theory of turbulence .....	257
J.B. Keller	
Some remarks on turbulence .....	273
E. Montroll	
Quantum optics, or quantum electronics .....	295
C.H. Townes	
Quantum statistical mechanics of systems with an infinite number of degrees of freedom .....	315
I. Prigogine	

## ASTROPHYSICS, QUASARS AND PULSARS

Introduction and topics in theoretical astronomy .....	335
E.E. Salpeter	
Survey of current problems in extragalactic astronomy .....	347
E.M. Burbidge	
Solar neutrino astronomy .....	359
W.A. Fowler	
Discrete extrasolar X-ray sources .....	371
B. Rossi	
Diffuse radiation in the high-energy region .....	393
P. Morrison	
Radio studies of galactic structure .....	401
B.F. Burke	
The magnetic field of the galaxy .....	417
F.G. Smith	
The peculiar A-type stars .....	423
R.P. Kraft	
Stellar populations and the evolution of the galaxy .....	449
R.P. Kraft	
Radio galaxies and quasars .....	459
F.G. Smith	
Observed properties of quasi-stellar objects .....	467
M. Schmidt	

### *Pulsars*

Introduction: opening talk on pulsars .....	471
J.G. Bolton	
Attempts at optical detection of pulsars .....	475
A.G.W. Cameron	
The nature of pulsars: survey of present views .....	477
T. Gold	

## GRAVITATIONAL THEORY AND COSMOLOGY

### *Cosmology*

Theoretical implications of the known facts about gravitation .....	485
W. Thirring	

Recent developments in observational cosmology .....	489
D.W. Sciama	
Observational cosmology: optical wavelengths .....	497
E.M. Burbidge	
Remarks on gravitation and cosmology .....	507
R.H. Dicke	
 <i>Gravitational Theory</i> 	
Remarks on the general principles of Einstein's gravitation theory ..	511
V.A. Fock	
General relativity: survey and experimental tests .....	515
R.H. Dicke	
General relativity theory and experiments .....	533
J. Weber	
The quantization of the gravitational field .....	539
P.A.M. Dirac	
 <i>Special Problems</i> 	
On gravitational collapse .....	545
R. Penrose	
Singularity in the general solution of the gravitational equations (short contribution) .....	557
E.M. Lifshitz	
Remarks on gravitational radiation .....	559
S. Weinberg	
Hamiltonian dynamics and positive energy in general relativity .....	563
S. Deser	
 Monitors .....	571
Author Index .....	572

## ***EDITORIAL NOTE***

*The papers and discussions incorporated in the proceedings published by the International Atomic Energy Agency are edited by the Agency's editorial staff to the extent considered necessary for the reader's assistance. The views expressed and the general style adopted remain, however, the responsibility of the named authors or participants.*

*For the sake of speed of publication the present Proceedings have been printed by composition typing and photo-offset lithography. Within the limitations imposed by this method, every effort has been made to maintain a high editorial standard; in particular, the units and symbols employed are to the fullest practicable extent those standardized or recommended by the competent international scientific bodies.*

*The affiliations of authors are those given at the time of nomination.*

*The use in these Proceedings of particular designations of countries or territories does not imply any judgement by the Agency as to the legal status of such countries or territories, of their authorities and institutions or of the delimitation of their boundaries.*

*The mention of specific companies or of their products or brand-names does not imply any endorsement or recommendation on the part of the International Atomic Energy Agency.*

**THEORY OF CONDENSED MATTER  
AND RELATED PROBLEMS**



# AN INTRODUCTION TO QUANTUM LIQUIDS: FERMI LIQUIDS AND COLLECTIVE MODES\*

D. PINES

Department of Physics, University of Illinois,  
Urbana, Ill., United States of America

## Abstract

AN INTRODUCTION TO QUANTUM LIQUIDS: FERMI LIQUIDS AND COLLECTIVE MODES. 1. Brief remarks on problems and methods in the quantum theory of condensed matter; 2. Quantum liquids; 3. The normal Fermi liquid; 4. Collective modes; 5. Conclusion.

## 1. BRIEF REMARKS ON PROBLEMS AND METHODS IN THE QUANTUM THEORY OF CONDENSED MATTER

As monitor, in this Symposium, for solid-state and condensed-matter physics, I was aware of the difficulties involved when I wrote to those invited to give the main lectures in this field: "..... It is important that each lecture should be readily intelligible to the non-specialist, and that it spell out what we have learned, what we are doing, and what are some of the important unanswered questions in the field under discussion. Wherever possible it would also seem useful to indicate relationships between the various sub-fields; the basic experimental facts, as well as the deep theoretical problems, should be discussed."

I begin my own paper by saying a few words about the organization of the talks in solid-state and condensed-matter physics. It has been nearly twelve years since the theoretical physics community last met together as a group (in Seattle, in 1956). It seems useful, therefore, to consider this symposium from the point of view of a participant who was at Seattle, and had at that time a modern overall view of theoretical physics. During the past decade our mythical theorist has been part of the increasing specialization characteristic of all of physics; it seems a fair assumption that by now he has only a vague notion of work in sub-fields outside his own. Given the impressive growth in solid-state physics in the past decade and the small amount of time available to us at this Symposium, we can only tell him a small fraction of what has transpired; the question is "which fraction?".

There have been, in general, two main lines of development in solid-state theory:

(1) Problems in which the periodicity of the solid as manifested via the potential field of the ions, plays no essential role. These include most of the applications of many-body theory to solid-state and condensed-matter physics. Significant progress has been made on the development of a theory of quantum liquids, on the theory of second-order phase transitions, investigations of magnetic impurities in non-magnetic metals, studies of solid-state plasma, etc.

---

\* This work was sponsored in part by the Air Force Office of Scientific Research, Office of Aerospace Research, United States Air Force, under AFOSR contract number AF-AFOSR-328-67.

(2) Problems in which periodicity plays an essential role. Among the most significant developments we may mention the accurate determination of the Fermi surface of electrons in metals, phonon dynamics, pseudo-potential theory, better theories of the transition metals, studies of localized excitations in solids, etc.

After discussing with a number of colleagues which of these many important developments might best be communicated to the non-solid-state theorist, it was decided to concentrate on those problems which are most closely related to those which arise in other fields of physics, especially nuclear theory, plasma theory, and astrophysics, and which may involve as well techniques not dissimilar to those used by the particle theorists. Accordingly, the papers will be largely on the problems cited in the first category above, that is, on quantum liquids, on phase transitions, and on other many-body problems which may have applicability beyond the immediate purview of the solid-state physicist.

I myself shall try to give a brief introduction and review of the work on quantum liquids, then proceed to discuss in somewhat more detail Landau's theory of normal Fermi liquids, and the description of collective modes in many-particle systems. In these Proceedings you will also hear about two of the most active fields of investigation in many-body theory: second-order phase transitions, from Fisher and Ferrell, and the behaviour of magnetic impurities in non-magnetic metals, the "Kondo" problem, which will be surveyed by Abrikosov. In addition Doniach and Pethick describe some "frontier" work on Fermi liquids, to wit the behaviour of nearly ferromagnetic Fermi liquids.

## 2. QUANTUM LIQUIDS

Quantum liquids are spatially homogeneous systems of strongly interacting particles at temperatures sufficiently low that effects of quantum statistics (i.e. degeneracy) are important. Only three genuine quantum liquids are found in nature:  $^3\text{He}$ , a Fermi liquid,  $^4\text{He}$ , a Bose liquid, and dilute solutions (< 6%) of  $^3\text{He}$  in  $^4\text{He}$ , a low density Fermi liquid. In both  $^3\text{He}$  and  $^4\text{He}$  the strong interaction between particles makes an accurate microscopic theory difficult, if not impossible. The dilute solutions of  $^3\text{He}$  in  $^4\text{He}$  are far more tractable from a microscopic point of view both because of the low concentration of  $^3\text{He}$  atoms and because the effective interaction between  $^3\text{He}$  atoms is, in fact, quite weak, since the  $^3\text{He}$  atoms are isotopic impurities in the  $^4\text{He}$  background.

There are, in addition, the honorary Fermi liquids: electrons in metals, semiconductors or semimetals under circumstances that the influence of the ionic field is weak. Electrons in metals represent another system in which the effects of particle interaction are strong. This is perhaps not surprising, since in the formation of a metal there is a close balance between the influence of the kinetic energy and that of the potential energy. Electrons in metals thus form an electron liquid, rather than an electron gas.

We should further distinguish between normal and superfluid Fermi or Bose liquids. A normal Bose or Fermi liquid is a degenerate system whose properties are not drastically altered by particle interaction in the sense that it retains the essential properties of a gas of free particles. Thus, a normal Fermi liquid has a sharp Fermi surface, specific heat proportional

to T, etc. Superfluids are best characterized by the existence of a macroscopically occupied single quantum state. This condensate is responsible for resistance free motion and, through its general variation in space and time, for a host of other fascinating phenomena in superfluids. Thus far superconductors are the only examples of superfluid Fermi liquids, though there remains the possibility of <sup>3</sup>He becoming superfluid at sufficiently low temperatures, while the <sup>3</sup>He portion of <sup>3</sup>He-<sup>4</sup>He solutions is predicted to become superfluid at a temperature of a few microdegrees. The only Bose superfluid is liquid He II, that is, <sup>4</sup>He below the  $\lambda$ -point (2.18°K). As reported by Anderson, liquid He II and superconductors possess many features in common, features which are related to macroscopic condensate motion.

Thanks to the work of very many people, substantial advances in the past decade have been made both in the development of fundamental concepts and specific techniques for dealing with quantum liquids. We have today a unified point of view and a language appropriate for the description of these many-particle systems. It is at first sight curious that despite the sizable interaction between particles, despite the fact that we are dealing with a quantum system, quantum liquids are much better understood than their classical counterparts. The explanation lies in the fact that under suitable circumstances one may get a complete description of the low-lying excited states of the system in terms of the concept of elementary excitations. At sufficiently low temperatures there are only a few such present; they interact weakly with each other and hence are relatively long-lived.

There are, in general, two kinds of elementary excitations: quasi-particles and collective modes; we now understand the relationship between these, and how both are affected by particle interaction. We have learned as well how to give a precise description of the linear response of a many-body system to weak external probes, and so possess a far better understanding of transport phenomena. Among the techniques which have been useful in studying many-particle systems, special mention should be made of Green's functions and Feynman diagrams, which are as indispensable for the many-body theorist as for the particle physicist. Sum rules, too, play an important role; these plus relatively simple assumptions concerning the relevant spectral densities, provide considerable physical insight into the behaviour of both normal and superfluid systems.

Of the work in the past twelve years on the theory of quantum liquids, four advances may be singled out for especial mention:

- (1) The development of a microscopic theory of superconductivity, due to Bardeen, Cooper, and Schrieffer.
- (2) The Landau theory of the macroscopic behaviour of normal Fermi liquids.
- (3) The great progress in the description of the macroscopic motion of superfluids, that is, condensate motion and its consequences (the Josephson effect, etc.).
- (4) The solution of various model problems — electrons at high densities and low temperatures, the dilute Bose gas, various one-dimensional systems.

In all these areas the theorists have been so successful as to bring about a kind of technological unemployment; in the view of many people none of the above-cited areas can properly be regarded as frontier in work on many-

particle systems. Still it would seem important that these advances be described in these Proceedings. More specifically, I myself shall now discuss the Landau theory, Schrieffer will discuss the microscopic theory of superconductivity, Anderson will describe our present understanding of the macroscopic behaviour of superfluids, and Lieb will report on solutions to one-dimensional problems.

### 3. THE NORMAL FERMI LIQUID

The Landau theory of the normal Fermi liquid<sup>1</sup> is of considerable interest in itself; it serves, as well, as a useful vehicle for introducing a number of concepts common to all quantum liquids.

Consider first a gas of non-interacting fermions of mass m. A particle of momentum  $\vec{p}$  has energy

$$\epsilon_p^0 = \frac{\vec{p}^2}{2m} \quad (1)$$

(spin indices will be suppressed unless needed for clarity). The system energy takes the form

$$E = \sum_{\vec{p}} n_{\vec{p}} \frac{\vec{p}^2}{2m} \quad (2)$$

where  $n_{\vec{p}}$ , the distribution function which characterizes the eigenstates of the system, is equal to 1 if the state  $\vec{p}$  is occupied, to zero otherwise. In the ground state of the system, all states of momentum less than  $p_F$  are occupied, those greater than  $p_F$  are empty. Excited states of the system are then easily specified with reference to the ground state, via the departure of the distribution function from its ground state value, that is

$$\delta n_{\vec{p}} = n_{\vec{p}} - n_{\vec{p}}^0 \quad (3)$$

The excitation energy is thus

$$E - E_0 = \sum_{\vec{p}} \frac{\vec{p}^2}{2m} \delta n_{\vec{p}} \quad (4)$$

Assume, now, that the interaction between particles is switched on adiabatically; the eigenstates of the ideal system are transformed into states of the real, interacting system; for a normal Fermi liquid the real ground state is adiabatically generated starting from some eigenstate  $n_{\vec{p}}^0$  of the ideal system. If we now add a particle of momentum  $\vec{p}$  to the ideal ground state distribution, and again turn on the interaction adiabatically, we generate an excited state of the real system, which we may identify as the real ground state plus a quasi-particle of momentum  $\vec{p}$ . The quasi-

---

<sup>1</sup> The exposition follows closely that given in Ref. [1] to which the reader is referred for references and further details.

particle corresponds to the bare particle of the non-interacting system plus the surrounding particle distortion brought about by the particle interaction; it is a bare particle dressed with a self-energy cloud; more formally, its energy may be found from the pole of the single particle Green's function for the system. We can further specify the eigenstates of the real system by the same functions,  $n_{\vec{p}}$ , used for the ideal system, but which now represent the distribution of quasi-particles.

For the interacting Fermi liquid, the energy of the system does not take the simple form, Eq.(2), but rather must be expressed in functional form,  $E[n_{\vec{p}}]$ . If we assume that the degree of excitation of the system is small, in other words, that

$$\alpha = \sum_{\vec{p}} \delta n_{\vec{p}} / N \ll 1 \quad (5)$$

then we may write

$$E[n_{\vec{p}}] = E_0 + \sum_{\vec{p}} \epsilon_{\vec{p}} \delta n_{\vec{p}} + O(\delta n^2) \quad (6)$$

where  $\epsilon_{\vec{p}}$ , the first functional derivative of  $E$ , represents the energy of a quasi-particle.

In the same way the free energy may be expanded as

$$F - F_0 = \sum_{\vec{p}} (\epsilon_{\vec{p}} - \mu) \delta n_{\vec{p}} + \frac{1}{2} \sum_{\vec{p}} f_{\vec{p}\vec{p}'} \delta n_{\vec{p}} \delta n_{\vec{p}'} \quad (7)$$

where  $\mu$  is the chemical potential and the second term on the right-hand side represents the effect of interaction between the quasi-particles. This latter term is the new feature of the Landau theory.  $f_{\vec{p}\vec{p}'}$ , which has the dimensions of an energy, describes the effective interaction between quasi-particles of momentum  $\vec{p}$  and  $\vec{p}'$ . We note that the energy required to create a quasi-particle of momentum  $\vec{p}'$  in the presence of a distribution of excited quasi-particles  $\delta n_{\vec{p}}$  is

$$\tilde{\epsilon}_{\vec{p}} = \epsilon_{\vec{p}} + \sum_{\vec{p}'} f_{\vec{p}\vec{p}'} \delta n_{\vec{p}'} \quad (8)$$

If further, we assume the system is slightly inhomogeneous so that  $\delta n_{\vec{p}}$  is a function of  $r$ , then we can define a local quasi-particle energy as

$$\tilde{\epsilon}_{\vec{p}}(r) = \epsilon_{\vec{p}} + \sum_{\vec{p}'} f_{\vec{p}\vec{p}'} \delta n_{\vec{p}'}(r) \quad (9)$$

The gradient of this energy

$$\vec{\nabla}_r \tilde{\epsilon}_{\vec{p}}(r) = \vec{\nabla}_r \left\{ \sum_{\vec{p}'} f_{\vec{p}\vec{p}'} \delta n_{\vec{p}'} \right\} \quad (10)$$

is the average force exerted by the surrounding medium on a quasi-particle  $p$ .

Given these definitions it is straightforward to find the distribution function of quasi-particles at temperature T

$$n_{\vec{p}}^0(T, \mu) = \frac{1}{1 + \exp(\frac{\epsilon_{\vec{p}} - \mu}{kT})} \quad (11)$$

and to calculate quantities such as the specific heat, spin susceptibility, and compressibility in terms of  $\epsilon_{\vec{p}}$  and the interaction between the quasi-particles. One finds, for example, that at very low temperatures, the specific heat is linearly proportional to the temperature, and that the ratio of the specific heat for the real liquid,  $C_V$ , to that for the free particle system,  $C_V^0$ , is simply

$$\frac{C_V}{C_V^0} = \frac{m^*}{m}$$

where  $m^*$  is the effective mass of the quasi-particle, and is given by

$$(\vec{\nabla}_{\vec{p}} \epsilon_{\vec{p}})_{p=p_F} = p_F / m^*$$

A transport equation for quasi-particles may be obtained by regarding the quasi-particles as independent, described by a classical Hamiltonian,  $\epsilon_{\vec{p}}(\vec{r}, t)$ . It takes the form

$$\frac{\partial}{\partial t} \delta n_{\vec{p}}(\vec{r}, t) + \vec{\nabla}_{\vec{p}} \delta n_{\vec{p}}(\vec{r}, t) \cdot \vec{\nabla}_{\vec{p}} - \vec{\nabla}_{\vec{p}} n_{\vec{p}}^0 \sum_{\vec{p}'} f_{\vec{p}\vec{p}'} \vec{\nabla}_{\vec{p}'} \delta n_{\vec{p}'}(\vec{r}, t) + \vec{F}_{\vec{p}} \cdot \vec{\nabla}_{\vec{p}} n_{\vec{p}}(\vec{r}, t) = I(\delta n_{\vec{p}}) \quad (12)$$

where  $\vec{F}_{\vec{p}}$  represents the external force felt by a quasi-particle, while  $I(\delta n_{\vec{p}})$  is a collision term which measures the rate of change of  $\delta n_{\vec{p}}$  due to collisions. The novel feature of this equation is the third term on the left-hand side which represents the interaction between excited particles and those in the ground state distribution; according to Landau this interaction may be reduced to an average macroscopic force, which is the gradient of the quasi-particle energy, as in Eq. (10).

A few words about the validity of the transport equation (12):

1. If  $\delta n_{\vec{p}}$  is regarded as the probability for finding a quasi-particle, quasi-hole pair, the equation is valid for frequencies  $\omega$ , and wave-vectors  $\vec{q}$ , such that

$$\begin{aligned} qv_F &\ll \mu \\ \omega &\ll \mu \end{aligned} \quad (13)$$

that is, for macroscopic fluctuations.

2. A second requirement is that the degree of excitation of the system be small, as spelled out in Eq. (5); this means that Eq. (12) is only valid at temperatures T such that

$$\kappa T \ll \mu \quad (14)$$

3. A little thought shows that the above restrictions are equivalent to taking into account all effects associated with single quasi-particle quasi-hole pairs; indeed the transport equation (12) (minus the external force field and collision terms) corresponds to the Bethe-Salpeter equation, and represents the multiple scattering of a particle-hole pair. It may be derived rigorously using field-theoretic methods.

It is relatively straightforward to solve this transport equation in either of two limits, the hydrodynamic or the collisionless regime. At any temperature  $T$ , there will be present a certain number of thermally-excited quasi-particles and quasi-holes; these interact with each other; the influence of those interactions is measured by a characteristic time  $\tau$  which represents the life-time of a quasi-particle against decay or the mean time between collisions. Suppose one considers a probe at some frequency  $\omega$ ; if, then

$$\omega\tau \ll 1 \quad (15)$$

the particles have ample time to scatter against each other during the period of the perturbation. The system can then be said to be in local thermodynamic equilibrium; one is in a hydrodynamic regime. If one considers, for example, solutions of (12) which correspond to density fluctuations, one finds the usual first sound mode, subject to the usual viscous damping, proportional to the coefficient of viscosity for the Fermi liquid.

Suppose, now, that  $\omega$  is large, and such that

$$\omega\tau \gg 1 \quad (16)$$

Under these circumstances the excitations don't have time to interact; there is no local thermodynamic equilibrium, and one is in a collisionless regime. Can one then have a sound mode? The answer is yes, provided the interaction between the quasi-particles,  $f_{pp}$ , is sufficiently repulsive. It is a quite different kind of sound mode in that the restoring force which brings it about is the averaged self-consistent field of the particles, in contrast to the collisions between quasi-particles responsible for first sound. Indeed, quasi-particle collisions act to damp this sound mode, which is known as zero sound.

Zero sound is but one of the many collective modes possible in a Fermi liquid; depending on the sign and size of the interaction,  $f_{pp}$ , one may have longitudinal or transverse spin waves, etc. In all cases the collective modes are found by solving the transport equation in the absence of external forces and in the collisionless limit. I shall return shortly to a more general discussion of these collective modes.

Thus far the Landau theory has been applied to liquid  $^3\text{He}$ , and, with the relevant extension to take into account electron-phonon interactions and the influences of the periodic ionic potential, to electrons in simple metals (the alkalis, aluminium, etc.). We here consider briefly the application of the theory to experiment in the case of liquid  $^3\text{He}$ .

For  $^3\text{He}$ , the region of temperatures at which the Landau theory begins to be applicable is below  $0.1^\circ\text{K}$ , and experimental measurements are now carried out in the millidegree region [2]. We discuss the following:

- (i) Macroscopic equilibrium properties
- (ii) Zero and first sound
- (iii) Transport coefficients.

First, according to theory, the spin susceptibility and compressibility should be independent of temperature, and are found to be so. The specific heat shows the familiar linear temperature variation; there is, in addition, a logarithmic term in the specific heat

$$CT^3 \ln T$$

whose existence was quite unexpected. Anderson [3] first called attention to the possible importance of a logarithmic term in fitting the experiments of Wheatley and his collaborators; Doniach and Engelsberg [4] and Berk and Schrieffer [5] showed on the basis of a model calculation that such a temperature dependence was plausible for an almost ferromagnetic liquid such as  ${}^3\text{He}$ . They ascribed its appearance to the presence of damped spin density collective modes, or paramagnons, about which Doniach reports in these Proceedings. It is now known that such a finite temperature correction is characteristic of all Fermi liquids [6].

From measurements of the spin susceptibility, compressibility, and the coefficient of the linear term in the specific heat, one learns about the strength of the interaction between the quasi-particles in  ${}^3\text{He}$ . More specifically, we note that for sufficiently low temperatures, it is legitimate to regard all quasi-particles as located on the Fermi surface; for an isotropic system the interaction  $f_{\overline{p}\overline{p}}$ , then depends only on the angle  $\xi$  between the quasi-particles, and on their relative spin orientations. On introducing the spin symmetric  $f_{\overline{p}\overline{p}}^s$ , and spin anti-symmetric  $f_{\overline{p}\overline{p}}^a$ , parts of the interaction, one may express the interaction in reduced dimensionless units,  $F_\ell^{s(a)}$ , according to

$$f_{\overline{p}\overline{p}}^{s(a)} = \sum_{\ell=0}^{\infty} [F_\ell^{s(a)} / \nu(0)] P_\ell(\cos\xi) \quad (17)$$

where  $\nu(0)$  is the density of states per unit energy.

The usual procedure is to take into account only the first few coefficients in this Legendre polynomial expansion.

The coefficients  $F_0^s$ ,  $F_0^a$ ,  $F_1^s$  are obtained from experimental measurements of the compressibility, spin susceptibility, and the coefficients of the linear term in the specific heat, and are given in Table I for two different values of the pressure. The large values of these parameters are a reflection of the strong interaction between the quasi-particles in  ${}^3\text{He}$ . With these, one can then attempt to explain experimental measurements on zero sound and transport in  ${}^3\text{He}$ .

TABLE I. LANDAU COEFFICIENTS FOR  ${}^3\text{He}$  \*

Pressure (atm)	$F_0^s$	$F_0^a$	$F_1^s$	$F_1^a$
0.28	10.5	-0.67	6.0	$\sim -0.5$
27	73.2	-0.73	13.8	$\sim -0.6$

\* The values of  $F_0^a$ ,  $F_0^s$ , and  $F_1^s$ , and  $F_1^a$  at high pressure are taken from Wheatley [2], while those for  $F_1^a$  at low pressure are due to Dy and Pethick [11].

The first evidence for zero sound came somewhat indirectly, from the measurements of Keen, Matthews, and Wilks [7] on the acoustic impedance of  ${}^3\text{He}$  as a function of temperature. Subsequently, Abel, Anderson and Wheatley [8] were able to observe it directly in a study of ultrasonic propagation in the millidegree range; they found a velocity of 194.4 m/s, some 3.5% larger than the first sound velocity of 187.2 m/s. Measurements of the zero sound velocity provide a check on the Landau theory, since it depends sensitively on the parameters  $F_0^s$  and  $F_1^s$ . Excellent agreement between the theory of Abrikosov and Khalatnikov [9] and experiment is found; agreement with their theory is also good for the maximum in the damping of the ultrasonic waves.

Measurements of transport properties (thermal conductivity, viscosity, spin diffusion, etc.) provide a further check on the Landau theory. More specifically, Dy and Pethick [10] have shown that when one assumes  $F_\ell^{s(a)} = 0$  for  $\ell \geq 2$ , and uses exact solutions of the transport equation, together with an interpolation formula for the scattering amplitude which satisfies the Pauli principle, good agreement between theory and experiment exists for the thermal conductivity, spin diffusion, and viscosity, in the low temperature limit, as shown in Table II. For  ${}^3\text{He}$ , because  $F_0^a$  is large and negative, there are important finite temperature corrections to the thermal conductivity, and spin diffusion coefficient; one has [11, 12]

$$\frac{1}{KT} - \left( \frac{1}{KT} \right)_{T=0} = AT \quad (18)$$

$$\frac{1}{DT} - \left( \frac{1}{DT} \right)_{T=0} = BT \quad (19)$$

where K and D are the thermal conductivity and spin diffusivity, and an expression for the coefficients A and B may be given in terms of Landau parameters; expressions of this form were first obtained from an approximate "paramagnon" calculation [13]. Assuming  $F_\ell^{s(a)} = 0$  for  $\ell \geq 2$ , Dy and Pethick use the thermal conductivity experiments of Wheatley and his collaborators [which show the temperature variation (18)] to determine  $F_1^a \approx -0.5$ ; this

TABLE II. COMPARISON OF THEORY AND EXPERIMENT FOR TRANSPORT PROPERTIES OF  ${}^3\text{He}$  [10]

Pressure (atm)	KT (erg/cm s)	$DT^2$ ( $\times 10^{-6}\text{cm}^2 \cdot \text{K}^2/\text{s}$ )	$\eta T^2$ ( $\times 10^{-6}\text{poise} \cdot \text{K}^2$ )
0.28			
Experiment	35	1.4	$\sim 2$
Theory	33	1.6	1.6
27			
Experiment	< 12	0.17	...
Theory	8.6	0.16	0.54

result is likewise consistent with the experimental measurements of the coefficient B in (19). To sum up, the Landau theory is thus seen to provide a consistent account of both equilibrium and non-equilibrium properties of  $^3\text{He}$ .

To conclude this discussion of the normal Fermi liquids, let me once again emphasize that the Landau theory is the sum of all the exact statements one can make about the low-lying excited states of the system, and its response to macroscopic (low  $\vec{q}$ , low  $\omega$ ) probes, which therefore involves only single quasi-particle quasi-hole pairs and collective modes. Good agreement with experiment is found for those regimes of temperature, wavenumber and frequency that one expects it to apply, both for  $^3\text{He}$  and for electrons in simple metals<sup>2</sup>. On the other hand, it has its limitations; it is not a microscopic theory, in that, for example, it does not specify the system response to a microscopic (high frequency or short wavelength) external probe.

#### 4. COLLECTIVE MODES

As I mentioned earlier, in the collisionless regime, collective modes in many-particle systems are brought about by the average self-consistent field of the particles acting in concert; this may be regarded as a polarization or molecular field which provides the restoring force responsible for the mode in question. For the purpose of illustration, I shall consider a particular class of collective modes, those which involve a fluctuation in the system density; examples are the plasma oscillations of electron systems and the zero sound mode of  $^3\text{He}$ . However, essentially everything I have to say can be transposed to the case of spin waves, or various kinds of transverse collective modes.

Information about the density fluctuation excitation spectrum may be found in the density-density response function. This correlation function measures the response of the system to an external scalar probe coupled directly to the system density, according to

$$\rho(\vec{r}, t)\varphi_{\text{ext}}(\vec{r}, t)$$

where  $\varphi_{\text{ext}}$  represents the external scalar field, and  $\rho$  is the system density. Examples of such scalar probes are neutron or electron beams, or an externally imposed sound wave. If now one Fourier-analyses everything in sight, the coupling takes the form

$$\rho^+(\vec{q}, \omega)\varphi_{\text{ext}}(\vec{q}, \omega) + \text{c. c.}$$

where  $\rho^+(\vec{q}, \omega)$  and  $\varphi_{\text{ext}}(\vec{q}, \omega)$  are the Fourier transforms in space and time of the density and probe respectively. The probe acts to induce a density fluctuation in the system; if it is weak, the system will respond linearly, so that the induced density fluctuation,  $\langle\rho(\vec{q}, \omega)\rangle$ , is proportional to the external field, according to

$$\langle\rho(\vec{q}, \omega)\rangle = \chi(\vec{q}, \omega)\varphi_{\text{ext}}(\vec{q}, \omega) \quad (20)$$

---

<sup>2</sup> For a review of work on the alkali metals, see Ref. [14].

$\chi(\vec{q}, \omega)$  is the density-density correlation function; an exact expression for  $\chi$  in terms of the exact eigenstates of the system is easily obtained using perturbation theory.

A collective mode appears as a pole in  $\chi$ , or, what is equivalent, as a peak in  $\chi''$ , the imaginary part of  $\chi(\vec{q}, \omega)$ , which is the spectral density for the density fluctuations. Where that peak is displaced from that part of  $\chi''$  which contains the major strength associated with the single-particle excitations, the pit will be sharp and represent a well-defined (or weakly damped) collective mode; on the other hand, where the "collective" peak falls at an energy which substantially overlaps the single-particle spectrum, the collective mode is so strongly damped as to be virtually indistinguishable from the single-particle modes.

There are many ways of calculating  $\chi(\vec{q}, \omega)$  — diagrams, equations of motion, Green's function methods, to name but a few. I shall not attempt to summarize any of these, but rather wish to consider the structure of  $\chi(\vec{q}, \omega)$  from a quite general point of view, one which makes evident circumstances under which one may expect to find collective modes for any many-particle system. The approach is one I developed a few years ago to treat zero sound in liquid helium [15]; it is, however, easily generalized to treat other kinds of collective modes.

The basic philosophy is simple; if one is interested in a possible collective mode, it is useful to begin by treating the force responsible for it on a special basis. In the present case, the force is longitudinal, and hence is derivable from a scalar polarization potential,  $\phi_{\text{pol}}(\vec{r}t)$ . The latter must, in turn, be related to the averaged self-consistent field associated with the density fluctuations,  $\langle \rho(\vec{r}, t) \rangle$ . In the case of an electron liquid, this relation is well known; when one Fourier-analyses both fields, it takes the simple form

$$\phi_{\text{pol}}(\vec{q}, \omega) = \frac{4\pi e^2}{q^2} \langle \rho(\vec{q}, \omega) \rangle \quad (21)$$

where  $4\pi e^2/q^2$  is the Fourier transform of the Coulomb potential, and the superscript indicates that we are referring to an electron system. It is appealing, therefore, to define the polarization potential for a neutral system by analogy to Eq. (21)

$$\phi_{\text{pol}}(\vec{q}, \omega) = f_q \langle \rho(\vec{q}, \omega) \rangle \quad (22)$$

The phenomenological quantity,  $f_q$ , thus measures the strength of the self-consistent density field; it has the dimensions of an energy, and in the limit of wavelengths greater than the interparticle spacing may be expected to reduce to a constant; thus

$$f_q = f_0 \quad (qr_0 \lesssim 1) \quad (23)$$

In the calculation of  $\chi(\vec{q}, \omega)$  it is straightforward to take explicit account of the presence of this polarization potential; to do so we again take a hint from the electron system. There it is convenient to recognize the primacy of polarization effects by considering the response of the system to the sum of the external field,  $\phi_{\text{ext}}$ , and the internal polarization field to which it gives

rise; one thereby defines a screened response function,  $\chi_{sc}(\vec{q}, \omega)$  according to [16]

$$\langle \rho^-(\vec{q}, \omega) \rangle = \chi_{sc}^-(\vec{q}, \omega) \{ \varphi_{ext}^-(\vec{q}, \omega) + \varphi_{pol}^-(\vec{q}, \omega) \} \quad (24)$$

$\chi_{sc}^-(\vec{q}, \omega)$  is so called because it measures the response to a screened external field (since the polarization field acts to screen out the influence of the external field). Let me remind you that  $\chi_{sc}$  is not exactly unfamiliar, whereas  $\chi$  measures the response to the displacement field,  $\vec{D}$ .  $\chi_{sc}$  measures the response to the effective, screened field,  $\vec{\epsilon}$ ; it is this latter response one considers in all elementary treatments of the subject, and  $\chi_{sc}$  is proportional to the conductivity as might be expected.

If we now write down the neutral analogue of (24), and make use of (22), we obtain the fundamental relation between  $\chi$  and  $\chi_{sc}$

$$\chi = \frac{\chi_{sc}}{1 - f_q \chi_{sc}} \quad (25)$$

In physical terms,  $\chi_{sc}$  represents that part of  $\chi$  which remains once all processes described by the polarization potential are taken into account; for the case of the Coulomb interaction it is the sum of all irreducible polarization diagrams. For neutral systems, a microscopic specification of  $\chi_{sc}$  is rather more complicated; we still shall refer to it as the irreducible polarization part of  $\chi$ .

The relation (25) is useful because one can make some rather general statements about  $\chi_{sc}$  and  $f_q$ , and from these draw significant conclusions concerning possible collective modes. From the explicit expression for  $\chi$ , in terms of system eigenstates, and the  $f$  sum rule [17], it can be shown that

$$\lim_{\omega \rightarrow \infty} \chi_{sc} = \frac{Nq^2}{m\omega^2} \quad (26)$$

$\chi_{sc}$  will take its asymptotic value, (26), as soon as  $\omega$  is large compared to any of the characteristic frequencies which appear in its spectral representation. If we denote the largest of such frequencies by  $\omega_{sc}$ , we can write

$$\chi(q, \omega) = \frac{Nq^2/m}{(\omega + i\eta)^2 - Nq^2 f_q/m} \quad (\omega > \omega_{sc}) \quad (27)$$

(on introducing the positive infinitesimal quantity  $\eta$  to allow for the appropriate retarded boundary conditions on  $\chi$ ).

This is our desired result; from it we conclude that there exists a pole in  $\chi$ , and hence a collective mode at

$$\omega_c = \left( \frac{Nq^2 f_q}{m} \right)^{\frac{1}{2}} \quad (28)$$

provided  $f_q$  is sufficiently strong, that is, provided

$$\omega_c > \omega_{sc} \quad (29)$$

Let us now pass to some specific examples. Again, consider the electron liquid; for this system

$$\omega_c = \sqrt{\frac{4\pi ne^2}{m}} = \omega_p \quad (30)$$

and is the usual frequency for plasma oscillation; on the other hand the maximum frequency for which  $\chi_{sc}$  has appreciable spectral density is

$$\omega_{sc} \approx qv_F \quad (\text{for a quantum plasma}) \quad (31a)$$

$$\omega_{sc} \approx q\sqrt{\frac{kT}{m}} \quad (\text{for a classical plasma}) \quad (31b)$$

$$\omega_{sc} \approx \Delta \quad (\text{for a superconductor}) \quad (31c)$$

where  $v_F$  is the Fermi velocity and  $\Delta$  is the energy gap. It follows at once that in the long wavelength limit plasmons of energy  $\omega_p$  are a well-defined collective mode in all three systems, since under these circumstances  $\chi_{sc}$  has taken on its asymptotic value (26), at frequencies far below  $\omega_p$ . Note that if the interaction is sufficiently strong, the collective mode exists at essentially the same frequency, no matter what statistics the particles obey.

For the neutral system, the relation (30) is replaced, in the long wavelength limit, by a zero-sound dispersion relation,

$$\omega_c = s_0 q \quad (qr_0 \lesssim 1) \quad (32)$$

where  $s_0$ , the velocity of the possible zero sound mode, is given by

$$s_0 = \sqrt{\frac{Nf_0}{m}} \quad (33)$$

on making use of (23). It follows that zero sound will be a well-defined collective mode for any many-particle system provided the condition (29) is satisfied, that is, provided

$$s_0 q > \omega_{sc} \quad (34)$$

For some of you the above discussion may seem hauntingly familiar; sixteen years ago David Bohm and I [18] used essentially the same language to discuss plasma oscillations and the possible existence of a sound mode in a neutral system. Our considerations were based on the random phase approximation, which led us to the correct conclusions for the case of the electron liquid; for the strongly interacting neutral system, our conclusions were qualitatively correct, but quantitatively in error; the RPA predicts that  $f_q$  is given by the Fourier-transform of the particle interaction

$$f_q = v_q = \int d\vec{r} e^{-i\vec{q}\cdot\vec{r}} v(r) \quad (35)$$

a result which is not valid in other than the weak coupling limit.

Consider next a strongly coupled normal Fermi liquid, one for which

$$F_0 = f_0/(E) \gg 1 \quad (36)$$

In this case the criterion (34) is well satisfied [ $\omega_{sc}$  being given by (31a)]; the predicted zero sound velocity is identical to that of the Landau theory. However the present theory permits one to go beyond the Landau theory: it predicts that zero sound should exist with essentially unchanged velocity in the superfluid state (a result subsequently verified by Leggett); it also predicts that at moderate wave-vectors ( $q \sim p_F/3$ , say) a zero-sound mode will persist at temperatures larger than the degeneracy temperature,  $T_F = E_F/\kappa$ , and that it will continue to exist until

$$\kappa T \approx \omega_{sc} \approx q s_0 = \frac{p_F s_0}{3} = \frac{p_F v_F}{3} \sqrt{\frac{F_0}{3}} = \frac{2}{3} \kappa T_F \sqrt{\frac{F_0}{3}} \quad (37)$$

Thus zero sound is predicted to exist under conditions that the Landau theory no longer works (because the quasi-particles are no longer well defined).

Put another way, for a strongly coupled Fermi liquid, one expects zero sound to exist, and to possess essentially the same velocity whether the system is normal or superfluid, degenerate or non-degenerate, so long as the condition (29) is satisfied. From this point of view, the interesting prediction of the Landau theory is not that zero sound exists for a strongly coupled system, but that it will likewise exist for a weakly coupled system, provided the temperature is low enough, the wavelength long enough.

Liquid  $^3\text{He}$  furnishes us with an example of a strongly coupled Fermi liquid, since one has

$$\begin{aligned} F_0 &\approx 10.5 \quad (0.17 \text{ atm}) \\ F_0 &\approx 73.2 \quad (27 \text{ atm}) \end{aligned} \quad (38)$$

Here the interesting question is whether zero sound exists in a temperature regime for which quasi-particles are not well defined, and the Landau theory no longer applies. Recently Shah [19] has carried out studies of the temperature dependence of the X-ray scattering cross-section at small angles; working at momentum transfers of  $\sim p_F/6$ , he finds strong evidence for a transition from first sound to zero sound as the temperature is reduced from  $0.8^\circ\text{K}$  to  $0.4^\circ\text{K}$ . Since at this latter temperature, the Landau theory has not yet begun to apply, this experiment may be regarded as providing confirmation for the above view of zero sound.

Liquid  $^4\text{He}$  is another system to which the above theory applies; in fact, the neutron scattering experiments of Woods [20] on variation with temperature of phonon energies in the collisionless regime provided the impetus for the theory. Before the experiment of Woods, it had generally been thought that in the collisionless regime, the existence of phonons in liquid He II was connected to its superfluidity, that is, to the presence of a condensate. Woods found, however, that in the superfluid regime the velocity of phonons of momentum  $0.38^{-1}\text{A}$  is essentially independent of  $\rho_s$ , the superfluid density; moreover neither the energy nor the lifetime of the excitation change

appreciably on going from He II to He I. From our present vantage point, such a result is to be expected; like  $^3\text{He}$ ,  $^4\text{He}$  is a strongly coupled system for which  $f_0$  may be expected to be so large that zero sound exists over a wide range of temperatures. Accordingly, one would expect to find it in the normal state, and to find its velocity little changed on going from the normal to the superfluid state (as was the case for plasmons in metals).

Classical liquids represent a fourth system to which the theory may be applied. Quite recently, Singwi, Sköld, and Tosi [21] have used a zero-sound approach to calculate the spectral function of the current-current correlation at short wavelengths and high frequencies. They apply the theory to explain neutron inelastic scattering results in liquid argon, and find good agreement with experiment.

In concluding this discussion, let me emphasize once again the following points:

- (1) Where the polarization potential is strong, one expects to find zero sound as a distinct excitation mode, be the liquid classical or quantum, neutral or superfluid, charged or neutral.
- (2) There exists at present no satisfactory microscopic calculation of  $f_q$  or  $\chi_{sc}$  for liquid  $^3\text{He}$  at elevated temperatures, or for liquid  $^4\text{He}$  in either the normal or superfluid state.
- (3) The zero-sound theory I have described may be directly transposed to other kinds of collective excitations; thus one is led to expect high frequency spin waves in a ferromagnetic above the Curie temperature, provided the spin polarization potential is strong enough, for just the same reason one has zero sound in He I; the existence of a condensate plays little role in a strongly coupled system.

## 5. CONCLUSION

As shown elsewhere in these Proceedings, the outlook, approaches and methods developed for the many-body problem in solid-state physics find a wide range of applicability in nuclear physics, in plasma physics, and, occasionally, even in particle physics. It is interesting to speculate on their utility in other fields — biology, for example. Might some of the techniques and attitudes we have developed for the many-body problem prove useful in molecular biology — not only in putting some existing theories in better order, but perhaps as well in devising new theories? And, to range still farther, might it not prove useful to consider certain social problems from a many-body point of view? Certainly cities, or ghettos, present us with a many-body problem. Will the social scientist of the future find it helpful to analyse interactions as strong or weak, and consider both collective and individual behaviour — both stable and unstable? Are there instabilities in societies which are signalled by the appearance of large-scale fluctuations? Perhaps just as we have had here the opportunity to learn something of the outlook of the molecular biologist, at some future symposium we'll have an opportunity to consider these, and related questions, in collaboration with our colleagues in the social sciences.

## R E F E R E N C E S

- [1] PINES, D., NOZIERES, P., *The Theory of Quantum Liquids* 1, W. A. Benjamin, New York (1966).
- [2] WHEATLEY, J. C., in *Quantum Fluids* (BREWER, D. F., Ed.) North-Holland Publishing Co., Amsterdam (1966) 183; *Phys. Rev.* 165 (1968) 304.
- [3] ANDERSON, P. W., *Physics* 2 (1965) 1.
- [4] DONIACH, S., ENGELSBERG, S., *Phys. Rev. Lett.* 17 (1966) 750.
- [5] BERK, N. F., SCHRIEFFER, J. R., *Phys. Rev. Lett.* 17 (1966) 433.
- [6] AMIT, D., KANE, J., WAGNER, H., *Phys. Rev.* 175 (1968) 313, 326.
- [7] KEEN, B. E., MATTHEWS, P. W., WILKS, J., *Physics Lett.* 5 (1963) 5.
- [8] ABEL, W., ANDERSON, A. C., WHEATLEY, J. C., *Phys. Rev. Lett.* 17 (1966) 74.
- [9] ABRIKOSOV, A. A., KHALATNIKOV, I. M., *Rep. Prog. Phys.* 22 (1959) 329.
- [10] DY, K. S., PETHICK, C. J. (to be published).
- [11] DY, K. S., PETHICK, C. J., *Phys. Rev. Lett.* 21 (1968) 876.
- [12] PETHICK, C. J., these Proceedings.
- [13] RICE, M. J., *Phys. Rev.* 159 (1967) 153.
- [14] RICE, T. M., *Phys. Rev.* 175 (1968) 858.
- [15] PINES, D., in *Quantum Fluids* (BREWER, D. F., Ed.) North-Holland Publishing Co., Amsterdam (1966) pp. 257-266.
- [16] Ref. [1] pp. 205-215.
- [17] Ref [1] pp. 90-93.
- [18] BOHM, D., PINES, D., *Phys. Rev.* 85 (1952) 338.
- [19] SHAH, N. P., *Phys. Rev. Lett.* 20 (1968) 1026.
- [20] WOODS, A. D. B., *Phys. Rev. Lett.* 14 (1965) 335.
- [21] SINGWI, K. S., SKÖLD, K., TOSI, M. P., *Phys. Rev. Lett.* 21 (1968) 881.

# PHASE TRANSITIONS AND CRITICAL PHENOMENA

MICHAEL E. FISHER

Baker Laboratory, Cornell University,  
Ithaca, N.Y., United States of America

## Abstract

PHASE TRANSITIONS AND CRITICAL PHENOMENA. The general theory of phase transitions and equilibrium critical phenomena is surveyed with special emphasis on rigorous results concerning the existence and non-existence of phase transitions in one-, two- and three-dimensional systems, and on the interrelations between the critical point exponents.

## 1. INTRODUCTION

I was originally asked to give a lecture regarding the "classical aspects of physics" and, indeed, essentially all that I want to say would come as no great surprise to the giants of 50 or 100 years ago if they could be with us here - I have in mind particularly Maxwell, Boltzmann, van der Waals and Gibbs, the founders, in fact, of the theoretical study of bulk matter in its "extensive" aspects. If we think of classical physics, and hence of bulk matter, the most outstanding characteristic is the existence of distinct phases - traditionally, gaseous, liquid and solid, although in practice other phases, or states of matter, in particular magnetic, superfluid, and superconducting, must now also be recognized. Initially it was natural, that after many centuries of every-day experience, the existence of phases was essentially taken for granted; it was really only at the time of van der Waals that a theoretical approach was made to the questions of the existence of phases, the way in which they relate to one another, and the characterization of the "phase transformations" which take place.

My talk will be divided into two parts: in the first part I shall mainly be concerned with the existence of phase transformations and, to some extent, with their nature. I will discuss the progress that has been made in the last ten years in the mathematical theory, stating what we know about the existence of phase transitions and, particularly, what we know about their non-existence in certain systems! One of the aspects of recent progress I want to stress is the use of really strict and rigorous mathematical methods to discuss some of these problems. In the second half of my talk I will move on to consider one of the most fascinating aspects of phase transitions: this is the detailed behaviour that one can observe in the vicinity of a "continuous phase transition" or what used to be called a "second-order order phase transition" (this latter name is in most cases rather inappropriate). The basic distinction here is well known: if at normal temperatures and pressures we boil a liquid, it changes abruptly into a vapour; if we condense a vapour it changes abruptly into liquid; these are first-order phase transitions; but we know there is, in fact, a critical point where the sharp distinction between liquid and vapour becomes meaningless. This is illustrated in Fig. 1 which shows the

temperature dependence of the densities,  $\rho_L$  and  $\rho_G$ , of coexisting liquid and gas. As the temperature  $T$  is raised the difference  $\rho_L - \rho_G$  becomes small and ultimately it vanishes continuously at the critical point  $(T_C, \rho_C)$ . As we pass through a critical point (say, at constant overall density) gas and liquid phases merge continuously while many physical properties exhibit "anomalies" both above and below  $T_C$ . As we will show, most of these anomalies have a universal character; thus, for example, plots like Fig. 1 have closely similar shapes (when normalized by  $T_C$  and  $\rho_C$ ) for a wide range of simple fluids; furthermore the universal shape is dominated by the occurrence of a definite mathematical singularity (or non-analyticity) at the critical point itself.

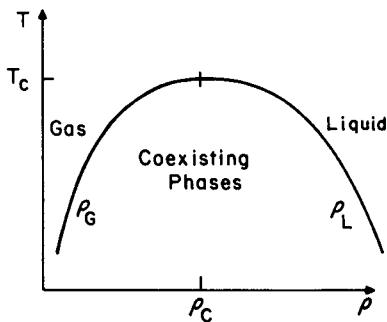


FIG. 1. Phase diagram of a fluid in the temperature-density plane indicating the critical point  $(T_C, \rho_C)$ .

## 2. THEORETICAL TOOLS

a) The traditional tool for the study of bulk matter and phase transitions is thermodynamics. Although we cannot manage without thermodynamics one should bear in mind that the use of thermodynamics is, in a way, an admission of weakness. Thermodynamics normally serves as a cloak for our ignorance of the detailed structure of a system or for our inability to analyze it theoretically. Frequently, in fact, "thermodynamic arguments" are dangerous since their assumptions are easily hidden and, even when visible, may appear innocent while in reality being profound and far-reaching.

b) Thus, rather than accepting thermodynamics directly we will call upon statistical mechanics; we will adopt the viewpoint that to discuss the equilibrium states of bulk matter, statistical mechanics is the basic tool. From the Hamiltonian  $\mathcal{H}$  of a system one may then both delve into the microscopic properties (which will be discussed in other lectures) and, through the standard formulae of statistical mechanics, calculate macroscopic properties. The "observables" which we compare with experiment are, firstly, what we may loosely call "equation of state data" e.g. the variation of the pressure with the density and the temperature, the anomalies in the compressibility and the specific heat and, secondly, those more detailed "correlations" or "fluctuations" in a system, which can be studied by making (elastic) scattering measurements, e.g. with light, X-rays or neutrons.

c) A third tool, which is becoming increasingly important, is non-equilibrium or time-dependent statistical mechanics. Here one again starts with the microscopic Hamiltonian but instead of making connections to thermodynamics one makes the connection to hydrodynamics. The "observables" are now, on the one hand, transport coefficients, like the thermal conductivity and viscosity and their state dependence, and, on the other hand, the time- and frequency-dependent correlations and fluctuations which may be studied by inelastic scattering experiments (particularly with neutrons). I will not discuss time-dependent or non-equilibrium phenomena in this talk but other speakers will consider them.

In all branches of science one may distinguish various levels of argument. I will describe firstly analyses which are rigorous and exact mathematically (such arguments are comparatively rare even in the physical sciences). In the present day, when the availability of large computers gives one the courage to make extensive numerical studies, one can often learn a lot from exact and precise numerical results; such has been the case in the theory of critical phenomena. Lastly, as always, one must consider arguments by analogy, by intelligent conjecture, and by, we hope, inspired speculation.

### 3. METHODS AND BASIC QUESTIONS

If, to start with, we agree that we are mainly interested in bulk matter which is describable by thermodynamics, we must discuss the so-called thermodynamic limit. In other words we consider a large physical system and ask for its asymptotic properties as it becomes of infinite extent. The traditional recipe for a system of  $N$  particles confined in a domain  $\Omega$  is firstly to calculate the partition function

$$Z(T, N, \Omega) = \text{Tr}_\Omega \{ e^{-\beta \mathcal{H}_N} \}, \quad \beta = 1/kT, \quad (1)$$

where

$$\mathcal{H}_N = T_N + U_N \quad (2)$$

is the Hamiltonian, formed from the kinetic energy  $T_N$  and the potential energy  $U_N$ . (In a system obeying classical mechanics  $\text{Tr}_\Omega$  denotes an appropriate integral over the  $N$ -particle phase space.) The free energy per particle,  $F(T, v)$ , at specific volume  $v = 1/\rho$  is then to be calculated from

$$-\frac{F(T, v)}{kT} = \lim_{\substack{V(\Omega) \rightarrow \infty \\ V/N \rightarrow v}} N^{-1} \ln Z(T, N, \Omega) \quad (3)$$

One of the points where progress has recently been made is in elucidating the circumstances in which one can really prove that there is such a unique limiting free energy. It has been known [1] that one can establish the existence of the free energy on the basis of essentially three conditions:

(a) Stability. The forces of interaction must have a "saturation" property (as familiar in the study of nuclear matter) which may be expressed by the existence of the lower bound

$$\langle \mathcal{E}_N \rangle \geq E_0(N) \geq -Nw_A, \quad (w_A > 0) \quad (4)$$

This ensures that the ground state energy  $E_0(N)$  does not go to  $-\infty$  faster than  $N$ , so that there is a maximum binding energy per particle and the system cannot "collapse" under its own forces.

A second, more technical condition is:

(b) Tempering. The forces of interaction must not be too repulsive at long distances. This may be expressed in terms of the potential energies of  $N+N'$  particles split into two groups of  $N$  and  $N'$ , respectively, with a minimum distance between the groups of  $R$ , by the condition

$$\Phi_{N,N'}(R) = U_{N+N'} - U_N - U_{N'} \leq NN'w_B/R^{d+\epsilon} \quad \text{for all } R \geq R_0 \text{ and fixed } \epsilon > 0, \quad w_B > 0 \quad (5)$$

Here  $d$  is the dimensionality of the system.

A final point is:

(c) Shapes of the container. If the shapes of the domains enclosing the system become too pathological (loosely, too great an area-to-volume ratio) proper thermodynamic behaviour does not occur.

On the basis of these and related conditions, rigorous mathematical discussions have been presented which lead to a number of valuable conclusions. One of the first results is to lay finally to rest a ghost that has haunted people since the time of Gibbs. Gibbs introduced a variety of ensembles; he left the grand-canonical ensemble until the end of his book (and as a consequence many text book writers have done likewise)! Unfortunately this has frequently left the impression that the different ensembles were on a really different footing: in particular near a phase transition it was not clear whether one got the same answers whichever ensemble one used. But now we know with certainty that if you calculate correctly you may use any ensemble you wish.

Another important conclusion concerns the convexity properties of the free energy (and all other thermodynamic potentials); these correspond to the conditions of thermodynamic stability (effectively the positivity of appropriate specific heats, compressibilities and susceptibilities). These convexity and stability properties are now seen to follow rigorously from the nature of the Hamiltonian for suitable forces. Among other things they show that an exact calculation can never yield the familiar Van Der Waals "loops", e.g. in the  $(p, v)$  plane.

A much-used and convenient theoretical idealization is to suppose the system of interest has only pairwise particle interactions with potential  $\phi(r)$ . In this case the stability and tempering conditions above are satisfied in  $d$  dimensions if

$$(i) \quad \phi(r) \geq C/r^{d+\epsilon}, \quad \text{as } r \rightarrow 0 \quad (6)$$

$$(ii) \quad |\phi(r)| \leq C'/r^{d+\epsilon'}, \quad \text{as } r \rightarrow \infty \quad (7)$$

where  $\epsilon$ ,  $\epsilon' > 0$  and  $C$  and  $C'$  are constants. These conditions are easily seen to be satisfied by the usual sort of Lennard-Jones ( $n, m$ ) interatomic forces with  $n = 6$  for the Van Der Waals tail and a strong repulsive core at the origin. Unfortunately, however, the condition (ii) excludes dipole-dipole forces, which are important in magnetism, since these decay only as  $1/r^3$  ( $d = 3$ ).

From a more fundamental point of view we would like to regard bulk matter as made up of nuclei and electrons rather than interacting atoms and molecules. Clearly the prime interaction is then the "bare" Coulomb interaction between the charges, which varies as  $1/r$  (generally, as  $1/r^{d-2}$ ) and hence violates both conditions (i) and (ii). A rigorous discussion of such a Coulomb system is hard and has not been fully completed. Recently, however, Dyson and Lenard [2] have established the stability of a system of interacting charges: given (a)  $N$  negative particles obeying Fermi-Dirac statistics (this is a particularly essential condition) and belonging to no more than  $q$  distinct species, the particles of each species having a finite, non-zero mass and a bounded charge, and (b) an arbitrary number of positive charges (of bounded charge but arbitrary statistics and masses, including  $n = \infty$ ), they prove there is a constant  $w_0$  such that

$$E_0 > - w_0 q^{\frac{2}{3}} N \quad (8)$$

As yet no general way has been found of avoiding the subsequent use of the tempering condition (b) and, indeed, it is not quite clear that the free energy of a system of interacting charges should inherit all the convexity properties of idealized atomic and molecular systems. Griffiths [3], however, has been able to prove the existence of the limiting free energy for what we may call completely symmetric systems. These include a system of magnetic spins with bilinear spin-spin interactions in zero external magnetic field, and a neutral "plasma" of "particles" and "antiparticles" (such that  $m^+ = m^-$  and  $e^+ = -e^-$ ) in zero external electric field.

#### 4. CHARACTERIZATION OF PHASE TRANSITIONS

The first essential theoretical clue to the characterization of a phase transition is the existence of a non-analyticity of the limiting free energy (or other thermodynamic potential) as a function of its arguments. Consider for example the free energy of a fluid which undergoes a first order transition as in condensation from gas to liquid. As indicated in Fig. 2 the free energy  $F(T, v)$  must have a linear portion as a function of  $v$  (corresponding to the region of coexistence of liquid and gas); such a linear region cannot be an analytic continuation of the rest of the function. The same transition would be described in terms of the pressure  $p$  and the chemical potential  $\mu$  as shown in Fig. 3. Here there is a discontinuous change in gradient at the transition point  $\mu_t$ . (The grand-canonical statistical mechanical ensemble will, of course, lead in the first place to this description.)

In the case of a continuous (or "second order") transition the non-analyticity may be more subtle. Thus the data for the specific heat of

liquid helium near its lambda point shown in Fig. 4 indicate unequivocally that the appropriate thermodynamic potential must contain a non-analytic term of a form close to  $-\text{sgn}\{T - T_C\}(T - T_C)^2 \ln |T - T_C|$ . The specific heats of many other materials, normal fluids, metallic alloys, ferromagnets and antiferromagnets, display very similar near-logarithmic singularities at appropriate critical points; in each case the free energy will contain a corresponding non-analyticity. (Experimentally, of course, a critical point or transition point is never indefinitely sharp; but in favourable circumstances and with sufficient skill critical points as sharp as  $\Delta T/T_C \sim 10^{-5}$  or  $10^{-6}$  have been observed in most classes of system.)

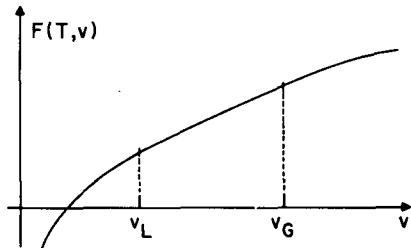


FIG. 2. Helmholtz free energy versus specific volume  $v$  indicating the linear portion corresponding to a first-order phase transition (condensation from gas to liquid).

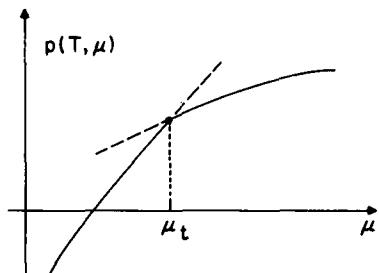


FIG. 3. Pressure versus chemical potential  $\mu$  showing the appearance of the same first-order transition as in Fig. 2.

A second most important, and in a sense more physical, characterization of a phase transition is in terms of an order parameter. One of Landau's important contributions was to stress this aspect of a phase transition. However, although an order parameter usually plays an important and natural role — often by way of describing a "broken symmetry" which characterizes the "ordered phase" — its occurrence does not seem to be universal. More generally one may focus attention on the long-range correlations (long range "order" if the correlations are of essentially infinite range) describing the system. To illustrate the concept of an order parameter and its usefulness in classifying phases and phase transitions I will first list a few concrete examples:

- (i) Scalar order parameter: Probably the conceptually simplest situation arises in a binary metallic alloy like  $\beta$ -brass ( $\sim 50\%$  Cu Zn). The intensity of the super-lattice line which appears in the X-ray diffraction pattern of the ordered phase is proportional to the square of the excess of atoms on the "right" sub-lattice over those on the "wrong" sub-lattice;

this difference constitutes the order parameter. In an anisotropic ferromagnet the spontaneous magnetization  $M_0(T)$  (observed in zero field) can point only along the (positive or negative) "easy axis". In the condensation of a gas one may properly regard  $\rho_L - \rho_G$  as the order parameter (although this already has some artificial aspects).

(ii) Vector order parameter: The spontaneous magnetization vector  $M_0(T)$  of an isotropic ferromagnet (like iron or nickel near  $T_C$ ) and the complex "condensate wave function"  $\langle\Psi\rangle$  in a superfluid or a superconductor are prime examples.

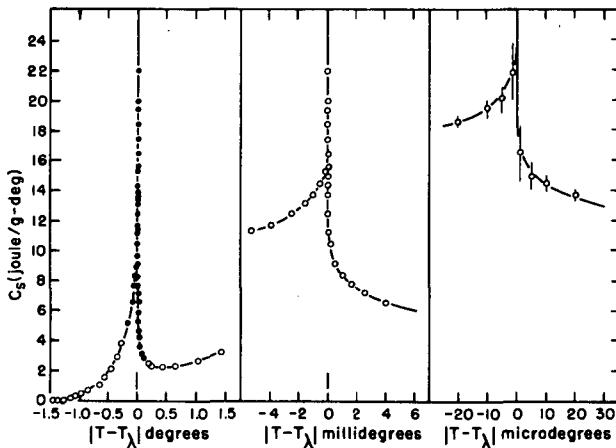


FIG.4. The specific heat of liquid helium (in coexistence with its vapour) through the lambda transitions. Note the increasingly magnified scales (after Fairbank et al. [42]).

(iii) Multi-dimensional order parameters arise in "metamagnets" where spiral orderings of variable pitch are observed. Other systems of layer, and linear ordering, such as occur in liquid crystals, etc., must also be remembered.

The way in which the order parameter enters the theory follows a general pattern which we may formalize as follows. (At each stage I will develop in parallel the more explicit expressions for a ferromagnet.) Firstly one discovers (or constructs) a locally defined mechanical variable (quantum-mechanically an operator)

$$\hat{\Psi} = \hat{\Psi}(\vec{r}) \quad (9a)$$

which represents the microscopic order; for a magnet this is just the local magnetization

$$\hat{M}(\vec{r}) = m \hat{S}(\vec{r}) \quad (9b)$$

where  $\vec{S}(\vec{r})$  is the spin variable at the site  $\vec{r}$ . The total order parameter for the system is defined by

$$\hat{\Psi}_{\text{tot}} = \int_{\Omega} \hat{\Psi}(\vec{r}) d\vec{r} \quad (10a)$$

$$\hat{M}_{\text{tot}} = \sum_{\vec{r}} m \hat{S}(\vec{r}) \quad (10b)$$

One can then define a conjugate "ordering" field  $\xi$ ; for a magnet this is simply the external magnetic field  $H$ , but, unfortunately, the conjugate field is not always physically realizable. It corresponds generally to an increment to the basic Hamiltonian  $\mathcal{H}$  given by

$$\Delta\mathcal{H} = - \int_{\Omega} \xi \Psi(\vec{r}) d\vec{r} \quad (11a)$$

$$= - \sum_{\vec{r}} m \vec{H} \cdot \vec{S}(\vec{r}) \quad (11b)$$

(It can be useful to allow  $\xi$  and  $H$  to depend on  $\vec{r}$ .) If the basic Hamiltonian has some definite symmetry one can usually choose a "symmetric" order parameter. Thus in the case of a scalar order parameter and a reflection symmetry we have

$$\mathcal{H}\{\hat{\Psi}(\vec{r})\} = \mathcal{H}\{-\hat{\Psi}(\vec{r})\}; \quad \mathcal{H}\{S_z(\vec{r})\} = \mathcal{H}\{-S_z(\vec{r})\} \quad (12)$$

In such a case the field  $\xi$  (or  $H_z$ ) is clearly a "symmetry breaking" field (and furthermore its presence normally destroys the transition as a function of, say, temperature).

The thermodynamic or "equilibrium" value of the order function can be derived from the free energy for general  $\xi$  as the "response" to the field i.e.

$$\Psi(\xi, T) = \langle \Psi \rangle = - \left( \frac{\partial F}{\partial \xi} \right)_T; \quad M(H, T) = - \left( \frac{\partial F}{\partial H} \right)_T \quad (13)$$

The existence of a phase transition may then be signalled by the existence of a "spontaneous order"  $\Psi_0$  which "spontaneously breaks the symmetry". Mathematically  $\Psi_0$  is best defined by a difference between left and right derivatives in (13) i.e.

$$\Psi_0(T) = \lim_{\xi \rightarrow 0^+} \frac{1}{2} [\Psi(\xi, T) - \Psi(-\xi, T)] \quad (14a)$$

$$M_0(T) = \lim_{H \rightarrow 0^+} \frac{1}{2} [M(H, T) - M(-H, T)] \quad (14b)$$

An alternative somewhat more general microscopic viewpoint is to define the correlation functions, say,

$$G(\vec{r}) = \langle \Psi(\vec{0}) \Psi(\vec{r}) \rangle; \quad \Gamma(\vec{r}) = \langle S_{\vec{0}}^z \cdot S_{\vec{r}}^z \rangle \quad (15)$$

Normally these will decay to zero [in the disordered phase this is a natural expectation based on the symmetry (12)]. A new ordered phase, however, may be characterized by a "long range order"

$$G(\infty, T) = \lim_{|\vec{r}| \rightarrow \infty} G(\vec{r}); \quad \Gamma(\infty, T) = \lim_{|\vec{r}| \rightarrow \infty} \Gamma(\vec{r}) \quad (16)$$

In normal circumstances one expects the long range order to be equal to the square of the "spontaneous order" i.e.

$$G(\infty, T) = [\Psi_0(T)/g]^2, \quad \Gamma(\infty, T) = [M_0(T)/m]^2 \quad (17)$$

where  $g$  is a constant of appropriate dimensions. When the distribution function for the order parameter can be defined it is similarly expected to be sharply "peaked" at the values  $\pm \Psi_0$ . Unfortunately it is hard to justify (17) rigorously even for simple models. (Indeed over-simple ideas concerning the double peaked distribution can be shown to be incorrect.) Furthermore one can imagine a phase transition characterized, say, by an infinite response to the ordering field ( $\partial^2 F / \partial \xi^2 \rightarrow \infty$  as  $\xi \rightarrow 0$  for  $T \leq T_C$  or  $\chi_T = \partial M / \partial H \rightarrow \infty$  as  $H \rightarrow 0$  for  $T \leq T_C$ ) but for which there is no spontaneous order i.e.  $\Psi_0(T) \equiv 0$  (or  $M_0(T) \equiv 0$ ). This would arise if, below  $T_C$ , the correlation function  $G(\vec{r})$  still decayed to zero as  $|\vec{r}| \rightarrow \infty$ , but only so slowly that the integral  $\int G(\vec{r}) d\vec{r}$  (which will generally be proportional to  $\partial^2 F / \partial \xi^2$ ) was divergent. Such "order-less" phase transitions may even be of practical significance (see below). Despite such difficult questions, the concept of an order parameter and the conceptual scheme outlined above has proved most fruitful in recent years - not least of all by bringing into clear focus the strong analogies between phase transitions in physically very different systems.

## 5. NON-EXISTENCE OF PHASE TRANSITIONS - ONE-DIMENSIONAL SYSTEMS

Moving on now from generalities let us discuss some particular instances where one can prove that phase transitions (or some of their correlates) do not occur. In the literature there frequently appear statements about the non-existence of phase transitions in one-dimensional systems which are a little misleading. I will thus try to state rather carefully some things that can really be proved rigorously - the main point is that one must impose extra conditions on the potentials. I will also show that if these conditions are not satisfied one-dimensional phase transitions may very well occur.

Firstly consider a system of particles which (a) obey classical mechanics, and which (b) interact with an infinitely repulsive hard core (so that only a finite number of particles can occupy a finite volume). Further we suppose (c) that the interactions are of a strictly finite range  $b$  in the sense that  $\Phi_{N,N'}(R)$  defined in (5) vanishes identically if  $R > b$ . Otherwise (d) the forces may be of 2-body, 3-body up to  $\ell_0$ -body character, for some fixed finite  $\ell_0$ . Then if the domain  $\Omega$  occupied by the particles is a cylinder of finite cross-section  $A$  (but of length  $L \rightarrow \infty$ ) there are no phase transitions at non-zero temperatures (more specifically the limiting free energy per particle as  $L \rightarrow \infty$  is analytic in  $T$  and  $\rho < \rho_{\max}$  and in the other parameters that enter the Hamiltonian analytically).

The proof of this result [4] follows an argument of van Hove [5]. Firstly one constructs an integral operator  $\underline{K}(A, T, \mu)$  which increases the length  $L$  of a finite cylinder by a distance  $b$ . In view of the conditions (b), (c) and (d) this operator entails only a finite number of variables and is non-negative and bounded. The largest eigenvalue  $\lambda_0(T, \mu)$  of  $\underline{K}$  determines the limiting thermodynamic potential, the pressure  $p(T, \mu)$ , as  $L \rightarrow \infty$ . By appropriate generalizations of Frobenius' theorem on positive matrices one knows that  $\lambda_0$  is non-degenerate (which enables one to rule out long-range order), and furthermore that  $\lambda_0$  depends analytically on the parameters entering  $\underline{K}$ . Thus while  $A$  is finite, so that the system remains effectively one-dimensional, there is no phase transition; at least one more infinite dimension is required for a phase transition to be possible. (In the limit  $A \rightarrow \infty$  the eigenvalue  $\lambda_0(A)$  may become asymptotically degenerate with, say,  $\lambda_1(A)$  and the proof of analyticity breaks down as the number of variables in  $\underline{K}$  increases without limit.)

The theorem extends to lattice systems with discrete local variables (e.g. lattice gases, see below) and to quantum mechanical systems of, for example, spins provided that all the variables actually appearing in the Hamiltonian commute (such systems are, of course, effectively "classical").

Recently Gallavotti, Miracle-Solé and Ruelle [6] have proved a somewhat different theorem for a completely one-dimensional (or linear) particle system with a pair interaction potential  $\phi(r)$  which is again assumed (a) to have an infinite hard core of diameter, say,  $a$ . A second condition (b) concerns the long range behaviour of  $\phi(r)$ : it is supposed that there is a decreasing function  $v(r) \geq 0$  such that

$$|\phi(r)| \leq v(r) \quad \text{for } r > a$$

and

$$\int_a^\infty r v(r) dr < +\infty \tag{18}$$

This means, in effect, that  $\phi(r)$  decays to zero more rapidly than  $1/r^2$ .

[Note that the weaker condition  $\int_a^\infty v(r) dr < +\infty$  is sufficient to ensure stability and tempering and hence the existence of a limiting free energy.] The condition (18) implies that the interaction energy between all particles to the left of some fixed point and all those to the right is always finite and bounded i.e.

$$|\Phi_{N,N'}| \leq |\Phi_{\max}| < \infty \quad \text{all } N, N' \tag{19}$$

This is the essential relation; it may be interpreted to mean that the total "interfacial energy" (or "surface tension") between two regions of opposite local order, in some appropriate sense, is always bounded thereby permitting the smooth break-up of an ordered region by natural fluctuations forming many "interfaces" [7]. In any event a rigorous argument based on (18) shows that the first derivatives of the thermodynamic potential  $p(T, \mu)$ , and all the many-body correlation functions (integrated against test functions) are continuous functions of  $T$  and  $\mu$ . In effect this rules out a phase transition of any finite order (in an  $n$ th order transition the  $n$ -body correlation functions are discontinuous); the result, however, is clearly somewhat weaker than a statement of complete analyticity.

For a one-dimensional "lattice gas" (in which a particle can occupy only the sites of a regular lattice to the exclusion of a second particle) Ruelle [8] had earlier proved a similar result which, however, also allowed for many-body interaction potentials  $\phi_\ell(r_1, r_2, \dots, r_\ell)$ . His condition for continuity of the density, energy and correlation functions is

$$\sum_{\ell \leq 2} \sum_{r_1=0} \dots \sum_{r_\ell < r_1 < \dots < r_\ell} r_\ell |\phi_\ell(r_1, \dots, r_\ell)| < +\infty \quad (20)$$

This condition again implies the bound (19) on the interaction energy. On the other hand if we violate the bound (19), phase transitions are possible: this may be seen in an exact calculation of the equation of state of a one-dimensional (lattice or continuum) gas in which, in addition to the hard repulsive cores, many-body forces of a particular type are introduced [9]. Specifically we postulate a "clustering distance"  $c$  and consider many-body potentials satisfying

$$\phi_\ell(r_1, r_2, \dots, r_\ell) \equiv 0$$

$$\text{whenever } r_{j+1} - r_j > c \text{ for any } j = 1, 2, \dots, \ell-1 \quad (21)$$

(Since the system is one-dimensional we may always assume  $r_1 < r_2 < \dots < r_\ell$ .) These potentials vanish unless  $\ell$  successive particles form a "cluster". They are thus strictly short range in the sense previously explained [ $\Phi_{N,N'}(R) \equiv 0$  if  $R > c$ ]; however we allow indefinitely large values of  $\ell$ . The simplest model is obtained when

$$\phi_\ell(r_1, r_2, \dots, r_\ell) = \text{constant} < 0$$

$$\text{provided } r_{j+1} - r_j < c \text{ for all } j = 1, 2, \dots, \ell-1 \quad (22)$$

For stability one must impose  $\ell\phi_\ell \rightarrow 0$  but Ruelle's condition (20) requires, in addition,  $\ell^2\phi_\ell \rightarrow 0$ . We may thus choose physically sensible interactions such that  $\ell^2\phi_\ell \rightarrow \infty$  (or  $\lim_{\ell \rightarrow \infty} \ell^2\phi_\ell > 0$ ). Such potentials violate (19) and indeed lead to first order phase transitions (and to a class of higher order transitions). More elaborate choices of  $\phi_\ell(r_1, \dots, r_\ell)$  yield a surprising range of different types of phase diagram.

One-dimensional many-body forces of the type (21) are not, of course, known in real systems. However, a fairly realistic pseudo-one-dimensional model of the helix-coil transition in a double-stranded polymeric molecule (such as a nucleic acid) can also yield a phase transition. Figure 5 shows the model schematically: ordered double-helical regions of the molecule alternate with disordered single-stranded closed loops. The ordered regions are favoured by some binding energy per link (properly a free energy) while the disordered loops are favoured statistically by their entropy. For a closed loop of  $n$  links of polymer chain, the reduced entropy  $S_n/k$  varies for moderate to large  $n$  as  $s_0 n - s_1 \ln n$  where the pure number  $s_1$  depends on the dimensionality of the space in which the loop wanders and on the steric self-hindrance of the polymer chain [10]; for real systems  $s_1 \approx 1.75$ . The model may be solved theoretically [9] (if the interferences between successive loops are neglected) and one finds that an infinitely long molecule will undergo a true phase transition as a function of temperature ("melting" of the ordered regions) whenever  $s_1 > 1$ .

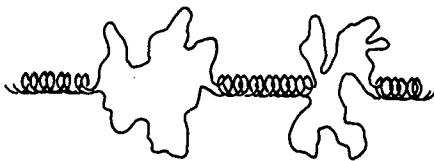


FIG.5. Schematic sketch of a double-stranded polynucleotide molecule with disordered and ordered, double helical, sections.

Finally we mention that none of the above proofs (or counter-examples) applies to a truly quantum-mechanical system; one suspects, however, that some rather similar theorems are still true.

#### Note added November 1968

Dyson has recently proved that a one-dimensional lattice gas with a pair interaction potential  $\phi_2(r)$  ( $r = na$ ,  $n = 0, 1, 2, \dots$ ) such that  $-\phi_2(r)$  is positive and monotonic decreasing (except that  $\phi(0) = +\infty$ ) will have a phase transition provided the sum

$$S = \sum_{n=1}^{\infty} (\ln \ln n) / n^3 \phi_2(na)$$

converges. When  $S$  is finite  $\phi_2(r)$  necessarily violates (18) or (20). Specifically if  $\phi_2(r) \sim 1/r^{1+\sigma}$  there is a phase transition if  $\sigma < 1$  but none if  $\sigma > 1$  ( $\sigma > 0$  is required for stability).

## 6. NON-EXISTENCE OF PHASE TRANSITIONS - CONTINUOUS SYMMETRY GROUPS

The proofs of the non-existence of a phase transition discussed above hinge in an essential way on the dimensionality of the system being unity. Recently, however, it has been discovered by Mermin, Wagner and Hohenberg [11, 12] that one can rigorously exclude certain phase tran-

sitions with a spontaneous order parameter (or long range order) if the Hamiltonian possesses a continuous symmetry. The gist of the argument is as follows.

We start with the following inequality due to Bogolyubov:

$$|\langle [\hat{C}, \hat{A}] \rangle|^2 \leq (kT)^{-1} \langle \frac{1}{2} \{ \hat{A}, \hat{A}^\dagger \} \rangle \langle [\hat{C}, \hat{\mathcal{H}}], \hat{C}^\dagger \rangle \quad (23)$$

where  $\hat{A}$  and  $\hat{C}$  are general quantum-mechanical operators,  $\hat{\mathcal{H}}$  is the Hamiltonian operator,  $\{, \}$  denotes the commutator, and  $\{, \}$  denotes the anticommutator. This inequality is just the Schwarz inequality for a vector space of operators with a scalar product defined by

$$\langle \hat{A}, \hat{B} \rangle = \sum_{n \neq m} \frac{A_{nm}^* B_{nm}}{\text{Tr} \{ \exp(-\beta \hat{\mathcal{H}}) \}} \frac{e^{-\beta E_n} - e^{-\beta E_m}}{E_m - E_n} \quad (24)$$

where  $E_n$  are the eigenvalues of  $\hat{\mathcal{H}}$ , and the matrix element  $A_{nm}$  etc. are taken in the corresponding diagonal representation.

Next we suppose there is a group of local transformations  $\hat{t}_\theta(\vec{r})$  with a continuous parameter  $\theta$ . A good example would be the rotation of an individual spin at  $\vec{r}$  in a magnet through an angle  $\theta$  about a chosen axis (say, the y axis). The corresponding uniform total transformation will be

$$\hat{T}_\theta = \int_{\Omega} d\vec{r} \hat{t}_\theta(\vec{r}) \quad (25)$$

In a magnet this would represent a uniform rotation of all the spins.

In many cases the Hamiltonian  $\hat{\mathcal{H}}\{\Psi(\vec{r})\}$  with local order parameter  $\Psi(\vec{r})$  will be invariant under the uniform transformation i.e.

$$\hat{T}_\theta \hat{\mathcal{H}}\{\Psi(\vec{r})\} = \hat{\mathcal{H}}\{\hat{t}_\theta(\vec{r})\Psi(\vec{r})\} = \hat{\mathcal{H}}\{\Psi(\vec{r})\} \quad (26)$$

We may refer to a system with such a Hamiltonian as "isotropic". An example would be a ferromagnet with spin-spin coupling of the usual Heisenberg form  $\vec{S}_i \cdot \vec{S}_j$ ; however, a magnet with only one axis of rotational symmetry would also qualify. We presume, of course, the absence of a symmetry breaking field  $\zeta$ , which, nonetheless, we may reintroduce for convenience.

The method is now to look at the infinitesimal generators  $\hat{c}(\vec{r})$  of the local transformations for which we have

$$\hat{t}_\theta(\vec{r}) = e^{i\theta \hat{c}(\vec{r})} \quad (27)$$

The operator

$$\hat{C}(\vec{k}) = \int_{\Omega} d\vec{r} e^{i\vec{k} \cdot \vec{r}} \hat{c}(\vec{r}) \quad (28)$$

then generates an infinitesimal "twist" of the system with a wavelength  $\lambda = 2\pi/|\vec{k}|$ . In the case of a ferromagnet the resulting state is essentially a "spin wave" and, indeed, the proof will develop along lines which are rather similar to old heuristic spin wave arguments which suggested that an isotropic ferromagnet would only exhibit a spontaneous magnetization if the dimensionality was at least  $d = 3$ .

By the symmetry (26) the uniform operator  $C(0)$  must commute with the Hamiltonian so that the double commutator in (23) then vanishes identically. The symmetry can be broken by coupling the field  $\xi$  to the operator  $\hat{\Psi} = \{\hat{C}(0), \hat{A}\}$  where  $\hat{A}$  is to be chosen so that  $\Psi$  is the desired order parameter. (This is normally straight-forward.) Evaluation of the double commutator of  $\hat{C}(k)$  and  $\hat{\mathcal{H}}$  for small  $k^2$  then yields

$$D(k) = \langle [\hat{C}, \hat{\mathcal{H}}, \hat{C}] \rangle = b^2 k^2 + \xi \Psi + O(k^2) \quad (29)$$

Since  $D(0)$  had to vanish by (26) this result could have been guessed on the grounds of continuity. The constant  $b^2$  is a measure of the range of the interactions which must be supposed to decay sufficiently rapidly with distance. (The pair interactions must have a second moment so that the Fourier transform of the long range part will vary as const. +  $b^2 k^2$  for small  $k$ ; in fact for a magnet with exchange energy  $J(\vec{r})$ , what enters is  $\hat{J}(\vec{k}) - \hat{J}(\vec{0})$ .)

Bogolyubov's inequality (23) can then be rearranged and integrated on  $\vec{k}$  to yield (after the use of a few inequalities which may vary with the particular problem)

$$|\Psi|^2 \int \frac{d\vec{k}}{D(\vec{k})} \leq (\text{constant})/kT \quad (30)$$

The "fluctuation integral" here is always finite when the field  $\xi$  does not vanish. In a three-dimensional system it remains finite as  $\xi \rightarrow 0$  and we obtain an interesting but not very useful bound on  $\Psi_0$  the spontaneous order. However since  $D(k) \propto k^2$  when  $\xi = 0$  the integral diverges at small  $k$  in a two- or one-dimensional system in zero field. Consequently as  $\xi \rightarrow 0$  we can obtain bounds of the form

$$\begin{aligned} \Psi(T, \xi) &\leq a/T^{\frac{1}{2}} |\ln \xi^2|^{\frac{1}{2}} \quad \text{for } d=2 \\ &\leq a\xi^{\frac{1}{3}}/T^{\frac{2}{3}} \quad \text{for } d=1 \end{aligned} \quad (31)$$

where  $a$  is constant. These results prove that  $\Psi$  must vanish as  $\xi \rightarrow 0$  and hence [see (14)] that there is no spontaneous order possible (at non-zero temperature) in an "isotropic" one- or two-dimensional system (with finite range forces).

We have pointed out that the argument proves the absence of ferromagnetic order in an isotropic magnet; it works equally well [11] to prove (a) the absence of anti-ferromagnetic ordering in one or two dimensions. By the use of gauge symmetry it shows that (b) superconductors and (c)

superfluids cannot exhibit the usual sort of "condensation" in one and two dimensions. Similarly, translational symmetry can be employed to prove (d) that one- and two-dimensional crystals cannot exist (in the sense that they cannot exhibit a true Bragg peak in the scattering of X-rays, etc.). Of course this last result is well known in the case of a purely harmonic crystal but the present proofs cover general (anharmonic) interactions.

It is worth remarking that the bounds (31) do not rule out the possible phase transition mentioned in our initial discussion of order parameters, in which the susceptibility  $\partial\Psi/\partial\zeta$  diverges to infinity as  $\zeta \rightarrow 0$  even though  $\Psi$  itself vanishes. Indeed there are some numerical indications that this interesting situation might occur in certain isotropic two-dimensional magnetic systems [13].

## 7. THE EXISTENCE OF PHASE TRANSITIONS

There are many heuristic and approximate theoretical arguments which suggest that systems described by certain types of Hamiltonian should indeed exhibit phase transitions. The rigorous negative results described above demonstrate that one must be rather wary of some of these arguments (particularly those of a "mean-field" character) but, on the positive side, we do now have a number of exact results and rigorous proofs that really establish the existence of phase transitions in particular systems. Some of these analyses gives us appreciable insight into "where the transition comes from". I will survey briefly the extent of our present knowledge:

### (a) Soluble models

In the first place there are a number of models for which the thermodynamic properties and also the correlation functions can be calculated explicitly in more or less complete mathematical detail.

- (i) One-dimensional models. We may mention again the many-body cluster interaction model of a linear gas and the double-stranded polymer models described already.
- (ii) Two-dimensional Ising ferromagnet or lattice gas. This model has proved to be one of the most fruitful models of a many-body system ever discovered. In the magnetic version one supposes that  $S = \frac{1}{2}$  spin variables are associated with every point (or site) of a regular space lattice, for example, a plane square lattice. Spins at sites  $i$  and  $j$  are coupled by an Ising interaction of the form  $-J_{ij} S_i^z S_j^z$ . The model is mathematically equivalent to a lattice gas with interaction potential  $\phi(r_{ij}) \propto -J_{ij}$ . It also describes quite well binary alloys, binary fluid mixtures, anti-ferromagnets, etc.

The Ising model with nearest-neighbour interactions on the square lattice was solved for zero magnetic field by Onsager in 1944 [14]. It was found to exhibit a continuous phase transition at a temperature  $T_c$  at which the specific heat diverged logarithmically to infinity (in striking contrast to the predictions of all previous approximate treatments). Below  $T_c$  it exhibits spontaneous magnetization and long range order.

Much work has been done on the Ising model since 1944 and a variety of other lattices have been solved. No one, however, has really been able to avoid the restriction to zero field or to planar lattices (without crossing interaction bonds). It is also unfortunate that despite the development of much beautiful mathematics the exact solutions are rather unilluminating from a physical viewpoint. The origin of the transition and especially the reasons for its particular nature remain quite mysterious.

- (iii) Two-dimensional hydrogen-bonded ferroelectrics. Very recently Lieb [15], has solved and he, Sutherland, C.N. Yang and C.P. Yang [16] have investigated in detail a most interesting class of two-dimensional models which enshrine the main physical principle responsible for the "residual" entropy of ice and for the behaviour of ferroelectrics of the potassium dihydrogen phosphate (KDP) class. This principle, the so-called "ice-rule" states that in a four-co-ordinated hydrogen-bonded crystal precisely two of the four hydrogen atoms (or protons) neighbouring each principal ion must lie close to the ion while the remaining two must lie far from it (i.e. close to the adjacent ions; notice that the rule allows many possible configurations of the protons). These models exhibit a range of fascinating and subtle phase transitions; I will not describe them further since Professor Lieb himself will talk about the work. It is again sad, however, that the mathematics seems to conceal as much as it reveals of the "working" of the model.

#### (b) General existence proofs

Rather more illuminating as regards the physical "origin" of the transition have been rigorous general proofs of the existence of a phase transition in the Ising model for all dimensionalities  $d \geq 2$ . The first real proof was given by Griffiths [17] following old heuristic arguments of Peierls. Significant further work has been done especially by the Russian workers Minlos and Sinai and Dobrushin [18]. What is actually proved is the existence of a spontaneous magnetization [in the sense of the definition (14)] in zero field below some temperature  $T_0$ , which is thus a lower bound to the critical temperature  $T_c$ . [Rather accurate rigorous upper bounds to  $T_c$  can also be found [19].

There are two basic "ingredients" used in the existing proofs. The first of these is:

- (i) The symmetry between positive and negative fields. Equivalently there is a hole-particle symmetry in the lattice gas so that the location of the transition i.e. the chemical potential  $\mu_t$ , is effectively given a priori. One knows physically that the real gas-liquid transition is not perfectly symmetric, so this feature of the proofs should not be essential. It seems, however, hard to dispense with.

Secondly we have:

- (ii) The surface-to-volume ratio of a droplet of  $\ell$  atoms. This seems to be the crucial physical feature leading to an abrupt first order transition or condensation. Thus consider for an example an anisotropic ferromagnet

in an ordered state with all spins pointing "up". If we wish to break the ordered state by overturning a group of  $\ell$  spins to form a "down" domain or "droplet" we have, in general, to do two types of work: firstly there is a "bulk" term proportional to  $\ell$  (due to the interaction with the external magnetic field); and secondly, there is a term arising from the interactions with the remaining "up" spins; this varies as  $s(\ell)$  the "surface area" of a cluster, where, clearly  $s(\ell)/\ell \rightarrow 0$  as  $\ell \rightarrow \infty$ . The bulk term vanishes at the transition point (in zero field) but the surface term remains. Furthermore, and this is a vital feature, if the dimensionality is two or more we must have  $s(\ell) \rightarrow \infty$  as  $\ell \rightarrow \infty$ . The resulting mechanism of the transition has been discussed heuristically by a number of authors (see Ref.[9]) but the rigorous proofs really show it is the prime physical "cause".

Conversely the previous non-existence proofs indicate that if  $s(\ell)$  does not diverge, or if there is no surface energy, an abrupt transition cannot occur.

### (c) The Van Der Waals or long range limit

The considerations above rested on the short range character of the basic forces of attraction causing the transition. This is indeed the normal situation in most real physical systems. However from a theoretical viewpoint it is instructive to consider a system with long-range but weak forces. Following detailed one-dimensional model calculations by Kac, Uhlenbeck and Hemmer [20] and a more general analysis by Lebowitz and Penrose [21] one may discuss a fluid with potentials of the form

$$\phi(r) = \phi_0(r) + \lambda^d \phi_1(\lambda r) \quad (32)$$

where  $\phi_0$  and  $\phi_1$  have the usual short range character. If we take the thermodynamic limit, and afterwards - this, of course, is an artificial and non-physical procedure - take the "long-range limit", namely  $\lambda \rightarrow 0$  with  $\alpha = \int \lambda^d \phi_1(\lambda \vec{r}) d\vec{r} = \int \phi_1(\vec{r}) d\vec{r}$  constant, we find that the resulting free energy exhibits a first order phase transition (of the classical van der Waals character) for a wide-range of potentials  $\phi_0$  and  $\phi_1$ . This result throws light on a number of commonly-used approximate techniques but its artificiality shows up particularly in one dimension where a transition can occur in the limit  $\lambda \rightarrow 0$  even though there is no transition for any non-zero value of  $\lambda$ . [Notice also that if  $\lambda \rightarrow 0$  before the thermodynamic limit is taken  $\phi(\vec{r})$  reduces simply to the short range potential  $\phi_0(\vec{r})$ .]

## 8. NATURE OF THE SINGULARITY AT CONDENSATION

We have seen that what characterizes a phase transition in a general way is the existence of a non-analyticity in the thermodynamic potential. Thus in the case of the condensation of a fluid, as illustrated in Fig. 3, the pressure  $p(T, \mu)$  as a function of the chemical potential  $\mu$  has a discontinuous gradient at the transition  $\mu_t$ . For  $\mu$  sufficiently smaller than  $\mu_t$ , however,  $p$  is known to be an analytic function of  $\mu$ , say  $p_<(\mu)$ . Now it is interesting theoretically to ask what happens as we analytically continue  $p_<(\mu)$  along the

real  $\mu$ -axis towards  $\mu_t$ . Will the transition point itself be a non-analytic point of  $p_c(\mu)$ ? Or will it be possible to analytically continue  $p_c(\mu)$  beyond  $\mu_t$  to describe a "natural" extension of the isotherm?

The traditional, if implicit, answer to this second question has been "Yes"; indeed the analytically continued isotherm has been supposed (for at least some way beyond  $\mu_t$ ) to represent a "metastable" but physically realizable extension of the low density phase i.e. a supersaturated vapour. This answer is also supported by many approximate calculations of mean field type and by the process, described above, of taking the long-range or van der Waals limit. However, this limit is very artificial and we should not trust the answer as regards systems with more realistic short-range forces. Indeed one finds that in the exactly soluble one-dimensional many-body cluster interaction models [9] there is a singularity in  $p_c(\mu)$  precisely at the transition point  $\mu_t$ . This confirms for the model a long-standing conjecture of Mayer that answers "Yes" to the first question and hence concludes that the condensation point (at fixed T) could in principle be found solely from the low density virial coefficients. (These serve to define  $p_c(\mu)$  completely.)

At present, however, these questions are open as regards two- or three-dimensional models with short range forces even in the simplest case of a nearest-neighbour lattice gas or Ising model. It is tempting, however, to speculate along lines inspired by the rigorous proofs of the existence of condensation in the Ising model. As we saw, a crucial feature was the relation between the surface and volume free energies of a "droplet" of  $\ell$  atoms of, say, liquid phase in the dilute vapour phase. If we ignore [9] interactions between droplets of different sizes, as is not unreasonable at low densities, we have

$$p/kT \approx \sum_{\ell=1}^{\infty} [q_{\ell}(T, \Omega)/V(\Omega)] z^{\ell} \quad (33)$$

where  $z = \exp(\mu/kT)$  and  $q_{\ell}(T, \Omega)$  is the partition function for a droplet in the domain of volume  $V(\Omega)$ . The form of  $q_{\ell}$  for large  $\ell$  is then conjectured, to involve a bulk contribution proportional to  $\ell$ , a surface term proportional to some average  $\bar{s}(\ell) \sim \ell^{\sigma}$  with  $0 < \sigma < 1$ , and various higher order geometrical terms. The analytical properties of  $p(\mu)$  are hence found to be similar to those of the generating function

$$P(x, z) = \sum_{\ell=1}^{\infty} \ell^{-\tau} x^{\ell} \sigma(z/z_t)^{\ell} \quad (34)$$

in which  $x=x(T) \rightarrow 0$  as  $T \rightarrow 0$  (and  $\tau$  is a geometrically determined exponent). A simple analysis shows that condensation occurs at  $z = z_t$  but, furthermore, that this point is an essential singularity of  $P(z)$  [and hence of the corresponding  $p_c(\mu)$ ]. At this condensation point all the derivatives of  $p_c(\mu)$  exist and are finite but nonetheless the function cannot be continued through  $\mu_t$ . (There may be an analytical continuation around  $\mu_t$  but this will not be real for real  $\mu$ .)

This argument is, of course, speculative although it embodies the essential feature needed in the rigorous proofs and, appropriately adapted, predicts correctly the lack of a phase transition in short range one-dimensional and "isotropic" two-dimensional systems. Future research should bring us a more certain answer and a deeper insight into the statistical mechanism of first order phase transitions.

## 9. CRITICAL POINT SINGULARITIES

Let us turn now from considering general existence questions to matters more intimately connected with experimental observation. At present this entails working at a lower level of mathematical rigour but the challenge to our theoretical ability and understanding will be correspondingly greater. Specifically I want to discuss phenomena in the vicinity of continuous phase transitions or critical points. We may take the critical point of a fluid (Fig. 1) and the Curie point of a ferromagnet as canonical examples, although many other classes of system are amenable both to theory and experiment.

Our first concern is to introduce the now rather extensive family of "critical exponents" used to describe the asymptotic behaviour of physical observables as a critical point is approached closely. We have already remarked (a) that the experimentally observed specific heat of helium near its lambda point  $T_c = T_\lambda$  diverges logarithmically with  $|\Delta T| = |T - T_c|$ , and (b) that the theoretically calculated specific heat of the two-dimensional Ising models behaved rather similarly. More generally we allow for a divergent specific heat  $C(T)$  by defining exponents  $\alpha$  and  $\alpha'$  such that

$$C(T) \approx (A/\alpha) [(\Delta T/T_c)^{-\alpha} - 1] + \dots \quad (T \rightarrow T_c^+) \quad (35)$$

For  $T \rightarrow T_c^-$  the exponent  $\alpha$  is replaced by  $\alpha'$ ;  $\alpha \rightarrow 0$  corresponds to a logarithmic divergence. For most real fluid and magnetic systems  $\alpha$  and  $\alpha'$  lie in the range 0 to 0.1. A second exponent  $\beta$  characterizes the way in which the spontaneous order vanishes at  $T_c$ , namely,

$$\Psi_0 \sim (\rho_L - \rho_G)/\rho_c \sim M_0(T) \sim (|\Delta T|/T_c)^\beta \quad (T \rightarrow T_c^-) \quad (36)$$

(As before  $\rho_L$  and  $\rho_G$  are the densities of coexisting liquid and gaseous phases while  $M_0$  is the spontaneous magnetization.) It has been found experimentally that  $\beta$  lies in the range 0.33 to 0.36, that is, quite close to  $\frac{1}{3}$ , for a wide range of gases, binary fluids, magnets, etc. A particularly careful and precise measurement for the antiferromagnet  $MnF_2$  by Heller and Benedek [22] gave  $\beta = 0.335 \pm 0.005$ . (We may refer further to recent reviews of the experimental and theoretical situation [22-24].)

The third letter of the Greek alphabet is reversed to describe the divergence of the initial isothermal susceptibilities

$$\chi_T = \left( \frac{\partial \Psi}{\partial \xi} \right)_{T, \xi=0+} \sim \left( \frac{\partial M}{\partial H} \right)_{T, H=0+} \sim K_T = \frac{1}{\rho} \left( \frac{\partial \rho}{\partial p} \right)_{T, \mu=\mu_t^+} \quad (37)$$

or, equivalently, the total fluctuation in the order parameter, as the critical point is approached. Thus we suppose

$$\chi_T \sim (\Delta T / T_c)^{-\gamma} \quad T \rightarrow T_c + \quad (38)$$

and similarly in terms of  $\gamma'$  for  $T < T_c$ . Experimentally magnetic materials are found to have values of  $\gamma$  clustering around  $\gamma = \frac{4}{3}$ . Amusingly, the measurements first used to demonstrate this fact were made nearly forty years earlier by Pierre Weiss [25]. For fluids and other materials values in the range 1.2 and 1.3 are again found.

A variety of other exponents have been defined and studied -  $\delta$ ,  $\Delta_k$ ,  $\Delta'_k$  ( $k = 1, 2, 3 \dots$ ) for other thermodynamic derivatives and  $\eta$ ,  $\nu$ ,  $\nu'$  etc. for the correlation functions. I will not list all the definitions here (the reader may consult Refs [23, 24] but will remark that a convenient and mathematically precise general definition of an exponent  $\lambda$ , for which we write  $f(x) \sim x^\lambda$  as  $x \rightarrow 0+$ , is provided by

$$\lambda = \lim_{x \rightarrow 0+} \frac{\ln f(x)}{\ln x} \quad (39)$$

In practice, of course, the exponent must characterize the behaviour of  $f(x)$  reasonably well for an accessible range of positive  $x$  if it is to be measurable.

## 10. CALCULATION OF CRITICAL EXPONENTS

Having recognized the existence of critical exponents and having noted that they appear to have at least some degree of "universality" we will inquire into their theory. The first observation to make is that the classically accepted theories of critical point phenomena - the theory of van der Waals, mean-field theories of magnetism, Landau's general treatment of "second order transitions", etc. - make quite incorrect predictions. Specifically all these theories assert that the exponents are integral multiples of  $\frac{1}{2}$ . Thus the coexistence or spontaneous magnetization curve is expected to be parabolic with  $\beta = \frac{1}{2}$  (in place of the observed  $\beta \approx \frac{1}{3}$ ); the susceptibility is expected to diverge like a simple pole with  $\gamma = 1$  (in place of  $\gamma \approx 4/3$ ). Similarly the specific heat is predicted to have a jump discontinuity in place of an infinite divergence. (Some modifications of classical theory lead, equally incorrectly, to  $\alpha = \alpha' = \frac{1}{2}$ .) Other examples are the predictions  $\delta = 3$ ,  $\nu = \frac{1}{2}$  in place of the observed values  $\delta \approx 4.5$ ,  $\nu \approx \frac{2}{3}$ .

One of the important factors leading to our present rejection of the classical approaches on a theoretical (as against an experimental) footing was Onsager's solution [14] of the plane Ising model, already referred to. Apart from the logarithmically divergent specific heat Onsager found that  $\beta = \frac{1}{6}$  - a result quite different from both the classical prediction  $\beta = \frac{1}{2}$  and the observations, on three-dimensional systems, of  $\beta \approx \frac{1}{3}$  [23]. From Onsager's work one can also conclude strictly that  $\gamma' \geq 1\frac{3}{4}$  and  $\delta \geq 15$ , where the equalities are almost certainly realized [23].

A particularly significant feature of all the results on plane Ising models is that the natures of the critical singularities i.e. the values of the exponents, are independent (a) of the lattice structure, provided the dimensionality remains  $d = 2$ , and (b) of the spatial isotropy or anisotropy of the interactions, provided the lattice remains "connected". These observations lead naturally to the conjecture that the critical singularities are determined mainly by the dimensionality and the short-range character of the forces, and only to a lesser extent by the details of the Hamiltonian. This idea is borne out fairly well by the further theoretical developments I shall now describe.

It is clear that an analytical solution of the three-dimensional Ising model would be very instructive. But it seems most unlikely to be forthcoming in this generation. Fortunately, however, the development of new theoretical methods has appreciably reduced the need for such an exact solution (although it would always be welcome!). I refer specifically to techniques pioneered by Domb, Sykes and their collaborators, based on a numerical analysis of sufficiently long power series expansions [23]. As a concrete example, consider the high temperature expansion of the susceptibility of the three-dimensional simple cubic Ising ferromagnet in powers of the variable  $v = \tanh(J/kT)$ , namely,

$$\begin{aligned} kT\chi_T &= 1 + 6v + 30v^2 + 150v^3 + 726v^4 + \dots \\ &\dots + 8306862v^{10} + 38975286v^{11} + \dots \end{aligned} \quad (40)$$

Needless to say, special graph-theoretical and combinatorial mathematics have had to be developed in order to compute correctly such lengthy perturbation expansions; high speed digital computers are a great help but by no means render the problem trivial. If one examines the coefficients  $a_n$  of this and similar power series, one finds they vary in a regular fashion; in particular the ratios of successive coefficients satisfy very closely the "law"

$$\mu_n = a_n/a_{n-1} = \mu_\infty [1 + (g/n) + \dots] \quad (41)$$

for the "moderately large" values of  $n$  available. If this asymptotic behaviour is presumed to continue for all  $n$  it is easily seen to imply that

$$\chi_T(T) \sim [1 - (T/T_c)]^{-\gamma} \quad \text{as } T \rightarrow T_c \quad (42)$$

where  $T_c$  and  $\gamma$  are determined by

$$v_c = \tanh(J/kT_c) = 1/\mu_\infty \quad \text{and} \quad \gamma = 1 + g \quad (43)$$

This extrapolation procedure can be tested on the two-dimensional Ising lattices where it is strikingly successful; the exact critical point  $kT_c/J$  is indicated to within a few parts in  $10^4$  while the exponent is determined

correctly to within 0.3 to 0.5%. With this justification one feels confident in applying the method in three dimensions where the apparent precision is quite comparable. One discovers that the simple cubic, body centered cubic face cubic lattice and tetrahedral Ising lattices all yield  $\gamma = 1.250 \pm 0.003 \approx 1\frac{1}{4}$ . The same result is also found if second or third neighbour interactions are included [23]. The method may be applied to other functions: thus for the specific heat Sykes and co-workers [26] have recently found  $\alpha = 0.125 \pm 0.015 \approx \frac{1}{8}$ ; it also applies to other models; in particular for the Heisenberg model of a ferromagnet one obtains  $\gamma \approx 1.37$  (with, it seems at present, some dependence on the spin) [23]. This last figure is in markedly better accord with experiments on ferromagnets than the long-standing Curie-Weiss prediction  $\gamma = 1$ .

The "ratio method" of analysing a power series explained above breaks down for irregular series like

$$\begin{aligned} M_0 = & 1 - 2x^{12} - 24x^{22} + 26x^{24} + 0 + 0 \\ & - 48x^{30} - 252x^{32} + 720x^{34} - \dots \\ & \dots + 631608x^{64} - 9279376x^{66} + \dots \end{aligned} \quad (44)$$

which represents the spontaneous magnetization of the f. c. c. Ising ferromagnet in terms of  $x = \exp(-4J/kT)$ . In such cases, however, Padé approximant techniques as introduced by Baker [27] have proved successful in continuing a power series beyond its radius of convergence and along the real axis up to the critical point singularity. By numerical evaluation of functions such as

$$\beta^*(T) = (T_c - T)(\partial/\partial T) \ln M_0(T) \rightarrow \beta \quad \text{as} \quad T \rightarrow T_c^- \quad (45)$$

it is estimated that  $\beta = 0.312 \pm 0.004 \approx 5/16$ . This is much closer to the values of  $\beta$ , around 0.33 to 0.36, observed for real fluids and magnets, than the classical result or than the value  $\beta = \frac{1}{8}$  for the planar Ising models. The importance of dimensionality is clearly evident.

The residual experimental-theoretical discrepancy of about 0.02 to 0.04 can be attributed to the fact that the lattice gas or ferromagnet Ising models certainly do not do full justice to the more "continuous" aspects of real fluids and magnets. On the other hand the discrete nature of the Ising model should be a rather better representation of binary metallic alloys such as  $\beta$ -brass; accordingly it is gratifying that recent experiments by Als-Nielsen and Dietrich [28] (using neutron diffraction) yield  $\gamma = 1.25 \pm 0.01$  and  $\beta = 0.305 \pm 0.010$  in complete agreement with the calculations for the three-dimensional Ising model.

## 11. THEORIES OF THE CRITICAL POINT

The numerical calculations of the critical point exponents just described are very satisfactory in as far as they reveal the properties of various models and successfully correlate with experiment. Obviously, however, one would like a "deeper" theory that gave analytic expressions for the various exponents or, at least, related one with another. A first attempt

in this direction may be based on the droplet picture of condensation we sketched above (in section 8) [9, 29]. If the approximations embodied in (33) and (34) are extended to higher fluid densities and temperatures (where they are certainly less well justified) one finds they predict the existence of a critical point (although a description of the liquid side for  $T \leq T_c$  is understandably lacking). The corresponding critical point exponents  $\alpha$ ,  $\beta$ , etc. turn out to be simple bilinear functions of the surface exponent  $\sigma$  and the geometrical exponent  $\tau$  introduced in the analysis of the droplet partition function. By elimination of  $\sigma$  and  $\tau$ , relations between the critical point exponents are discovered. The first of these [29] was

$$\alpha' + 2\beta + \gamma' = 2 \quad (46)$$

Happily this relation is confirmed exactly by the results for the plane Ising models; within the numerical uncertainties it is valid also in three dimensions. (Indeed the formula is verified by the classical exponent values.) Other relations such as  $\alpha = \alpha'$ ,  $\gamma = \gamma'$  and

$$\alpha' + \beta(1 + \delta) = 2 \quad (47)$$

also hold exactly in two dimensions and seem accurate in three dimensions (although there are some doubts concerning the symmetry of the exponents above and below  $T_c$ ). It is interesting that the relations (46) and (47) were later proved rigorously as inequalities (with  $\geq$  replacing  $=$ ) [30].

According to the droplet picture the surface exponent  $\sigma$  should be close to  $\frac{1}{2}$  for  $d = 2$  and close to  $\frac{2}{3}$  for  $d = 3$ . In terms of the thermodynamic exponents one finds  $\sigma = 1/(\beta + \gamma')$ ; for the plane Ising models this yields  $\sigma = 8/15 = 0.533$  while for the three-dimensional models  $\sigma \approx 0.61$  to 0.64. The agreement with the expected values is quite encouraging; but clearly the droplet picture is a gross simplification of the true situation even in a lattice gas, and the discrepancies although small, should be understood. Furthermore the value of the second exponent  $\tau$  is somewhat obscure even heuristically [9].

A more general, if rather less physical, hypothesis has been advanced by Widom [31] and others [23, 24]. To explain this thermodynamic scaling or homogeneity hypothesis consider the equation of state of a ferromagnet, namely,

$$M = \mathcal{M}(H, T) \quad (48)$$

from which the free energy can be found, up to a non-singular contribution, by integration. In place of a function  $\mathcal{M}(H, T)$  of two variables we postulate that as  $T \rightarrow T_c$  and  $H \rightarrow 0$  the equation of state can be written in the scaled form

$$\frac{M}{m(T)} = F_* \left( \frac{H}{h(T)} \right) \quad (49)$$

where

$$m(T) \sim |\Delta T|^\beta \quad \text{and} \quad h(T) \sim |\Delta T|^\Delta \quad (50)$$

The two parts of the function  $F$  apply for  $T > T_c$  and  $T < T_c$ , respectively, but must satisfy a matching condition at large arguments. An equivalent postulate is the assertion that the "singular part" of the free energy  $A(M, T)$  is a homogeneous function of degree  $2-\alpha$  in the variables  $\Delta T$  and  $|M|^{1/\beta}$ .

The two exponents  $\beta$  and  $\Delta$  (or  $\beta$  and  $2-\alpha$ ) serve to determine all the other thermodynamic exponents and, indeed, lead to precisely the same exponent relations (46), (47), as before. However, no indication as to the exponent values is given. On the other hand the general scaling hypothesis (49) is directly susceptible to experimental test; it has in fact been rather successful in reducing data taken on many isotherms to a single universal plot. We may illustrate this with the recent experiments of Kouvel and Comly on nickel which are shown in Fig. 6 [32]. Further tests have been [33] and will be forthcoming but at this stage the thermodynamic homogeneity postulate looks very promising. Unfortunately comparatively little success has been had theoretically in attempting to justify it on any grounds much deeper than its natural appeal!

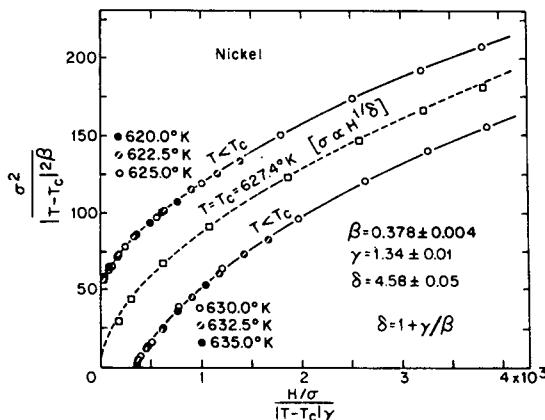


FIG. 6. A scaled plot of six near critical isotherms of nickel verifying the homogeneity hypothesis and exponent relations. Note that  $\sigma$  is used to denote the magnetization  $M$  (after Kouvel and Comly [32]).

## 12. THE PROPAGATION OF CORRELATION

So far in our survey of the theory of critical point phenomena I have focused attention on the thermodynamic observables. In many respects, however, the most interesting and fundamental aspect of a continuous phase transition concerns the behaviour of the microscopic fluctuations as represented in particular by the pair correlation functions  $G(\vec{r})$  or  $\Gamma(\vec{r})$ , defined in (15). The leading theoretical development has been the recognition of the importance of a correlation length or range,  $\xi = \kappa^{-1}$ , which characterizes the distance over which the net correlation functions,  $G(\vec{r}) - G(\infty)$  or  $\Gamma(\vec{r}) - \Gamma(\infty)$ , decay (substantially) to zero. In most disordered, and "non-isotropic" ordered systems with short range forces this decay is asymptotically exponential,  $\sim e^{-\kappa r}$ , as predicted by the original Ornstein-Zernike discussion [23]. In more abstract terms  $\kappa$ , the inverse range,

may be expressed in terms of the Fourier transform  $\hat{G}(\vec{k})$  of the net correlation function; if  $|\vec{k}| = k = k_\alpha$  is the location of the singularity of  $\hat{G}(k)$  in the complex  $k$ -plane which lies closest to the real axis then we have  $\kappa = \text{Im}\{k_\alpha\}$ . In the neighbourhood of (but not at) the critical point this singularity will generally be a simple pole, or rather, a conjugate pair of poles,  $1/(k^2 + \kappa^2)$ . Physically this corresponds to a simple Laplacian, or Coulomb-Yukawa-like law for the propagation of the correlations at long distances, i.e. the correlation is transferred via a local kernel; this is natural and "intuitive". An analogy may be made with field theory in which  $\hat{G}(\vec{k})$  corresponds to the propagator of a particle of mass  $m = \kappa$ .

Now the fluctuation relation or "sum-rule"

$$\chi_T = - \frac{\partial^2 F}{\partial \xi^2} \propto \int [G(\vec{r}) - G(\infty)] d\vec{r} \quad (51)$$

shows that if  $\chi_T \rightarrow \infty$  as the critical point is approached (which is so in all cases of practical interest), the range of correlation  $\xi$  must diverge to infinity. In a magnet for  $H=0$ , or a fluid at  $\rho=\rho_c$  etc., exponents  $\nu$ , and  $\nu'$ , are defined by  $\xi \sim \Delta T^{-\nu}$  as  $T \rightarrow T_c +$ , etc. At the critical point itself  $\xi = \infty$  and, therefore, the decay of correlation can no longer be exponential – rather, some form of power law is to be expected. The usual classical theories (Ornstein-Zernike and Landau theories, the random-phase approximation, etc.) suggest that the singularity remains a simple pole (or rather a coincident pair of poles) so that  $\hat{G}_c(k) \sim k^{-2}$  as  $k \rightarrow 0$ . For a three-dimensional system this implies that  $G(r) \sim 1/r$  as  $r \rightarrow \infty$ ; in other words the propagation of correlation is expected to remain Laplacian, or Coulomb-like, even at the critical point itself. (We may note that in field-theoretic terms the critical point corresponds to massless particles.)

This is a very far-reaching conjecture; and indeed there is increasingly strong evidence that for many systems it is incorrect! Without prejudicing the question we may write, for the correlations at the critical point [23]

$$\hat{G}_c(k) \sim 1/k^{2-\eta} \quad k \rightarrow 0 \quad \text{or} \quad G_c(r) \sim 1/r^{d-2+\eta} \quad r \rightarrow \infty \quad (52)$$

This defines exponent  $\eta$  which takes the value zero in the classical theories. Our first observation is that the exact results of Onsager and Kaufman on the correlation functions of the square lattice Ising model, since greatly extended by others [23], show unequivocally that  $\eta = \frac{1}{4}$  in that case. A numerical study of the three-dimensional cubic Ising lattices [34] indicates  $\eta \approx 0.055$  with an uncertainty of about 0.015 or smaller. This value of  $\eta$  is rather small but the experimental data on  $\beta$ -brass [28], which agreed closely with the Ising values of  $\gamma$  and  $\beta$ , seem to confirm it, as do more recent experiments on  $\text{Fe}_3\text{Al}$  [35]. Other neutron scattering experiments on magnetic materials – a good example is  $\text{RbMnF}_3$  [36] – also indicate comparable positive values of  $\eta$ .

An interesting light on this question is thrown by some inequalities first advanced by Buckingham and Gunton [37] which state

$$\eta \geq 2 - d\gamma' / (\gamma' + 2\beta) \quad \eta \geq 2 - d(\delta - 1) / (\delta + 1) \quad (53)$$

The proofs rest on some very general monotonicity and positivity properties of  $G(\vec{r})$  which are, in fact, quite rigorous for general ferromagnetic Ising models [38]. One consequence of (53) is that the classical exponent values  $\beta = \frac{1}{2}$ ,  $\gamma' = 1$ ,  $\delta = 3$ , are inconsistent with the Coulomb-like Ornstein-Zernike result  $\eta = 0$  in any dimensionality less than four! Furthermore, for many real systems studied experimentally one finds  $\delta \leq 4.7$  which, through (53), implies  $\eta \geq 0.05$ .

Our present theoretical understanding of this non-Laplacian, effectively non-local, propagation which sets in as the critical point is approached, is far from complete. Ideas which parallel those of the thermodynamic homogeneity or scaling hypotheses have, however, been developed. In the first place it is natural to postulate that the correlation length  $\xi$  is the only relevant length (up to multiplicative factors) which diverges at  $T_c$  and thence that the variation of  $G(\vec{r})$  scales with  $\xi$  in the neighbourhood of the critical point [23, 24]. From this we may deduce, for example, the relations

$$\gamma = (2 - \eta)\nu \quad \text{and} \quad \gamma' = (2 - \eta)\nu' \quad (54)$$

(the first of which has since been proved as an inequality for Ising models [4]). Together with the thermodynamic scaling hypotheses we then achieve what we may call a "three-exponent theory" of a critical point – all thermodynamic and pair correlation exponents can be expressed in terms of the trio:  $\beta$ ,  $\gamma$ , and  $\eta$ . This theory is confirmed by all known exact results for planar Ising models and correlates well with the numerical work on three-dimensional models, and experiments on real systems of the sort already referred to [23, 24].

A reduction of this "three-exponent theory" to a "two-exponent theory" has been achieved independently by Widom [31], Kadanoff [24, 39], Patashinskii and Pokrovskii [40], and Stell [41] in a very interesting series of developments. Perhaps the simplest way of characterizing these ideas is to say that the typical fluctuations of the order parameter within a volume in the system of diameter equal to the correlation length  $\xi$ , are held to have a magnitude which scales like the order parameter itself [23]. A direct consequence of this conclusion (or hypothesis) is that the inequalities (53) should, in fact, hold as equalities. Consequently once the dimensionality  $d$  is given, all critical exponents can be determined from a pair of them, say,  $\gamma$  and  $\delta$ .

Most encouragingly this two-exponent  $d$ -dependent theory seems to be completely correct for the planar Ising models. It runs into difficulties, however, with the numerical studies of the three-dimensional models where the rather firm estimates  $\delta \geq 5$  and  $\eta \geq 0.04$  are inconsistent with an inequality in (53). There are also theoretical difficulties concerning the situation for  $d > 4$  where other arguments suggest that all the classical exponent values should be achieved. The present experimental situation is not clear-cut (but, in as far as the data for  $\beta$ -brass and  $\text{Fe}_3\text{Al}$  are in close accord with the Ising model calculations for  $d = 3$ , they also raise doubts).

At this stage we are faced with a disagreement between non-rigorous theories and inexact calculations and experiments. For the present, at least, the question must remain open as a challenge to further thought and experiment. Ultimately we should hope for further reductions to a "one-

exponent theory" and thence to a complete theory in which all the exponents would be calculated from first principles, that is, from the Hamiltonian with its given dimensionality, symmetries and interactions.

### 13. CONCLUDING REMARKS

It has been my intention in surveying the statistical mechanics of phase transition and critical phenomena to convince you that this branch of physics, although nearly a century old, has not lost its vitality. While, on the one hand, the connection with rigorous, not to say pure mathematics has been secured and strengthened, the link with the subtle, and recalcitrant behaviour of real matter is still strong and actively developing. There are those who have foretold the early death of macroscopic physics as a satisfying intellectual discipline; I believe, however, that there still remains, for this field of science, a generation or so of fruitful life.

### A C K N O W L E D G E M E N T S

The support of the National Science Foundation and the Advanced Research Projects Agency through the Materials Science Center at Cornell University is gratefully acknowledged.

### R E F E R E N C E S

- [1] RUELLE, D., *Helv.phys.Acta* 36 (1963) 183, 789; FISHER, M.E., *Archs ration.Mech.Analysis* 17 (1964) 377.
- [2] DYSON, F.J., LENARD, A., *J.Math.Phys.* 8 (1967) 423; DYSON, F.J., *ibid.* 8 (1967) 1538.
- [3] GRIFFITHS, R.B., (in press).
- [4] FISHER, M.E., (unpublished).
- [5] VAN HOVE, L., *Physica* 16 (1950) 137.
- [6] GALLAVOTTI, G., MIRACLE-SOLE, S., RUELLE, D., *Physics Lett.* 26A (1968) 350.
- [7] See LANDAU, L.D., LIFSHITZ, E.M., *Statistical Physics*, Pergamon Press, London (1958) 482.
- [8] RUELLE, D., *Commun. Math.Phys.* 9 (1968) 267.
- [9] FISHER, M.E., *Physics* 3 (1967) 255.
- [10] FISHER, M.E., *J.chem.Phys.* 45 (1966) 1469.
- [11] MERMIN, N.D., WAGNER, H., *Phys.Rev.Lett.* 17 (1966) 1133; MERMIN, N.D., *J.Math.Phys.* 8 (1967) 1061.
- [12] HOHENBERG, P.C., *Phys.Rev.* 158 (1967) 383.
- [13] STANLEY, H.E., KAPLAN, T.A., *Phys.Rev.Lett.* 17 (1966) 913.
- [14] ONSAGER, L., *Phys.Rev.* 65 (1944) 117.
- [15] LIEB, E.H., *Phys.Rev.Lett.* 18 (1967) 692; 19 (1967) 108, 1046; *Phys. Rev.* 162 (1967) 162.
- [16] SUTHERLAND, B., *Phys.Rev.Lett.* 19 (1967) 103; YANG, C.P., *Phys. Rev.Lett.* 19 (1967) 586; SUTHERLAND, B., YANG, C.N., YANG, C.P., *Phys.Rev.Lett.* 19 (1967) 588.
- [17] GRIFFITHS, R.B., *Phys.Rev.* 136 (1964) A437.
- [18] DOBRUSHIN, R.L., *Proc.Berkeley Symp. on Probability Theory and Statistics* (1967) 73; MINLOS, R.A., SINAI, Ya.G., *Soviet Phys. Dokl.* 12 (1968) 688.
- [19] FISHER, M.E., *Phys.Rev.* 162 (1967) 480.
- [20] KAC, M., UHLENBECK, G.E., HEMMER, P.C., *J.Math.Phys.* 4 (1963) 216, 229.
- [21] LEBOWITZ, J.L., PENROSE, O., *J.Math.Phys.* 7 (1966) 98.
- [22] HELLER, P., *Rep.Prog.Phys.* 30 (1967) 731.
- [23] FISHER, M.E., *Rep.Prog.Phys.* 30 (1967) 615.
- [24] KADANOFF, L.P., et al., *Rev.mod.Phys.* 39 (1967) 395.

- [25] KOUVEL, J.S., FISHER, M.E., Phys.Rev. 136 (1964) A1626; WEISS, P., FORRER, R., Annls Phys. 5 (1926) 153.
- [26] SYKES, M.F., MARTIN, J.L., HUNTER, D.L., Proc.phys.Soc. 91 (1967) 671.
- [27] BAKER, G.A., Jr., Adv.Theor.Phys. I (1965).
- [28] ALS-NIELSEN, J., DIETRICH, O., Phys.Rev. 153 (1967) 706, 711, 717.
- [29] ESSAM, J.W., FISHER, M.E., J.chem.Phys. 38 (1963) 802.
- [30] RUSHBROOKE, G.S., J.chem.Phys. 39 (1963) 842; GRIFFITHS, R.B., Phys.Rev.Lett. 14 (1965) 623; J.chem.Phys. 43 (1963) 1958; see also Ref. [23].
- [31] WIDOM, B., J.chem.Phys. 43 (1965) 3892, 3898.
- [32] KOUVEL, J.S., COMLY, J.B., Phys.Rev.Lett. 20 (1968) 1237.
- [33] GREEN, M.S., VICENTINI-MISONI, M., LEVELT-SENGERS, J.M.H., Phys.Rev.Lett. 18 (1967) 1113.
- [34] FISHER, M.E., BURFORD, R.J., Phys.Rev. 156 (1967) 583; these calculations are being extended by M.A. MOORE, M. WORTIS and D. JASNOW, who still find a positive, although slightly lower, value for  $\eta$ .
- [35] GUTTMAN, L., SCHNYDERS, H.C., (to be published).
- [36] CORLISS, L.M., DELAPALME, A., HASTINGS, J.M., NATHANS, R., J.appl.Phys. 40 (1969) (in press).
- [37] BUCKINGHAM, M.J., GUNTON, J.D., Phys.Rev.Lett. 20 (1967) 143; Phys.Rev. (1968).
- [38] GRIFFITHS, R.B., J.Math.Phys. 8 (1967) 478, 484; and more recent work.
- [39] KADANOFF, L.P., Physics 2 (1966) 263.
- [40] PATASHINSKI, A.Z., POKROVSKI<sup>II</sup>, V.L., Zh.éksp.teor.Fiz. 50 (1966) 439 (transl. Soviet Phys. JETP 23 (1966) 292).
- [41] STELL, G., 1965 (unpublished; Phys.Rev.Lett. 20 (1968) 533).
- [42] FAIRBANK, W.M., BUCKINGHAM, M.J., KELLERS, C.F., Proc.5th Int.Conf.Low-temperature Physics, Madison, Wisconsin (1957) 50.

# MACROSCOPIC COHERENCE AND SUPERFLUIDITY

P. W. ANDERSON\*

Cavendish Laboratory,  
University of Cambridge,  
Cambridge, United Kingdom

## Abstract

MACROSCOPIC COHERENCE AND SUPERFLUIDITY. The problem of the coherence in a superfluid, namely a substance which possesses ODLRO in the Yang sense, is treated, with a particular account of the problem of dissipation in the superfluid. The relation between dissipation and quantized vortices is also explained, both for superfluids and superconductors.

My feelings at this point are best expressed by describing a cartoon which appeared in the New Yorker a number of years ago. It showed a father whose little boy, dressed in pyjamas, was sitting on his knee, looking up at him trustingly, and saying "Daddy, tell me again the story of how jazz came up the river from New Orleans". I have been asked to retell on this occasion our success story in the theory of superfluidity, a story which I believe is now mostly a source of fruitful ideas for other fields. I should like, in addition, to touch some of the boundary points of our knowledge, especially where the problems are of some real theoretical interest.

It seems most likely that almost all non-magnetic systems (at least) which have enough zero-point energy to remain fluid in their ground-states will, in the end, be found to be superfluids, in the sense that they spontaneously transform from the so-called normal Bose or Fermi liquid state into a state which has a certain definite type of correlation. It can certainly not be so proven, because we cannot imagine all the other types of possible ground-states for many-body systems which might have yet lower energy. The motivation for this ordering can be expressed in terms of what I shall say shortly, but a much deeper microscopic discussion, at least for metals, is given by Schrieffer in these Proceedings.

We thus just define a superfluid as a substance which has the type of correlation exhibited by liquid helium (postulated by Penrose and Onsager) and superconductors (BCS and Gor'kov) and named ODLRO by Yang:

$$\langle \psi^\dagger(\mathbf{r}) \psi(\mathbf{r}') \rangle = F^\dagger(\mathbf{r}) F(\mathbf{r}') + \text{short-range terms}$$

In the case of a metal we must define  $\psi = \psi_{\mathbf{e}\sigma} \psi_{\mathbf{e}-\sigma}$ , a field operator for pairs of electrons, not single particles as in helium.

This definition is compatible with a number of prejudices many-body theorists have since discarded, especially that the phase of the field operator is not a physical quantity and that coherence cannot exist between states of different particle numbers, since this mean value conserves N. However, this formulation is most inconvenient, if barely usable, for discussing the

\* Present address: Bell Telephone Laboratory, Murray Hill, N.J., United States of America.

coherence phenomena of superfluidity. In studying coherence phenomena for light, we use slits, half-silvered mirrors, etc. — all devices for separating out different pieces of light whose relative phase we want to test. In the case of superconductivity and superfluidity the great advance came when we invented some corresponding entities: Josephson junctions and their relatives, and the orifice, which can serve as switches connecting or disconnecting different samples of our particle field at will. Imagine, now, two buckets of superfluid connected by an orifice that can be opened or closed (Fig. 1). If this system has ODLRO, it is easy to see that, if  $r$  is on one side of the orifice and  $r'$  on the other, one must have coherence between two states

$$\psi_{NN'}: N_1 = N_1, N_2 = N' \text{, and } \psi_{N+1, N'-1}$$

because these are the states which are connected by  $\psi^\dagger(r) \psi(r')$ . Thus when we close the door and, at least in principle, can then handle the two parts independently, each part has an internal coherence which cannot be described by ODLRO alone, but which does involve coherence of states of different  $N$  and a real physical meaning for the phase. For instance, an experiment which can be done in principle at least, is to subject the two sides to different gravitational potentials and re-open the door, comparing phases by the resulting instantaneous current.

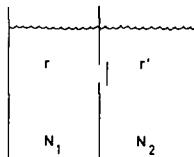


FIG. 1. Two buckets of superfluids connected by an orifice.

We can also get a glimpse of the motivation for the ODLRO from this picture. We can easily understand that when the orifice is open, there is a kinetic energy term in the Hamiltonian which transfers particles from one side to the other

$$\mathcal{H}_K = -T(C_1^\dagger C_2 + C_2^\dagger C_1)$$

and this term can be caused to have a negative mean value by having coherence. This coupling energy is a very central concept in superfluidity.

Although it is possible, by treating one's system as a part of a larger one, to avoid grasping this particular nettle, it seems far more convenient to do so and to satisfy the ODLRO condition by the radical assumption

$$\langle \psi^\dagger(r_1) \rangle = F^\dagger(r_1, t_1) \langle \psi(r_2) \rangle = F(r_2, t_2)$$

i.e. that the phase and amplitude of the quantum particle fields have macroscopic mean values in the superfluid.

What kind of state has this property of macroscopic quantum coherence? Clearly, it has a wave-packet state made up from different  $N$ -values

$$\Psi = \sum_N a_N e^{iN\varphi} \Psi_N$$

where by actual calculations we find

$$\frac{1}{V} \int d\tau \langle \psi(r) \rangle = \frac{1}{V} \left( \sum_N a_{N-1} a_N (\psi_{N-1}, \psi \psi_N) \right) e^{i\varphi}$$

For a perfect Bose gas, if the  $a_N$ 's spread over a wide enough range  $\Delta N \gg 1$ ,  $\langle \psi \rangle \sim (N/V)^{\frac{1}{2}}$ ; for actual superfluids, the value ranges from  $\sim 0.1$  for He to  $\sim 10^{-4}$  for superconductors.

We notice that the transformation coefficient from states with definite  $N$  to definite  $\varphi$  is  $e^{iN\varphi}$ . Except for minor corrections for small  $N$  (which do not alter the rest of this paper), it is precisely valid to treat  $N$  and  $\varphi$  as conjugate dynamical variables, the latter being treated as a true macroscopic statistical parameter of the system in the same sense as  $P$ ,  $V$ ,  $S$ ,  $T$  (and no less abstract than the latter pair). The Hamilton's equations expressing the motion of these two variables are the basic equations of superfluidity

$$\hbar \frac{d(\varphi)}{dt} = \langle \frac{\partial \mathcal{H}}{\partial N} \rangle = \mu$$

$$\hbar \frac{dN}{dt} = -\langle \frac{\partial \mathcal{H}}{\partial \varphi} \rangle$$

The second equation is the general equation ensuring the presence of supercurrents. If, for instance, there is indeed a coupling force across our orifice holding the phases of the two buckets together, a free energy

$$F \propto \frac{U_0(\varphi_1 - \varphi_2)^2}{2}$$

there is then the possibility of a current between them

$$\frac{dN_1}{dt} = -\frac{dN_2}{dt} = -U_0(\varphi_1 - \varphi_2)$$

It seems clear that supercurrents exist if and only if one has phase coupling: the phase is the only variable conjugate to  $N$ . This is the nearest thing there is to a rigorous characterization of superfluidity.

For electrons, the charge  $2eN$  is coupled to the electromagnetic field and the phase alone is not a gauge-invariant quantity but  $\nabla\varphi - 2eA/\hbar c$ . Thus supercurrents can flow in response to magnetic fields as well as to phase gradients. Incidentally, by a definition which can be checked at  $T = 0$  and for homogeneous samples against Galilean invariance, we may define, if we like,  $v_s = (\hbar/m)\nabla\varphi$  and  $n_s = \partial^2 U / \partial(\nabla\varphi)^2$  and then  $j = n_s v_s$ . Quantization of flux and of circulation are merely two manifestations of the elementary idea that  $\langle \psi \rangle$  is single-valued and thus  $\oint \nabla\varphi \cdot dS = 2n\pi$ .

It is the other of the superfluid equations which may be more interesting here.

In one form, it is the oldest existing description of superfluidity, i.e. London's acceleration equation

$$m \frac{dv_s}{dt} = m \frac{\hbar}{m} \frac{d}{dt} \nabla \phi = \nabla \mu = \vec{F}$$

This is the basic characteristic of superfluidity: a superfluid cannot sustain a true force — in the sense of a gradient of the thermodynamic potential — without undergoing acceleration. This clearly means that in the absence of vorticity we must have flow without force. Conversely, if we have a difference in chemical potential  $\mu_1 - \mu_2$  between two reservoirs of superfluid, their relation phases must be changing at precisely the rate

$$\frac{d\phi}{dt} = \frac{\mu_1 - \mu_2}{\hbar}$$

Note that, by at least two arguments, we can see that there are no intrinsic corrections to this equation. First, the derivation draws only on the meaning of  $\phi$  and the actual definition of  $\mu$  as  $dE/dn$ . Second, gauge invariance tells us that  $\hbar \partial\phi/\partial t - 2 \text{ eV}$  is the only way in which  $\phi$  can occur physically. Of course, we can make mistakes in measuring either  $\mu$  (temperature or contact potential differences, for instance) or  $d\phi/dt$  (we can count wrong), but that is an experimental problem.

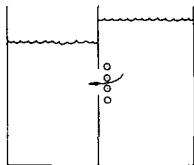


FIG. 2. Phase slippage due to vorticity.

There are two very similar ways the phase change can manifest itself. Most simple is a Josephson device, a connection between superconductors which exerts a coupling energy  $U \propto -\cos(\phi_1 - \phi_2)$  and thus a current  $\propto \sin(\phi_1 - \phi_2)$ . More generally interesting is the possibility of phase slippage due to vorticity. A quantized vortex is a line singularity around which  $\oint \nabla \phi \cdot d\mathbf{s} = 2\pi$ . One can see immediately that every time a vortex passes between two points, it changes their relative phase by  $2\pi$ : so we can sustain a  $\overline{\mu_1 - \mu_2}$  if at the same time we have a rate of motion of vortices of  $dN/dt = (\overline{\mu_1 - \mu_2})/\hbar$  (Fig. 2).

This phenomenon is behind all manifestation of dissipation in superfluids and superconductors. Incidentally, after this equation had been applied to superfluids, we found that it is actually a perfectly respectable — if undiscovered — theorem in classical perfect fluid hydrodynamics.  $\phi$  is simply a quantized version of a velocity potential, and going through precisely the same arguments for potential flow one finds

$$(\text{mean chemical potential difference}) = \text{mean transverse flow of vorticity}$$

Whether this has any hydrodynamic or magnetohydrodynamic applications, I don't know. Incidentally, the superconducting version also has a classical

equivalent:  $d\Phi/dt = E.M.F.$  (expressed in quantized flux units). As in the hydrodynamic case, it is important to realize that the relationship to  $dN/dt = \Delta\mu$  is tricky. This relationship of electromagnetism and hydrodynamics has a fascinating 19th-century ring to it.

The original and most elegant way of demonstrating and measuring the Josephson a.c. equation was suggested by Josephson in 1962 and carried out a year later for superconductors, two years later for superfluids; it uses the phenomena of entrainment of frequency. I want to mention a fact out of a book on non-linear mechanics about this. It appears that entrainment was discovered by Huyghens, who noticed that two clocks hanging on the same wall ran in precise synchronism. More mundanely, anyone familiar with TV knows that a strong a.c. signal applied to a non-linear oscillator has the capability of entraining the frequency of the non-linear oscillator in rather precise synchronism with the external signal, and this is in fact the most accurate way to measure  $\hbar/2e$  for electron pairs, or  $\hbar/m$  for helium atoms, using the Josephson equation. One can induce steps in an I-V characteristic at multiples of  $V = \hbar v/2e$  (Fig. 3). Unfortunately, one can also induce them at subharmonics and on occasion at absolutely arbitrary frequencies in ways which will only be understood when we have the actual non-linear mechanics under more precise control. But it is fairly easy to set up conditions where the intrinsic accuracy of this phenomenon — which is in principle a mechanism, if you like, rather than a measurement in the usual sense, and thus almost infinitely accurate — is far better than other uncertainties such as in absolute measurement of voltage.

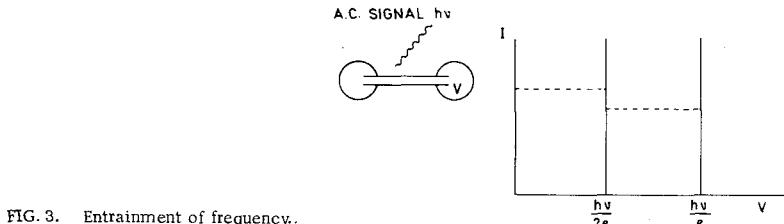


FIG. 3. Entrainment of frequency..

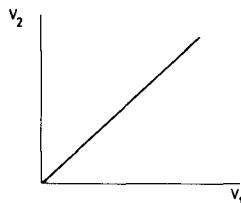


FIG. 4. Experiment of Ivar Giaever.

Let me return now to the statement that dissipation in superfluid flow is always the consequence of the motion of quantized vortices. A very beautiful demonstration of this fact was given recently by Ivar Giaever in one of those experiments which convince on the spot. He superposed two superconducting films in a magnetic field, separating them by an oxide layer which was quite thick enough to provide good insulation, but thinner than the distance  $(B/\Phi_0)^{1/2}$  between quantized matrices. When a current flowed through one (the primary), he observed a voltage in the second practically equal to that in the first: essentially a perfect "d.c. transformer" (Fig. 4).

This is not to imply that these dissipative effects are understood perfectly. In both superfluid helium and superconductors, one of the major theoretical problems remaining is the question of the detailed explanation of dissipation. In the case of helium, our worst problem is to understand how the vortices get there in the first place, a question which has been emphasized by Vinen. The energy of a quantized vortex per unit length is enormous, i.e.  $\sim(\hbar^2/me^3)\ln(R/a)$ , which for a vortex ring of only barely macroscopic size ( $\sim 100 \text{ \AA}$ ) is already tenths of an electron volt (as we verify from the Reif experiment). Thermal fluctuations of this magnitude are hopelessly rare and thus the nucleation of vorticity must be terribly difficult. One can only hope that the experimentalists have so far always been so clumsy as to introduce microscopically rough places in their apparatus or the like at which the nucleation takes place.

Incidentally, I definitely do not wish to imply that all problems of the microscopic theory and the detailed behaviour of liquid helium are solved — quite the contrary. There is a fairly respectable microscopic theory to which the number of contributors is simply too great to mention any of them here, which seems to give  $\sim 20 - 50\%$  accurate answers and nothing seriously wrong in principle, but many details remain to be worked out; I only mention vortex nucleation because it is the biggest puzzle at the moment. One hopeful sign for the microscopic theory is the experimental possibility of an equivalent to superconducting tunnelling — the velocity distribution for evaporation of particles from the superfluid surface, an experiment which John King and students at MIT are carrying out.

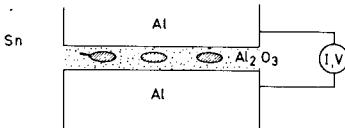


FIG. 5. Experiments on tin particles.

In the case of superconductors, since the work of Bardeen and Stephen on the flow of flux lines, and mine on their creep, the orders of magnitudes of the viscosity and other properties of flux line motion are not unreasonable. But both the qualitative two-fluid picture of Bardeen and Stephen and the beautiful perturbation calculation of Caroli and Maki seem to be hung up quantitatively at least on a rather old dilemma in the field — to what extent do the flux lines flow downstream with the electron gas? This (using our canonical equation saying that potentials are due to transverse vorticity flows) determines the Hall effect, and while it is now apparent both experimentally and theoretically that the Hall effect is of the same order of magnitude as the normal one, the fact is that Bardeen and Stephen give the wrong answer quantitatively, while Caroli and Maki do not answer the question at all.

One of the most interesting limiting cases for the phenomenon of superfluidity is that of very small particles. For nine or ten years we have had one example of superconductivity in very small particles, namely the pairing phenomenon in nuclear matter. However, nuclei are not easy things to work with; we can apply the equivalent to H-fields by rotating them, but we cannot do thermal or coherence experiments on them, nor can we go to the macroscopic limit of nuclear matter. Thus it is nice that Giaever and a co-worker

have come up with a good series of experiments on tin particles in the interesting range  $\sim < 10^2 \text{ \AA}$ . Giaever evaporates very tiny amounts of tin on to an oxidized Al substrate, which then forms separate globules of fairly sharply defined sizes  $\sim 15$  to  $150 \text{ \AA}$ . He then oxidizes again, covers all with Al, and measures the tunnelling characteristic (see Fig. 5).

The question is, of course, what sign is left of the superconductivity of Sn. There are basically two new energy parameters introduced by the smallness of the particles: the electrostatic energy per electron

$$\frac{e^2}{C} = \frac{e^2 d}{\epsilon A}$$

$\sim 10/R^2(\text{\AA}) \text{ eV} \sim 1 \text{ MeV for } 100 \text{ \AA} (10^\circ\text{K})$ , and the granularity of the energy levels,  $E_F a^3/R^3 \sim 100/R^3 \text{ eV}$ .

The first effect of the electrostatic energy is to destroy the coupling between particles — the Josephson current — since the electrostatic energy difference between states of different N will cause  $\Delta N \rightarrow 0$ ,  $\Delta\phi \rightarrow \infty$  when it is greater than the rather weak phase coupling.

One might wonder if electrostatics will prevent internal phase coupling. The idea that the internal coupling could also be broken up appears in the literature, but it is wrong. What is essential is the effective quasi-particle interaction, which includes the ionic and electronic charge clouds — the actual renormalized charge of a quasi-particle is zero, the charge all appearing at the particle surface. Thus indeed the tunnel characteristic shows evidence of superconductivity down to  $\sim 50 \text{ \AA}$ . How can it, when electrostatics alone introduces a gap in the spectrum of order  $10/R^2$ ? Because the Fermi levels of the neutral particles have a random distribution, and some few match the Al matrix perfectly and cancel the gap. Granularity should begin to be very effective at  $\sim 50 \text{ \AA}$ . This is rather a complicated effect and when it is really bad, of course, the superconducting gap becomes meaningless, since there is already a larger gap; but in determining an effective  $T_c$ , it does indeed weaken the tendency for pairs to break up, so  $T_c$  increases more or less as  $\sqrt{(\Delta^2 + \Delta_{\text{gran}}^2)}$  before disappearing (Parmenter). This is indeed observed.

I would like to end by making a few remarks on the generally interesting subject of "blue-sky" superconductors. It seems to be fashionable to write hopeful papers about the possibility of really high-temperature superconductors, preferably made from complicated organic molecules. I have no pipeline to the Deity so cannot pronounce on the existence of such things. I can, however, remark on the three most important points which have not been discussed in these papers.

(1) Organic materials are generally rather loose vibrationally and have large observed electron-vibration couplings. The incoherence introduced by such couplings and their influence on the effective quasi-particle interactions cannot be ignored.

(2) Any pairing strong enough to overcome this as well as the intrinsic "granularity" energy gaps of the order of at least  $0.1 - 1 \text{ eV}$  will have very large chemical effects. In particular, the structural chemistry of a substance with such pairing must be peculiar, and the delicate adjustment of bond lengths and angles which led to the Double Helix structure, for instance,

could not be useful for such a substance: to find an organic superconductor look for a structure which Pauling, Crick, et al. cannot solve.

(3) Large pairing energies may or may not mean superconductivity. The energy gap in He, for instance, is 20 eV, but  $T_c$  is 2°K. Both pairing and the zero-point kinetic energy which propagates the pairs must be present, and  $T_c$  is given by the smaller of the two.

# MICROSCOPIC THEORY OF SUPERCONDUCTIVITY

J. R. SCHRIEFFER

University of Pennsylvania,  
Philadelphia, Pa., United States of America

## Abstract

MICROSCOPIC THEORY OF SUPERCONDUCTIVITY. 1. Introduction; 2. Weak coupling pairing theory; 3. Excitation spectrum; 4. Coherence effects; 5. Strong coupling superconductivity; 6. Ferromagnetism and superconductivity; 7. Non-phononic mechanisms; 8. Conclusion.

## 1. INTRODUCTION

The phenomenon of superconductivity was discovered in 1911 when it was observed that the electrical resistivity of mercury wires vanished below a critical temperature  $T_c$ . Subsequent experiments have shown that roughly half the elements in the periodic table exhibit this phenomenon, with  $0 < T_c < 21^\circ\text{K}$  at present.

Theoretically, the problem involves the understanding of the electron field  $\psi_\sigma$  (fermion) and phonon field  $\phi_\lambda$  (boson) interacting with the coupling

$$H_{\text{int}} = \int \psi_1^\dagger \psi_2^\dagger \frac{e^2}{r_{12}} \psi_2 \psi_1 d\tau_1 d\tau_2 \\ + \int \psi_1^\dagger \psi_1 g(1,2) \phi_2 d\tau_1 d\tau_2 \quad (1.1)$$

Here,  $d\tau$  represents spatial integration plus polarization sum ( $\sigma = \pm \frac{1}{2}$  and  $\lambda = 1, 2, 3$  since the phonons are a vector field). Physically, the phonons are quantized crystal lattice vibrations.

The scale of energies involved is

$$E_F \equiv \text{Fermi energy} \approx 5 - 10 \text{ eV}$$

$$\omega_D \equiv \text{Debye (maximum phonon) energy} \approx 0.01 - 0.05 \text{ eV}$$

$$\text{Correlation energy} = |W_{0S} - W_0^{HF}| \approx 1 \text{ eV/electron}$$

$$\text{Condensation energy} = |W_{0S} - W_{0N}| \approx 10^{-8} \text{ eV/electron}$$

$$k_B T_c \sim 0 - 10^{-3} \text{ eV}$$

where  $W_{0S}$  and  $W_{0N}$  are the ground-state energies in the super and normal phases respectively (the latter being achieved by applying a magnetic field  $> H_{\text{critical}}$  which destroys superconductivity).  $W_0^{HF}$  is the ground-state energy within the Hartree-Fock approximation. Since correlation energies can at present be calculated to an accuracy of about 0.1 eV/electron, we see that a straightforward calculation of the condensation energy from the

difference of  $W_{0S}$  and  $W_{0N}$  is impossible. Thus, one must isolate those correlations which are essential in distinguishing the N and S phases and treat them carefully, the remaining correlations being parametrized in the theory. Fortunately, the Landau theory of Fermi liquids provides a convenient framework for this separation.

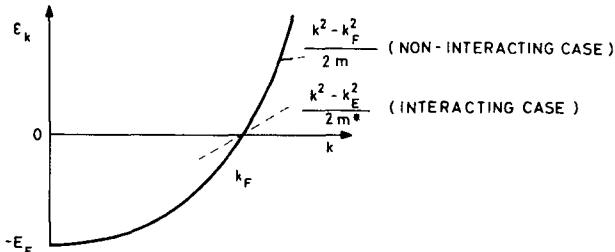


FIG.1. States of the interacting system characterized by a spectrum of one—"quasi-particle" states.

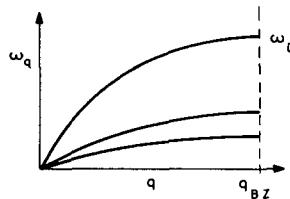


FIG.2. Phonon spectrum for a typical metal.

As discussed by Pines in these Proceedings, the Landau theory shows that the states of the interacting (normal) system can be characterized by a spectrum of one—"quasi-particle" states (or elementary excitations), as shown in Fig.1. We choose the origin of energy so that  $\epsilon_{k_F} = 0$ . The many-body effects renormalize the electron mass, with  $m^*/m \sim 1.1 - 2$  in typical cases. At zero temperature, the quasi-particle occupation numbers are

$$\langle n_{k\sigma} \rangle = \begin{cases} 1 & |k| < k_F \\ 0 & |k| > k_F \end{cases} \quad (1.2)$$

Because of the Pauli principle one can only add quasi-electrons to states  $|k| > k_F$  and add quasi-holes to states  $|k| < k_F$  at  $T = 0$ . The Landau theory breaks down when the spontaneous decay rate of an excitation  $\Gamma_k = 1/2\tau_k$  (due to phonon emission or particle-hole pair production) becomes comparable to the excitation energy  $\epsilon_k$ . The essential point is that for weak coupling superconductors the virtually excited Landau states which enter into making up the actual ground-state of the superconductor involve only low-lying excitations and these lifetime effects are not important.

The phonon spectrum for a typical metal is sketched in Fig.2. The spectrum cuts off at the Brillouin zone boundary  $q_{BZ}$  since there are only  $3N$  phonon modes in the system of  $N$  atoms moving in three dimensions.

There are of course residual interactions between the Landau quasi-particles even in the normal state. Moreover, if the system is superconducting at low temperature, the Landau theory only treats that part of the correlation problem corresponding to the normal state and the correlations (or interactions) involved in the condensation to the superconducting state remain to be handled by non-perturbative means.

Let the irreducible interaction between the Landau excitations be  $V$ . In Born approximation, this would be represented by a screened Coulomb interaction (with appropriately renormalized vertices) plus a one-phonon exchange potential, as shown in Fig. 3.

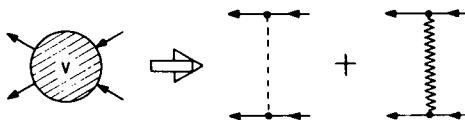


FIG.3. Representation of the irreducible interaction between the Landau excitations, in the Born approximation.

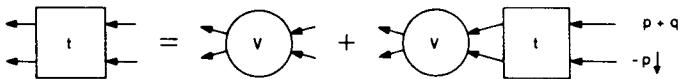


FIG.4. Bethe-Salpeter equation.

If one uses thermodynamic Green's functions

$$\langle \langle \dots \rangle \rangle \equiv \text{Tr } e^{-\beta(H-\mu N)} \langle \dots \rangle / \text{Tr } e^{-\beta(H-\mu N)} \quad (1.3)$$

( $\beta = 1/k_B T$ ), the critical temperature is given by the appearance of a pole at zero centre-of-mass energy and momentum ( $q_0, \vec{q}$ ) in the particle-particle  $t$ -matrix which satisfies the integral equation sketched in Fig. 4, i.e. the Bethe-Salpeter equation. This pairing instability was first discovered by Cooper who showed that the normal state is unstable for some  $T_c > 0$  so long as  $V$  is attractive near the Fermi surface, regardless of its strength. Since  $t$  becomes singular for  $\vec{q} \neq 0$  only for  $T < T_c$ , it seems clear that one should first resolve the  $\vec{q} = 0$  instability before proceeding to the  $\vec{q} \neq 0$  case. We note that spin triplet or singlet states will enter the instability, depending on whether the strongest attraction is in odd  $\ell$  or even  $\ell$  states.

## 2. WEAK COUPLING PAIRING THEORY

We define the  $\vec{q} = 0$  pair creation operator as

$$b_k^\dagger = c_{k\uparrow}^\dagger c_{-k\downarrow}^\dagger \quad (2.1)$$

The  $c_{k\sigma}^\dagger$  and  $c_{k'\sigma'}^\dagger$  create Landau quasi-electrons and quasi-holes. The Hamiltonian which describes the scattering of  $\vec{q} = 0$  quasi-particle pairs is:

$$H_p = \sum_{k, \sigma} \epsilon_{k\sigma} n_{k\sigma} + \sum_{kk'} V_{kk'} b_{k'}^\dagger b_k \quad (2.2)$$

where

$$V_{kk'} = \begin{matrix} k' \uparrow & & k \uparrow \\ & \swarrow \curvearrowright & \swarrow \\ -k' & & -k \downarrow \end{matrix} \quad (2.3)$$

In Eq.(2.2),  $\epsilon_{k\sigma} = k^2 - k_F^2 / 2m^*$ . In the limit of large volume and fixed density of electrons, the ground-state of  $H_p$  can be represented by

$$\Psi_0 = \prod_k (u_k + v_k b_k^\dagger) |0\rangle \quad (2.4)$$

To normalize  $\Psi_0$  we require

$$|u_k|^2 + |v_k|^2 = 1 \quad (2.5)$$

with  $v_k$  being determined below. The state (2.4) is remarkable in that it is not a state of a fixed number of particles  $N$ . If we expand  $\Psi_0$  in terms of its normalized fixed  $N$  projections

$$\Psi_0 = \sum_N a_N \Psi_{0N} \quad (2.6)$$

$a_N$  is non-zero for all even  $N$ . For cases of physical interest, the average number of particles/cm<sup>3</sup> is of the order  $10^{23}$  and we restrict  $\Psi_0$  so that

$$\langle \Psi_0 | N | \Psi_0 \rangle = N_0 \approx 10^{23} \quad (2.7)$$

There is a continuous manifold of ground-states  $\Psi_{0\varphi}$  which can be generated from  $\Psi_0$  according to

$$\Psi_{0\varphi} = \prod_k (u_k + e^{2i\varphi} v_k b_k^\dagger) |0\rangle \quad (2.8)$$

then

$$\Psi_{0N} = \text{const.} \int_0^{2\pi} e^{-2iN\varphi} \Psi_{0\varphi} d\varphi \quad (2.9)$$

The  $\varphi$  (phase) or  $N$  degeneracy of the ground-state simply reflects the fact that the ground-state energy is stationary to first-order changes in  $N$

so that we are free to make up such  $\Psi_0$  states to simplify the diagonalization of  $H$ .

By introducing a Lagrange multiplier  $\mu$  to handle the constraint (Eq. (2.7)), we find  $v_k$  by minimizing  $H_p - \mu N$

$$\frac{\delta}{\delta v_k} \langle \Psi_0 | H_p - \mu N | \Psi_0 \rangle = 0 \quad (2.10)$$

One finds

$$u_k^2 = \frac{1}{2} \left( 1 + \frac{\epsilon_k}{E_k} \right) \quad v_k^2 = \frac{1}{2} \left( 1 - \frac{\epsilon_k}{E_k} \right) \quad (2.11)$$

where

$$E_k = \sqrt{\epsilon_k^2 + \Delta_k^2} \quad (2.12)$$

and the so-called energy gap parameter  $\Delta_k$  satisfies

$$\Delta_k = - \sum_{k'} V_{kk'} \frac{\Delta_{k'}}{2E_{k'}} \quad (2.13)$$

For a simple model, one takes an attractive well in  $k$ -space

$$V_{kk'} = \begin{cases} -V |\epsilon_k| & \text{and } |\epsilon_{k'}| < \omega_D \\ 0 & \text{otherwise} \end{cases} \quad (2.14)$$

Then

$$\Delta_k = \begin{cases} \frac{1}{N(0)V} \omega_D e^{-\frac{1}{N(0)V}} & |\epsilon_k| < \omega_D \\ 0 & |\epsilon_k| > \omega_D \end{cases} \quad (2.15)$$

where  $N(0)$  is the density of states at the Fermi surface. Since  $\Delta$  shows an essential singularity as the coupling constant  $N(0)V \rightarrow 0$ , this result cannot be obtained by perturbation theory.

One finds that the condensation energy is

$$|W_{0S} - W_{0N}| = 2N(0) \Delta^2 = H_0^2 / 8\pi \quad (2.16)$$

where the second equality follows from thermodynamics. Since  $N(0)$  is directly measured by the specific heat and one knows the critical field  $H_0$  experimentally, Eq.(2.16) allows one to determine the gap parameter  $\Delta$  which, as we shall see below, is the energy gap in the excitation spectrum of the superconducting phase. This determination of  $\Delta$  agrees well with more direct measurements of this quantity.

We note that the N-projected pairing state  $\Psi_{0N}$  is given in configuration space by

$$\begin{aligned}\Psi_{0N} = & \mathcal{A} \phi(\vec{r}_1 - \vec{r}_2) \phi(\vec{r}_3 - \vec{r}_4) \cdots \\ & \cdots \phi(\vec{r}_{N-1} - \vec{r}_N) \uparrow_1 \downarrow_2 \uparrow_3 \downarrow_4 \cdots \uparrow_{N-1} \downarrow_N\end{aligned}\quad (2.17)$$

where  $\mathcal{A}$  is the antisymmetrization operator and the pair function  $\phi$

$$\phi(\vec{p}) = \text{const.} \sum_{\vec{k}} \frac{v_{\vec{k}}}{u_{\vec{k}}} e^{i\vec{k} \cdot \vec{p}} \quad (2.18)$$

is the same for all pairs. In this sense, the pairing state corresponds to a Bose condensation into the unique state  $\phi(\vec{r}_i - \vec{r}_j)$ . If we view the pair in terms of the relative co-ordinate  $\vec{p} = \vec{r}_i - \vec{r}_j$  and the centre-of-mass co-ordinate  $\vec{R} = (\vec{r}_i + \vec{r}_j)/2$  we see that  $\phi$  describes a condensation into a zero centre-of-mass momentum state, as in the case of superfluid  $^4\text{He}$ . An essential difference between superconductors and superfluids is the strong overlap of pair in space in the former. The range of  $\phi$  is  $\xi \approx 10^{-4}$  cm and on the average  $10^6$  pairs have their centres-of-mass falling within  $\xi^3$ . Thus, the superconductor is a case of very strong pair overlap in space, a fact which is important in producing a gap in the excitation spectrum of superconductors. Were there no pair-pair overlap and hence no pair-pair interactions, there would be a continuum of excited states, starting at zero energy corresponding to pair drift (rather than pair break-up, which requires a finite energy). It is in fact this very strong pair-pair interaction which suppresses fluctuations in which  $\vec{q} \neq 0$  pairs try to be formed. Since  $\vec{q} \neq 0$  pairs would have an extra phase factor  $e^{i\vec{q} \cdot \vec{R}}$  in their  $\phi$ , their absence corresponds to lack of phase fluctuations of the type considered by Anderson in these Proceedings. If the pairing energy is weak, as it is across a Josephson junction, one can produce phase differences  $\Delta\phi$  and pair currents between two superconductors. We also note that  $\Psi_0$  exhibits "long-range off-diagonal order":

$$\lim_{|\vec{r} - \vec{r}'| \rightarrow 0} \langle \Psi_0 | \psi_{\uparrow}^{\dagger}(r) \psi_{\downarrow}^{\dagger}(r) \psi_{\downarrow}(r') \psi_{\uparrow}(r') | \Psi_0 \rangle = f(r)^* f(r') \quad (2.19)$$

where  $f(r) = \text{const.}$  in this  $\vec{q} = 0$  pairing case.

### 3. EXCITATION SPECTRUM

As for the normal phase, the electronic excitations in the superconducting phase are fermions of momentum  $\vec{k}$ . By using the states

$$\begin{aligned}\Psi_{\vec{k}; \sigma}^{\text{el}} &= c_{\vec{k}; \sigma}^{\dagger} \Psi_0 \\ \Psi_{-\vec{k}; -\sigma}^{\text{hole}} &= c_{-\vec{k}; -\sigma} \Psi_0\end{aligned}\quad (3.1)$$

one finds that

$$-E_{k,\sigma}^{\text{el}} = E_{-k,-\sigma}^{\text{hole}} = \sqrt{\epsilon_k^2 + \Delta_k^2} \equiv E_k \quad (3.2)$$

These excitation spectra are plotted in Figs 5a and 5b.

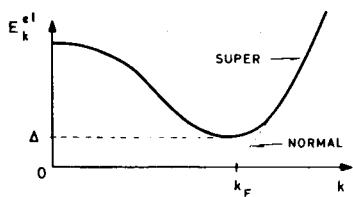


FIG. 5a. Excitation spectrum.

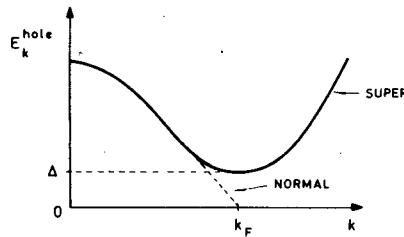


FIG. 5b. Excitation spectrum.

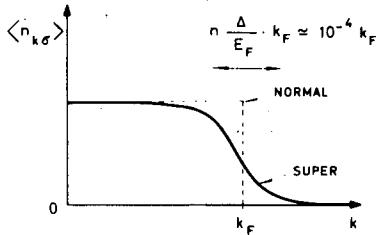


FIG. 5c. Excitation spectrum in the superconducting state.

For electrons, the normal phase spectrum has meaning only for  $|k| > k_F$  and for holes only for  $|k| < k_F$ . In the superconducting state both excitation spectra extend over all  $k$ , reflecting the smearing of the Fermi surface by the pairing interaction as shown in Fig. 5c. The minimum excitation energy is  $\Delta_{k_F} \equiv \Delta$ , so that there is a gap in the excitation spectrum. Standard methods for measuring  $\Delta$  include acoustic attenuation, electromagnetic absorption, electron tunnelling, etc.

By extending the treatment to finite temperature one finds the spectrum is of the same form, except that  $\Delta_k$  satisfies

$$\Delta_k = - \sum_{k'} V_{kk'} \frac{\Delta_{k'}}{2E_{k'}} \tanh \frac{\beta E_{k'}}{2} \quad (3.3)$$

Since  $\Delta_k = 0$  in the normal state,  $T_c$  can also be determined by the largest value of  $T$  for which a non-trivial solution of Eq.(3.3) occurs. This result agrees with that given by the instability analysis discussed above. The electronic specific heat is sketched in Fig. 6 and exhibits a second-order phase transition at  $T_c$ . Critical phenomena of the type discussed by

Fisher in these Proceedings are immeasurably small for bulk superconductors, although fluctuations can be observed in ultra-thin films, since long-range order presumably does not exist in two-dimensional superconductors.

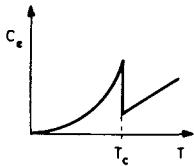


FIG. 6. Electronic specific heat.

#### 4. COHERENCE EFFECTS

In treating dynamical processes in superconductors it is important to note that in addition to a change of density of states in energy for quasi-particle excitations from  $N(0)$  in the normal state to

$$N_s(E) = N(0) \frac{d\epsilon}{dE} = N(0) \frac{E}{\sqrt{E^2 - \Delta^2}} \quad (4.1)$$

if  $\Delta$  is a constant, there is also a change of the one-body matrix elements. Two cases should be distinguished:

Case I: electron-phonon interaction

$$g_{q\lambda} \sum_{k-s} c_{k+q,s}^\dagger c_{k,s} (a_{q\lambda} + a_{q\lambda}^\dagger) \quad (4.2a)$$

Case II: electron-photon interaction

$$\frac{-e}{mc} \sum_{ks} c_{k+q,s}^\dagger c_{k,s} (\vec{r}\vec{k} + \vec{q}) \cdot \vec{A}_q \quad (4.2b)$$

electron-nuclear spin interaction

$$J\psi^\dagger(r) \vec{\sigma} \psi(r) \cdot \vec{I}_r \delta(r - R_r) \quad (4.2c)$$

In case I, the interaction is invariant under time reversal of the electron co-ordinates while it is odd under time reversal in case II. Since the scattering of a thermally excited quasi-particle from  $ks$  to  $k$ 's in the superconducting phase involves a coherent superposition of Landau transitions  $ks \rightarrow k's'$  and  $-k's' \rightarrow -ks$ , we see that the interference effects are of opposite sign in the two cases, being constructive in case II and destructive in case I. These effects are seen in a striking way in the acoustic attenuation rate shown in Fig. 7a where the coherence effect cancels the peak in the density of states (4.1) and the attenuation follows the Fermi distribution

of quasi-particles  $\alpha_S/\alpha_N = 2/(e^{\beta\Delta(\beta)} + 1)$ . For case II, sketched in Fig. 7b, the peak in the density of states leads to an increased value of  $\alpha_S/\alpha_N$  as T drops below  $T_c$ , eventually falling away as the quasi-particles condense out into the superfluid.

Another type of interference effect occurs when an electron is injected with energy E into a superconducting film of thickness L. In this case the electron enters into a linear combination of the degenerate states  $k_1$  and  $k_2$  on opposite sides of the Fermi surface. If  $(k_1 - k_2)L = n\pi$  ( $n$  - integer), a geometrical interference effect occurs and the injection rate is modulated by this interference (the Tomasch effect). These coherence effects give strong support for the detailed nature of the pairing correlations which are central to the theory.

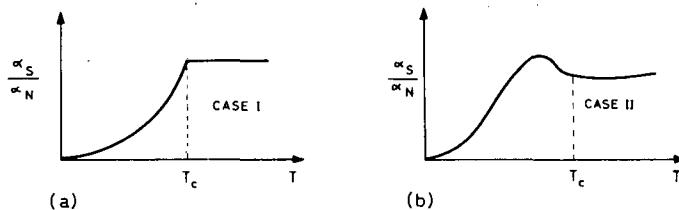


FIG. 7. Acoustic attenuation rates.

## 5. STRONG COUPLING SUPERCONDUCTIVITY

If the electron-phonon coupling is so strong that the Landau quasi-particle level widths  $\Gamma_k$  are comparable to the excitation energies  $\epsilon_k$  for states entering in an important way in representing the superconducting phase, one must resort to field theoretic techniques. This problem has been treated by Nambu and Gor'kov who introduced a two-component spinor field

$$\psi_k = \begin{pmatrix} c_{k\uparrow} \\ c_{-k\downarrow}^\dagger \end{pmatrix} \quad (5.1)$$

and the associated Green's function

$$G_{\alpha\beta}(k, t) = -i \langle T \left\{ \psi_{k\alpha}(t) \psi_{k\beta}^\dagger(0) \right\} \rangle \quad (5.2a)$$

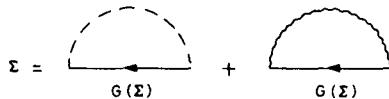
$$G_{11} = -i \langle T \left\{ c_{k\uparrow}(t) c_{k\uparrow}^\dagger(0) \right\} \rangle \equiv G \quad (5.2b)$$

$$G_{12} = -i \langle T \left\{ c_{k\uparrow}(t) c_{-k\downarrow}(0) \right\} \rangle \equiv F \quad (5.2c)$$

etc.

The "anomalous" Green's function F describes the pairing with the expectation value being taken in  $\psi_\varphi$  type states discussed above. Dyson's equation

$$G^{-1} = G_0^{-1} - \Sigma \quad (5.3)$$

FIG. 8. The lowest-order graphs for  $\Sigma$ .

holds for  $G$  where  $\Sigma$  can be calculated by the usual Feynman-Dyson rules if one uses self-consistently determined  $G$ 's in the internal lines. The lowest-order graphs for  $\Sigma$  are shown in Fig. 8. By generalizing a normal phase result of Migdal, Eliashberg showed that higher-order phonon graphs lead to corrections of the order of the speed of sound  $S$  divided by the Fermi velocity  $v_F$ ,  $S/v_F \approx \omega_D/E_F \sim 10^{-2} - 10^{-3}$  and can be neglected, regardless of whether the electron-phonon coupling constant is  $> 1$  or not. This is a truly remarkable result and allows one to solve this strongly coupled field problem to order  $10^{-2}$ , at least as far as the electron-phonon interaction is concerned. If one expands  $\Sigma$  in the Pauli matrices as

$$\Sigma(k) = (1 - Z(k)) k_0 \mathbb{1} + Z(k) \Delta(k) \tau_1 \quad (5.4)$$

(the  $\tau_2$  term can be set equal to zero by a rotation in the pseudo-spin space and the  $\tau_3$  term absorbed into  $\epsilon_k$ ) one finds  $Z$  and  $\Delta$  are given at zero temperature by the generalized energy gap equations

$$\Delta(E) = \frac{1}{Z(E)} \int_0^{\omega_c} \text{Re} \left\{ \frac{\Delta(E')}{\sqrt{E'^2 - \Delta^2(E')}} \right\} K_+(E, E') dE' \quad (5.5a)$$

$$[Z(E) - 1]E = \int_0^{\omega_c} \text{Re} \left\{ \frac{E'}{\sqrt{E'^2 - \Delta^2(E')}} \right\} K_-(E, E') dE' \quad (5.5b)$$

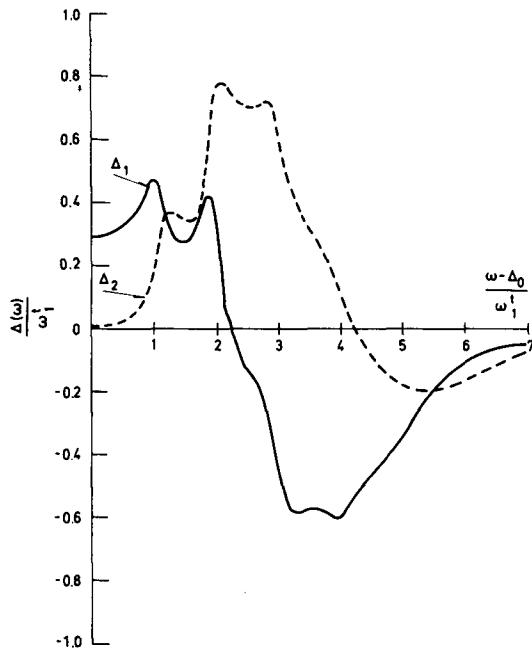
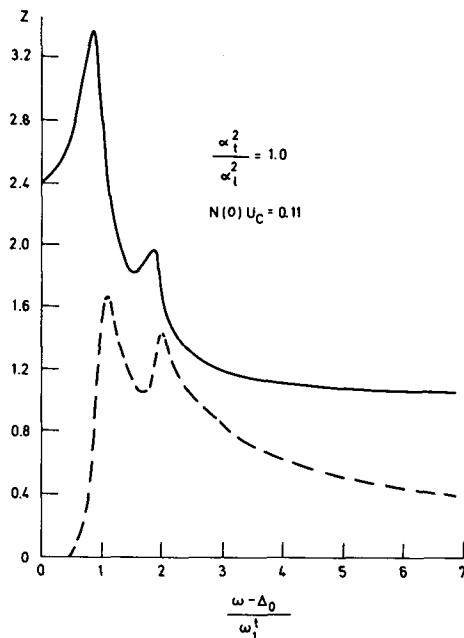
where the interaction kernels are

$$K_{\pm} = N(0) \int \alpha_{\omega}^2 \left\{ \frac{1}{E' + E + \omega + i\delta} \pm \frac{1}{E' - E + \omega + i\delta} \right\} F(\omega) d\omega \\ + \begin{cases} N(0) U_C & (+) \\ 0 & (-) \end{cases} \quad (5.5c)$$

and  $U_C$  is a Coulomb pseudo-potential which includes high-energy virtual transitions

$$U_C = \frac{V_C}{1 + N(0) V_C \log E_F/\omega_c} \quad (5.5d)$$

The cut-off  $\omega_c$  is to be chosen so that  $\omega_D \ll \omega_c \ll E_F$ . In Eq.(5.5c),  $\alpha_{\omega}^2$  is a momentum and polarization averaged electron-phonon coupling and  $F(\omega)$  is the phonon frequency distribution.

FIG. 9a. Real and imaginary parts of  $\Delta$ .FIG. 9b. Real and imaginary parts of  $Z$ .

In an attempt to explain tunnelling experiments of Rowell, Anderson and Thomas, Scalapino, Wilkins and I attempted to solve Eq.(5.5) for lead with  $F(\omega)$  fixed by the phonon spectrum determined by neutron inelastic scattering and by the general structure of the tunnelling curves. The results for the real and imaginary parts of  $\Delta$  and  $Z$  are plotted in Figs 9a and 9b. The rich structure and large imaginary (absorptive) amplitudes are directly reflected in the tunnelling conductance

$$\frac{(dI/dV)_S}{(dI/dV)_N} = \text{Re } \left. \frac{E}{\sqrt{E^2 - \Delta^2(E)}} \right|_{E=eV} \quad (5.6)$$

as shown in Fig. 10. The short dashed curve is the result of the weak coupling theory and shows no structure. The scale of variation is of the order 3-5% of the total density of states. These results give remarkably good support for the validity of the strong coupling theory. In fact, McMillan has used the tunnelling curves to invert the gap equations to find  $F(\omega)$  and has obtained new information on phonon spectra in this way. These results appear to be a major success for quantum field theory.

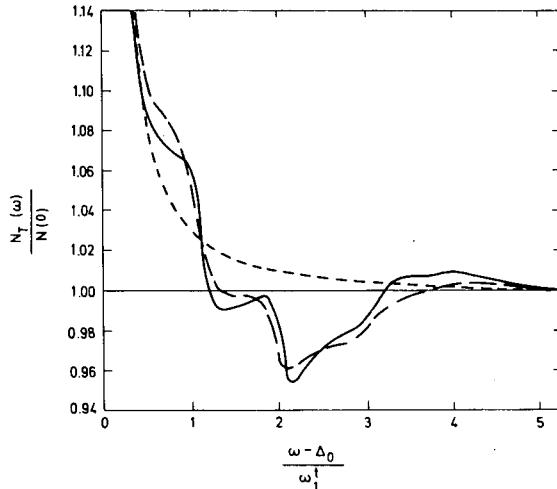


FIG.10. Tunnelling conductance.

## 6. FERROMAGNETISM AND SUPERCONDUCTIVITY

In analogy to superconductivity, ferromagnetism occurs at a temperature  $T_{fm}$  such that the spin-one particle-hole scattering amplitude develops a pole at zero centre-of-mass energy and momentum,  $q = 0$ , as sketched in Fig. 11.

For metals which are nearly ferromagnetic at low temperature,  $\tau$  has a large peak of its spectral weight along the "paramagnon" dispersion curve

$$q_0 = v_F \left( \frac{x}{x_p} \right) |\vec{q}| \quad (6.1)$$

where  $\chi$  is the observed spin susceptibility and  $\chi_p$  its Pauli value. If we approximate the irreducible particle-hole interaction by the screened Coulomb potential  $V$ , then  $\tau^{\uparrow\downarrow}$  is given by the ladder sum, and it is clear that in considering superconductivity in nearly ferromagnetic metals, particle-hole multiple scattering graphs will play an essential role. In particular, the effective pairing potential  $V_{\text{eff}}$  is now given by the sum of the graphs shown in Fig. 12.

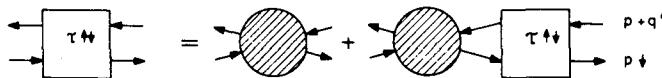


FIG.11. Spin-one particle-hole scattering amplitude.

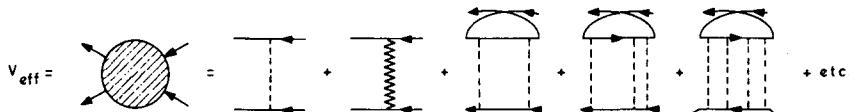


FIG.12. Effective pairing potential.

If we write  $V_{\text{eff}} = V_{\text{coul}} + V_{\text{phonon}} + V_{\text{paramagnon}}$  we can show that  $V_{\text{paramagnon}}$  is repulsive for spin-zero pairing and tends to suppress superconductivity. Physically, this is due to the fact that if we are trying to pair an up spin electron 1 with down spin electron 2 to form a condensed pair, electron 1 is surrounded by a cloud of predominantly up spin electrons whose spatial size extends to very large distances as the system tends toward ferromagnetism. Therefore, the down spin electron 2 is repelled by the ferromagnetic exchange interactions with this up spin cloud and a pair condensation is suppressed. While triplet spin pairing is enhanced by this process, imperfections scatter the electrons so as to break up functions  $\phi(r_1 - r_2)$  which have non-zero orbital angular momentum, and it is the odd  $\ell$  states which enter for triplet pairing, as a result of the Pauli principle. Calculations show that

$$N(0) V_{\text{paramagnon}} \approx + \text{const.} \log \frac{\chi}{\chi_p}$$

for  $\chi / \chi_p \gg 1$ . It seems likely that the paramagnons play a role in suppressing superconductivity near the end of the transition metal series, e.g. in Pd. A similar effect can occur if a metal is nearly antiferromagnetic, i.e. a pole of  $\tau$  near  $q_0 = 0$  in the complex  $q_0$ -plane for  $\vec{q} \neq 0$ , corresponding to a spin spiral instability. There is evidence for such an incipient antiferromagnetism in Sc, which may account for the absence of superconductivity in the beginning of the transition series (Sc, Y, Lu). A good deal more work is needed in this area to firmly establish the interference between magnetic and superconducting order.

In the papers by Doniach and Pethick in these Proceedings the concept of paramagnons is developed a good deal further.

## 7. NON-PHONONIC MECHANISMS

Over the years, there have been many suggestions of mechanisms other than the phonon attraction to produce the pairing condensation. In particular, Matthias has suggested a mechanism in which conduction electrons interact via virtual excitation of core electrons, i.e. a core polarization mechanism. He has cited the deviation of isotope effect index  $\alpha$

$$T_c \propto M^\alpha \quad (7.1)$$

in transition metals from  $-\frac{1}{2}$  as evidence for another pairing mechanism ( $M$  = isotopic mass). Calculations show that these deviations can be accounted for by the usual phonon and Coulomb mechanism if one takes proper account of the rapidly varying density of states in the d-bands and it is unlikely that core polarization plays an appreciable role in these metals.

Recently, it has been observed that uranium, in its face-centred cubic phase, has  $\alpha \approx +2$  and Matthias has argued that this is due to the enhancement of the core polarization mechanism for increasing  $M$  by the reduction of recoil of an ion as a whole during an electron-ion collision. A simple estimate based on the Debye-Waller factor shows that this effect alone leads to  $\alpha \approx 10^{-1} \rightarrow 10^{-2}$ , in strong disagreement with experiment. In uranium, an f-band is probably just above the Fermi surface and its position seems to be a strong function of pressure. A possible explanation of  $\alpha \approx +2$  on the basis of electronic polarization might involve a virtual excitation of electrons from near the Fermi level to the f-band. Since the zero point vibrations of the crystal lattice will modulate the position of the f-band and transition matrix elements to it, and since the quantities do not enter linearly in the effective interaction, the decrease of zero point amplitude with increasing  $M$  could conceivably be partially responsible for the large positive  $\alpha$ . One can, however, explain such values of  $\alpha$  on the basis of the conventional phonon and Coulomb mechanism if the f-band is near the Fermi surface, since the deviation of  $\alpha$  from  $-\frac{1}{2}$  in this model is strongly enhanced by a rapidly varying density of states near the Fermi surface. Further work must be done to resolve this question.

## 8. CONCLUSION

It has not been possible for me to discuss many areas of high current interest in superconductivity, such as type II behaviour, fluctuation phenomena in one- and two-dimensional systems. These questions are briefly touched upon in these Proceedings by Anderson and Fisher. While the microscopic theory is in exceptionally good over-all agreement with experiment, mathematical complexities have held back progress in understanding such things as dissipation in type II materials. At present, there do not appear to remain any deep mysteries in the problem of superconductivity,

although there remains a vast spectrum of challenging problems in applying the basic pairing correlation theory to various physical situations.

#### B I B L I O G R A P H Y

A fairly complete set of references to the literature can be found in:

SCHRIEFFER, J.R., *Theory of Superconductivity*, W.A. Benjamin Inc., New York, N.Y. (1964).  
RICKAYZEN, G., *Theory of Superconductivity*, Interscience Publishers, New York, N.Y. (1965).



# THEORY OF BOSE-FERMI QUANTUM LIQUIDS

I.M. KHALATNIKOV

Institute for Theoretical Physics

Moscow, Union of Soviet Socialist Republics

## Abstract

THEORY OF BOSE-FERMI QUANTUM LIQUIDS. A phenomenological theory of a mixture of Fermi and Bose liquids is presented here, similarly to Landau's procedure for Fermi liquids. We give a definition of the Fermi excitation energy in a superfluid liquid. An exact set of equations has been obtained which describes the properties of a Fermi-Bose liquid; the solutions in the acoustic range are discussed.

## INTRODUCTION

The solutions of  ${}^3\text{He}$  and  ${}^4\text{He}$  at low temperatures are cases of quantum liquids in which Fermi and Bose excitations are present at the same time. At temperatures below the temperature of degeneration of the Fermi excitations, the kind of interaction becomes essential and, in such a case, the fermionic part of the liquid forms a quantum Fermi liquid diluted in a superfluid liquid.

The phase diagram of  ${}^3\text{He}-{}^4\text{He}$  solutions shows, at temperatures around  $0.8^\circ\text{K}$ , a critical point below which the solutions separate into their phases. At  $T \rightarrow 0$  the separation line goes towards the pure substance ( ${}^3\text{He}$ ) on one side and on the other towards a point corresponding to a solution of 6% of  ${}^3\text{He}$  in  ${}^4\text{He}$  [1]. In this way, for  ${}^3\text{He}$  concentrations smaller than 6%, the solution does not separate and by lowering the temperature one always arrives at Fermi degeneration. This is in agreement with measurements of specific heat of such solutions, which depends linearly on the temperature [2]. Such solutions are, therefore, a particular Fermi-Bose quantum liquid. We will speak of a Fermi-Bose liquid in all the cases dealing with a superfluid Fermi liquid. Another example of Fermi-Bose liquid should be pure  ${}^3\text{He}$  liquid in the range of temperature where it shows superfluidity due to the Kupler coupling among non-zero momentum Fermi excitations [3]. Many properties of  ${}^3\text{He}$  in  ${}^4\text{He}$  solutions have been discussed in Refs [4 - 6]. The phenomenological theory of Fermi-Bose quantum liquid described in this paper follows the procedure indicated by Landau for Fermi liquids [7]<sup>1</sup>.

## 1. EXCITATION ENERGY

Let  $E'$  and  $P'$  be the energy and the momentum (per unit volume) of the liquid in the reference frame moving with the velocity of the super-

<sup>1</sup> In doing so, it is clear that we do not put limitations on the concentration in the Fermi phase. It is known that the parameter in the series expansion of the thermodynamic functions of a degenerate solution is the concentration to the power  $1/3$ . It is clear that for concentrations of  $\sim 6\%$  such a parameter is not small.

fluidity motion  $v_s$ . The energy in the static reference frame will be equal to

$$E = E' + P' \cdot v_s + \rho v_s^2 \quad (1.1)$$

and the momentum to

$$J = \rho v_s + P' \quad (1.2)$$

The same momentum  $J$  can be seen as the sum of the momentum of the Bose-phase of the liquid ( $\rho_1 v_s$ ) and of the total momentum of the excitations ( $\int p n_p d\tau$ )

$$J = \rho_1 v_s + \int p n_p d\tau \quad (1.3)$$

The density of the Bose-phase of the liquid  $\rho_1$  is given by<sup>2</sup>

$$\rho_1 = \rho - m \int n_p d\tau \quad (1.4)$$

where  $m$  is the mass of the Fermi particles.

By comparison of (1.2) and (1.3) we find

$$P' = \int (p - mv_s) n_p d\tau = \int p n_{p+mv_s} d\tau \quad (1.5)$$

In such a way the momentum of the excitation is expressed in the form of a variational derivative

$$p = \frac{\delta P}{\delta n_{p+mv_s}} \quad (1.6)$$

Similarly the energy of excitation is given by

$$\epsilon_p = \frac{\delta E'}{\delta n_{p+mv_s}} \quad (1.7)$$

in the frame for which  $v_s = 0$ .

The state of our system is defined by three functions: the density  $\rho$  or  $\rho_1$ , the velocity  $v_s$  and the distribution function  $n_{p+mv_s}$ . The energy of excitation can be defined as the variational derivative of the density of energy  $E$  multiplied by the distribution function, the other two functions  $\rho$  and  $v_s$  remaining constant.

---

<sup>2</sup> As for the Fermi liquid theory, the number of Fermi excitations in the  ${}^3\text{He}-{}^4\text{He}$  solutions is assumed equal to the number of Fermi particles.

In the case of the  $^3\text{He}$ - $^4\text{He}$  solutions the most natural variable is the density  $\rho_1$  and we define the excitation energy as the energy change of the liquid solution as a result of the introduction of one atom of  $^3\text{He}$  (for a given mass of the Bose liquid  $^4\text{He}$ ). In the case of a Fermi liquid, which exhibits superfluidity because of the coupling, the most natural variable will be the total density. The formulas will, thus, appear more compact. Finally we shall give the results for both cases.

Now we calculate the excitation energy  $H_{p+mv_s}$  in the presence of a superfluid motion with velocity  $v_s$  (it is obvious that in this case the momentum of the excitation is given by  $p+mv_s$ ). From Eq.(1.1) we obtain

$$H_{p+mv_s} = \frac{\delta E}{\delta n_{p+mv_s}} = \epsilon_p + p v_s \quad (1.8)$$

The energy  $\epsilon_p$  is a functional of the excitation density and can be expressed as a functional derivative

$$f(p, p') = \frac{\delta \epsilon}{\delta n_{p+mv_s}} \quad (1.9)$$

For small deviations from equilibrium one thus obtains

$$\epsilon_p = \epsilon_p^{(0)} + \int f(p, p') \delta n_{p+mv_s} d\tau' \quad (1.10)$$

where the energy  $\epsilon_p^{(0)}$  is the energy of excitation of the system in equilibrium while

$$\delta n_{p+mv_s} = n_{p+mv_s} - n_p^{(0)} \quad (1.11)$$

$n_p^{(0)}$  is the distribution function in equilibrium conditions. The fact that the variation must be defined in this way follows directly from expression (1.5), which can be rewritten as  $p' = \int p(n_{p+mv_s} - n_p^{(0)}) dt$ .

Let us now write the excitation energy  $H_p$  as a function of the momentum  $p$ . By means of Eqs.(1.8) and (1.9) and after separation of the terms which depend linearly on  $v_s$ , we obtain

$$H_p = \epsilon_{p+mv_s} + (p - mv_s, v_s) \approx \epsilon_p^{(0)} + \left( p - m \frac{\partial \epsilon}{\partial p}, v_s \right) + \int f(p - mv_s, p') \delta n_{p'+mv_s} d\tau' \quad (1.12)$$

For values of the momentum close to the Fermi limit, we obtain

$$H_p = \epsilon_p^{(0)} + \left( p - m \frac{\partial \epsilon}{\partial p}, v_s \right) - m \left( \frac{\partial \epsilon}{\partial p} v_s \right) \frac{F_1}{3} + \int f(p, p') (n_{p'} - n_p^{(0)}) d\tau' \quad (1.13)$$

We write the excitation energy  $\epsilon_p^{(0)}$  in the form

$$\epsilon_p^{(0)} = \epsilon_0 + \frac{p^2}{2m^*} \quad (1.14)$$

where  $m^*$  is the total effective mass due to the interaction of the Fermi particles with the Bose particles and with the other Fermi particles of the liquid,  $\epsilon_0$  is the zero excitation energy. For such a type of spectrum, we can deduce from expression (1.13)

$$H_p = \epsilon_0 + \frac{p^2}{2m^*} + (pv_s) \left( 1 - \frac{m}{m^*} (1 + \frac{F_1}{3}) \right) + \int f(p, p') (n_p - n_{p'}^{(0)}) d\tau' \quad (1.15)$$

The expression thus obtained differs from the one given in the paper of Bardeen, Baym and Pines [4] in the fact that the term  $p^2/2m^*$  contains the total effective mass and also in the coefficient at  $(pv_s)$ .

The value of the effective mass  $m^*$  is determined by the interaction of the Fermi excitations. The coefficient at  $pv_s$  meets a natural and necessary condition: in the pure Fermi liquid the expression  $1 - (m/m^*) (1 + F_1/3)$  equals zero and there is no superfluidity.

## 2. EQUILIBRIUM DISTRIBUTION FUNCTION

The equilibrium distribution function can be obtained by minimizing the entropy

$$S = - \int \{ (1 - n) \ln (1 - n) + n \ln n \} d\tau \quad (2.1)$$

for given values of the total energy  $E$ , of the number of excitations  $N = \int n d\tau$  and of the relative momentum  $P'$ . It is therefore necessary to minimize the functional

$$\phi = S - \beta(E - \mu_3 N - v_n P') \quad (2.2)$$

( $\beta$ ,  $\mu_3$  and  $v_n$  are the Lagrange factors with their normal meaning:  $\beta = 1/T$ ;  $\mu_3$  is the chemical potential, and  $v_n$  is the normal velocity). Changing the functional (2.2) by  $\delta n_{p+mv_s}$  we obtain

$$n_{p+mv_s}^{(0)} = 1 / \left( e^{\frac{H_p + mv_s - \mu_3}{T}} + 1 \right) \quad (2.3)$$

From Eq.(2.2) the chemical potential terms are

$$\mu_3 = \left( \frac{\partial E'}{\partial N} \right)_{S, P'} \quad (2.4)$$

whereas the thermodynamical equation has the following form:

$$dE' = TdS + \mu_3 dN + \frac{1}{m_4} \mu_4 d\rho + (v_n - v_s, dP') \quad (2.5)$$

$m_4$  is the mass of  $^4\text{He}$ .

In the case of density  $\rho_1$  as an independent variable instead of  $\rho$ , Eq.(2.5) takes the form

$$\begin{aligned} dE' &= TdS + \mu_3^{(1)} dN + \frac{1}{m_4^{**}} \mu_4 d\rho_1 + (v_n - v_s, dP') \\ \mu_3^{(1)} &= \mu_3 + \frac{m}{m_4} \mu_4 \end{aligned} \quad (2.6)$$

### 3. NORMAL DENSITY AND SPECIFIC HEAT

We shall now calculate the momentum of the Fermi excitations in the frame of reference in which the superfluid phase of the liquid is static. One has

$$P' = \int p n_{p+mv_s} d\tau \quad (3.1)$$

Let us take its equilibrium value as the distribution function. We find the change of energy of excitation in the presence of normal and superfluid shear. From Eqs.(1.8) and (1.10) we have the equation

$$\delta H = (p, v_s - v_n) + \int (p, p') \frac{\partial n^{(0)}}{\partial \epsilon'} \delta H' d\tau' \quad (3.2)$$

from which we obtain

$$\delta H = \frac{(p, v_s - v_p)}{1 + \frac{F_1}{3}} \quad (3.3)$$

From (3.1) and (3.3) we have

$$P' = \int p \frac{\partial n^{(0)}}{\partial \epsilon} \delta H d\tau = \frac{m^{**} N}{1 + \frac{F_1}{4}} (v_n - v_s) \quad (3.4)$$

The total momentum of the liquid is

$$J = \rho v_s + P' = \left( \rho - \frac{m^{**} N}{1 + \frac{F_1}{3}} \right) v_s + \frac{m^{**} N}{1 + \frac{F_1}{3}} v_n \quad (3.5)$$

The density of the normal component is therefore given by

$$\rho_n = \frac{m^* N}{1 + \frac{F_1}{3}} \quad (3.6)$$

and, obviously, not equal to the density of the Fermi component of the liquid  $mN$ .

The specific heat of the Fermi excitations can be calculated in a similar way to the Fermi liquid. We obtain

$$c = \frac{C}{N} = \gamma T, \quad \gamma = \frac{\pi^2}{3N} \left( \frac{d\tau}{d\epsilon} \right)_{\epsilon=\mu_3} = \left( \frac{\pi}{3N} \right)^{2/3} \frac{m^*}{\hbar^2} \quad (3.7)$$

The coefficient  $\gamma$  in (3.7) contains the total effective mass  $m^*$  of the excitation instead of  $m^*/(1 + \frac{F_1}{3})$  as is the case for (3.6).

#### 4. MOMENTUM AND ENERGY CONSERVATION LAW

The kinetic equation for the function  $n_{p+mv_s}$  can be written as

$$\frac{\partial n_{p+mv_s}}{\partial t} + \frac{\partial n_{p+mv_s}}{\partial x} \frac{\partial H_{p+mv_s}}{\partial p} - \frac{\partial n_{p+mv_s}}{\partial p} \frac{\partial H_{p+mv_s}}{\partial x} = I(n) \quad (4.1)$$

In order to obtain the complete set of equations describing the Fermi-Bose liquid we will make use of the conservation laws:

Continuity equation

$$\dot{\rho} + \text{div } J = 0 \quad (4.2)$$

Momentum conservation

$$J + \frac{\partial \Pi_{ik}}{\partial x_k} = 0 \quad (4.3)$$

where  $\Pi_{ik}$  is tensor of momentum flow.

When multiplying Eq.(4.1) by  $\rho_i$  and integrating over the phase volume (for simplicity we will indicate the operation of integration by a line) it should be taken into account that, if  $\int I d\tau = 0$ , we will have  $\int \rho I d\tau = \int (p + mv_s) I d\tau = 0$  and find

$$\frac{\partial P'_i}{\partial t} + \frac{\partial}{\partial x_k} \left( n_{p+mv_s} p_i \frac{\partial H_{p+mv_s}}{\partial p_k} \right) + n_{p+mv_s} \frac{\partial H_{p+mv_s}}{\partial x_i} = 0 \quad (4.4)$$

By subtracting Eq.(4.4) from Eq.(4.3) and by making use of the equation of continuity, we obtain

$$\begin{aligned} \rho \dot{v}_{si} - v_{si} \frac{\partial}{\partial x_k} (\rho v_{sk} + \overline{n_{p+mv_s} p_k}) + \frac{\partial}{\partial x_k} \left( \Pi_{ik} - \overline{n_{p+mv_s} p_i} \frac{\partial H_{p+mv_s}}{\partial p_k} \right) \\ - n_{p+mv_s} \frac{\partial H_{p+mv_s}}{\partial x_i} = 0 \end{aligned} \quad (4.5)$$

We transform the last equation by considering also the condition  $\text{rot } v_s = 0$ , and have

$$\begin{aligned} \rho \left( \dot{v}_{si} + \frac{\partial}{\partial x_i} \frac{v_s^2}{2} \right) + \frac{\partial}{\partial x_k} \left( \Pi_{ik} - \rho v_{si} v_{sk} - \overline{n_{p+mv_s} p_i} \frac{\partial H_{p+mv_s}}{\partial p_k} \right) \\ - v_{si} \frac{\partial}{\partial x_k} \overline{n_{p+mv_s} \rho_k} - n_{p+mv_s} \frac{\partial H_{p+mv_s}}{\partial x_i} = 0 \end{aligned} \quad (4.6)$$

We now transform the last term of Eq.(4.6)

$$\begin{aligned} \overline{n_{p+mv_s} \frac{\partial H_{p+mv_s}}{\partial x_i}} &= \frac{\partial}{\partial x_i} \overline{n_{p+mv_s} H_{p+mv_s}} - H_{p+mv_s} \frac{\partial n_{p+mv_s}}{\partial x_i} \\ &= \frac{\partial}{\partial x_i} \overline{n_{p+mv_s} H_{p+mv_s}} - \frac{\partial E}{\partial x_i} + \frac{\partial E}{\partial \rho} \frac{\partial \rho}{\partial x_i} + \frac{\partial E}{\partial v_{sk}} \frac{\partial v_{sk}}{\partial x_i} \\ &= \frac{\partial}{\partial x_i} \left( \overline{n_{p+mv_s}} - E + \frac{\partial E}{\partial \rho} \rho \right) - \rho \frac{\partial}{\partial x_i} \frac{\partial E}{\partial \rho} \\ &+ \frac{\partial}{\partial x_k} \left( v_{si} \frac{\partial E}{\partial v_{sk}} \right) - v_{si} \frac{\partial}{\partial x_k} \frac{\partial E}{\partial v_{sk}} = 0 \end{aligned} \quad (4.7)$$

By substituting (4.7) into (4.6) and by considering the equation

$$\frac{\partial E}{\partial v_{sk}} = \rho v_{sk} + \overline{n_{p+mv_s} p_k} \quad (4.8)$$

we obtain

$$\begin{aligned} \rho \left( \dot{v}_{si} + \frac{\partial}{\partial x_i} \frac{\partial E}{\partial \rho} \right) + \frac{\partial}{\partial x_k} \left( \Pi_{ik} - \overline{n_{p+mv_s} p_i} \frac{\partial H_{p+mv_s}}{\partial p_k} \right) \\ - \frac{\partial}{\partial x_k} \left( v_{si} \frac{\partial E}{\partial v_{sk}} \right) - \frac{\partial}{\partial x_i} \left( \overline{n_{p+mv_s} H_{p+mv_s}} - E + \rho \frac{\partial E}{\partial \rho} \right) = 0 \end{aligned} \quad (4.9)$$

From (4.9) we find the equation of superfluid motion and the expression for the tensor momentum flow

$$\dot{\vec{v}}_s + \nabla \left( \frac{\partial E}{\partial \rho} \right)_{n, v_s} = 0 \quad (4.10)$$

$$\begin{aligned} \Pi_{ik} = & n_{p+mv_s} p_i \frac{\partial \epsilon}{\partial p_k} + n_{p+mv_s} (p_i v_{sk} + p_k v_{si}) + \rho v_{si} v_{sk} \\ & + \delta_{ik} \left( \overline{n_{p+mv_s} \epsilon_p} - E' + \rho \frac{\partial E'}{\partial \rho} \right) \end{aligned} \quad (4.11)$$

By using (1.1), the equation of superfluid motion (4.10) can be rewritten as follows:

$$\dot{\vec{v}}_s + \nabla \left( \frac{\partial E'}{\partial \rho} + \frac{v_s^2}{2} \right) = 0 \quad (4.12)$$

We define the pressure by means of the usual relationship:

$$p = -E' + \frac{\partial E'}{\partial \rho} \rho + TS \quad (4.13)$$

and, consequently, we can write the equation for the tensor momentum flow

$$\begin{aligned} \Pi_{ik} = & n_{p+mv_s} p_i \frac{\partial \epsilon}{\partial p_k} + \overline{n_{p+mv_s} (p_i v_{sk} + p_k v_{si})} \\ & + \rho v_{si} v_{sk} + \delta_{ik} (\overline{n_{p+mv_s} \epsilon_p} - TS + p) \end{aligned} \quad (4.14)$$

To obtain the energy conservation law let us multiply Eq.(4.1) by  $H_{p+mv_s}$  and integrate over the phase volume; from this we obtain

$$H_{p+mv_s} \frac{\partial n_{p+mv_s}}{\partial t} + \frac{\partial}{\partial x_k} \left( \overline{n_{p+mv_s} H_{p+mv_s}} \frac{\partial H_{p+mv_s}}{\partial p_k} \right) = 0 \quad (4.15)$$

The time-derivative of the total energy is

$$\frac{\partial E}{\partial t} = H_{p+mv_s} \frac{\partial n_{p+mv_s}}{\partial t} + \frac{\partial E}{\partial \rho} \frac{\partial p}{\partial t} \frac{\partial E}{\partial v_{sk}} \frac{\partial v_{sk}}{\partial t} \quad (4.16)$$

and, by using (4.15), (4.2) and (4.10), we obtain

$$\begin{aligned}\frac{\partial E}{\partial t} &= -\frac{\partial}{\partial x_k} \left( n_{p+mv_s} H_{p+mv_s} \frac{\partial H_{p+mv_s}}{\partial p_k} \right) - \frac{\partial E}{\partial \rho} \operatorname{div} J - J \operatorname{div} \frac{\partial E}{\partial p} \\ &= -\operatorname{div} \left\{ n_{p+mv_s} H_{p+mv_s} \frac{\partial H_{p+mv_s}}{\partial p} + J \frac{\partial E}{\partial \rho} \right\}\end{aligned}\quad (4.17)$$

The term within the brackets is the flow of energy

$$Q = J \frac{\partial E}{\partial \rho} + n_{p+mv_s} H_{p+mv_s} \frac{\partial H_{p+mv_s}}{\partial p} \quad (4.18)$$

## 5. FIRST AND SECOND SOUND

Let us analyse the acoustic oscillations at low frequency of a Fermi-Bose liquid. In such limiting case hydrodynamics holds both for the liquid as a whole and for its Fermi component. However, for coherence, we will keep using the kinetic equation for the Fermi excitations. If  $n_1 = n_{p+mv_s} - n_p^{(0)}$  and  $\rho' = \rho - \rho_0$ , the deviations from the equilibrium values of  $n_{p+mv_s}$  and  $\rho$ , the velocity  $v_s$  will be an infinitesimal of the same order.

Let us suppose that all these variables are periodic functions of time and space co-ordinates:  $\exp(i\omega t - ikx)$ . Then Eqs (4.1), (4.2) and (4.10) yield, in the linear approximation, the following system:

$$i(\omega - kv)n_1 + ikv \frac{\partial n^{(0)}}{\partial \epsilon} \left( \frac{\partial \epsilon}{\partial \rho} \rho' + \rho v_s + \int f(p, p') n'_1 d\tau' \right) = 0 \quad (5.1)$$

$$i\omega \rho' - ik(\rho v_s + \int n_1 p d\tau) = 0 \quad (5.2)$$

$$i\omega v_s - ik \left( \frac{s^2}{\rho} \rho' + \int \frac{\partial \epsilon}{\partial \rho} n_1 d\tau \right) = 0; \quad v = \frac{\partial \epsilon}{\partial p}, \quad \frac{s^2}{\rho} = \frac{\partial^2 E}{\partial \rho^2} \quad (5.3)$$

Let us introduce, further, the following symbols and non-dimensional variables:

$$n_1 = \frac{\partial n^{(0)}}{\partial \epsilon} m^* v_F^2 \nu(x), \quad \kappa = \cos \theta, \quad F(x) = f(x) \left( \frac{d\tau}{d\epsilon} \right)_{\epsilon=\mu_3} = \sum F_n P_n (\cos x)$$

$\theta$  is the angle between the excitation momentum  $p$  and the wave vector  $k$ ,  $v_F = (\partial \epsilon / \partial \rho)_{\epsilon=\mu_3}$ ,  $\alpha = (\partial \epsilon / \partial \rho)_{\epsilon=\mu_3}$ ,  $\rho / m^* s^2$ ,  $u = \omega / kv_F$ ,  $\tilde{\rho}' = \frac{\rho'}{\rho}$ ,  $\tilde{v}_s = \frac{v_s}{v_F} = 0$

With the new variables the set of Eqs (5.1), (5.2) and (5.3) becomes

$$(u - x)\nu(x) + x(\alpha \tilde{\rho}' - \overline{F(x)\nu(x')}) + x\tilde{v}_s = I(\nu) \quad (5.4)$$

$$u\tilde{\rho}' - \tilde{v}_s + \nu_1 \frac{Nm^*}{\rho} = 0 \quad (5.5)$$

$$uv_s - \frac{s^2}{v_F^2} \tilde{\rho}' + \nu_0 3\alpha \frac{Nm^*}{\rho} \frac{s^2}{v_F^2} = 0 \quad (5.6)$$

The function  $\nu(x)$  can be expanded in a series of Legendre polynomials

$$\nu(x) = \sum \nu_n P_n(x) \quad (5.7)$$

$\nu_0$  and  $\nu_1$  in Eqs (5.4), (5.5) and (5.6) are, respectively, the zero and first harmonic.

Obviously, the set of equations that we just obtained is valid for all values of frequency. For small values of  $\omega$ , such that the collision time  $\tau \gg 1/\omega$ , hydrodynamic holds. In this case, the zero and first harmonic of the function  $\nu(x)$  are different from zero (second and higher harmonics are comparable with or larger than  $\omega\tau$ ). By imposing that  $P_0(x) = P_1(x)$  in Eq.(5.4) we obtain the equations for  $\nu_0$  and  $\nu_1$

$$uv_0 - \frac{1}{3} \left( 1 + \frac{F_1}{3} \right) \nu_1 + \frac{1}{3} \tilde{v}_s = 0 \quad (5.8)$$

$$-(1 + F_0)\nu_0 + uv_1 + \alpha \frac{s^2}{v_F^2} \tilde{\rho}' = 0 \quad (5.9)$$

The condition of compatibility of the set of Eq.(5.5), (5.6) and (5.8), (5.9) yields the following dispersion relations that define  $u$ :

$$\begin{aligned} u^4 - u^2 & \left\{ \frac{s^2}{v_F^2} \left[ 1 + \frac{Nm^*}{\rho \left( 1 + \frac{F_1}{3} \right)} \left( \alpha \left( 1 + \frac{F_1}{3} \right) + 1 \right)^2 - 1 \right] + \frac{1}{3} \left( 1 + \frac{F_1}{3} \right) (1 + \tilde{F}_0) \right\} \\ & + \frac{1}{3} \left( 1 + \frac{F_1}{3} \right) (1 + \tilde{F}_0) \frac{s^2}{v_F^2} \left( 1 - \frac{Nm^*}{\rho \left( 1 + \frac{F_1}{3} \right)} \right) = 0 \end{aligned} \quad (5.10)$$

where

$$\tilde{F}_0 = F_0 - 3\alpha^2 \frac{s^2}{v_F^2} \frac{Nm^*}{\rho} \quad (5.11)$$

In the case of small concentration of Fermi particles, Eq.(5.10) has the two following approximate roots:

$$u_1^2 = \frac{s^2}{v_F^2} \left\{ 1 + \frac{Nm^*}{\rho \left( 1 + \frac{F_1}{3} \right)} \left[ \left( \alpha \left( 1 + \frac{F_1}{3} \right) + 1 \right)^2 - 1 \right] \right\} \quad (5.12)$$

$$u_2^2 = \frac{1}{3} \left( 1 + \frac{F_1}{3} \right) \left( 1 + \tilde{F}_0 \right) \left[ 1 - \frac{Nm^*}{\rho \left( 1 + \frac{F_1}{3} \right)} \left( \alpha \left( 1 + \frac{F_1}{3} \right) + 1 \right)^2 \right] \quad (5.13)$$

Equation (5.12) yields the velocity of sound (first sound) whereas Eq.(5.13) yields the velocity of second sound. The second sound in the Fermi-Bose liquid is a sound that propagates in the Fermi component of the liquid. Equation (5.13), in first approximation, coincides with the expression yielding the velocity of sound in a Fermi liquid. The difference lies in the fact that instead of  $F_0$  one has  $\tilde{F}_0$ , renormalized as a consequence of the phonon interaction (5.11).

For the case of small concentrations of the Fermi component, it is more convenient to express the results through the variables  $\rho_1$  and  $N$  instead of  $\rho$  and  $N$ . In what follows we will use such more convenient variables.

## 6. ZERO SOUND

Let us now consider the case of sound at high frequencies (such that  $\omega\tau \gg 1$ ). In the kinetic equations one can, in such a case, neglect the integral of the collisions. For the solution of the kinetic equation, we will expand the distribution function in series of Legendre polynomials. Let us suppose, for simplicity, that only the first two harmonics of the function  $F(F_0$  and  $F_1$ ) are different from zero. The kinetic equation will thus yield two equations for the first two harmonics of the function  $v(v_0$  and  $v_1$ )

$$(1 + F_0 w)v_0 - \frac{1}{3} F_1 uwv_1 + \frac{s^2}{v_F^2} \alpha w \tilde{\rho}' + uw \tilde{v}_s = 0 \quad (6.1)$$

$$-F_0 uwv_0 + \frac{1}{3} \left( 1 + \frac{F_1}{3} + F_1 u^2 w \right) v_1 + \frac{s^2}{v_F^2} \alpha uw \tilde{\rho}' - \left( \frac{1}{3} - u^2 w \right) \tilde{v}_s = 0 \quad (6.2)$$

where

$$w = -1 + \frac{u}{2} \ln \frac{u+1}{u-1} \quad (6.3)$$

The conditions of compatibility of Eqs (6.1) and (6.2) with Eqs (5.1), (5.2), (5.3), (5.4), (5.5) and (5.6), yield the dispersion equation in the high-frequency region. Since for one solution the functions  $F_0$  and  $F_1$  are

small, the root of the dispersion equation is of the order of unity. In proximity of this root the dispersion equation has the following form:

$$1 + \left( \tilde{F}_0 + \frac{\tilde{F}_1 u^2}{1 + F_1/3} \right) w = 0 \quad (6.4)$$

where

$$\tilde{F}_0 = F_0 - \frac{3Nm^*}{\rho} \alpha^2 \frac{S^2}{v^2}, \quad \tilde{F}_1 = F_1 - \frac{Nm^*}{\rho} (\alpha + 1) \quad (6.5)$$

We keep, for  $F_1$ , in addition to the main term, also the terms of higher order with respect to  $Nm^*/\rho$  since they might contribute considerably in the case of the  ${}^3\text{He}$ - ${}^4\text{He}$  solutions.

Equation (6.4) defines the velocity of zero sound in the solution and has non-damped solutions only in the case

$$\tilde{F}_0 + \frac{\tilde{F}_1 u^2}{1 + \frac{F_1}{3}} > 0 \quad (6.6)$$

As far as  $\tilde{F}_0$  is concerned, it has negative values close to 1; the second term is difficult to evaluate: for  $u^2 \sim 1$  it is probably less than one.<sup>3</sup> In this case the dispersion equation does not have zero sound solutions. However, since  $F_1$  is not well known, the problem of the existence of the zero sound in the solution remains open (it is worth noticing the fact that the second term in  $\tilde{F}_1$  is, unfortunately, negative).

## 7. TWO DEFINITIONS OF THE EXCITATION ENERGY

Let us now write some relationships among the parameters of the theory for two different groups of variables. Up to now we have made use of:  $\rho$ ,  $v_s$  and  $\delta n_{p+mv_s} = n_{p+mv_s} - n_p^{(0)}$ . Let us now consider another group: the density of the component of Bose liquid  $\rho_1 = \rho - m \int n d\tau$ , the velocity  $v_s$  and let us define the excitation energy as the variational derivative of the energy  $E$  for  $\delta n_p = n_p - n_p^{(0)} - mv_s$ . We have

$$H_p = \frac{\delta E}{\delta n_p} \quad (7.1)$$

We will indicate with the index 1 all the derivatives calculated for constant values of  $\rho_1$ . In this way we obtain the following expression for the excitation energy:

$$H_p = \left( \frac{\delta E}{\delta n_p} \right)_\rho = \left( \frac{\delta E}{\delta n_p} \right)_{\rho_1} - m \frac{\partial E}{\partial \rho_1} = H_p^{(1)} - m \frac{\partial E}{\partial \rho_1} \quad (7.2)$$

---

<sup>3</sup> If  $u^2 \gg 1$ , then  $w = 1/3 u^2$  and Eq. (6.4) could have solutions only for large values of  $F_1$ , which, obviously, are not justified.

The second term in Eq.(7.2) arises from the change in reference; its physical meaning is understandable: the excitation energy for a given total density  $\rho$  and for a given density of the Bose component are not equivalent. Such a term is equal to the chemical potential of the Bose component, multiplied by the mass of the Fermi particles

$$m \frac{\partial E}{\partial \rho_1} = m \frac{\mu_4}{m_4} \quad (7.2')$$

The chemical potentials  $\mu_3$  and  $\mu_3^{(1)}$  are bound by the relation

$$\mu_3 = \mu_3^{(1)} - \frac{m}{m_4} \mu_4 \quad (7.3)$$

and therefore for the difference  $H_p - \mu_3$  we obtain

$$H_p - \mu_3 = H_p^{(1)} - \mu_3^{(1)} \quad (7.4)$$

The excitation energy, thus, if computed from the chemical potential, is invariant. The function  $f(pp')$  can be computed from (7.2) and we find

$$f(p, p') = \frac{\delta H_p}{\delta n_{p'}} = \frac{\delta \epsilon_p^{(1)}}{\delta n_{p'}} - 2m \frac{\partial \epsilon_p^{(1)}}{\partial \rho_1} + m^2 \frac{\partial^2 E}{\partial \rho_1^2} = f^{(1)}(p, p') - 2m \frac{\partial \epsilon_p^{(1)}}{\partial \rho_1} + m^2 \frac{\partial^2 E}{\partial \rho_1^2} \quad (7.5)$$

Let us introduce the following symbols:

$$\frac{\partial^2 E}{\partial \rho_1^2} = \frac{\partial}{\partial \rho_1} \frac{\mu_4}{m_4} = \frac{s_1^2}{\rho_1}, \quad \alpha_1 = \frac{\rho_1}{m^* s_1^2} \left( \frac{\partial \epsilon^{(1)}}{\partial \rho_1} \right)_{\epsilon^{(1)} = \mu_3^{(1)}} \quad (7.6)$$

Equation (7.3) yields the limit Fermi energy

$$F = F^{(1)} + \frac{3Nm}{\rho_1} \frac{s_1^2}{v_F^2} \left( -2\alpha_1 + \frac{m}{m^*} \right) \quad (7.7)$$

We can also easily find a relation between  $\alpha$  and  $\alpha^{(1)}$ :

$$\alpha = \frac{\rho}{m^* s^2} \frac{\partial \epsilon}{\partial p} = \frac{\rho}{m^* s^2} \left( \frac{\partial \epsilon^{(1)}}{\partial \rho_1} - m \frac{\partial E}{\partial \rho_1^2} \right) = \frac{\rho s_1^2}{\rho_1 s^2} \left( \alpha_1 - \frac{m}{m^*} \right) \quad (7.8)$$

For  $\tilde{F}_0$  we find

$$\tilde{F}_0 = F_0 - 3\alpha^2 \frac{s^2}{v_F^2} \frac{Nm^*}{\rho} = F_0^{(1)} + \frac{3Nm^*}{\rho_1} \frac{s_1^2}{v_F^2} \left[ \frac{m}{m^*} \left( -2\alpha_1 + \frac{m}{m^*} \right) - \frac{\rho_1 s^2}{\rho s_1^2} \left( \alpha_1 - \frac{m}{m^*} \right)^2 \right]$$

and since  $\partial^2 E / \partial \rho_1^2 = \partial^2 E / \partial \rho^2$  we will have  $s^2 / \rho = s_1^2 / \rho_1$  and

$$\tilde{F}_1 = F_0^{(1)} - \frac{3Nm^*}{\rho_1} \frac{s_1^2}{v_F^2} \alpha_1^2 = \tilde{F}_0^{(1)} \quad (7.9)$$

Let us write the velocity of sound for a weak solution as a function of its value for the pure solvent

$$s_1^2 = \rho_1 \frac{\partial}{\partial \rho_1} \frac{\mu_4}{m_4} = s_{10}^2 + \rho_1 \frac{\partial}{\partial \rho_1} \frac{\delta \mu_4}{m_4}$$

the quantity  $\delta \mu_4$  can be found from the single relation

$$\frac{1}{m_4} \delta \mu_4 = \frac{1}{m_4} \frac{\partial \mu_4}{\partial N} N = \frac{\partial \mu_3^{(1)}}{\partial \rho_1} = \alpha_1 \frac{m^* s_1^2}{\rho_1} N \quad (7.10)$$

In this way we obtain

$$s_1^2 \approx s_{10}^2 \left( 1 + \beta \frac{Nm^*}{\rho_1} \right), \quad \beta = \frac{\rho_1^2}{m^* s_{10}^2} \left( \frac{\partial^2 \epsilon}{\partial \rho_1^2} \right)_{\epsilon=\mu_3^{(1)}} \quad (7.11)$$

Similarly for  $s^2$

$$s^2 = \frac{\rho}{\rho_1} s_1^2 = s_{10}^2 \left[ 1 + \left( \beta + \frac{m}{m^*} \right) \frac{Nm^*}{\rho_1} \right] \quad (7.12)$$

With the new variables the first- and second-sound velocity are given by

$$U_1^2 = \frac{s_{10}^2}{v_F^2} \left\{ 1 + \frac{Nm^*}{\rho_1 \left( 1 + \frac{F_1}{3} \right)} \left[ \left( \alpha_1 \left( 1 + \frac{F_1}{3} \right) + \frac{\delta m}{m^*} \right)^2 + \beta \left( 1 + \frac{F_1}{3} \right) - \frac{\delta m}{m^*} \right] \right\} \quad (7.13)$$

$$U_2^2 = \frac{1}{3} \left( 1 + \frac{F_1}{3} \right) \left( 1 + \tilde{F}_0 \right) \left[ 1 + \frac{Nm^*}{\rho_1 \left( 1 + \frac{F_1}{3} \right)} \left( \alpha_1 \left( 1 + \frac{F_1}{3} \right) + \frac{\delta m}{m^*} \right)^2 \right] \quad (7.14)$$

$$\frac{\delta m}{m^*} = 1 - \frac{m}{m^*} \left( 1 + \frac{F_1}{3} \right)$$

A peculiar situation is the following: let us compute  $\tilde{F}_0$ . Similarly to the procedure of Ref.[4], we can show that  $F_0^{(1)}$  is approximately equal to

$$F_0^{(1)} = \frac{3Nm_4}{\rho_1} \frac{s_1^2}{v_F^2} \left( 2\alpha_1 - \frac{m_4}{m^*} \right) \quad (7.15)$$

and thus

$$\tilde{F}_0 = \tilde{F}_0^{(1)} = -\frac{3Nm^*}{\rho_1} \frac{s_1^2}{v_F^2} \left( \alpha_1 - \frac{m_4}{m^*} \right)^2 \quad (7.16)$$

The parameter  $\alpha_1 m^*/m_4$  is equal to the ratio of the relative volumes of the atoms  ${}^3\text{He}$  and  ${}^4\text{He}$ :

$$\frac{m^*}{m_4} \alpha_1 = \frac{\partial \mu_3}{\partial \rho_1} \frac{\rho_1}{m_4 s_1^2} = \frac{\partial \mu_3}{\partial \rho_1} \left/ \frac{\partial \mu_4}{\partial \rho_1} \right. = \frac{v_3}{v_4} \quad (7.17)$$

Equation (7.16), using also expression (7.17), yields the following numerical value for  $\tilde{F}_0$

$$\tilde{F}_0 = -1.15 c^{1/3} \quad (7.18)$$

For concentrations "c" equal to 6%,  $\tilde{F}_0 = -0.45$ . As is well known [8], the inequality  $1 + \tilde{F}_0 > 0$  is the condition of stability of the Fermi system relative to the fluctuations of the density of excitations. Violation of such an inequality would yield an infinite increase of the fluctuations of excitation density. This can be seen from Eq.(5.13) since, if  $1 + \tilde{F}_0 < 0$ , the velocity of the second sound would be imaginary. The fact that  $1 + \tilde{F}_0$  is small is not by chance and must result in an enhancement of some effects.

The physical meaning of the function  $\tilde{F}_0$  is easily understood. Where-  
as  $F_0$  is the partial derivative  $\delta \mu_3 / \delta N$  at constant density of the liquid  $\rho$ ,  $\tilde{F}_0$  is the derivative of  $\mu_3$  in equilibrium conditions, i.e., when  $\mu_4$  is constant. Consequently, we have

$$\tilde{F}_0 = \left( \frac{\partial \tau}{\partial \epsilon} \right)_{\epsilon=\mu_3} \left[ \frac{\partial \mu_3}{\partial N} + \frac{\partial \mu_3}{\partial p} \left( \frac{\partial p}{\partial N} \right)_{\mu_4} \right] = F_0 - \frac{3Nm^*}{\rho} \alpha^2 \frac{s^2}{v_F^2}$$

## 8. CONCLUSIONS

Up to now, the temperature was assumed to be so low that the contributions of Bose-type elementary excitations, such as phonons and rotons, could be neglected.

In a system like the solution of  ${}^3\text{He}$  in  ${}^4\text{He}$  at temperatures below the degeneration temperatures, i.e., in the region in which the model of Fermi liquid holds, the contribution of the phonons in all the thermodynamical variables is so small that it can be neglected. The normal density of phonons at  $0.1^\circ\text{K}$  is less than  $10^{-8}$ .

In the general case, conversely, if there are different kinds of excitations, both Fermi and Bose type, it is necessary, to carry out the

summation over all types of excitations in all the equations containing the integration over the phase volume of the excitations.

Up to now, the spins of the Fermi excitations have never been described: the spin can be treated with the same procedure followed in the theory of the Fermi liquid. Wherever one finds integrals as  $\int d\tau$  one must write  $\frac{1}{2} Sp_0 \int d\tau$  (for the case of spin  $\frac{1}{2}$ ) and one must consider that the function  $f_{\sigma\sigma'}(pp')$  also depends on the spins  $\sigma$  and  $\sigma'$ .

#### ACKNOWLEDGEMENT

I am very grateful to Professor A. Alekseev for a useful discussion.

#### REFERENCES

- [1] EDWARDS, D.O., BREWER, D.F., SELIGMAN, P., SKERTIC, M., YAQUB, M., Phys. Rev. Lett. 15 (1965) 773.
- [2] ANDERSON, A.C., ROACH, W.R., SANWINSKI, R.E., WHEATLY, J.C., Phys. Rev. Lett. 16 (1966) 263.
- [3] PITAIEVSKY, L.P., Soviet Phys. JETP 37 (1959) 1794; LORKIN, A., MIGDAL, A., Soviet Phys. JETP 44 (1963) 1703.
- [4] BARDEEN, J., BAYM, G., PINES, D., Phys. Rev. 156 (1967) 207.
- [5] BAYM, G., Phys. Rev. Lett. 18 (1967) 71.
- [6] KHALATNIKOV, I.M., Soviet Phys. JETP Lett. 5 (1967) 288.
- [7] ABRIKOSOV, A.A., KHALATNIKOV, I.M., Soviet Phys. Usp. 66 2 (1963) 177.
- [8] POMERANCHUK, J. Ya., Soviet Phys. JETP 35 (1958) 524.

# PARAMAGNETISM IN FERMI LIQUIDS\*

S. DONIACH  
Physics Department,  
Imperial College,  
London, United Kingdom

## Abstract

PARAMAGNETISM IN FERMI LIQUIDS. A summary is given of developments in the theory of the low-temperature properties of a paramagnetic Fermi liquid which is close to an instability to the ferromagnetic state.

This paper is a summary of recent developments in the theory of a normal Fermi liquid which is paramagnetic at zero temperature, but is nevertheless close to an instability to the ferromagnetic state.

As is well known in the theory of ferromagnetic metals, instabilities of this type can result, within a molecular field or random phase approximation (r. p. a.), from the effects of strong short-range repulsive forces acting between the fermions.

Examples of physical systems which behave as if they were close to the ferromagnetic transition are liquid  $^3\text{He}$  (which does not actually reach the ferromagnetic state but goes solid on compression beyond 27 atm) and certain transition metal alloys.

## SPIN FLUCTUATIONS

The interesting features of the theory which we wish to discuss result from a combination of the critical fluctuations which occur in any system just above a second-order phase transition and the sharpness of the Fermi surface in the Fermi liquid near absolute zero in temperature. It should be stressed that, in contrast to an insulating magnetic system, the Fermi liquid has the possibility of remaining in the paramagnetic state at  $T = 0$ , so that the critical fluctuations are not the same as those for a system with a finite critical temperature. The main effect [1, 2] in the Fermi liquid is that interaction of fermion excitations (Landau quasi-particles) with critical spin density fluctuations in the liquid leads to anomalously large temperature dependence of the thermodynamic and transport properties of the liquid at low temperatures.

The nature of the critical fluctuations may be obtained by studying the spin density space- and time-dependent response function

$$\chi^{-+}(\vec{r}, t) = -i\theta(t)\langle[\sigma^-(\vec{r}, t), \sigma^+(0, 0)]\rangle \quad (1)$$

\* Much of the material in this review also appears in: DONIACH, S., "Current status of liquid  $^3\text{He}$ : theory and experiment", 11th Int. Conf. Low Temperature (Proc. Conf. St. Andrews, 1968).

The tendency towards instability is shown by the large value of the static response, given in terms of the Fourier transform,  $\chi^+(q, \omega)$ , of Eq. (1) by

$$\chi_{\text{static}} = \lim_{\substack{q \rightarrow 0 \\ \omega \rightarrow 0}} \chi^+(q, \omega) \quad (2)$$

We denote the parameter  $\chi_{\text{static}}/\chi_{\text{Pauli}}^0$ , where  $\chi_{\text{Pauli}}^0$  is the value of  $\chi_{\text{static}}$  for the non-interacting system, by  $1/K_0^2$ . In terms of the Landau parameters of the system

$$\frac{1}{K_0^2} = \frac{m^*/m}{1 + F^a} \quad (3)$$

For liquid  ${}^3\text{He}$   $1/K_0^2$  varies from 9 at vapour pressure to 22 at 27 atm. For palladium metal  $1/K_0^2$  is about 10 and for certain Rh-Ni alloys values of up to 50 have been observed [3].

Corresponding to a large value of  $1/K_0^2$ , we may deduce from the phenomenological Ornstein-Zernike formula

$$\frac{\chi(q, \omega=0)}{\chi_{\text{Pauli}}} = \frac{1}{K_0^2 + \alpha(q/P_F)^2} \quad (4)$$

(where  $\alpha$  is a numerical constant) that local perturbations of spin density will produce a polarization of spin density in the liquid [4, 5, 6] with a long range of order  $\sqrt{\alpha}/K_0 P_F$ . Such polarizations have been observed around iron impurities in Pd metal by neutron diffraction measurement of magnetic form factors [7].

In order to study the effects of the fluctuations on the excitations spectrum, one requires the frequency dependence of  $\chi^+(q, \omega)$ . Both from the Landau phenomenological theory [8], valid at  $q/P_F \ll 1$ ,  $\omega/q V_F \ll 1$ , and, more approximately, from the random phase approximation [9], using a  $\delta$ -function model for the t-matrix for effective particle-hole scattering, one finds the following approximate form for the spectral function:

$$\text{Im } \chi^+(q, \omega) \approx \frac{i \frac{\pi}{2} \omega/q V_F}{\left[ K_0^2 + \alpha(q/P_F)^2 \right]^2 + \left[ \frac{\pi}{2} (1 - K_0^2) \omega/q V_F \right]} \quad (5)$$

which has a characteristic peak in the region of

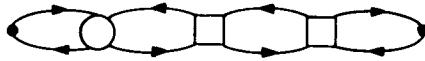
$$\omega = \frac{q}{P_F} K_B T_S \quad (6)$$

where  $T_S$  is a spin fluctuation or "paramagnon" excitation temperature

$$K_B T_S = \frac{4}{\pi} K_0^2 \epsilon_F \quad (7)$$

which tends to zero as the instability is approached. This peak represents a resonance in the particle-hole scattering amplitude (Fig. 1). A resonance of this type has been observed in paramagnetic nickel at high temperatures ( $T \sim 1.6 T_c$ ) using inelastic slow neutron scattering [10].

FIG. 1. Contributions to the particle-hole scattering amplitude.



## EQUILIBRIUM PROPERTIES

At zero temperature the main effect of the paramagnons is to enhance the quasi-particle effective mass. Using the  $\delta$ -function interaction model, in which the interaction has strength  $I$  (leading to  $K_0^2 = 1 - I N(0)$ , where  $N(0)$  is the density of unperturbed one-particle states, in r.p.a.), this enhancement may be estimated by using an extension [1, 2] of r.p.a. (Fig. 2) leading to

$$\frac{m^*}{m} \approx 1 + \log(1/K_0^2) \quad (8)$$

which diverges at the instability limit  $K_0^2 \rightarrow 0$ . Enhancements of this type, which contribute to the linear term,  $C_v \propto \gamma T$ , in the electronic specific heat,



FIG. 2. R.p.a. diagram representing emission and re-absorption of a paramagnon (wavy line) by a fermion.

have been observed in Ni-Rh alloys [3]. However, the above theoretical estimate is necessarily unreliable as the main contributions come from large momentum transfer to the spin fluctuations (5), over the range of which there is probably a bad approximation.

At finite temperatures the effect of the paramagnons on the free energy can be obtained in the extended r.p.a. [11] used for Fig. 2 as

$$\Delta F = \sum_q \frac{3}{4\pi} \int_0^\infty d\omega \left( n(\omega) + \frac{1}{2} \right) \text{Im} \log [1 - I \chi^0(q, \omega)] \quad (9)$$

where  $n(\omega)$  is the Planck function  $1/(e^{\beta\omega} - 1)$ .

In the low temperature limit this leads to the same result (Eq. (8)) as obtained from Fig. 2. The higher temperature corrections are of the form [1]

$$C_v \propto \gamma T + \left( \frac{T}{T_S} \right)^3 \log T/T_S + \dots \quad (10)$$

Unlike the linear term, the contribution to the  $T^3 \log T$  term comes mainly from the small  $q$  part of the integral [9], i.e. involves small momentum

transfers. This suggests that the coefficient of this term should be obtainable more generally from Landau theory than from the above approximation [12]. However, the  $T = 0$  Landau theory expression for the entropy

$$S = - \sum_p [f_p \log f_p + (1 - f_p) \log (1 - f_p)] \quad (11)$$

gives an answer [11] which disagrees with the coefficient of the  $T^3$  term in Eq. (10) if the quasi-particle energy obtained from Fig. 2 is used in the calculation. A generalization of Eq. (10) valid at finite temperature has not so far been found.

The use of approximation (9) for a system where the ferromagnetic instability occurs at a finite temperature  $T_c$  gives a specific heat singularity [13] of the form  $(T - T_c)^{-1/2}$  which is manifestly wrong — the approximation breaks down completely for finite temperature critical fluctuations. However, it can be used to estimate the specific heat contribution [14] due to spin-wave collective modes in an itinerant ferromagnet at low temperatures,  $T \ll T_c$ , when Eq. (9) gives the usual Bloch spin wave  $T^{5/2}$  contribution to the specific heat.

## TRANSPORT PROPERTIES

The contribution of the critical spin fluctuations to temperature dependence of the transport coefficients, particularly thermal conductivity  $K$  and spin diffusion  $D$ , may be obtained from a Boltzmann equation treatment [15, 16] in terms of a quasi-particle relaxation time

$$1/\tau = \int d\epsilon_2 \int d\epsilon'_1 \int d\epsilon'_2 \delta(\epsilon_1 + \epsilon_2 - \epsilon'_1 - \epsilon'_2) W \times F(\epsilon'_2)(1 - F(\epsilon'_1))(1 - F(\epsilon'_2))$$

given in terms of an average particle-particle scattering rate  $W$  appropriate to the transport coefficient of interest.  $W$  now contains contributions from paramagnon exchange (Fig. 3) between the quasi-particles. The resulting temperature dependence may be written in the form (applicable in the cases of  $K$  and  $D$ )

$$\frac{1}{T^2 \tau} = a \left( \frac{1}{T_S^2} \right) - b \left( \frac{T}{T_S^3} \right) + \dots$$

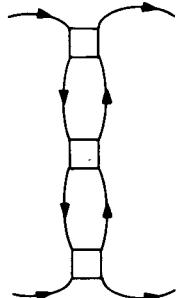


FIG. 3. Contributions to particle-particle scattering amplitudes.

This form of the temperature dependence [17] fits in beautifully with observations [18] of the transport coefficients of pure liquid  $^3\text{He}$  in the millidegree range.

As in the specific heat case, the leading contribution  $1/\tau \propto T^2$  comes from large momentum transfer scattering processes and cannot be estimated reliably. However, the higher temperature correction terms contain large contributions from the small momentum and energy transfer region and so can be more reliably estimated in terms of Landau parameters [15, 16].

## R E F E R E N C E S

- [1] DONIACH, S., ENGELBERG, S., Phys. Rev. Lett. 17 (1966) 750.
- [2] BERK, N.F., SCHRIEFFER, J.R., Phys. Rev. Lett. 17 (1966) 433.
- [3] BUCHER, E., BRINKMAN, W.F., MAITA, J., WILLIAMS, H.J., Phys. Rev. Lett. 18 (1967) 1125.
- [4] GIOVANNINI, B., PETER, M., SCHRIEFFER, J.R., Phys. Rev. Lett. 12 (1964) 736.
- [5] DONIACH, S., WOHLFARTH, E.P., Proc. R. Soc. A296 (1967) 442.
- [6] ABRIKOSOV, A.A., "Magnetic impurities in non-magnetic metals," these Proceedings.
- [7] LOW, G., HOLDEN, T., Proc. phys. Soc. 89 (1966) 119.
- [8] ABRIKOSOV, A.A., KHALATNIKOV, I.M., Rep. Prog. Phys. (Physical Society, London) 22 (1959) 329.
- [9] IZUYAMA, T., KIM, D.J., KUBO, R., J. Phys. Soc. Japan 18 (1963) 1025.
- [10] LOWDE, R., WINDSOR, C., Phys. Rev. Lett. 18 (1967) 1136.
- [11] BRENIG, W., MIKESKA, H.J., RIEDEL, E., Z. Physik 206 (1967) 439.
- [12] AMIT, D.J., KANE, J.W., WAGNER, H., Phys. Rev. Lett. 19 (1967) 425; and to be published.
- [13] IZUYAMA, T., KUBO, R., J. appl. Phys. 35 (1964) 1074.
- [14] BRINKMAN, W.F., ENGELBERG, S., Phys. Rev. 169 (1968) 417.
- [15] EMERY, V.J., Phys. Rev. 170 (1968) 205.
- [16] PETHICK, C.J., Physics Lett. 27A, (1960) 219; and to be published.
- [17] RICE, M.J., Phys. Rev. 159 (1967) 153; 162 (1967) 189.
- [18] WHEATLEY, J.C., Phys. Rev. 165 (1968) 304.



# TRANSPORT COEFFICIENTS OF A NORMAL FERMI LIQUID\*

C.J. PETHICK

Department of Physics,

University of Illinois,

Urbana, Ill., United States of America

## Abstract

TRANSPORT COEFFICIENTS OF A NORMAL FERMI LIQUID. An account is given of the difficulties encountered in the accurate calculation of transport coefficients for Fermi liquids, as a result of insufficient knowledge about the transition probability for collisions between quasi-particles and of uncertainties in the solutions of the transport equation.

In almost ferromagnetic Fermi liquids there are large finite-temperature corrections to the limiting low-temperature behaviour of the transport coefficients [1]. The occurrence of such corrections does not in itself signal a breakdown in Landau's theory of a normal Fermi liquid [2] and, in fact, they may be calculated rather precisely within the framework of Landau theory without invoking the concept of paramagnons.<sup>1</sup> These corrections are a general property of all Fermi liquids and are not peculiar to almost ferromagnetic systems.

In the past, the two things which have stood in the way of accurate calculations of transport coefficients for Fermi liquids have been, firstly, insufficient information about the transition probability for collisions between quasi-particles and, secondly, uncertainties in the solutions of the transport equation. As to the first difficulty, we shall see that the situation is more favourable for the finite-temperature corrections to the transport coefficients than it is for their values in the extreme low-temperature limit; the latter are determined by processes involving arbitrary momentum transfers, the scattering amplitude for which is not well known. However, calculations based on spin-fluctuation theory [4] show that processes in which the momentum transfer is small give rise to large finite-temperature corrections to the transport coefficients of almost ferromagnetic Fermi liquids. The observation which is crucial for the calculations described below is that the transition probabilities for small momentum transfer processes may be determined exactly in terms of the Landau parameters  $F_l^s$  and  $F_l^a$  discussed by Pines [2]; information about a number of these parameters may be obtained from measurements of equilibrium properties.

We now turn to the problem of solving the transport equation. For definiteness let us consider the calculation of the thermal conductivity.<sup>2</sup>

\* This research was sponsored by the Air Force Office of Scientific Research, Office of Aerospace Research, United States Air Force, under AFOSR contract/grant No. AF-AFOSR-328-67.

<sup>1</sup> This has also been pointed out by Emery [3]. A more complete list of references may be found in PETHICK, C.J., in *Lectures in Theoretical Physics 11*, Gordon and Breach, New York, to be published.

<sup>2</sup> We follow the calculation by Dy and Pethick [5].

The standard quasi-particle transport equation [2] has the form

$$\frac{\epsilon_1}{\kappa T} \vec{v}_1 \cdot \hat{u} n_1 (1 - n_1) = \frac{1}{\kappa T^2} \sum_{2,3,4} W_{12}^{34} n_1 n_2 (1 - n_3) (1 - n_4) \\ \times \delta_{\vec{p}_1 + \vec{p}_2, \vec{p}_3 + \vec{p}_4} \delta(\epsilon_1 + \epsilon_2 - \epsilon_3 - \epsilon_4)(\Phi_1 + \Phi_2 - \Phi_3 - \Phi_4) \quad (1)$$

where  $\epsilon_i$  is the energy (measured with respect to the chemical potential) and  $\vec{v}_i$  is the velocity of a quasi-particle of momentum  $\vec{p}_i$ ;  $\hat{u}$  is an arbitrary unit vector,  $n_i$  is the Fermi distribution function and  $n_i(1 - n_i)\Phi_i/\kappa T^3$  is the deviation from local equilibrium of the number of quasi-particles in the state  $i$ .  $W_{12}^{34}$  is the transition probability for the binary collision process  $1+2 \rightarrow 3+4$ , and for simplicity spin indices have been suppressed. After expressing the equation in terms of the dimensionless variables  $\epsilon_i/\kappa T$  it may be rewritten in a convenient vector notation as follows:

$$|X\rangle = I|\Phi\rangle \quad (2)$$

The vector  $|X\rangle$  corresponds to the left-hand side of the transport equation and the multiplication operation corresponds to summation over the states of a quasi-particle and multiplication by  $1/\kappa T$ . In this notation the thermal conductivity,  $K$ , is given by

$$KT = \langle X | \Phi \rangle$$

At low temperatures,  $I$  has an asymptotic expansion of the form  $I = I_0 + TI_1 + \dots$ , and by solving Eq.(2) by perturbation theory one can easily see that

$$\frac{1}{KT} - \frac{1}{KT} \Big|_{T=0} = T \frac{\langle \Phi_0 | I_1 | \Phi_0 \rangle}{\langle X | \Phi_0 \rangle^2} + \dots \quad (3)$$

where  $|\Phi_0\rangle$  is the solution of the equation  $|X\rangle = I_0|\Phi_0\rangle$ . In liquid  ${}^3\text{He}$  and almost ferromagnetic Fermi liquids, to which we apply our results, the contributions to  $I_1$  from small momentum transfer processes are particularly large as a result of spin-fluctuation effects, and we neglect other contributions to  $I_1$ . With the latter assumption,  $I_1$  may be expressed solely in terms of Landau parameters.  $|\Phi_0\rangle$  was found by solving the transport equation numerically<sup>3</sup> and  $\langle \Phi_0 | I_1 | \Phi_0 \rangle \langle X | \Phi_0 \rangle^2$  was calculated. Fortunately this quantity is relatively insensitive to the exact form of  $I_0$ , which is not well known.

The results in general are rather complicated, but for an almost ferromagnetic Fermi liquid ( $F_0^a \rightarrow -1$ ) the dominant term is given by

$$\frac{1}{KT} - \frac{1}{KT} \Big|_{T=0} \sim \frac{1215}{2} \pi \xi(5) \frac{m^*^3}{p_F^7} (A_0^a)^3 \gamma \kappa T$$

<sup>3</sup> Recently, analytical expressions for  $|\Phi_0\rangle$  have been obtained by Brooker and Sykes, and by Jensen et al. [6].

where  $m^*$  is the effective mass,  $p_F$  is the Fermi momentum,  $\xi(n)$  is the Riemann zeta function and  $A_0^a = F_0^a/[1+F_0^a]$ .  $\gamma$  is a quantity of order unity and depends only on  $|\Phi_0\rangle$ .

The  $(A_0^a)^3 T$  behaviour of the result may be understood in terms of the strong energy dependence of the scattering amplitude for processes involving a small momentum transfer; two powers of  $A_0^a$  come from the transition probability and, as we shall explain below, the remaining factor  $A_0^a T$  comes from the fraction of scattering processes for which dynamical screening is important. For an almost ferromagnetic Fermi liquid the dominant contribution to the scattering amplitude for small momentum transfer processes is of the form  $F_0^a/(1+F_0^a \chi(s))$ , where  $s = \omega/v_F q$ ,  $\omega$  being the energy transfer,  $q$  the momentum transfer and  $v_F$  the Fermi velocity.  $F_0^a$  plays the role of a bare interaction and  $1+F_0^a \chi(s)$  is a shielding factor.  $\chi(s)$  is given by  $1 - s/2 \ln(s+1)/(s-1) \approx 1 + i\pi/2 s$ , ( $s \ll 1$ ). For small values of  $s$  ( $\ll 1$ ) the effective interaction may be approximated by  $A_0^a/[1 + \frac{1}{2}i\pi A_0^a s]$ . This shows that for an almost ferromagnetic Fermi liquid ( $F_0^a \rightarrow -1$ , or  $A_0^a \rightarrow -\infty$ ) the static interaction is greatly enhanced by screening. However, the interaction is a very strong function of frequency, and the enhancement falls off rapidly when  $\omega \gtrsim v_F q / |A_0^a|$ . At a temperature  $T$  the energy transfer in a collision is typically of the order of  $kT$ , and therefore the frequency dependence of the interaction will be important for processes in which  $q \lesssim q_0 = |A_0^a|kT/v_F$ . The fraction of collision processes for which dynamical screening is important varies as  $(p_F)^{-1} \sum_{q(q \lesssim q_0)} q^{-2} \sim |A_0^a|kT/v_F p_F$ , since

the number of possible collisions for a given energy transfer and momentum transfer  $q$  varies as  $1/q^2$ . This latter result is obtained most easily by noting that the collision  $1+2 \rightarrow 3+4$  may be regarded as the scattering of a quasi-particle - quasi-hole pair from the state  $(1, 3)$  to the state  $(4, 2)$ . The density of initial and final pair states of a given energy each vary as  $1/q$ , which accounts for the weight factor  $1/q^2$  in the sum.

Although the results for an almost ferromagnetic Fermi liquid are not applicable to liquid  $^3\text{He}$ , there are large finite-temperature corrections to the thermal conductivity, and the experimental results are shown in Fig. 1. Making the popular assumption that all Landau parameters vanish for  $\ell \geq 2$  and using values of  $F_0^s$ ,  $F_1^s$  and  $F_0^a$  obtained from measurements of equilibrium properties, we were able to estimate the previously undetermined Landau parameter  $F_1^a$ . The value obtained is -0.47. Similar calculations for the finite-temperature contributions to the spin diffusion coefficient have also been performed; the value of  $F_1^a$  obtained from the experimental data is -0.39 which is not inconsistent with that obtained from thermal conductivity.

In summary we make three remarks:

(a) Although contributions to  $1/KT$  of order  $T$  and similar contributions to other transport coefficients were first found in studies of almost ferromagnetic Fermi liquids, such contributions are to be expected in all normal Fermi liquids.

(b) In the past, microscopic information obtained from transport coefficient measurements has been somewhat uncertain due to the use of approximate solutions of the transport equation. Now that exact solutions

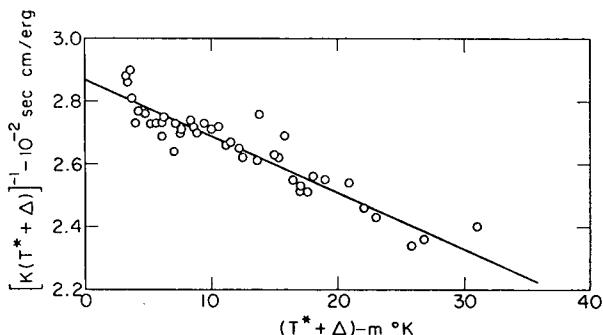


FIG.1. The thermal conductivity of liquid  ${}^3\text{He}$  at low pressure.

The data are those of Abel et al. [7].  $T^* + \Delta$ , where  $T^*$  is the magnetic temperature and  $\Delta = 0.3 \text{ m } ^\circ\text{K}$ , is an effective temperature which Abel and Wheatley [8] believe to be very close to the Kelvin temperature,  $T$ , in the range of temperatures of interest. Note how large the linear term is on a scale of the Fermi temperature,  $p_F^2/2m^*\kappa$ , which is  $\sim 1700 \text{ m } ^\circ\text{K}$ . The line shown is a fit to the data, and from its slope we find  $F_1^a = -0.47 \pm 0.14$ , where the quoted error is only that due to uncertainties in fitting a straight line to the data.

of the transport equation are available, measurements of transport coefficients for Fermi systems will provide valuable information about the interactions between quasi-particles.

(c) A number of problems related to finite-temperature contributions to the transport coefficients still remain to be solved, and we mention but two. Firstly, there is a need for an investigation of the limits of validity of the quasi-particle transport equation and secondly, contributions coming from processes in which the momentum transfer is not small should be considered in detail.

#### A C K N O W L E D G E M E N T

The work described here was performed in collaboration with Dr. K.S. Dy.

#### R E F E R E N C E S

- [1] DONIACH, S., these Proceedings.
- [2] PINES, D., these Proceedings; see also PINES, D., NOZIÈRES, P., *The Theory of Quantum Liquids* 1, W.A. Benjamin, Inc., New York (1966).
- [3] EMERY, V.J., Phys. Rev. 170 (1968) 205.
- [4] RICE, M.J., Phys. Rev. 159 (1967) 153; 162 (1967) 189.
- [5] DY, K.S., PETHICK, C.J., Phys. Rev. Lett. 21 (1968) 876.
- [6] BROOKER, G.A., SYKES, J., Phys. Rev. Lett. 21 (1968) 279.
- JENSEN, H.H., SMITH, H., WILKINS, J.W., Physics Lett. 27A (1968) 532.
- [7] ABEL, W.R., JOHNSON, R.T., WHEATLEY, J.C., ZIMMERMANN, W., Phys. Rev. Lett. 18 (1967) 737.
- [8] ABEL, W.R., WHEATLEY, J.C., Phys. Rev. Lett. 21 (1968) 597.

# MAGNETIC IMPURITIES IN NON-MAGNETIC METALS

A. A. ABRIKOSOV

Landau Institute for Theoretical Physics,  
Moscow, USSR

## Abstract

MAGNETIC IMPURITIES IN NON-MAGNETIC METALS. 1. Introduction; 2. localized spins; 3. ordering; 4. superconductivity (S); 5. superconductivity and ordering (SO); 6. Kondo effect (K); 7. Kondo effect and ordering (KO); negative magneto-resistance; 8. Kondo effect and superconductivity (SK).

## 1. INTRODUCTION

The subject of my communication seems to be rather narrow at first glance. However, the number of articles concerning various problems in this field during the last years is very large and in my opinion the influx of papers is increasing. The first reason for this is that there remain many mysterious phenomena. Apart from that, there are many cases where a small magnetic impurity in a non-magnetic metal can lead to a considerable change in its properties. For example, the low-temperature electronic specific heat increases several times and the thermoelectric power becomes 2-3 orders of magnitude larger than in a pure metal. These experimental facts have only recently been understood, but many questions remain unanswered.

## 2. LOCALIZED SPINS

At the present moment there is no doubt that all the peculiarities of dilute solutions of intermediate metal atoms in non-magnetic metals are connected with the fact that such atoms which, in the isolated state, possessed unclosed d- or f-shells and a finite spin, often retain this property when solved in a non-magnetic metal. Extensive literature is now available concerning the possibility of such localized spins in metals and the main ideas in this field are due to Friedel and Anderson. I have no opportunity to consider this question in detail, but it is necessary to say, at least, a few words on the subject. The following is an extract. When a magnetic atom is placed in a metal we can no longer speak of separate energy levels. Instead, we can say that there is an increase of the state density in some narrow interval of the conduction band. Of course, this happens only if the impurity energy level gets into the conduction band, but since this band is very wide, this happens often enough. These new states are rather similar to the original orbital states in an isolated atom; therefore, the electrons in such states spend a long time close to the impurity core and to each other and interact strongly. In an isolated atom the Coulomb interaction leads to the well-known Hund's rules, according to which an unclosed shell

must have the maximum spin possible. Since, as previously mentioned, the Coulomb interaction remains to a considerable extent, even when the impurity atom is placed in a metal, Hund's rules can also act in this case. Everything depends on the time which the electron spends close to the impurity. Quantitatively, the possibility of formation of localized spin in a metal is defined by the relation between the Coulomb interaction of the electrons in a d- or f-shell and the width of the former discrete energy level which reflects the possibility of the transition into the conduction band. How this happens and the changes that occur in the properties of conduction elections were demonstrated by Anderson for a very simple model. Let us write the Hamiltonian

$$\begin{aligned} H = & \sum_{\vec{p}\alpha} \epsilon_{\vec{p}} n_{\vec{p}\alpha} + \sum_{\alpha} \epsilon_d n_{d\alpha} + \frac{1}{2} \sum_{\alpha} U n_{d\alpha} n_{d-\alpha} \\ & + \sum_{\vec{p}\alpha} \left( V a_{\vec{p}\alpha}^{\dagger} d_{\alpha} + V^* d_{\alpha}^{\dagger} a_{\vec{p}\alpha} \right) \end{aligned} \quad (1)$$

Here  $a_{\vec{p}\alpha}$  and  $d_{\alpha}$  are destruction operators,  $n_{\vec{p}\alpha} = a_{\vec{p}\alpha}^{\dagger} a_{\vec{p}\alpha}$ ,  $n_d = d^{\dagger} d$ , and the energies  $\epsilon_{\vec{p}}$  and  $\epsilon_d$  are measured from the Fermi level, the impurity level  $\epsilon_d$  is assumed to be non-degenerate, and  $\epsilon_d < 0$ . The term containing  $U$  is a part of the Coulomb interaction between the electrons which serves to fulfil Hund's rules; the last term describes the mixing between the  $\epsilon_d$  level and the conduction band;  $\vec{p}$  is the quasimomentum. According to Anderson's estimates  $U \sim 10$  eV and  $V \sim 2$  eV. Therefore, to the first approximation, we can neglect the last term. It is easy to see that if there is one electron at the  $\epsilon_d$  level with spin up, another electron can only have the spin down. However, in this case its energy will be  $\epsilon_d + U$ . If this is a positive quantity, the energy of such an electron will be above the Fermi level and this state will remain empty. Obviously the same argument is valid if at the  $\epsilon_d$  level there is an electron with its spin down. So the term with  $U$  gives us a localized spin  $\frac{1}{2}$ . Of course, the  $V$ -mixing term makes the situation more complicated and there arises in fact a limited region in the plane of  $V$  and  $U$  where the localized spin is possible.

We shall leave these considerations and, using Anderson's Hamiltonian, consider instead the scattering of an s-electron at the impurity with a spin flip. Let the  $\epsilon_d$  level be occupied by an electron with spin up. A conduction electron with spin  $\frac{1}{2}$  and quasimomentum  $\vec{p}$  is scattered by the impurity; its momentum and spin become  $\vec{p}'$ ,  $\frac{1}{2}$ . This can proceed in two ways:

- (a) the electron  $\vec{p}, -\frac{1}{2}$  becomes  $d, -\frac{1}{2}$  and then the electron  $d, \frac{1}{2}$  becomes  $\vec{p}', \frac{1}{2}$ . The transition amplitude is then

$$- \frac{|V|^2}{\epsilon_{\vec{p}} - \epsilon_d - U}$$

- (b) the electron  $d, \frac{1}{2}$  becomes  $\vec{p}', \frac{1}{2}$  and then the first electron  $\vec{p}, -\frac{1}{2}$  becomes  $d, -\frac{1}{2}$ . In this case we get

$$- \frac{|V|^2}{\epsilon_d - \epsilon_{\vec{p}'}}$$

Assuming the scattering to be elastic and the metal to be uniform we have  $\epsilon_{\vec{p}'} = \epsilon_{\vec{p}}$ , and so the total amplitude becomes

$$\frac{|V|^2 U}{(U + \epsilon_d - \epsilon_{\vec{p}})(\epsilon_{\vec{p}} - \epsilon_d)}$$

Usually only the electrons close to the Fermi level are important, so  $\epsilon_{\vec{p}} \approx 0$ . Then we get a matrix element independent of  $\vec{p}$  and equal to

$$\frac{|V|^2 U}{(U + \epsilon_d)(-\epsilon_d)}$$

This is a positive quantity since for the existence of a localized spin it is necessary to have  $\epsilon_d < 0$ ,  $U + \epsilon_d > 0$  in any case.

Three things must be mentioned. First, the result is equivalent to an exchange interaction between the conduction electron and the impurity spin of the form

$$- \frac{J}{N} \sum_{\vec{p}\alpha} a_{\vec{p}\alpha}^\dagger \sigma_{\alpha\alpha}^i a_{\vec{p}'\alpha'} a_{\vec{p}'\alpha'}^\dagger S_{\beta\beta}^i d_\beta,$$

where

$$\frac{J}{N} = - \frac{2|V|^2 U}{(U + \epsilon_d) |\epsilon_d|}$$

Here  $N$  is the atom density of the host metal which is introduced for normalization. The second is that according to our derivation  $J < 0$ , that means anti-ferromagnetic coupling. Finally,  $\epsilon_d$  or  $U + \epsilon_d$  can lie close to the Fermi level and in this case  $J$  will not be a small quantity. This is essential since usually the  $(\vec{S}_1 \cdot \vec{S}_2)$  interactions are of true exchange nature and therefore several times less than the Fermi energy. The rule that  $J < 0$  is true if the considered interaction is the most important, i.e. always when  $J$  is not small. But in the case of rare-earth impurities (i.e. unclosed f-shells) the true exchange interaction of f- and s-electrons can also be of importance and the sign of  $J$  may be positive as well.

The equivalence between the Anderson Hamiltonian and the "exchange"  $\vec{\sigma} \cdot \vec{S}$  interaction was at first demonstrated by Kondo approximately in the same way as I did it here, and after that Schrieffer and Wolf obtained it rigorously by means of a canonical transformation.

Of course, the Anderson's Hamiltonian is the simplest model, but even if one assumes the level  $\epsilon_d$  to be degenerate and corresponding to an orbital momentum  $\ell$ , the Hamiltonian can nevertheless be transformed to the  $\vec{\sigma} \cdot \vec{S}$  form. In this case the interaction energy  $J(\vec{p} \cdot \vec{p}')$  becomes essentially dependent on the momenta  $\vec{p}$  and  $\vec{p}'$ . According to Kondo it is proportional to  $P_\ell(\cos \theta_{\vec{p}\vec{p}'})$ . B. Caroli (Orsay) has shown that this is essential for numerical comparison with various experimental data since different quantities contain different angle integrals of  $J$ . However, I shall not go

into that. For the sake of simplicity we shall assume  $J = \text{const}$  everywhere.

Now we shall pass to various phenomena connected with the localized spins. There are many of them, and it will be difficult to cover everything. Therefore, only a few which seem to be of principal importance will be mentioned. These subjects can be represented by a matrix (Fig. 1).

O		
OS	S	
OK	SK	K

FIG. 1. Matrix showing subjects connected with localized spins. O - ordering; S - superconductivity; K - Kondo effect.

### 3. ORDERING (O)

The interaction Hamiltonian between the localized spins and conduction electrons

$$H_{\text{int}} = - \frac{J}{N} \sum_n \psi_\alpha^\dagger(\vec{R}_n) \sigma_{\alpha\alpha}^i \psi_\alpha(\vec{R}_n) \hat{S}^i \quad (2)$$

where the sum is taken over all impurity atoms, leads to an interaction between the impurity spins themselves: an impurity spin polarizes the electrons and the latter polarize another impurity spin. If we take  $|J| \ll \epsilon_F$  we can consider only the lowest order of the perturbation theory. The interaction process of two impurity atoms by means of electrons can be described by a picture (Fig. 2).

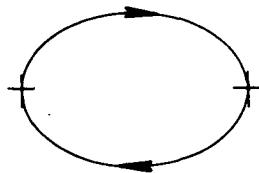


FIG. 2. Interaction of two impurity atoms by means of electrons. An impurity atom produces an electron-hole pair which vanishes, interacting with the other impurity atom.

At the same time, the figure is a Feynman graph. A simple calculation gives the well-known formula of Ruderman, Kittel, Kasuya, Yosida (RKKY)

$$H_{S_1 S_2} = - \left( \frac{J}{N} \right)^2 v(\epsilon_F) f(\vec{R}_1 - \vec{R}_2) \vec{S}_1 \cdot \vec{S}_2 \quad (3)$$

where  $\nu(\epsilon_F) = p_0 m / 2\pi^2$  is the state density at the Fermi level,  $p_0$  the Fermi momentum

$$f(R) = \frac{1}{4\pi} \left[ \frac{\sin 2p_0 R}{2p_0 R^4} - \frac{\cos 2p_0 R}{R^3} \right] \quad (4)$$

(for simplicity everywhere we put  $\hbar = 1$ ). The Fourier transform of the  $f(R)$ -function we call  $\chi(q)$ . It is equal to

$$\chi(q) = \frac{1}{2} + \frac{4p_0^2 - q^2}{8p_0 q} \ln \left| \frac{2p_0 + q}{2p_0 - q} \right| \quad (5)$$

Hence the interaction (3) decreases by its magnitude like  $1/R^3$  and at the same time is rapidly oscillating with a period  $1/2p_0$ , of the order of magnitude of interatomic distances. Since the impurity atoms are randomly placed, this interaction for some of the atom pairs is ferromagnetic and tends to put their spins parallel and for the others it is anti-ferromagnetic. What kind of ordering can arise from such an interaction? There are two different approaches to this problem. Each is supported by many research workers, but the question is far from clear.

One of these approaches is due to Blandin, Friedel and Marshall. The idea is that the oscillating and rapidly decreasing interaction cannot establish a long-range order in a metal, and therefore, a state is formed where the spins are more or less randomly oriented. This configuration at sufficiently low temperatures is stable but there is, of course, no sharp phase transition, and all the singularities of the thermodynamic quantities, such as specific heat or magnetic susceptibility, will have the form of smooth maxima. Every spin can be considered to be under the influence of some effective exchange field and to have the energy  $-\vec{S} \cdot \vec{Q}$  (the quantity  $\vec{Q}$  plays the role of  $g\mu_B \vec{H}$  where  $\vec{H}$  is some magnetic field and  $g$  is the gyromagnetic ratio). A distribution function of the magnitudes of the effective field is introduced  $p(Q)$  (obviously it does not depend on the direction of  $\vec{Q}$ ). After that it is assumed that all the spins can be treated as independent and all the quantities obtained must be averaged by  $Q$  with the distribution  $p(Q)$ .

Of course it is necessary to know the  $p(Q)$ . No-one has succeeded in calculating it. However, Klein and Brout did the calculation for the so-called Ising model, i.e. for an interaction  $S_{1z}S_{2z}$  instead of  $\vec{S}_1 \cdot \vec{S}_2$  and every spin was assumed to be  $\frac{1}{2}$ . Such a substitution is of course a great disadvantage of the calculation.

There exists however a more general dimensional argument due to Blandin. At large distances the amplitude of the RKKY interaction falls like  $1/R^3$ . The average distance between the impurities changes proportionally to  $c^{-\frac{1}{3}}$  ( $c$  is the concentration). Hence one can suppose that the function  $p(Q)$  has such a form that all the average energies (such as  $\sqrt{Q^2}$  or some characteristic temperatures) are proportional to  $1/\bar{R}^3$  or to  $c$ . That demands the distribution  $p(Q)$  to have the form

$$p(Q, T) = \frac{1}{c} f\left(\frac{Q}{c}, \frac{T}{c}\right) \quad (6)$$

This is already enough to have many consequences. For example, if  $p(0, 0)$  is finite, then at low temperatures the impurity parts of the specific heat and magnetic susceptibility are

$$\Delta C = \frac{\pi^2}{6} N c p(0, 0) \frac{4S^2}{2S+1} T \quad (7)$$

$$\Delta \chi = 2 p(0, 0) \mu^2 N c \quad (8)$$

where  $\mu$  is the magnetic moment of the impurity atom ( $g \mu_B S$ ),  $c$  the concentration,  $N$  the density of atoms of the host metal. According to Eq. (6) we have  $p(0, 0) \sim 1/c$ . It follows that  $\Delta O$  and  $\Delta \chi$  do not depend on the impurity concentration. Since  $\Delta C$  depends linearly on the temperature this term gives the impression that the linear electronic specific heat increases several times, and this increase does not depend on the impurity concentration. The same is also true for susceptibility. Of course, the concentration defines the "Curie temperature", i.e. the characteristic temperature values after which no ordering is left. Its order of magnitude is  $\theta \sim C J^2 / \epsilon_F$  and so the temperature interval where this enhancement exists is proportional to the concentration, but the enhancement itself does not depend on  $C$ .

Dividing  $\Delta C$  by  $\Delta \chi$  we obtain

$$\frac{\Delta C}{\Delta \chi} = \frac{\pi^2}{3} \frac{S^2}{\mu^2 (2S+1)} \quad (9)$$

No  $p(0, 0)$  is left here. If we assume  $\mu \approx 2\mu_B S$  we can define the impurity spin and compare it with the results for  $T \gg \theta$ , where, obviously, we have paramagnetism and the susceptibility is given by

$$\Delta \chi = \frac{4\mu_B^2 S(S+1)}{3T}$$

The results for Au + Fe, Cu + Mn obtained in Grenoble are in excellent agreement with these considerations. The Grenoble group (Dreyfus, Tournier, Souletie, et al.) also did many other measurements to verify the consequences of Blandin's law for  $p(Q)$ ; everywhere good agreement was achieved.

Also in favour of this model is the smoothness of all transitions (broad maxima instead of sharp singularities) and a large magnetization remanence. The latter is explained in the following manner. In a large magnetic field some definite spin orientations are established, i.e. some function  $p(\vec{Q})$  exists. In this case it is anisotropic. This function is "hard", which means that it will not follow immediately a decreasing field, so considerable remanence must exist.

However, there are two important objections. The first is that since  $Q$  is a randomly oriented vector its distribution function should be

$p(Q) dQ = p_1(Q) Q^2 dQ$ . And even if  $p_1(Q) \neq 0$ ,  $p(0) = 0$ . But this destroys completely all the success of this model. The other objection will be discussed a little later.

Let us now deal with the other approach. This is the so-called spin density wave (SDW) invented by Overhauser. Let us suppose that the conduction electrons are re-distributed in such a fashion that a periodic distribution of spin density is established in space (it may be, for example, a uniform distribution, i.e. ferromagnetism). Of course, this will demand an increase of the electron energy, but if we now average the Hamiltonian (2) over this new distribution we again obtain  $-\sum \vec{Q}(\vec{R}_n) \vec{S}_n$  where

$$\vec{Q}(\vec{R}_n) = \frac{J}{N} \langle \psi_\alpha^\dagger(\vec{R}_n) \vec{\sigma}_{\alpha\alpha'} \psi_{\alpha'}(\vec{R}_n) \rangle$$

This effective field has different orientations at different points but the impurity spin is always oriented in the proper direction, and hence the energy is decreased. This decrease can overcome the original increase of the electron energy. This can be shown to happen below some Curie point  $\theta$  having the order of magnitude  $cJ^2/\epsilon_F$ . In this case the transition must be sharp. The observed smooth transitions can be ascribed to the non-uniformity of concentration in a real sample.

What kind of SDW will actually be established? Obviously this is defined by the minimal energy. Let us suppose that the wave has a helicoidal form (i.e. the point of the  $\vec{Q}$ -vector moves along a spiral trajectory in space with a wave vector  $\vec{q}$ ). In this case the energy gain becomes (per unit volume)

$$\Delta E = - (cJ)^2 \nu(\epsilon_F) S^2 X(q) \quad (10)$$

where the function  $X$  is defined by formula (5). It was assumed here that the metal is uniform and the impurities are distributed entirely at random. The function  $X(q)$  is monotonous and its maximum value corresponds to  $q = 0$ . So in such a model ferromagnetic alignment is preferable. Of course, the electron-electron interaction can change this result, and Overhauser himself found this effect, but this can only happen if the Coulomb interaction is not sufficiently screened.

There exists, however, another point of greater importance. In this model we did not take into account the fact that the impurity atoms occupy definite positions in the crystal cells and therefore the averaging can only be done over the cells and not over the position in a cell. Accordingly we obtain

$$\Delta E = - (cJ)^2 \nu(\epsilon_F) S^2 \frac{1}{N} \sum_{\substack{\vec{R}_n \neq 0}} f(\vec{R}_n) \cos \vec{q} \cdot \vec{R}_n \varphi(\vec{R}_n) \quad (11)$$

where the sum is taken over the whole lattice, and the function  $\varphi(\vec{R}_n)$  accounts for the correlation between the cells where the impurity atoms can stick. The importance of this function is due to the fact that in the sum over  $\vec{R}_n$  the most important distances are of the order of interatomic distances, where  $\varphi(\vec{R}_n)$  can deviate strongly from unity, since the solid solution must

conserve a remanence of the correlation in the liquid state. However, since little can be said about this function it is usually taken to be unity. After that the sum becomes equal to

$$\lim_{P \rightarrow \infty} \left[ \sum_{|\vec{K}| < P} \chi(\vec{q} + \vec{K}) - \frac{1}{N} \int_{|\vec{K}| < P} \frac{d^3 \vec{K}}{(2\pi)^3} \chi(\vec{K}) \right]$$

where the  $\vec{K}$ -sum is taken over all the reciprocal lattice points. It is convenient to subtract the expression for the ferromagnetic case and after that the limit  $P \rightarrow \infty$  can be taken. We get

$$\Delta E - \Delta E_{\text{ferr}} = - (cJ)^2 \frac{p_0 m}{2\pi^2} S^2 \sum_{\vec{K}} \left[ \chi(\vec{q} + \vec{K}) - \chi(\vec{K}) \right] \quad (11')$$

Calculations for particular types of lattices and valence electron numbers per atom show that this sum has often a maximum at  $\vec{q} \neq 0$ . Moreover, here the energy of anisotropy was not taken into account and the latter will tend to bind the period of the wave to the lattice period. Therefore, it is probable that in reality we shall not have the  $\vec{q}$  corresponding to the minimum of (11') but another, corresponding to the nearest  $\vec{K}$ . Such a helicoid is the simplest example. Other types of SDW are of course also possible.

To our knowledge the physical properties of this model have not been studied well enough. In the case when the effective field changes its magnitude (e.g. plane SDW:  $Q^x = A \cos \vec{q} \cdot \vec{R}$ ,  $Q^y = Q^z = 0$ ) then in principle there is no difference from the previous approach since there is also some  $p(Q)$  and it can be shown that Blandin's correspondence principle is fulfilled. It may be that this is the only possible way to get  $p(0) \neq 0$ . However, if there is ferromagnetism or a helicoid then the effective field only changes its direction but not its magnitude and  $P(Q) \sim \delta(Q - Q_0)$ . One can argue that according to the previous consideration this contradicts the experiment since the impurity part of the specific heat cannot be linear in this case. However, there are in fact no contradictions.

Kondo was the first to mention that the presence of impurities changes the energy spectrum of conduction electrons (Fig. 3). In the vicinity of the Fermi energy in a region of the order  $\theta \sim cJ^2/\epsilon_F$  the effective mass is enhanced. Kondo introduced the effective field and his expression for  $\Delta m/m$  was proportional to  $\langle c/Q \rangle_{\text{av}}$ . Since  $Q$  itself is proportional to  $c$  so the change of the effective mass does not depend on concentration and is of the order of unity.

This results in an increase of the coefficient in the linear electronic specific heat. Therefore, the experimental data can be explained not in terms of the impurity part but as the enhancement of the effective mass of the electrons. Klein later mentioned that if Kondo's expression for the specific heat is averaged over  $p(Q)$  then the integral contains  $\int (P(Q)/Q) dQ$  and if  $p(0) \neq 0$  then this integral diverges logarithmically. Therefore, the specific heat becomes proportional to  $T \ln \theta/T$  which is not observed in experiment. This is the second argument against the short-range order approach.

In fact, Kondo's argument must be improved. The correction considered by him is the first of an infinite series. In the case of long-range order these corrections are the interaction of the electron with the spin waves, and the mass enhancement has the same origin as in the case of electron-phonon interaction. For ferromagnetic ordering the additional specific heat is proportional to  $T \ln \theta/T$  and this seems to argue against ferromagnetism. For all anti-ferromagnetic structures the correlation is proportional to  $T$ . In every case it does not depend on concentration. We do not know the exact answer for the random field model but it seems that Kondo's result is qualitatively true and that means a strong argument against the Blandin-Friedel-Marshall concept.

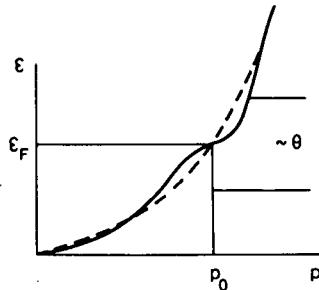


FIG.3. Changes in the energy spectrum due to the presence of impurities.

Finally, the experimental results regarding the measuring of the hyperfine field at the nucleus by means of the Mössbauer effect will be treated. The results of Violet and Borg for small impurities of iron in gold show that the iron atoms occupy definite positions in the crystal cells (seen from the sharp character of quadrupole splitting of the line at temperatures  $T \gg \theta$ ). Secondly, at low temperatures the results show that the effective field acting on every impurity atom is of the same value. It is also important that the line width does not change with concentration. It should be mentioned, however, that the transition temperature depends on concentration as  $c^\alpha$  where  $\alpha < 1$  (in different measurements different numbers are found). That may argue against the simple approach (e.g. the correlation of impurity positions may be important). The results for Cu Fe show that in this case there is not a δ-like form of  $P(Q)$ . This may be interpreted in favour of the short-range approach or it may mean that there is another type of SDW.

Finally, we have to mention that the first approach has been worked out in greater detail than the second one, and therefore it is difficult to draw any definite conclusions. We prefer the second approach which so far does not collide with any serious objections (the non-linear dependence of  $\theta$  on concentration in Au Fe contradicts both of the concepts but if this effect is real it could, in my opinion, be explained more easily in the framework of the long-range order approach).

#### 4. SUPERCONDUCTIVITY (S)

In this section only those phenomena will be discussed which are not connected with ordering. The measurements of Matthias and his group on

lanthanum with small rare-earth impurities have shown that superconductivity is very sensitive to the presence of localized spins. The critical temperature falls rapidly with increasing concentration of magnetic impurity and one per cent is sufficient to destroy superconductivity entirely. Ordinary impurities produce a rather small effect on the transition temperature. The first explanations of this effect are due to Herring, Matthias and Suhl, who also calculated the change of  $T_c$  for small concentrations. Afterwards, Gorkov and the author of this paper constructed a theory based on the method of temperature Green's functions which held for all concentrations (in the absence of ordering).

Physically, the effect of small magnetic impurities is connected with the spin-flip scattering of electrons which was mentioned at the beginning of this paper. As is well known, the electrons in a superconductor are bound in the so-called Cooper pairs with opposite spins. The dimensions of such pairs or the correlation length in a pure metal has the order of  $\hbar v/T_c$ , i.e.  $10^{-4} - 10^{-5}$  cm. Being scattered at the impurity, an electron entering such a pair can flip its spin, and the pair becomes unstable. So it is clear that a spin-flip scattering must suppress superconductivity. This happens in fact. The change of critical temperature was found to be

$$\ln \frac{T_{c0}}{T_c} = \psi\left(\frac{1}{2} + \frac{\rho}{2}\right) - \psi\left(\frac{1}{2}\right) \quad (12)$$

where  $\rho = 1/\pi\tau_s T_c$ ,  $1/\tau_s = (1/6)v(\epsilon_F)(J^2/N)cS(S+1)$  is the spin-flip collision time,  $\psi(x) = [\ln \Gamma(x)]'$ . For small concentrations this becomes

$$T_c = T_{c0} - \frac{\pi}{4\tau_s} \quad (13)$$

If the concentration becomes  $c_{crit}$  such that

$$\tau_{s crit} = \frac{1.1}{T_{c0}} \quad (14)$$

$T_c$  turns to zero.

The A-G theory produced the phenomenon of gapless superconductivity. At concentrations approximately equal to  $0.9 n_{crit}$  the gap in the energy spectrum disappears. This means that even for  $T = 0$  such a superconductor is able to absorb quanta of arbitrarily small energy and its specific heat will be linear at low temperatures. However, the property of zero resistance for an electric current will remain. The reason for this is that the so-called order parameter which measures the number of Cooper pairs in the Bose condensate must not always be proportional to the energy gap as it is in a pure superconductor. It may be said that, as in a non-ideal Bose gas at  $T = 0$ , not all the particles are in the condensate; in the case considered, some of the Cooper pairs are not at the level with the smallest energy. They are distributed among the higher levels, and at a certain

concentration they reach the level corresponding to the decay of a pair in two separate electrons. We must mention here that after our prediction it was found that the gapless superconductivity is not a rare phenomenon and can happen on various occasions.

Then, Reif and Woolf, by means of a tunnelling experiment, measured the state density and compared it with the theory. In the cases where the impurity has in fact a localized spin (e.g. Gd in Pb) the agreement is good. According to these measurements the gapless superconductivity begins at smaller concentrations than was predicted by the A-G theory. But Fulde and Maki have shown that this is possibly due to ordering.

Many calculations were performed afterwards on the properties of superconductors with magnetic impurities, e.g. the heat conductivity and sound absorption, but I am not able to discuss these in a short time.

## 5. SUPERCONDUCTIVITY AND ORDERING (SO)

From the formulas for the critical temperature it follows that  $T_c$  changes considerably for such concentrations when  $1/\tau_s \sim T_{c0}$ . But the order of magnitude of  $1/\tau_s$  is  $cJ^2/\epsilon_F$  and that is just  $\theta$ , the ordering temperature. Therefore, the intersection of the curves  $T_c(c)$  and  $\theta(c)$  must have the form shown in Fig. 4.

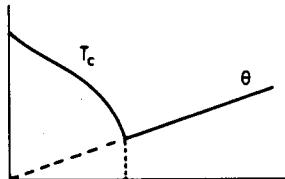


FIG. 4. Intersection of  $T_c(c)$  and  $\theta(c)$ .

A question arises: what will happen after the intersection? And, particularly, can there exist a phase which is superconducting and ordered at the same time? So far, this question was always solved only for the case of ferromagnetic ordering.

Gorkov and Rusinov have shown that the co-existence of superconductivity and impurity ferromagnetism is possible. If the spin-orbit interaction is strong enough (it produces a term  $\vec{\sigma} \cdot [\vec{p} \times \vec{p}']$  in the scattering amplitude even for non-magnetic impurities) then according to their prediction the curves  $\theta$  and  $T_c$  do not change after the intersection and all the area under them is occupied by the S-F phase. This prediction was corrected by Benneman, who drew attention to the change of the time  $\tau_s$  when the impurity spin becomes polarized by the exchange field. Since in this case the spin-flip scattering is suppressed (the total spin is conserved and the impurity spin is fixed by the exchange field) so the true  $T_c$  curve may go in this case even higher than the curve without ordering (Fig. 5). Fulde and Maki calculated the critical temperature and the main properties for the case of ferromagnetic ordering, strong spin-orbit coupling in the presence of an external magnetic field. They have found that this essentially leads to the change of  $1/\tau_s$  entering the A-G theory to the combination

$$\frac{1}{\tau_s} + \frac{\tau_{tr} v_F^2 eH}{3} + \frac{I^2 \tau_{so}}{2} \quad (15)$$

where the second term is the consequence of the curving of the pair trajectories by the external magnetic field,  $\tau_{tr}$  is the ordinary collision time entering the normal conductivity,  $v_F = p_0/m$  is the velocity at the Fermi level. The last term comes from the ordering. Here  $I = N_c J \bar{S}_z$  and  $\tau_{so}$  is the spin-orbit interaction collision time. The reduction of  $\tau_{so}$  reduces the effect of ordering. It is interesting that there exists a range of concentrations where the critical field temperature dependence has a maximum instead of a monotonous behaviour. This comes from the fact that the magnetization  $I$  depends on the external field. This was observed experimentally by Crow, Guerstin and Parks.

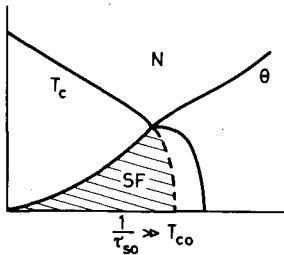


FIG.5. Intersection of  $T_c(c)$  and  $\theta(c)$  in case of suppression of spin-flip scattering.

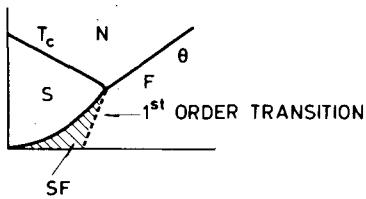


FIG.6. Intersection of  $T_c(c)$  and  $\theta(c)$  for small spin-orbit interaction.

In the opposite limit, when the spin-orbit interaction is small the co-existence region of superconductivity and ferromagnetism is much smaller and there exists a region of first-order phase transitions (Fig. 6). This case was studied in detail by de Gennes, Sarma and Cyrot. Here also a non-uniform slope of the  $H_c(T)$  curve is possible. All these cases have many interesting details which I have to skip. Instead of that I would like to mention two interesting phenomena.

The first is the formation of electron pairs with non-zero momentum. Fulde and Ferrell (and independently Larkin and Ovchinnikov) proposed the following idea. If a system of electrons is magnetized, e.g. by the impurity ferromagnetism, then at  $T = 0$  there will be a first-order phase transition at  $I = \Delta_0/\sqrt{2}$  if no scattering processes are considered. Here  $I = N_c JS$ , and  $\Delta_0$  is the superconducting gap. However, it occurs that at a larger value of  $I = 0.755 \Delta_0$  there exists a possibility of a second-order phase transition into a very peculiar superconducting phase where the Cooper pairs have a finite momentum and therefore  $\Delta$  depends on co-ordinates with a characteristic period  $q_0 = 2.4I/v_F$ . This new phase has many interesting features, e.g. if  $I \ll \Delta_0$  the specific heat depends on the temperature like  $1/\ln^3 \Delta_0/T$ .

Unfortunately, this phase seems to be impossible under real conditions because the free path length of the electrons must be larger than the period of the structure. This is impossible to have even in the case where there are no other impurities except the magnetic ones since the potential interaction of the electrons with impurities is usually of the order of the Fermi energy and the energy  $J$  is several times smaller. In the case of rare-earth impurities in lanthanum the potential interaction is small but there is a considerable spin-orbit interaction which very strongly suppresses this effect.

Now another hypothesis will be discussed. The previous one referred to the case of small  $\Delta$  which did not change the magnetism I itself. The one to be discussed now - belonging to Anderson and Suhl - refers to the opposite case, strong superconductivity and very weak magnetism, i.e. to the region of small impurity concentrations where  $\theta \ll T_c$ . In this case the superconducting characteristics can be assumed to be the same as in a pure metal. Anderson and Suhl calculated the function  $X(q)$  which, as already mentioned, defines the energy of the ordered state with a spiral SDW having a wave vector  $\vec{q}$ . It happened that for  $q < \Delta_0/v$  the function  $X(q)$  in this case falls and becomes zero for  $q = 0$ . Hence, even in a uniform model,  $X(q)$  is maximal not at  $q = 0$  but at a finite  $q$  which is of the order  $(p_0^2 \Delta_0/v)^{1/3} \sim (50 \text{ \AA})^{-1}$ . Hence in this case one can expect a SDW with such a period. This phenomenon was called by its authors "cryptoferromagnetism". Since the period is much greater than the interatomic distances one can expect that the non-uniformity of impurity concentration gives a smaller effect in smearing out the transition point and the transition becomes sharper than without superconductivity. According to a private communication of Professor Anderson this is confirmed by experiment.

## 6. KONDO EFFECT (K)

We now come to the most difficult cell of the matrix - the Kondo effect. Since the thirties it has been known that the resistance of noble metals as a function of temperature sometimes shows a minimum at low temperatures. Hence there exists a scattering mechanism which becomes more effective at low temperatures. In some cases a maximum was found below the minimum. Further investigation showed that this new resistance mechanism is due to magnetic impurities. The experiments were done with various combinations of host metals and impurities, e.g. Ag Mn, Cu Fe, Au Fe, Mg Mn, Cu Mn, Zn Mn, Au V, Au Mo and others. The minimum was observed for sufficiently low impurity concentrations, as a rule 0.1% and less. In 1956, Alexejevski and Gaidukov measured the galvanomagnetic properties of gold with a very small addition of iron. They have found that in a large interval the resistance is very well described by the formula

$$\rho = \rho_1 + \rho_2 \ln \frac{1}{T} \quad (16)$$

and then saturates at smaller temperatures. This indicates that to call this phenomenon Kondo effect is no more justified than, for example, to call ferromagnetism the effect of Weiss or Heisenberg. But it is so accepted in the literature by now that we will also use this name.

There have been many unsuccessful attempts to explain this phenomenon and only in 1964 did Kondo give the correct explanation. Usually, the part of resistance associated with the interaction  $J\vec{\sigma} \cdot \vec{S}$  was calculated in Born approximation. This was considered to be enough because  $J$  was assumed small (as already explained, this is not necessarily true). A constant resulted which was added to the ordinary residual resistance. Kondo calculated the next approximation for the scattering amplitude and he found that this correction has the relative order  $J/\epsilon_F$ . In  $(\epsilon_F/\epsilon)$  where  $\epsilon$  is the energy of the electron measured from the Fermi level. For  $J < 0$  the additional term produces an increase in the resistance.

The logarithm is very important for the following and therefore a simple derivation will be in order. Let us consider a scattering amplitude of an electron at the impurity due to the  $\vec{\sigma} \cdot \vec{S}$  interaction. In the initial state the electron has the momentum  $\vec{p}$ , spin projection  $\alpha$  and the impurity spin projection is  $M$ . In the final state we have  $\vec{p}'$ ,  $\alpha'$ ,  $M'$ . In the first Born approximation we have

$$A_{\vec{p}\alpha M}^{(1)\vec{p}'\alpha' M'} = -\frac{J}{N} \langle OM' | a_{\vec{p}', \alpha'} \sum_{\vec{p}_1 \vec{p}_1' \alpha_1 \alpha_1'} a_{\vec{p}_1, \alpha_1}^\dagger \vec{\sigma}_{\alpha_1 \alpha_1'} \cdot \vec{S} a_{\vec{p}_1', \alpha_1'} a_{\vec{p}, \alpha}^\dagger | OM \rangle$$

where  $|OM\rangle$  means the equilibrium state of the electrons and the impurity atom in the state  $M$ . This expression is equal to

$$A_{\vec{p}\alpha M}^{(1)\vec{p}', \alpha' M'} = -\frac{J}{N} (1 - n_{\vec{p}'}) (1 - n_{\vec{p}}) (\vec{\sigma} \cdot \vec{S})_{\alpha M}^{\alpha' M'} \quad (17)$$

To the second approximation we have

$$A_{\vec{p}\alpha M}^{(2)\vec{p}', \alpha' M'} = \left(\frac{J}{N}\right)^2 \langle OM' | a_{\vec{p}', \alpha'} \sum_{\vec{p}_1 \alpha_1 \vec{p}_1' \alpha_1'} a_{\vec{p}_1, \alpha_1}^\dagger \vec{\sigma}_{\alpha_1 \alpha_1'} \cdot \vec{S} a_{\vec{p}_1', \alpha_1'} \\ \times (E_{in} - H_0 + i\sigma)^{-1} \sum_{\vec{p}_2 \vec{p}_2' \alpha_2 \alpha_2'} a_{\vec{p}_2, \alpha_2}^\dagger \vec{\sigma}_{\alpha_2 \alpha_2'} \cdot \vec{S} a_{\vec{p}_2', \alpha_2'} a_{\vec{p}, \alpha}^\dagger | OM \rangle$$

A simple calculation gives

$$A_{\vec{p}\alpha M}^{(2)\vec{p}', \alpha' M'} = \left(\frac{J}{N}\right)^2 (1 - n_{\vec{p}}) (1 - n_{\vec{p}'}) \sum_{\vec{p}_1} \frac{1 - n_{\vec{p}_1}}{\epsilon_{\vec{p}} - \epsilon_{\vec{p}_1} + i\delta} (\vec{\sigma} \vec{S} \cdot \vec{\sigma} \vec{S})_{\alpha M}^{\alpha' M'} \\ - \sum_{\vec{p}_1} \frac{n_{\vec{p}_1}}{-\epsilon_{\vec{p}} + \epsilon_{\vec{p}_1} + i\delta} (\sigma_i \sigma_K S_K S_i)_{\alpha M}^{\alpha' M'}$$

Since  $\epsilon_{\vec{p}'} = \epsilon_{\vec{p}}$ , if instead of  $\vec{\sigma} \cdot \vec{S}$  there would have been potential scattering, then the principal value would not contain the distribution function  $n_{\vec{p}}$  at all.

But in the present case, since the operators  $\vec{S}$  do not commute we obtain

$$\begin{aligned} \vec{\sigma} \cdot \vec{S} \cdot \vec{\sigma} \vec{S} &= S(S+1) - \vec{\sigma} \vec{S} \\ \sigma_i \sigma_K S_K S_i &= S(S+1) + \vec{\sigma} \vec{S} \end{aligned} \quad (18)$$

Performing the integration over  $\vec{p}_1$  we keep in mind that all the energies are measured relative to the Fermi level. Assuming the integration limits to be symmetrical around  $\epsilon = 0$  and  $|\epsilon_{\vec{p}}| \ll \epsilon_F$  we get

$$\begin{aligned} A_{\vec{p}\alpha M}^{\vec{p}'\alpha' M'} &= \left(\frac{J}{N}\right)^2 (1 - n_{\vec{p}})(1 - n_{\vec{p}'}) \left[ -S(S+1) \delta_{\alpha\alpha'} \delta_{MM'} i\pi\nu(\epsilon_F) (1 - 2n_{\vec{p}}) \right. \\ &\quad \left. - (\vec{\sigma} \cdot \vec{S})_{\alpha M}^{\alpha' M'} \sum_{\vec{p}_1} \frac{1 - 2n_{\vec{p}_1}}{\epsilon_{\vec{p}} - \epsilon_{\vec{p}_1}} + (\vec{\sigma} \cdot \vec{S})_{\alpha M}^{\alpha' M'} i\pi\nu(\epsilon_F) \right] \end{aligned} \quad (19)$$

The second term in this expression gives the logarithm. The other terms will be discussed a little later. With logarithmic accuracy we obtain therefore

$$A_{\vec{p}\alpha M}^{\vec{p}'\alpha' M'} \approx 2 \left(\frac{J}{N}\right)^2 (1 - n_{\vec{p}})(1 - n_{\vec{p}'}) (\vec{\sigma} \cdot \vec{S})_{\alpha M}^{\alpha' M'} \ln \frac{\epsilon_F}{\max \xi |\epsilon_{\vec{p}}|, T_\xi} \quad (20)$$

So the result of this approximation is represented by the transformation

$$\frac{J}{N} \rightarrow \frac{J}{N} \left( 1 - 2 \frac{J}{N} \nu(\epsilon_F) \ln \frac{\epsilon_F}{\max \xi |\epsilon_{\vec{p}}|, T_\xi} \right) \quad (21)$$

The scattering amplitude begins to depend logarithmically on the energy of the electron. Since the electrons contributing to conductivity have the energy of the order of  $T$ , the resistance to this approximation will have the form

$$\rho = \rho_0 \left( 1 - \frac{4J}{N} \nu(\epsilon_F) \ln \frac{\epsilon_F}{T} \right) \quad (22)$$

where  $\rho_0$  is the Born value. This is Kondo's result. In the temperature region where this correction is small, the formula (22) gives a good description of experimental data.

We shall make now a few comments. In the Born approximation the resistance did not depend on the sign of  $J$  but here it does and it increases if  $J < 0$ , i.e. if anti-ferromagnetic interaction takes place. A simple physical explanation was given by Anderson. The higher Born approximations take into account the correlation of the positions of the electron

and the impurity atom. If  $J < 0$  the electron tends to approach the impurity with the opposite spin orientation, but in this case the possibility exists of a spin-flip scattering. This means that a reaction channel is present which is absent for the parallel spin orientation. Hence, the scattering amplitude for  $J < 0$  will be larger than for  $J > 0$ .

Now from the way in which this result was obtained it follows that two things are essential. First of all the sharpness of the Fermi surface. Therefore, if  $|\epsilon_p| \ll T$  the scattering amplitude becomes independent of energy. The second important thing is the non-commutativity of the spin operators. Of course, this takes place only if the spins are not polarized by an internal exchange field or external magnetic field. If such a polarization exists then the change of the spin orientation is connected with change of energy. In this case, into the denominator will enter a term  $\pm Q$ . Hence, if  $Q \gg T$ ,  $|\epsilon|$  the logarithmic integral will be limited from below by  $Q$ . So in this case we obtain the  $\ln(\epsilon_F/Q)$  in the resistance. In other words, the growth of the logarithmic term in the resistance with decreasing temperature in all cases is possible only until  $Q \ll T$ . As I have already said many times,  $Q$  may be either the exchange field (it is of the order of  $\theta$ , the Curie temperature) for the Zeeman splitting  $g\mu_B H$  where  $H$  is the external field and  $g$  the gyromagnetic ratio. We shall assume at present that there is no external field and that the concentration is so small that  $\theta \sim cJ^2/\epsilon_F$  is much less than the temperature where the logarithmic correcting having the relative order  $(J/\epsilon_F) \ln(\epsilon_F/T)$  becomes of the order of unity.

Before proceeding to this case I want to say a few words regarding the so-called giant thermopowers. I shall remind you what is meant by that. If we apply to a metal an electric field and a temperature gradient there will be an electric current

$$\vec{j} = \sigma \vec{E} + \beta \nabla T$$

If the circuit is disconnected  $J = 0$  and  $E = (-\beta/\sigma) \nabla T$ . Now since  $E = \nabla \varphi$  where  $\varphi$  is the potential, we can write this as  $d\varphi/dT = -\beta/\sigma$ . That means that  $-\beta/\sigma$  is the potential difference per degree. From the kinetic equation with a given collision time we get

$$\frac{d\varphi}{dT} = -\frac{\beta}{\sigma} = -\frac{\int \tau(\epsilon) v^2 \epsilon \frac{\partial n}{\partial \epsilon} v(\epsilon) d\epsilon}{eT \int \tau(\epsilon) v^2 \frac{\partial n}{\partial \epsilon} v(\epsilon) d\epsilon} \quad (23)$$

where  $\tau(\epsilon) = 1/w(\epsilon)$  is the collision time or the reciprocal scattering probability,  $v$  the velocity,  $v$  the state density,  $n$  the Fermi distribution function (the energy is measured from  $\epsilon_F$ ). The integrals with the function  $\partial n/\partial \epsilon$  are taken according to the rule

$$\int F(\epsilon) \frac{\partial n}{\partial \epsilon} = -F(0) - \frac{\pi^2 T^2}{6} \left( \frac{\partial^2 F}{\partial \epsilon^2} \right)_0 - \dots$$

if the function  $F$  varies slowly near  $\epsilon = 0$  where  $\partial n / \partial \epsilon$  has the form of a  $\delta$ -function. Using this rule we obtain

$$\frac{d\phi}{dT} = \frac{\pi^2 T}{3 e} \left[ \frac{\partial}{\partial \epsilon} \ln (v^2 \tau \nu) \right]_0 \quad (24)$$

This quantity has the order of magnitude  $T/e\epsilon_F$ . At temperatures  $\sim 1^\circ K$  this gives  $10^{-8}$  V/deg. This number is in fact observed for pure metals. However, if  $\tau$  contains a term proportional to  $2n(\epsilon) - 1$  which varies rapidly in the region  $\epsilon \approx 0$ , the general rule does not apply. Moreover this term is odd in  $\epsilon$  and since  $\partial n / \partial \epsilon$  is an even function, the numerator in the general formula becomes much larger than in the case of an even  $\tau$ . The derivative  $(\partial / \partial \epsilon)(2n - 1)$  has the order of magnitude  $1/T$  and if we put that into (24) the result becomes independent of  $T$ . The terms containing  $(2n-1)$  in the scattering amplitude arise from the residues at the poles in the second approximation. We have seen an example when we derived the logarithmic correction. But it is more difficult to get such a term in the scattering probability. Kondo has shown that such terms arise from interference between the exchange and potential scattering. If we assume both amplitudes to be isotropic we get

$$\frac{d\phi}{dT} = \frac{4\pi^2}{e} \nu(\epsilon_F) J V \left( \frac{\rho_{ex}}{\rho} \right) \quad (25)$$

where  $\rho_{ex}/\rho = J^2 S(S+1)/[V^2 + J^2 S(S+1)]$ . Here it is supposed  $J \ll V$ . The order of magnitude becomes  $J^3/eV\epsilon_F^2$ . If we take  $V \sim \epsilon_F$ ,  $J \sim 0.2\epsilon_F$ , we get  $10^{-6}$  V/deg. This is the typical number obtained by experiment. Contrary to the pure metal,  $d\phi/dT$  does not depend on temperature. This is also found in experiment apart from very low temperatures where the thermopower decreases in magnitude. This can be due to ordering or to the formation of a quasi-bound state, which will be discussed later. The sign of  $d\phi/dT$  is defined by the relation of the signs of  $J$  and  $V$ . For the majority of alloys exhibiting Kondo effect it is negative. On the high temperature side the magnitude of the thermopower decreases when ordinary scattering begins to increase because of the phonons.

It is also worth mentioning that before this explanation due to Kondo there have been other explanations based on interference scattering by ordering. These explanations are most probably false since the giant thermopower exists above the ordering temperature.

Let us now return to the logarithmic correction to the scattering amplitude. If the concentration is sufficiently low then by decreasing the temperature we can reach the situation where the correction becomes of the order of unity. What will happen in this case? Suhl and the present author, using different methods, summed up the main terms. To the logarithmic accuracy the result (for  $T = 0$ ) was

$$\frac{J}{N} \rightarrow \frac{J}{N} \left( 1 + 2 \frac{J}{N} \nu(\epsilon) \ln \frac{\epsilon_F}{|\epsilon|} \right)^{-1} \quad (26)$$

If  $J > 0$  this expression can be used everywhere, but if  $J < 0$  the amplitude becomes infinite at some energy. The same is true for resistance. At a certain temperature

$$T_K = a \epsilon_F e^{-\frac{N}{2|J|\nu(\epsilon_F)}} \quad (27)$$

where  $a$  is some constant of the order of unity, the resistance becomes infinite. This temperature was called the Kondo temperature. It is clear that at temperatures of the order of or lower than  $T_K$  the logarithmic accuracy is not enough. Two ideas were proposed for solving this problem.

One method is due to Suhl and Wong. Maleiev and Ginzburg (Leningrad) are working in the same direction. The two latter research workers probably gave the clearest formulation of this method. It is assumed that the scattering amplitude as a function of the complex energy variable is analytic in the plane with cuts along the positive real semi-axis from 0 to  $\infty$ . Then the unitarity condition for the scattering matrix is used:

$$i(T^\dagger - T) = T^\dagger T \quad (28)$$

On the right-hand side, there is a sum over intermediate states having, apart from an electron, an arbitrary number of electron-hole pairs. It is assumed that these many-particle states do not give any essential contribution. On this assumption the unitarity condition becomes an equation for the  $T$ -matrix. A solution is found which has the analyticity properties stated above and whose limiting value at large energies is just the perturbational expression (26).

The disadvantages of this approach are the following:

- (a) There is some doubt on the limitation of the intermediate states. In the framework of perturbation theory it is possible to obtain equations which have the necessary accuracy. However, I did not succeed in showing that they coincide with Suhl's or Maleiev's equations more than in the logarithmic approximation.
- (b) The assumption of analytical properties of the scattering amplitudes is, in fact, not unique.

The results of this approach which I shall call the "unitarity approach" are the following. For  $J < 0$  at  $T = 0$  and  $\epsilon \rightarrow 0$  the scattering cross-section reaches the unitarity limit

$$\sigma_{\text{eff}} = \frac{4\pi}{K^2} (2\ell + 1) \quad (29)$$

Here the factor  $(2\ell + 1)$  takes into account the fact that the "Kondo effect" can occur in every partial amplitude (with a certain  $\ell$ ) and, of course, the most important amplitude is that with which this happens first.

With decreasing temperature the curve  $\rho(T)$  tends gradually to this limit according to the law

$$\rho_0 - \rho(T) \sim \begin{cases} \frac{1}{\ln \frac{T_K}{T}} & \ell \neq 0 \\ \frac{1}{\ln^2 \frac{T_K}{T}} & \ell = 0 \end{cases} \quad (30)$$

There exists an additional specific heat proportional to the impurity concentration. It has a maximum at some temperature of the order of the Kondo temperature and for  $T \rightarrow 0$  goes to zero as  $1/\ln^4(T_K/T)$ . The heat resistivity can be obtained from the electric resistivity by means of the Wiedemann-Franz law. The thermoelectric power was also found, the rather complicated result will not be quoted. The magnetic properties were not considered. All these predictions are in qualitative agreement with experiment but quantitatively the logarithmic law for  $\rho_0 - \rho(T)$  seems to contradict the experimental data.

The other idea was first stated by Nagaoka, although now this approach has been developed in greater detail by Yosida and the author of this paper (closely related are the works of Kondo and Anderson). It starts from the fact noticed that the behaviour of the scattering amplitude for  $J < 0$  resembles very closely the behaviour of the scattering amplitude of two electrons with attractive interaction. It is well known that in the latter case this indicates that the proposed ground state is unstable and it corresponds to the superconducting interaction. It is natural to suppose that also in the case considered some "quasi-bound state" of the electron with the localized spin of the type of the Cooper pairs in a superconductor is formed. In fact, this is a collective phenomenon which leads to some new correlation between the electrons and the localized spin. But, in the same way as for the superconductor, this gives in many respects the impression that there were true bound states.

We shall not discuss everything in this field. In most cases the works based on perturbation theory do not take into account all the necessary graphs. This refers particularly to the articles of Nagaoka himself for spin  $\frac{1}{2}$ . Therefore, there seems to be no sense in solving Nagaoka's equations more precisely than Nagaoka himself did in his first articles (Hamann for example). On the other hand, there exist calculations based on variational approaches. But, as in all such cases, the accuracy of such an approach is difficult to establish. What are the results of the "bound state" approach? Here it is difficult for the author to resist the temptation to quote his own results. The technique of this calculation is based on the representation of the spin operators by means of some fictitious "pseudo-fermions"

$$\hat{S}^i = \sum_{\beta\beta'} a_\beta^\dagger S_{\beta\beta'}^i a_{\beta'} \quad (31)$$

The  $S_{\beta\beta'}^i$  are the matrices of the spin in question, e.g. the Pauli matrices for  $S = \frac{1}{2}$ , and on the right-hand side we have, in fact, bilinear expression

(e.g. for  $S = \frac{1}{2}$ ,  $S^z = \frac{1}{2}(a_{\frac{1}{2}}^\dagger a_{\frac{1}{2}} - a_{-\frac{1}{2}}^\dagger a_{-\frac{1}{2}})$ ). Although the introduction of such quasi-fermions means an introduction of some unphysical states (the physical are those where one of the particle numbers  $n_\beta$  is unity and all the others are zero) they can be excluded when necessary. The introduction of quasi-fermions allows the use of ordinary field theory methods. In particular, the scattering amplitude above the Kondo temperature is obtained by summing up some particular two-dimensional sequence of Feynman graphs which we call "parquet". The examples are listed in (a), (b), (c), (d), whereas graph (e) is the simplest graph not belonging to the "parquet" (Fig. 7). Here the dashed lines are the propagators of the quasi-fermions and the solid lines those of the electrons. The "parquet" graphs with a slight modification of the vertices and the quasi-fermion propagators, in fact, give the scattering amplitude with the required accuracy and not only with the logarithmic one. Since we doubt that the corresponding equation coincides with Suhl's, the results obtained with the latter are liable to be doubted.

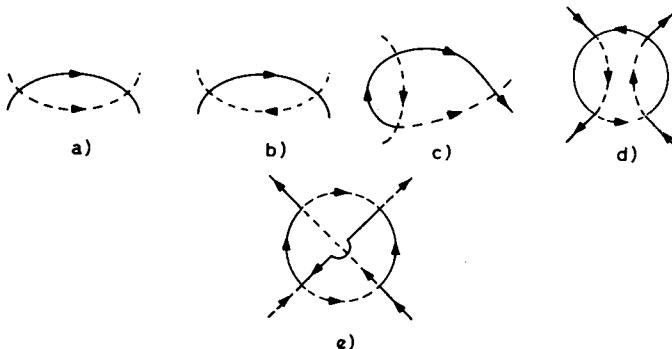


FIG. 7. Feynman graphs belonging to a "parquet" (a - d); (e) simplest non-parquet Feynman graph.

The formation of the quasi-bound states is taken into account by introducing "averages" of the type  $\langle \psi_\alpha a_\beta \rangle$ . There are four of them since instead of  $a$  and  $\psi$  there can be  $a^\dagger$  and  $\psi^\dagger$ . Although at first glance such quantities are unsuitable since they do not conserve the number of electrons, they are, in fact, matrix elements for the transition of an electron into this bound state. This resembles the case of superconductors where such quantities mean a transition of free electrons into a bound Cooper pair. Here, however, the situation is somewhat different since the "parquet" is a much more complicated sequence than the superconductivity "ladder". At  $T = 0$  it is possible to show that the bound state can be formed only for  $J < 0$  with the net spin  $S = \frac{1}{2}$ . The binding energy is equal to

$$\Delta E = -K \epsilon_F \left( \frac{N}{\nu(\epsilon_F) |J|} \right)^{S-1} \exp \left( -\frac{N}{2\nu(\epsilon_F) |J|} \right) \quad (32)$$

where the constant  $K$  is of the order of unity. Unfortunately, the problem here can only be solved with logarithmic accuracy and there is no certainty that  $K$  is finite. However, for  $S = 1$  there is an essential simplification. In

this case the parquet is reduced to a ladder. Therefore, the problem becomes of the same complexity as superconductivity.

The results for this case are the following. There exists a transition temperature  $T_K$  which is connected to the binding energy at  $T = 0$  with the relation

$$|\Delta E| = \frac{8}{3\gamma} T_K \quad (33)$$

where  $\gamma = 1.78$ . The bound state is destroyed by the magnetic field  $H_K$  which has the following asymptotic values:

$$g\mu_B H_K \approx \Lambda \left[ 1 - (2\nu^2(\epsilon_F) \left( \frac{J}{N} \right)^2 \ln \frac{T_K}{T})^{-1} \right]; \quad T \rightarrow 0 \quad (34)$$

$$g\mu_B H_K \approx \pi \sqrt{\frac{6}{7} \xi(3) T_K (T_K - T)} \quad T \rightarrow T_K$$

where  $\Lambda$  is of the order of Fermi energy ( $T_K$  depends on  $\Lambda$ :  $T_K = (2\Lambda\gamma/\pi) \exp(-N/2\nu(\epsilon_F)|J|)$ ). Hence it is practically impossible at sufficiently low temperatures to destroy this state by the magnetic field.

The additional specific heat has the form

$$\begin{aligned} \frac{\Delta C}{cN} &= 3.16 \frac{T}{T_K}; \quad T \ll T_K \\ \frac{\Delta C}{cN} &= 1.08 + 0.34 \frac{T - T_K}{T_K}; \quad T_K - T \ll T_K \end{aligned} \quad (35)$$

The linear specific heat at low temperatures indicates that there is no true bound state. It is more natural to speak of the existence of a maximum in the state density just at the Fermi level. The width of this maximum is of the order of  $T_K$ , and its magnitude is gradually reduced with increasing temperature. The approximation used here is probably incorrect close to  $T_K$  and so the discontinuity of  $\Delta C$  is not real.

The magnetic moment is

$$\begin{aligned} \frac{\Delta M}{cN} &= \frac{2}{3} \frac{(g\mu_B)^2 H}{T}; \quad T \geq T_K, \quad g\mu_B H \ll T \\ \frac{\Delta M}{cN} &= \frac{2}{27} \frac{(g\mu_B)^2 H}{T}; \quad T \ll T_K, \quad g\mu_B H \ll T \\ \frac{\Delta M}{cN} &= \frac{3}{7} g\mu_B + \frac{16\gamma}{7\pi^2} \frac{(g\mu_B)^2 H}{T_K} \left( \ln \frac{3\pi T_K}{2\gamma g\mu_B H} + \frac{1}{2} \right); \quad T \ll g\mu_B H \ll T_K \end{aligned} \quad (36)$$

So at small fields ( $g\mu_B H \ll T$ ) it looks as though the effective magnetic moment becomes three times smaller at  $T \ll T_K$  but does not turn to zero.

For  $S = \frac{1}{2}$ , because of the parquet situation, it is impossible to obtain all the quantities but it is possible to express the magnetic moment in terms of the order parameter  $R$  which must be of the order of  $T_K$ .

$$\frac{\Delta M}{cN} = \frac{(g\mu_B)^2 H}{R(0)}; \quad g\mu_B H, \quad T \ll T_K \quad (37)$$

(for an arbitrary relation between  $H$  and  $T$ ).

The same result was obtained by Ishii and Yosida. According to this, for  $S = \frac{1}{2}$  there exists only polarizability of the complex but no intrinsic magnetic moment. But I have already said that the assumption that  $R(0)$  is finite only does not contradict the logarithmic approximation, and one must keep this in mind.

Regarding the electric resistance for  $S = 1$ , it is possible to say that it approaches the unitarity limit according to the law

$$\rho = \rho(0) \left[ 1 - \alpha \left( \frac{T_K}{T} \right)^2 \right]; \quad \alpha \sim 1, \quad T \ll T_K \quad (38)$$

The constant  $\alpha$  is difficult to calculate since it demands all the parquet. We must mention that the results for the resistance and specific heat qualitatively correspond to the first articles of Nagaoka. The other point is that, although all the results quoted above qualitatively correspond to experiment, we are not extremely confident in all steps of the derivation. We shall not discuss it because it demands some technical details.

Let us say a few more words about the other results. From a variational procedure Anderson found the law  $\chi \sim T^{-\frac{1}{2}}$  and  $\Delta C \sim T^{\frac{1}{2}}$  for  $S = \frac{1}{2}$  and Hamann and Bloomfield obtained  $\Delta C \sim T^{0.57}$  from the exact solution of the Nagaoka equation for  $S = \frac{1}{2}$ . Using the latter method, Murata and Wilkins obtained the law  $1 / \ln^2(T_K/T)$  for the approach of resistivity to the unitarity limit at  $T \rightarrow 0$ . These are the most important results which we have mentioned before presenting the objections to these methods.

Furthermore, there exists an unproved opinion that for every value of impurity spin the ground quasi-bound state must be spinless, i.e. the electron cloud necessarily compensates the impurity spin. We doubt it since we could not find logarithmic singularities in graphs with two electron lines for  $S = 1$ , but Yosida states that a binding with spin  $S = 1$  of an electron and a hole which compensate entirely the impurity spin gives a binding energy twice that of a single electron.

What is the experimental situation? The measurements of the resistivity by Daybell and Steyert in Cu Fe and Cu Cr indicate that for  $T \rightarrow 0$   $\rho(T)$  saturates and reaches a finite value according to the  $T^2$  law. The limiting cross-section value for Cr is 2.4 times larger than in Fe and if we believe it to be the unitarity limit, which comes out from all the theories, then certainly the Kondo effect takes place not necessarily only for the s-wave scattering.

As to the susceptibility measured in the same experiments, one can see from it that the magnetic moment is reduced with decreasing temperatures but it is difficult to say whether it tends to zero. Anderson says that the results are in good agreement with the law  $\Delta X \sim 1/\sqrt{T}$  but we think that they also do not disagree with the author's results, which we described although the latter are true only for  $S = 1$  and, therefore, not comparable with experiments in the general case.

From the formula for the Kondo temperature it follows that for  $|J| < \epsilon_F$  we have  $T_K \ll \epsilon_F$ , but as already mentioned,  $J$  is not necessarily small; this is confirmed by the experiments of Kume with AuV, were  $T_K$  happens to be in the vicinity of 300°K. By the way, in these experiments, due to the possibility of measuring in a large temperature interval, the  $T^2$  law for the resistivity has been well confirmed. The magnetic susceptibility is closer to the law  $T^{-0.2}$ , but since all the theoretic calculations are performed, indeed, only for the case  $|J| \ll \epsilon_F$ , it is possible that the results of Kume should not be compared with the existing theories. However, the  $T^2$  law for the resistance, if it exists, must be of a very general nature since it is the consequence of the Fermi statistics. Therefore, we think that the AuV results are additional evidence against the "unitarity" and the exact Nagaoka solution where a logarithmic law is predicted.

One more interesting experimental result was obtained in the work by Frankel and others. The hyperfine field at the iron nucleus was measured in a dilute solution Cu Fe. It appeared that although the field is saturated with decreasing temperature according to the law  $H_{hf} = H_{sat} Bg(g\mu_B H/T)$  ( $Bg$  being the Brillouin function) the field  $H_{sat}$  itself depends on the applied field up to very large values of the latter. The authors conclude that the bound state is not destroyed at low temperatures, at least up to the fields  $g\mu_B H > 4-5 T_K$ . This corresponds to the result regarding critical field at low temperatures, which has already been discussed.

We have discussed on purpose all the main concepts since none of them can be assumed to be absolutely true. We should also like to mention Schrieffer's idea that all the transitional metal atoms in non-magnetic metals have a localized spin but some of them have the Kondo temperature higher than the melting point and at a lower temperature the electronic cloud completely screens the impurity spin. This could remind us of the expression "Everybody is baldheaded but some of the bald heads are covered with hair".

Finally, we shall discuss another result having relation to the Kondo effect. It concerns the form of the tunnelling characteristics at small potential differences. If two metals are separated by a dielectric layer of several interatomic dimensions thick tunnelling of electrons becomes possible. However, the  $I(V)$  curve sometimes shows the increase of effective resistance in the vicinity of  $V = 0$ . Appelbaum and Anderson, Zavadovski and Soljom explained it by the presence of magnetic impurities in the insulating layer. In this case the tunnelling Hamiltonian can be shown to have a  $\sigma S$  term and the interference of this term with the exchange scattering without tunnelling produces the same Kondo logarithm. At low temperatures the logarithmic integral is limited from below by eV. This gives the dependence of the effective resistance on the potential difference proportional to  $\ln \epsilon_F/eV$  until  $eV \sim T$ . These considerations are in good agreement with the results of Wyatt, Rowell and others. Mezej, in particular, introduced controlled quantities of chromium in an Al-Insulator Al

tunnelling contact and has shown definitely that the maximum is due to the magnetic impurities.

## 7. KONDO EFFECT AND ORDERING (KO): NEGATIVE MAGNETO-RESISTANCE

As mentioned before, the external or internal fields ordering the spins suppress the Kondo effect. This means that below  $T < Q$ , where  $Q$  is the exchange effective field of the order of  $\theta$ ;  $g\mu_B^H S$  is the log and becomes  $\ln \epsilon_F/\theta$  or  $\ln \epsilon_F/\mu_B g H S$  and saturates.

However, the polarization of spins has also another consequence. The scattering amplitude  $\vec{\sigma} \cdot \vec{S}$  gives a scattering cross-section proportional to  $S(S+1)$ , but in the case of polarized spins from  $\vec{\sigma} \cdot \vec{S}$  only  $\sigma_z S$  remains. Therefore, the scattering probability becomes proportional to  $S^2$  instead of  $S(S+1)$ . This effect can also be described as suppression of spin-flip scattering. So, apart from the saturation of the log there is also the reduction of the factor coming from  $\vec{\sigma} \cdot \vec{S}$  in the resistance associated with magnetic impurities.

If the ordering is ferromagnetic or caused by the external field then there exists also another phenomenon - the interference between the potential scattering and the  $\vec{\sigma} \cdot \vec{S}$  interaction. Indeed, for an electron with its spin along the field we get the scattering amplitude  $V - JS$  and for an electron with the opposite spin  $V + JS$  (of course for  $T \ll Q$ ). The corresponding scattering probabilities are proportional to  $V^2/2JSV + (JS)^2$ . This probability has to be averaged over all the impurities. If the ordering has an anti-ferromagnetic or random character then the interference term will vanish, but if there is ferromagnetic ordering or a polarization by an external magnetic field it will remain. For the corresponding collision times we get

$$\tau_{\pm} \sim \frac{1}{V^2} \pm \frac{2JS}{V^3} + \frac{4(JS)^2}{V^4} - \frac{(JS)^2}{V^4}$$

(we have assumed  $J \ll V$ ). The third term here has interference origin. Since the conductivities due to + and - electrons are added, the full conductivity will be proportional to

$$\frac{\tau_+ + \tau_-}{2} \sim \frac{1}{V^2} + \frac{3(JS)^2}{V^4}$$

and hence the resistance to

$$\rho \sim V^2 - 3(JS)^2$$

In the absence of interference we would get

$$\rho \sim V^2 + (JS)^2$$

Here we did not account for the Kondo effect. If  $Q > T_K$ , we obtain for  $T \ll Q$

$$\rho \sim V^2 + (JS)^2 \left[ 1 + \frac{2J}{N} \nu(\epsilon_F) \ln \frac{\epsilon_F}{Q} \right]^{-2} - 4(JS)^2 \quad (39)$$

So in the second term at  $T \sim Q$  the log saturates and  $S(S+1)$  is changed to  $S^2$ . Apart from that there appears the third negative term. All this leads to the fact that in the temperature dependence of the resistance the ordering gives not only saturation at decreasing temperature but also a maximum at the temperature  $T_{\max} \sim Q$ . We would like to stress here that the maximum has two origins and therefore also appears if the ordering is anti-ferromagnetic. Of course, the ordering temperature must be lower than the  $T_{\min}$  at which the resistance minimum takes place. If  $Q$  is due to ordering then  $T_{\max} \sim \theta$  and is therefore proportional to the impurity concentration. If, however, it is due to the external field, then  $T_{\max} \sim H$ .

Apart from that, it should be mentioned that the external field polarizing the spins must certainly lead to the decrease in the part of the resistance due to the  $\delta$ - $S$  interaction. This decrease may be larger than the increase of the resistance due to the curving of the electron trajectories and the net effect may be negative magneto-resistance.

All these phenomena, the maxima, their position and the negative magneto-resistance are observed in experiment and are in good agreement with the above considerations. We shall not present the rather complicated formulas, which also cannot be considered very rigorous if the spin polarization only is not due to the external field since, as has been said already, the type of ordering is not very clear. I would just like to mention that if  $\mu_B g H \ll T$ , the correction to the resistance will obviously be proportional to  $-(\mu_B g H/T)^2$  or  $\rho(H) - \rho(0) \sim -(\Delta M)^2$  where  $\Delta M$  is the impurity part of the magnetic moment.

Until now we considered the case  $T_K < \theta$ . Now, what happens in the opposite case? Of course, in the case when the Kondo effect leads to the formation of bound states with zero net spin, there is no problem. However, if the spin is not screened entirely then this question exists, but, of course, there is no sense in solving it before the screening of the spin has been clarified.

## 8. KONDO EFFECT AND SUPERCONDUCTIVITY (SK)

In this direction there are several works. Almost all of them consider the small-concentration limit. Probably the most certain result was obtained by Fowler and Maki by means of some modification of Suhl's theory. The same result was obtained by myself for  $S = 1$  by means of the method described earlier. According to these results the quasi-bound state cannot appear if  $T$  is smaller than the superconducting transition temperature for a pure metal. If, however,  $T_K$  is larger than this temperature  $T_c$  decreases and by order of magnitude

$$\frac{\delta T_c}{T_{c0}} \sim - \frac{c\epsilon_F}{T_K} \quad (40)$$

For comparison we mention the fact that if  $T_{c0} \gg T_K$  then the change of the superconducting transition temperature at small concentration has the order of magnitude

$$\frac{\delta T_c}{T_{c0}} \sim -\frac{cJ^2}{\epsilon_F T_{c0}} \quad (41)$$

Therefore, according to these results the  $T_c$  must be very much reduced by the Kondo effect. However, there exists an opposite statement by Ginsburg (Leningrad) coming from the unitarity approach. According to him the transition temperature rises with impurity concentration if  $T_c \ll T_u$ . The author does not understand this work well enough to draw any conclusions. For physical reasons this result seems to be very doubtful.

To finish with this problem, we can say that if  $T_{c0} \gg T_K$  and there are no bound states formed, nevertheless the Kondo corrections seem to have an effect on the superconducting properties. If these corrections are introduced in the spin-flip scattering time  $\tau_s$ , which defines the properties of superconductors containing magnetic impurities; the agreement between the theory and experiment improves.

As mentioned at the beginning of the paper, it was not possible to consider all the problems; e.g. the question of magnetic impurities in the so-called almost ferromagnetic metals, such as palladium, was completely left out. This problem comprises many interesting phenomena as, e.g., giant magnetic moments, the ferromagnetic transition under the influence of a small ferromagnetic impurity, and many other phenomena whose treatment would cover much more space in this book. Finally we wish to express the hope that we could give the impression that small magnetic impurities are not a small problem.

# SOLID-STATE PROBLEMS FOR PARTICLE PHYSICISTS

P.C. MARTIN

Lyman Laboratory of Physics,  
Harvard University,  
Cambridge, Mass., United States of America

## Abstract

SOLID-STATE PROBLEMS FOR PARTICLE PHYSICISTS. Similar techniques have played an important role in recent progress in particle physics, solid-state physics, and statistical mechanics. On the basis of this similarity it is possible to codify current solid-state and statistical problems for the particle physicist. Within this framework, the utility and applicability for solid-state problems of different approaches to particle physics are discussed and contrasted.

Since the last theoretical physics conference in Japan, I think it is fair to say that some of the most noticeable changes in solid-state and statistical theory have come about through and been clarified by field theoretic techniques. There are still a few reactionaries who resist the use of these tools but their number is dwindling as their mean age increases. Strikingly, it is the same kinds of techniques which worked well in electrodynamics, but were to some extent abandoned in strong interaction physics, which have proven successful here. I suppose part of the reason for this difference is that many-particle theorists are better versed in the more venerable techniques. However, I suspect that because we are concerned with somewhat different physical questions, we may not be so far off base in using these old-fashioned tools.

In this general introduction to the subject, I shall try to summarize the kinds of questions with which we are concerned, the sorts of techniques which have been brought to bear, and the difficulties we face. Clearly, these difficulties are in large measure mathematical. At one level, we believe we know the Hamiltonians of the systems which concern us or could find them out, and it is only a problem of finding the consequences of our equations. At a practical level, however, we must concern ourselves with isolating and calculating tractable and measurable aspects of the behaviour without solving insoluble mathematical problems. For these purposes, the field theoretic language has turned out to be particularly appropriate. Indeed, a tenable way to organize this summary of the problems for discussion is in terms of the degree of field correlations necessary to explain them. Such an organization can obviously be condemned as not doing justice to the diverse physics of these systems; nonetheless, the common threads have proven useful and do tie the subject together more closely than one might suspect. They may also serve as a guide for the elementary particle and field theorists in making appropriate translations.

Let me begin by observing that in many-particle systems we are concerned with states which have considerably lower symmetry than the favourite state of particle physicists - the Lorentz invariant vacuum. In

even the simplest systems we consider - normal fluids - there is typically a mass, and the effect of a transformation on this massive state is to produce a different one, with a different momentum. Our vacuum has translational but not Galilean invariance.

In many of the systems we consider, the state also has sufficient symmetry so that the dynamical field which describes it, e.g. the particle field  $\psi$ , the electromagnetic field  $B$ , or in a system of spins, the magnetization  $M$ , has vanishing expectation value. However, this is not always the case. In many-particle physics, states with a high degree of symmetry, but in which the quantum field has finite expectation value, play a crucial and central role. (The possibility that the expectation value of the field be non-vanishing in the vacuum state has, of course, been discussed in elementary particle physics, but, as I understand it, the tadpole diagrams are mainly a curiosity there.) The most familiar example is a ferromagnetic system in equilibrium. In contrast with a paramagnet, the appropriate translationally invariant state for a ferromagnet is one in which the magnetization is non-vanishing even when there is no external field. As Fisher reports in these Proceedings, it appears that most, if not all, phase transitions are conveniently characterized and studied below the transition in terms of this kind of state where the field associated with the ordered phase has a non-vanishing expectation value. In the equilibrium states of these ordered systems, the expectation value of the field, called the order parameter, is time-independent; in non-equilibrium states it varies with time. An example of a time-dependent state in which a quantum field,  $A$ , has a time-varying expectation value is a laser which is continuously pumped. Anderson and Schrieffer deal in these Proceedings with some of the more esoteric order parameters whose non-vanishing values are the fundamental basis for superfluidity and superconductivity, and with the techniques for giving them, like the electromagnetic field in a laser, a time-varying value (e.g. the A.C. Josephson effect).

Just as most of elementary particle physics is concerned with states in which the expectation value of the field vanishes, much of many-particle physics is concerned with normal systems - systems in which the field has vanishing expectation value but non-vanishing fluctuations.

In translationally invariant equilibrium systems, the spectrum of these fluctuations serves to define the elementary excitations. In the Lorentz invariant vacuum, of course, this spectrum gives the masses associated with the field. In normal many-particle systems, however, where the momentum relative to that of the medium plays a role, the spectrum depends independently on wavelength and frequency. Formally, we may say that the spectral function defined by

$$\langle \psi^\dagger(r,t) \psi(r',t') \rangle \equiv \int \frac{d\omega}{2\pi} \int \frac{dk}{(2\pi)^3} B(k\omega) e^{ik \cdot (r-r') - i\omega(t-t')}$$

takes the form

$$B(k\omega) = B(c^2 k^2 - \omega^2)$$

in vacuum because of Lorentz invariance but is a function of energy and momentum in many-particle systems. The function  $B(k\omega)$  can be thought of as the relative probability for creating particles of momentum  $\hbar k$  and

energy  $\hbar\omega$  by an oscillating external source of particles with unit amplitude, wave number  $k$ , and frequency  $\omega$ .

In elementary particle physics, if the field  $\psi$  describes a single stable neutral particle, for each  $k$ , a plot of  $B(k\omega)$  will look as shown in Fig. 1(a). When the particle is charged and electromagnetic effects are taken into account, the continuum is not separated from the elementary excitation and  $B$  is described as in Fig. 1(b). Likewise, when the particle is long-lived but unstable, the spectral function contains a Lorentzian whose width is related to the lifetime in a background (Fig. 1(c)).

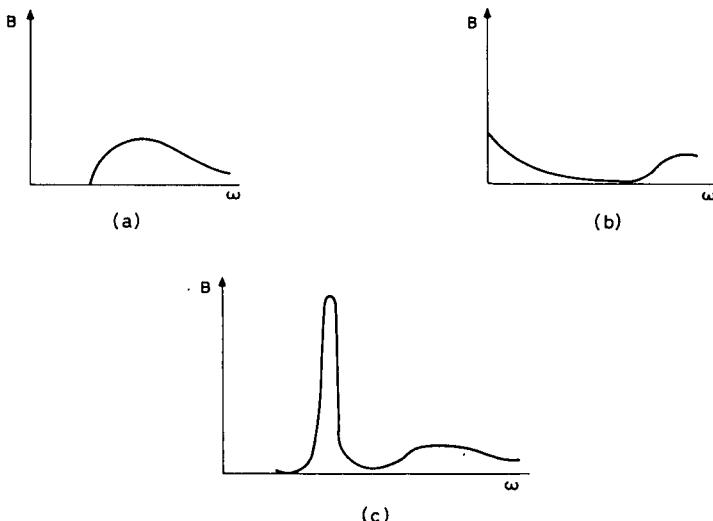


FIG. 1.  $B(k,\omega)$  where (a) the field  $\psi$  describes a single stable neutral particle, (b) the particle is charged, and (c) the particle is long-lived but unstable.

In many-particle physics we always are involved with the last situation, the one of least concern in particle physics. Indeed, in many cases there may not even be Lorentzian peak in  $B$ , that is, the excitations do not necessarily live a long time compared with their interactions. For this reason, S-matrix techniques do not seem so well suited to the problems we wish to consider. Particles do not have a very well-defined energy shell to which they must return outside of their interaction region, and we must deal with the entire energy-momentum spectrum.

There are, however, certain instances in which certain excitations are relatively long-lived and in which these excitations together with their interactions – and possibly their interaction with the coherent background of the ordered state (see Khalatnikov, these Proceedings) – determine a large number of relevant properties. This appears to be the case at vanishing temperature for certain values of  $k \sim k_F \sim (M)^{1/3} [6\pi^2/(2S+1)]^{1/3}$  in Fermi systems and when  $k \sim 0$  in Bose systems. The simplified treatments which can be made in terms of these "quasi-particles" with their renormalized energies and observed renormalized interactions constitute the Landau theories of Fermi and Bose fluids which Pines dis-

cusses in these Proceedings. The application of these theories to  ${}^3\text{He}$  and  ${}^4\text{He}$ , respectively, constitutes one major achievement of many-particle physics. Quantitative agreement among various experiments has been achieved (see Pethick, these Proceedings), although none of the parameters are calculated from the two-body potential (i. e. the analogy of the bare coupling constant).

Because the medium may be looked upon as a heavy particle which defines a frame, the analysis of the propagation of a single particle, or field, in a medium has a degree of structure similar to that of an elastic scattering amplitude in elementary particle physics (i. e. it depends on two scalar variables). In fact, it is possible to show that almost all properties that we can think of measuring – be they thermodynamic, hydrodynamic, or the more detailed information obtained by neutron or light scattering – are measurements which can be described in terms of the correlation of two fields or two currents. For example, it is not difficult to prove that essentially all the interesting physical information about a one-component classical system can be determined from the particle current correlation function

$$\langle j_k(r,t)j_\ell(r',t') \rangle = \int \frac{dk}{(2\pi)^3} \int \frac{d\omega}{2\pi} S_{k\ell}(k\omega) e^{ik \cdot (r-r') - i\omega(t-t')}$$

In particular, the thermodynamic parameters, the transport coefficients (viscosity, thermal conductivity, etc.) and the properties of neutron scattering are all contained. Few experiments require us to, or enable us to, measure more complicated correlations.

Thus, the scattering experiments, which are so central to elementary particle physics are not something we normally concern ourselves with, nor is the elastic scattering amplitude a particularly meaningful or useful concept for the rather broad resonances which constitute our particles. While the higher correlations are not very accessible experimentally, they clearly are important in setting up a calculational scheme. The calculation of the correlation function for a pair of fields or currents involves us in an analysis of correlations having a larger number of fields or currents. In the approximations which perturbative field theoretic techniques generate for these functions, certain important properties are maintained to a given order of the perturbation parameter (current conservation (Ward identities), antisymmetry or symmetry with respect to interchange of identical fermions, bosons, and so forth). How to maintain all of these identities in strong coupling approximation schemes is unclear; indeed, it seems as if the requirement that all these identities hold exactly may generate, as unitarity and causality appear to generate in elementary particle physics, the complete theory. Nonetheless, certain techniques have been developed. These techniques bear a striking resemblance to the techniques which have been employed recently to deduce sum rules and low energy theorems in elementary particle physics. For example, the oldest sum rule in many-particle physics, is the Placzek sum rule for neutron scattering. It is merely a statement of the fact that non-relativistically

$$[j_k(r,t), n(r',t')] = \nabla n(r,t) \delta(r-r')$$

The consequences of the algebra of the field for a system of spins

$$S_i(r,t) \equiv \sum_{\alpha} S_i(t) \delta(r - r^{\alpha}(t))$$

$$[S_i(rt), S_j(r't)] = \epsilon_{ijk} S_k(rt) \delta(r - r')$$

and of the algebra of the infinite component field

$$f(rpt) \equiv \int d\bar{r} \psi^{\dagger} \left( r + \frac{\bar{r}}{2} t \right) \psi \left( r - \frac{\bar{r}}{2} t \right) e^{\frac{i p \cdot \bar{r}}{\hbar}} = \frac{1}{(2\pi\hbar)^3} \sum_{\alpha} \delta(p - p^{\alpha}(t)) \delta(r - r^{\alpha}(t))$$

describing the distribution of particles in phase space can be employed to deduce many properties of these correlation functions and vertices.

Likewise, the question of maintaining approximately various crossing symmetries, those for identical particles, and those (in the case of fermions with momentum near the Fermi surface, at vanishing temperature) relating the scattering of particles with energy  $\epsilon - \epsilon_F$  and holes with energy  $-(\epsilon - \epsilon_F)$ , imposes restrictions on the approximation scheme. The simplest aspect of this symmetry is the source of the difference between the original Cooper pair suggestion and the BCS theory of superconductivity. More complex aspects of this symmetry, which require treating various channels on the same footing, have led us to consider the so-called parquet diagrams which were previously studied in connection with quantum electrodynamics. It appears that they may be required to systematically attack the magnetic impurity problems (see Abrikosov, these Proceedings). If the impurities are so rare that we can consider them as separated, and if the interactions between electrons can be neglected, then the electron scattering states can be well defined. Consequently, the simplifications that result when one deals with real rather than virtual particles, and which I have argued are not usually possible in many-particle systems, can be made. Crossing can then be taken into account on the energy shell. As a result, S-matrix techniques can be employed in this problem; Suhl mentions them in these Proceedings in addition to reporting on his more recent studies.

In a certain very restricted sense, we may say that we understand how to discuss all these problems although the calculations are difficult and there is need for better technique. We would particularly like to develop more tractable schemes for systematically dealing directly with fields like  $S_i(rt)$  and  $f(rpt)$  in terms of which the Hamiltonians and observables can be expressed, but for which we have current commutation relations rather than field commutation relations and for which the standard field theoretic techniques cannot be directly applied.

There is another class of problems, however, about which we have much less of a clue. How does one understand the behaviour of the correlations in the field when order is about to set in? Are there approximate theories which can be systematically improved and which explain how fluctuations become larger in range and longer in time until at the critical temperature they extend over infinite distance and last forever so that a state in which the field is non-vanishing becomes appropriate? In brief, how does one understand the complicated infra-red problem associated with

the second order phase transition? In this area we have some suggestions of broad outlines we did not have five years ago, but we do not yet have a theory. In these Proceedings Ferrell and Fisher comment on some of these ideas and Doniach comments on similar effects the interactions can produce at low temperatures when they are almost but not quite strong enough to produce a phase transition at zero temperature. The continuous transition and certain aspects of the first order transition and metastability appear to be the fundamental unsolved many-body problems and we are not presented with their solution here. However, Lieb reports in these Proceedings on the solution of certain special two-dimensional models. Only time will tell the extent to which these exactly soluble models in fewer dimensions shed light on the general problem.

# FIELD THEORY OF PHASE TRANSITIONS\*

R.A. FERRELL

University of Maryland,

College Park, Maryland, United States of America

and

University of Paris,

Orsay, France

## Abstract

**FIELD THEORY OF PHASE TRANSITIONS.** A continuum formulation of the problem of phase transitions is developed from the Van Der Waals-Orstein-Zernike form of the free energy density. The mean field approximation can be systematically corrected for fluctuations by a perturbative or "weak coupling" expansion. This expansion breaks down close to the transition where the fluctuations diverge. A new "strong-coupling" approach is developed for this regime, based upon a saturation property of the higher correlation functions. In this way a derivation is given of the scaling laws of Widom and Kadanoff. The general idea of the "geometrization" of the thermodynamic functions by associating all of the critical variation with a correlation length which tends to infinity is also applied to the dynamical properties. The divergence in the electrical conductivity in the "paraconductive" range of a superconductor just above its transition temperature is discussed in detail. Finally, the method of dynamical scaling is explained in connection with the similar problem of the divergence in the thermal conductivity just above the  $\lambda$ -point of liquid helium.

## 1. INTRODUCTION

The problem of phase transitions and critical phenomena is one involving an infinite number of degrees of freedom. This is evident from the fact that the variable describing the transition (e.g. the mass density in the classical gas, the magnetization density in the ferromagnet, etc.) is free to take on different values at every different point in space. Thus the problem is in fact a problem of classical field theory. A further important aspect which has become abundantly clear from the experimental results of the last few years is that the dependence of the various relevant observable quantities (e.g. compressibility for the gas and susceptibility for the ferromagnet) are not analytic functions of the thermodynamic variables at the critical point [1]. This means that the theoretical solution of the problem has to be of a non-perturbative nature. As non-perturbative solutions of field theoretic problems are rather scarce, the problem may have some general theoretical interest outside the phase transition and many-body area.

Perhaps the simplest and best-known phase transition theory is that of Weiss for the ferromagnetic transition. Here the interaction between neighbouring spins is represented by an effective magnetic field  $\lambda m$ , where  $m$  is the magnetization density and  $\lambda$  is a constant of the material, proportional to the strength of the spin-spin interaction. Thus the total field  $H$  acting upon an individual spin is the sum of this self-field  $\lambda m$  and the applied external field  $H_{ex}$ , and satisfies the equation

$$H = \chi_c^{-1} m + b m^3 = H_{ex} + \lambda m \quad (1.1)$$

---

\* This work was supported in part by the US Air Force Office of Scientific Research and by the US Office of Naval Research.

The Curie paramagnetic susceptibility  $\chi_c(T) = C/T$ , where  $T$  is the absolute temperature and  $C$  is the Curie constant, expresses the linear response of an individual spin to the total field, as shown by the part of the curve of Fig. 1 near the origin. The magnetization saturates at large fields, as illustrated by the horizontal dashed line, which introduces convexity into the curve. This is represented by the cubic term in Eq.(1.1), which is valid only for relatively small values of  $m$  and  $H$ . (For strong fields, more terms of the Taylor's series expansion would have to be included.)

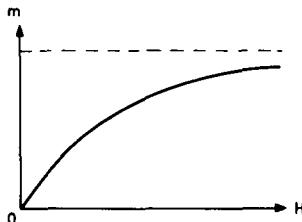


FIG.1. Magnetization versus applied magnetic field for non-interacting spins.

The slope of the graph near the origin is the linear Curie susceptibility and the departure from true linear behaviour can be approximated by including a cubic term.

Equation (1.1) is simplified by collecting all of the linear terms, to give

$$H_{ex} = \chi^{-1} m + b m^3 \quad (1.2)$$

where the susceptibility is now given by

$$\chi^{-1} = \chi_c^{-1} - \lambda \quad (1.3)$$

This is a kind of "Dyson equation" with  $\lambda$  playing the role of a self-energy or "mass operator". Substituting the Curie law for  $\chi_c$  and introducing the Curie temperature  $T_c = \lambda C$  yields

$$\chi^{-1} = (T - T_c)/C \quad (1.4)$$

As illustrated by Fig. 2, the effect of the feedback produced by the self-field  $m$  is to shift the temperature scale for the susceptibility so that the singularity in the Curie susceptibility at  $T = 0$  now falls at the finite temperature  $T_c$ .

As the temperature is reduced below  $T_c$ , the divergence in the susceptibility signifies that the spin system requires no external field in order

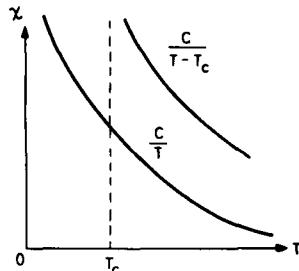
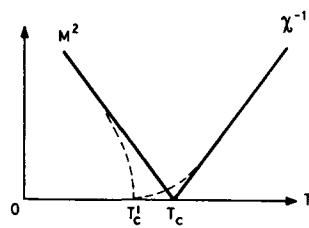


FIG.2. Paramagnetic susceptibility versus temperature.

The mean field approximation has shifted the susceptibility of non-interacting spins by taking into account the interaction in an average way, and has shifted the curve by the Curie temperature  $T_c$ .

FIG.3. Spontaneous magnetization and magnetic stiffness versus temperature.

The straight lines show the expected linear dependence of the square of the spontaneous magnetization,  $M^2$ , in the ordered state of broken symmetry below the transition temperature, and the reciprocal of the paramagnetic susceptibility above the transition temperature. The dashed curves show the deviations found experimentally which give an infinite slope to the magnetization graph and a vanishing slope to the stiffness, at the shifted transition temperature  $T'_c$ .



to acquire polarization. The strength of the spontaneous magnetization in this ordered state of broken symmetry is given by

$$M^2 = (T_c - T)/Cb \quad (1.5)$$

as shown in Fig. 3. The straight lines representing Eqs (1.5) and (1.4) meet at the transition temperature  $T_c$ . Unfortunately, experiment, as shown by the dashed curves, does not bear out these simple predictions of the Curie-Weiss theory. The order does begin to set in at the temperature, say  $T'_c$ , at which the "stiffness"  $x^{-1}$  drops to zero, but the approach of these curves to  $T'_c$  cannot be fit by straight lines.  $x^{-1}$  is found experimentally to have vanishing slope at  $T'_c$ , while  $M^2$  sets in with infinite slope. A fit close to  $\Delta T = T - T'_c = 0$  can be obtained of the form

$$x \propto \Delta T^{-\gamma} \quad (1.6)$$

and

$$M^2 \propto |\Delta T|^{2\beta} \quad (1.7)$$

The critical exponents are found to satisfy the constraint  $\gamma + 2\beta = 2$  and to have numerical values very close, if not exactly equal, to the simple fractions  $\beta = 1/3$  and  $\gamma = 4/3$ . A similar discrepancy is encountered between the temperature dependence of the specific heat predicted by the Curie-Weiss theory and the logarithmic singularity actually observed.

The central task posed to the theoretician by these discrepancies is, first of all, to explain the cause of the breakdown of the simple theory. Having done this and having opened the door to more complicated behaviour, the second step is to establish that the actual non-analytic critical behaviour to be expected is not so complicated after all, but is of the relatively simple form of Eqs (1.6) and (1.7).

## 2. MEAN FIELD APPROXIMATION

Mathematically equivalent forms of the Weiss theory of ferromagnetism can be applied to many different physical systems, so that it is useful to generalize the notation.<sup>1</sup> Instead of  $m$ , let us consider a generalized order

<sup>1</sup> For a review of the mean field theory see Kadanoff [2a]. For some rigorous inequalities independent of the mean field approximation see Josephson [2b], and references cited by him.

parameter  $\eta$  and its conjugate field  $\mu$  (which replaces  $H_{\text{ex}}$ ). The analogue of Eq.(1.2) is obtained by minimizing the free energy density

$$F_{\text{local}} = \frac{a}{2} \eta^2 + \frac{b}{4} \eta^4 - \mu \eta \quad (2.1)$$

where the stiffness is denoted by  $a(T) = \chi^{-1}(T)$  and spatial gradients of  $\eta$  are for the moment neglected. For  $T > T_c$  the quadratic term is stable, but becomes unstable for  $T < T_c$  as shown in Fig. 4. The resulting ambiguity in the minimization gives rise to the breaking of the symmetry which the system exhibits in its high-temperature disordered phase.

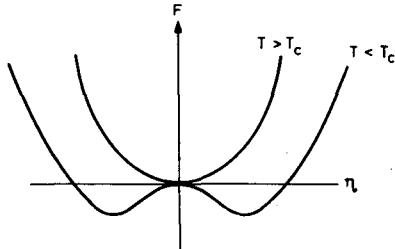


FIG.4. Free energy versus order parameter for the symmetric state ( $T > T_c$ ) and for the state of broken symmetry ( $T < T_c$ ).

In the latter case the minimization of the energy leads to two equivalent alternative solutions for the equilibrium state.

Because  $\eta(\vec{x})$  is free to have spatial variation it is necessary to correct the feedback effect of the self-field for the finite range of the forces. In an isotropic medium, the self-field at point  $\vec{x} = \vec{x}' - \vec{r}$  is given generally by an integral over the non-local kernel  $\lambda(r)$

$$\begin{aligned} & \int \lambda(|\vec{x} - \vec{x}'|) \eta(\vec{x}') d^3x' \\ &= \int d^3r \lambda(r) \left[ \eta(\vec{x}) + \frac{1}{2} (\vec{r} \cdot \nabla)^2 \eta \Big|_{\vec{x}} + \dots \right] \\ &\cong \lambda \eta(\vec{x}) + Z^{-1} \nabla^2 \eta(\vec{x}) \end{aligned} \quad (2.2)$$

The linear term in the Taylor's series expansion vanishes by symmetry, and higher order terms have been neglected. The zeroth and second moments of the kernel have been written as

$$\lambda \equiv \int d^3r \lambda(r) \quad (2.3)$$

and

$$\begin{aligned} Z^{-1} &\equiv \frac{1}{6} \int d^3r r^2 \lambda(r) \\ &\equiv R^2 T_c a' \end{aligned} \quad (2.4)$$

which defines an effective force range  $R$  ( $a' \equiv da/dT$ ). The non-local

correction in Eq.(2.2) contributes to the free energy density the term  $(\nabla \eta)^2/2 Z$  so we have finally

$$\begin{aligned} F' &= F_{\text{local}} + (\nabla \eta)^2 / 2 Z \\ &= \frac{a}{2} \eta^2 + \frac{b}{4} \eta^4 + \frac{Z^{-1}}{2} (\nabla \eta)^2 - \mu \eta \end{aligned} \quad (2.5)$$

The problem of the phase transition and critical phenomena is now entirely contained in the mathematical problem of the evaluation of the partition function

$$Z = \sum_{\eta} e^{-\beta \mathcal{F}_{\Omega}} \quad (2.6)$$

where

$$\mathcal{F}_{\Omega} = \int_{\Omega} d^3 x F \quad (2.7)$$

as a sum over all arbitrary configurations arrived at by independent variation<sup>2</sup> of the order parameter  $\eta(\vec{x})$  at all different points  $\vec{x}$  within a given volume  $\Omega$ . ( $\beta = T^{-1}$  and we use a temperature unit such that Boltzmann's factor equals unity.) A typical such configuration is shown in Fig. 5. If a particular configuration involves too rapid spatial variation of  $\eta(\vec{x})$  the  $(\text{grad } \eta)^2$  term in the free energy will lead to a small Boltzmann factor and a small probability for such a configuration. Once  $Z_{\Omega}$  has been obtained as a function of  $T$  and  $\mu$  in the limit  $\Omega \rightarrow \infty$ , all of the static equilibrium properties of the system can be obtained by appropriate differentiation.

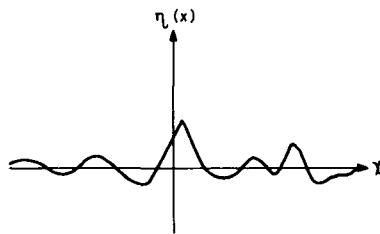


FIG.5. Spontaneous fluctuations in the order parameter as a function of position.

The order parameter is free to vary kinematically independently at each point of space. The partition function is the sum over all such configurations.

For example, the mean value of the integrated order parameter

$$M \equiv \int_{\Omega} d^3 x \eta(\vec{x}) \quad (2.8)$$

is given by

$$\langle M \rangle = T \frac{\partial \ln Z_{\Omega}}{\partial \mu} \quad (2.9)$$

<sup>2</sup> Because of the free variation in  $\eta(\vec{x})$ , the present continuum formulation of phase transitions differs in an essential way from the Ising model, where the order parameter is subject to the constraint  $|\eta| = 1$ . For a review of calculations of the Ising model see Fisher, these Proceedings, and also Ref. [3].

where the angular brackets will denote throughout this paper the statistical average weighted according to the Boltzmann factor occurring in the partition function. Proceeding further we find for the fluctuation  $\delta M = M - \langle M \rangle$

$$\begin{aligned}\langle \delta M^2 \rangle &= \left( T \frac{\partial}{\partial \mu} \right)^2 \ln Z_\Omega \\ &= T \frac{\partial \langle M \rangle}{\partial \mu} = \Omega T \frac{\partial \langle \eta \rangle}{\partial \mu} \\ &= \Omega T \chi\end{aligned}\quad (2.10)$$

and the third moment

$$\begin{aligned}\langle \delta M^3 \rangle &= \left( T \frac{\partial}{\partial \mu} \right)^3 \ln Z_\Omega \\ &= \Omega T^2 \frac{\partial \chi}{\partial \mu}\end{aligned}\quad (2.11)$$

The next step leads to

$$\begin{aligned}T \frac{\partial}{\partial \mu} \langle \delta M^3 \rangle &= T \frac{\partial}{\partial \mu} Z_\Omega^{-1} \sum_{\eta} \delta M^3 e^{-\beta \mathcal{F}_\Omega} \\ &= Z_\Omega^{-1} \sum_{\eta} \delta M^3 \cdot M e^{-\beta \mathcal{F}_\Omega} \\ &\quad - 3 \frac{\partial \langle M \rangle}{\partial \mu} Z_\Omega^{-1} \sum_{\eta} \delta M^2 e^{-\beta \mathcal{F}_\Omega} \\ &\quad - Z_\Omega^{-2} T \frac{\partial Z_\Omega}{\partial \mu} \sum_{\eta} \delta M^3 e^{-\beta \mathcal{F}_\Omega} \\ &= \langle \delta M^3 \cdot M \rangle - 3 T \frac{\partial \langle M \rangle}{\partial \mu} \langle \delta M^2 \rangle \\ &\quad - T \frac{\partial \ln Z_\Omega}{\partial \mu} \langle \delta M^3 \rangle \\ &= \langle \delta M^4 \rangle - 3 \langle \delta M^2 \rangle^2 \equiv \langle \delta M^4 \rangle_c\end{aligned}\quad (2.12)$$

where the subscript signifies the cumulant. Thus, the n-th derivative is the n-th order cumulant

$$\langle \delta M^n \rangle_c = \left( T \frac{\partial}{\partial \mu} \right)^n \ln Z_\Omega \quad (2.13)$$

By substitution from Eq.(2.8), relations of the form of Eq.(2.13) express a connection between the thermodynamic properties (the right-hand member) and the microscopic correlation functions of the general type  $\langle \delta\eta(\vec{x}_1) \delta\eta(\vec{x}_2) \cdots \delta\eta(\vec{x}_n) \rangle$ . The most important of these is that given by Eq.(2.10), which essentially states the connection noted by Einstein between the fluctuations in an equilibrium system and the linear response function for the system. If we Fourier-transform the fluctuation of the order parameter

$$\delta\eta(\vec{x}) = \Omega^{-\frac{1}{2}} \sum_{\vec{k}} \eta_{\vec{k}} e^{i\vec{k} \cdot \vec{x}} \quad (2.14)$$

and define the Green's function by

$$\begin{aligned} G(|\vec{x} - \vec{x}'|) &\equiv \langle \delta\eta(\vec{x}) \delta\eta(\vec{x}') \rangle \\ &= (2\pi)^{-3} \int g(k) e^{i\vec{k} \cdot (\vec{x} - \vec{x}')} d^3 k \end{aligned}$$

where  $g(k) = \langle |\eta_{\vec{k}}|^2 \rangle$ , we find that Eq.(2.10) becomes

$$\int d^3 r G(r) = g(0) = T \chi \quad (2.15)$$

The susceptibility is readily generalized, by considering the response to a field of wave number  $k$ , to the wave number dependent form

$$\begin{aligned} \chi(k) &= \frac{1}{a + Z^{-1} k^2} \\ &= \frac{Z}{\xi^{-2} + k^2} \end{aligned} \quad (2.16)$$

where the correlation length is

$$\xi = (a Z)^{-\frac{1}{2}} \quad (2.17)$$

With this generalization, Eq.(2.15) becomes

$$g(k) = T \chi(k) = \frac{Z T}{\xi^{-2} + k^2} \quad (2.18)$$

so that the inverse Fourier transform gives the Yukawa form

$$G(r) \approx \frac{Z T}{4\pi} \frac{1}{r} e^{-r/\xi} \quad (2.19)$$

a result of Ornstein and Zernike.

Because  $a \rightarrow 0$  as  $T \rightarrow T_c$ , the correlation length must approach infinity as

$$\xi \propto \Delta T^{-\nu} \quad (2.20)$$

where the critical exponent, according to mean field theory, is  $\nu = 1/2$ . This is in disagreement with experiment which indicates  $\nu = 2/3$  close to the critical point. The derivation of this result will be discussed in Section 5. In spite of this quantitative discrepancy we want to emphasize here the following two qualitative features of the Green's function which we believe to have a validity beyond the mean field approximation and to prevail also in an exact calculation.

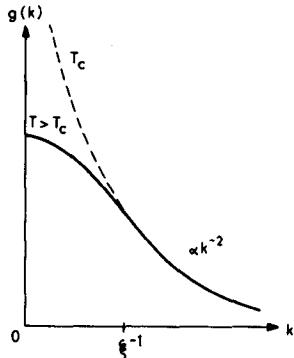


FIG. 6. Fourier transform of the order parameter Green's function  $g(k)$  as a function of wave number  $k$ .

As the critical point is approached, the correlation length  $\xi$  becomes infinity and the Green's function approaches the Coulomb form (dashed curve). For fixed  $k$  no significant change occurs after the correlation length has exceeded the wavelength  $2\pi k^{-1}$ .

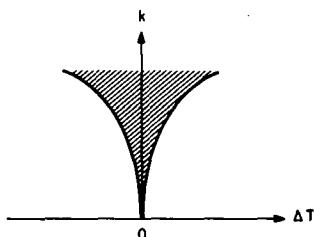


FIG. 7. Region of temperature continuity, as shown by the shaded region.

For given wave number  $k$ , the width of the shaded region is proportional to  $k^{1/\nu}$  where  $\nu = 1/2$  from the mean field approximation and  $2/3$  from the Widom-Kadanoff scaling laws. Within the shaded region only smooth temperature dependence is possible.

### (a) Temperature continuity for $k \neq 0$

This refers to the dependence of  $G(r)$  upon  $\xi$  for fixed  $r$ . As  $\xi \rightarrow \infty$  the tail of the Yukawa function sweeps past the separation  $r$ , producing a strong variation in  $G(r)$ . But once  $\xi \gg r$ , no further significant change occurs in  $G(r)$ . This is especially clearly seen in the Fourier transform, Eq. (2.18), where the denominator is dominated by  $\xi$  for  $\xi^{-1} > k$  but is essentially independent of  $\xi$  when  $\xi^{-1} < k$ . This behaviour is illustrated by the curves of Fig. 6. The temperature width of this region of smooth temperature dependence has the  $k$ -dependence

$$\Delta T \propto k^{1/\nu} \quad (2.21)$$

as shown by the shaded region of Fig. 7.

## (b) Coulomb core

For values of  $r$  less than  $\xi$ , the Green's function can be approximated by its  $\xi \rightarrow \infty$  limit, the Coulomb function

$$G_c(r) = \frac{Z T}{4\pi r} \quad (2.22)$$

In Section 4 reasons are presented for believing that the same simple radial dependence is also possessed by the exact critical Green's function. The effect of finite  $\xi$  can then be treated qualitatively in Eq.(2.15) as a cut-off of the Coulomb tail at radius  $r \approx \xi$  giving the general connection, valid beyond the mean field approximation, between the susceptibility and the correlation length

$$\begin{aligned} \chi &= \beta g(0) = \beta \int d^3r G(r) \\ &\approx 4\pi\beta \int_0^\xi dr r^2 G_c(r) \\ &\approx Z \int_0^\xi dr r = \frac{1}{2} Z \xi^2 \end{aligned} \quad (2.23)$$

This simple geometrical picture, illustrated in Fig. 6, yields the proportionality between the susceptibility and the square of the correlation length. As the precise numerical relationship, we fix the coefficient of  $\xi$  from Eq.(2.18), so that

$$\chi = Z \xi^2 \quad (2.24)$$

Thus, the variation in the susceptibility is completely determined once that of the length  $\xi$  is established, and vice versa. This "geometrization" of the thermodynamic equation of state is expressed most compactly by the relationship between the critical exponents

$$\gamma = 2\nu \quad (2.25)$$

In concluding this section on the mean field approximation, it is worthwhile to note from dimensional considerations the possibility of forming a fundamental length from the basic parameters of the free energy density, Eq.(2.5). First we observe that a trivial change of scale by a factor  $\alpha$  in the definition of the order parameter induces the transform  $b \rightarrow b\alpha^4$  and  $Z^{-1} \rightarrow Z^{-1}\alpha^2$ , so that the combination  $bZ^2$  remains invariant. This quantity has the dimensions of the reciprocal of energy times length so that

$$\xi_0 = (T_c b Z^2)^{-1} \quad (2.26)$$

is a length. As will be seen in Section 4, the criterion for the applicability of the mean field approximation is  $\xi \ll \xi_0$ . When  $\xi > \xi_0$ , the mean field approximation is grossly incorrect, and the equation of state has to be obtained by an entirely different approach, that of strong coupling which is discussed in Section 5. It should further be noted that the condition  $\xi \ll \xi_0$  is equivalent to the criterion derived by Ginzburg [4] for the ordered state, for the validity of the mean field approximation.

### 3. SPECIFIC HEAT

The total free energy contained within volume  $\Omega$  is  $-T \ln Z_\Omega$ . To obtain the critical specific heat, we need study only the temperature dependence which enters through the difference  $\Delta T = T - T_c$ . Thus we can replace the temperature derivative by  $\partial/\partial T = \partial/\partial \Delta T = -\partial/\partial T_c$  to obtain the critical entropy

$$\begin{aligned} S_\Omega &= - \frac{\partial T \ln Z_\Omega}{\partial T_c} \\ &= - \frac{a^4}{2} \int_{\Omega} d^3x \langle \eta^2 \rangle \\ &= - \Omega \frac{a^4}{2} \langle \eta^2 \rangle \end{aligned} \quad (3.1)$$

(For the sake of simplicity we restrict ourselves throughout this section to the so-called "critical isochore",  $\mu = 0$ .) Substitution from Eq. (2.18) gives

$$\begin{aligned} \langle \eta^2 \rangle &= (2\pi)^{-3} \int \langle |\eta_{\vec{k}}|^2 \rangle d^3k \\ &= \frac{1}{2\pi^2} \int g(k) k^2 dk \\ &= \frac{Z T}{2\pi^2} \int \frac{k^2 dk}{k^2 + \xi^{-2}} \\ &= \frac{Z T}{2\pi^2} \int dk - \frac{Z T}{2\pi^2 \xi^2} \int_0^\infty \frac{dk}{k^2 + \xi^{-2}} \\ &= \frac{Z T}{2\pi^2} \int dk - \frac{Z T}{4\pi \xi} \end{aligned} \quad (3.2)$$

The first term depends upon the interaction range  $R$ , which is the shortest wavelength for which the above continuum expressions can be used. But

aside from depending upon this "Debye" cut-off, this term has no critical dependence. The second term alone determines the critical specific heat and is independent of the cut-off. For this reason, the upper limit of the integration has been set equal to  $\infty$ .

Equation (3.2) has been derived in the above manner to exhibit explicitly the fact that the critical specific heat comes only from the long wavelength Fourier components of  $\eta$ . The same result can be obtained more readily, however, by noting that the required expression is simply the autocorrelation function of the order parameter taken at the same point. Expansion of the exponential factor then yields

$$\begin{aligned} \langle \eta^2 \rangle &= \lim_{r \rightarrow 0} G(r) = \lim_{r \rightarrow 0} \frac{ZT}{4\pi r} \left( 1 - \frac{r}{\xi} + \frac{r^2}{2\xi^2} - \dots \right) \\ &= \lim_{r \rightarrow 0} G_c(r) - \frac{ZT}{4\pi \xi} \end{aligned} \quad (3.3)$$

This alternative derivation demonstrates the well-known fact that some care has to be exercised in using the terms "short distance" and "short wavelength" interchangeably. The critical variation of the Green's function at short distances comes entirely from the long wavelength components, as exhibited by the last term in Eq.(3.3).

The critical portion of the specific heat per unit volume is now obtained by substituting the mean field approximation, Eq.(2.17),

$$\begin{aligned} C &= \Omega^{-1} T dS/dT \\ &= \frac{ZT^2 a^4}{8\pi} \frac{d\xi^{-1}}{dT} \\ &= \frac{T^2 (Z a^4)^{3/2}}{16\pi} \Delta T^{-\frac{1}{2}} \\ &= \frac{T a^4}{16\pi b} \frac{\xi}{\xi_0} = \frac{\Delta C}{8\pi} \frac{\xi}{\xi_0} \end{aligned} \quad (3.4)$$

where  $\Delta C$  is the jump in the specific heat at  $T_c$  when fluctuations are neglected entirely. The last line results from making use of Eq.(2.26). This result can be put into a different form which emphasizes more explicitly the long wavelength properties of the system. By again replacing  $\partial/\partial T$  by  $-\partial/\partial T_c$  we obtain

$$\begin{aligned} C &= \frac{T a^4}{2} \frac{\partial \langle \eta^2 \rangle}{\partial T_c} \\ &= \frac{a^4}{4} \int d^3 r G_2(r) \end{aligned} \quad (3.5)$$

$$= \Delta C \frac{2^{-1} b}{T_c} \int d^3 r G_2(r)$$

where the "two-particle" Green's function is

$$G_2(|\vec{x} - \vec{x}'|) \equiv \langle \delta \eta^2(\vec{x}) \delta \eta^2(\vec{x}') \rangle \quad (3.6)$$

The fluctuation in the quadratic field is defined as  $\delta\eta^2(\vec{x}) \equiv \eta^2(\vec{x}) - \langle\eta^2\rangle$ .

An explicit expression for the specific heat now follows from the decoupling approximation ( $\vec{r} = \vec{x} - \vec{x}'$ )

$$\begin{aligned} G_2(r) &= \langle \delta\eta^2(\vec{x}) \delta\eta^2(\vec{x}') \rangle \\ &\approx 2 \langle \delta\eta(\vec{x}) \delta\eta(\vec{x}') \rangle^2 \\ &= 2 G(r)^2 \end{aligned} \quad (3.7)$$

Substituting the mean field Yukawa expression for the Green's function (Eq. (2.19)) into this factored form now gives

$$\begin{aligned} C &= \Delta C \frac{\frac{2}{3}b}{T_c} \int G_2(r) d^3r \\ &= \Delta C \frac{4\pi b}{T_c} \int_0^\infty dr r^2 \left( \frac{ZT}{4\pi r} \right)^2 e^{-2r/\xi} \\ &= \frac{\Delta C}{4\pi \xi_0} \int_0^\infty dr e^{-2r/\xi} \\ &= \frac{\Delta C}{8\pi} \frac{\xi}{\xi_0} \end{aligned} \quad (3.8)$$

which is exactly the result of Eq. (3.4).

Thus we see that the critical variation in the specific heat as well as in the susceptibility is a long wavelength property of the system and can be expressed explicitly in terms of the correlation length  $\xi$ . This geometrization can be effected for all equilibrium thermodynamic functions of the system. The critical variation of such a function follows as soon as the radial dependence of the relevant critical correlation function is determined. If the latter follows the power law  $r^{-n}$ , the corresponding thermodynamic function has the critical variation

$$\int_{r \leq \xi} d^3r r^{-n} \propto \int_0^\xi dr r^{-n+2} \propto \xi^{-n+3}$$

As discussed in Section 2, the susceptibility varies as  $\xi^2$  because of the Coulomb core ( $n=1$ ). The approximation  $G_2 \approx G^2$  would suggest  $n=2$  for the specific heat problem. But the validity of this approximation is limited to  $\xi \ll \xi_0$ . The true critical  $G_2$  can be expected to differ from this and in fact has to be assumed to be  $n=3$  to yield the experimentally observed critical variation,  $C \propto \ln \xi$ . For a smooth fit at  $r = \xi_0$  of the  $r^{-3}$  tail for  $\xi > r > \xi_0$

onto the  $r^{-2}$  portion of  $G_2$  for  $\xi_0 > r > R$ , the coefficient of  $r^{-3}$  has to be of the order of  $(4\pi)^2 T_c/b$ . Inserted into Eq.(3.5) this would yield the logarithmic singularity

$$C \approx \frac{\Delta C}{2\pi} \int_{\xi_0}^{\xi} \frac{dr}{r} = \frac{\Delta C}{2\pi} \ln \frac{\xi}{\xi_0} \quad (3.9)$$

Equation (3.9) signifies that, if the interaction of the fluctuations is such as to produce a logarithmically singular specific heat, then the strength of the logarithmic singularity can be expected to be determined by the size of the jump which would be predicted theoretically on the basis of neglecting the fluctuations completely.

#### 4. WEAK COUPLING

For  $\xi < \xi_0$  the mean field approximation can be systematically improved by taking the non-linear term  $b\eta^4/4$  into account as a perturbation. For simplicity, we continue to restrict ourselves, as in the preceding section, to  $\mu = 0$  and  $\Delta T > 0$ . For this case the mean field approximation ignores the non-linear term entirely. The remaining free energy  $\mathcal{F}_\Omega^{(0)}$  is a diagonal quadratic form in the Fourier components  $\eta_{\vec{k}}$ . Consequently, the thermal average determined by the zero-order partition function

$Z_\Omega^{(0)} = \sum_{\eta} \exp(-\beta \mathcal{F}_\Omega^{(0)})$  of any product  $\eta_{\vec{k}_1} \eta_{\vec{k}_2} \cdots \eta_{\vec{k}_n}$  which we will now denote by  $\langle \eta_{\vec{k}_1} \eta_{\vec{k}_2} \cdots \eta_{\vec{k}_n} \rangle$  must vanish unless all of the wave numbers  $\vec{k}_i$  add pairwise to zero. The remaining factor in the true partition function  $Z_\Omega$  can be treated by a Taylor's series expansion in powers of  $b$ , yielding

$$\begin{aligned} Z_\Omega &= \sum_{\eta} \exp \left[ -\beta \mathcal{F}_\Omega^{(0)} - \frac{b\beta}{4} \int \eta(\vec{x})^4 d^3x \right] \\ &= \sum_{\eta} \exp \left[ -\beta \mathcal{F}_\Omega^{(0)} \right] \sum_{n=0}^{\infty} \left( \frac{-b\beta}{4} \right)^n \frac{1}{n!} \\ &\quad \times \int \cdots \int d^3x_1 \cdots d^3x_n \eta(\vec{x}_1)^4 \cdots \eta(\vec{x}_n)^4 \end{aligned} \quad (4.1)$$

Because  $\mathcal{F}_\Omega^{(0)}$  is a diagonal quadratic form and does not induce any cross-correlations among the  $\eta_{\vec{k}}$ 's, it is possible to establish a simple graphical scheme for the computation of the correlations induced by the  $n > 0$  terms of Eq.(4.1). Some of the graphs through fourth order are exhibited in Fig. 8. A line corresponds to the zero-order pairing of two  $\eta$ -factors (i.e.  $\langle \eta(\vec{x}) \eta(\vec{x}') \rangle_0 = G_0(|\vec{x} - \vec{x}'|)$ ), while a junction of four lines corresponds to the  $\eta^4$  interaction, with strength  $-\beta b/4$ . The number of permutations possible in ascribing the  $n$  variables of integration  $\vec{x}_1, \dots, \vec{x}_n$  to the  $n$  junctions of an  $n$ -th order graph is  $n!$  and just cancels the  $1/n!$  factor in the expansion of the exponential function.

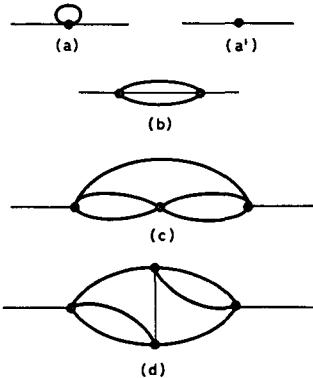


FIG. 8. Weak coupling perturbation graphs, for the self-energy corrections to the order parameter Green's function.

(a') is a counter-term cancelling the critical part of the forward scattering graph (a); (b) corrects the single-particle line, while (c) involves a correction to the two-particle lines. In fourth order both single-particle and two-particle corrections enter, as well as the intrinsic three-particle correction shown in (d).

With these preliminaries, it is easy to see that Dyson's theorem holds, so

$$g(k)^{-1} = g_0(k)^{-1} - \sum (k) \quad (4.2)$$

The "self-energy"  $\Sigma(k)$  is the sum over all proper graphs, of the type shown in Fig. 8. In the language of particle propagators, Fig. 8(a') shows the counter-term which cancels the  $T = T_c$  value of the forward scattering graph Fig. 8(a), which itself can be absorbed in a redefinition of  $G_0$  by a renormalization of the transition temperature. The second-order graph Fig. 8(b) has a weak logarithmic dependence upon the ratio  $\xi/R$ . The higher graphs introduce no further dependence upon the cut-off. Figure 8(c) can be regarded as a correction to one of the two-particle propagators in Fig. 8(b), while Fig. 8(d) shows an intrinsic three-particle effect. There are three other fourth-order graphs, but these are simply two-particle and single-particle corrections to Fig. 8(b).

A great deal of work remains to be done in evaluating these higher order graphs. It is, however, easy to estimate their order of magnitude by noting that four lines emanate from each junction. As each internal line is counted twice, the number of internal lines in a graph of order  $n$  is  $2n-1$ . Thus we obtain from dimensional considerations

$$\begin{aligned} \sum_n & \propto (-b\beta)^n \int \dots \int (d^3 x)^{n-1} G_0^{2n-1} \\ & \propto (-b\beta)^n (Z T)^{2n-1} \xi^{n-2} \\ & \propto a\beta (-b T Z \xi)^n \\ & = a\beta (-\xi/\xi_0)^n \end{aligned} \quad (4.3)$$

where we have substituted from Eq.(2.26). (In all our considerations of critical phenomena we ignore the difference between  $T$  and  $T_c$ , except when the difference  $\Delta T$  stands by itself.) As  $a\beta$  is the mean field approximation to the reciprocal Green's function, it is clear from Eqs (4.2) and (4.3) that

the fractional correction to  $g_0^{-1}$  can be reliably regarded as small only for  $\xi/\xi_0 \ll 1$ . When  $\xi$  becomes comparable with  $\xi_0$ , considerable deviation can be expected from the mean field behaviour. Consequently, as mentioned in Section 2, the range of validity of the mean field approximation is

$$\xi \ll \xi_0$$

It is of interest to exhibit the explicit dependence of the critical region width upon the interaction range. Substitution of Eq.(2.4) gives

$$\xi_0 = T_c \frac{a'^2}{b} R^4 = 2 \Delta C R^4 \quad (4.4)$$

Expressed as a width in the reduced relative temperature  $\tau \equiv \Delta T/T_c$ , this becomes

$$\begin{aligned} \tau_0 &= \frac{a}{T_c a'} = (T_c a' Z \xi_0^2)^{-1} \\ &= (R/\xi_0)^2 \\ &= (2 \Delta C R^3)^{-2} \end{aligned} \quad (4.5)$$

In spin language this is an inverse-square dependence upon the number of neighbours (i.e. number of spins within the range of interaction,  $R$ ).

A further point which we note in concluding this section is the increase in the width of the critical region for a system with plane parallel boundaries of finite thickness  $L$ , where  $R < L < \xi_0$ . The critical fluctuations are then entirely two-dimensional and the Green's function is  $TZ/L$  times a logarithmic factor which does not contribute any dimensional factors but which cuts off the spatial integrations at the correlation length. A further modification is the replacement of  $b/4$  by  $bL/4$  as the coefficient of the  $\eta^4$  term in the free energy. With these changes it is easily seen that the  $n$ -th order graphs are proportional to  $(\beta bL)^n (TZ/L)^{2n} \xi^{2n} = (\xi^2/\xi_0^2)^n$ , where the two-dimensional critical length is the geometric mean

$$\xi_2 = (L \xi_0)^{1/2} \quad (4.6)$$

This in turn determines the temperature width

$$\begin{aligned} \tau_0^{(2)} &= (T_c a' Z \xi_2^2)^{-1} \\ &= \frac{R^2}{\xi_0 L} \\ &= (2 \Delta C L R^2)^{-1} \end{aligned} \quad (4.7)$$

inversely proportional to the two-thirds power of the number of neighbours. Numerical application of this formula to superconducting films is given in Section 6.

## 5. STRONG COUPLING

The perturbative or weak coupling method of the preceding section is no longer useful for taking into account the non-linear interactions when  $\xi$  exceeds  $\xi_0$ . In this case it is necessary to return to the basic thermodynamic relationships of the type of Eq. (2.13), as these have a validity independent of the strength of the non-linearity. As we shall see, in this way it is possible to obtain some useful relationships in the "strong-coupling" regime.

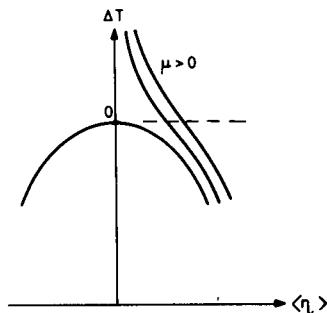


FIG. 9. Dependence of the long range order parameter on both temperature  $\Delta T$  and applied field  $\mu$ .

The horizontal line  $\Delta T = 0$  is the critical isotherm along which variations of  $\mu$  are related to higher order correlation functions of the order parameter field.

To illustrate the method, it is useful first to study the regime  $\xi \ll \xi_0$ , where the mean field approximation is valid. In order to concentrate on the effect of only one variable,  $\mu$ , we restrict ourselves for the moment to the critical isotherm  $\Delta T = 0$ , the horizontal dashed line in Fig. 9. Curves of constant  $\mu$  cross the isotherm at various equilibrium values of  $\langle \eta \rangle$ , so that from Fig. 9 we see that the equation of state along the isotherm reduces to a functional relationship between  $\mu$  and  $\langle \eta \rangle$ . Minimization of the free energy of Eq. (2.5) for  $a = 0$  gives the mean field approximation

$$\mu = b \langle \eta \rangle^3 \quad (5.1)$$

This yields in turn

$$\chi = \frac{\partial \langle \eta \rangle}{\partial \mu} = \frac{1}{3} b^{-1/3} \mu^{-2/3} \quad (5.2)$$

Because of this power law dependence upon  $\mu$ , each successive cumulant, aside from a numerical factor, is determined according to Eq. (2.13) by an additional factor  $T/\mu$  so we have approximately for any positive integer  $n$

$$\begin{aligned} \left| \frac{\langle \delta M^{n+1} \rangle_c}{\langle \delta M^n \rangle_c} \right| &\approx \frac{T}{\mu} \\ &\equiv \xi \langle \delta M^2 \rangle^{1/2} \end{aligned} \quad (5.3)$$

As indicated in Eq. (5.3), this ratio can be reduced to a dimensionless quantity  $\xi$  by dividing by the standard deviation  $\langle \delta M^2 \rangle^{1/2}$  determined from

Eq.(2.10). For the volume  $\Omega$  we choose  $\xi^3$ , the smallest volume for which the thermodynamic formulae apply, and which is assumed to be essentially independent of n. Neglecting numerical factors of order unity we obtain the ratio.

$$\xi = \frac{T_c / \mu}{(T_c \Omega \chi)^{1/2}} \frac{T_c^{1/2} \mu^{-1}}{Z^{1/2} \xi^{5/2}} \quad (5.4)$$

Substitution from Eq.(5.2) of the mean field equation of state

$$\mu^{-1} = 3^{3/2} b^{1/2} Z^{3/2} \xi^3 \quad (5.5)$$

then gives

$$\begin{aligned} \xi &= T_c^{1/2} b^{1/2} Z \xi^{1/2} \\ &= (\xi / \xi_0)^{1/2} \end{aligned} \quad (5.6)$$

Dimensionless measures of the deviation of the statistical distribution  $P(M)$  of  $M$  from a Gaussian curve, as illustrated in Fig.10, are given by the coefficient of skewness

$$C_1 \equiv \langle \delta M^2 \rangle^{-3/2} |\langle \delta M^3 \rangle| \quad (5.7)$$

the coefficient of excess

$$C_2 \equiv \langle \delta M^2 \rangle^{-2} |\langle \delta M^4 \rangle_c| \quad (5.8)$$

and for the  $n$ -th cumulant generally by

$$C_n \equiv \langle \delta M^2 \rangle^{\frac{n}{2}-1} |\langle \delta M^{n+2} \rangle_c| \quad (5.9)$$

Equation (5.4) gives in the mean field approximation

$$C_n \approx \xi^n = (\xi / \xi_0)^{n/2} \quad (5.10)$$

as shown in Fig.11. For  $\xi \ll \xi_0$  we see that the high order correlations are weak but become important as  $\xi \rightarrow \xi_0$ . In fact, Eq.(5.10) even predicts that the  $C_n$  become greater than unity for  $\xi > \xi_0$ . This mean field result violates the following important inequality which we now proceed to prove

$$C_n \lesssim 1 \quad (5.11)$$

It will suffice for our purposes to prove Eq.(5.11) only for the skewness coefficient ( $n = 1$ ). Suppose that inequality (5.11) is violated such that

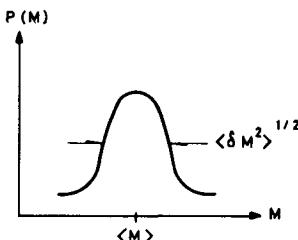


FIG.10. Probability distribution  $P(M)$  of the fluctuating order parameter integrated over a finite volume,  $M$ .

The variance  $\langle (\delta M^2) \rangle$  is proportional to the product of the volume and the static susceptibility.

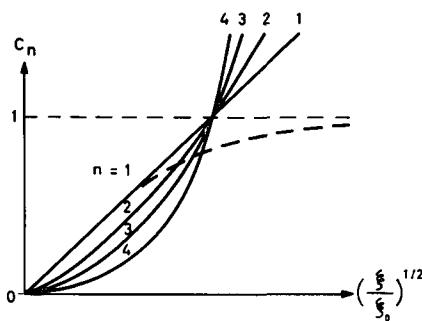


FIG.11. Dimensionless coefficients giving the deviation of Fig.10 from a Gaussian normal distribution.

$n=1$  and 2 are the coefficients of skewness and excess, respectively. The corrections are calculated with the mean field approximation and violate the upper bound shown by the horizontal dashed line, for values of the correlation length  $\xi$  greater than the characteristic length  $\xi_0$ . This illustrates the complete breakdown of the mean field approximation in this region. The dashed line shows the behaviour which corresponds to the strong coupling approach and which leads to the Widom-Kadanoff scaling laws.

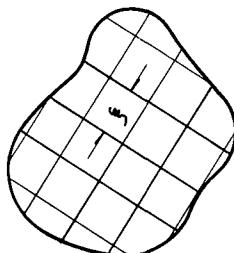


FIG.12. Division of a volume  $\Omega'$  into independently fluctuating cells of the order of the correlation length.

Because of the lack of correlation between different cells, the central limit theorem applies to a variable defined over the entire volume  $\Omega'$ .

$C_1 = K \gg 1$ . Then we can return to the definition of  $C_1$  and choose a new volume  $\Omega' = K^2 \Omega = K^2 \xi^3$ . This yields a new skewness coefficient  $C_1' = C_1 (\Omega / \Omega')^{1/2} = C_1 / K = 1$ . But this is in contradiction to the fact that  $\Omega'$  can be divided up as shown in Fig.12 into  $K^2$  cells of the order of  $\xi^3$ . By the basic nature of the correlation length these cells are uncorrelated with one another so that we sum over cells,  $M = \sum_{i=1}^K M_i$ , with  $M_i$  uncorrelated with  $M_j$  for  $i \neq j$ . Hence the central limit theorem applies and yields a Gaussian distribution for  $M$ , with necessarily  $C_1' \ll 1$ . It follows that the original assumption of  $C_1 \gg 1$  is false, and that this coefficient is bounded in accordance with Eq.(5.11).

The first conclusion from Eq.(5.11) is that the mean field equation of state is grossly in error for  $\xi > \xi_0$  and cannot be applied in this region. A more quantitative application follows from the strong coupling assumption that Eq.(5.11) "saturates" and becomes an equality for  $\xi \gg \xi_0$ , as indicated by the dashed curve in Fig.11. With this assumption we have an equation which effectively replaces the equation of state. Setting  $\xi = 1$  in Eq.(5.4) we obtain

$$\xi = T_c^{1/5} Z^{-1/5} \mu^{-2/5} \quad (5.12)$$

and

$$\begin{aligned} \chi &= \frac{\partial \langle \eta \rangle}{\partial \mu} = Z \xi^2 \\ &= Z \left( T_c^{1/2} Z^{-1/2} \mu^{-1} \right)^{4/5} \\ &= T_c^{2/5} Z^{3/5} \mu^{-4/5} \end{aligned} \quad (5.13)$$

Integration yields

$$\langle \eta \rangle = 5 T_c^{2/5} Z^{3/5} \mu^{1/5} \quad (5.14)$$

where the critical exponent [5]  $\delta = 5$  replaces the mean field value of  $\delta = 3$  from Eq.(5.1). This result expresses some of the content of the so-called "static scaling laws" of Widom [6] and Kadanoff [7]. The present treatment has the advantage that it gives an estimate of the numerical coefficient as well as the critical exponent.

Returning to Eq.(5.3) we can reinterpret the strong coupling result  $\xi = 1$  in terms of the following relationship between the  $n$ - and  $n+1$ -point correlation functions:

$$\begin{aligned} &\int d^3x_1 \cdots d^3x_n d^3x_{n+1} \langle \delta\eta(\vec{x}_1) \cdots \delta\eta(\vec{x}_n) \delta\eta(\vec{x}_{n+1}) \rangle \\ &= Z^{1/2} T_c^{1/2} \xi^{5/2} \int d^3x_1 \cdots d^3x_n \langle \delta\eta(\vec{x}_1) \cdots \delta\eta(\vec{x}_n) \rangle \end{aligned} \quad (5.15)$$

It is interesting to note that this ratio could be obtained purely formally by imagining that the dependence upon the additional variable  $\vec{x}_{n+1}$  is proportional to the square root of the two-point Green's function so that the integration over  $\vec{x}_{n+1}$ , cut off at  $\xi$ , is of the order of  $Z^{1/2} T_c^{1/2} \xi^{5/2}$ . In other words, the extra factor of  $\delta\eta(\vec{x}_{n+1})$  in the  $n+1$ -point function, relative to the  $n$ -point function, is effectively equal to the square root of the two-point function.

Just as Eq.(5.14) answers the question of the dependence of the variation of the correlation length along the isotherm in the strong coupling regime when the fluctuations and non-linear interaction dominate, a similar treatment determines the variation along the isochores. In this case we are dealing with the cumulants of the integrated quadratic field

$$\langle \delta Q^n \rangle_c = \left( \frac{T}{a'} \frac{\partial}{\partial T_c} \right)^n \ln Z \quad (5.16)$$

for  $n \geq 2$ , where

$$Q \equiv \int_{\Omega} d^3x \eta^2(\vec{x}) \quad (5.17)$$

The case  $n = 2$  has already appeared in Eq.(3.5) and can be rewritten as

$$\langle \delta Q^2 \rangle = \Omega \frac{T_c}{2b} \frac{C}{\Delta C} \quad (5.18)$$

where the specific heat per unit volume  $C$  now includes a non-critical background term, generally of order  $\Delta C$  or  $T_c a'^2 / 2b$ , in addition to the term exhibiting critical behaviour. It is convenient to introduce the logarithmic derivative  $\mathcal{L} \equiv |\partial \ln C / \partial \Delta T|$  so that the skewness coefficient becomes (upon neglecting numerical coefficients and taking the volume equal to the critical value  $\Omega = \xi^3$ )

$$\begin{aligned} C_1^{(2)} &= \langle \delta Q^2 \rangle^{-3/2} |\langle \delta Q^3 \rangle| \\ &= \langle \delta Q^2 \rangle^{-3/2} \frac{T}{a'} \left| \frac{\partial \langle \delta Q^2 \rangle}{\partial T_c} \right| \\ &= \langle \delta Q^2 \rangle^{-1/2} \frac{T}{a'} \mathcal{L} \\ &= \Omega^{-1/2} (T_c b Z)^{1/2} (Z a')^{-1} \mathcal{L} \\ &= \xi^{-3/2} \xi_0^{-1/2} (Z a')^{-1} \mathcal{L} \end{aligned} \quad (5.19)$$

In the case  $\xi \ll \xi_0$ , the critical portion is only a small fraction of the total specific heat and we can substitute Eq.(3.4) to find  $\mathcal{L} \approx \Delta T^{-1} \xi / \xi_0$ . According to Eq.(5.19) this yields

$$C_1^{(2)} \approx (\xi / \xi_0)^{3/2} \quad (5.20)$$

As expected, the skewness coefficient is small in the weak coupling region  $\xi \ll \xi_0$  but becomes of the order of unity for  $\xi \approx \xi_0$ .

Clearly, the mean field approximations cannot be used for  $\xi > \xi_0$  as Eq.(5.20) would then yield  $C_1^{(2)} > 1$ , which would violate the inequality of Eq.(5.11). Making the strong coupling assumption that Eq.(5.11) saturates in this region at  $C_1^{(2)} = 1$ , and taking the critical specific heat from experiment to be logarithmic, without further theoretical investigation, we obtain  $\mathcal{L} \approx \Delta T^{-1} = T_c^{-1} \tau^{-1}$  and

$$\begin{aligned} C_1^{(2)} &= 1 \approx \xi^{-3/2} \xi_0^{-1/2} (T_c Z a')^{-1} \tau^{-1} \\ &\approx (\xi / \xi_0)^{3/2} (\tau_0 / \tau) \end{aligned} \quad (5.21)$$

Here we have introduced the expression for  $\tau_0$ , the temperature width of the critical region from Eq.(4.5). This yields for the correlation length the critical behaviour

$$\xi \approx \xi_0 (\tau / \tau_0)^{-\nu} \quad (5.22)$$

where  $\nu = 2/3$  instead of the mean field value of  $1/2$ .

It is interesting to note that, just as in the case of the  $\delta\eta(x)$  fluctuations, this strong-coupling result is formally equivalent to having the correlation function for the  $\delta\eta^2(x)$  fluctuations of order  $n+1$  equal to that of order  $n$  times the square root of the two-point function.

We close this discussion of the strong coupling approach by noting that useful relations can be obtained from the various higher order cross-correlations of the fluctuations in linear and quadratic fields,  $\delta\eta(x)$  and  $\delta\eta^2(x)$ . In this way one can cross the critical isotherm and arrive at the curve of spontaneous broken symmetry,  $|\langle\eta\rangle| \propto |\Delta T|^\beta$ . The mean field result  $\beta = 1/2$  of Eq.(1.5) is replaced by the strong coupling critical exponent  $\beta = 1/3$ , in good agreement with experimental data. But for the sake of brevity we do not enter into the details of the calculation here.

## 6. DYNAMICAL PROPERTIES

The dynamical properties and transport coefficients, as well as the static response functions discussed above, exhibit non-analytic critical behaviour near phase transitions. The above work is extended to include the dynamics by replacing the static by the time-dependent Green's function. Some of the critical phenomena exhibited by a real field have been discussed by Kadanoff and Swift [8] and by Mountain and Zwanzig [9]. In order to illustrate the basic idea it is more convenient to study here the case of a pair of coupled real fields  $\eta_{1,2}$ , which can be combined into the single complex field  $\eta = \eta_1 + i\eta_2$ . We approach the critical point along  $\mu = 0$  so  $\langle\eta\rangle = 0$ . Introduction of absolute values into Eq.(2.5) then converts it to the standard Landau-Ginzburg form. In the presence of a vector potential  $\vec{A}(x)$  the non-local term has to have the form  $|(\nabla - iq\vec{A}/c)\eta|^2/2Z$  in order to be gauge-invariant ( $c$  is the velocity of light and  $q$  is the charge carried by the field). The current density is given by the variational derivative of the free energy  $\mathcal{F} = \int d^3x F$

$$\begin{aligned} \vec{J}(\vec{x}) &= -c \delta\mathcal{F}/\delta\vec{A}(\vec{x}) \\ &= -\frac{iq}{2Z} \left[ \eta^* \left( \nabla - i \frac{q}{c} \vec{A} \right) \eta \right. \\ &\quad \left. - \eta \left( \nabla + i \frac{q}{c} \vec{A} \right) \eta^* \right] \\ &= Z^{-1} q (\eta_1 \nabla \eta_2 - \eta_2 \nabla \eta_1) \end{aligned} \quad (6.1)$$

where we have set  $\vec{A} = 0$  and have reintroduced the real fields.

Simultaneous fluctuations of the fields produce fluctuations in the current which according to the fluctuation-dissipation theorem imply a static long wavelength conductivity

$$\begin{aligned} \sigma &= \frac{1}{6T} \int d^4\tau_{21} \langle \vec{J}(2) \cdot \vec{J}(1) \rangle \\ &= -\frac{q^2}{3Z^2T} \int d^4\tau_{21} \langle \eta_1(2) \nabla \eta_2(2) \cdot \eta_1(1) \nabla \eta_2(1) \\ &\quad - \eta_1(2) \nabla \eta_2(2) \cdot \nabla \eta_1(1) \eta_2(1) \rangle \\ &= -\frac{2q^2}{3Z^2T} \int d^4\tau_{21} \langle \eta_1(2) \nabla \eta_2(2) \cdot \nabla \eta_1(1) \eta_2(1) \rangle \end{aligned} \quad (6.2)$$

The integration is over all space-time separations of the points 1 and 2. In the weak coupling regime we can factor the integrand in terms of the time-dependent Green's function

$$\begin{aligned} G(2,1) &\equiv \langle \eta_1(2) \eta_1(1) \rangle \\ &= \langle \eta_2(1) \eta_2(2) \rangle \end{aligned} \quad (6.3)$$

Thus Eq.(6.2) becomes

$$\begin{aligned} \sigma &= -\frac{2q^2}{3Z^2T} \int d^4\tau_{21} \langle \eta_1(2) \nabla \eta_1(1) \rangle \\ &\quad \cdot \langle \nabla \eta_1(2) \eta_2(1) \rangle \\ &= \frac{2q^2}{3Z^2T} \int d^3r dt |\nabla_{\vec{r}} G(\vec{r}, t)|^2 \end{aligned} \quad (6.4)$$

where  $\vec{r}$  and  $t$  now denote the space and time separations, respectively.

If the Fourier components of  $\eta_{1,2}$  of wave number  $k$  relax at rate  $\Gamma_k$  we can write

$$G(\vec{r}, t) = (2\pi)^{-3} \int d^3k g(k) e^{-\Gamma_k |t|} e^{i\vec{k} \cdot \vec{r}} \quad (6.5)$$

where  $g(k)$  is the Fourier coefficient of the static Green's function. Substituting Eq.(6.5) into Eq.(6.4) we obtain

$$\sigma = \frac{q^2}{12\pi^3 Z^2 T} \int d^3 k \, k^2 g(k)^2 \int_{-\infty}^{+\infty} dt e^{-2\Gamma_k |t|}$$

$$= \frac{q^2}{3\pi^2 Z^2 T} \int_0^{\infty} dk \, k^4 g(k)^2 \Gamma_k^{-1} \quad (6.6)$$

If we further assume  $\Gamma_k = \lambda_0 g(k)^{-1}$ , as can be explicitly verified in the microscopic theory ( $\lambda_0$  is a  $k$ -independent proportionality factor), and substitute from Eq. (2.18) we find

$$\sigma = \frac{q^2}{3\pi^2 Z^2 T \lambda_0} \int_0^{\infty} dk \, k^4 g(k)^3$$

$$= \frac{q^2 Z T^2}{3\pi^2 \lambda_0} \int_0^{\infty} dk \, \frac{k^4}{(k^2 + \xi^{-2})^3} \quad (6.7)$$

$$= \frac{q^2 Z T^2}{16\pi \lambda_0} \xi$$

Equation (6.7) can be applied to superconductors for which  $q$  is twice the elementary charge because of Cooper pairing of the electrons. As conductivity measurements are generally carried out on thin films, say of thickness  $L$ , the three-dimensional integration has to be replaced by  $2\pi/L$  times a two-dimensional integral. Including a factor  $3/2$  to allow for the restriction of the polarization of the current fluctuations within the plane of the film yields a net correction factor for two-dimensionality of  $2\xi/L$ , so

$$\sigma_{2\text{-DIM}} = \frac{e^2 T^2 Z}{2\pi L \lambda_0} \xi^2$$

$$= \frac{e^2 T^2}{2\pi L \lambda_0 a} \quad (6.8)$$

$$= \frac{e^2 T}{2\pi L} \Gamma_0^{-1}$$

Here we have reintroduced the long wavelength relaxation time  $\Gamma_0^{-1} = \lambda_0^{-1} g(0) = T/(\lambda_0 a)$ , which is of the order of and has the temperature dependence of  $\Delta T^{-1}$ . (We use units in which Planck's constant equals  $2\pi$ .) Thus we are led to  $\sigma_{2\text{-DIM}} \propto e^2 L^{-1} \tau^{-1}$ , where  $\tau = \Delta T/T_c$  is the reduced relative temperature. By including the numerical coefficients from the BCS theory, Schmidt [10] and Aslamazov and Larkin [11] have found good agreement with the measurements of Glover [12] and of Strongin et al. [13].

Equation (6.8) can be compared with the standard Drude expression for the normal state conductivity,

$$\begin{aligned}\sigma_{D.C.} &= n e^2 \ell / p_F \\ &= e^2 p_F^2 \ell / 3\pi^2\end{aligned}\quad (6.9)$$

where  $n$ ,  $m$ ,  $p_F$  and  $\ell$  are the electron density, mass, Fermi momentum, and mean free path, respectively. Thus the ratio of the two conductivities is of the order of

$$\frac{\sigma_{2-DIM}}{\sigma_{D.C.}} \approx \frac{e^2 / L\tau}{e^2 p_F^2 \ell} = (p_F \ell)^{-1} (p_F L)^{-1} \tau^{-1} \quad (6.10)$$

This is approximately unity for a reduced relative temperature

$$\tau \approx (p_F \ell)^{-1} (p_F L)^{-1} \quad (6.11)$$

which has typically the value  $10^{-4}$  for a film 500 Å thick and of mean free path 5 Å. The inverse proportionality to  $L$  has recently been verified [14]. It is worthwhile to note that a value of the correlation length follows from substitution of Eq.(6.11) into Eq.(2.17). From the BCS theory we know that the specific heat jump at  $T_c$ ,  $\Delta C = T_c a'^2 / 2b$ , equals, up to a numerical factor, the normal state Sommerfeld specific heat, or  $T_c$  times the density of single-particle energy levels per unit energy and volume. Therefore, we can replace  $p_F^2/v_F$  by  $a'^2/b$ . Furthermore, the range of interaction is  $R \approx (v_F \ell / T_c)^{1/2}$ , so  $Z^{-1} = a' T_c R^2 \approx a' v_F \ell$ . Thus we obtain from Eq. (6.11)

$$\begin{aligned}\xi^2 &= (a' \tau T_c Z)^{-1} \\ &\approx \frac{p_F^2 \ell L}{a' T_c Z} \\ &\approx \frac{a' v_F \ell L}{b T_c Z} \\ &\approx \frac{L}{b T_c Z^2} = L \xi_0^2 = \xi_2^2\end{aligned}\quad (6.12)$$

where  $\xi_2$  is the two-dimensional critical correlation length determined in Eq.(4.6).

This result signifies that, as soon as the transition has been approached sufficiently closely from above ("paraconductivity" region), so that the fluctuations have contributed to the conductivity an amount equal to the D.C. background value, then the interactions of the fluctuations can no longer be

neglected. For  $\xi > \xi_2$  the fluctuations have to be treated self-consistently and deviations can be expected from the Curie-Weiss temperature dependence  $\tau^{-1}$  predicted by Eq.(6.8). It is clear in retrospect that the estimate of Ferrell and Schmidt [15], which involved the total effect of the fluctuations of all wave numbers, was not related to the width of the critical region but instead was essentially the renormalization of the transition temperature associated with the graphs of Figs 8(a) and (a'). Although this is in principle an observable temperature shift, it is not as unambiguously identifiable and hence not as interesting from an experimental point of view as the width of the critical region itself. In any case, the estimate has to be modified for the two-dimensionality of the films. When this is done, one finds a long wavelength logarithmic divergence of the transition temperature renormalization, suggesting that the long wavelength fluctuations completely suppress the transition. This result would seem to be compatible both with the theoretical impossibility of long range order for dimensionality lower than three [16, 17, 18] and with the experimental behaviour of the films. The resistance data [12, 13, 14] generally show a "foot" for small resistance values, with a temperature derivative of the resistance which vanishes with decreasing temperature. Therefore, there is not a definite temperature at which the resistance vanishes, and there exists the interesting possibility that a very small residual resistance may remain at all temperatures.

The preceding work has to be modified in the critical region. The dynamics so far have been studied only in the weak coupling approximation, and temperature dependence can be expected to be different in strong coupling. A method of "dynamical scaling" has been established for this purpose. We will here give only a very brief sketch of this method as a detailed exposition of it has recently been published [19]. The method can be applied to any system which possesses an infinite degeneracy in the ordered state and consequently has excitations of the Goldstone boson type. (The superconductor, because of the long-range Coulomb force, does not have a critical Goldstone boson, but in its place possesses the plasmon. For this reason, the superconductor has to be studied as a special case and the general discussion given here does not apply to it in all details.)

It is generally true that the frequency of the Goldstone bosons has a temperature dependence which is proportional to that of the order parameter itself or, in other words, proportional to  $|\Delta T|^\beta$  where  $\beta = 1/3$ . The method of dynamical scaling simply uses the temperature continuity for finite wave number which was enunciated in Section 2 and appears as Eq.(2.21). Thus the temperature dependence has to deviate from  $|\Delta T|^\beta$  at a value of temperature given by Eq.(2.21). This is just the shaded region of temperature continuity shown in Fig. 7. Then the temperature-dependent factor becomes replaced by the square root of the wave number. Continuing to pass to higher temperatures and away from the ordered phase, we must then find that the frequency spectrum which was originally described by the Goldstone bosons (i.e. more or less sharp lines) has merged into a single central line, the so-called Landau-Placzek central peak. The width of the central peak is  $Dk^2$ , where the thermal diffusion coefficient D must have a temperature dependence such that it fits smoothly to the frequency width estimated from the Goldstone bosons by using the temperature continuity and passing smoothly into the shaded region of Fig. 7.

As a specific application we consider superfluid He II. The Goldstone boson is the second sound vibration and for wave number  $k$  has the frequency

$$\begin{aligned}\omega_k &= C_2 k \\ &\propto |\Delta T|^{1/3} k \\ &\rightarrow k^{3/2} \\ &\rightarrow \Delta T^{-1/3} k^2\end{aligned}\tag{6.13}$$

The first arrow indicates entry into the shaded region of Fig. 7 and the second arrow indicates leaving it towards higher temperatures. Thus we see in this case that the temperature continuity requirement leads to a divergence in the thermal conductivity of liquid He I just above the  $\lambda$ -point. Similarly, the underlying two-fluid relaxation rate, assuming  $k$  independence and assuming that it has to join onto the Goldstone boson spectrum in the shaded region of Fig. 7, necessarily has a linear dependence upon  $\Delta T$ . These predictions have experimental consequences in the He I thermal conductivity and in the damping of first sound in He II, and have received confirmation in both cases.

The critical properties of the isotropic antiferromagnet can also be related to the Goldstone bosons in the ordered state, which are, in this case, the spin waves. Here there is no fundamental theory, analogous to the two-fluid model for helium, which requires the temperature dependence of the Goldstone boson frequencies to be that of the order parameter (in this case, the sub-lattice magnetization). However, making this assumption, which is borne out by experiment, Halperin and Hohenberg [20] have noted that the mathematical properties of the antiferromagnet become identical to those of liquid helium in the vicinity of the  $\lambda$ -point. Consequently, one expects the spectral width at the transition temperature (i.e. the Neel temperature) to vary as the  $3/2$  power of the wave number. This prediction has been confirmed by inelastic neutron scattering [21].

In the ferromagnet the Tahir-Kheli [22] sum rule can be used to show that the Goldstone boson frequency, or, in other words, the spin-wave frequency, is proportional to

$$\begin{aligned}\omega_k &\propto \langle \eta \rangle k^2 \\ &\propto |\Delta T|^{1/3} k^2 \\ &\rightarrow k^{5/2}\end{aligned}\tag{6.14}$$

where we have again replaced the temperature variation by the wave number dependence specified by Eq. (2.21). Recent neutron diffraction measurements seem to bear out this prediction [23].

From these examples it is clear that the geometrization already employed for the static properties of a system undergoing a phase transition

can easily be extended, with non-trivial results ensuing, to the dynamical or transport properties. From this point of view it would seem that the basic mathematical difficulties posed by the phase transition problem are already exhibited by the static properties, or the equal-time correlation functions. Once a complete detailed theory of the static properties is established, it seems fairly likely that its dynamical consequences should then follow without serious difficulty.

## 7. SUMMARY

The basic mathematical structure of the problem of phase transitions consists in a hierarchy of multi-point Green's functions. It is possible to derive differential equations relating the lower order Green's functions to the higher order ones. For the sake of brevity we have not done this, particularly because a frontal attack on the solution of these equations has not been successful up to the present time.

But it is possible to take into account the higher order correlations by a weak coupling approach which is discussed in Section 3. There it has been shown that the higher order graphs which correspond to the higher order Green's functions have only a small effect if one is sufficiently far from the phase transition. Thus, in a sense, in this range the hierarchy truncates automatically, and the weak coupling approach provides a practical means of calculating small corrections to expressions derived from the mean field approximation. A further result of the weak coupling expansion is the identification of a characteristic correlation range and corresponding temperature width for the critical region. Within this critical region the higher correlations become dominant and the hierarchy of Green's functions cannot and must not be truncated.

Within the critical region a strong coupling approximation is possible, as described in detail in Section 5. This approximation consists in establishing a relationship between successive orders of the higher order correlation functions. Expressed roughly, this relationship is essentially that the ratio of successive Green's functions is the square root of the two-point function. With this connection between the higher order correlation functions the scaling laws of Widom [6] and Kadanoff [7] can be obtained.

Throughout the paper, every critical variation is referred to the correlation length. In this way the thermodynamic and statistical mechanical problems of deriving the equation of state are converted into the geometrical problem of how strong do the various correlation functions reach out into configuration space and how far out do they reach. The basic aspect of this geometrization is that when one studies the behaviour of a finite wavelength property, then one should expect no significant variation in this property once the correlation length has become larger than the wavelength being studied. This principle of temperature continuity is equivalent to the familiar statement that a finite system cannot show a sharp phase transition. When this principle is extended to the time-dependent Green's functions, by the assertion that the spectrum for a finite wave number should not show significant temperature variation within a temperature range of the critical point which is proportional to the  $3/2$  power of the wave number, it becomes the dynamical scaling approach and gives results for the critical variation of the transport coefficients.

In conclusion, the essential problem of phase transitions is encountered in the question of what are the radial dependencies of the critical static correlation functions, i.e. the correlation functions precisely at the critical point. With the Coulomb core assumption, the Green's function for the order parameter itself is essentially already known. We have made this reasonable assumption for the sake of concreteness, but the discussion given in this paper is easily generalized to any other radial dependence for this two-point correlation function. In any case, the principal difficulty which remains to be solved concerns the critical correlation function for the quadratic field. This corresponds to the specific heat, which is known experimentally to diverge logarithmically (or possibly slightly more singularly) as the critical point is approached. Thus a complete mathematical solution of the phase transition problem has to provide a critical correlation function for the square field which varies inversely as the cube (or a slightly lower power) of the separation. This remains as the "hard core" unsolved problem.

### R E F E R E N C E S

- [1] Proc. Conf. Critical Phenomena, Washington, 1965 (GREEN, M. S., SENGERS, J. V., Eds). HELLER, P., Rep. Prog. Phys. 30 (1967) 731.
- [2a] KADANOFF, L.P. et al., Rev. mod. Phys. 39 (1967) 395.
- [2b] JOSEPHSON, B.D., Proc. Phys. Soc. Lond. 92 (1967) 276.
- [3] FISHER, M.E., J. math. Phys. 5 (1964) 944; Rep. Prog. Phys. 30 (1967) 615.
- [4] GINZBURG, V.L., Fizika tverd. Tela 2 (1960) 2; Soviet Phys. solid St. 2 (1960) 1824.
- [5] GREEN, M.S., VICENTINI-MISSONI, M., LEVELT SENGERS, J.M.H., Phys. Rev. Lett. 18 (1967) 1113.
- [6] WIDOM, B., J. chem. Phys. 43 (1965) 3898; 3892.
- [7] KADANOFF, L.P., Physics 2 (1966) 263.
- [8] KADANOFF, L.P., SWIFT, J., Phys. Rev. 165 (1968) 310.
- [9] MOUNTAIN, R., ZWANZIG, R., J. chem. Phys. (Feb. 1968).
- [10] SCHMIDT, H., "Fluctuations in a superconductor above  $T_c$ ", preprint, Inst. Max von Laue-Paul Langevin (to be published).
- [11] ASLAMAZOV, L.G., LARKIN, A.I., Physics Lett. 26A (1968) 238.
- [12] GLOVER, R.E., III, Physics Lett. 25A (1967) 542.
- [13] STRONGIN, M., KAMMERER, O.F., CRON, J., THOMPSON, R.S., FINER, H.L., Phys. Rev. Lett. 20 (1968) 922.
- [14] NAUGLE, D., private communication (to be published).
- [15] FERRELL, R.A., SCHMIDT, H., Physics Lett. 25A (1967) 544.
- [16] FERRELL, R.A., Phys. Rev. Lett. 13 (1964) 330.  
De WAMES, R.E., LEHMAN, G.W., WOLFRAM, T., Phys. Rev. Lett. 13 (1964) 749.
- [17] RICE, T.M., Phys. Rev. 140A (1965) 1889.
- [18] HOHENBERG, P.C., Phys. Rev. 158 (1967) 383.
- [19] FERRELL, R.A., MENYHARD, N., SCHMIDT, H., SCHWABL, F., SZEPFALUSY, P., Ann. Phys. 47 (1968) 565; Phys. Rev. Lett. 18 (1967) 891.
- [20] HALPERIN, B.I., HOHENBERG, P.C., Phys. Rev. Lett. 19 (1967) 700.
- [21] NATHANS, R., MENZINGER, R., PICKART, S., Proc. Magnetism Conf., Cambridge, Mass., September 1967; J. appl. Phys. (1968).
- [22] TAHIR-KHELI, R., preprint.
- [23] COLLINS, M.F., NATHANS, R., PASSELL, L., SHIRANE, G., Bull. Am. phys. Soc., Series II 13 (1968) 616, and private communication.

# LOCALIZED MAGNETIC MOMENTS IN METALS

H. SUHL

University of California, La Jolla,  
San Diego, Calif., United States of America

## Abstract

LOCALIZED MAGNETIC MOMENTS IN METALS. Although there has been some progress in the understanding of the so-called s-d model of conduction electron-impurity interaction in magnetic alloys, the model itself is subject to certain basic weaknesses. In this paper an attempt is described to derive the magnetic properties of such an impurity starting with a more basic model. No explicitly spin-dependent forces are assumed. The impurity is characterized by an ordinary potential, and the electrons are allowed to interact. Quantities such as the susceptibility are then calculated by many-body perturbation theory. The usual divergence (i.e. the Hartree-Fock condition for the emergence of a local moment) is avoided by renormalization of the propagators. A Curie law then results for high temperatures and adequate interaction strength, etc. At very low temperatures, the susceptibility saturates.

Very often, a magnetic impurity in a metal is treated as a well-defined agglomerate of electrons, coupled together to give a definite spin  $S$ , and, in some sense or other, distinguishable from the conduction electrons<sup>1</sup>. One can hardly question that such a description is at least approximately correct in cases where the electrons composing  $S$  belong to a shell whose energy lies well below the bottom of the conduction band. The matter is very much in doubt, however, if no such well-defined shell exists, and the magnetic moment arises from a virtual level whose energy lies within the conduction band (often near the Fermi level) in accordance with the descriptions of Friedel, Anderson, Wolff, Blandin and others.

In that case it seems more logical to approach the problem from the viewpoint of spin density fluctuations. Such fluctuations are always present, even in a pure metal, where they are known as paramagnons<sup>2</sup>. They have many interesting manifestations; historically, the first was the so-called Pauli paramagnetism, according to which the spin susceptibility of a degenerate electron gas goes to a constant as the temperature approaches zero. It can be shown that a rigorous expression for the spin susceptibility (in the absence of spin orbit coupling and anisotropies) must have the form

$$x \sim \langle \vec{S}^2 \rangle / T$$

where  $\vec{S}$  is the total spin of the system and  $\langle \vec{S}^2 \rangle$  denotes the thermal average of its square. The fact that  $x \rightarrow \text{constant}$  as the temperature  $T$  tends to the absolute zero must therefore mean that  $\langle \vec{S}^2 \rangle$  is proportional to  $T$ . Also, since the only other energy in the problem is the Fermi

<sup>1</sup> For details of the consequences of this model see the paper by Abrikosov in these Proceedings.

<sup>2</sup> See the paper by Doniach in these Proceedings.

energy  $\epsilon_f$ , it is evident on dimensional grounds that

$$\langle \vec{S}^2 \rangle \sim N \frac{k' T}{\epsilon_f}$$

where  $k$  is Boltzmann's constant.

That  $\langle S^2 \rangle$  is proportional to  $N$ , the total number of particles, is evident from the fact that  $\vec{S}^2 = \sum_{ij} \langle \vec{s}_i \cdot \vec{s}_j \rangle$ , a sum of products of the individual electron spins, whose correlation with each other drops off sufficiently rapidly, so that  $\sum_{ij} \langle \vec{s}_i \cdot \vec{s}_j \rangle$  is of the same order as  $\sum_i \langle \vec{s}_i^2 \rangle$ , namely  $N$ .

But the formula for  $\langle \vec{S}^2 \rangle$  implies that each atomic cell has a mean square spin of order  $kT/\epsilon_f$  (if there is about one electron per cell). This number is small at low temperatures (at 1°K it is between  $10^{-5}$  and  $10^{-4}$  for most metals). The important point is that it is finite rather than an infinitesimal like  $1/N$ . Therefore the occurrence of a localized magnetic moment in a region of the order of one cell should be viewed as a question of degree rather than principle. As a suitable impurity is introduced into the cell and its potential  $V$  is imagined as being slowly turned on, the small thermal value of  $\langle s^2 \rangle$  in the cell should gradually increase as the result of electron-electron interaction  $v$ . Eventually, for sufficiently strong  $v$  and for sufficiently well chosen  $V$ , even the qualitative dependence of  $\langle s^2 \rangle$  on  $T$  may change, and the Pauli-like susceptibility of the cell will turn into a Curie law.

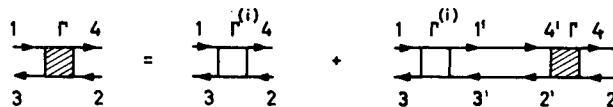


FIG. 1. The Bethe-Salpeter equation in the appropriate particle-hole channel.

Mathematically one may describe the susceptibility in terms of an exact two-particle Green's function, viewed along the appropriate particle-hole channel. Only the vertex part (not the factorable part) of that Green's function is of interest here. The vertex itself must be evaluated by perturbation theory. The formal summation of diagrams for the vertex leads to the Bethe-Salpeter equation which, in the relevant channel, has the form of Eq.(1) and is depicted in Fig. 1.

$$\Gamma(1234) = \Gamma^{(i)}(1234) + \Gamma^i(1'3'31') \mathcal{G}(1'4') \mathcal{G}(2'3') \Gamma(4'2'2'4) \quad (1)$$

The numerals stand for four-momentum and spin. Primed numerals are summed over. The energy-component of the four-momentum is of the form  $\omega_n = (2n + 1)\pi kT$ , where  $n$  is an integer, in accordance with the Matsubara finite temperature Green's function technique. The sum

of the energies of the two particles entering  $\Gamma$  equals that of the two outgoing particles; the same is not true for the three-momentum because of the presence of the impurity. For this reason the one-particle propagator  $\mathcal{G}$  depends on two three-momenta, though the fourth components are equal. Finally  $\Gamma^{(i)}$  is the vertex irreducible in the channel under consideration. In the absence of spin-orbit coupling, the Bethe-Salpeter Eq.(1) may be decomposed into singlet and triplet parts by writing

$$\Gamma(1234) = \Gamma_s \frac{1}{2} \delta_{13} \delta_{24} + \Gamma_t (\delta_{14} \delta_{23} - \frac{1}{2} \delta_{13} \delta_{24})$$

(where the subscripts of the Kronecker deltas denote spin orientation) and similarly for  $\Gamma^{(i)}$ .  $\Gamma_s$  and  $\Gamma_t$  now only depend on the orbital and energy variables, and both satisfy Eq.(1) with  $\Gamma^{(i)}$  replaced by  $\Gamma_s^{(i)}$  and  $\Gamma_t^{(i)}$  respectively. Only  $\Gamma_t$  is needed in the susceptibility calculation.

If bare propagators (uncorrected for electron-electron interaction) are used for  $\mathcal{G}$  in the Bethe-Salpeter equation, and if the irreducible vertex part is simply replaced by the bare electron-electron coupling, it is found that for sufficiently large coupling  $v$ , and for an impurity potential  $V$  causing a resonance sufficiently close to the Fermi level, the solution of Eq.(1) for  $\Gamma_t$  (a Fredholm integral equation) develops a pole at zero energy transfer  $\omega_3 - \omega_1 = v = 0$  and arbitrary three-momentum transfer. The condition for the appearance of this pole may be derived in an entirely equivalent way from the Hartree-Fock theory and was first found on that basis by Wolff. Essentially the same condition, also based on the Hartree-Fock theory, was given by Anderson using his well-known model. The appearance of the pole is taken to indicate the onset of a local magnetic moment. (It is the localized analogue of the onset of free-electron ferromagnetism in the uniform medium. In that case, for invariance reasons, the pole first appears at zero four-momentum transfer.) For coupling strengths beyond critical, the Hartree-Fock theory no longer permits calculating the susceptibility which becomes infinite at onset. Quite aside from this difficulty, it is hard to believe that a "small" system, like the immediate vicinity of an impurity, undergoes discontinuous changes as a function of the system parameters. One is therefore led to inquire what further minimal complications must be introduced in the calculation to permit smooth variation of the susceptibility everywhere in the parameter space.

Evidently it is necessary to use a more exact one-particle propagator in Eq.(1). A large vertex  $\Gamma$  is reflected in a large self-energy part  $M(12)$ . This, in turn, leads to a small  $\mathcal{G}(12)$ , through Dyson's equation. A small  $\mathcal{G}$  tends to reduce the size of the kernel of the integral Eq.(1). Writing the solution of Eq.(1) purely formally as  $\Gamma_t = v/(1 - v \mathcal{G} \mathcal{G})$  we see that a small  $\mathcal{G}$  tends to give a small  $\Gamma_t$  (the pole condition is, formally,  $v \mathcal{G} \mathcal{G} = 1$ ). We thus see the possibility of obtaining a self-regulating solution by simultaneously solving Dyson's Eq.(2) together with the Bethe-Salpeter Eq.(1)

$$\mathcal{G}^{-1} - \mathcal{G}_0^{-1} = - M \quad (2)$$

If we disregard the requirement of conservation of spin density and of antisymmetry of  $\Gamma$  with respect to the in- or out-variables<sup>3</sup> the self-energy contribution coming from  $\Gamma_t$  may be written

$$M(12) = \sum_{34} G(34) \Gamma(1342) \quad (3)$$

apart from counting factors. This result is obtained by closing the hole line on the extreme right of Fig. 1 on itself (Fig. 2). But aside from a small term, this is the same as Fig. 3, i.e. Eq.(3).

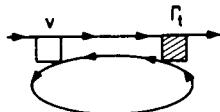


FIG. 2. The dominant contribution to the self-energy.

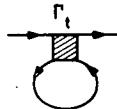


FIG. 3. The dominant contribution to the self-energy is also approximately equal to this diagram.

A complete solution of the non-linear coupled Eqs (1), (2) and (3) is of course very difficult. Further progress is possible, however, if it is assumed that the important variable is the energy rather than the three-momentum. If the orbital variables are simply averaged over,  $G$  and  $\Gamma$  respectively become functions of  $\omega_n = (2n + 1)\pi kT$  and  $v_n = 2n\pi kT$  (the energy transfer) only. The equations that must then be solved are

$$\Gamma(\nu) = -v / \{1 + (vkT) \sum_{\omega} G(\omega) G(\omega - \nu)\} \quad (4)$$

$$\frac{1}{G_0(\omega)} - \frac{1}{G(\omega)} = -\frac{3}{4} kT \sum_{\nu} \Gamma(\nu) G(\nu + \omega) \quad (5)$$

An approximate solution of these may be found by recalling that, if bare propagators are used,  $\Gamma_t(\nu = 0)$  is the first vertex to go to infinity. This suggests neglecting  $\Gamma_t(\nu \neq 0)$ . Then Eq.(5) is a quadratic for  $G(\omega)$ . Its solution in terms of  $\Gamma(0)$  may be substituted in (4), which, with  $\nu = 0$ , then gives an equation for the constant  $\Gamma(0)$ . From  $\Gamma(0)$  one may directly calculate the spin susceptibility, and one finds a smooth change-over from a Pauli-type to a Curie-type law as the parameters (for example,  $v$ ) are changed.

<sup>3</sup> A more elaborate theory can be constructed that takes these requirements into account. At this time we know only that restoration of spin conservation does not greatly change the results presented here.

However, this simple solution fails at low temperatures, where the spacing between successive  $\nu$ 's becomes small, and there is no justification for retaining  $\Gamma(0)$  only. A complete numerical analysis shows (for the case of an impurity potential causing a resonance at the Fermi level) that at a sufficiently small  $T$ , the susceptibility flattens off. The larger the  $\nu$ , the lower the temperature at which the deviation begins. Thus the system apparently tends to revert to a singlet state at sufficiently low temperatures but, of course, with a rather high Pauli susceptibility per impurity (Fig. 4).

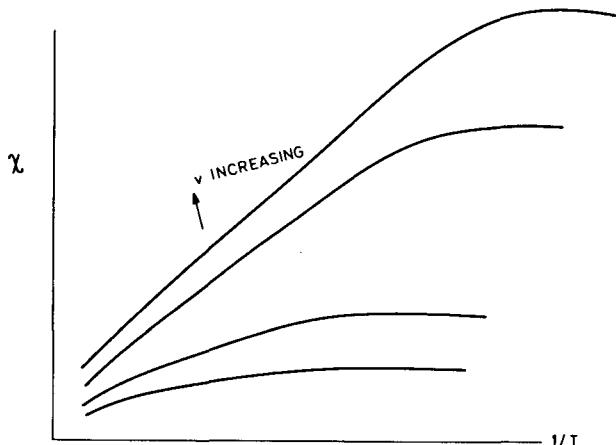


FIG. 4.  $\chi$  against  $1/T$  with increasing  $v$ .

It is also possible to calculate the resistivity. For this purpose one needs the value of the T-matrix at  $\omega = 0$ . It can be shown that this T-matrix is related to the mass operator by the analogue of the Lippmann-Schwinger equation

$$T(\omega) = V + M(\omega) + \{V + M(\omega)\} \mathcal{G}_0(\omega) T(\omega)$$

where  $\mathcal{G}_0(\omega)$  is the propagator for the pure medium. Thus, to find  $T(0)$  one needs  $M(0)$ . But  $M$  is defined only at the "odd" points along the imaginary axis. Therefore it is necessary either to transform Eqs (4) and (5) to integrals along the real axis, or else to analytically continue the numerical solution found in connection with the susceptibility, from the odd points on the imaginary axis to  $\omega = 0$ .

This latter procedure was adopted and resistivity curves were found much like those from the "definite spin" model, according to which a well-localized spin interacts with the conduction electron spin density. At high temperatures the resistivity is constant. As the temperature is lowered, the resistivity begins to rise, approximately linearly on a logarithmic temperature scale, and flattens off again at approximately the same temperature at which the susceptibility flattens off. The difference between the high and low temperature plateaux is the greater, the larger the  $v$ .

These results confirm in a general way what has frequently been inferred from the "definite spin" model: that some kind of singlet state forms at very low temperatures.

#### A C K N O W L E D G E M E N T S

The work described above is the joint effort of the author and Drs M. J. Levine, T. V. Ramakrishnan and R. A. Weiner.

# SURVEY OF THE ONE-DIMENSIONAL MANY-BODY PROBLEM AND TWO-DIMENSIONAL FERROELECTRIC MODELS

E.H. LIEB

Mathematics Department,

Massachusetts Institute of Technology,

Cambridge, Mass., United States of America

## Abstract

SURVEY OF THE ONE-DIMENSIONAL MANY-BODY PROBLEM AND TWO-DIMENSIONAL FERROELECTRIC MODELS. In the first part of the paper a survey is given of some of the exactly soluble many-particle problems in one-dimension, both classical and quantum mechanical. In the second part, the properties of exactly soluble two-dimensional ferroelectric problems are summarized.

## INTRODUCTION

I shall summarize what has been done on the one-dimensional many-body problem and I shall also mention the two-dimensional ferroelectric models which were solved last year [1]. The discussion will necessarily be very brief. For more details on the former topic I refer to the book written by D.C. Mattis and myself in 1965 [2].

## 1. ONE-DIMENSIONAL MODELS

I shall begin by citing some reasons for studying exactly soluble one-dimensional models. Perhaps the first reason, historically speaking, is their pedagogic value which, it is certain, will continue to be great just as it has been in the past. I can hardly believe, for instance, that there breathes a solid-state physicist whose subconscious reaction to band theory is not to think of the one-dimensional Kronig-Penney model of 1931. Likewise, more complicated and more recent models of many-body systems will, I am sure, ultimately find their way into the textbooks of the future.

One-dimensional models, however, also have a value to theoretical physics itself, as distinct from students of physics. On the one hand they provide a test of our mathematical tools and on the other they broaden our intuitive understanding of the ways in which interacting systems can be fundamentally different from non-interacting systems.

Admittedly, one-dimension is different from three. It is a pale reflection of the real world and asking what we can learn from studying it is a bit like asking what one can learn about life by reading King Lear or a history book or the Bible. The answer depends on how much prior knowledge of the subject and how much receptivity to nuance and suggestion we bring to the study. In other words, sophistication is a sine qua non rather than a hindrance to the study of one-dimensional systems.

In one-dimension we sacrifice some structure to obtain exact solutions to problems. We expect, and indeed find, that one-dimensional systems are simpler than their three-dimensional counterparts. Although the word pathology is frequently used, it is a bad one to describe this state of affairs, for three-dimensions is not so much different as it is more complex. Since one-dimension is simpler than three, it is important to ask the right questions, assuming that one is not dealing exclusively with critical phenomena.

It is probably fair to say that an approximation scheme or concept that cannot come to grips with the trivialities of one-dimensional life must find itself on the defensive. While it is possible in principle for an approximation scheme to have explicit topological arguments built into it, this is rarely the case, and we are, therefore, entitled to ask that the scheme acquit itself well in one-dimension.

It will perhaps be asked why it is so necessary to test mathematical concepts and approximations. The reason, it seems to me, is a profound one. The so-called many-body problem is in a curious state of development, almost unparalleled in the history of physics. It is a field in which the dynamics are completely understood whether they be quantum or classical. Nevertheless, there exists a chasm between the dynamical equations and their consequences. One has the comfortable feeling that Maxwell's equations, for example, completely tied up the subject of electrodynamics, at least in principle. The same may be said of Newton's equations or of the equations of the special theory of relativity, but the N-particle Schrödinger equation is distinguished by the fact that it ties up little except government funds. A direct, concentrated attack on the many-body problem is largely a post-war development in which the mathematics has been characterized by uncontrolled approximations based largely on intuition. While the intuition of great physicists is not to be despised, we cannot rely on it forever as an infallible guide to truth. We have about reached the time when the mathematical questions we want to ask can only be decided in an objective fashion, and a decent respect for the opinion of mathematicians requires that if we cannot give water-tight proofs, we at least sharpen our intuition as much as possible.

It is here that a catalogue of exactly soluble one-dimensional models could play a decisive role in broadening our intuitive perception. Most of our intuition of the real world is, it must be said, based on non-interacting systems. It is the ultimate aim of almost any textbook on the many-body problem to make real systems look as far as possible like a collection of non-interacting sub-systems, usually called phonons or elementary excitations. One chooses a non-interacting system as a base system for which all interactions are regarded as minor perturbations and by a Herculean effort one transforms this non-interacting system into another quasi-non-interacting system. It is surely conceivable that important truths might be lost this way.

A second complementary dimension of our intuition could very well be one-dimensional systems, for these need not suffer from forced analogies with non-interacting systems. In short, since our connection with the real world must be through tenuous threads tied to unreal simplistic worlds and since a fully interacting one-dimensional many-body system is not patently more unreal than a non-interacting three-dimensional

system, it can perhaps be useful as a second anchor for those threads extending from our minds to objective reality.

It might be worthwhile mentioning, at this point, one or two examples of the manner in which well-understood interacting systems can play an important role in our thinking. The first example might be the Kronig-Penney model mentioned before. To be sure this is not an example of an interacting system, but it is an example of band theory when there are non-trivial bands present. Can it be denied that one does not have a proper intuitive grasp of the significance and meaning of bands until one actually works out the example of the square well or delta function, periodic potential in one-dimension. To be sure, no very great surprises are manifested by the Kronig-Penney model, nor are they manifested by the other examples of soluble, one-dimensional periodic potentials that have been discovered since. The significance of this model is that it allows us a concrete visualization of how a band might look, and it is not rejected even though it is well known that three-dimensional bands in real substances are vastly more complicated than the Kronig-Penney bands.

A far more sophisticated, far more complicated, far more important model is the two-dimensional Ising model which, it should be remembered, is made soluble because of the fact that it can be reduced to a one-dimensional problem along one of the two lattice directions. At any rate, that is how Onsager first solved it. The Ising model, it will be recalled, parallels in a simple fashion two important physical systems: one is a lattice of interacting spins; the other is a classical gas of interacting particles. When the Ising model is thought of in the latter context it is referred to as the lattice gas. No one would suggest for a moment that a real gas, even a two-dimensional real gas, is the same as a lattice gas. Nevertheless, I need hardly remind you of the great theoretical interest that the Ising model has for statistical mechanicians and that this interest has grown with the passage of time rather than decreased. It may be fairly said that one of the important theoretical problems before us at the present time, one which has no intrinsic interest other than as a stepping stone to the understanding of real gases, is the following: How can we understand the phase transition of the Ising model, especially near the critical point, in a simple manner, that is to say, without completely solving the problem for a finite lattice and passing to the limit of an infinitely large system. The simple  $1/8$  power law discovered by Yang and Onsager for the spontaneous magnetization as a function of temperature, and which was derived only after a fantastic amount of analytic complexity, still defies what may be called an intuitive understanding. In short, the Ising model is one of the important corner-stones upon which our present poor understanding of real gases rests. By itself it is an academic exercise. But in conjunction with profound ideas it is extremely relevant. In like manner, other models of interacting systems could be of equal importance although it must be admitted that for most of them we shall not be so fortunate as to have them two-dimensional.

How much work has, in fact, been done on one-dimensional models of many-body problems? Mattis and I were surprised to discover a bibliography of approximately 230 papers (up to 1965), each one concerned with some exact statement about physically reasonable models in one-dimension. This does not count papers on field theory, which

comprise several dozen, nor does it count a large number of papers on the statistical theory of energy levels which surely is the one-dimensional theory par excellence, nor does it count a generous number of papers on the testing of approximation theories in one-dimension which would likely double the total number. Occasionally, especially during the war and shortly thereafter, one finds some duplication. This occurs, no doubt, because when the solutions to these models were discovered in moments of reverie, it was probably assumed that no one else would bother his mind with such useless trivia. As time went on, however, and the solution of more and more complicated one-dimensional problems became a real challenge, this duplication tended to be rare. Recently, however, in 1961 to be exact, there did occur the remarkable simultaneous discovery by Lenard and by Prager of the exact partition function for the one-dimensional Coulomb gas. Their discovery was announced simultaneously at the same meeting of the American Physical Society.

Aside from the occasional duplication, a survey of these many papers reveals a remarkable diversity of physical content and a great variety of mathematical ideas and I will turn now to a survey of some of them. Since it is obviously impossible to review everything, I must apologize to those whose work I am forced to omit.

The first broad category is in classical statistical mechanics. In 1936, Tonks was the first to publish, I won't say discover, the partition function of the hard core gas in one-dimension. Hard cores in one-dimension behave a bit like hard cores in three-dimensions at a density close to tight packing. That is to say, they simply give rise to an excluded volume. There is not a great deal to be learned from this model although the fact that all the correlation functions can be computed exactly, together with the fact that these correlation functions are non-trivial, makes the model a useful one for testing those approximation theories which are based on correlation functions. In addition, the model is historically important because it was probably the first example of the fact that the Mayer cluster series could diverge before one reached a phase transition. That is to say, the Mayer activity series for this gas has only a finite radius of convergence for the density, even though there is no phase transition at any value of the density or temperature other than zero.

In 1942, Takahashi discovered the partition function for a gas of particles with hard cores plus any finite potential, repulsive or attractive, of sufficiently short range that any one particle could interact only with its nearest neighbours. That is to say, the range of the entire potential was less than twice that of the hard core alone. Again no surprises occur, and no phase transition occurs, although by choosing the parameters of the potential wisely, one could achieve a set of isotherms that look remarkably like that of a real gas, i.e. isotherms which have a fairly flat, although not strictly flat portion.

In discussing classical statistical mechanics, it is possible to take the narrow view that there is nothing to be learned because of the well-known dictum that no phase transition can occur in one-dimension. There are several comments to be made about this. The first is that even though no phase transition can occur at finite temperature, zero temperature frequently has all the properties of a phase transition point. That is to

say, the thermodynamic functions can become singular at  $T = 0$ . A one-dimensional gas at temperature other than zero can look very much like a real gas above its transition temperature or boiling point. Likewise, a one-dimensional magnet at temperatures other than zero can look, in many respects, like a real magnet above its Curie point. The second comment is that a rigorous proof that no phase transition can occur in one-dimension has, strictly speaking, not been given, for the simple reason that it is impossible to define the theorem with sufficient precision. It has been proved by Van Hove that a one-dimensional gas with finite range forces can have no phase transition – and that is all.

If we cheat a little bit, however, it is possible to have a one-dimensional system with a phase transition. Consider a gas with the hard core potential mentioned above, together with an attractive potential, and ask what happens as the range of the potential is made longer and longer and the height of the potential is made smaller and smaller in such a manner that the total integrated value of the potential is kept constant. If one passes to the limit of an infinitely weak, infinitely long-range potential after passing to the bulk, or thermodynamic, limit of an infinitely large system, it is quite possible to obtain a phase transition. Historically, Kac evaluated the partition function of the hard core gas together with an exponential potential of arbitrary range. Baker realized that extending the range to infinity would result in a phase transition, which turned out to be of the Van Der Waals type. It is, indeed, curious that one obtains exactly the Van Der Waals type isotherm for a one-dimensional system. Subsequently, Lebowitz and Penrose were able to give a completely rigorous proof of what with a little hindsight is an obvious fact: namely, if one wishes to pass to the limit of an infinitely long-range, infinitely small attractive potential, then it should presumably not be necessary to go through the intermediate step of computing the partition function for potentials of finite range and then passing to the limit in tedious detail. In short, molecular field theory – which would have been naively applied by the average solid-state physicist unaware of the niceties of statistical mechanics – ought to give the right answer in the limit of an infinitely long-range potential. Lebowitz and Penrose proved that it is enough to know the partition function of a system with a finite-range potential, such as a hard core potential alone, in order to calculate in a very simple way what will happen if one adds on an attractive potential and takes the limit mentioned above. Their theorem is applicable to a wide variety of potentials (both the short-range part and the long-range part) and to any number of dimensions. For this reason, the Kac-Baker model probably does not belong, strictly speaking, in the one-dimensional category, even though it was first worked out there. It is a fact, nevertheless, that the only system with a finite-range potential for which one could work out the partition function exactly as a principal ingredient in the Lebowitz-Penrose theorem, was, in point of fact, the one-dimensional hard core gas or else the one-dimensional Takahashi gas. In that sense, the Kac-Baker model continues to be one-dimensional, but not because of any intrinsic property of the model.

One would be inclined to think that the same analysis should be applicable to quantum systems, say a Bose gas. While it is possible to compute the energy levels, wave functions and partition function of a hard

core Bose gas in one-dimension, it is not possible to calculate the partition function when a finite exponential attractive potential is added. Nevertheless, the Lebowitz-Penrose theorem encourages us to think that molecular field theory might be applicable to this system, too. That is to say, if we want to add an infinitely long-range, infinitely small potential, we could do so merely by knowing the partition function of the base system, and without the necessity of calculating the partition function for finite values of the potential, as was done originally by Kac for the classical gas. All that is required then, in order to have a model of a one-dimensional Bose gas with a first order phase transition, is a quantum-mechanical extension of the Lebowitz-Penrose theorem, and this I have been able to prove.

The third important model in classical statistical mechanics is the solution of the equilibrium Coulomb gas problem by Lenard and Prager separately, as was mentioned above. In one-dimension, unlike in three-dimensions, it is permissible to consider both the problem of one-component gas with a positive background and the two-component gas with a 50/50 mixture of positive and negative particles. In three-dimensions, it will be recalled, the latter system does not have a finite partition function unless quantum mechanics be invoked. It was precisely the latter system that Lenard and Prager solved. Baxter subsequently solved the former problem, that is, the one-component gas with background. When the two solutions are compared it turns out that there is no very important difference between them, as was to be expected. To my mind one of the most interesting results of this analysis is the following fact, whose significance I confess I do not fully comprehend. There are, as always, two regimes, weak coupling and strong coupling. For the former, the Debye-Hückel theory, which is a non-perturbative theory, performs well and correctly. That is to say, with a comparatively small amount of work, although with a great appeal to intuition, it gives the correct first term in an asymptotic expansion of any of the thermodynamic quantities in terms of the coupling constant. For strong coupling, on the other hand, there is no analogous theory – simple or otherwise. If the exact solutions are examined, however, it is found that to obtain the partition function it is necessary to calculate an eigenvalue of the Mathieu equation. For strong coupling, it is a very easy matter to obtain an asymptotic expansion of the eigenvalue in terms of the coupling constant. For weak coupling, however, it requires a mathematical tour-de-force to obtain even the first few terms of an asymptotic expansion in terms of the coupling constant. In short, the exact solution is simple for strong coupling and complicated for weak coupling, whereas perturbation theory (and the word is used here advisedly) is feasible for weak coupling and impossible for strong coupling. This curious feature, namely, that strong coupling is in some sense simpler than weak coupling, is met with again in the exact solution of a Bose gas system which will be mentioned later.

Another important topic which has had much attention in recent years is the subject of the disordered chain of harmonic oscillators. The problem is to discover the average spectrum of a chain of oscillators whose masses are randomly distributed. Such a system is supposed to have some relevance to the theory of liquid metals and to the theory of semi-amorphous solids such as glass. Be that as it may, in 1953, Dyson discovered an equation which, could it be solved, would give the required

average spectrum. The equation is a curious difference equation which is probably the most non-linear ever to appear in mathematical physics, and its solution has escaped us to the present day. Schmidt found a similar, but slightly simpler, equation and the solution to that, too, has not been forthcoming. It is possible to show that for certain values of the parameters, the solution to Dyson's equation is a function that has zero derivative except for a set of measure zero and has an infinite number of points where the derivative does not exist and these points have an infinite number of points of accumulation which in turn have an infinite number of points of accumulation and so forth. A very curious function indeed!

There has been a considerable amount of work on trying to find the spectrum by numerical methods both in one- and two-dimensions. There have also been some special theorems developed that apply strictly to the one-dimensional system. These theorems, which were discovered by Matsuda, Borland and others, are really all one common theorem which says that there are so-called special frequencies for which one can, in fact, compute the spectrum exactly without having to solve Dyson's or Schmidt's equation.

These theorems apply equally well to the physically different, but mathematically similar, problem of electron energy bands in disordered crystals. It is this subject which is directly related to the theory of liquid metals. For both disordered oscillators and disordered periodic potentials there is another curious theorem which has been virtually proved, although in a non-rigorous manner, by Borland. The theorem states that, whereas for an ordered system all the modes are non-localized and have the character of periodic functions extending throughout the whole lattice, for the disordered system all the modes are localized. No attempt has been made to find out exactly how localized "localized" means, but it does seem to be the case that for an infinite crystal any mode will ultimately die out exponentially fast, both to the left and the right. It is believed that a similar phenomenon can occur for three-dimensional structure although there it is generally agreed that only some, and not all, of the modes are localized. The situation is vaguely reminiscent of the fact that one light mass in an otherwise perfect crystal will always cause one mode to pop out of the continuum in one-dimension, whereas it does not always do so in three-dimensions.

Having discussed the classical many-body problem and the quantum mechanical one-body problem, we must say a few words about the quantum mechanical many-body problem. With all the current interest in liquid helium and Bose gases it would certainly be desirable to have a model, even a one-dimensional one, of an interacting Bose gas that could be completely analysed. Such a model was introduced in 1963 by Liniger and myself. The problem of one-dimensional bosons interacting via a delta function potential can be solved exactly in the sense that one can find all the wave functions and energy levels. The wave functions are somewhat complicated and it has so far been impossible to compute correlation functions from these wave functions, because to do so involves an N-fold integration which is virtually as complicated as the N-fold integrations which appear in classical statistical mechanics. Nevertheless, the spectrum can be elucidated and it is found that if the coupling constant is not too large then Bogolyubov's theory gives an excellent

account of the ground-state energy as a function of the coupling constant, and also gives an excellent account of at least one part of the excitation spectrum. As far as the ground-state energy is concerned, however, we found a phenomenon similar to that discussed above in connection with the Coulomb gas in one-dimension, namely, that the exact solution is very easy to find in the case of strong coupling and very difficult to find in the case of weak coupling. There is, apparently, an essential singularity at zero coupling constant, and the nature of that singularity has so far eluded us. Bogolyubov's well-known theory, on the other hand, while completely incapable of handling the strong coupling situation, very nicely gives the first two terms in an asymptotic expansion of the energy in terms of the coupling constant for weak coupling. It does this in terms of a very simple integral and the whole calculation can be done in a few lines. But why it works is difficult to understand. The Bogolyubov integral is arrived at by an uncontrolled approximation. We have not been able to derive that integral from the exact answer and, secondly, one of the principal ingredients in the Bogolyubov analysis is the concept of condensation into the zero momentum mode. It is not known whether this condensation ever occurs, in fact, in any number of dimensions, although there is very good reason for believing that it does not occur in one-dimension. At any rate, it is known that for the present model under discussion, there is certainly no condensation for infinite coupling constant, that is to say, for the case of hard core bosons in one-dimension. This was shown by Schultz. The situation, therefore, can be summarized this way: it is very likely, although not certain, that one of the principal hypotheses of Bogolyubov's calculation is not satisfied for the present model of bosons interacting with a repulsive delta function potential. Nevertheless, the Bogolyubov ansatz, together with a small turn of the crank, gives the right answer, and this right answer is very difficult to obtain by an exact calculation.

The situation is similar for the excitation spectrum, although here Bogolyubov's theory does not acquit itself quite so well. It seems to be the case that if one looks at the exact spectrum and tries to regard it as a collection of non-interacting phonons, insofar as this is possible, it is found that the simplest way to describe the spectrum is in terms of two kinds of phonons. It must be admitted, however, that one could describe the spectrum in terms of only one kind of phonon if one wished, but one would then have the anomalous situation of having certain low-lying states, which are obviously quite important states, regarded as an infinite number of phonons of vanishing small momentum. In order to overcome this intolerable situation we found it useful, and also natural, to introduce two kinds of phonons. This result, or conclusion, if it is justified, is quite possibly one of the most important lessons to be learned from one-dimensional models. For it seems to say, that an interacting system is in some sense basically different from a non-interacting one and, as we already know from the theory of superconductivity, certain features can appear in the spectrum of an interacting system which have no counterpart for a non-interacting system. The new spectrum cannot be transformed into the old one simply by switching off the coupling constant, because there is apparently an analytic singularity at zero coupling constant.

The real test of whether the double spectrum has any meaning would be to try to compute the neutron scattering from the delta function gas.

This involves calculating a correlation function but, as was mentioned above, this has so far proved to be impossible to do. One can, however, state what happens for infinite coupling constants, that is, for the hard core gas. In this case the correlation function required for neutron scattering can be calculated exactly and easily, and it is found that while the neutron scattering function calculated from the correlation function looks quite reasonable from the physical point of view, the correlation function itself, regarded as a function in the complex frequency plane looks quite different from what is normally supposed to be the topography of such functions. That is to say, it is normally supposed that the correlation function in the complex plane has a cut on the positive real axis and that an analytic continuation onto the second Riemann sheet yields a simple pole lying near the real axis. The real part of the pole is taken to be the frequency of a quasi-particle and the imaginary part is taken to be the reciprocal lifetime of the quasi-particle. The function that one can, in fact, calculate for the hard core Bose gas in one-dimension does not have this property. What it has is a finite, rather than an infinitely long, cut on the real axis and there are no poles or any other singularities on the second Riemann sheet. The two branch points associated with this finite cut turn out to be exactly the two spectra mentioned above and we are therefore entitled to assume, at least tentatively, that these two spectra have some relevance and are not simply ad hoc inventions.

In any event, it is clear that effort spent in calculating the one-particle and two-particle correlation functions for this gas when the coupling constant is finite would, very likely, be well repaid. As I said above, the wave functions are completely known and calculating the correlation function is nothing more, indeed it is nothing less, than a difficult exercise in combinatorial analysis. I very much hope to see it carried through one day.

What about interacting fermions? Here there are two models available. One is the delta function gas mentioned above, but it is far more difficult to solve for fermions than for bosons, for the two spin states per particle mean that one is effectively dealing with a two-component gas. In 1965 Flicker and I solved the problem when two spins are down and N-2 are up and the algebraic complexity was horrifying. Within the last year, however, Gaudin was able to solve the general case and his solution must be counted as one of the most beautiful examples of mathematical elegance in the physics literature. At the moment the problem has been reduced to solving a rather non-linear set of algebraic equations. Until these are untangled we are unable to describe the spectrum in detail.

There also exists a discrete version of the model, namely the short range, one band model of electrons hopping between nearest neighbour Wannier states in a crystal. Very recently, Wu and I solved this model by adapting Gaudin's solution and we found that there is no Mott conductor-insulator transition in the ground-state as the strength of the delta function repulsion is increased. For a band less than half full (i.e. when the number of electrons is less than the number of atomic sites) the system is always a conductor. But when the band is exactly half filled the system is always an insulator, except for the special case of zero repulsion when it is a conductor. Here again we see that an arbitrarily small interaction can fundamentally alter a system's properties.

The second interacting fermion problem that can be solved generally is the so-called Luttinger model. It is an adaptation of the field theoretic Thirring model to the problem of the N-electron Schrödinger equation. For good or bad one introduces the basic assumption that the kinetic energy is linear in the momentum, rather than quadratic. Thus, thinking in terms of perturbation theory, denominators which conserve momentum automatically conserve energy. Thus, the perturbation series can be summed to all orders. One does not, in fact, make use of perturbation theory; one solves the problem directly. The solution given by Luttinger in his original paper was not correct, however, and Mattis and I, in the process of writing our book, realized that something was amiss. We then were able to solve the problem exactly and found that the spectrum of interacting fermions in the Luttinger model had in it plasma modes which, in itself, is not so surprising. But these plasmons were exact in the sense that they had infinite lifetime and did not interact with the fermion modes. Moreover, and perhaps even more surprising, is the fact, which was subsequently pointed out by Overhauser, that the plasma modes exhaust the entire spectrum. There are, in fact, no fermion modes left. Another interesting property of this model is that the Fermi surface disappears for sufficiently strong interaction strength.

Amusing as it undoubtedly is, the Luttinger model is probably not completely satisfactory because of the stringent assumption made about the kinetic energy. The delta function model, when it is completely solved, will probably have much more meaningful things to say.

Among the important quantities to be calculated for a system of interacting electrons are the magnetic properties. In this context, I should like to mention a theorem proved by Mattis and myself. It is a very general theorem and is not restricted to any particular model although it is restricted to one-dimension. It will be recalled that the general theory of ferromagnetism is predicted on the assumption that ordinary forces together with the Pauli principle can cause the spins of electrons to line up parallel to each other. That is to say, it is not necessary to invoke spin-spin forces or spin-orbit forces in order to account for ferromagnetism. We have proved the following very general theorem: Given a system of N-electrons in one-dimension, subject to any completely arbitrary potential, including interparticle potentials and fixed single-particle potentials, it is always true that the ground-state of the entire system has zero total spin, and that the ground-state of the entire spectrum exclusive of the zero spin states belongs to  $s=1$ , and so forth. In one-dimension, it is very easy to compute exchange integrals which would, if one believed in them, indicate that the ground-state could be ferromagnetic. What the theorem says is that exchange integrals can be a very misleading guide to the magnetic properties of an electron system. This theorem, therefore, illustrates one of the very important uses of considering one-dimensional many-body systems. It serves, so to speak, as a guide to our consciences, for it tells us that if the conventional theory of ferromagnetism is to be correct, any calculation predicting ferromagnetism must, perforce, have built into it some three-dimensional topology. It is not enough merely to compute integrals. One must show in an explicit way how three-dimensions comes into the calculation, either through the degeneracy of three-dimensional atomic functions or in some other manner. It should be pointed out, by the way, that we are not here

concerned with the question of ferromagnetism in the sense of a thermodynamic phase transition. We are concerned only with the much simpler question of the tendency of the system. That is to say, does the system prefer to have its spins lined up or disaligned. In other words, if one believes that a Heisenberg type model could represent the magnetic properties of a system, is the sign of the so-called exchange constant to be taken positive or negative?

Speaking of Heisenberg models, it will be recalled that in 1931 Bethe solved the problem of the one-dimensional Heisenberg chain. At least he did so in principle. Since that time, a tremendous amount of work has been devoted to the model, but all we know at the present time is the ground-state energy and wave function, and the low-lying excited states. We do not yet know how to calculate the partition function of the system, even though the partition function of the corresponding Ising model can be calculated in two lines. In recent years, however, certain modifications have been made to the model which are interesting. On the one hand, the problem has been analysed for the anisotropic case and, on the other hand, it has been analysed for the ground-state and low-lying states when the total spin is not taken to be either zero or  $\frac{1}{2}n$ , but is anything in between.

It is also known that the ground-state energy and wave function of the Heisenberg chain, when considered as a function of the anisotropy parameter, has two singularities. The significance of these, apart from their intrinsic physical significance, is that they illustrate the fact that transitions can occur in one-dimensional quantum systems if one studies the zero temperature limit. The Heisenberg chain is also related, curiously enough, to a two-dimensional classical statistical mechanics problem which we shall turn to next.

## 2. FERROELECTRIC MODELS

In 1944 Onsager solved the two-dimensional Ising model, and since then the model has been studied extensively because it was the only model with short range forces that could be solved exactly and had a phase transition. Onsager's approach was to use the transfer matrix which describes the manner in which two neighbouring rows of the lattice interact. It is a matrix whose indices label states, or configurations, of a row and whose value is the Boltzmann factor appropriate to the two states on the two successive rows. The problem of computing the partition function reduces to the problem of finding the largest eigenvalue of the transfer matrix.

A single row can be thought of as a one-dimensional system and the transfer matrix is like a Hamiltonian, for it operates on some state to produce other states. By a certain amount of judicious juggling, the transfer matrix for the Ising problem can be represented as an exponential of a quadratic form in fermion operators and so can be diagonalized by elementary methods.

The two-dimensional ferroelectric models we shall now describe were solved last year [1] and represent the second class of soluble statistical mechanical models with nearest neighbour interactions and with phase transitions. They, too, are solved by the transfer matrix method, but here the transfer matrix is quite a bit more complicated than a quadratic

form. In fact, the transfer matrix turns out to be related to the Heisenberg chain problem mentioned above.

The most elementary of these problems is the ice problem, due to Pauling. At zero temperature, ice has an entropy (called the residual entropy) caused by the fact that each hydrogen atom can be in one of two positions between each pair of oxygen atoms. Picture a square lattice whose vertices represent oxygen atoms. On each bond we place one hydrogen atom one third of the way from either of the two oxygens. If there are  $N$  oxygens there will be  $2N$  hydrogens and the entropy would be

$$S = k \ln 2^{2N} = Nk \ln 4$$

This is too large experimentally; instead of  $\ln 4$  we should have  $\ln(1.5)$ . Pauling then proposed the so-called "ice condition": The only allowed configurations are those for which each oxygen is surrounded by two hydrogens close to it and two hydrogens far from it.

To visualize the constraint more clearly we can represent a configuration of hydrogen atoms by arrows drawn on the bonds of the lattice, the direction of each arrow signifying the position of the hydrogen relative to the bond midpoint. The problem is then to count the number of ways of drawing arrows on the bonds so that precisely two arrows point into each vertex.

This problem can be solved by the transfer matrix method and the result, for large  $N$ , is

$$S = Nk \ln (4/3)^{3/2}$$

which agrees very well with experiment since  $(4/3)^{3/2} = 1.5396$ .

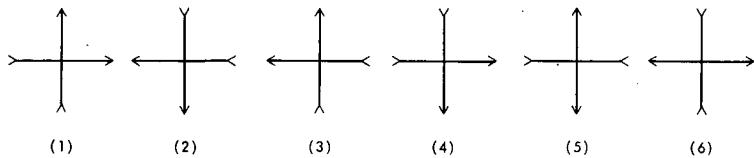
The ferroelectric models are obtained by modifying the basic ice problem as follows: At each vertex, six arrow configurations are allowed by the ice condition. These are shown in Fig. 1. To each of these we assign an energy and hence, by simple addition, we assign a total energy to an entire configuration. The partition function is then

$$Z = \sum_{\text{allowed configurations}} \left[ -\frac{1}{kT} \text{ total energy} \right]$$

Two "classical" choices are also shown in Fig. 1. The first is the F model of Rys which favours those vertices having zero polarization and hence is a model of an antiferroelectric. (Note that  $\epsilon > 0$ .) The other is the KDP (potassium dihydrogen phosphate) model of Slater. It favours two configurations which have equal and opposite polarizations and hence is a model of a ferroelectric. Finally, if we add an external electric field,  $E$ , in the vertical direction there is an additional energy because of the various polarizations of the six vertices. The problem can still be solved when this external field is present and we recall that the Ising model cannot be solved with an external magnetic field.

For both models we expect the following:

At high temperatures there will be disorder, but at low temperatures an



	<u>Vertex Energies</u>					
Ice Model	○	○	○	○	○	○
F Model	€	€	€	€	○	○
KDP Model	○	○	€	€	€	€
External electric field	+E	-E	+E	-E	○	○

FIG. 1. The six vertex configurations allowed by the "ice condition" are shown, together with their associated energies appropriate to the three models. Note that  $\epsilon > 0$ . The last row shows the energy to be added when an external electric field in the vertical direction is present. The arrows indicate the position of a hydrogen atom relative to the bond midpoint.

ordered phase will predominate. Furthermore, the onset of order should not be smooth but should happen abruptly at some definite temperature,  $T_c$ . Why this should be so is far from obvious, and it is the central problem of the theory of phase transitions.

Indeed both models have a critical temperature and they are identical for zero electric field

$$\epsilon = k T_c \ln 2$$

Beyond that the two models are completely different and we shall list a few of the more important properties:

KDP: 1. With zero electric field there is a first order phase transition with a latent heat and an infinite specific heat (for  $T > T_c$ ) which behaves like  $(T - T_c)^{-\frac{1}{2}}$ .

2. With an electric field the phase transition becomes second order (i. e. the latent heat disappears). Moreover,  $T_c$  increases with E.

3. For  $T > T_c$ , the polarization increases with E in the usual S-shaped manner except that it saturates at unity for a finite value of E. For  $T < T_c$  the polarization is unity for all values of E. Needless to say, this behaviour is quite different from the Ising model.

F: 1. With zero electric field there is an infinite order phase transition. This means that every derivative of the free energy with respect to temperature is bounded and continuous at  $T_c$ . Nevertheless, a power series expansion about  $T_c$  has a zero radius of convergence. Furthermore, there is a natural boundary (i. e. a cut without a second Riemann sheet) in the T plane.

2. With an electric field the transition becomes second order. For  $T > T_c$  the specific heat behaves like  $(T - T_c)^{-\frac{1}{2}}$  while for  $T < T_c$  the specific heat is unchanged and hence smooth. Moreover,  $T_c$  decreases with  $E$  and vanishes for a finite value of  $E$ .
3. For  $T > T_c$  the polarization increases with  $E$  in the usual S-shaped manner and saturates only for infinite  $E$ . For  $T < T_c$ , the polarization is identically zero until  $E$  exceeds a critical value (which depends upon  $T$ ) and then begins to rise with an infinite initial slope. The Ising antiferromagnet does not have this property. For non-zero  $H$  the polarization may be small, but it does not vanish identically. Furthermore, this behaviour appears to be in paradoxical contradiction to No. 1 above, because if we vary the temperature we seem to see an almost non-existent phase transition, while if we vary the electric field we see that the system is really quite "locked in" below  $T_c$ .

#### R E F E R E N C E S

NOTE: I apologize to all authors whose works should properly have been referenced in this paper, and hope that referring the reader to the more careful bibliography [2] will be a satisfactory substitute.

- [1] LIEB, E.H., Phys. Rev. Lett. 18 (1967) 692, 1046; 19 (1967) 108; Phys. Rev. 162 (1967) 162.
- [2] LIEB, E.H., MATTIS, D.C., Mathematical Physics in One-Dimension - Exactly Soluble Models of Interacting Particles, Academic Press, Inc., New York (1966).

# THREE EXAMPLES OF ONE-DIMENSIONAL SYSTEMS\*

E. W. MONTROLL

Department of Physics and Astronomy,  
University of Rochester,  
Rochester, N. Y., United States of America

## Abstract

THREE EXAMPLES OF ONE-DIMENSIONAL SYSTEMS. The following one-dimensional problems are discussed: (a) the vibrations of arrays of coupled springs and masses composed of two atomic species of different masses randomly distributed; (b) thermodynamics of an atmosphere in a gravitational field; and (c) theory of motion of a line of vehicular traffic. The discussions of topics (a) and (c) are reviews. Topic (b) concerns a one-dimensional gas with nearest neighbour intermolecular forces under the influence of the gravitational field. The main new result is that, at any altitude, the relation between the pressure and density is the same as it would be in a container of gas in the absence of an external field.

This lecture is concerned with three different types of one-dimensional systems. In each case it is exceedingly difficult to derive clear two- or three-dimensional generalizations. However, as will be seen, one does get some physical insight into the problems by considering one-dimensional systems.

The three topics presented are: (a) the vibrations of arrays of coupled springs and masses composed of two atomic species of different masses randomly distributed; (b) thermodynamics of an atmosphere in a gravitational field; and, (c) theory of motion of a line of vehicular traffic. The discussion of topics (a) and (c) are brief reviews of old subjects; that of topic (b) is new.

## 1. FREQUENCY SPECTRUM OF RANDOM LATTICE [1-6]

Consider a one-dimensional chain of atoms connected by springs. Let us suppose that the chain is composed of two atomic species of different mass according to a pre-assigned sequence but, for simplicity, suppose that the spring which couples neighbouring masses has a force constant  $\gamma$  which is independent of the masses which are connected. A surprise which occurs in the analysis of this problem comes from the fact that the frequency spectrum of the normal vibrations of the chain fluctuates very violently when plotted as a function of frequency in the case that the masses are arranged randomly.

It is easy to show that when the masses alternate on the chain ( $m$  being the light mass and  $M$  the heavy one) with equal number of light and heavy masses (in the limit as the number of lattice points becomes infinite), the normal mode frequencies are

$$\frac{\omega_{\pm}^2(\theta)}{\gamma} = \left( \frac{1}{m} + \frac{1}{M} \right) \pm \left\{ \left( \frac{1}{m} + \frac{1}{M} \right)^2 - \frac{4 \sin^2 \theta}{mM} \right\}^{\frac{1}{2}} \quad (1a)$$

---

\* This work was partially supported by the Air Force Office of Scientific Research Grant No. AF OSR 1314-67.

with

$$0 < \theta < \pi/2 \quad (1b)$$

The + branch of frequencies is the usual high frequency optical branch while the - branch is the low frequency acoustical branch. These branches have the following restrictions

$$\frac{2\gamma}{m} < \omega_+^2 < 2\gamma \left( \frac{1}{m} + \frac{1}{M} \right) \quad (2a)$$

$$0 < \omega_-^2 < \frac{2\gamma}{M} \quad (2b)$$

The frequency distribution function  $g(\omega^2)$  which is defined so that  $g(\omega^2) d\omega^2$  is the fraction of normal mode frequencies between  $\omega^2$  and  $\omega^2 + d\omega^2$  is plotted in Fig. 1. In the limit as the heavy mass becomes infinite, the two branches become very narrow with the acoustic branch shrinking to a point and corresponding to the very low frequency vibrations of the very heavy masses which drag all light masses along with them. The high frequency optical spectrum is shrunk to the single frequency which would correspond to that of the light mass vibrating between two rigid walls.

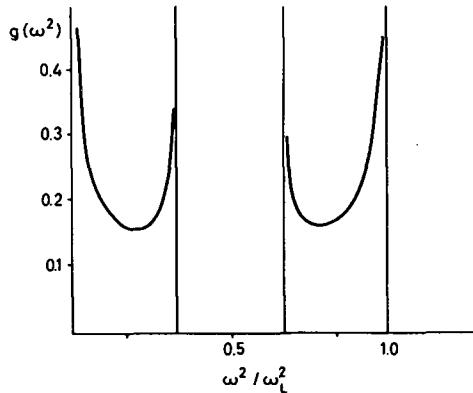


FIG. 1. Frequency spectrum of chain of alternating masses with mass ratio two to one.  $g(\omega^2)$  is given in arbitrary units.

Now let us consider the general case in which the heavy atoms are not placed periodically, but appear at lattice points  $h_1, h_2, h_3, \dots$  so that the number of light atoms between the successive pairs of heavy ones are respectively  $s_1 = h_2 - h_1 - 1, s_2 = h_3 - h_2 - 1$ , etc. We might also introduce a function  $G(n)$  which is the number of runs of  $n$  light atoms which appear in the chain. In the alternating light and heavy atom case discussed above  $G(1) = N/2$  ( $N$  being the total number of atoms in the chain), and  $G(n) = 0$  for other integers,  $n$ .

To get a qualitative feeling for the effect of randomizing the chain, we examine the limit in which the mass of the heavy atoms becomes infinite.

Then the various sequences of light atoms are decoupled from each other so that each sequence of light atoms behaves as though it were connected to rigid walls at each end of the chain. The normal mode frequencies in that case are known to be

$$\omega = \omega_L \sin[\pi k/(n+1)] \quad k = 1, 2, \dots, n \quad (3)$$

for a chain of  $n$  light atoms with  $\omega_L^2 = 4\gamma/m$ .

Every normal mode frequency of the set (3) can be associated with a couple  $(k, n)$ . Many of these frequencies have the same numerical value even though they correspond to different couples. For example, the couples

$$(k, n) = (1, 1), (2, 3), (3, 5) \dots, (\ell, 2\ell-1), \dots \quad (4a)$$

all correspond to the frequency

$$\omega = \omega_L \sin \pi/4 \quad (4b)$$

The weight, or number of times this frequency occurs when  $n = 1$  is  $G(1)$ , when  $n = 3$  is  $G(3)$ , etc., so that the total number of times this frequency appears is

$$W(1/2) = G(1) + G(3) + G(5) + \dots = \sum_{\ell=1}^{\infty} G(2\ell - 1) \quad (4c)$$

The frequencies

$$\omega = \begin{cases} \omega_L \sin \frac{\pi}{2} \left(\frac{1}{3}\right) \\ \omega_L \sin \frac{\pi}{2} \left(\frac{2}{3}\right) \end{cases} \quad (5a)$$

appear respectively from the sets

$$(k, n) = \begin{cases} (1, 2), (2, 5), (3, 8), \dots, (\ell, 3\ell-1), \dots \\ (2, 2), (4, 5), (6, 8), \dots, (2\ell, 3\ell-1), \dots \end{cases} \quad (5b)$$

Clearly the number of each type depends on the denominator,  $n$ . Hence, the weight of each of the two frequencies (5b) is

$$W\left(\frac{1}{3}\right) = W\left(\frac{2}{3}\right) = \sum_{\ell=1}^{\infty} G(3\ell - 1) \quad (5c)$$

In a similar manner one sees that for every ratio

$$k/(n+1) = p/q \quad (6a)$$

where  $p/q$  is reduced to lowest terms, there is a frequency

$$\omega = \omega_L \sin \frac{\pi}{2} \left( \frac{p}{q} \right) \quad (6b)$$

which appears from the couples

$$(p, q-1), (2p, 2q-1), (3p, 3q-1), \dots \quad (6c)$$

and, therefore, with the weight

$$W(p/q) = \sum_{\ell=1}^{\infty} G(\ell q-1) \quad (6d)$$

Every infinite mass corresponds to a zero frequency mode. Hence, if  $c$  is the concentration of light atoms,  $(1-c)N$  is the number of zero frequency modes.

Now let us consider several periodic cases. If A represents an atom of mass  $m$ , and B one of mass  $M$ , then the alternating sequence discussed at the beginning of this section is ABABAB..., so that  $G(1) = N/2$  while  $G(n) = 0$  if  $n = N/2$ . Then in the infinite mass limit  $N/2$  frequencies are zero and the only possible value for  $p/q$  in Eq. (6b) is  $1/2$ , so the weight  $W(1/2) = G(1) = N/2$  is appropriate for the frequency  $\omega = 2\sqrt{\gamma/m} \sin \pi/4 = \sqrt{2\gamma/m}$  in agreement with Eq. (2).

The case AABBAABB... has one non-vanishing  $G(n)$  value,  $G(2) = N/4$ . There are, therefore,  $N/2$  frequencies equal to zero, and  $N/4$  equal to  $\omega = 2\sqrt{\gamma/m} \sin \pi/6 = \sqrt{\gamma/m}$ , and  $N/4$  equal to  $\omega = 2\sqrt{\gamma/m} \sin \pi/3 = \sqrt{3\gamma/m}$ . If the  $M = \infty$  condition is relaxed in these periodic cases, each frequency given above becomes a band and each band has the typical shape of the frequency distribution function of a monatomic 1D lattice.

The case of a random distribution is quite different. We will find that the spectrum is everywhere dense and everywhere discontinuous in the frequency range between 0 and  $\omega_L = (4\gamma/m)^{1/2}$ . The number of runs of A's of length  $n$  in a random sequence is  $Nc^n(1-c)^2$ , where  $c$  is the concentration of light atoms. Hence  $G(n) = Nc^n(1-c)^2$  and

$$W(p/q) = \sum_{\ell=1}^{\infty} NC^{\ell q-1}(1-c)^2 = N(1-c)^2 c^{q-1} / (1-c^q) \quad (7)$$

To see that the spectrum is everywhere dense in the interval  $(0, \omega_L)$ , we note that for any possible frequency  $\omega$

$$\left( \frac{p}{q} \right) = \frac{2}{\pi} \sin^{-1} (\omega/\omega_L) \quad (8)$$

As  $\omega/\omega_L$  ranges from 0 to 1, the right-hand side of Eq. (8) takes on all values between 0 and 1. However, since any number between zero and one can be approximated to any required accuracy by a rational number, we can find a fraction  $(p/q)$  such that the associated frequency is as close as we wish to any number in the range  $(0, \omega_L)$  as we set out to show.

The everywhere discontinuous character of the spectrum follows by comparing the weight  $W(p/q)$  to that at a neighbouring  $p'/q'$ ,  $W(p'/q')$  as  $p'/q' \rightarrow p/q$ . As a fraction reduced to lowest terms approaches  $p/q$  (without equalling  $p/q$ ), its denominator  $q'$  gets larger. For example, consider the approximations to  $1/2$ ,  $n/(2n+1)$  as  $n \rightarrow \infty$ . Since

$$W(p'/q')/W(p/q) \approx c^{q'-q} \quad (9)$$

we see that as  $(q'-q) \rightarrow \infty$  as is required for  $p'/q' \rightarrow p/q$ , the ratio of the weights approaches zero. Since this is true for every  $p/q$ , the spectrum is everywhere discontinuous.

When the  $M = \infty$  condition is relaxed, we would still expect the spectrum to oscillate wildly even if it becomes continuous. This is borne out by machine calculations by Dean which are exhibited in Fig. 2. A systematic way of relaxing the  $M = \infty$  condition is to successively consider the variation in the spectrum which develops when a sequence of light atoms is broken into two parts by a heavy atom (now with a finite mass) but with chain ends attached to infinite mass particles, and then to consider a sequence of light atoms broken into three sequences by two heavy atoms, etc. This, essentially, gives a perturbation theory in the main ratios. It will be discussed elsewhere.

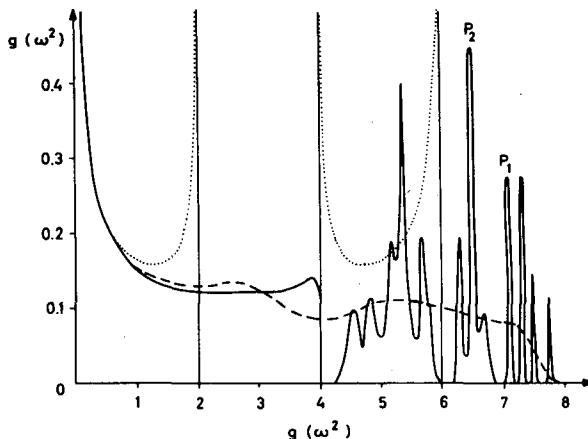


FIG. 2. The spectrum of squared frequencies of a randomly disordered one-dimensional lattice for which the mass ratio is 2 and the two constituents are present in equal amounts; — result of a machine calculation by Dean [2] for a chain of 32 000 particles; --- the 20-moment approximation to the spectrum by Domb et al. [1], the spectrum of an alternating (ordered) diatomic linear chain.

The 17 non-vanishing frequencies of greatest weight in a random chain of 50% light atoms and 50% infinite mass atoms are as follows:

<u>p/q</u>	<u><math>N^{-1}W(p/q)</math></u>	<u><math>p/q</math></u>	<u><math>N^{-1}W(p/q)</math></u>
1/2	1/6	1/5, 2/5, 3/5, 4/5	1/62
1/3, 2/3	1/14	1/6, 5/6	1/126
1/4, 3/4	1/30	1/7, 2/7, 3/7, ..., 6/7	1/254

2. STATISTICAL MECHANICS OF A ONE-DIMENSIONAL "ATMOSPHERE"  
IN A UNIFORM GRAVITATIONAL FIELD

2.1. Introduction

The general formalism of statistical mechanics applies most directly to uniform systems in the thermodynamic limit (i.e. number of particles  $N \rightarrow \infty$  and container volume  $V \rightarrow \infty$  in a manner such that  $N/V = \rho$  has a prescribed value). On the other hand, certain applications involve open systems in which such limits are not appropriate. An example of such a case is a planetary atmosphere in which gravitational forces keep the atmosphere close to the planet, a high density existing close to the surface and the density diminishing with height. If the base area under the atmosphere is finite, the approach to the thermodynamic limit would imply the development of an atmosphere of infinite weight above any element of area. To see what is involved in this type of problem, we consider a one-dimensional "tube" of particles which have strong short range repulsions and weaker longer range attractive forces between molecules so that the particles cannot pass through each other. We limit ourselves to a model in which only adjacent pairs interact with each other. An expression will be found for the density at the level of the  $j$ -th particle in the tube and for the pressure at the same point. It will be shown that these two quantities are related by the same equation of state that one calculates by the standard methods of statistical mechanics. An expression is also found for the average level at which the  $j$ -th particle can be found and for the fluctuations in separation distance between adjacent pairs of particles.

2.2. General formulation of problem

Consider a "chain" of  $N$  particles of equal mass, in a constant gravitational field. Let  $\phi(r_{j-1} - r_j)$  be the potential energy of interaction between two adjacent particles ( $j-1$ ) and  $j$  when their distance of separation is  $(r_j - r_{j-1})$ . We number the lowest particle on the chain 1, the next 2, etc. If we postulate interactions to exist only between neighbouring particles, then the total potential energy of interaction of the assembly of particles is

$$V(r_1, r_2, r_3, \dots, r_N) = \phi_0(r_1) + \sum_{j=2}^N \phi(r_j - r_{j-1}) + mg(r_1 + r_2 + \dots + r_N) \quad (1)$$

$g$  being the acceleration due to gravity,  $\phi_0(r_1)$  the interaction between the lowest particle and the "ground". This sum can be rewritten as

$$V(r_1, r_2, \dots, r_N) = \sum_{j=0}^{N-1} \phi_j(r_{j+1} - r_j) \quad r_0 \equiv 0 \quad (2a)$$

where

$$\begin{aligned}
 \Phi_0(r_1 - r_0) &= mg N r_1 + \varphi_0(r_1) \\
 \Phi_1(r_2 - r_1) &= mg(N-1)(r_2 - r_1) + \varphi(r_2 - r_1) \\
 \Phi_2(r_3 - r_2) &= mg(N-2)(r_3 - r_2) + \varphi(r_3 - r_2) \\
 &\quad \cdot \cdot \cdot \cdot \cdot \cdot \cdot \\
 \Phi_j(r_{j+1} - r_j) &= mg(N-j)(r_{j+1} - r_j) + \varphi(r_{j+1} - r_j) \\
 \Phi_{N-1}(r_N - r_{N-1}) &= mg(r_N - r_{N-1}) + \varphi(r_N - r_{N-1})
 \end{aligned} \tag{2b}$$

The configurational partition function of our chain is then (still defining  $r_0 \equiv 0$ , and letting  $\beta = 1/kT$ )

$$Z_c(\beta, N) = \int_0^\infty dr_1 \int_{r_1}^\infty dr_2 \int_{r_2}^\infty dr_3 \dots \int_{r_{N-2}}^\infty dr_{N-1} \exp \left\{ -\beta \sum_{j=0}^{N-1} \Phi_j(r_{j+1} - r_j) \right\} \tag{3}$$

The two quantities which interest us are the pressure and density as a function of altitude. The average  $\langle r_{j+1} - r_j \rangle \equiv s_j$  represents the average space per particle at the average point where the  $j$ -th particle is located. Hence

$$\rho(r_j) = 1/s_j \tag{4}$$

is the number of particles per unit length which is just the density at the location of particle  $j$ .

The mean height of particle  $j$  above the ground is

$$\begin{aligned}
 \langle r_j \rangle &= \langle r_1 + (r_2 - r_1) + (r_3 - r_2) + \dots + (r_j - r_{j-1}) \rangle \\
 &= s_0 + s_1 + \dots + s_{j-1}; \quad s_0 \equiv \langle r_1 \rangle
 \end{aligned} \tag{5}$$

The average of  $(r_{j+1} - r_j)$  is

$$s_j = Z_c^{-1} \int_0^\infty \int_{r_1}^\infty \dots \int_{r_{N-1}}^\infty (r_{j+1} - r_j) \exp \left\{ -\beta \sum_{j=0}^{N-1} \Phi_j(r_{j+1} - r_j) \right\} dr_1 dr_2 \dots dr_N \tag{6}$$

We start to calculate  $Z_c(\beta, N)$  by letting  $r = r_N - r_{N-1}$ . Then the  $r_N$  integral becomes

$$\int_{r_{N-1}}^\infty dr_N \exp \{ -\beta \Phi_{N-1}(r_N - r_{N-1}) \} = \int_0^\infty \exp \{ -\beta \Phi_{N-1}(r) \} dr \tag{7}$$

Continuing this process, integral by integral, we find

$$Z_c(\beta, N) = \prod_{\ell=0}^{N-1} \int_0^{\infty} \exp \{ -\beta \Phi_{\ell}(r) \} dr \quad (8)$$

The numerator in  $s_j$  is obtained in the same way and is a product of integrals just as  $Z_c(\beta, N)$ , except that the  $j$ -th factor is  $\int_0^{\infty} r \exp \{ -\beta \Phi_j(r) \} dr$  instead of  $\int_0^{\infty} \exp \{ -\beta \Phi_j(r) \} dr$ . Hence

$$s_j = \int_0^{\infty} r e^{-\beta \Phi_j(r)} dr / \int_0^{\infty} e^{-\beta \Phi_j(r)} dr \quad (9)$$

so that the density is

$$\rho_j = 1/s_j = \int_0^{\infty} e^{-\beta \Phi_j(r)} dr / \int_0^{\infty} r e^{-\beta \Phi_j(r)} dr \quad (10)$$

The pressure can be calculated in a similar manner. Suppose that the  $(j+1)$ -st and  $j$ -th particles are replaced by two of the same mass which are connected by a spring with a spring constant  $k$ . Then the pressure is directly related to the change in length of the spring. We let  $r_{j+1} - r_j = b + \xi$  where  $b$  is the equilibrium distance between the  $(j+1)$ -st and  $j$ -th particles and the potential energy of interaction

$$\phi(r_{j+1}, r_j) = \phi_0 + \frac{1}{2} k \xi^2 \quad (11)$$

Then the average force, i.e. the pressure, on the spring is measured by

$$\langle F \rangle = \left\langle -\frac{\partial \phi}{\partial \xi} \right\rangle = -k \langle \xi \rangle = -k \langle r_{j+1} - r_j - b \rangle \quad (12)$$

where, in the averaging process we now choose

$$\Phi_j(r_{j+1} - r_j) = \phi_0 + \frac{1}{2} k (r_{j+1} - r_j - b)^2 + mg(N-j)$$

and leave all other  $\Phi_{\ell}$  in the original form (Eq. (2b)). Then, following the same pattern of calculation adopted for  $s_j$  we find

$$\begin{aligned} \langle F \rangle &= -k \frac{\int_0^{\infty} (r-b) \exp \{ -\frac{1}{2}\beta k - \frac{1}{2}\beta k(r-b)^2 - mg\beta(N-j)r \} dr}{\int_0^{\infty} \exp \{ -\frac{1}{2}\beta k(r-b)^2 - \beta mg(N-j)r \} dr} \\ &= -k \frac{\int_{-b}^{\infty} \xi \exp \{ -\frac{1}{2}\beta k\xi^2 - mg\beta(N-j)\xi \} d\xi}{\int_{-b}^{\infty} \exp \{ -\frac{1}{2}\beta k\xi^2 - mg\beta(N-j)\xi \} d\xi} \end{aligned} \quad (13)$$

If we choose the spring constant to be very large, the limits of integration can be extended to cover the range  $(-\infty, \infty)$ , since the main contributions to the integrands in Eq. (13) come from the neighbourhood of  $\xi = 0$ . Then one finds the pressure to be

$$\langle F \rangle = mg(N-j) \equiv P_j \quad (14)$$

which is the weight of the atmosphere over the  $j$ -th particle, as one would expect.

To determine the pressure and density as a function of height one finds the altitude of the  $j$ -th particle from Eqs (5) and (6)

$$\langle r_j \rangle = \sum_{l=0}^{j-1} \left\{ \int_0^{\infty} r \exp[-\beta \Phi_j(r)] dr \middle/ \int_0^{\infty} \exp[-\beta \Phi_j(r)] dr \right\} \quad (15)$$

one then identifies the pressure (14) and density (10) with this height.

### 2.3. Hard sphere molecules

An understanding of the above equations can easily be derived from studying the special case of hard sphere molecules of diameter  $a$ , so that

$$\Phi(r) = \begin{cases} 0 & \text{if } r > a \\ \infty & \text{if } r < a \end{cases} \quad (16)$$

Then

$$\begin{aligned} \int_0^{\infty} \exp(-\beta \Phi_j) dr &= \int_a^{\infty} \exp[-\beta mg(N-j)r] dr \\ &= [\beta mg(N-j)]^{-1} \exp[-[a\beta mg(N-j)]] \end{aligned} \quad (17)$$

and

$$\int_a^{\infty} r \exp(-\beta \Phi_j) dr \middle/ \int_a^{\infty} \exp(-\beta \Phi_j) dr = a + kT/mg(N-j)$$

so that the average density at the mean position of the  $j$ -th atom is

$$\rho_j = \left\{ a + \frac{kT}{mg(N-j)} \right\}^{-1} = \left\{ a + kT/P_j \right\}^{-1} \quad (18)$$

This shows that the average pressure and density at the average location of any particle are related by

$$P_j = kT/(\ell_j - a) \quad (19)$$

where  $\ell_j = 1/\rho_j$  is the specific length per particle at  $r_j$ . This relation is, however, just the equation of state of a hard sphere one-dimensional gas

confined to a line of length  $N\ell$ ;  $P=kT/(\ell-a)$  first derived by Tonks [7], as we set out to prove.

Furthermore, the average altitude of the  $j$ -th particle (see Eq. (5)) is

$$\langle r_j \rangle = ja + (kT/mg) \sum_{n=1}^j (N-n)^{-1} \quad (20)$$

$$= ja + (kT/mg) [\psi(N) - \psi(N-j)] \quad (21)$$

where  $\psi(x)$  is the logarithmic derivative of the gamma function

$$\psi(x) = \Gamma'(x)/\Gamma(x) \quad (22a)$$

which has the property

$$\psi(x+n) - \psi(x) = \frac{1}{x} + \frac{1}{x+1} + \dots + \frac{1}{x+n-1} \quad (22b)$$

when  $x$  is an integer, say  $n$

$$\psi(n) \equiv \gamma + \sum_{k=1}^{n-1} k^{-1} \quad (22c)$$

where  $\gamma$  is Euler's  $\gamma = 0.57721\dots$

It is well known that for large  $Z$

$$\psi(Z) = \log Z - \frac{1}{2Z} - \frac{1}{12Z^2} - \dots$$

Hence

$$\psi(N) - \psi(N-j) = -\log(N-j)/N + \frac{1}{2N} (j/N)/(1-j/N) + \dots \quad (23)$$

If  $N=O(10^8)$  and  $(j/N) \approx f$  is considered to be a finite fraction of the molecules, say  $0.01 < f < 0.99$ , then the second term in Eq. (23) is negligible compared to the first and we can write

$$\langle r_j \rangle = ja - (kT/mg) \log[1 - (j/N)] \quad (24)$$

When  $(j/N)$  is small as one would find in the lower part of the atmosphere

$$\langle r_j \rangle \sim j[a + kT/Nmg] \quad (25)$$

so that the average space per particle is  $a+(kT/Nmg)$ , the sum of its diameter and a term which depends on the ratio of the mean kinetic energy of the particle and the total weight of the atmosphere.

## 2.4. Arbitrary short-ranged pair potential

These results are easily generalized to the case of molecules with repulsive cores and short-range attractive potentials. To best appreciate the results given in Section 2 let us review some aspects of the theory of a chain of particles with interaction between successive pairs only in the absence of a gravitational field when the particles are confined to a linear box of length  $L$ . This problem has been discussed by many authors (see Refs [8, 9]).

The configurational contribution to the standard partition function in this case is

$$Z_c(L, \beta, N) = \int_0^L dr_N \dots \int_0^{r_3} dr_2 \int_0^{r_2} dr_1 \exp \left[ - \sum_{j=0}^N \beta \phi(r_{j+1} - r_j) \right] \quad (26)$$

where  $\phi(r_1 - r_0)$  and  $\phi(r_{N+1} - r_N)$  represent interactions of the end particles with the walls at the end of the box and may have a different form from the other  $\phi$ 's. We set  $r_0 \equiv 0$  and  $r_{N+1} = L$  and neglect this difference, which is irrelevant when  $N$  is large.

The thermodynamic properties of this assembly are better discussed in the isobaric ensemble rather than in the canonical ensemble. The isobaric partition function is defined by

$$Z_I(\beta, N, P) = \int_0^\infty Z(T, N, L) e^{-LP\beta} dL \quad (27)$$

the Laplace transform of the ordinary partition function,  $Z(T, N, L)$ , which is the product of Eq. (26) and the kinetic energy factor  $(2\pi\hbar^2 mkT)^{3N/2}$ . The chemical potential and volume (here length  $\langle L \rangle$ ) are related to the isobaric partition function by

$$\mu N = -kT \log Z_I(\beta, N, P)$$

and

$$\langle L \rangle = -kT(\partial/\partial P) \log Z_I(\beta, N, P) \quad (27a)$$

$\langle L \rangle$  is identified as our box length).

Now let  $\mathcal{L}(f)$  be the Laplace transform of the function  $f$ . By the Laplace Faltung theorem

$$\mathcal{L} \left[ \int_0^x f_1(x-x_1) f(x_1) dx_1 \right] = \mathcal{L}(f_1) \mathcal{L}(f_2) \quad (28)$$

It follows that the Laplace transform of the  $N$ -fold integral formula (26) for  $Z_c$  is, neglecting boundary conditions,

$$\mathcal{L}(Z_c) = \left\{ \int_0^\infty \exp[-Pr/(kT)] \exp[-\phi(r)/(kT)] dr \right\}^{N-1} \quad (29)$$

Hence the isobaric partition function  $Z_j$  for the assembly is expressed as a product of isobaric partition functions for independent pairs of particles and, for large  $N$ ,

$$N^{-1} \ln Z_i = \frac{3}{2} \log(2\pi h^2 m kT) + \log \int_0^\infty \exp [-(Pr + \varphi)/(kT)] dr \quad (30)$$

If the length  $\ell = L/N$  is given, then  $P$  is obtained from the equation of state

$$\begin{aligned} \ell &= -\beta^{-1} \frac{\partial}{\partial P} \log \int_0^\infty \exp [-\beta(Pr + \varphi)] dr \\ &= \frac{\int_0^\infty r \exp [-\beta(rP + \varphi)] dr}{\int_0^\infty \exp [-\beta(rP + \varphi)] dr} = \langle r \rangle_{Av} \end{aligned} \quad (31)$$

where  $\langle r \rangle_{Av}$  is the average value of  $r$  weighed according to the normalized distribution function indicated above.  $P$  is a monotone decreasing function of  $\ell$  in any  $(P, \ell)$  isotherm, since

$$(\partial P / \partial \ell) = -kT \{ \langle (r - \bar{r})^2 \rangle_{Av} \}^{-1} < 0$$

Since  $\partial P / \partial \ell \neq 0$ , no liquid-gas equilibrium and no phase transition can exist in our one-dimensional model.

If we introduce Eq. (2b) into Eq. (10) and write  $mg(N-j) \equiv P_j$  as exhibited in Eq. (14) we note that Eq. (10) becomes

$$\frac{1}{\rho_j} = \int_0^\infty r \exp [-\beta \{\varphi(r) + r P_j\}] dr / \int_0^\infty \exp [-\beta \{\varphi(r) + r P_j\}] dr \quad (32)$$

Hence if we identify the linear density  $\rho_j$  at the location of the  $j$ -th particle as the reciprocal of  $\ell_j$ , the mean space available to the  $j$ -th particle, then Eq. (32) is the equivalent to Eq. (31) so that at the location of every particle the standard equation of state for particles in a box is applicable to relating the local density and pressure in our particles in a gravitational field.

Campagner [10] has also considered this problem recently, but his main results are derived in the limit case as the particle mass and Boltzmann's constant become vanishingly small while  $Nm$  and  $Nk$  remain fixed (the number of particles  $N \rightarrow \infty$ ).

### 3. REMARKS ON VEHICULAR TRAFFIC

#### 3.1. Introduction

In his introductory paper in these Proceedings Professor Pines states that perhaps experience with many-body systems would lead to some insight

into sociological many-body problems. While I did not originally intend to dwell on such problems, his remarks have tempted me to discuss what I consider to be one of the simplest problems of social interaction and yet one which has many characteristics of more complicated situations. It is the problem of the motion of vehicles which form a single line of traffic on an open road, but under conditions in which passing is very difficult. This situation is an important one because even on multiple lane highways a traffic problem arises only under high density conditions, those under which very little passing occurs.

This problem is the "hydrogen atom" of behavioural science because each vehicle can be characterized by the value of only one variable as a function of time, its acceleration. The interaction with other vehicles is mainly a nearest neighbour one, each driver mainly interacting with the vehicle in front of him. The line of traffic is one-dimensional. In most other aspects of group behaviour, more variables are important and the interaction graph is more complicated.

Following the usual practice of physicists, one might first consider the "free particle", the driver-vehicle combination (which we refer to as the DV) in the absence of other traffic. While the driver may wish to proceed with some average velocity,  $v$ , there are always fluctuations. The acceleration distribution function might be used to characterize the DV. Experimentally, the distribution is essentially Gaussian. On a good road, a typical value for the width of the Gaussian is about  $0.3 \text{ ft/s}^2$ . In traffic this width may be multiplied by a factor of ten. Drivers might be compared by their acceleration distribution function width when driving under similar conditions. Roads might be compared by comparing the widths for a given driver on them.

### 3.2. Equations of motion [11, 12]

In the presence of other vehicles, the acceleration of a given DV has the random component discussed above (which we shall neglect for the moment) and an interaction component which depends on the motion of the preceding vehicle in a line. On an open road under rather dense traffic conditions, the traffic stream progresses with some mean velocity  $v$  and some average space per vehicle,  $a$  (which we define to be the distance between the front bumper of a vehicle and that of the vehicle behind it). In a co-ordinate system, moving with the velocity  $v$ , each vehicle under stable conditions undergoes small vibrations with the inter-vehicle spacing varying by a small amount about  $a$ . We let  $x_j(t)$  represent the deviation of the position of the  $j$ -th vehicle from where it should be at time  $t$  (the traffic is postulated to be moving to the right with the lead car numbered 1, the next 2, etc.).

The basic scheme for investigating the motion of a line of traffic is to introduce a set of stimulus-response equations, the stimulus being the interaction between successive vehicles and the response being the acceleration of the vehicles. In the linear small oscillation from equilibrium regime, a typical such set of equations might be

$$\ddot{x}_j(t) = \mu [x_{j-1}(t) - x_j(t)] \quad (1)$$

where  $\mu$  is a proportionality constant. This corresponds to the  $j$ -th car accelerating when the spacing is greater than the mean spacing, and de-

celerating when it is smaller — at first glance a rather reasonable assumption. One way of testing the validity of such a stimulus-response relation is to see how a line of traffic would respond to small fluctuations in the motion of the lead car. Since we know that rather stable lines of traffic exist, Eq.(1) must not amplify small fluctuations in the motion of the lead car. Let us Fourier-analyse the motion of the lead car and consider the component which corresponds to frequency  $\omega$ :

$$x_1(t) = x_1 \exp i\omega t$$

The resulting response of the  $j$ -th car is

$$x_j(t) = x_j \exp i\omega t$$

It can be shown that there is a resonance at  $\omega^2 = \mu$  so that

$$x_j = [1 - (\omega^2/\mu)]^{-n} x_1$$

which means that Eq.(1) describes an amplifier for small disturbances and is not an appropriate equation to describe the interaction between vehicles. Another law of following which might be more suitable is

$$\dot{v}_j(t) = \lambda_0 [v_{j-1}(t) - v_j(t)] \quad (2)$$

which means that the  $j$ -th vehicle accelerates or decelerates according to whether the relative velocity of the  $(j-1)$ -st and  $j$ -th vehicles is increasing or decreasing. It can be shown that any frequency component of the fluctuations of the lead vehicle is damped out when Eq.(2) is valid. Indeed, it can be shown that if  $\ddot{x}_j(t)$  has a component which is any even derivative of  $(x_{j-1} - x_j)$  then a resonance exists, while a right-hand side of Eq.(2) which is composed of only odd derivatives of this difference leads to a stable small vibration stimulus-response model. Since it is doubtful that a driver can observe third derivatives of the spacing, Eq.(2) seems like a reasonable set of equations to start from.

One objection to these equations is that they represent an immediate response to a fluctuation. In reality, there is a time lag which has three components. It requires a time  $\Delta_1$  to recognize that a fluctuation exists, a time  $\Delta_2$  to make a decision on how to respond to the fluctuation, and a time  $\Delta_3$  for the driver and the vehicle to put the response into effect. If the total time lag  $\Delta = \Delta_1 + \Delta_2 + \Delta_3$  is the same at all times, then

$$\dot{v}_j(t) = \lambda_0 [v_{j-1}(t - \Delta) - v_j(t - \Delta)] \quad (3)$$

It might be better to introduce a weight function  $\sigma(\Delta)$  for time lags. However, if this function has a single peak, the results are essentially the same as those for a single time lag. With an equation such as (3), we would expect instabilities to result when both  $\Delta$  and  $\lambda_0$  are large, for in that case a strong response would be given to an event which occurred in the distant past. Since even the sign of the fluctuation might have changed during the time lag period, one might be giving the wrong response. It has, in fact, been shown that the condition for stability is

$$\Delta \lambda_0 < \frac{1}{2} \quad (4)$$

A number of car following experiments have been performed and Eq. (3) seems to be very good for high density driving conditions. Typical values for  $\lambda_0$  and  $\Delta$  for American driving with medium size American cars are, respectively,  $0.35 \text{ s}^{-1}$  and  $1.5 \text{ s}$ . Bus drivers have somewhat smaller values of  $\Delta$  and some drivers tested violate the inequality (4). However, as long as long runs of such drivers do not appear in real traffic, the line is stable.

When traffic is not so dense, it is expected that  $\lambda_0$  should be a function of the separation distance, since two cars which are very widely separated cannot influence each other, i.e.  $\lambda_0$  should vanish as the separation distance becomes infinite. It has been observed experimentally that the response strength parameter  $\lambda_0$  is inversely proportional to the separation distance, so that

$$\dot{v}_j(t) = \lambda \frac{v_{j-1}(\tau) - v_j(\tau)}{x_{j-1}(\tau) - x_j(\tau)_{\tau=t-\Delta}} \quad (5)$$

This formula is consistent with car following experiments which yield a continuous record of the quantities involved. From it, as will be shown below, one can also deduce an "equation of state" for traffic which can be compared with experiment.

### 3.3. Equation of state of traffic [13, 14]

We now use Eq. (5) to derive a relationship between the mean density and mean velocity of a line of traffic. Equation (5) can be integrated to yield

$$\begin{aligned} v_j(t_1) &= \lambda \log \{x_{j-1}(t_1 - \Delta) - x_j(t_1 - \Delta)\} \\ &= v_j(t_2) - \lambda \log \{x_{j-1}(t_2 - \Delta) - x_j(t_2 - \Delta)\} \end{aligned} \quad (6)$$

The local separation distance at the  $j$ -th car represents the length available per car which is the reciprocal of the local density which is the number of cars per unit length which we set equal to  $\rho_j(t)$  with

$$1/\rho_j(t) = x_{j-1}(t) - x_j(t) \quad (7)$$

Our required equation of state is obtained by averaging over a long line of traffic. We define the arithmetic mean velocity by

$$v(t) = \frac{1}{N} \sum_k^N v_j(t) \quad (8a)$$

and the geometric mean density by

$$\log \rho(t) = \frac{1}{N} \sum_j \log \rho_j(t) = \frac{1}{N} \log [\rho_1(t) \rho_2(t) \dots \rho_N(t)] \quad (8b)$$

Then summing over Eq. (6) we find

$$v(t_1) + \lambda \log \rho(t_1 - \Delta) = v(t_2) + \lambda \log \rho(t_2 - \Delta) = \text{constant} \quad (9)$$

since  $t_1$  and  $t_2$  are independent variables. When the traffic stream is changing its character only slowly, the time lag  $\Delta \approx 1.5$  s is irrelevant so that  $t_1 - \Delta$  and  $t_2 - \Delta$  can be replaced respectively by  $t_1$  and  $t_2$ . The value of the constant is determined from the fact that at close packing, i.e. bumper-touching-bumper state, the velocity is zero. If  $\rho_0$  is the density in this close packed state, we find that at any time the relationship between the average velocity and average density is

$$v = \lambda \log(\rho_0 / \rho) \quad (10)$$

The flow rate of the traffic in cars per unit time past a given point,  $q$ , is given by  $q = \rho v$ . Hence from Eq. (10) we find the flow rate density relation to be

$$q = \rho v = \lambda \rho \log \rho_0 / \rho \quad (11)$$

This curve is plotted in Fig. 3 with certain observational data [15] taken from the traffic flow in Lincoln Tunnel in New York City.

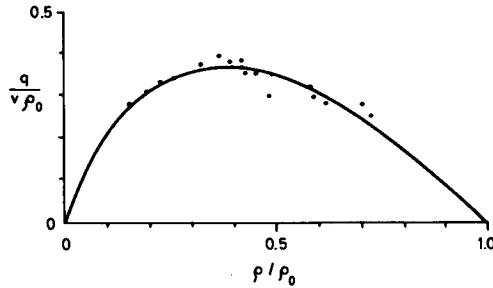


FIG. 3. Variation of flow with density. The curve is from Eq. (11).  $\rho_0 = 228$  cars/mile,  $v = 17.2$  min/h.

In closing we might ask if there are any features of the above analysis which are more broadly applicable than to our special example. Consider an individual who is attempting to perform some task. The performance might be characterized by one or more parameters which we suppose to be continuous variables (we are not concerned here with processes which involve a sequence of yes or no decisions; these may or may not fit into the scheme discussed below but, in any case, they require more discussion and precise definition). It usually has some distribution function just as the acceleration of our DV, even in the absence of interactions with other individuals. When individuals interact, there are some regimes in which their performance is going smoothly so that a small vibration theory can be developed provided that some graph of interactions can be constructed. In an industrial or government hierarchy, the organization chart gives a first approximation to the graph. Our three components to total time lags exist in almost all human

interaction processes. In a world of "eager beavers" who react strongly to all stimuli, the conditions for stability analogous to Eq. (4) are not always recognized and instabilities often arise from reacting too quickly and too strongly to long existing but only recently recognized difficulties. The stimulus-response equations would generally be vector rather than scalar equations when the  $k$ -th component of the  $j$ -th individuals response vector would correspond to the  $k$ -th variable required to characterize his behaviour. The analogue of our equation of state would generally follow from the non-linear equations which would describe the individual's performance over a broad range of stimuli.

## R E F E R E N C E S

- [1] DOMB, C., MARADUDIN, A. A., MONTROLL, E. W., WEISS, G. H., Phys. Rev. 115 (1959) 18.
- [2] DEAN, P., Proc. R. Soc. A254 (1960) 507.
- [3] HORI, J., Spectral Properties of Disordered Chains and Lattices, Pergamon Press (1968).
- [4] MATSUDA, H., Prog. theor. Phys., Japan 31 (1964) 161.
- [5] PAYTON, D. N., III., VISSCHER, W. M., Phys. Rev. 154 (1967) 802; 156 (1967) 1032.
- [6] ROSENSTOCK, H. B., McCILL, R. E., Vibrations of Disordered Solids (preprint).
- [7] TONKS, L., Phys. Rev. 50 (1936) 955.
- [8] TAKAHASI, H., Proc. phys.-math. Soc. Japan 24 (1942) 60.
- [9] VAN HOVE, L., Physica 16 (1950) 137.
- [10] CAMPAGNER, A., Physics Lett. 21 (1968) 627.
- [11] CHANDLER, R. E., HERMAN, R., MONTROLL, E. W., Ops Res. 6 (1958) 165.
- [12] HERMAN, R., MONTROLL, E. W., POTTS, R. B., ROTHERY, R., Ops Res. 7 (1959) 86.
- [13] GAZIS, D. C., HERMAN, R., POTTS, R. B., Ops Res. 7 (1959) 499.
- [14] MONTROLL, E. W., Proc. First Symp. Engineering Applications of Random Function Theory and Probability, J. Wiley and Sons (1963) 231.
- [15] GREENBERG, H., Ops Res. 7 (1959) 79.



# SOME APPLICATIONS OF PATH INTEGRALS AND DIAGRAMMATIC METHODS TO CHEMICAL PHYSICS

P.G. de GENNES  
Faculté des Sciences,  
91 Orsay, France

## Abstract

SOME APPLICATIONS OF PATH INTEGRALS AND DIAGRAMMATIC METHODS TO CHEMICAL PHYSICS. 1. Introduction. 2. Denaturation of DNA. 3. Lamellar structures.

## 1. INTRODUCTION

The configurations of a long flexible molecule can be viewed as the trajectories of a non-relativistic particle. This analogy was first noted by Edwards [1]. To understand it, let us first consider the average configuration of a flexible chain under inhomogeneous external forces. We think of the chain as a collection of beads located at points  $\vec{r}_1, \vec{r}_2, \dots, \vec{r}_n$ . A small elongation of the  $(m_1 m + 1)$  unit (i.e. a finite  $\vec{r}_{m+1} - \vec{r}_m$ ) implies a reduction of entropy amounting to

$$\frac{3[\vec{r}_{m+1} - \vec{r}_m]^2}{2a^2}$$

where  $a$  is the r.m.s. elongation of one unit (and Boltzmann's constant is taken to be unity). This gives a free energy

$$G = -TS = \sum_m \frac{3T}{2a^2} [\vec{r}_{m+1} - \vec{r}_m]^2$$

and a force on the  $m$ -th bead

$$\frac{\partial G}{\partial \vec{r}_m} = \frac{3T}{a^2} (\vec{r}_{m+1} - 2\vec{r}_m + \vec{r}_{m-1}) \approx \frac{3T}{a^2} \frac{\partial^2 \vec{r}}{\partial m^2}$$

The average configuration of the chain is obtained by balancing this against the external force  $F(r_m)$  applied on the  $n$ -th bead

$$\vec{F}(\vec{r}_m) = - \frac{3T}{2a^2} \frac{\partial^2 \vec{r}}{\partial m^2}$$

This is the equation for the trajectory of a non-relativistic particle of

mass  $3T/2a^2$ ,  $m$  is a force field  $F$  playing the role of an (imaginary) time.

There are a number of problems from which the average chain configuration, obtained from this equation, leads to non-trivial first approximations. This is true in particular for the excluded volume problem discussed by Edwards [1] where the force  $F$  is due in fact to a repulsion between all chain units and must accordingly be made self-consistent.

In certain cases, however, knowledge of the average chain configuration is not enough and we must ascertain the weight of all configurations where a chain (of  $m$  units) starts from point 1 and ends at point 2. For an ideal flexible chain, this weight  $G_m(1, 2)$  is ruled by a Schrödinger-like equation

$$-\frac{\partial G_m(12)}{\partial m} = -\frac{a^2}{6} \nabla_{(2)}^2 G_m + \frac{V(2)}{T} G_m$$

where  $\vec{V(r)}$  is the potential from which  $F$  is derived. Again  $m$  plays the role of an imaginary time. Thus a number of problems in quantum mechanics find their counterpart in the statistical mechanics of long chains. A few examples will be presented in the next sections.

## 2. DENATURATION OF DNA

Deoxyribonucleic acid (DNA) is the vector of the genetic code in nearly all living beings. One DNA molecule is made of two strands, each of them carrying a well-defined sequence of four bases (A, T, G, C) (Fig.1). In the normal room-temperature configuration the two strands build up a rigid double helix, in which the bases of different strands are linked in pairs. These bonds are highly selective (A can pair only with T, and G with C). The binding energy per pair is typically  $1/3$  eV. However, a mild raise in temperature "denatures" the sequence: in some regions the two strands become separated and behave as flexible coils (Fig.2). This implies an enormous increase in entropy and explains why the transition occurs at comparatively low temperatures ( $\gtrsim 400^\circ K$ , to be compared with a binding energy of  $\sim 3000^\circ K$ /pair).

With natural DNA, in this partially denatured state, there is essentially only one way of putting the two strands in register. The  $n$ -th base on strand one can link only with the  $n$ -th unit on strand two: in our quantum mechanical picture, the diagram of Fig.2 corresponds to two particles with an attractive, instantaneous (equal  $n$ ) interaction of finite range. (The lower the temperature, the stronger the attraction.) In three dimensions we know that such an interaction leads to a bound state only when it exceeds a certain threshold. Here this means that there is a well-defined transition temperature  $T_m^1$  (i.e. a singularity in the partition function for two infinitely long strands). Below  $T_m$  there is long-range order in the following sense: If we know that the two strands are linked at  $(n-n)$ , there is a finite probability of finding another link at  $(m-m)$ , even when  $|n-m| \rightarrow \infty$ .

This is an interesting example of a strict phase transition occurring with long molecules. In practice there are various physical complications,

---

<sup>1</sup> See e.g. Ref. [2] (where a different approach is used) and the references quoted therein.

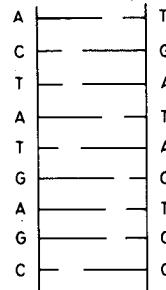


FIG. 1. Schematic representation of a double-stranded DNA with complementary base pairing. In the actual sequence the two strands are twisted round one another (double helix).



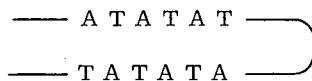
FIG. 2. Partial denaturation of a natural DNA. The helical regions are represented by two parallel linear segments. In the coiled regions the two strands are separated.

in particular the two types of pairs (AT and GC) do not have the same binding energy and since they are distributed more or less at random, this broadens the transition. These effects are discussed by Montroll in these Proceedings.

To avoid these complications, one may then study certain synthetic nucleic acids where the sequence of bases has some simple periodicity<sup>2</sup>. Here I shall restrict my attention to one particular family of such acids, the so-called alternant copolymers, an example of which is the sequence

... A T A T A T ...

With such a sequence a single strand can fold on itself and make a double helix ("hairpin" helix)



When such a hairpin is partially denatured it takes the branched structure shown in Fig.3. The overall size (the radius of gyration) of such a branched molecule may be measured by various means, and the qualitative aspects of the results are shown in Fig.4b. The size is minimum in the melting region. Why?

<sup>2</sup> For a review of synthetic polynucleotides see Ref. [3].

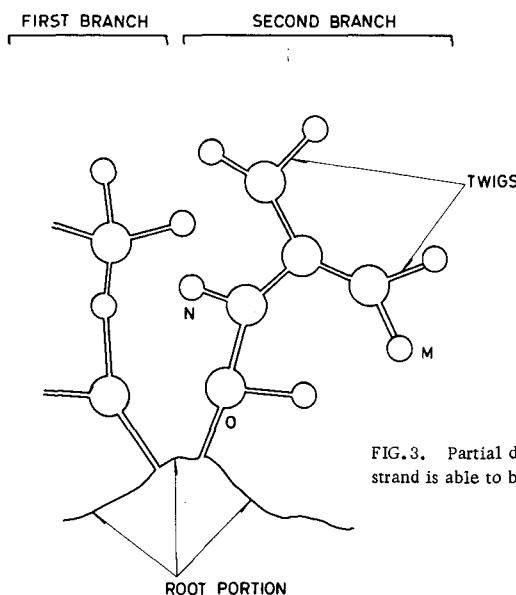


FIG. 3. Partial denaturation of an alternant copolymer — a single strand is able to build up double helices with itself.

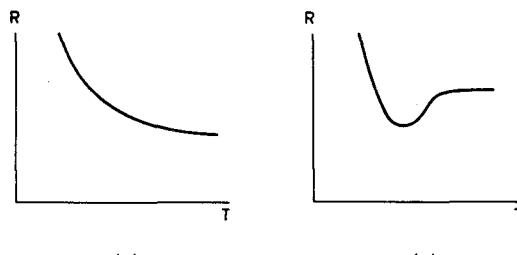


FIG. 4. Radius of gyration of a DNA molecule as a function of temperature in the "melting" (denaturation) region. a) Natural DNA, b) alternant copolymer. (Both plots are only qualitative.)

The answer can be obtained from a summation of the relevant diagrams [4] weighted as shown in Fig. 5. Each coil portion has a free propagator  $G$ . A helical portion of  $M$  bonds has a factor  $s^M$  where  $s$  is the equilibrium constant of the hairpin  $\leftrightarrow$  coil reaction ( $s \gtrsim 1$  below the melting temperature). Finally, for each helical rod we must also include a small factor  $\eta$ , expressing the fact that the first bond in the helix is much weaker than the others (some of the van der Waals attractions between stacked bases are lacking).

The final results concerning the size of the molecule can be obtained by a comparatively simple argument: The diagrams exemplified in Fig. 3 are formally identical with the diagrams which occur in the scattering of a low-energy electron by random impurities in a metal<sup>3</sup>, a helical portion

<sup>3</sup> A good review is found in Ref. [5].

FIG.5. Weighting factors for a typical diagram contributing to the partition function.

$$1 \text{ } \begin{array}{|c|} \hline \text{---} \\ \hline \end{array} 2 = 1 \text{ } \begin{array}{|c|} \hline \text{---} \\ \hline \end{array} 2 + 1 \text{ } \begin{array}{|c|} \hline \text{---} \\ \hline \end{array} \text{---} \text{---} 2 + \dots$$

$$= 1 \text{ } \begin{array}{|c|} \hline \text{---} \\ \hline \end{array} 2 + 1 \text{ } \begin{array}{|c|} \hline \text{---} \\ \hline \end{array} \text{---} \text{---} 2$$

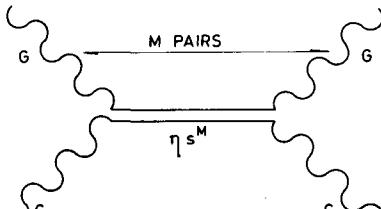


FIG.6. The Bethe-Salpeter equation for the two-point correlation function. Each individual propagator is already renormalized (lateral "twigs" are allowed).

corresponding to two successive scatterings by the same impurity. In this language: a) The average end-to-end distance  $R_{12}$  is the range of the one-particle propagator, and this is bounded by the mean free path  $\ell$ . Thus for a long chain  $R_{12} \sim \ell$  is finite and independent of the length of the chain. b) The "root" in Fig.3 is thus rather small and all the extension takes place in the "branches". To study the size of a branch we note that it has a cyclic structure. We are then led to consider the diagrams of Fig.6 where two points are connected by two chain portions. These diagrams occur in the electron problem for the density-density correlation function, i.e. for transport properties. The corresponding Bethe-Salpeter equation describes the diffusion of one electron colliding on successive impurities. The diffusion coefficient  $D = 1/3 v\ell$  where the electron velocity  $v$  is related to the energy  $E$  by  $E = 1/2 mv^2$ . The range of the kernel of Fig.6 is then given by the diffusion law

$$R \sim \sqrt{Dt}$$

From the uncertainty principle  $E \sim 1/t$ ,  $v \sim t^{1/2}$ ,  $D \sim t^{1/2}$  and  $R \sim t^{1/2}$ . Returning to the molecular problem where the analogue of  $t$  is the number  $N$  of bases along the chain ( $N \sim 10^5$ ) we find  $R \sim N^{1/2}$ . This explains the kink in Fig.4b. At low T (rigid helix)  $R \sim N$ . At high T (ideal coil)  $R \sim N^{1/2}$ . Thus  $R$  has a minimum in the transition region.

### 3. LAMELLAR STRUCTURES

Soaps are made up of water + fatty acid ion + cation. In certain concentration ranges they take up the lamellar structure shown in Fig.7 [6]. A single lamella may have some (rough) structural analogy with a cellular membrane. Many physical experiments cannot be carried out on membranes because the amount of matter available is too small (membrane thickness  $\sim 80 \text{ \AA}$ ). This difficulty does not occur with soap-like structures and this is one of the motivations for this study.

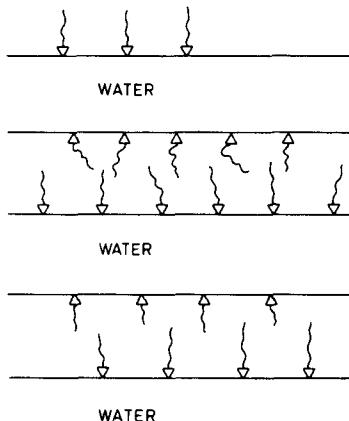


FIG. 7. Schematic structure of the lamellar plane of soaps. The arrows denote the ionic part of a fatty acid chain.

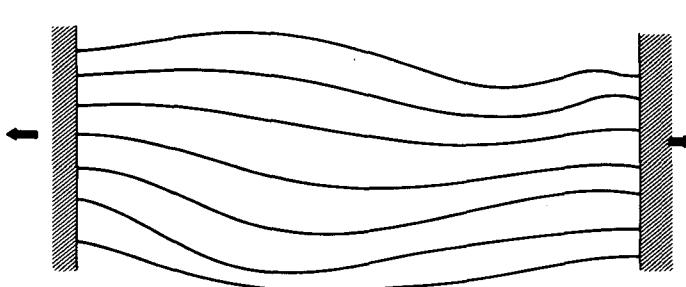


FIG. 8. Two-dimensional set of flexible chains under tension. Successive chains are not allowed to cross each other.

We often find lamellar structures with rather large water layers (up to 30 Å). The forces between lipidic units separated by such distances should be very small (electrostatic forces are screened out efficiently by the cations). Yet the X-ray experimentalists observe rather well-defined Bragg reflections measuring the lamellar periodicity. Why?

Consider the following two-dimensional model (Fig. 8). Flexible lines of negligible thickness are attached between two plates and put under a tension  $F$ . This provides a preferred axis (which we call the  $t$  axis) and we can study the X-ray intensities  $I(q)$  for a wave vector  $q$  normal to the  $t$  axis (parallel to the  $x$  axis). There is no force between the lines, except that they cannot cross each other.

This model can be solved if we interpret each line configuration  $x(t)$  as a particle trajectory [7]. For instance, with one single line, the statistical weight of a configuration linking points  $(x', t')$  and  $(x'', t'')$  is

$$G(x', t'; x'', t'') = \int \mathcal{D}x(t) \exp \left[ -\frac{F}{T} \int_{t'}^{t''} \sqrt{1 + \left( \frac{dx}{dt} \right)^2} dt \right]$$

$$\approx \text{const} \int \mathcal{D}x(t) \exp \left[ -\frac{F}{2T} \int_{t'}^{t''} \left( \frac{dx}{dt} \right)^2 dt \right]$$

where  $\int \mathcal{D}x(t)$  is an integral over paths and where we have gone directly to the "non-relativistic" (high tension) limit.  $G$  has exactly the form of a free-particle propagator. Similarly for  $N$  lines the statistical weight is an  $N$  particle propagator and the requirement of no intersection can be satisfied by imposing Fermi statistics. The problem is thus brought into correspondence with a one-dimensional gas of free fermions and can be solved exactly. There is no long-range order (no delta function in  $I(q)$ ) but a very singular type of short-range order:  $I(q)$  diverges logarithmically at the  $q$  value corresponding to the average interline distance. This corresponds to the well-known Kohn anomaly of the density-density response function in a free fermion gas [8].

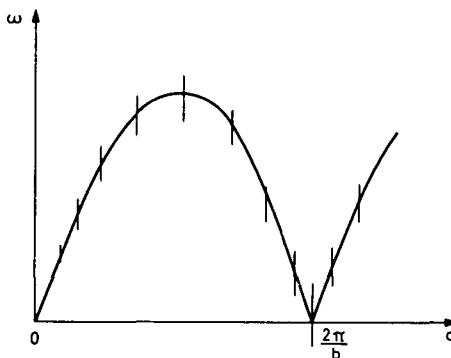


FIG. 9. Dispersion relation for density fluctuations in a lamellar structure for a wave vector  $q$  normal to the plane of the lamellae. In the region  $q \sim 2\pi/b$  the frequency spectrum is expected to be rather narrow.

Extrapolating boldly to three dimensions, these results suggest that the experimental X-ray peaks observed in soaps are compatible with very weak interlayer forces, but are not necessarily to be associated with a strict long-range order.

"Quasi-long-range order" should also occur in the conventional smectic liquid crystals<sup>4</sup> with compact lamellar structures and strong interlayer forces. Although this goes somewhat beyond the scope of the present paper, it may be physically interesting to discuss briefly some collective motions in such systems. Consider now specifically the density fluctuations  $\rho_q$  with a wave vector  $q$  normal to the layers. If we had an exact lamellar periodicity the corresponding spectrum of longitudinal excitations  $\omega(q)$  would be periodic in  $q$ . With quasi-long-range order the spectrum should broaden somewhat and shift as shown in Fig. 9. The frequency width of the dynamic form factor

$$S(q, \omega) = \langle |\rho(q, \omega)|^2 \rangle$$

at fixed  $q \approx 2\pi/d$  (where  $d$  is the average layer thickness) should be very narrow. Physically, this means that since the fluctuations in  $\rho(q = 2\pi/d)$  are large, they can only relax slowly. In fact, similar (but weaker)

<sup>4</sup> For an introductory review on liquid crystals see, for instance, Ref. [9].

narrowing effects have been predicted and observed by neutron inelastic scattering in conventional isotropic liquids [10]. Neutron experiments should be done on selected (deuterated) smectic systems where the scattering is dominantly coherent.

To conclude, we have met a few typical examples of statistical problems involving chains and layers — in polymers, in liquid crystals and even in some systems of (slight) biological interest. Some of these problems are essentially unsolved; in some other cases analogies with quantum mechanics have been helpful. It seems to me that a growing number of theoretical physicists should enter the field.

#### R E F E R E N C E S

- [1] EDWARDS, S.F., Proc. phys. Soc. 85 (1965) 613.
- [2] ZIMM, B.H., J. chem. Phys. 33 (1960) 1349.
- [3] MICHELSON, A.M., MASSOULIE, J., GUSCHLBAUER, W., Progress nucleic Acid Res. mol. Biol. 6 (1967) 83.
- [4] de GENNES, P.G., Biopolymers, to be published.
- [5] RICKAYZEN, G., Proceedings of the Bergen summer school (1960).
- [6] LUZZATI, V., in Biological Membranes (Chapman, D., Ed.), Academic Press, N.Y. (1968).
- [7] de GENNES, P.G., J. chem. Phys., to be published.
- [8] KOHN, W., Phys. Rev. Lett. 2 (1959) 393.
- [9] CHISTIAKOV, J.G., Soviet Phys. Usp. 9 (1967) 551.
- [10] de GENNES, P.G., "Neutron scattering by 'normal' liquids", Inelastic Scattering of Neutrons in Solids and Liquids (Proc. Symp. Vienna, 1960), IAEA, Vienna (1961) 239.

PLASMA PHYSICS, TURBULENCE,  
QUANTUM OPTICS  
AND STATISTICAL MECHANICS



# PLASMA PHYSICS: GENERAL SURVEY

M.N. ROSENBLUTH

Institute for Advanced Study,

Princeton, N.J., United States of America

## Abstract

PLASMA PHYSICS: GENERAL SURVEY. A high-temperature, relatively dilute, gaseous plasma is considered. First, general topics are briefly discussed: the equations of motion (Vlasov's, Fokker-Planck's MHD), the problem of small amplitude oscillations in an infinite, homogeneous plasma, the Landau damping, and the motion of a charged particle in a given magnetic field. Attention then is concentrated on the problem of the stable confinement of a plasma by means of a magnetic field. Open and closed geometries are both examined and their MHD instabilities discussed, showing how they can be eliminated. Then the problem of microinstabilities is studied. A general discussion starting with quasi-thermodynamic considerations allows to individuate the most dangerous microinstabilities present in the open and closed geometries. These are briefly discussed.

This paper is confined to the physics of high-temperature, relatively dilute, gaseous plasmas with only peripheral references to solid-state plasmas where the approximations, the unknowns and the problems of particular interest are often rather different.

To begin with, we may note that quantum effects are unimportant in this regime as the energy levels of various excitations will be expected to correspond to many quanta of energy. Thus  $\hbar\omega_p \ll kT$  where  $\omega_p$  is the electron plasma frequency  $(4\pi n e^2/m)^{1/2}$  and  $\omega_c$  the cyclotron frequency. Similarly, interparticle forces are rather weak and of a completely known character, i.e. the Lorentz forces with the fields determined through Maxwell's equations from the charge and current density given by particle positions and velocities. These forces are weak, as usually  $e^2 n^{-1} \ll kT$ .

At this point – having defined the subject to be the classical limit of a system of weakly interacting particles – it may sound rather trivial, and indeed the statistical mechanics and equilibrium properties of such plasmas are relatively simple as we shall briefly sketch later. Nonetheless, it is perhaps of interest as the most rigorously solved interacting many-body problem. As must be quite evident, however, from the lectures on astrophysical plasmas at this symposium, the usual situations of interest are far from equilibrium and hence complex questions do arise. One reason for this is, of course, the difficulty of any non-linear problem such as fluid mechanical turbulence, but even the linear theory is quite rich. The reason for this is the long-range nature of the electromagnetic forces which allows a great variety of basic waves to exist whose properties may be determined by quite non-local geometrical factors. These waves have a collective property, i.e. while the interactions between individual particles may be quite weak, if many particles move coherently, they may have a large effective charge and the fields they produce may be quite significant, in particular they may be able to maintain the postulated large-scale coherent motion.

We return now to the problem of statistical equilibrium [1]. As we have seen earlier, the mean potential energy is much less than the mean

kinetic energy. This fact suggests an expansion procedure in which we let  $e, m \rightarrow 0, n \rightarrow \infty$  in such a way that the characteristic densities  $n_e, n_m$  remain unchanged.

The governing equation in the classical limit is the  $6N$ -dimensional Liouville equation for the distribution function  $D$

$$\left\{ \frac{\partial}{\partial t} + \sum_{i=1}^N \vec{v}_i \cdot \frac{\partial}{\partial \vec{x}_i} + \frac{e}{m} \left[ \vec{E}(\vec{x}, t) + \frac{\vec{v}_i}{c} \times \vec{B} \right] \cdot \frac{\partial}{\partial \vec{v}_i} \right\} D = 0 \quad (1)$$

where, if we consider only Coulomb interparticle forces

$$\vec{E}_i(\vec{x}, t) = -e \sum_{j \neq i}^N \frac{\partial}{\partial \vec{x}_i} \frac{1}{|\vec{x}_i - \vec{x}_j|} + \vec{E}^{\text{ext}} \quad (2)$$

To reduce the equation and apply our expansion we use the BBGKY method of defining  $S$ -particle distribution functions by

$$f_S(X_1, \dots, X_S, t) = \int D dX_{S+1}, \dots, dX_N \quad (3)$$

The appropriate equations for  $f_S$  may be generated by taking moments of Eq.(1) and are found of course to depend on  $f_{S+1}$ . We now solve this equation by an ansatz suggested by what we have previously noted - that interparticle correlations are weak, i.e. we take

$$\begin{aligned} f_S &= \prod_{i=1}^S f(X_i, t) + \lambda \sum_{\text{pairs}} \left[ \prod f(X_i, t) \right] P(X_j, X_k, t) \\ &\quad + \lambda^2 \sum_{p,p} \left[ \prod f(X_i, t) \right] P(X_j, X_k, t) P(X_\ell, X_m, t) \\ &\quad + \lambda^2 \sum_{\text{triplets}} \left[ \prod f(X_i, t) \right] T(X_j, X_k, X_\ell, t) \\ &\quad + \dots \end{aligned} \quad (4)$$

where  $f$  is the one-particle distribution function,  $P$  is the pair correlation function, and  $T$  the triplet correlation.  $\lambda$  is the small discreteness expansion parameter (it is equivalent to  $e^2 n^{-\frac{1}{2}} / kT$  or  $(n \lambda_D^3)^{-1}$  or  $\lambda_d / \lambda_{\text{coll}}$ ) and the functions  $f$ ,  $P$  and  $T$  are themselves expanded, i.e.

$$\begin{aligned} f &= f_0 + \lambda f_1 + \lambda^2 f_2 + \dots \\ P &= P_0 + \lambda P_1 + \lambda^2 P_2 + \dots \\ T &= T_0 + \lambda T_1 + \lambda^2 T_2 + \dots \end{aligned} \quad (5)$$

To lowest order, a deceptively simple equation appears for  $f_0$ : the so-called Vlasov, or collisionless Boltzmann equation

$$\left\{ \frac{\partial}{\partial t} + \vec{v} \cdot \frac{\partial}{\partial \vec{x}} + \frac{e}{m} \left[ \vec{E} + \frac{\vec{v}}{c} \times \vec{B} \right] \cdot \frac{\partial}{\partial \vec{v}} \right\} f_0 = 0 \quad (6)$$

where

$$\vec{E} = ne \int \left\{ \frac{\partial}{\partial \vec{x}} \frac{1}{|\vec{x} - \vec{x}'|} \right\} f_0 (\vec{x}', t) d\vec{x}'$$

Note that all microscopic field fluctuations arising from the discrete nature of the particles disappear in this approximation.

This equation describes a plasma in the continuum limit where our chopping procedure has been carried to completion and particle individuality is lost. Note that many equilibrium solutions are possible, e.g. in the absence of a magnetic field any  $f(\vec{v})$  or in the presence of a uniform field any  $f(|\vec{v} \times \vec{B}|, (v \cdot \vec{B}))$ .

In the next approximation we obtain an equation for the pair correlation function  $P_0$  which in turn determines  $f_1$ , the "collisional" modification of  $f$ . The equation for  $P$  is very complicated but has been solved ingeniously by Rostoker in terms of a simpler test particle problem. The latter is defined by sending an infinitely massive test charge,  $\rho = \delta(\vec{x} - \vec{x}_0 - \vec{v}_0 t)$  into a Vlasov plasma described by Eq.(6). This poses a quite tractable problem for the shielding cloud  $G(\vec{X} | \vec{X}', t)$  which forms around the test particle. Rostoker's superposition principle states that the pair correlation function may be found by a superposition of test particle solutions, i.e. that

$$\begin{aligned} P(\vec{X} | \vec{X}', t) &= f(\vec{X}, t) G(\vec{X} | \vec{X}', t) + f(\vec{X}', t) G(\vec{X}' | \vec{X}, t) \\ &\quad + n \int d\vec{X}'' f(\vec{X}'', t) G(\vec{X}'' | \vec{X}, t) G(\vec{X}'' | \vec{X}', t) \end{aligned} \quad (7)$$

This says that the pair correlation between particle  $X$  and particle  $X'$  is given by the sum of the correlations arising, first, from  $X$  being the test particle and  $X'$  being the field particle, second,  $X'$  being the test particle and  $X$  being the field particle, and third, from both being field particles around a test particle  $X''$ . This completes the solution in zeroth and first orders of the plasma many-body problem.

We may note here several relevant properties of this solution. First, the correlation term  $f_1$  corresponds to the addition to the Vlasov equation (Eq.(6)) of terms corresponding to a generalized Fokker-Planck collision operator. While an ad hoc small impact parameter cut-off is still required in the Coulomb logarithm, due to neglect of quantum effects, the large impact parameter is correctly given by the Debye shielding cloud which forms around the test particle. In addition, this cloud has elements corresponding to plasma wave emission by the test particle which con-

tribute modestly to the collision operator. Needless to say, the Boltzmann distribution is the equilibrium solution on the long collisional time scale.

A rather interesting experimental check on this picture of the plasma as made up of shielded freely moving particles comes from radar back-scatter off-electrons in the ionosphere [2]. Here the experimental results were at first quite surprising since they showed that the reflected waves had a Doppler broadening characteristic of the ion (rather than the electron) thermal spread. This can be easily understood in terms of the shielded test particle picture described above, since the waves scatter from electron density fluctuations. As the radar wavelength is long compared to the shielding cloud dimensions (a Debye length  $v_e/\omega_p$ ) it scatters from the net number of electrons within the cloud. As electrons move very rapidly compared to ions, the ions do not shield the electrons. Hence, a shielded electron consists of an electron test particle and the absence of an electron in the cloud. Hence, it has no total electron number and will not scatter the wave. An ion test particle, on the other hand, is shielded by the absence of half an ion and the presence of half an electron as both species contribute to shielding the slow ion. Hence, to the radar the test ion looks like half an electron thus explaining the experimental Doppler broadening. Other detailed predictions [3] of the theory, such as small scattering by plasma waves, have been observed.

While this generalized Fokker-Planck equation is almost certainly adequate for the discussion of collisionless plasmas, there still remain open questions of interest and possible generalizations. For example, Sandri has shown that logarithmic divergences occur in higher order so that the basic convergence has not been proven. Oberman and Rogister have shown, particularly for plasmas with unstable or weakly stable distributions, how the formalism of the many-body problem links into the formalism of non-linear wave studies.

We may also note that as the gravitational force is much like the Coulomb force we might expect galactic dynamics to be governed by a similar many-body theory regarding the stars as "electrons" or "ions". One difference arises from the fact that there is only one type of particle, with attractive forces. Another difficulty is that galactic dimensions, as determined by the Jeans criterion for gravitational equilibrium, are of the order of the "Debye length". Hence justification for use of the Vlasov or Fokker-Planck equation is unclear.

Finally, we point out that a similar derivation by way of the density matrix for  $f(p, q)$  has been given in the quantum case by Klimontovich. Again the principal difficulty for solid-state considerations would appear to be the smallness of  $n\lambda_D^3$ .

Creeping up on the topic of plasma excitations, let us consider next the subject of linearized plasma waves in an infinite homogeneous magnetized plasma. We proceed from the linearized Vlasov equation [4] putting

$$\frac{df_1}{dt} = \frac{e}{m} \left( \vec{E}_1 + \frac{\vec{v}}{c} \times \vec{B}_1 \right) \cdot \frac{\partial}{\partial \vec{v}} f_0 \quad (8)$$

where  $f_1$  is a linearized perturbation  $\exp[i(\omega t - \vec{k} \cdot \vec{X})]$ ,  $f_0$  is the equilibrium and the total derivative means the rate of change along the orbit given by

the unperturbed fields. In the case of an infinite homogeneous plasma with or without a DC magnetic field the integrals may be readily performed, the resulting charges and currents  $\rho = e \int f_1 d^3 \vec{v}$ ,  $\vec{j} = e \int f_1 \vec{v} d^3 \vec{v}$  calculated, and Maxwell's equations then yield a dispersion relation

$$\vec{E}(\vec{k}, \omega) \cdot \vec{E}_1 = 0 \quad (9)$$

We will not pursue the algebra further here, but before discussing the basic kinds of excitations let us remark on one mathematical peculiarity. It is apparent that when we perform the integration indicated in Eq.(8), factors like  $(\omega - \vec{k} \cdot \vec{v})$  will appear and provide singular denominators for the velocity integration which determines the currents. Their interpretation was provided by Landau [5], who pointed out that if one solved an initial value problem rather than a normal mode problem, one could see that the denominators correspond to a resonant interaction between the wave and particles moving at the same phase velocity as the wave. This interaction leads to the so-called Landau damping of the waves. Actually, this "damping" may also correspond to growth if the distribution function has a positive slope. The existence of Landau damping has been confirmed in a remarkable series of experiments by Malmberg, O'Neill, Wharton and Gould. In fact, they have further demonstrated that the beams of resonant particles which persist after the fields have decayed allow for the formation of plasma wave echoes, much like the spin echoes observed in solid-state physics. The Landau damping of plasma waves has also been observed in solid-state plasmas where it is quite traumatic due to the sharp edge on the Fermi distribution.

Even the simple dispersion relation (Eq.(9)) for a uniform infinite plasma gives rise to a great multiplicity of waves. Among these we mention the following:

(1) If a plane layer of charge is displaced a distance  $\vec{x}$  from its equilibrium position, a charge density  $n_e$  and a restoring electric field  $4\pi n_e \vec{x}$  arise. This leads to the equation of motion

$$m \ddot{\vec{x}} = -4\pi n_e e^2 \vec{x}$$

and to plasma oscillations at the frequency  $\omega_p = (4\pi n_e e^2/m)^{1/2}$ . A more detailed treatment leads to wave dispersion and Landau damping as discussed above.

(2) The current response to a high-frequency transverse wave yields the approximate result

$$\omega^2 = \omega_{pe}^2 + k^2 c^2 \quad (10)$$

showing that electromagnetic waves with a frequency below the plasma frequency cannot propagate - the basis of ionospheric plasma studies.

(3) In the presence of a uniform magnetic field resonant denominators  $(\omega - nw_c)^{-1}$  arise which drastically modify the plasma dispersion function. In particular, they lead to a great number of "windows" and anomalous waves lying near cyclotron harmonics.

(4) A very interesting regime is obtained in the low-frequency, long wavelength limit, i.e.  $\omega \ll \omega_c, \omega_p$ ;  $k \frac{v}{\omega_c} \ll 1$ . This is the magnetohydrodynamic (MHD) limit which we discuss in some detail for general geometries not restricting ourselves to infinite homogeneous plasmas. An approximate treatment [6] in this regime is given by taking velocity moments of the Vlasov equation (6).

Number conservation yields

$$\frac{\partial n_j}{\partial t} + \vec{\nabla} \cdot \left\{ n_j \vec{u}_j \right\} = 0 \quad (11)$$

where  $\vec{u}$  is the mean velocity and  $j$  indicates the species (ions or electrons), while momentum conservation gives

$$\rho \left[ \frac{\partial \vec{u}}{\partial t} + \vec{u} \cdot \vec{\nabla} \vec{u} \right] + \vec{\nabla} \cdot \vec{P} - ne \left[ \vec{E} + \frac{\vec{u}}{c} \times \vec{B} \right] = 0 \quad (12)$$

Here we note that the equation for the first velocity moment  $\vec{u}$  involves the second moment  $\vec{P}$  and evidently the moment procedure will never rigorously close. However, we are able to obtain approximate closure by taking advantage of the special regime in which we are interested. The large cyclotron frequency, small gyroradius limit corresponds formally to treating  $1/e$  as a small expansion parameter. This is a tricky point as the MHD equations are recovered by treating both frequency and wave number as first order small. The validity of such formal ordering scheme rests on no firm basis and indeed other orderings yield different sorts of waves. Poisson's equation then tells us that  $n_i = n_e + O(1/e)$  and, moreover, the momentum conservation equation for each species individually requires Ohm's law

$$\vec{E} + \frac{\vec{u}}{c} \times \vec{B} = 0 \quad (13)$$

i.e. the collisionless plasma is a very good conductor in whose rest frame the electric field must nearly vanish. Physically this means that the plasma remains "frozen to magnetic field lines"; as the field lines move due to the Faraday law the plasma moves with them.

A detailed look at the microscopic behaviour of the plasma shows that Ohm's law follows from the fact that in crossed electric and magnetic fields, particle orbit guiding centres drift according to the rule

$$\vec{u} = c \frac{\vec{E} \times \vec{B}}{B^2}$$

Adding the momentum equation for ions and electrons gives finally the pressure balance law

$$\rho \left[ \frac{\partial \vec{u}}{\partial t} + \vec{u} \cdot \vec{\nabla} \vec{u} \right] + \vec{\nabla} \cdot \vec{P} = \vec{j} \times \vec{B} \quad (14)$$

where the nature of the spiralling particle orbits tells us further that

$$\vec{P} = p_{\perp} \vec{I} + (p_{\parallel} - p_{\perp}) \frac{\vec{B} \vec{B}}{B^2} \quad (15)$$

For closure we still require a rule for the time rate of change of the components of the pressure tensor. The correct law is rather complicated even in this limit but for the special case of flow only in the plane perpendicular to  $\vec{B}$  the two-dimensional adiabatic law

$$\frac{d}{dt} (p \rho^{-2}) = 0 \quad (16)$$

is correct and an adequate indication for more complicated flows. It will be observed that the Faraday induction law

$$\frac{\partial \vec{B}}{\partial t} = \vec{\nabla} \times (\vec{u} \times \vec{B}) \quad (17)$$

indicates that the magnetic pressure  $B^2/8\pi$  obeys an adiabatic law similar to Eq.(16). Moreover, just as the pressure tensor is now anisotropic, the magnetic force term

$$\vec{j} \times \vec{B} = \vec{B} \cdot \vec{\nabla} \vec{B} - \vec{\nabla} \left( \frac{B^2}{2} \right) \quad (18)$$

displays an anisotropic nature. However, except for these anisotropies, Eqs (11) to (18) describe a system much like the ordinary fluid dynamical equations, valid only in the restricted regions of frequency and wavelength I have mentioned.

Two principal characteristic types of waves emerge from these equations:

Torsional waves in which the first term in Eq.(18) is dominant, corresponding to a tension in the magnetic field lines. This leads to waves propagating along the magnetic field with characteristic velocity  $\omega/k_{\parallel} = B/\sqrt{4\pi\rho}$ , the Alfvén speed.

Compressional waves roughly perpendicular to  $\vec{B}$  where the second term in Eq.(18) is dominant. These act much like the sound waves of ordinary hydrodynamics. It is worth noting the force  $-\nabla B^2/2$  which tells us that plasma may be pushed into regions of low magnetic field.

Just as fluid dynamics, especially in the turbulent regime, is by no means a completely solved subject, so also the MHD equations are not at all fully explored and may indeed be of predominant importance for many astrophysical applications, where the relevant distances and times are so large that only these macroscopic modes are of great importance. Nonetheless it is the more subtle aspects and subtle types of solutions of the Vlasov equation which are of prime interest to plasma physics

aficionados. It is also these subtle microinstabilities which pose the primary threat to the possibility of stable plasma confinement and thermonuclear fusion.

Returning to our basic Eq.(8) it is apparent that a deeper understanding of the nature of plasma waves depends on a good understanding of particle orbits in varying electric and magnetic fields. As we may hope that at least the equilibrium is slowly varying in space and time, a detailed orbit theory in this adiabatic limit,  $\frac{v}{\omega_{ci}} \frac{\nabla B}{B} < 1$ ,  $\omega \ll \omega_c$ , is called for and has been provided by Kruskal [7] who notes that the particle orbit may be considered as composed of three hierarchies of motion characterized by very different time scales

- (a) particle gyro-spiralling  $t \sim 1/\omega_c$
  - (b) particle motion along field lines  $t \sim L / V$
  - (c) particle drifts due to field gradients  $t \sim \omega_c L^2 / V^2$
- (19)

where  $L$  is a characteristic distance of the system in which the fields change.

The nature of the particle drifts is easily seen by referring to the picture of a particle spiralling in an inhomogeneous magnetic field

$$\downarrow \vec{\nabla} B + \text{curly arrow} \odot B$$

with

$$\vec{v}_D = \frac{m v_\perp^2}{eB^4} \vec{B} \times \vec{\nabla} \frac{B^2}{4} + \frac{m v_\parallel^2}{eB^4} \vec{B} \times (\vec{B} \cdot \vec{\nabla}) \vec{B} \quad (20)$$

As it is clearly de rigueur at this symposium to mention the name of Einstein, I will point out that Kruskal's method of solution goes back to a famous solution of the problem of a pendulum of slowly varying length alluded to by Einstein at a Solvay congress. Einstein pointed out that there existed an adiabatic invariant for this problem, namely, the energy divided by the frequency. Similarly, corresponding to the three particle frequencies of oscillation, there exist three adiabatic invariants

$$\text{the magnetic moment} \approx \frac{E_\perp}{\omega_c} \approx \frac{mv_\perp^2}{2\omega_c} \sim \frac{v_\perp^2}{B}$$

$$\text{the action integral } J = \oint v_\parallel dl \sim \oint \sqrt{E - \mu B} dl \quad (21)$$

the total flux enclosed by the drifting orbit

These adiabatic invariants are effectively constants of the motion for any motion which occurs more slowly than their respective time scales. Thus the magnetic moment is conserved for all perturbations at frequency less than the particle gyrfrequencies.

Using these adiabatic invariants, a detailed theory of particle orbits may be derived. It should be pointed out that the existence of such adia-

batic invariants occurs in many branches of physics, and their existence is necessary for the correspondence principle to hold.

In principle, then, it should be possible to solve Eq.(8) for any possible linearized mode around a given equilibrium and determine, for example, the stability of such an equilibrium. In practice, given a possibly rather complicated distribution function  $f_0$ , the intricate nature of the orbits, and the great freedom allowed by the Vlasov equation, this is a rather hopeless task, and one must seek further insight into the relevant factors determining particle dynamics.

From this point on, then, I will concentrate on the question which to me has been the primary fascination in the study of plasma physics: can one find a stable confined plasma? Here "confined" is defined to mean finite in extent and with pressure vanishing at all walls. This is evidently a crucial question for fusion possibilities, but it also presents us with a very concrete challenge regarding our understanding of all possible dynamical plasma modes. When we understand them, we shall probably be able to control them by proper field configuration, either eliminating instabilities completely, or at least reducing them to a harmless mean turbulence level.

This is a complicated and detailed subject and I trust the reader will bear in mind that even more than heretofore I will speak in oversimplified, incomplete half-truths.

We know immediately that no true thermodynamic equilibrium plasma confinement can exist, since in that case  $f \sim \exp(-H/kT) \sim \exp(-e\phi/kT + mv^2/2kT)$ . The potential can only confine one species of particle and we see that the Hamiltonian does not depend on the static magnetic field. Thus, at high densities, where the plasma must be neutral, the only possible thermodynamic equilibrium is uniform in space, hence not confined. However, under high-temperature low-density conditions, collisions are very rare as is collisional diffusion across the magnetic field and we may, therefore, legitimately restrict our question to the collisionless regime where confined collisionless equilibrium solutions of the Vlasov equation can easily be seen to exist and the crucial question is whether, flying almost in the face of conventional thermodynamics, such equilibria are stable.

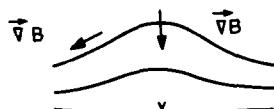


FIG.1. Mirror confinement scheme.

We turn first to the question of possible confined equilibria. Such equilibria can evidently be found by superposing particle orbits in the given static field and then choosing orbits such that in addition charge neutrality obtains. It is clear that any static magnetic field will confine a charged particle in the two dimensions perpendicular to  $B$  and on the short time scale (19a). Next arises the question of confining it on the second time scale, i.e. how to confine its motion along magnetic field lines. Here there exist two separate topological classes of possible equilibria - open-ended and closed confinement schemes. The first is shown schematically in Fig.1, where field lines are shown around an axis of symmetry. The centre of the plasma confinement region (indicated by an X) is a minmax point of  $B$ , the field increasing as we move

longitudinally but decreasing as we move radially outwards. Using the constancy of the magnetic moment and of energy we see that

$$v_{\parallel}^2 = \frac{2}{m} (E - \mu B) \quad (22)$$

$B$  increases as the particle moves along the field line and hence its velocity parallel to  $B$  will decrease. If its magnetic moment is sufficiently large,  $v_{\parallel}$  will vanish and the particle be reflected back to the centre. Hence parallel confinement is achieved by the magnetic mirror. Note that not all particles will be confined. If we plot in velocity space near the mid-plane (Fig.2) we see that Eq.(22) splits up velocity space as shown into a region where the magnetic moment is sufficient for particle confinement and a "loss cone" where it is not. Hence any equilibrium in an open-ended system is perforce anisotropic. Realization of the possibility of such a particle confinement geometry led in fact to a prediction of the existence of the Van Allen trapped radiation belts prior to their discovery. Finally, we must enquire as to confinement on the long time scale (19c). We may note that in an axisymmetric system the drifts are all in the azimuthal direction so that the orbit is confined or, alternatively, that the constancy of  $J = \oint \sqrt{E - \mu B} dl$  restricts the particle to orbiting only along a certain set of flux lines, namely those which differ only in azimuthal angle. The argument concerning the constancy of  $J$  of course is not restricted to axisymmetric systems.

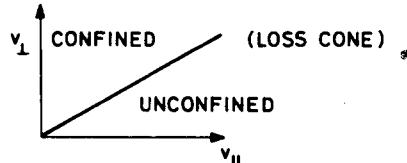


FIG.2.  $v_{\perp}$  versus  $v_{\parallel}$ .

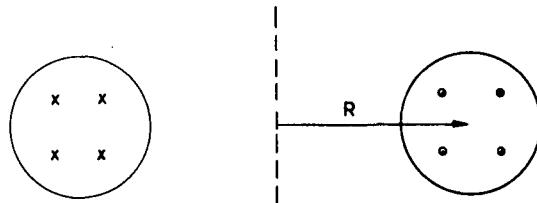


FIG.3. Toroidal confinement.

We turn now to the other possible topology for confinement, the toroidal or closed scheme (Fig.3). Here longitudinal confinement is obtained by returning the field lines on themselves by wrapping them around an axis of symmetry. Note that  $B = B_{\varphi} \approx 1/R$ . Now, however, when we examine confinement on the drift time scale we see that due to the radial gradient of  $B$ , ions will drift upwards and electrons downwards and confinement cannot be achieved. This can be remedied [8] by wrapping

helical coils around the torus whose field introduces a rotational transform into the field as shown in Fig.4. Here, if we trace a field line around the torus starting from position 1, we find that after one transit it has rotated by an angle  $i$ , and does so on each subsequent transit thus tracing out a nearly ergodic flux surface. The particle's longitudinal motion now carries it along the flux surface and we note that its continual upward drift now corresponds for half the cycle to a radially inwards motion and for the other half to a radially outwards motion. It is in fact not hard to show that the drifts exactly cancel and the orbit remains confined (Fig.5). Hence, equilibria can exist with plasma density varying from flux surface to flux surface.

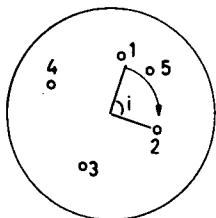


FIG.4. Rotational transform.

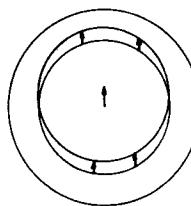


FIG.5. Compensating drifts.

We turn now to the stability of these confinement schemes and it is natural to look first at the question of magnetohydrodynamic stability. Here we see an immediate advantage for the sheared toroidal system with its ergodic flux surfaces. To see this we return to Ohm's law (Eq.(13)) and consider a situation in which we have only a small amount of confined plasma, i.e. for which  $p \ll B^2/8\pi$ . In this case it is clear that any perturbation which disturbs the magnetic field must raise the field energy (which is a minimum being determined by external coils) by an amount which is large compared to the small reservoir of plasma energy. Hence we make the electrostatic approximation  $\vec{E} = -\nabla\phi$  for the perturbation, and note that Eq.(13) tells us that  $\vec{B} \cdot \nabla\phi = 0$  which means the potential is constant over an ergodic flux surface. This excludes the possibility of any radial motions developing and hence no instability can occur. Note that this MHD stability rests on fairly weak ground. In particular, any small correction terms to Ohm's law might allow potentials with long wavelengths along the magnetic field to develop. We see later that such drift waves are indeed possible. Moreover, as the plasma pressure builds up, instability also becomes possible and violent MHD instabilities have been observed in pinch discharges.

Returning to the open topology, we see that now there is no inhibition against developing perturbation potentials which are constant along field lines but vary azimuthally  $\sim \exp(im\theta)$ . These will transport plasma radially and whether they are stable or unstable depends on whether the plasma energy increases or decreases by this convection. The answer to this question may be guessed from the earlier remark that plasmas tend to be forced out of regions of high  $|B|$ . Alternatively, one notes that the constancy of  $v_\perp^2/B$  indicates that if the bulk of the plasma can be moved to a region of lower field strength its energy will be lowered and instability will ensue. As we have remarked earlier, in a simple

mirror geometry of the type under discussion the field strength does decrease on the average with radius and we might expect instability to ensue. These flute instabilities have indeed been seen in mirror experiments and lead to violent plasma loss. An easy cure exists however. If we add current-carrying bars as shown, running roughly along the original field lines, then they produce a poloidal field which adds to the basic mirror field. This poloidal field evidently increases outwards and, if it is strong enough, can make a net outward increase of field intensity. In fact, the centre of the plasma can now be made a point of true minimum  $B$ . In this case we see that any plasma motion which conserves the magnetic moment and which displaces the bulk of the plasma outwards must increase its energy. Hence the plasma should be stable.

This was dramatically demonstrated in classic experiments [9] by Ioffe in 1962 who showed that as the currents in these Ioffe bars (Fig.6) increased by about 10% through the point where a minimum  $B$  was realized, the plasma immediately quieted down and its lifetime increased by a factor of  $10^4$ . Note that this minimum  $B$  hydrodynamic stability is based on firm energetic grounds. In any event, MHD instabilities seem both understood and curable and we must now consider the bewildering array of other modes which become possible when we try to go beyond these simple approximations.

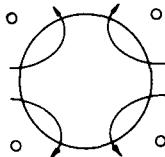


FIG.6. Ioffe bars.

Fortunately, an important simplification may be made from quasi-thermodynamic considerations. Pursuing an approach originated by Fowler [10], let us consider the free energy sources available in a confined plasma - the distinction between energy and free energy being that constraints on the motion, e.g. Liouville's theorem, conservation of adiabatic invariants, Ohm's law, etc., may not permit all the plasma energy content to be tapped. Basically there are three ways in which a confined plasma differs from thermodynamic equilibrium and hence three potential sources of plasma free energy. These are:

- (1) magnetic free energy arising from distortion of the confining field;
  - (2) anisotropy free energy arising from the deviation of the velocity distribution function from isotropy. As previously mentioned, this is an inevitable concomitant of open confinement schemes;
  - (3) expansion free energy arising from the localization of the plasma.
- (1) With regard to the magnetic free energy, we note that as the plasma currents modify the confining fields and as the unmodified fields represent the minimum magnetic energy state compatible with the confining coils, then indeed the plasma distortion of the field must represent an increase of energy available as free energy for the expulsion of the plasma.

We note, however, that the magnetic free energy is quadratic in the plasma density, or in the dimensionless quantity  $\beta \sim 8\pi p/B^2$ . To see

this, consider that the plasma currents and hence the field distortion are both linear in  $\beta$ . Hence when the plasma is expelled the induction electric field is of order  $\beta$  as is the current and hence the work done and the energy liberated are of order  $\beta^2$ . We will content ourselves with the low  $\beta$  limit in which the magnetic free energy is small compared to the other sources. We have already noted that in this limit only electrostatic perturbations can be unstable.

(2) The magnitude of the anisotropy free energy obviously depends on the system being considered. For a magnetic mirror system with a quite anisotropic distribution we might expect that the free energy per cubic centimetre is of the order of the energy density. In a toroidal system the distribution may be nearly isotropic and we find only residual free energy sources of this type perhaps of the order  $(a_i/L)^2 p$ . An important point to be made here is that in order to isotropize the distribution function it is evidently necessary to destroy the invariance of the magnetic moment. This in turn implies a high frequency perturbation of frequency greater than or equal to the ion gyrofrequency with wavelength less than or comparable to a gyroradius.

(3) In considering expansion free energy we must realize that several categories of constraints on the motion may exist. These are weak constraints such as Ohm's law which often rules out pure hydrodynamic modes but may allow for more subtle modes. However, one exact constraint is provided by Liouville's equation, telling us that volume in phase space is conserved. This implies a law of the type  $E\rho^{1-\gamma} \approx \text{constant}$ , since a lowering of the energy implies a consequent increase of volume. Consider now the release of expansion free energy by an instability which has the effect of smoothing out an inhomogeneous pressure profile as shown in Fig. 7, i.e. a homogenizing over a distance  $\Delta r < L$ , the characteristic density fall-off distance.

The adiabatic law now tells us that

$$\Delta E \sim p \left( \frac{\Delta r}{L} \right)^2 \quad (23)$$

This free energy must also be supplying the energy of motion for the instability, i.e.

$$\rho u^2 < p \left( \frac{\Delta r}{L} \right)^2$$

On the other hand, if we are considering a mode of frequency  $\omega$ , it must have an amplitude  $\sim \Delta r$  in order to liberate the free energy and hence we conclude that

$$\rho \omega^2 (\Delta r)^2 < p \left( \frac{\Delta r}{L} \right)^2$$

or

$$\omega^2 < \frac{v_i^2}{L^2} \ll \omega_{ci}^2 \quad (24)$$

In other words, expansion free energy can only be liberated by slow motions. The second inequality in Eq.(24) results from the assumption of a system which is many gyroradii in dimension.

We may summarize this discussion as follows:

Open-ended minimum B systems are completely stable against low-frequency motions which conserve the magnetic moment. Hence expansion free energy is not available in such systems. On the other hand, anisotropy free energy is always present because of the loss cone in velocity space. This may, however, be tapped only by high-frequency short wavelength disturbances. This suggests that the methods of geometrical optics may be used where the infinite medium dispersion relation applies locally and so a modified WKB method is used to treat an actual finite geometry. This may still be quite complicated, involving complex dispersion relations in highly anisotropic media.

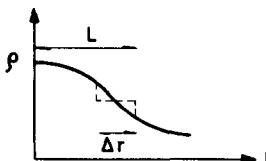


FIG. 7. Smoothing out an inhomogeneous pressure profile.

Toroidal systems on the other hand may have a nearly isotropic pressure in which case little anisotropy free energy is available but it is then impossible to rule out expansion free energy as a possible source of instability. However, such expansion free energy can only be liberated at low frequencies. Here a much simplified form of the Vlasov equation may be obtained by appropriately averaging over the fast particle gyration. While far more tractable than the original equation, the so-called drift kinetic equations are still formidable and must be solved in rather complicated geometries.

Of course, it must be borne in mind that even though the free energy for instability is present, there may still be other constraints which do not allow an actual unstable mode to develop. Hence we must finally come back to a normal mode analysis, although now on the basis of the simplified equations we have discussed above.

While it is beyond the scope of this paper to attempt a complete discussion of these complex questions, perhaps a brief discussion is appropriate. For the open-ended systems the anisotropy in velocity space leads to a version of the well-known two-stream instability [11]. Thus, if we consider the one-dimensional velocity distribution  $g = \int f(v_x^2 + v_y^2, v_z^2) dv_x dv_z$  it is easy to see that because of  $f(0, v_z^2) \equiv 0$  (the loss-cone condition)  $g$  must have a shape as shown in Fig.8. Thus for waves of an appropriate phase velocity we have a positive slope to the distribution and inverse Landau damping, i.e. instability may occur. On further detailed examination it appears that these waves should occur at moderate densities near harmonics of the ion cyclotron frequency. Moreover, they tend to be convective, running along the magnetic field lines, an effect that is aggravated for actual finite geometries by field and density variations. While the infinite medium instabilities are well

understood, it is the role of this convection which will give actual critical parameters for possible stable confinement schemes and this remains the principal problem to be studied in this area. Typically, it appears that lengths greater than several hundreds of gyroradii will lead to instability even for minimally anisotropic distributions.

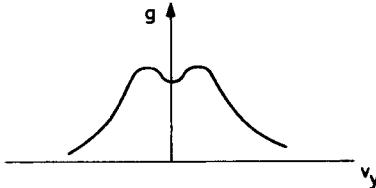


FIG. 8. Shape of  $g$  (versus  $v_y$ ).

For closed systems we have been led to look for non-hydrodynamic low-frequency waves [12]. If we consider an electrostatic disturbance with

$$v_{\text{the}} > \frac{\omega}{k_n} > v_{\text{thi}}$$

then the response of the slow-moving ions will be to move across the magnetic field with the  $E \times B/B^2$  drift while the faster electrons average out the drift and assume a Boltzmann distribution along field lines. This leads to the characteristic drift waves with

$$\omega \approx k_i a_i v_{\text{thi}} \frac{1}{n} \frac{dn}{dr} \quad (25)$$

A formal ordering of the Vlasov equation may be used to study this regime more precisely. Here we set  $\omega$  and  $k_n$  as of second order in the smallness parameter  $a_i/L$ . It is for this reason that the MHD equations which are derived with a different ordering do not yield these waves. It is further found that a small destabilizing Landau damping of these waves may occur  $\sim (k_i^2 \alpha_i^2)$ . We see that these waves are characterized by short wavelengths  $\sim a_i$  across the lines of force and rather long wavelengths along the lines of force. Due to their rather short perpendicular wavelengths we could expect them to remain of fairly small amplitude and lead to a kind of turbulence in the plasma. If we very crudely estimate the diffusion due to such turbulence, putting  $D \approx \gamma/k_i^2$ , we find under the worst conditions a predicted diffusion

$$D = \frac{c T_e}{e B} \quad (26)$$

a formula predicted by Bohm on dimensional grounds 20 years ago. (It can be seen that if we suppose that  $D \sim 1/e$  in accordance with our ordering on dimensional grounds, the Bohm result follows.) It might be noted that this diffusion is several orders of magnitude greater than

can be tolerated in a fusion reactor, although often of negligible interest astrophysically. However, when the details of a realistic geometry are considered, including the effects of particles trapped along the field lines, we find considerable modification of these results. In particular by introducing strong shear ( $di/dr$ ) it becomes difficult for the long wavelength drift waves to occur and considerable reduction in diffusion is predicted. Placing the plasma in a favourable magnetic well, i.e. having  $|B|$  generally larger on outer flux surfaces is also helpful, as discussed earlier (see Eq.(18)). Again much remains to be done on the matter of convection and localization of these unstable waves, especially in the presence of shear.

Recent experiments confirm the general outlines of the above theoretical outlook. Hydrodynamic instabilities have been suppressed. However, open-ended confinement systems tend to produce instability bursts with consequent particle loss at multiples of the ion gyrofrequency, while low-frequency turbulence with the properties of the drift waves discussed above is seen in toroidal systems as is anomalous diffusion at nearly the Bohm rate. As the experiments have been modified in directions indicated for theoretical stabilization some control over the instabilities has been achieved with consequent reduction in particle loss, e.g. diffusion of 0.1 Bohm has been achieved. However, such modifications are difficult and expensive and only a very rough approximation to production of plasma of known and theoretically desirable qualities is now possible. As a result it is too early to claim experimental proof either for the possibility of confinement or as confirmation of our theoretical ideas. We can only say that some qualitative improvement in confinement has occurred when proceeding in the direction predicted by theory.

Needless to say, a great limitation on our ability to interpret experiments, especially crude and unstable experiments, is the difficulty of developing a non-linear theory; this matter is dealt with by Dupr e  in these Proceedings.

#### REF E R E N C E S

- [1] ROSTOKER, N., ROSENBLUTH, M.N., Physics Fluids 3 (1960) 2.
- [2] BOWLES, K.L., Phys. Rev. Lett. 1 (1958) 454.
- [3] ROSENBLUTH, M.N., ROSTOKER, N., Physics Fluids 5 (1962) 776.
- [4] SIMON, A., in Plasma Physics, IAEA, Vienna (1965) 163.
- [5] LANDAU, L.D., J. Phys. USSR 10 (1946) 25.
- [6] OBERMAN, C., in Plasma Physics, IAEA, Vienna (1965) 103.
- [7] KRUSKAL, M., in Plasma Physics, IAEA, Vienna (1965) 91.
- [8] KRUSKAL, M., in Plasma Physics, IAEA, Vienna (1965) 115.
- [9] IOFFE, M.S., in Plasma Physics, IAEA, Vienna (1965) 421.
- [10] FOWLER, T.K., GUEST, G.E., (Proc. Conf. Culham, 1966), IAEA, Vienna (1966) 383.
- [11] POST, R.F., ROSENBLUTH, M.N., Physics Fluids 9 (1966) 730.
- [12] SACDEEV, R.Z., in Plasma Physics, IAEA, Vienna (1965) 555.

# NON-LINEAR PLASMA PHYSICS

T.H. DUPREE

Department of Nuclear Engineering,  
Massachusetts Institute of Technology,  
Cambridge, Mass., United States of America

## Abstract

NON-LINEAR PLASMA PHYSICS. 1. Nature of non equilibrium effects; 2. Linear theory; 3. Non-linear theory; 4. Wave-wave coupling; 5. Weak wave-particle interaction; 6. A single non-linear wave; 7. A few non-linear waves.

## 1. NATURE OF NON-EQUILIBRIUM EFFECTS

Plasma instabilities arising from collective motion frequently produce a plasma state whose properties are totally different than those at equilibrium. To illustrate the importance of non-linear collective processes due to plasma instabilities, we first consider the equilibrium situation [1, 2]. In this case, one may regard the plasma as being composed of quasi-particles and normal modes. A quasi-particle consists of a bare electron or ion plus a shielding cloud of other particle with a radius of the order of a Debye length  $\lambda_D$ .  $\lambda_D$  is equal to the average thermal velocity,  $v_T$ , divided by the plasma frequency  $\omega_p$ .  $\omega_p$  is equal to  $4\pi n q^2 / m$  where  $n$  is the average number density and  $q$  and  $m$  the charge and mass of a particle. At or near equilibrium the electric energy density per unit wave number interval,  $|E(k)|^2$ , due to the thermal fluctuation in normal modes and quasi-particle density, is shown in Fig. 1. The total electric energy density is

$$\langle E^2 \rangle = \int |E(k)|^2 dk = (1/2) nm v_T^2 \lambda_D^{-3} \quad (1)$$

The ratio of electric to kinetic energy density is

$$\frac{\langle E^2 \rangle}{nm v_T^2} = \frac{1}{n \lambda_D^3} \quad (2)$$

The dimensionless parameter  $1/n \lambda_D^3$  is the number of particles in a cube whose edge is a Debye length. As we shall see, this parameter characterizes the rate of relaxation and transport processes in a plasma due to the thermal fluctuations compared to the rate for collective processes.

For example, the velocity diffusion coefficient is given by:

$$D_v = \frac{\langle \Delta v^2 \rangle}{\Delta t} = \frac{q^2 \langle E^2 \rangle}{m^2} \omega_p^{-1} \quad (3)$$

where we have assumed the characteristic fluctuation time,  $\Delta t$ , of the electric field to be  $\omega_p^{-1}$ . The velocity relaxation rate (collision frequency) is given by

$$\nu_c = \frac{D_v}{v_T^2} = \frac{1}{n\lambda_D^3} \omega_p \quad (4)$$

The rate of particle diffusion across a magnetic field of strength  $B$  is given by

$$D_{\perp} = \frac{\langle \Delta v_{\perp}^2 \rangle}{\Delta t} = \frac{c^2 \langle E^2 \rangle}{B^2} \omega_p^{-1} \quad (5)$$

where we have assumed the cross-field speed is  $cE/B$  and the fluctuation time is again  $\omega_p^{-1}$ . Using Eqs (1) and (4) this becomes

$$D_{\perp} = \nu_c a_c^2 \quad (6)$$

where  $a_c$  is the cyclotron radius  $mv_T c/qB$ .

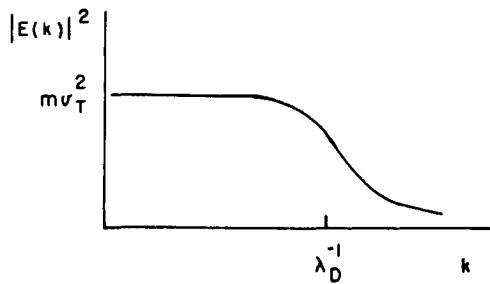


FIG. 1. The equilibrium electric energy density per unit wave number interval.

As a third example of an important transport process we consider the current density  $j$  induced by an applied constant electric field  $E_0$ . One can estimate this by assuming that a particle achieves a speed  $qE_0/m\nu_c$  between each collision, thus,

$$j = n \frac{qE_0}{m\nu_c} \quad (7)$$

In each case the transport or relaxation process depends on  $\nu_c$  which according to Eq.(4) is  $1/n\lambda_D^3$  times smaller than  $\omega_p$ . In a great many fully ionized plasmas of interest  $(n\lambda_D^3)^{-1}$  is a very small quantity. For instance, in the solar corona it is about  $10^{-8}$  and in proposed controlled thermonuclear devices it is of order  $10^{-6}$ . Thus we see that processes induced by thermal fluctuations are exceedingly slow compared to collective pro-

cesses which for the purpose of this illustration we may regard as having a rate  $\omega_p$ . More generally, the rate of collective processes is characterized by the frequency of the dominant linear instability.

It is a fact that many plasmas of interest are sufficiently far from thermal equilibrium so that formulas (4), (6), and (7) are not even qualitatively correct. This occurs because plasma instabilities cause  $\langle E^2 \rangle$  to be much larger than the equilibrium values shown in Fig. 1. To illustrate this point, we shall describe several dramatic experimental results.

In the first experiment [3] a 10 keV electron beam was directed at a low temperature plasma which was only 0.01 mean free paths thick, i.e.  $0.01 v_T/v_c$ . Accordingly, the thermal fluctuations should have produced no significant effects. In fact, however, periodic bursts of radiation at the plasma frequency were observed coupled with a large reduction in beam current out of the plasma. Also observed were 100 keV X-rays indicating the presence of particles whose energies were greatly in excess of that of either plasma or beam. Finally, the burst was accompanied by large increases in light emitted by the plasma indicating an overall plasma heating by the beam.

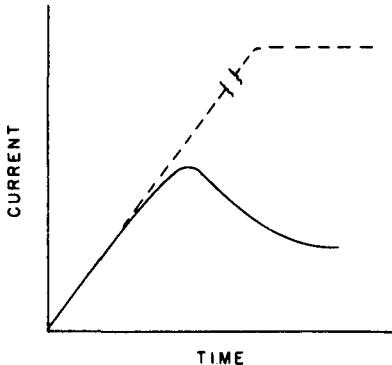


FIG. 2. The electron current as a function of time (from Ref. [5]).

In another recent experiment [4, 5] a constant electric field was applied to a plasma in which the ion temperature was much less than the electron temperature. From simple collision theory one would expect the ions and electrons to accelerate in opposite directions until the collisions impose a limiting current given by Eq.(7) and shown by the dotted line in Fig. 2. Actually, the current is limited at a much lower level as shown by the solid line in Fig. 2. The average electron speed is limited by the value at which the ion acoustic wave becomes unstable.

As a final example of the importance of non-linear collective effects we can cite the many laboratory confinement devices which have demonstrated, to the dismay of fusion researchers, that the diffusion rate across a magnetic field is in many cases much larger than that given by relation (6). Frequently, it is of the order of the so-called Bohm value:

$$D_{\perp} = a_c^2 \Omega_c \quad (8)$$

where  $\Omega_c$  is the cyclotron frequency.

## 2. LINEAR THEORY

We now consider the collective processes which arise from instabilities and which can dominate the collisional or thermal processes. The basic equation for the plasma distribution function  $f(\vec{r}, \vec{v}, t)$  is the Vlasov equation with a collision term:

$$\left( \frac{\partial}{\partial t} + \vec{v} \cdot \frac{\partial}{\partial \vec{r}} + \frac{q}{m} \vec{E} \cdot \frac{\partial}{\partial \vec{v}} \right) f(\vec{r}, \vec{v}, t) = \left( \frac{\partial f}{\partial t} \right)_{\text{collision}} \quad (9)$$

The right-hand side accounts for collisional effects. As discussed earlier,  $(\partial f / \partial t)_{\text{collision}}$  is usually assumed to be of order  $1/n\lambda_D^3$  times the order of the left-hand side of Eq.(9). It can therefore be neglected when collective effects dominate collisional effects. Maxwell's equations are used to determine  $\vec{E}$  (and  $\vec{B}$ ) in terms of the current and charge due to  $f$ .

By taking various velocity moments of Eq.(9) one can generate a set of moment equations which are adequate to describe the small amplitude (linear) development of certain (non-resonant) instabilities. However, the non-linear evolution may not be describable by the moment equation and for many instabilities even the linear theory requires the full Vlasov equation. This is an exceedingly complicated non-linear equation and only a few, simple exact solutions have been obtained.

Generally, the main effort has been made to obtain linearized solutions. Thus, one sets

$$f = \langle f \rangle + \tilde{f} \quad (10)$$

where  $\langle f \rangle$  is an average, slowly varying (in space and time) quantity, and  $\tilde{f}$  is a small perturbation describing a wave. One assumes  $\tilde{f} \ll \langle f \rangle$  and then neglects the term of order  $\langle f \rangle^2$  in Eq.(9). Putting

$$\tilde{f} \propto \exp(i\vec{k} \cdot \vec{r} - i\omega t) \quad (11)$$

we obtain in a well-known way a dispersion relation relating  $\omega$  and  $\vec{k}$ :

$$D(\vec{k}, \omega) = 0 \quad (12)$$

When  $\omega_k$  has a positive imaginary part for real  $\vec{k}$ , the wave is unstable and according to the linearized theory will grow forever at an exponential rate. The ultimate fate of such an instability must, of course, be determined by a non-linear theory.

The free energy that drives the instabilities arises from various types of inhomogeneities or anisotropies in the phase space distribution of plasma. Obviously, in a closed system the growth of the waves must be accompanied by a reduction in the inhomogeneity or anisotropy. Thus, it is clear even from the linear theory that plasma instabilities must lead to velocity relaxation and spatial transport. The exact account of such processes requires a non-linear theory but many aspects such as the general nature and approximate magnitude of the transport can often be deduced from the linear theory.

It is also an important experimental fact that many aspects of the non-linear state are frequently very similar to the predictions of linear theory. Thus, the spectrum of the non-linear state may contain just those frequencies and wave numbers predicted to be unstable by the linear

theory. Unlike the situation in high-Reynolds-number fluid turbulence, the non-linear forces usually do not destroy the linear mechanics of the waves. For these reasons it is easy to understand the primacy of linear theory in current theoretical plasma research. Furthermore, the extensive and complicated spectroscopy of linear plasma waves is far from complete, and there is not yet satisfactory agreement between theory and experiments.

### 3. NON-LINEAR THEORY

Like most non-linear theories, the plasma version [6, 7] consists of various and somewhat unrelated patches which are applicable in certain areas. Thus, one has a variety of non-linear problems, such as turbulence, shock waves, solitons, stochastic heating and acceleration, spatial transport and velocity relaxation, to mention a few. In many, but not all, non-linear problems, one is concerned with a state whose linear approximation is unstable. A linear instability leads to large wave amplitudes which then make non-linear effects important.

It is not surprising that many of the methods commonly employed in non-linear plasma problems are some form of perturbation theory. For example, the most popular theory is known as weak-turbulence theory. In this picture, one regards a turbulent plasma as a gas of a large number of random phase waves and particles interacting weakly with each other. The interaction can be calculated with a variety of methods which are all fundamentally equivalent. For instance, we could use the time-honoured methods of quantum-mechanical perturbation theory as if we were calculating phonon scattering. This leads to time-proportional transitions between wave and particle states. Or one can simply expand the solution of the Vlasov equation in powers of the electric field taking proper account of secular terms. Or one can generate a linked hierarchy of equations for the correlation functions  $\langle ff\dots \rangle$ . The solution of the truncated hierarchy of equations can be shown to have the same physical content as perturbation theory.

The term "weak-turbulence theory" may, in fact, be a misnomer since many people do not regard the state as being turbulent. In this connection, one can show that the validity of the theory rests on the wave-wave and wave-particle interaction time being short enough to ensure that no significant change in wave amplitude or particle orbit occurs during this time. This is definitely not the case in fluid turbulence.

Unfortunately, it is not clear that the conditions for the validity of weak-turbulence theory are generally realized in a great number of interesting plasma problems. For example, in a single finite-amplitude wave one has particles trapped in the potential troughs of the wave. The trapped particles generate half-integer powers of the potential in the particle distribution function – an effect which cannot be obtained from perturbation theory. The solution of this particular problem is obtained by transforming to the wave frame and using conservation of energy to find the particle orbits exactly. However, for most large-amplitude waves the proper theoretical approach is indeed obscure.

If there are only a very few waves interacting, the phases are not random and the appropriate theory for wave-wave interaction is essentially

that used in non-linear optics, or the theory of parametric amplifiers, or that of the van-der-Pol equation. As more waves are added to the problem, the recurrence time becomes essentially infinite, the phases become random, and one uses the asymptotic perturbation theory familiar from quantum mechanics.

For large wave amplitudes it appears that techniques which are reminiscent of those used in fluid turbulence theory may be appropriate. This means that one must compute wave and particle interactions by using the non-linear corrections to the unperturbed values of wave frequency and particle orbits. It is, in a certain sense, a renormalized theory. However, in most cases, the non-linear state of plasma does not appear to be so completely determined by non-linear processes as in high-Reynolds-number fluid turbulence.

A method of studying non-linear problems which is half-way between theory and experiments is the simulation of simple plasma configurations by computer experiments. In these experiments, one usually integrates the equation of motion for all particles and fields. Because one can control virtually every initial parameter in the problem, many general properties of the non-linear motion can be deduced. This technique is apparently becoming one of the principal means of studying non-linear problems.

As already mentioned, non-linear problems commonly develop as a result of a linear instability. In such cases the primary question is what mechanism ultimately limits the growth. At least four different processes have been discussed in the literature. They are: (a) The coupling of energy among the waves in order that energy from the linearly unstable portion of the spectrum will flow to the stable and damped part of the spectrum. This is the primary mechanism in fluid turbulence. (b) The free energy available to drive the instability may become exhausted. This can occur in the course of time in a closed system or in space as in the case of a beam impinging on the surface of a semi-infinite plasma. This mechanism forms the basis of the so-called quasi-linear theory. (c) Non-linear Landau damping. This is a wave-particle resonance of higher order than the linear theory which in principle can couple energy from waves back to particles. (d) Modification of the linear growth mechanism due to wave induced-particle scattering. In a rough way, this effect is similar to the eddy viscosity familiar from fluid turbulence. In some cases these effects are not distinct and tend to merge with each other. And, of course, when better theories are developed we shall probably find the list incomplete. It is worth-while to make a few specific remarks about some of the effects.

#### 4. WAVE-WAVE COUPLING

Some insight can be gained by comparing mode coupling in a fluid and a plasma. The Fourier transform of the Navier-Stokes equation for an incompressible fluid is

$$\left( \frac{\partial}{\partial t} + k^2 \nu \right) u_i(\vec{k}) = \sum_{j, l, \vec{k}'} M_{ijl} u_j(\vec{k}') u_l(\vec{k} - \vec{k}') \quad (13)$$

$\nu$  is the kinematic viscosity and  $\vec{u}$  is the macroscopic fluid velocity. The Reynolds number is the ratio of the non-linear to the linear rate of change of  $u$ :

$$R = \frac{Mu}{k^2 \nu} = \frac{u}{kv} \quad (14)$$

where we have assumed  $M$  to be of order  $k$ .

In the initial range where  $k^2$  is small,  $R$  can be very large, for example of order 2000. For such large Reynolds numbers, obviously one cannot solve Eq.(13) by treating the non-linear term as a perturbation. Various techniques have been devised which have been partly successful in solving Eq.(13), but the fluid turbulence problem is still considered to be in an unsatisfactory state. However, in the inertial range a simple dimensional argument due to Kolmogorov indicates that the energy spectrum should have a characteristic  $k^{-5/3}$  dependence.

The analogous mode coupling equation in a plasma is

$$\left( \frac{\partial}{\partial t} + \omega_{\vec{k}} \right) E_i(\vec{k}) = \sum_{j, l, \vec{k}'} C_{ijl} E_j(\vec{k}') E_l(\vec{k} - \vec{k}') \quad (15)$$

The Reynolds number for Eq.(15) is

$$R = \frac{cE}{\omega} \approx \frac{\bar{n}}{n} \quad (16)$$

$\bar{n}/n$  is the ratio of density fluctuation to average density. Actually,  $R$  is equal to  $\bar{n}/n$  only in the simplest cases. If  $\omega_{\vec{k}}$  is approximately real, and  $R \ll 1$ , the time-asymptotic solution of Eq.(15) shows that energy can be coupled among three waves only if

$$\omega_{\vec{k}} = \omega_{\vec{k}'} + \omega_{\vec{k}-\vec{k}'} \quad (17)$$

This is simply conservation of energy.

For  $R \approx 1$ , the wave lifetime is shortened because of non-linear interaction. Therefore, the spectral density of each wave is no longer proportional to  $\delta(\omega - \omega_{\vec{k}})$  but is broadened. Similarly, the constraint (17) need only be satisfied to within the broadening.

We can now point out some of the essential differences between fluid and plasma turbulence. First of all, since  $\omega_{\vec{k}}$  is determined by the linear dispersion relation and usually not "small" like  $k^2 \nu$ , the Reynolds number for a collisionless low- $\beta$  plasma is likely to be of the order of, or less than, unity and certainly much less than typical fluid values of 2000. Secondly, the frequency constraint (17) which is trivially satisfied for inertial range fluid turbulence where  $\omega = k^2 \nu \approx 0$ , severely restricts the plasma waves which can couple. For these two reasons, one would not expect plasma turbulence to resemble fluid turbulence.

For  $R \ll 1$  Eq.(15) can be solved by some form of perturbation theory. If there are many waves and a short coherent wave-wave interaction time then one can assume random phases and use weak turbulence theory. If there are only a few waves then the coherent interaction time is not short, the wave phase cannot be treated as random and one must use a van-der-Pol-type method.

There is ample experimental evidence that mode coupling occurs in plasma but no real evidence that it has provided stabilization in any given case. For example, type-II radio bursts from the solar corona are undoubtedly the coupling of two plasma waves to create an electromagnetic wave.

## 5. WEAK WAVE-PARTICLE INTERACTION

Another important difference between non-linear processes in fluids and plasmas is the importance of the microscopic level of plasma. Detailed velocity space structures cannot exist in a fluid because of the high collision rate, but can be very important in a plasma. Micro-instabilities in a plasma feed on the particle kinetic energy via the wave-particle interaction. These interactions cause the wave to grow initially in accordance with linear theory and cause an appropriate modification in the plasma distribution function  $\langle f(\vec{r}, \vec{v}, t) \rangle$  describable by what is called quasi-linear theory which is part of the weak-turbulence theory. Since the growth rates depend on  $\langle f(\vec{r}, \vec{v}, t) \rangle$ , growth may cease if the distribution function is sufficiently modified.

According to the quasi-linear theory [8, 9] the velocity scattering is described by the diffusion coefficient

$$D_v = \frac{q^2 \langle E^2 \rangle}{m^2} \tau_{wp} \quad (18)$$

In formula (18)  $\tau_{wp}$  is the auto-correlation time of the force as experienced by a particle as it moves through the plasma wave. If a particle moves at constant velocity (unperturbed) then  $\tau_{wp}^{-1}$  is approximately equal to the spectral spread in the Doppler shifted frequency

$$\tau_{wp}^{-1} \approx \Delta(\omega - kv) \quad (19)$$

For a wave-particle resonance to occur and thus for  $D_v$  to be non-zero a constraint such as

$$\omega - \vec{k} \cdot \vec{v} = 0 \quad (20)$$

must be satisfied for some wave in the spectrum. This is the wave-particle analogy to expression (17) and for real  $\omega$  simply expresses conservation of energy and momentum between wave and particle. In the simplest cases the growth (or damping) rate is proportional to the rate of change

with respect to energy of the population of particles which are resonant with the wave:

$$\text{Im } \omega \propto \frac{\partial \langle f \rangle}{\partial v^2} \quad (21)$$

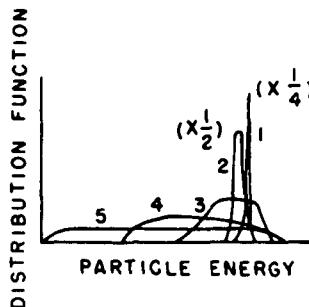


FIG. 3. The average velocity distribution at various stages in the temporal development of a two-stream instability (from Ref. [10]).

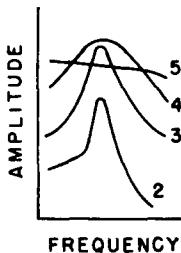


FIG. 4. The wave spectrum associated with Fig. 3 (from Ref. [10]).

A considerable number of calculations of  $D_v$  and its effect on the distribution function and wave growth in various geometries and for various types of waves are in the literature. There are also a much lesser number of relevant experiments which indicate that the quasi-linear theory or, at least, its general qualitative features are correct. For example [10], Fig. 3 shows the velocity distribution function at various stages in the temporal development of a two-stream instability. Curve 1 is the initial distribution function, etc. Figure 4 shows the corresponding wave spectrum as a function of frequency. As predicted by quasi-linear theory one finds that unstable waves develop in the region where  $\partial \langle f \rangle / \partial v^2 > 0$ . Velocity space flattens the beam distribution and moves it toward the velocity origin to conserve energy. Ultimately a quasi-steady state is reached in which  $\partial \langle f \rangle / \partial v^2 = 0$  over a wide region and a spectrum of waves remains whose phase velocities span the same region.

A more complicated instability known as the loss cone instability is depicted in Fig. 5. This distribution function is typical of mirror confinement devices. Particles can lose energy and cause waves to grow if they

diffuse into the loss cone along the dotted path on which  $\partial \langle f \rangle / \partial v^2 > 0$ . When (and if) the loss cone is filled the waves will cease growing.

The wave-particle interaction as described by  $D_v$  is also fundamental to other important plasma phenomena such as turbulent heating, enhanced thermalization, and stochastic acceleration.

Wave-particle resonances can occur at arbitrarily high order. For example, the second order version of expression (20) is

$$\omega_1 \pm \omega_2 - (\vec{k}_1 \pm \vec{k}_2) \cdot \vec{v} = 0 \quad (22)$$

This interaction exchanges energy between two waves and a group of resonant particles. The velocity diffusion coefficient will be proportional to  $E^4$  and therefore the wave growth or damping rates will be proportional to  $E^2$ . Generally speaking, it appears that higher-order wave and particle resonances are not very important because plasma Reynolds numbers are usually small.

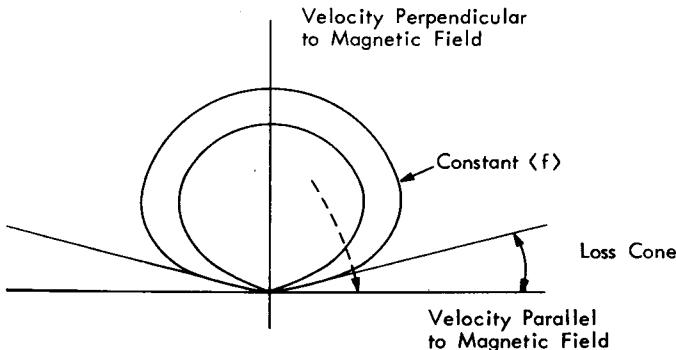


FIG. 5. Velocity distribution and diffusion path for the loss cone instability.

## 6. A SINGLE NON-LINEAR WAVE

As the amplitude of a wave is increased, new effects become important which are non-perturbative in nature. To illustrate this point, we consider the damping of a single coherent large amplitude wave. Figure 6 shows a recent experimental measurement [11] of wave damping as a function of time (a function of distance in the actual experiment). For small amplitudes (curve 1) the damping is exponential with a rate proportional to  $\partial \langle f \rangle / \partial v^2$  in accordance with the linear theory (21). However, for larger amplitudes (curve 2) the initial Landau damping gives way to a damped oscillation.

This behaviour, which was predicted theoretically [12, 13], is explained as follows. Particles whose energy in the wave frame is less than the wave potential,  $\phi$ , can be trapped in the troughs of the wave. The trapping criterion is

$$\left( v - \frac{\omega}{k} \right) < \left( \frac{q\phi}{m} \right)^{1/2} \quad (23)$$

This criterion is similar to expression (20) except that the resonance has been broadened by  $(q\phi/m)^{1/2}$ . The trapping occurs in a time  $\tau_{TR}$  of order

$$\tau_{TR} \approx \frac{\lambda}{(q\phi/m)^{1/2}} \quad (24)$$

$\tau_{TR}$  is the time required for the wave to cause a change in the particle orbit of one wavelength.

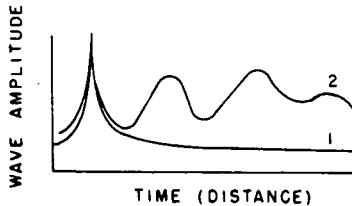


FIG. 6. Wave amplitude as function of time for a single finite amplitude wave (from Ref. [11]).

The change in wave amplitude must be equal to the negative of the change in resonant (trapped) particle energy. For  $t \gg \tau_{TR}$  all damping must stop since all resonant particles will be moving (on the average) at the wave speed. Thus for  $t \ll \tau_{TR}$  the linear theory is valid. For  $t > \tau_{TR}$  particle oscillations in the potential wells cause the amplitude to oscillate with a period  $\tau_{TR}$  as shown in curve 2. Since the wells are not parabolic, trapped particles have different oscillation times depending on their energy. Therefore for  $t \gg \tau_{TR}$ , the trapped population phase mixes to a steady state, the amplitude oscillation ceases, and a steady state is reached.

This steady-state mode has been predicted theoretically [14]. By properly specifying the trapped and untrapped particle distribution function, one can obtain virtually any sort of constant — in the wave frame — potential profile. In the linear picture this corresponds to superposing normal modes which all have the same phase speed. For example, one can have a pulse which is finite in spatial extent. If the upstream and downstream states are the same then such a pulse is usually called a soliton. If the two states are different, owing to particle reflection or trapping by the potential of the pulse, then the pulse can be regarded as a type of collisionless shock wave.

Pulses which have a constant profile are superpositions of plane waves which all move at the same speed. Therefore, the coherent interaction time between such plane waves is infinite. It follows that the interaction between these waves for a finite amplitude pulse cannot be calculated from perturbation theory. How such pulses develop naturally, and whether or not they are stable, are currently subjects of speculation. In analogy to compressional waves in a fluid, one can compute the steepening of wave fronts for waves which have no dispersion ( $\omega/k = \text{constant}$ ). However, for sufficiently small wave lengths, almost all plasma waves show dispersion. Dispersion will cause the shorter wavelength Fourier components to run ahead or lag behind the shock front. Therefore the steepness or thickness

of the front becomes limited by the minimum wavelength for no dispersion. In addition such waves may be either globally or locally unstable. A local instability in the front could produce small-scale turbulence and a dissipative mechanism which would play the same role as collisions in a fluid shock.

## 7. A FEW NON-LINEAR WAVES

We have seen some of the wave-particle aspects peculiar to large-amplitude wave forms. If, however, the individual Fourier components have sufficient dispersion, then they will move through each other and an incoherent or turbulent situation may develop. Because of the large amplitude, and the small number of waves, the weak-turbulence theory is inadequate. However, one can predict some of the new physical effects by comparing with the large-amplitude coherent case.

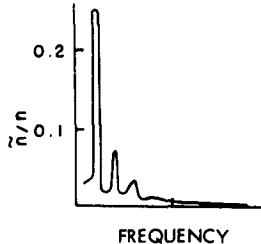


FIG. 7. Wave spectrum for a drift wave instability (from Ref.[15]).

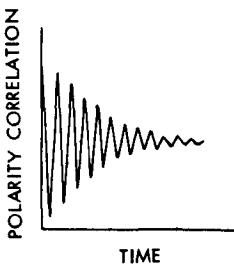


FIG. 8. Polarity correlation function for a drift wave (from Ref.[15]).

At the risk of oversimplifying, one can divide the problem into wave-wave and wave-particle interactions. It is likely that the wave-wave interaction for a large number of large-amplitude random-phase waves, i.e.  $R \gg 1$ , can be treated in a manner similar to that for Navier-Stokes turbulence. However, this situation does not appear to be common in plasma turbulence. From a number of experiments, one can infer that the Reynolds number for wave coupling in a plasma is generally not larger than unity, which is contrary to the fluid case. Characteristically, one observes the predominance in the spectrum of waves which are linearly unstable. A typical spectrum for unstable drift waves is shown in Fig. 7 [15]. It is

obviously not very similar to a Kolmogorov spectrum. Clearly, only a few waves are present and their narrow line width implies a long wave lifetime. In fact, phase correlation measurements show that wave lifetimes are many oscillation periods. Figure 8 shows the polarity correlation function for the drift waves of Fig. 7. In this case, the wave lifetime which is the reciprocal of the non-linear damping rate, is of the order of 10 oscillation times. Since the linear growth rates are of the order of the wave frequency, it is apparent that the coherent wave-wave interaction time is not small compared to the growth time as required by the weak-turbulence theory. The stabilization mechanism acts coherently with the linear growth mechanism such that the wave lifetime apparently depends on the algebraic sum of the two rates. If this were precisely true, one would expect the wave lifetime to become infinite at non-linear marginal stability, i.e.  $\text{Im } \omega = 0$ . Experimentally, this is reflected by a long phase-correlation time and a delta-function spectral density. This is to be contrasted to stabilization by the random-phase wave-wave coupling of weak turbulence theory in which the waves are constantly being created by a linear mechanism and destroyed by an essentially different process, i.e. one not coherent with the linear growth process. In this case, the wave lifetime (or phase-correlation time) at marginal stability would be equal to the reciprocal linear growth rate.

A mode coupling calculation based on a van-der-Pol-type analysis which is appropriate for a few waves with long coherence times has recently been performed [16]. This calculation reproduces many of the observed experimental features [17] except that the wave amplitudes for non-linear stability are much larger than those observed.

The wave-particle interaction, on the other hand, appears to be more efficient in stabilizing. Generalizing from the case of a stationary wave, we may anticipate that one effect of finite-amplitude turbulence will be to broaden the resonance (20) by an amount equal to the reciprocal of the time required for the waves to cause a deviation in a particle's trajectory of one wavelength. This effect may be regarded as a random Doppler shift due to the wave-induced component of particle motion. The random spatial motion of particles can modify wave growth as mentioned earlier. In some cases (very low frequency waves) this effect is analogous to an eddy viscosity. A more accurate picture is that the broadened resonance permits additional particles to resonate or exchange energy with the wave. If  $\partial\langle f \rangle / \partial v^2 < 0$  for these particles, then they can absorb energy from the wave and stabilize it.

The increase in the number of resonant particles will imply an increase in scattered particles and, therefore, an increase of various transport properties of the plasma. To illustrate this latter point, we consider particle motion across a magnetic field for the case in which the wave frequency  $\omega$  is much less than the cyclotron frequency. In this case, the instantaneous cross-field speed is  $cE/B$  and if  $\tau_{wp}$  is the electric field fluctuation time as seen by a moving particle, the particle's spatial diffusion perpendicular to the magnetic field is

$$D_{\perp} = \frac{c^2 \langle E^2 \rangle}{B^2} \tau_{wp} \quad (25)$$

The coherence time  $\tau_{wp}$  is in principle the same which appeared in expression (18). In the previous (weak turbulence) case we used an un-

perturbed orbit so that  $\tau_{wp}$  was given by relation (19). However, it is clear that for sufficiently large  $\langle E^2 \rangle$  the trapping time  $\tau_{TR}$  will be less than  $[\Delta(\omega - \vec{k} \cdot \vec{v})]^{-1}$  in which case  $\tau_{wp}$  becomes equal to  $\tau_{TR}$ .  $\tau_{TR}$  is the time to scatter one wavelength. In this instance, it will not be given by formula (24) but, instead, by

$$\tau_{wp} \approx \frac{\lambda B}{c E_{rms}} \quad (26)$$

Using this value in the formula (25) for  $D_\perp$  we obtain

$$D_\perp \approx \frac{c E_{rms}}{k_\perp B} \quad (27)$$

The dependence of  $D_\perp$  on  $E_{rms}$  rather than  $\langle E^2 \rangle$  as in the weak turbulence theory has been observed experimentally [18] as shown in Fig. 9. Not only the  $E$  dependence but the numerical value of  $D_\perp$  are in agreement with formula (27).

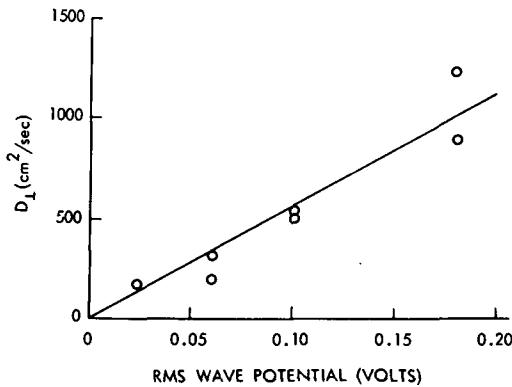


FIG. 9. Spatial cross-field diffusion coefficient as a function of the rms wave potential (from Ref. [18]).

If the wave phase velocity is much greater than the thermal velocity then virtually all particles become resonant when

$$\tau_{wp}^{-1} \approx \omega \quad \text{or} \quad k_\perp^2 D_\perp \approx \omega \quad (28)$$

$k_\perp$  is the perpendicular wave number. Since  $\partial \langle f \rangle / \partial v^2$  must be negative when averaged over all particles the criterion (28) should set an upper limit to the amplitude of the turbulence. This would indicate a diffusion coefficient of order

$$D_\perp \approx \frac{\omega}{k_\perp^2} \quad (29)$$

which is approximately equal to the Bohm value (8) when  $\omega$  is equal to the drift wave frequency.

The non-linear saturation criterion (28) predicts an rms electric field of

$$\frac{k_{\perp c} E_{rms}}{B} \approx \omega$$

which produces a density fluctuation of order unity in agreement with experiments:

$$\frac{\tilde{n}}{n} \approx \frac{2q E_{rms}}{k_{\perp} m v_T^2} \approx 1 \quad (30)$$

The ratio of linear to non-linear coherence times for wave-particle interaction can be used to define a Reynolds number for wave-particle interaction in analogy to the wave-wave Reynolds number (16).

There are, of course, many important developments in non-linear plasma theory which have not been mentioned. The topics discussed were chosen to indicate some of the emphasis of current research. This is a very active field, and future research work will undoubtedly witness many surprises and hopefully lead to a more adequate understanding of non-linear phenomena.

## R E F E R E N C E S

- [1] ROSTOKER, N., Nucl. Fusion 1 (1961) 101.
- [2] TIDMAN, D., MONTGOMERY, D., Plasma Kinetic Theory, McGraw Hill, N.Y. (1964).
- [3] SMULLIN, L., GETTY, W., Phys. Rev. Lett. 9 (1962) 3.
- [4] FIELD, E., FRIED, B., Physics Fluids 7 (1964) 1937.
- [5] DeGROOT, J., MacKENZIE, K.R., Phys. Rev. Lett. 20 (1968) 907.
- [6] KADOMTSEV, B.B., Plasma Turbulence, Academic Press, London (1965).
- [7] SAGDEEV, R.Z., GALEEV, A.A., Lectures on the Non-Linear Theory of Plasma, I.C.T.P., Trieste, preprint 1c/66/64.
- [8] DRUMMOND, W.E., PINES, D., in Plasma Physics and Controlled Nuclear Fusion Research, Nucl. Fusion-Suppl. Part 3, (1963) 1049.
- [9] VEDENOV, A.A., VELIKHOV, E.P., SAGDEEV, R.Z., in Plasma Physics and Controlled Nuclear Fusion Research, Nucl. Fusion-Suppl. Part 2, (1962) 465.
- [10] LEVITSKY, S.M., SHASHURIN, I.P., Zh. éksp. teor. Fiz. 52 (1967) 350 [translation: Soviet Phys. JETP 25 (1967) 227].
- [11] MALMBERG, J.H., WHARTON, C.B., Phys. Rev. Lett. 19 (1967) 775.
- [12] O'NEIL, T., Physics Fluids 8 (1965) 2255.
- [13] ALTSHUL, L.M., KARPMAN, V.I., Zh. éksp. teor. Fiz. 49 (1965) 515 [translation: Soviet Phys. JETP 22 (1966) 361].
- [14] BERNSTEIN, I.B., GREEN, J.M., KRUSKAL, M.D., Phys. Rev. 108 (1957) 546.
- [15] BUCHELNKOVA, M.S., SALIMOV, R.A., EIDEIMAN, Yu.I., Zh. éksp. teor. Fiz. 52 (1967) 837 [translation: Soviet Phys. JETP 25 (1967) 548].
- [16] STIX, T.H., Phys. Rev. Lett. 20 (1968) 1422.
- [17] HENDEL, H.W., COPPI, B., PERKINS, F., POLITZER, P.A., Phys. Rev. Lett. 18 (1967) 439.
- [18] EASTLUND, B.J., JOSEPHY, K., LEHENY, R.F., MARSHALL, T.C., Phys. Fluids 9 (1966) 2400.



# CONTROLLED THERMONUCLEAR RESEARCH

W.B. THOMPSON

University of California,  
La Jolla, San Diego, Calif.,  
United States of America

## Abstract

CONTROLLED THERMONUCLEAR RESEARCH. 1. Introduction; 2. Energetic considerations; 2.1. Nuclear energy release; 2.2. Containment times; 2.3. Radiation losses (Bremsstrahlung) synchrotron radiation, electron temperature, transparency; 3. Merits of fusion reactors; 4. Plasma confinement; 5. Microinstabilities.

## 1. INTRODUCTION

One of the greatest practical achievements of modern physics has been the tapping of nuclear energy and the subsequent development of the fission reactor. One of the great challenges is that of the controlled release of nuclear energy by the fusion of light nuclei. This is the great source of natural energy, through nuclear fusion in stellar interiors, and there are many arguments that make it the ideal artificial source of energy.

That such energy can be artificially released in an uncontrolled, catastrophic, fashion is a fact of which we are all painfully aware; but the problem of a controlled usable release presents many new interesting — and possibly even solvable — problems.

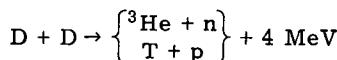
## 2. ENERGETIC CONSIDERATIONS

### (a) Nuclear energy release

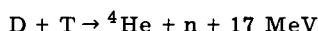
Preliminary to my discussion of this process must be basic energy considerations, which are dominated by the fact that fusion reactions occur between charged particles. It follows that a major role in the low-energy cross-sections is played by the Coulomb barrier, which contributes a factor

$$\exp\left(-\frac{Z_1 Z_2 e^2}{\hbar} \sqrt{\frac{m_r}{E}}\right) \sim \exp\left(-\frac{10}{\sqrt{E}} Z_1 Z_2\right)$$

where E is in keV. From the Z dependence it also follows that the important reactions are those starting with the isotopes of hydrogen



$$\sigma \sim 0.2 \text{ b at } 1 \text{ MeV}$$



$$\sigma = 4 \text{ b at } 100 \text{ keV}$$

The DT cross-section is so large at low energies that our attempt to produce a fusion reactor is challenged.

In considering how these exothermic nuclear reactions may be exploited, we must first consider the role of elastic collisions, which are determined by the Rutherford cross-section

$$\sigma_R \sim 8\pi \left( \frac{e^2}{mv^2} \right) \log \Lambda$$

where the logarithmic factor represents the usual consequence of the long range of the force, and is usually  $\sim 10$ :

$$\sigma_R \approx 10^5 / E^2 \text{ barn with } E \text{ in keV}$$

Thus, for energies in the keV range [ $E < 100$  keV]  $\sigma_R \gg \sigma_N$ , which implies that particle energies will be fairly well randomized before much nuclear energy is released; hence fusion energy is most likely realizable only through thermonuclear reactions.

Rates of energy release then require integration of cross-sections over a Maxwellian distribution. A crude steepest-descent evaluation yields the form

$$P_N \sim n^2 \langle \sigma v \rangle = n^2 a T^{-2/3} \exp(-b T^{-1/3})$$

Numerically

$$P_N \sim 10^{-26} n^2 T^{2/3} \exp(-18.8 T^{-1/3}) \text{ watt/cm}^3$$

(with  $T$  in keV: 1 keV  $\sim 10^7$  deg) for the DD-reaction

$$\sim 10^{-23} n^2 T^{-2/3} \exp(-19.9 T^{-1/3}) \text{ for the DT-reaction}$$

This immediately shows that the energy release is negligible for  $T$  much less than 1 keV.

#### (b) Containment times

The production of any thermonuclear power requires a considerable investment of energy in heating the fuel, and unless at least this investment is recovered there is little hope of a power gain. True, the energy in the particles is recoverable and could be recirculated, but the necessarily finite efficiency of this process results in the loss of a significant fraction of the circulating power.

Since the particles necessarily have energies in the few keV range, and the nuclear energy released is of order a few MeV per particle pair, it is clear that a reasonable fraction of the fuel, about 1%, must be burnt

up before the particles leave the system. Thus, if the burn-up rate of a particle is  $\tau \approx n\langle\sigma v\rangle$ , and the mean life  $\tau$ , then  $\tau \approx 10^{-2}$ . However,  $\sigma \approx 10^{-25}$  and  $v \approx 10^8$ , hence  $n\tau \approx 10^{15} \text{ s/cm}^3$ . This is, of course, somewhat smaller for the DT reaction,  $\sigma \approx 10^{-24}$  and  $\mathcal{E}_r \approx 20 \text{ MeV}$  and  $\approx 10^{14}$ .

### (c) Radiation losses

If the system were optically thick, radiation losses would occur only from the surface, but they would occur at the fantastic rate of  $\sigma T^4 \approx 10^{15} \text{ T}^4 \text{ watt/cm}^2$  ( $T$  in MeV). Optical depth, however, is determined by the Thomson cross-section  $\sigma \sim 10^{-25} \text{ cm}^2$  and the density  $n$ . The number density  $n$ , however, is limited by supportable pressures and even if we assume a pressure of  $10^6$  atmospheres, surely an upper limit, then, at  $T = 10 \text{ keV}$ ,  $n \sim 10^{17}$ , and a radiation mean free path of  $\lambda \approx 10^8 \text{ cm}$ , the systems are optically thin. In that case, radiation occurs directly from the electrons and we must consider the two main volume sources of radiation (see (i) and (ii) below).

#### (i) Bremsstrahlung

The thermonuclear fuel must necessarily be accompanied by enough electrons to neutralize the electric charge; these electrons will be hot, perhaps in equilibrium with the ions, and will suffer accelerations due to inter-particle forces. On collision, they radiate at a rate  $d\mathcal{E}/dt = (2/3)(e^2/c^3)a^2$ , and the fuel (a plasma) radiates power at a rate

$$P_b = \frac{2}{3} \frac{e^2}{c^3} a^2$$

However,  $a^2 = \sum_i Z_i^2 \frac{e^4}{m_i^2 r_i^4}$  where  $Z_i$  is the ionic charge and  $r_i$  the distance to the  $i$ -th ion. This is obtained by summing  $a^2 = \sum_{i,j} \frac{Z_i Z_j e^4}{m_i^2 m_j^2} \frac{\vec{r}_i \cdot \vec{r}_j}{r_i^3 r_j^3}$

and using the random phase approximation on the product  $\vec{r}_i \cdot \vec{r}_j^*$ , the sum  $n \int \frac{d^3 r}{r^4}$  diverges classically as  $1/r$ ; however, the electronic charge can be localized no better than the de Broglie length, hence

$$\frac{1}{r_{\min}} = \frac{\langle mv \rangle}{\hbar} \quad \text{and}$$

$$P_b = \frac{2}{3} n^2 \left( \frac{e^2}{mc^2} \right)^2 \frac{e^2}{\hbar c} mc^2 \langle v \rangle$$

$$= 5 \times 10^{-31} n^2 T^{1/2}$$

At low temperatures this exceeds the nuclear production rate, but as temperature goes up,  $P_N$  increases more rapidly than  $P_B$  and exceeds

it at  $T \approx 4$  keV for D-T, and at  $T \approx 30$  keV for D-D. At 50 keV, the D-D energy production is twice the bremsstrahlung loss.

### (ii) Synchrotron radiation

To obtain the confinement times  $\tau$  required for burn-up, it is usually considered necessary to have a strong magnetic field threading the reacting plasma. In that case, however, the electrons experience a further acceleration  $a = \omega_c v = (eB/mc)v$ , and there is a further volume radiation

$$P_{\text{syn}} = \frac{2}{3} n \frac{e^2}{mc^3} a^2 = \frac{2}{3} n \left( \frac{e^2}{mc^2} \right)^2 \left( \frac{v}{c} \right)^2 B^2 c$$

The ratio of synchrotron radiation to bremsstrahlung is

$$\frac{P_{\text{syn}}}{P_b} = \frac{\hbar c}{e^2} \frac{B^2}{nmc^2} \frac{v}{c} = \frac{0.05 T^{3/2}}{\beta}$$

where  $\beta = (8\pi nkT)/B^2$ , and represents the ratio of thermal to magnetic energy. At present, methods of plasma confinement are usually effective only at small values of  $\beta$ , certainly  $\beta < 1$ . Thus, for  $T = 40$  keV

$$\frac{P_{\text{syn}}}{P_b} = \frac{12}{\beta}$$

while at  $T = 4$  keV

$$\frac{P_{\text{syn}}}{P_b} = \frac{0.4}{\beta}$$

and synchrotron radiation is very damaging to the D-D system. There are some factors, however, that are able to reduce this damage.

### (iii) Electron temperature

These radiation losses depend only on the electron temperature, hence if the ion temperature were much greater than the electron temperature, we might have over-estimated the loss.

If the electrons are heated only by collisions with the ions, which is a good approximation if the plasma is quiescent, then

$$\begin{aligned} \frac{dT_-}{dt} &= n \langle \sigma v \rangle \frac{m_-}{m_+} [T_+ - T_-] \\ &= 8\pi n \left( \frac{e^2}{mv^2} \right)^2 \frac{m_-}{m_+} [T_+ - T_-] \end{aligned}$$

If electrons are heated in this way, but lose energy by synchrotron radiation and remain much cooler than the ions, then  $T_e$  is given by

$$8\pi \log \Lambda n \left( \frac{e^2}{mc^2} \right)^2 \left( \frac{c}{v_-} \right)^3 c \frac{m_-}{m_+} kT_+ = \frac{2}{3} \left( \frac{e^2}{mc^2} \right)^2 \left( \frac{v_-}{c} \right)^2 B^2 c$$

$$\left( \frac{v_-}{c} \right)^5 = \frac{3}{2} \log \Lambda \left( 8\pi \frac{n k T_+}{B^2} \right)$$

$$T_e \approx 50 \beta_+^{2/5} \text{ keV}$$

Even with the electron temperature limited in this way, however, the synchrotron losses are serious for the D-D system or the low- $\beta$  D-T system.

#### (iv) Transparency, etc.

A remaining feature is the self-absorption of synchrotron radiation. Although the plasma is transparent to the high-frequency bremsstrahlung, the synchrotron radiation comes out in frequencies where the plasma transmission effects are important; hence self-absorption effects are important, and the radiation is emitted from the surface rather than the volume of the plasma; therefore, if the plasma volume is great enough, these losses may be rendered unimportant. Moreover, the radiation, in the r.f. and infrared ranges, may easily be reflected and losses further reduced.

Calculation of the opacity is rather a delicate matter and, although straightforward, must be done with some care. When it is known, the necessary minimum scale length can be calculated. For a D-D system in which the active products T and  ${}^3\text{He}$  are also burned, the scale lengths are quite modest for high  $\beta$ . Indeed, if  $\beta = 0.4$ ,  $L = 16.7$  cm (Drummond). These simple energetic arguments are quite encouraging for a D-T system. They are, however, a good deal less optimistic for D-D; but with help of reflecting walls, or better, the development of a system in which the magnetic field penetrates only the edge of the plasma so that  $\beta \gg 1$  throughout most of the system, even that system is energetically feasible.

### 3. MERITS OF FUSION REACTORS

(a) If an economical D-D reactor could be produced, the arguments for it are quite overwhelming, provided only the cost, complexities, and size of the necessary apparatus remain within reasonable bounds. Fuel is essentially limitless, easily separated and universally available. The end product of the process is helium, and although a reasonable flux of neutrons and X-rays would be produced ( $10^{18}$  neutrons per second and megawatt) the problem of shielding and heat transfer, while formidable, does not seem excessively so. There is no stage where a large charge of fuel is required, at most a few grams, and even that of low reactivity, so that the risk of radioactive contamination of the environment is very small.

Thus, a D-D reactor could realize all the hopes of atomic power: negligible fuel costs, cleanliness, and safety.

(b) The D-T reactor is a good deal less attractive. In the first place, T does not occur naturally, but must be obtained from



i.e., we must employ the neutrons produced in the TD reaction. Since this reaction yields only one neutron/triton, it is necessary to breed neutrons using the reaction



Therefore, we need a neutron-breeding tritium-producing blanket as part of the system. Since only light elements are required in this it is probably possible to design a system that will avoid the build-up of long-lived radioactive products, and although the active charge is now radioactive, the total charge remains small, and of modest half-life, so that the safety and contamination hazards remain small.

On the other hand, the fuel argument is no longer too sound. It is true that even  ${}^6\text{Li}$ , the scarce isotope (7.4%), is more abundant than U, but not much so; and when the difference between the 250 MeV/U and the 20 keV/ ${}^6\text{Li}$  is taken into account, there is little advantage in fuel economy of a T-D system over a fission reactor, especially since breeding has been demonstrated as feasible.

The contamination argument, however, remains valid, and will certainly assume increasing significance as the use of atomic energy spreads.

Moreover, the technology associated with a T-D reactor has not yet been developed, and it may be that a thermonuclear reactor will be significantly simpler than a breeder reactor, and able to produce energy more economically. It will almost certainly provide less of a hazard.

Thus, even if controlled thermonuclear research cannot proceed beyond the D-T system, there are still strong arguments in favour of developing such a device. Moreover, in carrying out this development, our knowledge of plasma physics may reach the point where a D-D system can be developed. It must be remembered, however, that a D-D system will probably be quite different from one involving T-D, and working on a much smaller energy margin, will call for much more detailed knowledge of the properties of hot plasma.

A major property of any controlled thermonuclear reactor is that it involves an energy cycle: A very considerable investment of energy must be put into the fuel and its containing fields in order to produce any significant nuclear energy, and a great deal of energy perpetually escapes from the fuel and must be replaced; thus, the efficiency of transmitting energy to and receiving energy from the plasma is of central importance, a problem which is much less serious for most prime movers. This implies that success will require skill in manipulating plasma and that this skill must be based on extensive knowledge of plasma properties which is only acquired by broadly based scientifically motivated research!

#### 4. PLASMA CONFINEMENT

Of the schemes for plasma confinement that have been suggested (confinement by inertia; by electric fields; intersecting beams) only magnetic confinement seems to be very hopeful. Because of the existence of a volume force  $\vec{j} \times \vec{B}$ , which could be written  $-\nabla B^2/8\pi$  for straight field lines, a magnetic field can maintain a plasma pressure, and, thus, confine the plasma if the field is shaped properly. When so confined, the plasma will, of course, diffuse across the field lines, and if the diffusion is classical, the diffusion coefficient is

$$D \sim v r_L^2 \sim r_L^2 \langle n \sigma v_\theta \rangle = n v_\theta^3 \frac{mc^2}{e^2 B^2} \left( \frac{e^2}{mv_\theta^2} \right)^2 8\pi \log \Lambda$$

$$\approx 10^{-4} n/B^2 T^{1/2} \text{ cm}^2 \text{s}^{-1}$$

and the diffusion time  $\tau \sim R^2/D \approx 1 \text{ s}$  at  $n = 10^{17}$  for  $B = 100 \text{ kg}$  and  $T = 100 \text{ keV}$ , an adequately long time.

It is on the realization of the magnetic confinement of a hot dense plasma that most controlled thermonuclear research has been concentrated.

The problem of producing a magnetic field of a geometry able to confine plasma, while formidable in general, has a number of simple solutions. One class of fields relies on the use of closed and nested surfaces of magnetic flux, which at the same time contain the currents and are isobars of the confined plasma. Since flux and field lines are divergence-free the simplest of such configurations is a torus; and the simplest of all a self-constricted current loop, the toroidal pinch.

A second class of confining fields relies on the constancy, in slowly varying fields, of the "magnetic moment"  $\mu = \frac{1}{2}mv_\perp^2/B$ , generated by a charged particle as it rotates with a frequency  $\Omega = eB/mc$  in a circle of radius  $r_L = v_\perp/\Omega$

$$\mu = I_A = \frac{\Omega}{2\pi} \frac{e}{c} \pi r_L^2 = \frac{1}{2} \frac{e v_\perp^2}{c \Omega} = \frac{1}{2} m \frac{v_\perp^2}{B}$$

A field gradient operating on this moment drives the particle toward regions of weak field; hence, the magnetic field between a pair of Helmholtz coils will confine charged particles irrespective of sign at the field minimum between them. This is the basis of the "magnetic mirror".

These two simple devices, like most other simple confining geometries, are bedeviled by a host of instabilities, the plasma changing from its simple form to some more complicated one, separating charges, producing electric fields  $E$ , and drift motions  $\vec{v} = \vec{E} \times \vec{B}/B^2$  across the magnetic field.

To understand such motions we must have a picture of plasma dynamics. In the limit of small Larmor radius  $r_L \rightarrow 0$  the velocity of the plasma is exactly the drift  $\vec{v} = \vec{E} \times \vec{B}/B^2$  and this has as its consequence  $\vec{E} + \vec{v} \times \vec{B} = 0$  so that  $(D/Dt) \int B dA = D\Phi/dt = \phi (\vec{E} + \vec{v} \times \vec{B}) = 0$ , i.e. the plasma must move such as to conserve the magnetic flux. If, then, the magnetic field is strong and nearly a vacuum field (low- $\beta$  plasma), the specific volume

$\tau$  of a plasma element is specified by the flux element and the field geometry  $\tau = \int dl/B$ . We must now ask under what circumstances the internal energy of the fluid can be tapped by an interchange of fluid elements, just as in examining the problem of thermal convection; again we find the result that for stability the entropy gradient must be parallel to the specific volume gradient. For most of the simplest geometries, however, this inequality goes the other way ( $\nabla s \nabla \tau < 0$ ).

The first scheme adopted to overcome this difficulty was shear stabilization. If one considers a toroidal system in which there is an axial field (around the major axis), as well as an azimuthal one (around the minor axis), so that the field lines are helices lying on toroidal surfaces, then radially adjacent field lines are topologically inequivalent, and can only be interchanged if some flux breaking occurs. Enough shear, it was thought, would stabilize the plasma. Experiments intended to implement this, were, however, unsuccessful, and after some investigation, their failure was laid to a rapidly growing resistive instability. The plasma is, of course, not perfectly conducting, but the destruction of magnetic flux proceeds at a rate of  $\eta \partial^2 B / \partial x^2 \sim (\eta/L^2) B$ ; hence, if  $\eta/L^2$ ,  $\eta$  being the resistivity, is small compared to  $v/L$ , i.e., if the magnetic Reynolds number is high, then resistivity is negligible. However, in a stability calculation all scale lengths  $L$  must be included, and for some of these resistive dissipation, which permits reconnection of lines of flux, may release large amounts of energy. Instabilities are found to grow as  $\eta^{1/3}$ , hence remaining important even at small  $\eta$ .

The initial failure of shear stabilization left the possibility of making  $\nabla s \nabla \tau > 0$  by making the flux volume decrease outward. One simple geometry, the cusp obtained by reversing the current in one mirror coil, had long been known, but it suffers from an important disadvantage: since  $B = 0$  at the centre, the magnetic moment  $\mu$  is no constant of motion, and the losses are essentially geometric, determined by the fraction of the surface that permits escape. If the width on the line cusp is  $d$  then the total loss through the ends and the line cusp is  $\sim (4\pi dR/4\pi R^2) (nv_\theta/R)$  and the life-time  $\tau = R^2/dv_\theta$ . The smallest  $d$  could be  $\sim r_L^+$  and  $\tau \approx (n^2/N_\theta^+ N_\theta^-) (eB/mc) \sim (R^2 B/P) \times 2 \times 10^{-10}$ . If  $B = 10^5$ ,  $P = 1$  keV we have  $\tau = 1$  at  $R = 200$  cm. Experiments, however, have succeeded only in producing (for  $d \sim r_L^+$  and, therefore,  $\tau = 1$  at  $R = 1600$  cm) a volume of  $10^7$  litres; since a thermonuclear reactor would require densities of  $10^{15}$  cm $^{-3}$  during the confinement time, the output would be at least  $10^{-30}$  n $^2 \approx 1$  watt/cm $^3$ . Hence a total output exceeding  $10^9$  W seems to be unrealistic. At  $R = 200$  cm, this is reduced to 6 MW, which is quite reasonable.

A conceptually simple, though technically difficult modification of the cusp can overcome the field-zero problem. It is merely necessary to put a single current-carrying conductor along the cusp axis, thus producing a toroidal volume at which the magnetic field has a non-zero minimum. This stuffed cusp is a simple example of a magnetic well.

A somewhat easier device to fill is a magnetic mirror supplemented by an external set of axial rods each carrying a current directed opposite to that in its neighbours. These are the celebrated Joffe bars.

A number of experiments of this kind have been performed, the plasma being produced in a wide variety of ways; gun injection, neutral particle injection, breakdown in situ, and at low  $\beta$ , the plasma containment has been greatly improved. No sign of the flutes which characterize the unstable

plasma are seen, and the plasma is confined for times up of several milliseconds, 1000 times the growth time for rapid instabilities.

## 5. MICROINSTABILITIES

The successful development of a hydrodynamically stable plasma, which behaves in a theoretically predicted way, has been the most cheering event in controlled thermonuclear research for some time. It does not, however, mean that the pathway to the thermonuclear reactor is clear. Indeed, initial experiments at very low densities of  $10^8$ - $10^9$  cm $^{-3}$  were encouraging, but as density was increased and other causes of plasma loss were removed, it was discovered that although there were gross motions, particles were still being lost by some anomalous diffusion process.

Such anomalous diffusion effects have been commonly observed in low- $\beta$  containment devices. Although the plasma does not exhibit any gross motions, particles diffuse across the fields at a rapid rate; the diffusion coefficient is not given by  $r_L^2 v$ , but  $r_L^2 \Omega = (kT/2B)c \sim 10^4 T/B$  (Bohm's formula). This gives life-times of order  $\tau_B = BR^2/10^9 T$ . For  $B = 10^5$ ,  $R = 10$ ,  $T = 1$  we have  $\tau = 10^{-2}$  which is about two orders of magnitude too small for adequate containment; in many experiments  $\tau_B = 10^{-3}$  s is observed.

Associated with the appearance of anomalous diffusion, there is usually a fair amount of noise in electrostatic fluctuations, both of which phenomena are effects of microinstabilities, which draw their energy not from the expansion of the plasma, nor from the untwisting of magnetic field lines, but from the non-thermal equilibrium shape of the distribution function.

Microinstabilities are intimately associated with the phenomenon of Landau damping, the collisionless extraction of energy from a longitudinal wave in a plasma. Particles whose initial speed is close to that of a wave, find themselves riding in the wave, and if travelling slightly slower than the wave, being systematically accelerated to its speed. Of course, particles travelling faster than the wave are also slowed down, but if, as is the case, in thermal equilibrium  $\vec{v} \cdot \partial f / \partial \vec{v} < 0$  there is a net transfer of energy from wave to particle (surf riding). This process has been subject to theoretical discussion for some time, but has only recently been convincingly demonstrated in laboratory experiments.

If the distribution function is non-equilibrium, it may have an increasing derivative (more exactly  $\int d^2 v_1 \partial f / \partial v_{\parallel} > 0$ ) in which case energy is given from the particles to the waves, whose amplitude thereupon grows. Since a distribution function can have such a positive gradient for a wide number of reasons, and since a non-uniform magnetized plasma can sustain a remarkable spectrum of waves, it is not surprising that the catalogue of these microinstabilities is long, and since their analysis is woefully complex, and experiments difficult and uncertain, it is also not surprising that it has taken a long time to survey them. Moreover, these instabilities, which are essentially features of the collisionless plasma, may be interfered with in a variety of ways by collisions, sometimes rendered stable, and sometimes made unstable.

A fortunate feature is that very many indeed of the microinstabilities are found to be prevented by the construction of a magnetic well. In open-

ended (mirror) systems, however, there is a serious one left, the loss cone instability. Any system which depends on  $\mu$  for trapping particles clearly cannot trap those for which  $\mu = 0$ , hence, the equilibrium distribution must have  $\partial f / \partial^{\frac{1}{2}} v^2 > 0$  somewhere. If we consider a wave with phase velocity  $u$  propagating across the field, we can reduce the Landau term to

$$\int dv_{\parallel} \int_{-u}^{+u} \frac{\partial f}{\partial^{\frac{1}{2}} v_{\perp}^2} \frac{v_{\perp}^2 dv_{\perp}}{\sqrt{v_{\perp}^2 - u^2}}$$

For small  $\mu$ , this is either zero or positive. The associated wave propagates between the mirrors, growing as it goes. Its development has been demonstrated in experiments on jets of plasma with anisotropic velocity distributions.

Since the wave grows as it proceeds from one end of a mirror machine to the other, where it is fairly well absorbed in the mirror, it can be controlled by making the device short enough, at least for low- $\beta$  systems. This, of course, limits the size of a mirror trap, which in earlier analysis was assumed to be arbitrarily long. Even this expedient may not be effective at higher  $\beta$ , and a demonstration of a fully stabilized open-line magnetic well at significant  $\beta$  has not yet been given.

An alternative approach is to try to deform a toroidal system into a magnetic well. For the mirror type system, it is possible to produce field minima, from which the magnetic field increases in all directions. For closed systems, this is not possible, and one must be content with systems in which some average of the magnetic field increases:  $\oint dl/B$  or if  $V(\phi) \equiv$  volume enclosed by axial flux  $\phi$ ,  $d^2V/d\phi^2 > 0$ . The simplest of such systems is probably the multipole, in which a current is produced on a set of internal conductors, and imaged in an appropriately shaped boundary.

Experiments performed on toroidal multipoles have been remarkably successful, in that stable plasmas with no detectable fluctuations have been confined until swept into the support rods of the conductors. This time, however, is still short (a millisecond) and the plasma density confined is low ( $\sim 10^{10}$ ), thus in actual confinement, these geometries have not done much better than open wells. However, the magnetic fields are much lower,  $\sim 2000$  g instead of 60 kg and the values of  $\beta$  1000 times greater (though still low).

The support rods, which are essential parts of certain multipole devices, prevent proper plasma containment. There are geometries of a more complex form, usually with a helical axis, in which no such supports are necessary; but no experiments have yet been reported in any of these.

One possible consequence of these enhanced fluctuations is that even though the enhanced diffusion is tolerable, there may be a serious increase in the amount of synchrotron radiation, particularly if the collective oscillations introduce coherence. This problem has not yet been examined, but is probably not serious, at least for the toroidal systems, where the associated frequencies are low.

The stability in average minimum  $B$  wells depends on the approximate conservation of magnetic flux, and one might again ask if finite conductivity introduces new instabilities. It is found to do so, the instability now depending on the distance separating regions of good and bad curvature; growth

rates, however, scale with  $\eta$  not  $\eta^{1/3}$ , and this renders such effects less significant than in shear stabilized systems. But they again limit the lengths of closed magnetic wells. It is disconcerting to find that some of these instabilities are enhanced by increasing the well depth.

The closed wells, at first glance, should not be susceptible to the loss cone instability; but it has been pointed out that any variations of the magnetic field strength along the field line lead to particle mirroring, particles being trapped in a low-field region. If this is also a region of bad curvature (locally unstable) a local flute, driven by the trapped particles, can grow unstable. This imposes further limitations on field geometry. Finally, sufficiently large temperature gradients can lead to a further slowly growing instability.

At present, it is these toroidal wells that are the white hope of the controlled thermonuclear research projects, and they certainly do show most promise; mostly because of the loss cone instability problem. Some attention has been given to the problem of filling the loss cone by using time varying fields, the r.f. plug. Early experimental results were not promising and the amount of power needed is disconcerting, but this should probably be re-examined.

Finally: the question of the cusp is of great importance. If the most optimistic theoretical leak size could be realized, it is one of the few systems that might confine a plasma with only a small internal field and make possible a D-D reactor.

Next to stable confinement, the most important problem is that of plasma heating. Indeed, it is possible that low- $\beta$  magnetic wells have been experimentally successful less because of the stabilization scheme than because of the use of diffuse high-temperature plasma. The crucial criticism of the idealized stability theory was the result of the hard-core experiment, and the inner side of the hard core depends not on shear, but on minimum B for stabilization.

In this field, the most interesting developments have been those concerned with non-linear heating processes. While the most dramatic experiments, which involve putting enormous voltages across low-density plasmas, and producing violent electron beam instabilities, are a long way from detailed theoretical interpretation, this research has been stimulated and has stimulated theoretical research on the non-linear behaviour of plasmas, another major activity. It is not yet clear that a turbulence-based heating process will not lead to a plasma exhibiting turbulent diffusion. The hope, however, is that the turbulence will decay before any damage is done by diffusion.

Perhaps the most dramatic developments in plasma physics and controlled thermonuclear research have been outside the field of plasma physics itself. The development of the laser has enormously increased the power of optical diagnostics and has suggested new methods (evaporation of solids by focussed laser pulses) of filling traps with hot, pure plasma. The development of high-field superconductors and of practical superconducting magnets has promised to completely transform the energy balance of low- $\beta$  reactors — a steady, strong, magnetic field can be maintained at an extremely modest expenditure of energy. Another development — that of the heat pipe — might prove a necessary complement to this.

The most important development, however, has probably been in the development of a scientific attitude toward the physics of plasmas. Gradually,

theory and experiment are coming together. No longer is plasma physics a combination of disjointed parts, unverified and unintelligible theory, and uninterpretable experiment. Careful experiments have given unambiguous support to some aspects of the theory, and increasing realism among theorists has led to verifiable predictions. No longer is the uniform infinite plasma enough nor are the current voltage characteristics a result.

The most distressing element has been a dwindling support for plasma physics. The scientific developments of recent years have been highly encouraging; stable plasmas have been produced in accordance with theoretical predictions, thus indicating that, at least, the low- $\beta$  plasma is becoming intelligible. Furthermore the low- $\beta$  system has become a practical reactor possibility, thanks to the superconducting coil which has made magnetic energy virtually cost-free; but at the same time popular interest in and support for controlled thermonuclear research, has, at least in some quarters, seriously diminished. This disillusionment, which has no scientific or factual basis, is unfortunately timed, and we can only hope it will lead to no serious postponement of the realization of controlled thermonuclear power.

# THE PROBLEM OF PLASMA CONFINEMENT

B. COPPI

Institute for Advanced Study and Princeton University,  
Princeton, N.J., United States of America

## Abstract

THE PROBLEM OF PLASMA CONFINEMENT. An elementary presentation of the problems overcome and to be solved in the process of understanding the collective effects that determine the dynamics of low-density plasmas in strong magnetic fields is given. The agreement between theory and experiments that appears to characterize a new phase of this field of research is pointed out.

As is well known, the effort to obtain thermonuclear fusion reactions in a non-explosive and controlled form started almost simultaneously with the development of the hydrogen bomb. At that time there was a great deal of expectation and it was felt that the problem was only a technological one and that an imminent solution was possible. On these grounds, a number of devices for the magnetic confinement of a hot plasma were proposed and realized.

The theoretical basis was not very wide. Then the experimental evidence did not support the expectations. In fact, it soon became apparent that plasma was being lost at a rate which was much faster than that predicted by ordinary diffusion due to collisions between particles. The way in which plasma escaped differed from device to device, but it was evident that in all cases the plasma exhibited unexpected collective effects which caused the so-called "anomalous particle loss". This was a setback from a practical point of view but on the other hand it widened the perspectives of a new field of research.

Now I will limit myself to discussing the problem of confinement, that is, of containing a plasma in a magnetic field configuration for a time not too short in comparison with the time of "classical" diffusion, due to inter-particle collisions. To be more precise, we refer to Fig.1, where a confined plasma is represented by a one-dimensional model with a density gradient in the  $\hat{x}$  direction and a strong magnetic field in the  $\hat{z}$  direction. This configuration is not in thermodynamic equilibrium and it will tend to expand at a rate which can be represented by the diffusion equation

$$\frac{\partial n}{\partial t} + \frac{\partial}{\partial x} \mathcal{D}_c \frac{\partial n}{\partial x} = 0 \quad (1)$$

where  $\mathcal{D}_c$  is the "classical" diffusion coefficient and  $n$  is the particle density. Considering that the ions are dragged by the electrons while diffusing and that an electron moves by a step of the order of its gyro-radius  $a_e$  in a collision with an ion, we have, by considerations of random walk,

$$\mathcal{D}_c \approx v_{ei} a_e^2 \quad (2)$$

Here  $\nu_{ei}$  is the electron-ion collision frequency, so that  $\nu_{ei} \approx n e v_{the} \bar{\sigma}$ ,  $\bar{\sigma}$  being the average electron cross-section proportional to  $v_{the}^{-4}$ ,  $v_{the}$  being the electron thermal velocity,  $a_e \approx v_{the}/\Omega_e$ , and  $\Omega_e = eB/m_e c$  the electron gyrofrequency. So we have

$$\mathcal{D}_c \propto \frac{n}{B^2 T_e^{1/2}} \quad (3)$$

$T_e$  being the electron temperature. It is clear that  $\mathcal{D}_c$  has a favourable dependence on the magnetic field intensity and on the electron temperature for high-temperature plasmas in a strong magnetic field.

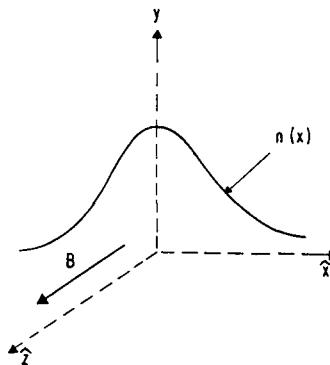


FIG.1. Sketch of a plasma confined in a strong magnetic field.

On the other hand, the experimental evidence in most of the tested confinement schemes has been for effective diffusion coefficients, introduced empirically to measure the particle losses that exhibit a quite different dependence on  $B$  and  $T_e$ . This is usually referred to as "anomalous" diffusion. Generally speaking, it has been found that the higher the electron temperature, the larger is the effective particle loss. From this we can roughly deduce the conclusion that collective effects in which the thermal electrons play an important part are mainly responsible for the high diffusion rates that have been observed. A typical expression has been empirically given by Bohm [1] for a diffusion coefficient

$$\mathcal{D}_B \approx \frac{1}{16} \frac{c T_e}{e B} \quad (4)$$

that has represented the outcome of a variety of confinement experiments [2]. Therefore, in recent years, a great deal of theoretical and experimental effort has gone into trying to identify and control the collective effects that can give rise to this type of diffusion. We refer to Fig.2 to summarize the achievements and what we think remains to be done in order to increase the present levels of plasma confinement.

1. Here we call macroscopic (fluid) modes those given by the equation of magnetohydrodynamics and not requiring knowledge of the one-particle distribution function  $f(\vec{r}, \vec{v}, t)$  but of the first three moments of it, i.e.

$n = \int f d^3 v$ ,  $\vec{u} = \int f \vec{v} d^3 v$ , and  $\Pi P = \int f v \vec{v} d^3 v$ . The main modes of this type are similar to those occurring in a container of two fluids of different density with the heavy fluid above the light one [3]. In a plasma containment system the light fluid is replaced by the magnetic field and the heavy fluid by plasma, the role of the gravity versus the density gradient being played by the magnetic field curvature. These types of mode have been quite well identified experimentally and have been suppressed. The suppression has been achieved by generating in various ways a magnetic field curvature which is favourable to stability in the same way as an upward gravity would be for the two-fluid container mentioned above, or by "nesting" the magnetic field, i.e. creating magnetic "shear", so that plasma cannot expand without distorting the magnetic field lines and, by increasing the magnetic energy, generating a restoring effect. The experimental attainment of these conditions has shown dramatic increases of the contained plasma [4].

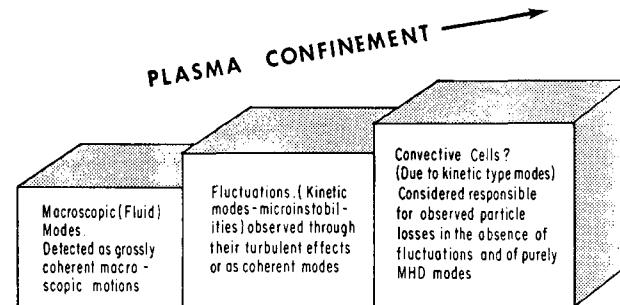


FIG.2. Theoretical and experimental difficulties met in trying to achieve plasma confinement.

2. The suppression of large-scale unstable motion, i.e. the overcoming of the first step in Fig.2, has brought to light the existence of fluctuations of particle density and electric field. These are considered the effects of modes that, roughly speaking and taking them as solutions of approximate linearized equations, have a real part of the frequency larger than its imaginary part, generally giving rise to a growth of the wave. Modes of this kind are not given by magnetohydrodynamics but require more refined theories that include the effects of finite gyration radius of the particles, wave-particle resonances (Landau damping effects), direct influence of particle-particle collisions, etc. The basic equations are the collisionless Boltzmann equation or the Fokker-Planck equation, since explicit consideration of the velocity space is now important. A great deal of experimental and theoretical work has been devoted to trying to identify those modes that appear to produce considerable energy and particle loss from a confined plasma. As a consequence, this has generated a wide interest in the study of non-linear interactions for these types of mode [5]. In particular, a great deal of theoretical experience has been

gained from solid-state physics [6] in order to construct kinetic equations for waves and for particles in the presence of fluctuations, needed to describe the effects of a statistical ensemble of modes in the so-called random phase approximation. We remark, however, that this is not always a good approximation of the experimental evidence [7], as discrete well-identified modes are observed in several significant cases. In parallel with the non-linear analysis, the investigation of modes giving rise to fluctuations has been carried out quite extensively in the linear approximation for a variety of physical regimes. On this basis, fairly precise predictions have been made concerning the dependence of the level of fluctuations on complex magnetic field configurations or on the relative influence of the particle-particle interactions versus wave-particle interactions [8]. As an example, the theoretical anticipation that a plasma, in which the particle mean free path is considerably longer than the length of the magnetic field lines inside it, is less subject to anomalous diffusion than a collisional plasma with short mean free paths, has been confirmed in a number of experimental situations<sup>1</sup>.

3. In this way we have now arrived at the last step where we have experiments in which there are no macroscopic (fluid) modes present and no fluctuations, but still a particle loss higher than that expected from classical diffusion is observed [10, 11]. Our present explanation for this is that there are modes left in our catalogue of plasma collective effects that are not given by the fluid (MHD) approximation and require starting from a kinetic equation but in a linearized approximation exhibit an imaginary part of the frequency larger than its real part. Therefore, we do not expect that they should be observed in the form of fluctuations, but rather that they may give rise to convective patterns leading the particles away from the confinement regions. Modes that are of special interest in this respect are those connected with the existence of trapped particles [12] in a configuration with varying magnetic field. More precisely, suppose for simplicity that we have a configuration with closed magnetic lines of force. Then some particles will circulate around, sampling the entire length of the lines of force, and some, the trapped ones, will oscillate around local wells (Fig.3). The trapped particles [13] can be shown to behave like another species, similar to the way impurities (ions of heavier mass than the average) or spots of cold ions (with temperature lower than the average) would behave in a plasma imbedded in a constant magnetic field [14]. So new kinds of mode with growth rate considerably larger than the frequency of oscillation and relatively large wavelengths have been found [12, 13, 15]. However, at this stage there is not enough experimental information nor enough theoretical analysis to make the convective cells conjecture more certain and to have safe indications on how to overcome this last difficulty.

Finally, before discussing in more detail some of the most typical experiments that are being carried out, I would remark that it has come to light quite evidently that often a plasma is more sensitive to the way in which it is formed than to the geometric and magnetic features of the configuration in which it is intended to be confined. In fact, we have seen

---

<sup>1</sup> A paper by Ellis and Eubank [9] presents a comparison of experimental measure of anomalous particle losses in the Model Etude stellarator for collisional and collisionless regimes in the presence or in the absence of fluctuations.

that, for instance, depending on whether the particle mean free path is long or short, a plasma can respond in a quite different way to the same magnetic confinement configuration, so that a topological classification, such as we usually give of the various types of experiments, is somewhat inadequate. In addition, the possible diagnostics of a contained plasma, for instance a given method to determine the electron temperature [16, 17], may be strongly dependent on the type of collective effects to which it is sensitive, and affect the over-all evaluation of the experiment.

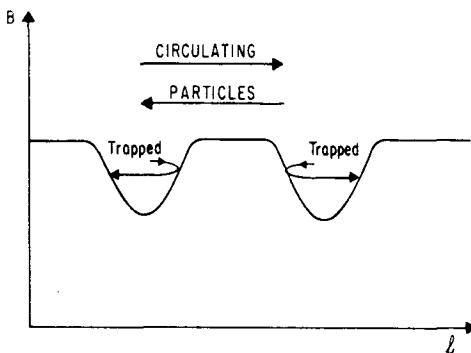
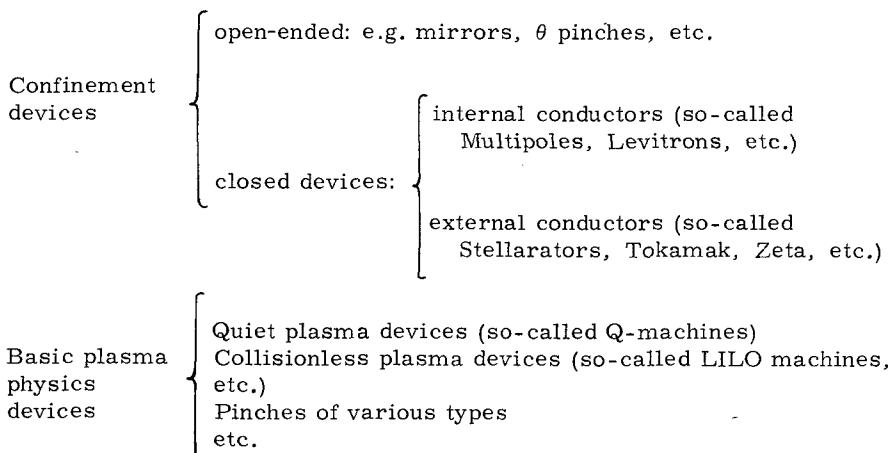


FIG.3. Sketch of a varying magnetic field profile indicating the existence of trapped particles in its local wells.

Then we can roughly say that the experimental program is centred on



The first category of devices is intended to be a viable approach to the problem of containing a plasma in a magnetic field with a sufficiently high temperature and for a sufficiently long time. The second type of devices is aimed at testing the theory, giving information on transport coefficients, modes of various types, non-linear effects, turbulence, shock waves, etc.

A simple illustration of a mirror device and a toroidal type of device is given in Figs 4 and 5. Here we can make two statements.

1. Open-ended devices of mirror type can be effectively made stable to fluid-like types of mode, as an absolute magnetic well configuration (minimum magnetic field at the point of maximum density, corresponding to directing the gravity upward for the two-fluid analogic containers mentioned above) can be realized in them as indicated in Fig.4. On the other hand, since particles with relatively large longitudinal velocity are lost through the open ends, the particle distribution function has a hole (is of loss-cone type) and is anisotropic. This makes mirrors particularly vulnerable to high-frequency velocity space instabilities. The observable effects are bursts of particles and radiation around the ion cyclotron frequency, and limitation of the particle density level that can be achieved.

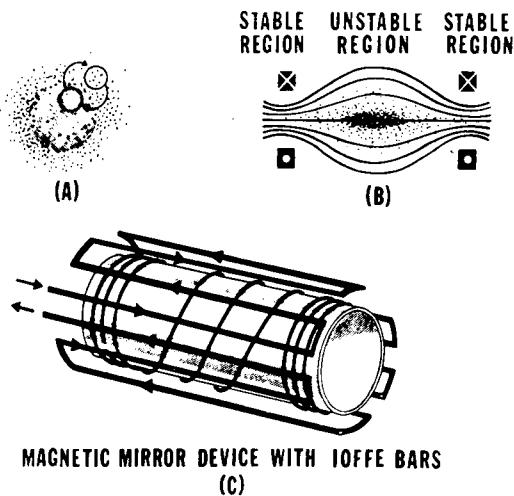


FIG.4. (A) Sketch of the fluid (interchange) instability taking plasma from the inner high-density region to the outer low-density region. (B) Regions of hydromagnetic stability in a simple configuration, without a stabilizing multipolar field. (C) Current conductors generating a multipolar minimum-B field in a mirror configuration.

2. Toroidal devices are generally immune from the loss-cone feature or strong anisotropy of the distribution function that besets mirrors. On the other hand, in order to be made stable against MHD fluid-like modes or other types of modes with macroscopic features, they require complex magnetic field configurations (see, for instance, Figs 5 and 6). This makes the theory and the interpretation of the experiments quite difficult. Therefore, a preliminary analysis of the modes and the kinds of effect that are expected in devices of this kind is often carried out on ad hoc models that contain all the main physical elements of a realistic experiment but have quite simplified geometrical features. In spite of these difficulties a number of predictions followed by favourable experimental indications have been produced recently. In addition, the need to achieve the closest correspondence between theory and experiments has opened

the way to the development of a generation of toroidal devices with internal conductors and axial (azimuthal) symmetry, of which the octopole represented by Fig.6 is an example.

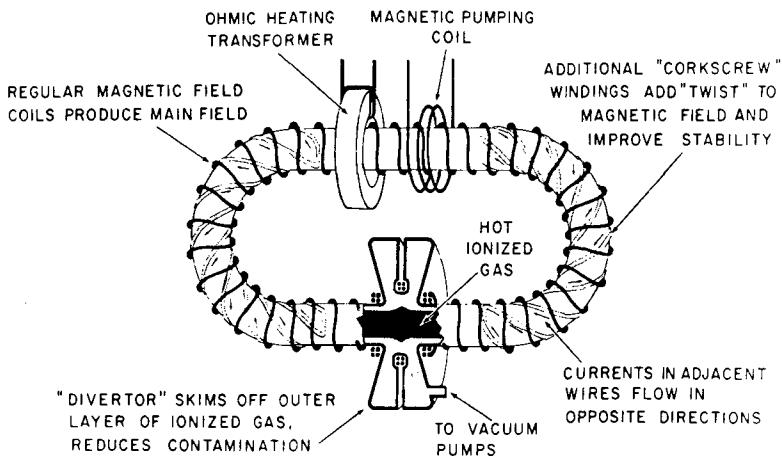


FIG.5. Schematic of a toroidal device with current conductors, generating the confining field, outside the plasma (stellarator).

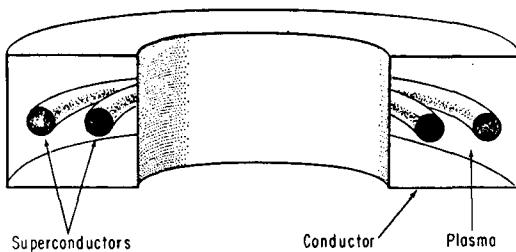


FIG.6. Schematic of a toroidal device (octopole) with current conductors, generating the confining field, inside the plasma.

Specifically, the internal conductors give the flexibility needed to realize a variety of magnetic configurations and offer some of the simplest tests of the theoretical predictions that are available. Already, as the early versions of these devices have come into operation, they have given theoreticians several good ideas, but the best information is expected from experiments with internal superconductors. These in fact lead to eliminate from the present devices any external support or current feeder to the inner conductors that destroys the symmetry and can cause spurious particle loss. Of course, these devices cannot be used as they are to achieve fusion, as it is difficult to imagine a superconductor working at about zero temperature and being surrounded by a plasma at thermonuclear temperature! Therefore, with the advent of these experiments the distinction between fusion devices and basic plasma physics devices has

become more and more difficult to make. In fact, while the latter type of devices has produced valuable information and induced theoretical work [8] relevant to plasma confinement research, the attainment of large volumes of plasmas, free from impurities to a high degree, in fusion devices such as stellarators has made them suitable for interesting experiments of basic plasma physics [17].

In conclusion, I would say that there are reasons for solid optimism for those who work in plasma physics with an interest in its application to the problem of controlled nuclear fusion. The fact is that many new collective effects predicted by the theory are now being found experimentally, the experiments become interpretable, and the theoreticians acquire a deeper interest in the experimental observations. It is not wise to make predictions on the time scale and on the outcome of this path of research, but given the direction of sound scientific understanding it has taken, it appears that the effort to achieve a new source of energy is also giving a substantial contribution to the physics of our days.

## R E F E R E N C E S

- [1] BOHM, D., in *The Characteristics of Electrical Discharges in Magnetic Fields* (GUTHRIE, A., WAKERLING, R.K., Eds), Mc Graw-Hill, New York (1949).
- [2] STODIEK, W., ELLIS, R.A., Jr., GORMAN, J., *Nucl. Fusion Suppl.*, Part 1 (1962) 193.
- [3] CHANDRASEKHAR, S., *Hydrodynamic and Hydromagnetic Stability*, Oxford Clarendon Press (1961) 428.
- [4] GOTTF, U.V., IOFFE, M.S., TEHKOVSII, V.G., *Nucl. Fusion Suppl.*, Part 3 (1962) 1045.
- [5] SAGDEEV, R.Z., GALEEV, A.A., *International Centre for Theoretical Physics Rep. IC/66/64*, Trieste (1966).
- [6] PINES, D., SCHRIEFFER, J.R., *Phys. Rev.* 125 (1962) 804.
- [7] COPPI, B., LAVAL, G., PELLAT, R., ROSENBLUTH, M.N., *International Centre for Theoretical Physics, Preprint IC/65/88*, Trieste (1966); *Nucl. Fusion* 6 (1966) 261.
- [8] COPPI, B., HENDEL, H., PERKINS, F., POLITZER, P., *Proc. Symp. Quiescent Plasma*, Frascati, CNEN, Rome, Italy (1967) 201.
- [9] ELLIS, R.A., EUBANK, H.P., *Int. Symp. Fluctuations and Diffusion in Plasmas*, Princeton University (1967), paper D7 (unpublished).
- [10] FORSEN, H., KERST, D., LENCIOMI, D., MEADE, D., MILLS, F., MOLVIK, A., SCHMIDT, J., SPROTT, J., SYMON, K., in *Plasma Physics and Controlled Nuclear Fusion Research (Proc. Conf. Novosibirsk, 1968)* 1, IAEA, Vienna (1969) 313.
- [11] YOSHIKAWA, S., BARRAULT, M., HARRIES, W., MEADE, D., PALLADINO, R., Von GOELER, S., in *Plasma Physics and Controlled Nuclear Fusion Research (Proc. Conf. Novosibirsk, 1968)* 1, IAEA, Vienna (1969) 403.
- [12] KADOMTSEV, B.B., POGUTSE, O.P., *Soviet Phys. JETP* 24 (1967) 1172.
- [13] GALEEV, A.A., SAGDEEV, R.Z., WONG, H.V., *Physics Fluids* 10 (1967) 1553; *International Centre for Theoretical Physics, Preprint IC/66/100*, Trieste (1966).
- [14] COPPI, B., FURTH, H.P., ROSENBLUTH, M.N., SAGDEEV, R.Z., *Phys. Rev. Lett.* 17 (1966) 377; *International Centre for Theoretical Physics, Preprint IC/66/79*, Trieste (1966).
- [15] COPPI, B., ROSENBLUTH, M.N., RUTHERFORD, P., *Princeton Plasma Physics Laboratory, MATT-611* (1968); *Phys. Rev. Lett.* (to be published).
- [16] ARTSIMOVICH, L.A., BOBROVSKII, G.A., MIRNOV, C.B., RAZUMOVA, K.A., STRELKOV, B.C., *Atomn. Energ.* 22 (1967) 259; translated in *Soviet atom. Energy* 22 (1967) 325.
- [17] DIMOCK, D., MAZZUCATO, E., *Phys. Rev. Lett.* 20 (1968) 713.

# SURVEY OF THE THEORY OF TURBULENCE\*

J.B. KELLER

Courant Institute of Mathematical Sciences,  
New York University,  
New York, N.Y., United States of America

## Abstract

SURVEY OF THE THEORY OF TURBULENCE. 1. Introduction; 2. Stability theory; 3. Validity of the Navier-Stokes equations; 4. Bifurcation theory and the solutions for  $R > R_c$ ; 5. Reynolds stress, eddy viscosity and mixing length; 6. Statistical theory of turbulence; 7. Status of the statistical theory of turbulence; Appendix: Stochastic equations and a theory of turbulence.

## 1. INTRODUCTION

Experiments show that at sufficiently low velocities the flow of a fluid is laminar, i.e. it is smooth streamline flow. It may be steady or it may change slowly if the external conditions are changed. At sufficiently high velocities the flow is turbulent, i.e. it is chaotic or irregular, changing rapidly in time even when the external conditions are constant. Whether the flow is laminar or turbulent is determined by the value of the dimensionless parameter  $R = UL/\nu$ , called the Reynolds number. It is constructed from a typical flow velocity  $U$ , a typical length  $L$  of the flow configuration and a typical value of the kinematic viscosity  $\nu$ . In terms of the viscosity coefficient  $\mu$  and the fluid density  $\rho$ ,  $\nu = \mu/\rho$ . For each flow configuration there is a critical value  $R_c$  of the Reynolds number such that the flow is laminar for  $R \leq R_c$  and turbulent for  $R > R_c$ .

The first problem of the theory of turbulence is to account for the "phase" transition of a flow from the laminar state to the turbulent state, and to provide a procedure for determining the transition point  $R_c$ . This aspect of the theory has been successfully developed and may be regarded as settled in principle. It is called stability theory. The explanation it offers is that for  $R \leq R_c$  the laminar state is stable, while for  $R > R_c$  it is unstable. Therefore when  $R > R_c$ , any small fluctuation in the initial velocity of the flow or in the velocity of a boundary of the fluid will grow and convert the laminar flow into a turbulent one. This theory is described in Section 2.

A second problem of the theory of turbulence is to determine the behaviour of flows with  $R$  slightly larger than  $R_c$ . The behaviour of such flows is relatively regular and their description is therefore relatively simple. This is because these flows can be viewed as laminar flows combined with a few unstable modes having small amplitudes. As a consequence, this part of the theory is also developing satisfactorily, although it has not been developed extensively. It is called non-linear stability theory. Some aspects of this theory are described in Section 4.

\* The research in this paper was supported by the National Science Foundation under Grant No. GP-8062.

The main problem of turbulence theory is to describe flows with  $R$  much larger than  $R_c$ . The turbulent motion in such flows is said to be fully developed. Although the development of a theory of fully developed turbulence is the major goal of turbulence theory, that goal is nowhere near being achieved. The difficulty is not due to the inapplicability of the macroscopic equations of fluid dynamics, as is explained in Section 3. The status of this theory is described in Sections 5 - 7, and a recent theory is presented in the Appendix.

Fully developed turbulence occurs in the fluid flows studied in oceanography, meteorology, hydraulic engineering, civil engineering, mechanical engineering, astrophysics, etc. The lack of a theory of such flows is a major barrier to progress in all these fields. It is fair to say that the problem of devising an adequate theory of fully developed turbulence is the major unsolved problem of classical physics.

## 2. STABILITY THEORY

To determine  $R_c$  for a fluid flow, stability theory employs a macroscopic description of the fluid. Thus the state of the fluid is described by its velocity  $\vec{u}(\vec{x}, t)$ , its pressure  $p(\vec{x}, t)$ , its density  $\rho(\vec{x}, t)$  and its temperature  $T(\vec{x}, t)$ . These quantities are assumed to satisfy the equations of conservation of mass, energy, linear momentum and angular momentum and the equation of state of the fluid. To make these equations determinate, the stress tensor occurring in the momentum is expressed in terms of  $p$  and  $\nabla \vec{u}$  by the Navier-Stokes formula. Similarly the heat flux vector, which occurs in the energy equation, is expressed in terms of  $\nabla T$ . In this way the angular momentum equation is satisfied identically and there remain six scalar equations for  $p$ ,  $\rho$ ,  $T$  and the three components of  $\vec{u}$ . Together with appropriate initial conditions at  $t = t_0$  and conditions on the boundary of the domain containing the flow, they determine the flow. Presumably they determine it uniquely for all  $t \geq t_0$ , although this has not been proved in general.

When the boundary conditions and external forces are independent of time, there may exist a solution  $\vec{u}_0(\vec{x})$ ,  $p_0(\vec{x})$ ,  $\rho_0(\vec{x})$ ,  $T_0(\vec{x})$  which is also independent of time. If so, this is a laminar solution. To determine its stability, we consider the variational equations, which are linear equations obtained from the Navier-Stokes and the other equations governing the flow. They can be obtained from those equations by setting  $\vec{u} = \vec{u}_0(\vec{x}) + \epsilon \vec{u}_1(\vec{x}, t) + O(\epsilon^2)$ ,  $p = p_0(\vec{x}) + \epsilon p_1(\vec{x}, t) + O(\epsilon^2)$ , etc. Then the terms linear in  $\epsilon$  yield the variational equations for the quantities  $\vec{u}_1$ ,  $p_1$ ,  $\rho_1$  and  $T_1$ . The coefficients in these equations are independent of  $t$  since they involve only the laminar flow  $\vec{u}_0(\vec{x})$ ,  $p_0(\vec{x})$ , etc. Therefore these equations, together with the boundary conditions, are translationally invariant in  $t$ . As a consequence, they have a complete set of solutions which are eigenfunctions of translation in  $t$ , i.e. exponential functions of  $t$ . In other words, the variational equations can be solved by separation of variables, with  $\vec{u}_1(\vec{x}, t) = e^{\alpha t} \vec{U}(\vec{x})$ ,  $p_1(\vec{x}, t) = e^{\alpha t} P(\vec{x})$ , etc.

The exponent  $\alpha$  is determined as an eigenvalue of the resulting linear problem for  $\vec{U}(\vec{x})$ ,  $P(\vec{x})$ , etc. Usually there are a denumerable set of eigenvalues  $\alpha_j$ ,  $j = 1, 2, \dots$ , and they are complex. They depend upon the

laminar flow which occurs in the coefficients of the linear equations. Therefore in particular they depend upon the Reynolds number  $R$  of the basic flow, so we may write  $\alpha_j(R)$ ,  $j = 1, 2, \dots$ . The laminar flow is said to be infinitesimally stable if  $\text{Re } \alpha_j(R) < 0$  for all  $j$ . It is unstable if  $\text{Re } \alpha_j(R) > 0$  for some  $j$ . Thus the critical Reynolds number  $R_c$  is the least upper bound of the values of  $R$  for which the flow is stable. Alternatively it is the smallest value of  $R$  for which  $\text{Re } \alpha_j(R) = 0$  for some  $j$ . We shall label the  $\alpha$ 's so that this value of  $j$  is unity. (Essentially the same considerations apply when the spectrum is continuous.)

According to this definition, the first mode or eigenfunction of the variational problem is damped for  $R < R_c$ , undamped at  $R = R_c$ , and exponentially growing for  $R > R_c$ . There are other critical values of  $R$  which we may call  $R_2, R_3$ , etc. at which the second, third, etc. mode changes from being damped to growing exponentially. We may label the  $\alpha$ 's so that the  $R_j$  form an increasing sequence with  $R_1 = R_c$  being the smallest. Only  $R_c$  is important in determining the transition from laminar to turbulent flow.

The values of  $R_c$ , calculated in the manner just described, agree with experimentally measured values in a variety of problems. This confirms the belief that stability theory accounts for the transition and determines  $R_c$  in principle. This theory is applied to many problems in the books of Lin [1] and Chandrasekhar [2].

### 3. VALIDITY OF THE NAVIER-STOKES EQUATIONS

In many circumstances temperature gradients in a flow are so small that the heat flux is negligible compared to the mechanical flux of energy. Then the flow can be treated as adiabatic. If in addition the entropy density is initially constant, it remains constant, so the flow is also isentropic. Then the energy equation is identically satisfied and the equation of state can be put in its adiabatic form.

Very often the flow velocity is small compared to the sound speed. Then compressibility can be ignored and the density of the fluid can be treated as a constant.

When the flow is both isentropic and of constant density  $\rho$ , the Navier-Stokes equations are

$$\vec{u}_t + (\vec{u} \cdot \nabla) \vec{u} - \nu \Delta \vec{u} = -\frac{1}{\rho} \nabla p + \vec{f} \quad (1)$$

$$\nabla \cdot \vec{u} = 0 \quad (2)$$

This is a set of four scalar equations for  $p$  and the three components of  $\vec{u}$ , the external force  $\vec{f}$  being assumed given. Most of stability theory and turbulence theory are based upon these equations.

Agreement with experiment has shown that Eqs (1) and (2) are valid in stability theory. It may be asked whether they are valid for the description of turbulent flow. An examination of the smallest scales of motion in actual turbulent flows shows that they are many orders of magnitude larger than the mean free path of a molecule in the fluid. Therefore a macroscopic or continuum description of the fluid is adequate. An examination of the

temperature gradients and density variations in actual turbulent flows shows that in most cases they are small enough for the simplified equations (1) and (2) to be valid. This reasoning is confirmed by the agreement of predictions based upon Eqs (1) and (2) with various measurements on turbulent flows.

#### 4. BIFURCATION THEORY AND THE SOLUTIONS FOR $R > R_c$

If  $R$  is slightly larger than  $R_c$ , linear stability theory shows that the first mode of the variational problem will grow exponentially. As soon as its amplitude becomes large, linear theory ceases to be valid and the non-linear equations must be used. They show that the amplitude of the unstable mode does not increase indefinitely. Instead, as  $t$  increases, the amplitude approaches an asymptotic value proportional to  $(R - R_c)^{1/2}$ . Thus the flow tends to a new steady state consisting of the laminar flow combined with the unstable mode, the latter having a small amplitude.

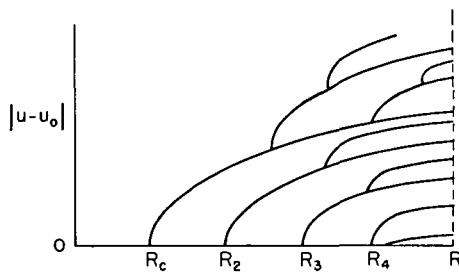


FIG.1. Sketch of the norm of  $|u - u_0|$  as a function of  $R$ . The curves are not calculated but just drawn in accordance with intuitive ideas.

This analysis shows that there are two time-independent flows for  $R > R_c$ , the laminar flow and the flow just described, and they become identical at  $R = R_c$ . Thus the point  $R = R_c$  is a bifurcation point, i.e. a point at which one solution – the laminar one for  $R \leq R_c$  – splits into two solutions. The study of this splitting of solutions is called bifurcation theory in mathematics. Presumably the second solution continues to exist for all  $R > R_c$  although that has not been proved. Additional steady solutions also bifurcate from the laminar solution at the higher critical points  $R_j$ ,  $j > 1$ . All these solutions are sketched in Fig. 1 in which the norm of  $\vec{u} - \vec{u}_0$  is shown as a function of  $R$ . (The norm may be the maximum value of  $|\vec{u} - \vec{u}_0|$  or any other norm.)

It is generally believed that bifurcation also occurs on the solution branches which split off from the laminar solution, then from the new branches, etc. (see Fig. 1). Thus each new steady flow becomes unstable for  $R$  large enough and a new stable solution replaces it.

Hopf [3] has constructed a model set of equations analogous to the Navier-Stokes equations, which can be solved explicitly. As a parameter  $R$  is increased, new solutions come into existence each time  $R$  passes through a critical value  $R_j$ . Each solution is a sum of  $j$  oscillatory trigonometric

functions, so it is almost periodic. The larger  $R$ , the more terms there are in the solution, so the less regular it is.

On the basis of intuitive considerations, supported by this model, it is generally believed that turbulent motion develops with increasing  $R$  in the same way as do the model solutions. This development is sometimes described by saying that the number of steady solutions increases with  $R$  and that the turbulent flow "jumps" from one of these solutions to another. Of course the turbulent flow cannot "jump", but it may approximate one solution for a while, then another, etc.

In non-linear stability theory, the time-dependent and the steady solutions for  $R$  in the vicinity of  $R_c$  are determined by appropriate perturbation methods. The deviation of the flow from the laminar flow is treated as small. The perturbation results appear to be in agreement with experiment.

## 5. REYNOLDS STRESS, EDDY VISCOSITY AND MIXING LENGTH

Fully developed turbulent flow occurs for  $R \gg R_c$ . To describe it we follow Reynolds and express  $\vec{u}(\vec{x}, t)$  and  $p(\vec{x}, t)$  as the sum of a mean flow  $\langle \vec{u}(\vec{x}, t) \rangle \langle p(\vec{x}, t) \rangle$  and a fluctuating flow  $\vec{u}'(\vec{x}, t)$ ,  $p'(\vec{x}, t)$ . We also write the force  $\vec{f}$  as the sum of  $\langle \vec{f}(\vec{x}, t) \rangle$  and  $\vec{f}'(\vec{x}, t)$ . We insert  $\vec{u} = \langle \vec{u} \rangle + \vec{u}'$ ,  $p = \langle p \rangle + p'$  and  $\vec{f} = \langle \vec{f} \rangle + \vec{f}'$  into Eqs (1) and (2) and then average these equations, assuming that  $\langle \vec{u}' \rangle = \langle p' \rangle = 0$ . In this way we obtain the Reynolds equations for the mean flow, which are

$$\langle \vec{u} \rangle_t + (\langle \vec{u} \rangle \cdot \nabla) \langle \vec{u} \rangle - \nu \Delta \langle \vec{u} \rangle = -\frac{1}{\rho} \nabla \langle p \rangle + \nabla \cdot \tau + \langle \vec{f} \rangle \quad (3)$$

$$\nabla \cdot \langle \vec{u} \rangle = 0 \quad (4)$$

Here the tensor  $\tau$ , called the Reynolds stress, has the components

$$\tau_{ij}(\vec{x}, t) = \langle u_i^j(\vec{x}, t) u_j^i(\vec{x}, t) \rangle \quad (5)$$

The Reynolds equations are of the same form as the Navier-Stokes equations except for the additional Reynolds stress term. Since the Reynolds stress is undetermined, the Reynolds equations are inadequate for the determination of the mean flow. We may say that they are not "closed". To be closed, they must be supplemented by an equation for  $\tau_{ij}$ .

The first closure hypothesis was that  $\tau_{ij}$  is related to  $\langle u_i \rangle$  in the same way as the viscous stress tensor is related to  $u_i$ , i.e.

$$\tau_{ij} = \nu_e \frac{1}{2} (\partial_j \langle u_i \rangle + \partial_i \langle u_j \rangle) \quad (6)$$

Here  $\nu_e$  is a new coefficient called the eddy viscosity. When Eq.(6) is used in Eq.(3) it becomes

$$\langle \vec{u} \rangle_t + (\langle \vec{u} \rangle \cdot \nabla) \langle \vec{u} \rangle - (\nu + \nu_e) \Delta \langle \vec{u} \rangle = -\frac{1}{\rho} \nabla \langle p \rangle + \langle \vec{f} \rangle \quad (7)$$

Equations (7) and (4) are a determined set of equations for the mean flow, provided the constant  $\nu_e$  is given. Generally it is determined to obtain

agreement with experiment. (Experiment shows that usually  $\nu_e \gg \nu$  so that molecular viscosity is usually negligible compared to eddy viscosity.)

In about 1911 both L. Prandtl and G.I. Taylor proposed the mixing length theory for the determination of  $\tau_{ij}$ . They assumed that the Reynolds stress is produced by momentum transfer from high momentum regions to low momentum regions, just as viscous stress is produced. In analogy to the mean free path of a molecule, they introduced the mixing length  $\lambda$ , which is the average distance over which an eddy transfers momentum. Then their closure hypothesis for  $\tau_{ij}$  was the same as Eq. (6), but  $\nu_e$  was expressed as the product of  $\lambda$  and a typical fluctuating velocity, for example

$$\nu_e = \lambda \sqrt{\langle (\vec{u}')^2 \rangle} \quad (8)$$

Of course  $\lambda$  and  $\langle (\vec{u}')^2 \rangle$  are both unknown, so this does not close the system of equations. However, it indicates that  $\nu_e$  may vary with position and time. If so, then  $\nu_e$  might be expressed in terms of  $\langle \vec{u} \rangle$  and its derivatives. For example, an expression of this kind for  $\nu_e$  with the correct dimension is

$$\nu_e = c \langle \vec{u} \rangle^{3/2} (\partial_j^2 \langle \vec{u} \rangle)^{-1/2} \quad (9)$$

Here  $c$  is a dimensionless constant.

We can now close the Reynolds equations (3) and (4) by using Eq.(6) for  $\tau_{ij}$  and some expression such as Eq.(9) for  $\nu_e$ . This closure yields equations for the mean flow. It has been possible to obtain fair agreement between the solutions of these equations and experiment in many cases. Of course a constant such as  $c$  in Eq.(9) has to be adjusted differently in each case. This is the only theory now in use for the determination of the mean flow. There is a large body of empirical knowledge about the best form of  $\nu_e$  to use in various circumstances and about the numerical values of the constants in it.

An obvious alternative to the closure based upon the assumption (6) is to obtain a differential equation for  $\tau_{ij}$ . We can do this easily by first differentiating the definition (5), which yields

$$\partial_t \tau_{ij} = \langle u'_j \partial_t u'_i + u'_i \partial_t u'_j \rangle \quad (10)$$

Then we note that  $\partial_t u'_i = \partial_t u_i - \partial_t \langle u_i \rangle$ . We can obtain  $\partial_t u_i$  from Eq.(1) and  $\partial_t \langle u_i \rangle$  from Eq.(3). Upon using these expressions in Eq.(10) and rearranging the result, we obtain

$$\partial_t \tau_{ij} + \langle u_k \rangle \partial_k \tau_{ij} - \nu \Delta \tau_{ij} + \tau_{ik} \partial_k \langle u_j \rangle + \tau_{jk} \partial_k \langle u_i \rangle = \sigma_{ij} \quad (11)$$

Here  $\sigma_{ij}$ , which we shall call the source of the Reynolds stress, is defined by

$$\sigma_{ij} = \partial_k \langle u'_i u'_j u'_k \rangle + 2\nu \langle (\partial_k u'_i)(\partial_k u'_j) \rangle + \rho^{-1} \langle u'_j \partial_i p' + u'_i \partial_j p' \rangle - \langle u'_j f'_i + u'_i f'_j \rangle \quad (12)$$

It also follows from Eq.(5) that  $\tau_{ij} = \tau_{ji}$  and

$$\partial_i \tau_{ij} = \partial_j \tau_{ij} = 0 \quad (13)$$

The new equations (11) and (13) for  $\tau_{ij}$ , together with Eqs (3) and (4) for  $\langle \vec{u} \rangle$  and  $\langle p \rangle$ , are still not closed because  $\sigma_{ij}$  occurs in Eq.(11). It may be expected that a more accurate theory would be obtained by making a closure hypothesis about  $\sigma_{ij}$ . This possibility has not been explored adequately. The resulting equations would be analogous to H. Grad's thirteen moment fluid dynamic equations, obtained from the Boltzmann equation.

A closure hypothesis asserts the existence of a more or less universal relation among various properties of the mean flow. In view of its universality, it must hold in the simplest turbulent flow. If the flow is too simple, the relation will be trivial because the terms in it will vanish, while if the flow is too complex, it cannot be analysed. Therefore we consider the simplest flow in which the quantities under consideration are not trivial. Then the relation between those quantities in this flow determines the universal relation.

This consideration provides a method for obtaining closure relations by an analytical or experimental study of a simple flow. It could be used to obtain a relation between  $\tau_{ij}$  and  $\nabla \langle \vec{u} \rangle$ , if such a relation exists, and this would provide a closure of the Reynolds equations. Alternatively, if Eq.(11) is used for the determination of  $\tau_{ij}$ , this method could be used to obtain a closure relation expressing  $\sigma_{ij}$  in terms of  $\tau_{ij}$ ,  $\langle u_i \rangle$ , and some derivatives of  $\tau_{ij}$  and  $\langle u_i \rangle$ .

## 6. STATISTICAL THEORY OF TURBULENCE

The averages occurring in the Reynolds equations have sometimes been interpreted as space or time averages. Usually, however, they are interpreted as ensemble averages. To make this interpretation precise, we must introduce a statistical mechanical theory of the solutions of the Navier-Stokes equations. Therefore we must first formulate an initial-boundary value problem for these equations, which we shall now do.

Let us consider a domain  $D$  with a boundary  $\partial D$  and let us prescribe the velocity  $u$  throughout the domain at  $t = 0$  and on the boundary for  $t \geq 0$ :

$$\vec{u}(\vec{x}, 0) = \vec{g}(\vec{x}), \quad \vec{x} \text{ in } D \quad (14)$$

$$\vec{u}(\vec{x}, t) = \vec{h}(\vec{x}, t), \quad \vec{x} \text{ on } \partial D, \quad t \geq 0 \quad (15)$$

In order that these conditions be compatible with the incompressibility condition  $\nabla \cdot \vec{u} = 0$ , and with each other, the given functions  $\vec{g}$  and  $\vec{h}$  must satisfy the conditions

$$\nabla \cdot \vec{g} = 0, \quad \vec{x} \text{ in } D \quad (16)$$

$$\int_{\partial D} \vec{h} \cdot d\vec{s} = 0, \quad x \geq 0 \quad (17)$$

$$\vec{g}(\vec{x}) = h(x, 0), \quad x \text{ on } \partial D \quad (18)$$

The initial-boundary value problem is to find  $\vec{u}(\vec{x}, t)$  and  $p(\vec{x}, t)$  for  $\vec{x}$  in  $D$  and  $t \geq 0$ , satisfying the Navier-Stokes equations (1) and (2), the initial

condition (14) and the boundary condition (15). The force  $\vec{f}(\vec{x}, t)$  in Eq.(1) and the functions  $\vec{g}$  and  $\vec{h}$  in Eqs (14) and (15) are assumed to be given and to satisfy Eqs (16) to (18). For suitable domains and suitable prescribed functions, this problem has a solution which is unique up to an additive constant in  $p$ .

The statistical theory of turbulence consists in permitting the initial velocity  $\vec{g}(\vec{x})$  and possibly the boundary velocity  $\vec{h}(\vec{x}, t)$  and the external force  $\vec{f}(\vec{x}, t)$  to be random functions with a prescribed joint probability distribution. Then the solution  $\vec{u}(\vec{x}, t)$ ,  $p(\vec{x}, t)$  of the above problem will also be random. The first objective of the theory is to determine the probability distribution of the solution. The second objective is to use this distribution to calculate various statistical properties of the solution, such as its mean, its two-point correlations, etc. These two problems are strictly mathematical problems, once the probability distribution of the data is given. The first is analogous to the determination of the N-particle distribution function in statistical mechanics, which involves solving the Liouville equation. The analogous equation for the statistical theory of turbulence has been derived by Hopf [4] under the condition that only the initial velocity is random and the flow is unbounded. Since the velocity field is a function, its probability distribution is a functional defined on a function space and Hopf's equation is a functional differential equation. (It is derived for the characteristic functional, which is the Fourier transform of the probability distribution functional.)

Up to now, Hopf's equation has been of as little use in turbulence theory as Liouville's equation has been in statistical mechanics. All the results of practical value in statistical mechanics have been obtained from equations for the one- or two-particle distribution functions,  $f_1$  and  $f_2$ . An equation for  $f_1$  can be obtained by integrating the Liouville equation, but it involves  $f_2$ . Various closure schemes for eliminating  $f_2$  have been devised for different circumstances by Boltzmann, Vlasov, Bogolyubov, Lenard and Balescu and others. All of them involve hypotheses about  $f_2$  which have never been proved and which cannot be exactly correct. Similarly Kirkwood and Born and Green have derived equations for  $f_2$  by integrating the Liouville equation and then making assumptions about  $f_3$ .

In the same way, almost all of the work in the statistical theory of turbulence has concerned the derivation of equations for the mean velocity  $\langle u_i \rangle$  and the two-point velocity correlation tensor  $R_{ij}(\vec{x}, t, \vec{x}', t')$  defined by

$$R_{ij}(\vec{x}, t, \vec{x}', t') = \langle u_i(\vec{x}, t) u_j(\vec{x}', t') \rangle \quad (19)$$

When  $\vec{x} = \vec{x}'$  and  $t = t'$  this becomes the Reynolds stress  $\tau_{ij}$  which we have already considered. Just as in the case of  $\tau_{ij}$ , the equations for  $R_{ij}$  obtained from the Navier-Stokes equations involve triple correlations. An attempt to obtain equations for them, leads to an infinite hierarchy of equations. Therefore various closure procedures have been devised for obtaining a finite closed set of equations for  $R_{ij}$  and  $\langle u_i \rangle$ .

The study of  $R_{ij}$  (with  $t = t'$ ) was initiated in 1935 by G.I. Taylor, who introduced the concept of homogeneous isotropic turbulence. For such turbulence it follows from invariance that  $\langle u_i \rangle = 0$  and  $R_{ij}(\vec{x}, t, \vec{x}', t') = R_{ij}(\vec{x} - \vec{x}', t) = F(\vec{r}, t) r_i r_j + G(\vec{r}, t) \delta_{ij}$  where  $\vec{r} = \vec{x} - \vec{x}'$  and  $F$  and  $G$  are scalar functions. These two scalar functions are related by

$4rF + r^2 F_r + G_r = 0$  as a consequence of the condition  $\partial_i R_{ij} = 0$ , which follows from the incompressibility condition  $\partial_i u_i = 0$ . Thus for homogeneous isotropic turbulence,  $R_{ij}$  is determined by one scalar function.

Because homogeneous isotropic turbulence is apparently so much simpler than other kinds of turbulence, most of the theoretical analyses have been concerned with it. One of the most important results of this analysis concerns the spectral energy density  $E(k, t)$ , defined by

$$E(k, t) = \frac{1}{\pi} \int_0^\infty R_{ii}(r, t) kr \sin kr dr \quad (20)$$

It is based either on dimensional analysis and a similarity hypothesis, or else on an assumption about energy flow from larger to smaller wave-numbers. If the Reynolds number is large enough, this analysis predicts that there is a range of  $k$  for which  $E$  has the form

$$E(k, t) \sim c(t) k^{-5/3} \quad (21)$$

This result, which has had some experimental confirmation, was obtained by Kolmogoroff [5] in 1941 and also by Obukhoff [6], von Weizsäcker [7] and Onsager [8].

Experiments show that the probability distribution of the fluid velocity at a point in a turbulent flow is nearly Gaussian. This fact has been the basis for the closure procedures of Millionshtchikov [9], Heisenberg [10], Proudman and Reid [11], Chandrasekhar [12] and Tatsumi [13]. In one form or another these authors assumed that fourth-order moments of the velocity are related to second-order moments in the same way as if the process were Gaussian. This attractive idea has led to equations whose solutions are unsatisfactory because they can yield negative values of quantities like energy density, which must be non-negative. As a result these closure procedures have come into disrepute.

Recently the nearly Gaussian character of turbulence has been made the basis of a new closure procedure by Siegel et al. [14]. They proposed to represent the velocity by a Wiener-Hermite expansion. This is a representation of a random function by a series analogous to a series of Hermite polynomials multiplied by a Gaussian function. They used it to obtain a closed set of equations for the two-point correlation function for a scalar equation proposed by J. Burgers as a model of the Navier-Stokes equations. G. Deem (unpublished) has used their method on the Navier-Stokes equations to obtain an equation for  $R_{ij}$  for homogeneous isotropic turbulence. Numerical solutions of this equation seem to be physically acceptable because they do not yield negative energy densities, and the equations have some desirable invariance properties.

One of the best-known closure methods is that of Kraichnan [15] who obtained a closed set of equations for  $R_{ij}(\vec{x}, t, \vec{x}', t')$  in homogeneous isotropic turbulence. His derivation neglects phase correlations of triplets of spatial Fourier components of the velocity except for those triplets which can interact directly, i.e. those which conserve momentum. This is called the direct interaction approximation. Kraichnan [16] also derived the same equations from a stochastic model which consisted of an ensemble of flows with stochastic couplings among them. This second derivation

guarantees that the equations cannot yield unphysical results. On the other hand, it also indicates that the quantitative results will be inaccurate because they are correct for a problem with the additional random couplings. In fact the equations predicted the exponent  $-5/2$  in Eq.(21) instead of  $-5/3$ . As a result Kraichnan [17] obtained a new set of more complicated equations which do predict  $k^{-5/3}$  in Eq.(21). But this new set has not been derived from a model, so there is no guarantee that it will not yield unphysical results.

Other closure methods have been devised using perturbation expansions and partial resummations, usually with the aid of diagrams.

To illustrate closure schemes and the kind of equations they lead to, a method devised by Keller [18] is presented in the Appendix, together with the equations it yields.

## 7. STATUS OF THE STATISTICAL THEORY OF TURBULENCE

There is no doubt that turbulence can be adequately described by a statistical theory based upon the Navier-Stokes equations. This conclusion is supported by a great many comparisons between observations and theoretical conclusions derived from those equations. These conclusions are generally kinematic relations, and sometimes even dynamic relations, among statistical quantities. However, there is not yet any statistical theory which can deduce the properties of turbulence from first principles.

One fundamental difficulty is that the probability distribution of the velocity field, which in principle can be found by solving Hopf's equation, depends upon the prescribed initial probability distribution. It is not known how to prescribe the initial distribution. Presumably there is some probability distribution which is approached after some time by the solution starting from almost any initial distribution. This ultimate distribution may then be called stable, and it is that distribution which the theory seeks. Of course that distribution depends upon the size and shape of the domain, the boundary conditions, etc.

The ultimate probability distribution is not an equilibrium distribution. This is because the Navier-Stokes equations are dissipative, and the only equilibrium state is the state of rest. As a consequence, there are no analogues of the results of equilibrium statistical mechanics which are based upon the Gibbs distribution. There is an equilibrium distribution for non-viscous fluids (i.e. fluid flows with infinite Reynolds numbers) and this could be used as a basis for an equilibrium theory. However, so far no method has been proposed for finding the stable distribution.

Every attempt at formulating a theory of turbulence has led to a system of equations for  $R_{ij}$  by some closure procedure, and in most theories only homogeneous isotropic turbulence is considered. The resulting equations are so complex that only a very limited number of solutions have been obtained from any of them, and they have been obtained by extensive numerical methods utilizing high speed computers. Many theories have led to equations which have not been solved at all. Because of the unphysical results yielded by some theories, a new viewpoint about closure schemes has developed. It is that the scheme should not be emphasized, but rather that the resulting equations should be required to

satisfy a certain number of conditions. Typical conditions are Galilean invariance, mass and momentum conservation, positive energy density, etc. Of course this viewpoint does not lead to a particular set of equations, but limits the admissible ones. Probably any correct closure scheme would automatically yield equations satisfying all these conditions. It is presumably believed that the correct set of equations will yield that  $R_{ij}$  which corresponds to a stable probability distribution of the kind considered above, although this question has not been discussed.

There is a growing belief that the study of homogeneous isotropic turbulence is a blind alley. This belief is due in part to the fact that a satisfactory theory of this type of turbulence has not been developed so far. As a consequence, this study has not shed much light upon the more realistic inhomogeneous anisotropic turbulence. This belief is also based upon the hope that the interaction between the mean flow and the turbulent fluctuations may be more important than the interaction of fluctuations with themselves. Only the latter occurs in homogeneous turbulence. Therefore it may be the case that the study of more realistic turbulence will be easier than the study of homogeneous turbulence. In any case, the latter study can yield no results about interactions between the mean flow and the fluctuations.

It is sometimes suggested that the turbulence problem will be solved when sufficiently large and fast computers become available. Then perhaps actual turbulent solutions of the Navier-Stokes equations will be computed. For this to be feasible the mesh spacing will have to be small compared to the smallest significant eddy size or wavelength of the flow and the time steps will have to be correspondingly small. In addition the computational method will have to be very stable so that numerical instability will not mask the hydrodynamic instability. As a consequence of these stringent requirements, it will require several orders of magnitude improvement in computer size and speed to make such calculations possible. Even then it is doubtful whether such computations will provide what may be called a theory of turbulence.

The current theories of turbulence attempt to surmount the computational difficulties just described by seeking equations for smoother functions, such as  $R_{ij}$ . The equations for these quantities can be solved numerically with a much coarser mesh than is necessary for the turbulent velocity itself. Much larger time steps can also be used. However, these advantages are completely offset by the increase in the number of independent variables from the four variables  $\vec{x}, t$  in the case of the turbulent velocity to the eight variables  $\vec{x}, t, \vec{x}', t'$  in the case of  $R_{ij}$ . Even with a mesh having only ten points along each co-ordinate axis,  $10^8$  mesh points would be needed. For higher order correlations the number of variables and mesh points is correspondingly greater. Only by restricting attention to the homogeneous isotropic case has it been possible to reduce the number of independent variables in  $R_{ij}$  to two. For more realistic turbulence no such reduction is possible.

These pessimistic observations suggest to me that some new ideas and insights are needed in the theory of turbulence. Despite the importance of the subject and its long history, it has been given relatively little attention by theoreticians. Therefore it is likely that with more intensive investigation the new ideas and insights will be found.

## Appendix

## STOCHASTIC EQUATIONS AND A THEORY OF TURBULENCE

For each  $\alpha$  let  $L(\alpha)$  be a linear operator which maps elements  $u$  of a space  $S$  into itself. We call  $L(\alpha)$  a stochastic operator if  $\alpha$  is a random variable with probability distribution  $p(\alpha)$ . For a given element  $g$  of  $S$  we consider the stochastic equation

$$L(\alpha) u = g \quad (1)$$

If  $L$  is invertible for each  $\alpha$ , as we assume, the solution of Eq.(1) is the random element

$$u(\alpha) = L^{-1}(\alpha) g \quad (2)$$

We shall seek the mean solution  $\langle u \rangle$ , defined by

$$\langle u \rangle = \int u(\alpha) p(\alpha) d\alpha \quad (3)$$

From Eq.(2) we obtain

$$\langle u \rangle = \langle L^{-1} \rangle g \quad (4)$$

We shall find it advantageous to multiply Eq.(4) on the left by  $\langle L^{-1} \rangle^{-1}$  to obtain

$$\langle L^{-1} \rangle^{-1} \langle u \rangle = g \quad (5)$$

Although Eq.(5) is an exact equation for  $\langle u \rangle$ , it is practically useless because of the difficulty of computing  $\langle L^{-1} \rangle^{-1}$ , which occurs in it.

To make Eq.(5) useful, we assume that  $L(\alpha)$  is the sum of a non-random invertible operator  $M$  and a relatively small random operator  $V(\alpha)$

$$L(x) = M + V(\alpha) \quad (6)$$

Then we have, using the binomial theorem,

$$L^{-1} = [M(1 + M^{-1}V)]^{-1} = (1 + M^{-1}V)^{-1} M^{-1} = \sum_{n=0}^{\infty} (-M^{-1}V)^n M^{-1} \quad (7)$$

Taking the mean of Eq.(7) yields

$$\langle L^{-1} \rangle = \sum_{n=0}^{\infty} \langle (-M^{-1}V)^n \rangle M^{-1} \quad (8)$$

Finally, inverting both sides of Eq.(8), again using the binomial theorem, leads to

$$\langle L^{-1} \rangle^{-1} = M \left[ 1 + \sum_{n=1}^{\infty} \langle (-M^{-1}V)^n \rangle \right]^{-1} = M \sum_{q=0}^{\infty} \left[ - \sum_{n=1}^{\infty} \langle (-M^{-1}V)^n \rangle \right]^q \quad (9)$$

Upon using Eq.(9) in Eq.(5) we obtain the following explicit equation for  $\langle u \rangle$ :

$$M \sum_{q=0}^{\infty} \left[ - \sum_{n=1}^{\infty} \langle (-M^{-1} V)^n \rangle \right]^q \langle u \rangle = g \quad (10)$$

Let us set  $\langle V \rangle = 0$ , which can be done without loss of generality by defining  $M = \langle L \rangle$ . Then if terms of order  $(M^{-1} V)^3$  are neglected, Eq.(10) becomes simply

$$(M - \langle VM^{-1}V \rangle) \langle u \rangle = g \quad (11)$$

Exactly the same procedure can be used to obtain equations satisfied by higher moments of  $u$ . We shall illustrate it by obtaining an equation for the second moment. In doing so it is convenient to place a bar over certain operators and vectors and to use the convention that a barred operator acts only on barred vectors and an unbarred operator acts only on unbarred vectors. This permits us to interchange the orders of certain non-commuting quantities without committing an error. Thus we may rewrite Eq.(2) in the form

$$\bar{u} = \bar{L}^{-1} \bar{g} \quad (12)$$

Now multiplying  $u$  given by Eq.(2) by  $u$  given by Eq.(12), we obtain an element of the product space; it is given by

$$u\bar{u} = Lg\bar{L}^{-1} \bar{g} = \bar{L}\bar{L}^{-1} g\bar{g} \quad (13)$$

The second form of the right side of Eq.(13) is obtained by using the convention introduced above. Now averaging Eq.(13) yields

$$\langle u\bar{u} \rangle = \langle \bar{L}\bar{L}^{-1} \rangle g\bar{g} \quad (14)$$

Multiplying Eq.(14) by  $\langle \bar{L}\bar{L}^{-1} \rangle^{-1}$  yields an equation for  $\langle u\bar{u} \rangle$

$$\langle \bar{L}\bar{L}^{-1} \rangle^{-1} \langle u\bar{u} \rangle = g\bar{g} \quad (15)$$

In the same way as we evaluated  $\langle L^{-1} \rangle^{-1}$ , we can evaluate  $\langle \bar{L}\bar{L}^{-1} \rangle^{-1}$ . Then Eq.(15) becomes

$$\bar{M} \sum_{q=0}^{\infty} \left[ -M \sum_{n+j=1}^{\infty} \langle (-M^{-1} V)^n M^{-1} (-\bar{M}^{-1} \bar{V})^j \rangle \right]^q M \langle u\bar{u} \rangle = g\bar{g} \quad (16)$$

When  $\langle V \rangle = 0$  and terms of order  $(M^{-1} V)^3$  are neglected, Eq.(16) becomes

$$[\bar{M}M - \langle VM^{-1}V \rangle M - \bar{M}\langle \bar{V}\bar{M}^{-1}V \rangle - \bar{M}\langle VM^{-1}\bar{M}^{-1}\bar{V} \rangle M] \langle u\bar{u} \rangle = g\bar{g} \quad (17)$$

Equations (11) and (17) have been used to analyse wave propagation in random media. When the random inhomogeneities, represented by  $V$ , have long-range correlations, it is necessary to replace these equations

by other ones appropriate to that case. The reason for this is that the operator  $M^{-1}$  in Eqs (11) and (17) represents propagation in the unperturbed medium, whereas long-range propagation must be described by the modified propagator appropriate to the perturbed medium. To this end we rewrite Eq.(8) in the form

$$M^{-1} = \langle L^{-1} \rangle - \sum_{n=1}^{\infty} \langle (-M^{-1} V)^n \rangle M^{-1} \quad (18)$$

This equation can be solved by iterations for  $M^{-1}$  in terms of  $\langle L^{-1} \rangle$ , so we shall write the solution as  $M^{-1}(\langle L^{-1} \rangle)$ . We observe that as it stands, Eq.(18) yields

$$M^{-1}(\langle L^{-1} \rangle) = \langle L^{-1} \rangle + O(M^{-1} V) \quad (19)$$

If  $\langle V \rangle = 0$ , the error term in Eq.(19) is  $O[(M^{-1} V)^2]$ .

When  $M^{-1}(\langle L^{-1} \rangle)$  is used in Eqs (10) and (16), equations appropriate to the case of long-range correlation result. If terms of order  $(M^{-1} V)^3$  are ignored and  $\langle V \rangle = 0$ , these equations simplify to the following, which are obtained by using Eq.(19) in Eqs (11) and (17) :

$$[M - \langle V \langle L^{-1} \rangle V \rangle] \langle u \rangle = g \quad (20)$$

$$[\bar{M}M - \langle V \langle L^{-1} \rangle V \rangle M - \bar{M} \langle \bar{V} \langle \bar{L}^{-1} \rangle \bar{V} \rangle - \bar{M} \langle V \langle L^{-1} \rangle \langle \bar{L}^{-1} \rangle \bar{V} \rangle M] \langle \bar{u} \rangle = \bar{g} \quad (21)$$

To complete these equations, an equation for  $\langle \bar{L}^{-1} \rangle$  is needed. To obtain it we begin with the identity

$$\langle L^{-1} \rangle^{-1} \langle L^{-1} \rangle = I \quad (22)$$

We now use Eq.(9) for  $\langle L^{-1} \rangle^{-1}$  in Eq.(22), replacing  $M^{-1}$  in it by  $M^{-1}(\langle L^{-1} \rangle)$ . This yields

$$M \sum_{q=0}^{\infty} \left\{ - \sum_{n=1}^{\infty} \langle [-M^{-1}(\langle L^{-1} \rangle) V]^n \rangle \right\}^q \langle L^{-1} \rangle = I \quad (23)$$

This is a non-linear equation for  $\langle L^{-1} \rangle$ . If we use Eq.(19) for  $M^{-1}(\langle L^{-1} \rangle)$ , set  $\langle V \rangle = 0$  and omit terms of order  $(M^{-1} V)^3$ , we obtain from Eq.(23)

$$[M - \langle V \langle L^{-1} \rangle V \rangle] \langle L^{-1} \rangle = I \quad (24)$$

Once  $\langle L^{-1} \rangle$  is obtained from Eq.(24),  $\langle u \rangle$  and  $\langle \bar{u} \rangle$  can be found from Eqs (20) and (21), which are linear equations.

To apply the preceding considerations to derive a theory of turbulence, we begin by writing the Navier-Stokes equations in the form

$$\begin{bmatrix} \partial_t - \nu \Delta - \vec{u} \cdot \nabla \\ \nabla \cdot \end{bmatrix} \begin{bmatrix} \vec{u} \\ p \end{bmatrix} = \begin{bmatrix} \vec{f} \\ 0 \end{bmatrix} \quad (25)$$

The matrix operator  $L$  is random because it contains the random velocity  $u$ . We define  $M$  and  $V$  by

$$M = \begin{bmatrix} \partial_t - \nu \Delta - \langle \vec{u} \rangle \cdot \nabla & \nabla \\ \nabla & 0 \end{bmatrix} \quad (26)$$

$$V = \begin{bmatrix} (\vec{u} - \langle \vec{u} \rangle) \cdot \nabla & 0 \\ 0 & 0 \end{bmatrix} \quad (27)$$

We can now use these definitions in Eq.(11) to obtain equations for  $\langle \vec{u} \rangle$  and  $\langle p \rangle$ . We see that the operator  $\langle VM^{-1}V \rangle$  in this equation will involve second moments of  $\vec{u}$  and  $p$ . These second moments are solutions of Eq.(17), which is non-linear because the coefficients also involve the same second moments. Thus the equations (11) and (17) constitute a closed non-linear system of equations for the first and second moments of  $\vec{u}$  and  $p$ . In these equations  $M^{-1}$  is an integral operator with a time-dependent Green's function as kernel. As a consequence, the equations involve the two-point two-time velocity and pressure correlations.

If the correlations are long range, we use Eqs (20), (21) and (24) instead. This is a closed system for the mean, the second moments and the operator  $\langle L^{-1} \rangle$  or its kernel, since it can be written as an integral operator. If we consider homogeneous isotropic turbulence in an unbounded domain and first eliminate  $p$ , the resulting system is similar to Kraichnan's original direct interaction equations.

## R E F E R E N C E S

- [1] LIN, C.C., *The Theory of Hydrodynamic Stability*, Cambridge University Press (1955).
- [2] CHANDRASEKHAR, S., *Hydrodynamic and Hydromagnetic Stability*, Oxford (1961).
- [3] HOPF, E., A mathematical example displaying features of turbulence, *Communs pure appl. Math.* 1 (1948) 303.
- [4] HOPF, E., Statistical hydromechanics and functional calculus, *J. rat. Mech. Analysis* 1 (1952) 87.
- [5] KOLMOGOROFF, A.N., The local structure of turbulence in an incompressible viscous fluid for very large Reynolds numbers, *Dokl. Acad. Sci. U.S.S.R.* 30 (1941) 301.
- [6] OBUKHOFF, A.M., On the distribution of energy in the spectrum of turbulent flow, *Dokl. Acad. Sci. U.S.S.R.* 32 (1941) 19; *Izv. Akad. Nauk S.S.R.*, Ser. Geogr. Geofiz. 5 (1941) 453.
- [7] von WEIZSÄCKER, C.F., Das Spektrum der Turbulenz bei grossen Reynolds'schen Zahlen, *Z. Phys.* 124 (1948) 614.
- [8] ONSAGER, L., The distribution of energy in turbulence (abstract only), *Phys. Rev.* 68 (1945) 286.
- [9] MILLIONHTCHIKOV, M., On the theory of homogeneous isotropic turbulence, *Dokl. Acad. Sci. U.S.S.R.* 32 (1941) 615.
- [10] HEISENBERG, W., Zur statistischen Theorie der Turbulenz, *Z. Phys.* 124 (1948) 628.
- [11] PROUDMAN, I., REID, W.H., On the decay of a normally distributed homogeneous turbulent velocity field, *Phil. Trans. R. Soc., London, Ser. A* 247 (1954) 163.
- [12] CHANDRASEKHAR, S., A theory of turbulence, *Proc. R. Soc., London, Ser. A* 229 (1955) 1.
- [13] TATSUMI, T., The theory of decay process of incompressible isotropic turbulence, *Proc. R. Soc., London, Ser. A* 239 (1957) 16.
- [14] SIEGEL, A., IMAMURA, T., MEECHAM, W.C., Wiener-Hermite expansion in model turbulence in the late stage, *J. math. Phys.* 6 (1965) 707.

- [15] KRAICHNAN, R.H., The structure of isotropic turbulence at very high Reynolds numbers, *J. Fluid Mech.*, 5 (1959) 497.
- [16] KRAICHNAN, R.H., Dynamics of nonlinear stochastic systems, *J. math. Phys.*, 2 (1961) 124.
- [17] KRAICHNAN, R.H., Lagrangian-history closure approximation for turbulence, *Physics Fluids* 8 (1965) 575.
- [18] KELLER, J.B., "Linear stochastic equations and a theory of turbulence", *Geophysical Fluid Dynamics Lecture Notes*, Woods Hole Oceanographic Institute, Woods Hole, Mass. 1 (1966) 132; "A survey of the theory of wave propagation in continuous random media", *Proc. Symp. Turbulence in Fluids and Plasmas*, Brooklyn Polytechnic Institute, Brooklyn, New York (1968).

# SOME REMARKS ON TURBULENCE \*

E. W. MONTROLL

Department of Physics and Astronomy,  
University of Rochester,  
Rochester, N. Y., United States of America

## Abstract

SOME REMARKS ON TURBULENCE. 1. Historical introduction; 2. Fluctuations in fluid flow and the scattering of light by such fluctuations; 3. Correlation functions and the statistical theory of turbulence; 4. On the  $(\omega, k)$  representation of the Navier-Stokes equation; 5. Response of two-dimensional incompressible fluids to periodic driving forces.

## 1. HISTORICAL INTRODUCTION

By the eve of World War I, the Thomson electron, the Rutherford atom, the quantum theory of Planck, Einstein and Bohr, the radioactivity discoveries of Becquerel and the Curies, the X-rays of Bragg and von Laue and the relativity theory of Einstein composed the forefront of physics. The difficult unsolved problems of classical physics were neglected and forgotten by most physicists. The topics discussed so far in this symposium on contemporary physics are the natural developments of the above-listed forefront subjects. Even the activities of our biophysical colleagues could not have started without technological advances in crystallography, radioactive tracer techniques and the electron microscope.

The great schools of physics at Cambridge (under J. J. Thomson), Munich (under Sommerfeld), Göttingen (under Born), Copenhagen (under Bohr), Leiden (under Lorentz and Ehrenfest), Berlin (under Planck and Einstein), and Zurich (under Debye), were flourishing and ready in the early 1920s for the development of quantum theory and modern physics. Most of our speakers and listeners can trace their scientific ancestry to these schools.

The subject reviewed by Professor Keller and myself in this and the preceding paper is completely separate from other branches of modern physics. It was developed by a few of the black sheep of the scientific families residing in the great schools. Perhaps G. I. Taylor at Cavendish, von Karman (and, on the experimental side, Prandtl) at Göttingen and J. Burgers after leaving Leiden to go to Delft, were enticed, as were other black sheep of fifty years ago who left the family farms and counting houses to pursue visions of man's conquest of the sky and space. Of their many researches on fluid physics, those which are of interest here lie in the field of turbulence.

The originator of our subject, Osborne Reynolds (the teacher of J. J. Thomson at Manchester before J. J. moved to Cambridge where, at Reynolds' suggestion, he got interested in microscopic physics), was a broad and imaginative scientist. While his ingenious models of the universe were neglected, his simple experiments, such as the one described below (see Fig. 1) exposed a complicated phenomenon which has remained one of the last mysteries of classical physics.

\* This work was partially supported by the Air Force Office of Scientific Research Grant No. AFOSR 1314-67.

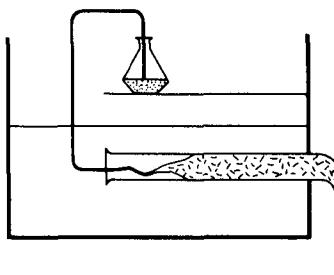


FIG. 1. Reynolds' original stability experiment. At the critical Reynolds number, the motion of the dark thread of dye marker loses its threadlike character and the fluid in the outflow tube becomes uniformly coloured by the dye.

Reynolds connected a long tube to a reservoir and the nature of the flow of fluid through the tube was observed by introducing a fine "thread" of dye into the mouth of the tube. At low velocities, the thread of dye remained thin and parallel to the axis of the tube. The flow velocity was slowly increased until at a certain critical velocity the dye thread oscillated wildly and to within a short distance from the entrance to the tube the dye spread uniformly through the tube. After repeating the experiment with a number of flowing fluids in tubes of various diameters, Reynolds noted that the onset of the wildly oscillating or turbulent state of flow in the tube could be characterized by a critical value of the dimensionless quantity

$$R_1 = UL\rho/\mu = UL/\nu \quad (1)$$

(which was later called the Reynolds number by Sommerfeld). Here  $U$  is the flow velocity in the tube,  $L$  the tube diameter,  $\rho$  the fluid density,  $\mu$  the viscosity and  $\nu = \mu/\rho$  the so-called kinematic viscosity. A critical Reynolds number was also associated with the transition from turbulent to laminar flow when one slowly decreased the velocity from a high value to a low one. This critical Reynolds number, which is generally called  $R_2$ , has the value 2300. It is not very sensitive to the roughness of the pipe. The critical Reynolds number in going from laminar to turbulent flow is more sensitive to the manner in which the experiment is done. While it is usually 2000, it can be increased somewhat in a quasi-stable state if the velocity is changed very slowly and great care is taken not to disturb the flow in any way. An  $R_1$  as large as  $10^5$  has been achieved; however, in this state the slightest disturbance will cause the flow to become turbulent.

The dye marker has also been used effectively by Prandtl and others to observe the flow pattern behind obstacles such as spheres, cylinders, and airplane wing sections. Excellent photographs of these fields are given in Ref.[1].

The basic equation which governs processes involving viscous fluids is the Navier-Stokes equation which, in the incompressible flow case, has the non-linear form

$$\frac{\partial u}{\partial t} + u \cdot \nabla u = - \frac{1}{\rho} \nabla p + \nu \nabla^2 u \quad (2a)$$

$$\text{with } \nabla \cdot \mathbf{u} = 0 \quad (2b)$$

where the vector  $\mathbf{u}(r, t)$  represents the velocity of the fluid at  $(r, t)$ , and  $p(r, t)$  is the pressure at that point. It is rather remarkable that when describing a fluid flowing in a pipe, this equation should imply a drastic variation in the characteristics of the flow when  $R = 2000$ . In the theory of other phase transitions, the change occurs when the dimensionless quantity  $\epsilon/kT$  ( $\epsilon$  being some appropriate interaction energy) has a specified critical value which is of  $O(1)$ .

Flows around obstacles also yield critical Reynolds numbers. Suppose one wishes to draw an object through a fluid and that a steady drag force  $D$  is required to maintain a steady velocity  $v$ . Then the dimensionless drag coefficient  $C_D$  is defined by

$$C_D = D / (\frac{1}{2} \rho v^2 S) \quad (3)$$

$\rho$  being the density of the fluid and  $S$  the area of the body. In the case of a sphere of radius  $a$  (choosing  $S = \pi a^2$  and defining the Reynolds number as  $R = 2av/\nu$ ), the experimental curve shown in Fig. 2 exhibits a sudden drop by almost an order of magnitude at  $R_C \approx 5 \times 10^5$ . In the small Reynolds number regime Stokes law gives a ratio  $C_D = 12/R$ , while Oseen found the first correction for larger Reynolds numbers and Goldstein [2] obtained the series expansion

$$C_D = \frac{12}{R} \left\{ 1 + \frac{3}{8} R - \frac{19}{1280} R^2 + \frac{71}{2560} R^3 - \frac{30179}{34406400} R^4 + \frac{122519}{560742400} R^5 \dots \right\}$$

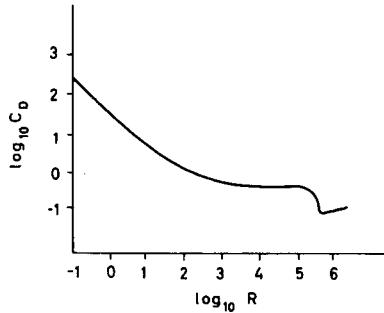


FIG. 2. Variation of track coefficient  $C_D$  of a sphere with Reynolds number.

This equation seems to agree fairly well with experimental data in the range  $R \leq 4$  which does not begin to approach the interesting range of  $R_C$ . It would be interesting to calculate another dozen or more terms in this series and attempt to adapt the methods used in the investigation of the Ising model of phase transitions to estimate critical Reynolds numbers and to analyse the behaviour of  $C_D$  in the neighbourhood of  $R_C$ .

Since the turbulent regime which is established when the Reynolds number exceeds its critical value is characterized by large fluctuations in local flow

velocities, one is interested in observing some of the statistical properties of these fluctuations.

We now discuss the recently developed light scattering techniques for the experimental investigation of both spontaneous and driven fluctuations in fluids.

## 2. FLUCTUATIONS IN FLUID FLOW AND THE SCATTERING OF LIGHT BY SUCH FLUCTUATIONS

There are two classes of fluctuations from mean behaviour of a fluid, those which develop spontaneously even in a fluid at rest, and those which are induced by various driving forces. Driven fluctuations have been observed traditionally by two techniques, the first through the introduction of a dye marker in the fluid to make the stream lines more visible, as was discussed above, and the second through the employment of a hot wire anemometer in the fluid. This is essentially a resistance thermometer. If the fluid is moving past a wire rapidly, it cools it faster than if it moves slowly past it. Hence, by measuring the temperature or the amount of heat which must be put into a wire to keep its temperature constant, the flow rate can be deduced. This section will be more concerned with the recently developed experimental techniques of scattering light from fluid elements and examining the flow from the Doppler shift in the frequency of the scattered light as well as from the line shape of the scattered light due to the dispersion in Doppler shifts in the fluctuating velocity field.

Three important classes of spontaneous fluctuations are:

- (a) Density fluctuations
- (b) Temperature fluctuations
- (c) Concentration fluctuations in multicomponent fluids

The density fluctuations can be harmonically analysed: a particular wave number component is not different from a sound wave of that wave number propagating through the fluid. Therefore, it has a particular velocity and frequency associated with it. While propagating, it is also damped (i.e. it has an associated "life time") by an amount which depends on the viscosity and the thermal conductivity which characterize the dissipative properties of the fluid. To a first approximation, the equation which describes propagation of density fluctuations is the wave equation ( $\rho$  being the deviation in density from its average value, and  $c$  the velocity of sound) in the absence of dissipation

$$\partial^2 \rho / \partial t^2 = c^2 \nabla^2 \rho \quad (4)$$

Temperature fluctuations decay according to Fourier's law and, therefore, through the heat equation

$$c_p \partial T / \partial t = \lambda_T \nabla^2 T \quad (5)$$

( $c_p$  being the specific heat per unit volume and  $\lambda_T$  being the thermal conductivity), while concentration fluctuations decay according to Fick's law and, therefore, the diffusion equation ( $D$  being the diffusion constant)

$$\partial C / \partial t = D \nabla^2 C \quad (6)$$

Typical driven fluctuations in a fluid are:

- (a) Sound waves
- (b) Effect of boundaries and obstacles on flow
- (c) Turbulence
- (d) Convection currents due to temperature gradients

We now discuss the scattering of light by a number of the fluctuations listed above.

A sound wave of wavelength  $\Lambda_s$  propagating in a fluid serves as an optical grating for a light wave which impinges on the fluid. Hence a light wave of wavelength  $\lambda$  suffers a Bragg reflection through an angle  $\theta$  which is related to  $\lambda$  and  $\Lambda_s$  by

$$\frac{2 \sin(\theta/2)}{\lambda} = \frac{1}{\Lambda_s} \quad (7)$$

The scattering is inelastic so that there is an exchange of energy between the phonon of frequency  $\omega_s(k)$  of the sound wave and that of the light photon of frequency. From the conservation of energy, the shift in circular frequency  $\Omega(k)$  of the scattered photon is

$$\hbar\Omega(k) = \hbar(\omega - \omega_0) = \pm \hbar\omega_s(k) = \pm 2\pi\hbar v_s/\Lambda_s \quad (8)$$

since the phase velocity of the sound wave is given by  $v_s = 2\pi\omega_s/\Lambda_s$ . Hence, from the Bragg condition, the frequency shift of the scattered photon is

$$\Omega(k)/2\pi = 2v_s \lambda^{-1} \sin(\theta/2) \quad (9)$$

The two choices of sign exist because the Bragg condition can be satisfied by a sound wave propagating either toward or away from the observer. In a fluid, all sound waves are longitudinal so that the sound velocity  $v_s$  depends only on the frequency.

The splitting formula (9) was checked experimentally a number of years ago by Debye and Sears [3] who, through a transducer, generated ultrasonic waves of known frequencies in several fluids and observed the frequency of the light scattered by the sound waves.

In the case of spontaneous density fluctuations, the splitting is called Brillouin scattering. It was first observed by Gross [4] and, in recent years, has been investigated in detail by Benedek and collaborators [5]. They have not only been able to resolve the Brillouin doublet, but they have also been able to examine its line shape. The line is Lorentzian and the line width, which is inversely proportional to the life time of the phonon, depends on the dissipation of the density waves through the viscosity and thermal conductivity of the fluid. It can be shown that the line width is

$$q^2\Gamma = \frac{1}{2} \left[ \frac{\mu}{\rho_0} + \frac{\lambda_T}{\rho_0} \left( \frac{1}{C_v} - \frac{1}{C_p} \right) \right] q^2 \quad (10)$$

$$|q| = (2\pi/\lambda)\sin(\theta/2)$$

where  $\mu$  is the viscosity,  $\rho_0$  the average density, and  $\lambda_T$  the thermal conductivity of the fluid.  $C_p$  and  $C_v$  are, respectively, its specific heats at constant

volume and pressure. Light scattering techniques are especially important for the investigation of the transport quantities  $\mu$  and  $\lambda_T$  close to a liquid-gas phase transition, properties which are otherwise difficult to measure.

The resolving power of the spectroscope required for the investigation of the line shapes described above is  $r = \omega_0/\Delta\omega \sim 10^{12}$  which is about 6 or 7 orders of magnitude more than can be provided by the best spectrometers. Furthermore, the displacement of the Brillouin doublets is comparable to the line width of standard pre-laser light sources. The availability of laser sources has been exploited by Benedek and collaborators at MIT, Alpert [6], Cummins and collaborators at Columbia [7], and Chiao and Stoicheff [8] at Toronto for detailed line shape investigations.

The resolving power difficulty is eliminated by optical heterodyning. One geometrical arrangement in which the scattered signal is beat with the incident beam is shown in Fig. 3. Part of the light from a laser source A goes through the half silvered mirror B to the mirror C, while the reflected beam from B is again reflected from the mirror D and finally scattered through an angle  $\theta$  by the scattering medium at E. The scattered signal is beat with the part of the incident beam which is reflected from C. The two signals are picked up together in the photomultiplier tube at F. Clearly the small difference in wavelength of the two signals is amplified in the beating process so that the wavelength of the beat signals is much larger than the difference in wavelength of the scattered and unscattered light.

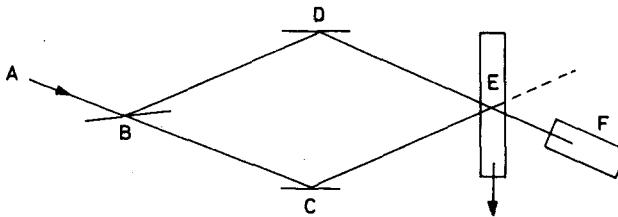


FIG. 3. A geometrical arrangement for optical heterodyning.

The complete spectrum of the light scattered by spontaneous fluctuations includes the broadening of the central beam, as well as the development of the Brillouin doublet. The shape of the central peak is also Lorentzian. It is due to temperature fluctuations and was first explained by Landau and Placzek [9]. The line width is

$$q^2\Gamma = \lambda \rho_0 c_p q^2 \quad (11)$$

Since concentration fluctuations decay according to the diffusion equation (6), which has the same form as the heat equation, one would expect the line width of the central peak in a two-component system with low concentration of one component to be proportional to the diffusion constant. Indeed, it can be shown to be  $Dq^2$ . Experiments have been performed on low concentration polymer solutions. In the case of spherical scatterers such as polystyrene spheres, one can verify the validity of Stokes law [10] which relates the diffusion constant to the viscosity  $\mu$  and the sphere radius  $a$  by

$$D = T k_B / 6a\mu \quad (12)$$

Yeh and Cummins [11] have pointed out that scattering techniques can be used to probe driven flow fields. The basis of the technique is the Doppler shift in the frequency of the scattered light. A shift was observed for velocities as low as 0.007 cm/s.

Since a pure fluid far from its transition temperature is not a good scatterer, it is desirable to enhance the scattered signal by introducing good scattering centres into the moving fluid. Those chosen by Yeh and Cummins were uniform size polystyrene spheres of diameter 0.557  $\mu\text{m}$ . They were used in the form of a colloidal suspension in water at the concentration 1 part in 30 000 parts of  $\text{H}_2\text{O}$ . They do not effect the flow. The collimated laser beam used (from a He-Ne laser with an incident beam of wavelength 6328 Å) had a diameter 0.16 cm. A geometry was employed in which the incident beam was propagated parallel to the axis of the tube containing the flowing fluid. The Doppler shift was observed in light scattered 30° to the tube axis while the beam was shifted to successive positions along the tube-diameter to probe the flow field as a function of distance from the tube axis.

The theoretical velocity profile of a fluid undergoing laminar flow in a tube is given by

$$v(r) = \Delta p(R^2 - r^2) / 4\nu\ell$$

$r$  being the distance from the centre of a tube of radius  $R$ ,  $\nu$  the kinematic viscosity of the fluid,  $\ell$  the length of the tube, and  $\Delta p$  the pressure difference between the ends. The agreement between the classical theory and the experiments of Yeh and Cummins is shown in Fig. 4.

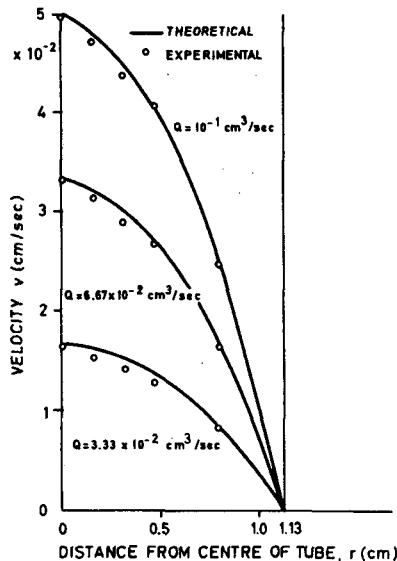


FIG. 4. Velocity profile of flow in a tube as observed by the Doppler shift in the scattered light.  $Q$  is the flow velocity through the tube. Data are taken from the experiments of Yeh and Cummins.

The method is applicable to other types of flows in which the velocity along the direction of the incident laser beam is not so constant. Since any variation in direction or magnitude of the flow within the scattering volume will produce a broadening of the Doppler shifted line, both the mean flow rate and the velocity gradient of the flow in the scattering region could be measured. This might provide a useful scheme for investigating flows which are difficult to analyse theoretically, but which are of considerable importance. Two such cases are relevant in the understanding of the flow of blood in arteries. They are flow in tapered tubes and flow in elastic tubes.

When the transition from laminar to turbulent flow occurs, the velocity distribution function of fluid elements can be expected to experience considerable broadening. The spectrum of the Doppler shifted scattered light would be broadened in a similar manner. The method of Yeh and Cummins has been applied by Goldstein and Hagen [12] to tube flow in the turbulent regime. Their beam geometry, which is shown in Fig. 5, is somewhat different from that of Yeh and Cummins in that incident beam is not taken to be along the flow axis.

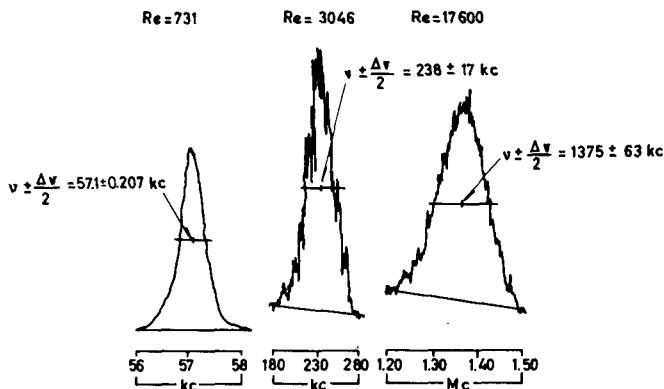


FIG. 5. Line width of scattered beam. Notice that the line width in the turbulent regime ( $Re > 12000$ ) is much greater than that under linear flow conditions. (From Goldstein and Hagen).

The spectrum analyser of the beat signals is scanned and integrated with an RC time constant of 3.5 s. If there is a time-varying velocity field with a period much shorter than 3.5 s, the intensity of the beat signal of given frequency is proportional to the fraction of time during which the fluid has a velocity which corresponds to that frequency. In turbulent flow this occupation time distribution is proportional to the probability distribution function of velocities in the fluid. This is, of course, correct only when the line width for a steady flow is negligible compared with the line width in the turbulent flow. Typical output records are shown in Fig. 5 to be essentially Gaussian. Note that the line is narrow in the laminar and broad in the turbulent regime, as one would expect.

The Gaussian line width is plotted as a function of Reynolds number in Fig. 6. It is rather constant and small until very close to the critical Reynolds number and slowly varying but large in the high Reynolds number turbulent region. The line width is greatest very close to the critical

Reynolds number where patches of laminar and turbulent fluid occur together in the scattering zone either at the same time or within very small time intervals of each other.

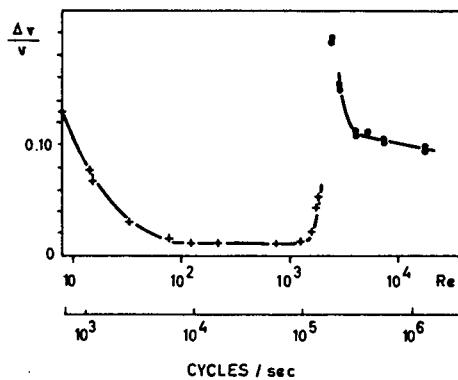


FIG. 6. Variation of line width with Reynolds number.

### 3. CORRELATION FUNCTIONS AND THE STATISTICAL THEORY OF TURBULENCE<sup>1</sup>

Since the motion of fluid elements is quite chaotic in the turbulent flow regime, it is felt that a statistical theory should be developed to describe that regime. While good ideas have appeared, attempts at such a theory have not been entirely successful. One generally considers homogeneous turbulence as it develops in an unbounded region. The non-linear character of the Navier-Stokes equation has suggested to most workers in the field that machine calculations will probably have to be made. For reasons which will be discussed in some detail in Section 4, numerical solutions of the NS equation in the large Reynolds number regime become unstable. Hence the starting point of the theory is usually certain differential equations for correlation functions, which is a quite reasonable starting point for a statistical theory.

The central quantity of the theory is the correlation tensor with elements ( $\alpha, \beta = 1, 2, 3$  referring to  $x, y, z$  components of the velocity)

$$R_{\alpha\beta}(r, t) = \langle u_\alpha(x+r, t) u_\beta(x, t) \rangle \quad (13)$$

where the bracket corresponds to some average. It may be a time average, or an ensemble average, or an average over a number of choices of  $x$  at which measurements might be made. In some theories, both the time and space variables are displaced from each other in the basic correlation functions. It can be shown that most of the interesting questions concerning homogeneous turbulence can be discussed in terms of  $R_{\alpha\beta}$  or its Fourier transform

$$\Phi_{\alpha\beta}(k, t) = \frac{1}{(2\pi)^3} \int R_{\alpha\beta}(r, t) \exp(ik \cdot r) d^3r \quad (14)$$

<sup>1</sup> For an excellent review of this field see Ref.[13].

It has been shown by von Karman and Hayworth that the  $R_{\alpha\beta}$ s are not independent of each other, but that they all depend on a single scalar function only. Robertson [14], from general tensor analysis, has discussed the relationship between elements of correlation function tensors of arbitrary order. This topic is discussed in considerable detail in Batchelor's book [15].

The program of the statistical theory of turbulence is then to seek a differential equation for  $R_{\alpha\beta}$  or  $\Phi_{\alpha\beta}$ . Before sketching how this is done by starting with the Navier-Stokes equations, we introduce a few relevant formulae and background ideas.

Since we are interested in velocity correlation functions, there is some merit in expressing the pressure in the Navier-Stokes equation as a function of velocities. This is easy for an incompressible fluid. Let us take the divergence of Eq. (2a) and employ Eq. (2b). Then

$$-\frac{1}{\rho} \nabla^2 p = \nabla \cdot \{u \cdot \nabla u\} \quad (15a)$$

This is analogous to Poisson's equation of electrostatics and can be solved by Green's function techniques to yield

$$\rho^{-1} p(r, t) = \frac{1}{4\pi} \int \frac{\partial^2 \{u_\alpha(r', t)u_\beta(r', t)\}}{\partial r'_\alpha \partial r'_\beta} \frac{d^3 r'}{|r - r'|} \quad (15b)$$

The usual summation convention on repeated indices is to be applied. When Eq. (15b) is substituted into Eq. (2a), we obtain an equation which contains only velocities

$$\begin{aligned} \frac{\partial u_\alpha}{\partial t} - \nu \nabla^2 u_\alpha &= - u_\beta \frac{\partial u_\alpha}{\partial r_\beta} \\ &- \frac{1}{4\pi} \int \frac{\partial^2 \{u_\alpha(r', t)u_\beta(r', t)\}}{\partial r'_\alpha \partial r'_\beta} \frac{\partial}{\partial r} \left\{ \frac{1}{|r - r'|} \right\} d^3 r' \end{aligned} \quad (15c)$$

It is convenient to Fourier-decompose the flow field. To this end we write

$$u(r, t) = \frac{1}{(2\pi)^3} \int a(k, t) e^{ik \cdot r} d^3 k \quad (16)$$

so that the total kinetic energy of the fluid is

$$\begin{aligned} E(t) &= \frac{1}{2} \rho \int u^2(r, t) d^3 r \\ &= \frac{1}{2} \frac{\rho}{(2\pi)^3} \int |a(k, t)|^2 d^3 k \end{aligned} \quad (17)$$

by Parseval's theorem.

If the fluid is driven in a homogeneous way so that no direction is preferred,  $a(k, t)$  depends only on the scalar  $k$  and

$$E(t) = \frac{1}{2} \frac{\rho}{(2\pi)^3} \int 2\pi k^2 |a(k, t)|^2 dk \quad (18)$$

An energy  $\epsilon(k)$  can be defined so that  $\epsilon(k, t)dk$  is the energy contained in waves with wave number between  $k$  and  $k+dk$ . Then

$$\epsilon(k, t) = \frac{1}{2} \rho \frac{k^2}{(2\pi)^2} |a(k, t)|^2 \quad (19)$$

Incidentally, the rate of dissipation of energy in an incompressible viscous fluid is

$$-\frac{dE(t)}{dt} = -\frac{d}{dt} \int \frac{1}{2} \rho u^2 d^3r = \mu \int (\text{curl } u)^2 d^3u \quad (20)$$

In the turbulent regime, large eddies (which correspond to small wave numbers) are developed by some driving process. These eddies are unstable and decay into smaller eddies which, in turn, decay into still smaller eddies (which correspond to larger wave numbers,  $k$ ). At the very small scale of eddy sizes, one is concerned with the molecular motion which defines the temperature of the system. As is shown in Fig. 7 in  $k$ -space, the driving disturbance gives rise to laminar instability, energy then being transferred

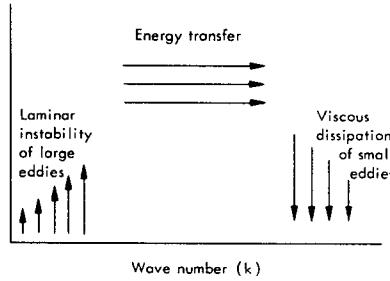


FIG. 7. Schematic mechanism for transfer of energy from small to large wave numbers.

from small to large wave numbers through the non-linearity in the Navier-Stokes equation and, finally, the cascade of energy transfer is terminated through the dissipation process which heats the fluid.

Kolmogoroff and, independently, Onsager, Obukhoff, and von Weizsäcker discussed the energy spectrum in the intermediate energy transfer regime by making several postulates about the cascade process and applying dimensional analysis. It was first assumed that the energy transfer process depended on two parameters, the kinematic viscosity of the fluid  $\nu$ , and the rate of energy decay  $\epsilon$ . All lengths and velocities can then be expressed in units of the

only quantities with the dimensions of length and velocity which can be constructed from  $\nu$  and  $\epsilon$ ,

$$\eta = (\nu^3/\epsilon)^{1/4} \quad (21a)$$

$$v = (\nu\epsilon)^{1/4} \quad (21b)$$

The spectrum then follows from three further postulates:

- (a) The turbulent motion in the regime of medium energy sizes is isotropic and independent of the nature of the large driven eddies;
- (b) The decay process at the very small eddy size regime is governed by the viscosity of the fluid only; and
- (c) The cascade process of going from small to large wave numbers in the intermediate regime is independent of the viscosity and, therefore, depends only on the energy decay rate  $\epsilon$ .

In a homogeneous isotropic fluid, the mean square velocity difference between two points separated by a distance  $r$  is

$$\langle [u(r+R) - u(R)]^2 \rangle = u^2 g_1(r) \quad (22)$$

where  $u$  is the root mean square velocity of the fluid,  $u^2 = \langle u^2 \rangle$  and  $g_1(r)$  is a dimensionless function whose  $r$ -dependence we wish to investigate. If  $u$  is expressed in terms of the velocity scale variable  $v = (\nu\epsilon)^{1/4}$ , we can rewrite Eq. (22) as

$$\langle [u(r+R) - u(R)]^2 \rangle = (\nu\epsilon)^{1/2} g_2(r/\eta) \quad (23)$$

The  $r$ -dependence of the function  $g_2(r/\eta)$  is obtained by invoking postulate (c) to the effect that the right-hand side of Eq. (23) must be independent of the viscosity. The only way this is possible is for  $g_2(r/\eta)$  to be a power of  $r/\eta$ , say the  $\lambda$  power, so that

$$\langle [u(r+R) - u(R)]^2 \rangle = D(\nu\epsilon)^{1/2-3\lambda/4} \epsilon^{\lambda/4} r^\lambda \quad (24)$$

which can be independent of  $\nu$  only if  $\lambda = 2/3$ , in which case ( $D$  being a dimensionless constant)

$$\langle [u(r+R) - u(R)]^2 \rangle = D(\epsilon r)^{2/3} \quad (25)$$

which is the Kolmogoroff result in the inertial range.

The energy spectrum  $\epsilon(k, t)$  is essentially the one-dimensional Fourier transform of Eq. (25). If we assume that

$$\int r^\alpha e^{irk} dr = ck^\beta \quad (26)$$

where  $c$  is a dimensionless constant, then

$$k^{-(\alpha+1)} \int rk e^{irk} d(rk) = ck^\beta$$

Since the integral over  $(kr)$  is dimensionless, we see that  $\alpha+\beta = -1$ , so that if  $\alpha = 2/3$ ,  $\beta = -5/3$  and

$$\epsilon(k, t)/\rho \sim \epsilon^{2/3} k^{-5/3} \quad (27)$$

which was first derived by Obukhoff by using the method discussed in the derivation of Eq. (25).

The systematic investigation of the correlation functions (13) and of the spectrum starts with the identity

$$\begin{aligned} \frac{\partial}{\partial t} R_{\alpha\beta}(R) &= \langle u_\beta(r+R, t) \frac{\partial}{\partial t} u_\alpha(r) \rangle \\ &+ \langle u_\alpha(r) \frac{\partial}{\partial t} u_\beta(r+R) \rangle \end{aligned} \quad (28)$$

One then substitutes into this equation the expression for  $\partial u_\alpha(r)/\partial t$  that follows from the Navier-Stokes equation (15c); similarly for  $\partial u_\beta/\partial t$ . Clearly, after this is done, the right-hand side of the equation contains  $R_{\alpha\beta\gamma}$  (i.e. three-point correlation functions as well as  $R_{\alpha\beta}$ ). An equation must then be found for  $R_{\alpha\beta\gamma}$ . Again one finds an expression  $\partial R_{\alpha\beta\gamma}/\partial t$  similar to Eq. (28) and again one uses Eq. (15c) to find the various time derivatives of the velocity which must be inserted in the equation. The right-hand side, however, contains four-point correlation functions  $R_{\alpha\beta\gamma\delta}$  as well as the three-point ones.

This process can be continued indefinitely yielding a hierarchy of equations which connect a correlation function of a given order with one of the next higher order. This hierarchy is discussed in more detail by Keller in these Proceedings, but I shall make a few remarks on it here. Chandrasekhar [16] and others have terminated the set by assuming some relation between several of the lower order correlation functions. For example, Chandrasekhar assumed that turbulence is generated by some Gaussian random process. A certain relationship exists between second and fourth order correlation functions and one has only two equations from which a single one for  $R_{\alpha\beta}$  can be derived. Among other things Chandrasekhar was able to derive the Kolmogoroff spectrum for the inertial range and, indeed, to extend it into a new form

$$E(k) \sim \left( \frac{8\epsilon}{9K_H} \right)^{2/3} k^{-5/3} \left\{ 1 + \frac{8\nu^3}{3\epsilon K_H^2} k^4 \right\}^{-4/3}$$

$K_H$  being a numerical constant.

While many useful results seemed to come from breaking the chain of equations as has been done by a number of authors, Kraichnan [17] has recently pointed out that in the zero viscosity limit energy is not conserved in these equations and, for both vanishing and finite viscosities, it may even become negative. He has attempted to correct this fault, but in doing so the theory has become rather complicated and unclear.

Incidentally, the hierarchy is quite analogous with Born-Green, Yvon, Kirkwood, Bogolyubov equations which relate molecular distribution functions,

and to the multiparticle Green's function hierarchy of quantum mechanical many-body theory.

#### 4. ON THE $(\omega, k)$ REPRESENTATION OF THE NAVIER-STOKES EQUATION

The Navier-Stokes equations (2a and 2b), the energy dissipation equation (20), and the appropriate boundary conditions at fluid surfaces are the basis for the investigation of any properties of the flow of viscous incompressible fluids. A form of these equations which I believe to be more suggestive to those more experienced in quantum field theory or many-body physics is that obtained by introducing the Fourier components of the flow field (16). The Fourier expansion of the pressure field is

$$p(r, t) = \frac{1}{(2\pi)^3} \int p(k, t) e^{ik \cdot r} d^3k \quad (29)$$

The first step for obtaining this equation is the replacement of the pressure by an appropriate function of the velocity. We see from Eqs (19a), (29), and (16)

$$\frac{k^2 p(k, t)}{\rho} = - \frac{1}{(2\pi)^3} \int [k' \cdot a(k-k', t)] [k \cdot a(k', t)] d^3k' \quad (30)$$

The incompressibility condition (2b) is equivalent to the statement

$$k \cdot a(k, t) = 0 \quad (31)$$

If we substitute Eq. (30) into Eq. (29) and introduce the resulting expression and Eq. (16) into Eq. (2a), we find the  $k$  space representation [18] of Eq. (2a) to be (as  $V \rightarrow \infty$ )

$$\begin{aligned} \frac{\partial a(k, t)}{\partial t} + \nu k^2 a(k, t) &= F(k, t) \\ &- \frac{i}{(2\pi)^3} \int_{-\infty}^{\infty} \int \int d^3k' \{k \cdot a(k', t)\} \{a(k-k', t) - kk' \cdot a(k-k', t)/k^2\} \end{aligned} \quad (32)$$

where  $F(k, t)$  is the Fourier transform of the force which is driving the fluid.

Let us further Fourier-analyse the time dependence of the flow pattern so that

$$a(k, t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{i\omega t} a(k, \omega) d\omega \quad (33)$$

Then

$$a(k, \omega) = \frac{1}{i\omega + k^2\nu} \left\{ F(k, \omega) - \frac{i}{(2\pi)^4} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} d\omega' d^3k' [k \cdot a(k', \omega')] [a(k-k', \omega-\omega') - kk' \cdot a(k-k', \omega-\omega') k^{-2}] \right\} \quad (34)$$

This non-linear integral equation is a basic equation through which homogeneous turbulence can be investigated. In this form, one can express the problem in a language which is more familiar to high-energy and many-body physicists than is the language of fluid dynamics.

The energy cascade associated with Eq. (19) is difficult to investigate because of the vector character of Eq. (34) as well as its non-linear form. The strategy adopted is to examine a scalar non-linear equation which is in fact the two-dimensional form of Eq. (34) to see how one should proceed to extract information from such equations without being tangled up in polarization problems. While it is well known that many of the characteristics of turbulence do not exist in 2D flows, it is hoped that some useful mathematical experience can be obtained from studying the 2D problem.

The nice feature of 2D incompressible flow is that the flow field is completely described by a scalar potential  $\phi$  which is related to the velocity components  $u_1$  and  $u_2$  (i. e. the velocities in the  $x$  and  $y$  direction) by

$$u_1 = -\partial\phi/\partial x_2 \quad \text{and} \quad u_2 = \partial\phi/\partial x_1 \quad (35)$$

The functions  $u_1$  and  $u_2$  when written in this form clearly satisfy the 2D equation of continuity

$$\frac{\partial u_1}{\partial x_1} + \frac{\partial u_2}{\partial x_2} = 0 \quad (36)$$

The 2D Navier-Stokes equation has the form

$$\frac{\partial u_1}{\partial t} + u_1 \frac{\partial u_1}{\partial x_1} + u_2 \frac{\partial u_1}{\partial x_2} = -\frac{1}{\rho} \nabla p + \nu \nabla^2 u_1 + F_1(r, t) \quad (37a)$$

$$\frac{\partial u_2}{\partial t} + u_1 \frac{\partial u_2}{\partial x_1} + u_2 \frac{\partial u_2}{\partial x_2} = -\frac{1}{\rho} \nabla p + \nu \nabla^2 u_2 + F_2(r, t) \quad (37b)$$

If we operate on Eq. (37a) with  $\partial/\partial x_2$  and on Eq. (37b) with  $\partial/\partial x_1$  and subtract the second equation from the first, the introduction of Eq. (35) yields

$$-\frac{\partial}{\partial t} \nabla^2 \phi + \begin{vmatrix} \frac{\partial}{\partial x_1} \nabla^2 \phi & \frac{\partial}{\partial x_2} \nabla^2 \phi \\ \frac{\partial \phi}{\partial x_1} & \frac{\partial \phi}{\partial x_2} \end{vmatrix} = -\nu \nabla^2 (\nabla^2 \phi) + \frac{\partial F_1}{\partial x_2} - \frac{\partial F_2}{\partial x_1} \quad (38)$$

This can be written also as two equations by defining  $f$  by

$$\nabla^2 \phi = f \quad (39a)$$

so that

$$\frac{\partial f}{\partial t} - \nu \nabla^2 f = \frac{\partial \phi}{\partial x_2} \frac{\partial f}{\partial x_1} - \frac{\partial \phi}{\partial x_1} \frac{\partial f}{\partial x_2} - \frac{\partial F_1}{\partial x_2} + \frac{\partial F_2}{\partial x_1} \quad (39b)$$

The  $k$ -space representation is obtained by setting  $F$ ,  $\phi$ , and  $f$  respectively equal to the Fourier transforms of  $f$ ,  $\phi$ , and  $F$ , so that

$$\Phi(k, t) = \int \phi(r, t) e^{ik \cdot r} d^2 r, \text{ etc.} \quad (40)$$

Then

$$-k^2 \Phi(k, t) = F(k, t) \quad (41)$$

and

$$\begin{aligned} \frac{\partial \Phi(k, t)}{\partial t} + \nu k^2 \Phi(k, t) &= -i(k \times F)_z / k^2 \\ &+ \frac{1}{(2\pi)^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (k'^2/k^2) [k \times k']_z \Phi(k', t) \Phi(k-k', t) d^2 k' \end{aligned} \quad (42)$$

we define  $[a \times b]_z = a_x b_y - a_y b_x$ .

The 2-D form of Eq. (34) in terms of  $\Phi(k, \omega)$  defined by

$$\Phi(k, \omega) = \int \phi(r, t) e^{i(k \cdot r - \omega t)} d^2 r dt \quad (43)$$

is

$$\begin{aligned} \Phi(k, \omega) &= -\frac{i(k \times F)_z}{k^2(i\omega + \nu k^2)} \\ &+ \frac{1}{(2\pi)^3} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{[k \times k']_z k'^2}{k^2(i\omega + \nu k^2)} \Phi(k', \omega') \Phi(k-k', \omega-\omega') d^2 k' d\omega' \end{aligned} \quad (44)$$

The Fourier components of the velocity are related to those of the scalar potential by Eq. (35)

$$U_1(k, \omega) = -ik_2 \Phi(\omega, k) \quad (45a)$$

$$U_2(k, \omega) = ik_1 \Phi(\omega, k) \quad (45b)$$

so that the energy associated with  $(k, \omega)$  is

$$\begin{aligned}\epsilon(k, \omega) &= \frac{1}{2} \rho \left\{ |U_1(k, \omega)|^2 + |U_2(k, \omega)|^2 \right\} \\ &= \frac{1}{2} \rho k^2 |\Phi(k, \omega)|^2\end{aligned}\quad (46)$$

In the next section we will discuss the distribution of energy into various wave numbers as obtained from a special driving force. One point which will become clear is that there is a dispersion relation between  $\omega$  and  $k$ . In the case which will be discussed

$$\omega(k) = (k_1 + k_2)\omega_0 \quad (47)$$

where  $\omega_0$  is the frequency of the periodic driven disturbance of small wave number. Clearly, as  $k = (k_1, k_2)$  becomes large, the frequency becomes large; that is, high frequencies are associated with small eddies.

This observation implies that it is very difficult to solve the original  $(r, t)$  representation of the Navier-Stokes equation numerically in the high Reynolds number regime. The standard way of using a high-speed computer to solve a partial differential equation is to approximate the PDE by a difference equation operation in space and time. The value of  $u(r, t)$  at one time interval is related to that at the previous time interval  $t - \Delta$ , the value of  $\Delta$  being fixed in the calculation. In the high Reynolds number regime, eddies of large wave number would be generated. Since these correspond to high frequencies, they also correspond to very short periods of oscillation in the fluid flow, eventually with the period becoming short compared with the time interval chosen for the calculation grid. This would lead to an instability in the iterative scheme for the solution of the equation.

## 5. RESPONSE OF TWO-DIMENSIONAL INCOMPRESSIBLE FLUIDS TO PERIODIC DRIVING FORCES

We have started a program of investigating the energy cascade process through  $k$ -space by driving a fluid with two Fourier exponentials

$$f_0(r, t) = A \sum_{j=1}^2 \alpha_j \exp(i(t\omega_0 - rq_j)) \quad (48)$$

which is a special case of the driving force

$$f_1(r, t) = A \sum_{j=1}^2 \{\alpha_j \exp(i(t\omega_0 - rq_j)) + \alpha_{-j} \exp(-i(t\omega_0 - rq_j))\} \quad (49)$$

Here the vectors  $q_1$  and  $q_2$  are fixed vectors as is the frequency  $\omega_0$ . When  $\alpha_j = \alpha_{-j}$  this corresponds to driving the fluid with two periodic plane waves

at an angle  $\theta$  with

$$\cos\theta = (q_1 \cdot q_2) / \sqrt{q_1^2 q_2^2} \quad (50)$$

The non-linear Navier-Stokes equation is difficult to solve in this case. It is, however, even instructive to set  $\alpha_{-1} = \alpha_{-2} = 0$ , in which case one obtains a certain recurrence formula which makes it easy to find the energy as a function of  $k$ .

In  $(\omega, k)$  space, the quantity  $(k \times F)_z$  required in Eq. (42) is

$$(k \times F)_z = (k \times A)_z \{ \delta(\omega - \omega_0) [\alpha_1 \delta(k - q_1) + \alpha_2 \delta(k - q_2)] + \delta(\omega + \omega_0) [\alpha_{-1} \delta(k + q_1) + \alpha_{-2} \delta(k + q_2)] \} \quad (51)$$

When this is substituted into Eq. (44), one might attempt to find the solution of Eq. (44) by iteration. One starts with energy being pumped into points in  $k$ -space  $\pm q_1$  and  $\pm q_2$ . The first iteration transfers energy into the four points  $\pm q_1 \pm q_2$ , the second introduces the new points  $\pm q_1 \pm 2q_2$ ,  $\pm 3q_1$ ,  $\pm 3q_2$ , and  $\pm 2q_1 \pm q_2$ . By continuing this process, one generates the complete set of lattice points  $\ell_1 q_1 + \ell_2 q_2$  with  $\ell_1 = 0, \pm 1, \pm 2, \dots$  and  $\ell_2 = 0, \pm 1, \pm 2, \dots$ .

We then seek a solution of Eq. (44) of the form

$$\Phi(k, \omega) = \sum_{\ell_1=-\infty}^{\infty} \sum_{\ell_2=-\infty}^{\infty} a(\ell_1, \ell_2) \delta(k - \ell_1 q_1 - \ell_2 q_2) \delta(\omega - [\ell_1 + \ell_2] \omega_0) \quad (52)$$

When this is substituted into Eq. (44) and coefficients of like  $\delta$ -functions on both sides of the resulting equation are equated, we find that  $a(\ell_1, \ell_2)$  satisfies

$$\begin{aligned} a(\ell_1, \ell_2) &= -i\beta_1 \delta_{\ell_2, 0} \delta_{\ell_1, 1} / g_{1,0} - i\beta_2 \delta_{\ell_2, 1} \delta_{\ell_1, 0} / g_{0,1} \\ &\quad - i\beta_{-1} \delta_{\ell_2, 0} \delta_{\ell_1, -1} / g_{-1,0} - i\beta_{-2} \delta_{\ell_2, 0} \delta_{\ell_1, -1} / g_{0,-1} \\ &\quad + \lambda \sum_{n_1, n_2=-\infty}^{\infty} a(n_1, n_2) a(\ell_1 - n_1, \ell_2 - n_2) (\ell_1 n_2 - \ell_2 n_1) (n_1 q_1 + n_2 q_2)^2 / f_{\ell_1 \ell_2} g_{\ell_1 \ell_2} \end{aligned} \quad (53)$$

where

$$\lambda \equiv (q_1 \times q_2) / (2\pi)^3 \quad (54a)$$

$$f_{\ell_1 \ell_2} \equiv (\ell_1 q_1 + \ell_2 q_2)^2 \quad (54b)$$

$$g_{\ell_1 \ell_2} = i(\ell_1 + \ell_2) \omega_0 + \nu(\ell_1 q_1 + \ell_2 q_2)^2 \quad (54c)$$

$$\beta_j = \alpha_j (q_j \times A)_z / q_j^2 \quad \text{with} \quad q_{-j} \equiv -q_j \quad (54d)$$

If one considers the special  $\beta_{-1} = \beta_{-2} = 0$ , then  $a(\ell_1, \ell_2) \equiv 0$  if  $\ell_1 < 0$  or  $\ell_2 < 0$ ; i. e. only the points in the positive quadrant are occupied because none of the driving vectors point in the negative  $\ell_1$  or  $\ell_2$  directions. In this case, one can derive a recursion formula which gives  $a(\ell_1, \ell_2)$  for a given vector  $(\ell_1, \ell_2)$  in terms of  $a$ 's whose  $(\ell'_1, \ell'_2)$  are closer to the origin.

Certain scaling factors are immediately apparent from the quadratic form of Eq. (53). Suppose that  $\ell_1, \ell_2$  is neither  $(1, 0)$  nor  $(0, 1)$ . Then it is clear that we can express  $a(\ell_1, \ell_2)$  in the form

$$a(\ell_1, \ell_2) = [a(1, 0)]^{\ell_1} [a(0, 1)]^{\ell_2} \{b(\ell_1, \ell_2)\} \quad (55a)$$

where

$$b(1, 0) = b(0, 1) = 1 \quad (55b)$$

For  $(\ell_1, \ell_2)$  vectors other than  $(0, 1)$  and  $(1, 0)$ ,  $b(\ell_1, \ell_2)$  satisfies

$$b(\ell_1, \ell_2) = \sum_{n_1=0}^{\ell_1} \sum_{n_2=0}^{\ell_2} b(n_1, n_2) b(\ell_1 - n_1, \ell_2 - n_2) f_{n_1 n_2} / f_{\ell_1, \ell_2} g_{\ell_1, \ell_2} \quad (56)$$

An alternative form for this equation can be obtained by making the change in variable  $n'_i = \ell_i - n_i$  in the summation, then removing the prime and averaging the resulting expression with Eq. (56). Then

$$\begin{aligned} b(\ell_1, \ell_2) &= \frac{1}{4} \sum_{n_1=0}^{\ell_1} \sum_{n_2=0}^{\ell_2} b(n_1, n_2) b(\ell_1 - n_1, \ell_2 - n_2) \\ &\times [\ell_1(2n_2 - \ell_2) - \ell_2(2n_1 - \ell_1)] [(2n_1 - \ell_1)q_1 + (2n_2 - \ell_2)q_2] (\ell_1 q_1 + \ell_2 q_2) / g_{\ell_1, \ell_2} \end{aligned} \quad (57)$$

which has the vector form

$$b(\ell) = \frac{1}{4} \sum_0^{\ell} b(n) b(\ell - n) [\ell \times (2n - \ell)]_z \hat{\ell} \cdot (2\hat{n} - \hat{\ell}) / \hat{\ell} \cdot \hat{g}(\ell) \quad (58a)$$

where

$$\hat{\ell} \equiv \ell_1 q_1 + \ell_2 q_2 \quad \text{and} \quad \ell \equiv (\ell_1, \ell_2) \quad \text{so that} \quad [\ell \times m]_z = \ell_1 m_2 - \ell_2 m_1 \quad (58b)$$

For later reference we note that

$$a(1, 0) = -i\beta_1/g_{1,0} = -i\beta_1/(i\omega_0 + \nu q_1^2) = -(\beta_1/\omega_0)/(1 - iR_1^{-1}) \quad (59a)$$

$$a(0, 1) = -i\beta_1/g_{0,1} = -i\beta_2/(i\omega_0 + \nu q_2^2) = -(\beta_2/\omega_0)/(1 - iR_2^{-1}) \quad (59b)$$

where  $R_j$  is the Reynolds number defined by

$$R_j = (\omega_0 / |q_j|) (1 / |q_j|) / \nu = \omega_0 / \nu q_j^2 \quad (60)$$

when  $\ell_1 = \ell_2 = 1$ , we find

$$b(1, 1) = \left\{ \frac{q_2^2 - q_1^2}{2i\omega_0 + \nu(q_1 + q_2)^2} \right\} (q_1 \times q_2)_z / (2\pi)^3 \quad (61)$$

Equation (57) can be used as a recurrence formula to obtain  $b(\ell_1, \ell_2)$  from those  $b$ 's which are closer to the origin. For example, one finds from Eq. (57) that

$$b(\ell, 1) = \lambda b(\ell_1 - 1, 1) \{(\ell - 2)q_1 + q_2\} (\ell q_1 + q_2) / g_{\ell, 1} f_{\ell, 1} \quad (62a)$$

$$b(1, \ell) = \lambda b(1, \ell - 1) \{q_1 + (\ell - 2)q_2\} (q_1 + \ell q_2) / g_{1, \ell} f_{1, \ell} \quad (62b)$$

Starting with  $\ell = 1$ , one can find all  $b(\ell, 1)$  and all  $b(1, \ell)$ . It is easily shown, for example, that (if  $\nu \neq 0$ )

$$b(\ell, 1) = \left( \frac{\lambda}{\nu q_1^2} \right) \prod_{\alpha=1}^2 \left\{ \frac{\Gamma(1+\Omega_\alpha[1]) \Gamma(1+\xi_\alpha[1]) \Gamma(\ell+1+\eta_\alpha[1])}{\Gamma(1+\eta_\alpha[1]) \Gamma(\ell+1+\Omega_\alpha[1]) \Gamma(\ell+1+\xi_\alpha[1])} \right\} \quad (63)$$

where

$$\Omega_1[m] + \Omega_2[m] = iR_1 + zm \quad \text{and} \quad \Omega_1[m] \Omega_2[m] = iR_1 m + m^2 q_2^2 / q_1^2 \quad (64a)$$

$$\eta_1[m] + \eta_2[m] = zm - 2 \quad \text{and} \quad \eta_1[m] \eta_2[m] = m^2 (q_2^2 / q_1^2) - zm \quad (64b)$$

$$\xi_1[m] + \xi_2[m] = zm \quad \text{and} \quad \xi_1[m] \xi_2[m] = m^2 q_2^2 / q_1^2 \quad (64c)$$

with

$$Z = 2q_1 \cdot q_2 / q_1^2 \quad (64d)$$

It is also easy to show that

$$\begin{aligned} b(\ell, 2) &= 4\lambda b(1, 0)b(\ell-1, 2)[(\ell_1-2)q_1+2q_2]^2 (\ell_1 q_1 + 2q_2) / g_{\ell, 2} f_{\ell, 2} \\ &\quad + \frac{1}{2} \lambda \sum_{n_1=0}^{\ell} b(n_1, 1)b(\ell-n_1, 1)(2n_1-\ell) q_1 \cdot (\ell_1 q_1 + 2q_2) / g_{\ell, 2} f_{\ell, 2} \end{aligned} \quad (65)$$

Since  $b(\ell, 1)$  is known from Eq. (63),  $b(\ell, 2)$  can be obtained from this recurrence formula by setting  $b(0, 1) \equiv 1$  as defined above. By continuing this process, all  $b(\ell_1, \ell_2)$  can be obtained.

The lines of constant amplitude  $|a(\ell_1, \ell_2)|^2$  are plotted in Fig. 8 for the set of parameters

$$[a(1, 0)]^4 = [a(0, 1)]^4 = 0.004$$

$$q_1^2 = 10, \quad q_2^2 = 20, \quad R_1 = 10, \quad \text{and} \quad \nu = 5 \times 10^{-5} \text{ m}^2/\text{s}$$

$$\omega_0 = 5 \times 10^{-3}/\text{s}$$

These machine calculations were programmed by B. Parekh. It is interesting to note that three peaks exist in the range of  $(\ell_1, \ell_2)$  considered, one at approximately  $(10, 12)$ , another at  $(20, 24)$ , and a third at  $(30, 36)$ . The reason for these relative locations of the peaks will become apparent from the remarks made below.

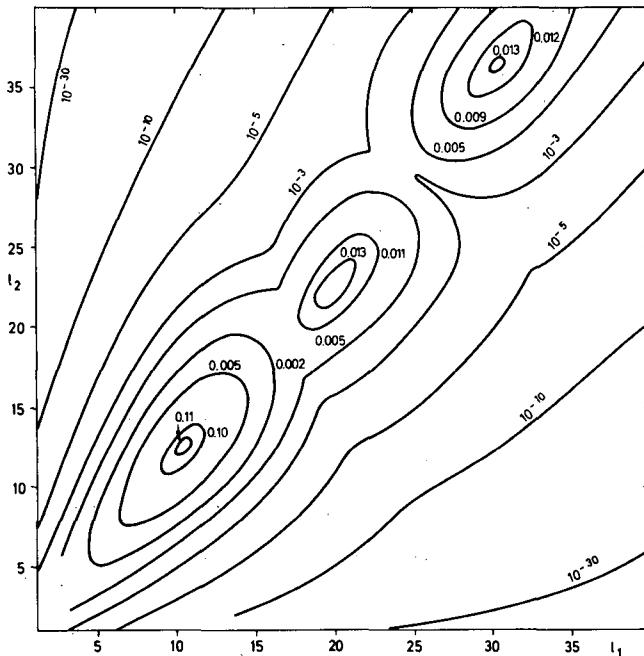


FIG. 8. An example of curves of constant amplitude  $|a(\ell_1, \ell_2)|$  associated with various wave numbers. The choice of parameters is  $[a(1, 0)]^4 = [a(0, 1)]^4 = 0.4 \times 10^{-4}$ ,  $q_1^2 = 10$ ,  $q_2^2 = 20$ ,  $R_1 = \omega_0/v\nu q_1^2 = 10$ ,  $\nu = 5 \times 10^{-5} \text{ m}^2/\text{s}$  and  $\omega = R\nu q_1^2 = 5 \times 10^{-3} \text{ s}$ . Numerical calculations were made by B. Parekh.

If we examine Eq. (65) to see which combination

$$b(n_1, 1)b(\ell - n_1, 1) \quad (66)$$

gives the biggest contribution to  $b(\ell, 2)$ , we find that when  $n$  and  $\ell$  are large enough so that the Stirling approximation can be applied to Eq. (63), the

largest contribution comes from the choice  $n_1 = \frac{1}{2}\ell \pm 1$ . If the factor  $(2n_1 - \ell)$  did not appear in Eq. (65), the largest would come from the choice  $n_1 = \frac{1}{2}\ell$ . In the Stirling approximation, each  $b(n_1, 1)$  can be expressed as  $b(\ell/2, 1)$  multiplying by a correction factor which goes to zero rapidly as  $| \frac{1}{2}\ell - n_1 |$  increases. The series in Eq. (66) can be summed so that

$$b(\ell, 2) \sim \lambda \{ q_2^2 - [q_1 + \ell q_2]^2 \} (\pi\ell/2)^{\frac{1}{2}} \{ b^2(\frac{1}{2}\ell, 1)/8f_{\ell, 2}g_{\ell, 2} \} \{ 1 + O(1/\ell) \} \quad (67)$$

One consequence of this formula is that if  $|b(\ell, 1)|$  has a maximum value for  $\ell = \ell^{(0)}$ , then  $|b(\ell, 2)|^2$  has a maximum at  $(\ell^{(0)}, 2)$ . It seems that, when  $\ell_1$  and  $\ell_2$  are both large, Eq. (67) can be generalized to

$$b(\ell_1, \ell_2) \sim F_{\ell_1 \ell_2} b^2(\frac{1}{2}\ell_1, \frac{1}{2}\ell_2) \quad (68)$$

where  $F_{\ell_1 \ell_2}$  is a slowly varying function of  $(\ell_1, \ell_2)$ . If  $b(\frac{1}{2}\ell_1, \frac{1}{2}\ell_2)$  has a maximum at  $(\ell_1, \ell_2) = (\ell_1^{(0)}, \ell_2^{(0)})$ , then  $b(\ell_1, \ell_2)$  would also have a maximum at  $(\ell_1^{(0)}, \ell_2^{(0)})$ . Equation (68) would then be the basis for the observation concerning the locations of the maxima in Fig. 8. A detailed investigation of these ideas will be published elsewhere.

Now let us consider the more physically interesting case with  $\beta_1 = \beta_{-1}$  and  $\beta_2 = \beta_{-2}$  in Eq. (53). An interesting question to which the answer is not known is whether an equation such as (68) is still valid even when feedback exists from which high wave numbers and low wave numbers combine to produce eddies of intermediate wave numbers. If it is true, then the feedback effect will be of only secondary importance.

## R E F E R E N C E S

- [1] PRANDTL, L., TIETJENS, O. G., *Hydro and Aeromechanics* 2, Dover (1957).
- [2] GOLDSTEIN, S., Proc. R. Soc. 123A (1929) 225.
- [3] DEBYE, P., SEARS, F. W., Proc. natn. Acad. Sci. 18 (1932) 409;  
See also LUCAS, R., BIQUARD, P., C. r. hebd. Séanc. Acad. Sci., Paris 194 (1932) 2132.
- [4] GROSS, E. F., Z. Phys. 63 (1930) 685; Naturwissenschaften 18 (1930) 717; Nature, Lond. 126 (1930) 201, 400, 603.
- [5] BENEDEK, G. B., GREYTAH, T., Proc. Instn elect. Engrs 53 (1965) 1623;  
FORD, N. C., JR., BENEDEK, G. B., in *Phenomena in the Neighbourhood of the Critical Point* (Proc. Conf. Washington, 1965) 150.
- [6] ALPERT, S. S., in *Phenomena in the Neighbourhood of the Critical Point* (Proc. Conf. Washington, 1965);  
ALPERT, S. S.; YEH, Y., LIPWORTH, E., Phys. Rev. Lett. 14 (1965) 486.
- [7] CUMMINS, H. Z., KNABLE, N., YEH, Y., Phys. Rev. Lett. 12 (1964) 150.
- [8] CHIAO, R. Y., STOICHEFF, B. P., J. opt. Soc. Am. 54 (1964) 1286.
- [9] LANDAU, L., PLACZEK, G., Z. Sovjetunion 5 (1934) 172.
- [10] DUBIN, S. B., LUNACEK, J. H., BENEDEK, G. B., Proc. natn. Acad. Sci. 57 (1967) 1164.
- [11] YEH, Y., CUMMINS, H. Z., Appl. Phys. Lett. 4 (1964) 176.
- [12] GOLDSTEIN, R. J., HAGEN, W. F., Physics Fluids 10 (1967) 1349.
- [13] LIN, C. C., REID, W. H., Handb. Phys. 8 2 (1963) 438.
- [14] ROBERTSON, H. P., Proc. Camb. phil. Soc. 36 (1940) 209.
- [15] BATCHELOR, G. K., *The Theory of Homogeneous Turbulence*, Cambridge (1953).
- [16] CHANDRASEKHAR, S., Proc. R. Soc. Lond., Ser. A 229 (1955) 1; Phys. Rev. 102 (1956) 941.
- [17] KRAICHNAN, R. H., Physics Fluids 8 (1965) 575; 9 (1966) 1728, 1884.
- [18] ONSAGER, L., Nuovo Cim. Suppl. 6 (1949) 279.
- [19] MONTROLL, E. W., *Boulder Lectures in Theoretical Physics* 10A (1967) 531.

# QUANTUM OPTICS OR QUANTUM ELECTRONICS\*

C. H. TOWNES

University of California,  
Berkeley, Calif., United States of America

## Abstract

QUANTUM OPTICS OR QUANTUM ELECTRONICS. Introduction; 1. Maser-type amplification; 2. Achievable characteristics of maser oscillators; 2.1. Intensity; 2.2. Pulse length; 2.3. Frequency limits; 3. Theory of maser oscillators; 4. Non-linear optics - induced transparency; 5. Non-linear optics - general comments; 6. Parametric conversion of electromagnetic waves; 7. Other variations of non-linearities.

## INTRODUCTION

The theory of quantum optics goes back to the photo-electric equation and the Bohr atom, and its basis was essentially finished with the development of quantum mechanics of electromagnetic interactions and quantization of the electromagnetic field some decades ago. Modern aspects of quantum electrodynamics play almost no role in the field. What is new are methods of producing and controlling electromagnetic waves which tap, for experimental and technological purposes, previously unutilized aspects of the theory, and which make possible the experimental examination of new regimes. These new regimes, such as very high intensity, sensitive amplification, and extraordinary spatial and temporal coherence, bring out a wide variety of new effects which, while involving no fundamentally new principles, have stimulated a considerable extension and elaboration of the basic theory.

The new regimes are illustrated by noting improvement in four dimensions:

- (1) Time coherence, where an improvement of about  $10^6$  has been achieved. Thus, whereas light could previously be made to interfere with itself for path length differences of about 1 m, it is now adequately monochromatic to use path differences of  $10^6$  m, assuming any of sufficient stability could be found.
- (2) Directivity, where a factor of about  $10^6$  has also been gained. Thus, light from a 2 W laser has such spatial coherence that it can be directed to a small area on the moon and easily seen there. Recently, such a beam sent from the Los Angeles area was detected on the lunar surface, and was orders of magnitude brighter there than all the lights of Los Angeles.
- (3) Sensitivity, where a factor of about 100 improvement in detection of microwaves has been made, and similar or larger factors seem possible in much of the infra-red region.

---

\* Work partially supported by the U.S. Army and by the National Aeronautics and Space Administration.

- (4) Intensity, where the total light intensity per unit area achievable has been increased about a factor of  $10^9$ . The effective radiation temperatures, i.e. the intensity per unit area per solid angle per frequency interval, has been increased by a factor of  $10^{15}$  to  $10^{20}$ .

While terms such as quantum optics or quantum electronics are suggestive, neither describes very adequately what I believe most useful to discuss here — a variety of recent developments in the control of electromagnetic waves and study of their interactions with matter. These developments have made much more evident the unity between light, produced characteristically by atomic transitions, and radio waves, normally produced by the techniques of electronics. The great intensities available make light seem even more classical than before, while more refined control of electromagnetic radiation and new methods require increasing use of quantum mechanics, even at the longer radio waves.

The most important single element in our new control over electromagnetic waves is the ability to use atomic and molecular transitions to produce coherent amplification and generation and thus both to refine the performance of such devices and to extend their operating frequency range by several orders of magnitude, or past the visible frequencies. Masers and lasers are the common embodiments of such amplification by stimulated emission.

## 1. MASER-TYPE AMPLIFICATION

Coherent amplification implies the amplification without frequency distortion of a wave having a frequency defined with arbitrary exactness. Such amplification cannot result from interaction between a wave and any matter in thermal equilibrium since the radiation is at a very high temperature and cannot be increased in intensity by matter at a finite temperature. Hence for amplification, matter must be clearly used under other conditions, and an ensemble of atoms or molecules distributed between two discrete energy levels can escape having a definable (positive) temperature by two methods:

- (1) There may be more elements in the upper than in the lower state, giving an "inverted" population. This corresponds to a Boltzmann distribution of negative absolute temperature.
- (2) The elements may be in mixed (upper and lower) states, with population inversion, but with phase correlation between individual elements rather than the randomness required by temperature equilibrium.

Most attention has been directed to case (1) — inverted population. However, either situation can successfully amplify by stimulated emission, and frequently both occur in a given device, as illustrated in Fig.1.

In the case of an inverted population, with no impressed phase correlation, the rate of growth of electromagnetic energy travelling through material at resonance is given by

$$\frac{d(E^2)}{dt} = E^2 \left[ \frac{4\pi^2 \mu^2 \nu (N_U - N_L)}{h \Delta \nu} \right] \quad (1)$$

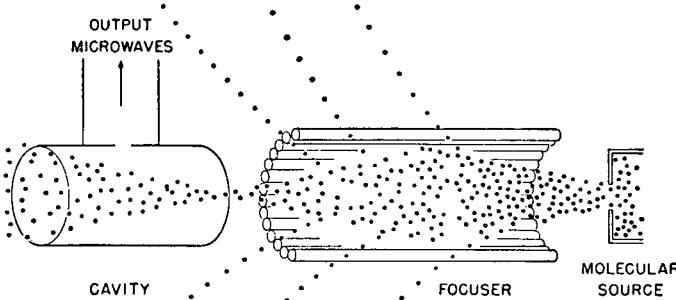


FIG.1. The ammonia (beam-type) maser. Molecules diffuse from the source into a focuser where the excited molecules (open circles) are focused into a cavity and molecules in the ground-state (solid circles) are rejected. A sufficient number of excited molecules will initiate an oscillating electromagnetic field in the cavity, which is emitted as the output microwaves. Because of energy given to the field, a large fraction of the molecules return to the ground-state toward the end of their transit through the cavity. On first entering the cavity, the molecular population is inverted. Near exit from the cavity, they show phase correlation but not necessarily population inversion.

where  $\mu$  is the matrix element for the transition,  $h$  Planck's constant,  $N_U$  and  $N_L$  the density of molecules in the upper and lower states, respectively, and  $\Delta\nu$  the molecular resonance half-width. If the radiation is contained in a resonant cavity with a fractional rate of loss of energy  $1/t$  due to wall resistance or radiation from the cavity, then the threshold condition for a self-sustained oscillation to occur is

$$N_U - N_L \geq \frac{\Delta\nu}{\nu} \frac{h}{8\pi^2\mu^2 t} \quad (2)$$

or

$$N_U - N_L \geq \frac{\Delta\nu}{\nu} \frac{h \Delta\nu_c}{4\pi \mu^2}$$

where  $\Delta\nu_c$  is the half-width of the cavity resonance. Case (2), which we shall ignore for a while, involves no such threshold, but can generate some coherent electromagnetic energy regardless of how great the losses. It amplifies only those waves, of course, which have suitable relation in phase with the molecular oscillators.

Expression (2) indicates that, if the ratio  $\Delta\nu/\nu$  is approximately constant, e.g. for Doppler broadening, and the cavity wall losses are not strongly dependent on frequency, then it is not more difficult to produce a coherent oscillator at high frequency than at low frequencies — a result which is essentially correct for frequencies as high as the ultraviolet where other effects, to be discussed shortly, come into play.

This type of amplification and oscillation, partly because of its very elementary nature, gives us, in a certain sense, the most sensitive conceivable detection and the most perfect electromagnetic sources.

Their quality is ultimately limited by the uncertainty principle, for, if a wave is amplified by a large factor, the number of quanta becomes so great that it behaves essentially classically, and one may in principle measure the electric and magnetic fields of the output with arbitrary fractional accuracy. If the amplifier relates the output fields uniquely to the much smaller input fields, then it allows an arbitrarily precise measure of the conjugate variables  $E$  and  $H$ . Hence, one must conclude that the amplifier cannot produce a perfect representation, i.e. it must be noisy.

While the phase of an electromagnetic wave does not correspond precisely to a quantum mechanical operator, it may be approximately so represented when the number of quanta is appreciably greater than unity. Hence the uncertainty relation between  $E$  and  $H$  may be reasonably well expressed by the rather useful form

$$\Delta n \Delta \phi \geq \frac{1}{2} \quad (3)$$

Here  $\Delta n$  is the uncertainty in the number of photons in an electromagnetic signal and  $\Delta \phi$  the uncertainty in radians of the phase. Any amplification or detection must satisfy such a relation, a fact having interesting aspects in non-linear optics, which will be discussed later.

Normal photodetectors satisfy (3) by giving us no information about phase, so that the uncertainty  $\Delta n$  in numbers of photons counted can approach zero. Maser or laser amplification, of the general type outlined above, does reproduce the phase of an amplified wave, and hence must give both phase and intensity fluctuations. Theoretically, for such a system with optimum design and  $N_U \gg N_L$ , there is essentially the same fractional uncertainty for each.  $\Delta n = \sqrt{n}$  and  $\Delta \phi = 1/2\sqrt{n}$ , just satisfying the uncertainty limit. Actual experimental arrangements come rather close to this limit. Various other systems, involving something like maser amplification in combination with photodetection, can in principle give any ratio of  $\Delta n$  to  $\Delta \phi$ , still preserving the uncertainty relation (3).

The fluctuations required by uncertainty are introduced into the amplifying system by spontaneous emission; hence a fixed minimum in the ratio of spontaneous emission to stimulated emission is required for any amplifying device which reproduces phase. This corresponds to the relation between Einstein's A and B coefficients.

Perhaps the easiest conceptual amplifier to consider is a wave packet of electromagnetic energy travelling through a medium of inverted population, with the gain given by Eq.(1). The probability  $P_{n,m}$  of having  $m$  photons at a time  $t$  if there were  $n$  photons initially is then given by

$$\frac{dP_{n,m}}{dt} = a[N_U(m-1) + 1] P_{n,m-1} + aN_L(m+1) P_{n,m+1} - a[(N_U + N_L)m + 1] P_{n,m} \quad (4)$$

This represents simply a photon cascade, with both production of new photons and absorption of some of those already present. In the first term,  $aN_U(m-1)$  corresponds to the probability of stimulated emission when there are  $m-1$  photons, and  $a$  to the probability of spontaneous emission. The

second term corresponds to absorption of a photon when there are  $m+1$  and the third term to any process occurring when there are  $m$  photons. It is clear that such a cascade, like a cosmic ray shower, will have a calculable spread in intensity after a large number of processes have taken place. Normal cosmic ray showers can have somewhat smaller spread because they do not maintain the phase of the initial particle and hence have no need for spontaneous emission.

Spontaneous emission associated with the amplifying mechanism also prevents maser-type oscillators from being perfectly coherent. Spontaneously emitted radiation mixes with the wave already present and produces phase fluctuations. Any amplitude fluctuations are normally much suppressed because oscillations build up to the point where the amplitude is limited by slow-acting non-linearities, while the phase is free to vary. The spectral half-width at half maximum for such a phase-fluctuating oscillator under steady conditions is

$$\Delta\nu_{\text{osc}} = \frac{2\pi h\nu}{P} (\Delta\nu_{\text{eff}})^2 \frac{N_U}{N_U - N_L} \quad (5)$$

or

$$\Delta\nu_{\text{osc}} = \frac{2\pi h\nu}{P} (\Delta\nu_{\text{eff}})^2 \quad \text{for } N_U \gg N_L$$

Here  $P$  is the power generated by the oscillator and  $\Delta\nu_{\text{eff}}$  depends on the molecular resonant width  $\Delta\nu$  and the resonator width  $\Delta\nu_c$  as

$$\frac{1}{\Delta\nu_{\text{eff}}} = \frac{1}{\Delta\nu} + \frac{1}{\Delta\nu_c} \quad (6)$$

Depending on conditions, the fractional half-width  $\Delta\nu_{\text{osc}}/\nu$  given by Eq.(5) is typically in the range  $10^{-16}$  to  $10^{-20}$ , and smaller than one can usually measure because of small, experimentally unavoidable changes with time of various parameters of the system. Oscillators have been demonstrated, however, in both the radiofrequency and the optical range with  $\Delta\nu_{\text{osc}}/\nu$  as small as about  $10^{-13}$ . Furthermore, spontaneous emission fluctuations given by Eq.(5) have been demonstrated at optical frequencies by operating laser oscillators at sufficiently low power levels.

It should be noted that use of feedback, as provided by a resonator, not only reduces the physical size of apparatus needed to obtain a desired gain (in the microwave range, gain lengths from expression (1) are typically tens of metres), it also profoundly affects other characteristics. By contrast with the very narrow frequency width given by Eq.(5), amplification of a travelling wave packet by photon cascade is much broader in response. As an amplifier, the latter still gives a performance limited only by the uncertainty principle, but as a source of power it is not very coherent, higher gain  $G$  in the centre of the molecular resonance simply emphasizing the central frequencies and giving a width  $\Delta\nu'$  which is

$$\Delta\nu' = \Delta\nu \left( \frac{\log 2}{\log G} \right)^{\frac{1}{2}} \quad (7)$$

Consider now the resonator. It is a strictly classical affair. However, recent developments have brought some additional understanding of both detailed and qualitative features of resonators, particularly those with only two reflecting walls and with other surfaces open or very lossy. Computer solutions have been worked out, for example, for the low-loss resonances of two plane-parallel mirrors, and for some curved mirror configurations which are related. Parallel mirrors represent a so-called unstable configuration where a single mode, that corresponding most closely to a plane wave travelling perpendicularly to the mirror surface, has appreciably smaller loss rates than other modes. In this case, oscillations build up in essentially a single mode, even though the dimensions are perhaps  $10^4$  times larger than the optical wavelength so that  $10^{10}$  or more modes may exist. Because the energy output can be many watts, and oscillation occurs in a single or in a few modes only, the quantum numbers for the field oscillators are as high as  $10^{20}$  or  $10^{30}$ . This makes such an oscillator about as close to a classical situation as a 10 W radio oscillator operating at 1 MHz. It has been interesting, but not surprising, to see that in some of the early work with maser and laser oscillators, electrical engineers accustomed to classical radiofrequency fields where  $h\nu$  is negligible found their intuitions frequently better than those of physicists accustomed to think of individual quanta.

While calculations for such a resonator give some variation in phase across the mirror, the wave is still remarkably close to a plane wave. Some of it transmitted through the mirror surface, as the emitted beam of a laser, hence gives light whose angular divergence is limited only by the laws of diffraction rather than by thermodynamic considerations of a fixed amount of power per unit area per unit solid angle. Spontaneous emission must produce some relative fluctuations in phase as a function of position on the mirror surface, as well as the fluctuations in time which are associated with the spectral width. These have not yet been calculated, presumably because they are very small compared with the diffraction width in any real case.

Lenses or other optical surfaces can transform the wave from its approximately planar form to any other ideal form, such as a plane wave or arbitrarily large diameter or a source limited in size only by the light wavelength.

## 2. ACHIEVABLE CHARACTERISTICS OF MASER OSCILLATORS

We have already, with the above, considered some of the limiting characteristics of quantum-mechanical amplifiers and oscillators — their detection sensitivity limits and the limits of temporal and spatial coherence. There are other important properties which will be reviewed briefly.

### 2.1. Intensity

The possibility of amplification means that there is no specific limit on intensity, but rather the practical limits of available energy and methods of containment. Powers as high as  $10^{12}$  W have been obtained by emitting energy in pulses as short as  $10^{-11}$  s. Such power cannot normally be trans-

mitted through materials at intensities greater than about  $10^{10}$  W/cm<sup>2</sup> without the onset of some of the instabilities associated with non-linearities which are discussed below.

### 2.2. Pulse length

The shortest pulse length achievable in principle is approximately equal to  $1/\Delta\nu$ , where  $\Delta\nu$  is the width of the atomic or molecular resonance. In most solids and liquids,  $\Delta\nu$  is less than a few tens of wave numbers, and in fact pulses as remarkably short as about  $10^{-12}$  s (a 1/30 cm long wave-train!), which approximately correspond to this spectral width, have been obtained. However, certain dyes which have been demonstrated to give gain have apparently usable spectral widths of thousands of wave numbers, so that one might eventually obtain pulse lengths still one hundred times shorter.

### 2.3. Frequency limits

It has been seen from expression (2) that threshold conditions for oscillation do not appear more difficult to meet at very high frequencies than at lower frequencies; however, there is a subsidiary problem which is not apparent from this condition and which becomes important at frequencies higher than about the ultraviolet region. This concerns the power required to maintain the threshold condition in the face of spontaneous emission, which tends to decrease  $N_U$  at a rate equal to

$$\frac{dN_U}{dt} = - \frac{64\pi^4 \nu^3 \mu^2 N_U}{3hc^3} \quad (8)$$

Thus, the power required to maintain oscillation for a fixed value of  $\Delta\nu/\nu$  and fixed cavity decay rate increases as  $\nu^4$ . Numerically, this power is a few milliwatts in the optical range, but as high as  $10^{12}$  W in the short X-ray region. While such power can in principle be supplied, at least by another laser, materials will not withstand it for long and the problem of X-ray maser oscillators is very difficult. So far, such oscillators have been extended from frequencies of about  $10^5$  Hz up to about  $2 \times 10^{15}$  Hz, or 1500 Å. It should also be noted that maser oscillation and related instabilities seem now to be a rather common phenomenon. Many hundreds of different systems, using solids, liquids and gases, powered by almost every conceivable form of energy, and operating over almost any wavelength within the range mentioned in expression (3) have been built. A powerful naturally occurring amplifying system has been discovered in our galaxy, where a large number of regions clearly produce OH in inverted states and amplify its characteristic microwave radiation.

## 3. THEORY OF MASER OSCILLATORS

The theory of maser (or laser) oscillators is somewhat more complex than that for maser amplifiers because saturation of the available supply of molecular energy is an inherent part of the oscillator problem, and hence one must deal with a varying molecular distribution as well as a

varying field. However, various models representing a number of stages of detail have been successfully developed. A semiclassical model which treats the molecular transitions quantum-mechanically and rather exactly, but uses a classical field and classical cavity losses is quite successful because, as we have seen above, the field contains so many quanta per mode that it is very close to the perfect classical case. For this model, excited and ground-state molecules may be introduced into the cavity where a classical field is assumed to be present and their time development worked out by well-known methods. From this, one can calculate the real and imaginary parts of the macroscopic induced polarization. The polarization provides energy to the field and also, with the resonant cavity characteristics, determines the frequency at which the field can acquire maximum strength, i.e. the eventual oscillator frequency. For a steady-state oscillator, there needs only to be self-consistency between the assumed steady-state field, its classical losses, and its calculated gain by stimulated emissions. Development of oscillations or other time variations of the field can also be calculated if they are much slower than the molecular relaxation time. The limitation in oscillator amplitude by non-linearities can be well treated in this model and, through these non-linearities, also the competition between various modes of oscillation or various polarizations. Frequently, one particular polarization and frequency suppresses gain and oscillation of others.

The semiclassical model does not directly give effects of spontaneous emission nor, hence, the resulting fluctuation phenomena. Noise in the form of thermal radiation from the cavity walls or exterior can be introduced into the equations, and gives fluctuations of much the same character as spontaneous emission. In fact, spontaneous emission itself may be simply introduced as a small perturbing field and valid calculations of the fluctuation phenomena made. However, for careful treatment of spontaneous emission, especially during the initiation of oscillation, a quantum mechanical treatment of the field as well as the molecules is desirable.

There are a number of approaches to a more completely quantum mechanical calculation; Scully and Lamb [1] have given a particularly lucid discussion from a density matrix calculation. Equations for the diagonal elements of the density matrix, giving the probability of finding  $m$  photons, have a form rather similar to that of Eq.(4), except that the gain due to molecular transitions is reduced by a saturation factor when the number of photons is high. These equations assume the common case that the rate of change of field strength is slow compared with molecular relaxation, and under this assumption the build-up of oscillation and other dynamical behaviour have been computed.

Very commonly, such oscillators emit short pulses rather than a steady-state oscillation, and various questions of non-linearities, spatial distribution of the fields, and relaxation mechanisms between molecules, or between molecules and various degrees of freedom of the amplifying medium, must be examined carefully in order to obtain a detailed description of such dynamic behaviour. Not infrequently, the pulses are so short and intense that the approximation of slow changes in the field by comparison with molecular relaxation is no longer valid. Treatment of these problems is still rather incomplete.

The nature of the photon distribution in a steadily oscillating laser deserves some additional comments. First, while the number of photons

in the oscillating field is not definite, it is strongly peaked about a particular value  $n$ , with a fractional variance  $1/\sqrt{n}$  for the ideal case where there is not much loss. This is in considerable contrast with black-body radiation, which would have a peak probability of zero photons, regardless of the temperature. Furthermore, for a short time the field may be approximately described by a displaced ground-state wave function

$$|\psi(q, t)|^2 = \frac{\alpha}{\pi^{\frac{1}{2}}} e^{-\alpha^2(q - a \cos \omega t)^2}$$

so that diagonal elements of the density matrix are

$$\rho_{nn} = \frac{\alpha^{2n}}{n!} e^{-\alpha^2}$$

This is the so-called "coherent state", representing an oscillating wave packet of frequency  $\omega$  and amplitude  $a$  with minimum uncertainty. In time, spontaneous emission or thermal noise, if present, produces random variations in the phase. This corresponds to the small spread in frequency  $\omega$ , as noted in Eq.(5) which gives the frequency band width.

The usefulness of coherent states in describing such a system and its statistics has been emphasized by Glauber [2], who has given extensive discussion of the statistics and defined appropriate correlation functions for examining the coherence of the field.

#### 4. NON-LINEAR OPTICS – INDUCED TRANSPARENCY

There are many developments, such as new types of spectroscopic measurements, production of  $\gamma$ -rays by inverse Compton scattering, or holography (a new method of processing waves), which come out of our new control over light. However, the most striking general class of phenomena of novelty and interest to physicists have been those due to the very high intensities available, the general assemblage of which is usually called non-linear optics. Non-linear optics treats cases where there is a dependence of electric or magnetic susceptibilities on optical field strengths.

The first such effect to be considered here, induced transparency, involves no property of matter not considered above, yet it was quite unexpected. As light intensity is increased, any absorbing medium may become saturated if induced transitions are more rapid than relaxation. What was not expected was complete transparency of the material at the resonant absorption frequency, with its molecules left in the ground-state after passage of a pulse. This result was found by McCall and Hahn [3] by computer study of transmission. They showed that if a pulse of coherent light is short compared to the relaxation time (but not necessarily short compared to the inverse line width), and the electric field has a form

$$E = \frac{\hbar}{|\mu|r} \operatorname{sech}\left[\frac{1}{\tau}\left(t - \frac{z}{v}\right)\right]$$

then it can be propagated at an absorption resonance without loss of energy or change of shape. Here  $t$  and  $z$  are time and propagation distance,  $v$  the velocity,  $\tau$  an arbitrary time constant, and  $\mu$  the dipole moment for the

absorbing transition. Such a pulse takes a molecule from the ground-state to the upper state and back again, leaving it with no net energy. Induced transmission sets in only after the intensity reaches a certain level, after which it is the stable form of pulse propagation, all other pulse forms being distorted into the shape given above. Induced transmission has been well confirmed experimentally.

## 5. NON-LINEAR OPTICS - GENERAL COMMENTS

To illustrate other surprises which occur in the new regime of very intense, highly directed light beams, consider Fig.2, which is a cross-section of the intensity of a laser beam after transmission through a few centimetres of transparent liquid. The beam, with power flow of about 1 MW, entered the liquid as a close approximation to a plane wave with a beam diameter of a few millimetres. Figure 2 shows, on a magnified scale, part of the hundreds of tiny circular filaments of light, each spot representing about 20 kW of power transmitted temporarily as a tiny sub-beam of diameter only about  $4\mu\text{m}$ , or 6 wavelengths. This behaviour is not completely understood but, along with a wide variety of other new effects, one can understand many of its features in terms of non-linear indices of refraction.

Ordinary optics dealt with a wave equation of the form

$$\nabla^2 E = \frac{1}{c^2} \frac{\partial^2 (\epsilon \mu E)}{\partial t^2} \quad (9)$$

where  $\epsilon$  and  $\mu$  are the electric and magnetic permittivities. We are accustomed to the cases where  $\epsilon$  and  $\mu$  are functions of frequency (dispersion), functions of position (refraction), and even functions of time (Doppler effect, or moving optical systems). Non-linear optics is concerned with the cases where  $\epsilon$  or  $\mu$  are also functions of the field strength  $E$ . The resulting phenomena are much more complex because variation of  $\mu\epsilon$  with position or time is now determined by solutions for  $E$  of the wave equation itself.

For simplicity, and as a good approximation to most actual cases, we shall assume  $\mu$  is a constant and equal to unity, so that non-linear properties are demonstrated by  $\epsilon$  alone.  $x = (\epsilon - 1)/4\pi$  is, of course, the electric susceptibility. It can be expanded as a Taylor series (except for cases where ionization or other catastrophic changes are induced by the field) and the polarization per unit volume written

$$\vec{P} = x_1 \cdot \vec{E} + x_2 : \vec{E} \vec{E} + x_3 : \vec{E} \vec{E} \vec{E} + \dots \quad (10)$$

This expression assumes that the polarization responds instantly to the field, which is not always an adequate assumption, but covers a wide range of phenomena. The first term is the usual polarization of a dielectric material. The second coefficient,  $x_2$ , as well as all  $x_n$  for even  $n$ , is zero in all those materials having a centre of inversion, since under such inversion  $\vec{P}$  must change sign. The third term is usually not zero and is of

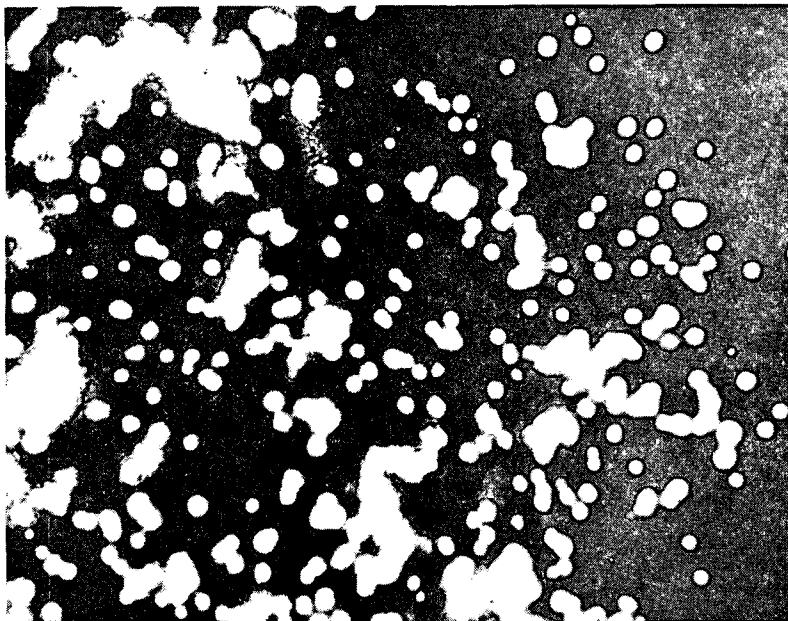


FIG. 2. Highly magnified cross-section of intense laser beam after traversing about 10 cm of  $\text{CS}_2$ . The beam, originally nearly uniform and a few millimetres in diameter, has split up into many tiny filaments, each about  $5 \mu\text{m}$  in diameter.

greatest importance for non-linear effects in the majority of materials. It is a fourth-rank tensor. Relative size of terms in the expansion depends, of course, on the particular material involved. However, for atoms the Taylor expansion is essential in powers of  $E/(e/a_0^2)$ , where  $e/a_0^2$  represents the internal atomic fields. Hence a field of many millions of volts/cm is required in order to make the successive terms comparable in size. For simplicity, we shall pick a particular direction and polarization so that tensor properties of  $\chi$  may be neglected, and write a particular component of the polarization as

$$\mathbf{P} = \chi_1 \mathbf{E} + \chi_2 \mathbf{E}^2 + \chi_3 \mathbf{E}^3 + \dots \quad (11)$$

Non-linearities of the form given by (11) were not previously unknown, but could be detected only with static or very low-frequency fields because only these could be made sufficiently strong. Both the Kerr effect and electrostriction represent  $\chi_3$  terms. The electro-optic effect, occurring in crystals without inversion symmetry, is due to the second term. Another interesting, still better-known example of  $\chi_2$  is the photoelectric emission, which establishes a low-frequency polarization in a circuit as the result of light intensity, or hence the square of an optical field. Even for normally isotropic materials,  $\chi_2$  can be non-zero for this case because a surface, which is required for a photoelectric emission, destroys the possibility of inversion symmetry.

## 6. PARAMETRIC CONVERSION OF ELECTROMAGNETIC WAVES

Consider now the general behaviour of an intense laser beam travelling through matter. The beam itself is at a very high temperature,  $10^{20}$  °K multiplied by plus or minus a few powers of ten. The material is, of course, at a much lower temperature. Hence there must be a flow of energy from the beam to various modes of excitation of the material. The normal means for such energy transfer to establish equilibrium is through non-linear coupling. The most powerful and rapid energy transfer occurs when there is an instability, so that some particular mode builds up exponentially, obtaining a large fraction of the beam energy and thus becoming violently excited — if this is an acoustical mode of a solid material, for example, the excitation is frequently violent enough to crack the material.

An enormous variety of effects spring from the non-linear equation (9). We shall first pay particular attention to various simple instabilities due to specific non-linearities which can occur in isolatable form. One such case involves three electromagnetic fields. If all fields are travelling in the  $z$  direction, they may be written

$$E_j = \vec{\mathcal{E}}_j(r) e^{i(\omega t - k_j z)} \quad (12)$$

where  $j = 1, 2$  or  $3$ , and  $\mathcal{E}_j$  is assumed to vary slowly with  $z$  by comparison with the optical wavelength  $2\pi/k_j$ . The result is three equations coupled by the non-linearity, and of the form

$$\begin{aligned} \frac{d\mathcal{E}_1}{dz} &= \frac{-i\omega_1}{2\sqrt{\epsilon}} \chi_2 \mathcal{E}_3 \mathcal{E}_2^* e^{-i(k_1 - k_2 - k_3)z} \\ \frac{d\mathcal{E}_2^*}{dz} &= \frac{i\omega_2}{2\sqrt{\epsilon}} \chi_2 \mathcal{E}_1 \mathcal{E}_3^* e^{i(k_1 - k_2 - k_3)z} \\ \frac{d\mathcal{E}_3}{dz} &= \frac{-i\omega_3}{2\sqrt{\epsilon}} \chi_2 \mathcal{E}_1 \mathcal{E}_2 e^{i(k_1 - k_2 - k_3)z} \end{aligned} \quad (13)$$

Here  $\omega_1$  has been taken equal to  $\omega_2 + \omega_3$ , corresponding to conservation of energy. For the coupling to have large effects,  $\Delta k = k_1 - k_2 - k_3$  must also be small, corresponding to conservation of momentum, and we shall assume it zero. Consider the case where  $\mathcal{E}_1$  is very large,  $\mathcal{E}_2$  small and  $\mathcal{E}_3$  initially zero. The solution then gives a flux of photons in the two weak fields of

$$n_2 = n_2(0) \cos h^2 \left( \frac{4\pi\chi_2}{c\sqrt{\epsilon}} \sqrt{\omega_2\omega_3} z \right) \quad (14)$$

$$n_3 = n_2(0) \sin h^2 \left( \frac{4\pi\chi_2}{c\sqrt{\epsilon}} \sqrt{\omega_2\omega_3} z \right)$$

where  $n_2(0)$  is the flux at  $z = 0$ . This case is known as parametric amplification, since the field  $\mathcal{E}_2$  can be amplified indefinitely.

A second interesting case is where the highest frequency field,  $\mathcal{E}_1$ , is initially zero,  $\mathcal{E}_2$  large, and  $\mathcal{E}_3$  small. The solution is then

$$\begin{aligned} n_3 &= n_3(0) \cos^2\left(\frac{4\pi\chi_2}{c\sqrt{\epsilon}}\sqrt{\omega_1\omega_3}z\right) \\ n_1 &= n_3(0) \sin^2\left(\frac{4\pi\chi_2}{c\sqrt{\epsilon}}\sqrt{\omega_1\omega_3}z\right) \end{aligned} \quad (15)$$

This is called up-conversion, because low-frequency photons ( $n_3$ ) are converted into high-frequency photons ( $n_1$ ) with the total number conserved.

It is instructive to note that both of these processes preserve phase information. However, while parametric amplification can be an excellent method of amplification, it increases the total number of photons and hence must have the noise required by uncertainty and produced by spontaneous emission. Up-conversion does not amplify the number of photons and introduces no noise by spontaneous emission. This process is actually quite attractive as a method of detecting infra-red radiation because by it the infra-red can be converted, quantum for quantum, into visible light which can be seen or photographed. Furthermore, because the conversion can maintain spatial relations between rays and hence reproduce a field, it appears especially attractive for infra-red astronomy, assuming technical problems can be happily solved.

This raises the question whether up-conversion can occur without parametric amplification (or down-conversion) which would produce noise. It can, if the non-linear material is dispersive, since then there can be momentum matching ( $\Delta k = 0$ ) for up-conversion, but mismatching ( $\Delta k \neq 0$ ) for parametric amplification. When there is a large mismatch, the down-conversion does not build up exponentially although it has a small fractional amplitude of maximum value  $[(4\pi\chi_3/c\sqrt{\epsilon})(\sqrt{\omega_2\omega_3}/\Delta k)]^2$ . Thus spontaneous breakup of a quantum of frequency  $\omega_1$  into  $\omega_2$  and  $\omega_3$  cannot occur because momentum cannot thus be conserved. In this and other ways interchange of energy between the coupled waves corresponds to particle decays, but with the difference that the coupling between two waves can be changed at will by changing material and the strength of the large field.

The magnitude of frequency conversion where there is imperfect momentum matching can be strongly affected by boundary conditions. For example, a change of material at a surface produces a boundary which can radiate converted waves. This type of production at surfaces has been discussed in detail by Bloembergen [4], along with a wide variety of other non-linearities.

The coupled equations (13) are essentially similar to those encountered in many areas of physics, such as geophysics, acoustics, and plasma physics; they are characteristic of parametric coupling. The driving term in these equations allows amplification of a wave of a particular frequency and phase. The population of states need not be inverted, for this is an example of case (2) discussed at the beginning, where phase coherence occurs. Only a few particular solutions for these equations have been examined here but there are many others of interest to non-linear optics.

An additional large class of phenomena occur because the susceptibility can vary with time due to change of some co-ordinate of the optical medium. To illustrate this,  $\chi_2$  will be assumed zero (as in any isotropic medium which is not optically active),  $\chi_3$  negligibly small and  $\chi_1 = \chi_1(0) + aq$ .

The polarization energy is hence  $W_q = -\frac{1}{2}[\chi_1(0) + aq]\vec{E} \cdot \vec{E}$ , and there is a force driving  $q$  of  $F_q = -\partial W/\partial q = +\frac{1}{2}a E^2$ . Since the displacement  $q$  from equilibrium is normally proportional to the force,  $q = k F_q$ , the polarization is

$$\vec{P} = \chi_1(0)\vec{E} + aq\vec{E} = \chi_1(0)\vec{E} + \frac{a^2 k}{2} \vec{E} \cdot \vec{E} \vec{E} \quad (16)$$

Thus there is an effective value of  $\chi_3$  which is  $a^2/2k$  for any case where a co-ordinate of the medium changes the susceptibility, which includes almost all modes of the medium.

A well-known example is the Raman effect, associated with molecular vibration, where the molecular polarizability is

$$\alpha = \alpha_0 + \frac{\partial \alpha}{\partial q} q \quad (17)$$

and  $q$  is a vibrational co-ordinate. If there is a strong field  $E_0$  of frequency  $\omega_0$  and a weak field  $E_{\pm 1}$  of frequency  $\omega_0 \pm \omega_q$ , where  $\omega_q$  is the vibrational frequency, then the molecular vibration acquires the amplitude

$$q = \frac{\pm \tau / 2 \partial \alpha / \partial q E_0 E_{\pm}}{m \omega_q}. \text{ Here } m \text{ is the reduced mass and } \tau \text{ the relaxation}$$

time of the vibration. In this case, the vibration is driven by the product of the fields  $E_0 E_{\pm}$ , and there is a polarization  $aqE_0$  representing a driving term for the fields  $E_+$  or  $E_-$ . Thus, if  $E_0$  is very large and the other fields smaller, we have another form of parametric oscillation, with gain for the field  $E_-$  which is of lower frequency, and up-conversion but no gain for the field  $E_+$  of higher frequency. The phenomenon is also a stimulated Raman effect, which can convert most of the field  $E_0$  into coherent molecular vibrations and a strong Raman field  $E_-$ .

The intimate relation between population inversion and phase coherence can be illustrated by looking at the stimulated Raman effect as maser action from a virtual level, as illustrated in Fig.3. The initial level, created by the high-frequency field  $E_0$  and sharing the population of the ground-state, is occupied by a larger number of molecules than the excited vibrational state, so that maser action can occur.

Once stimulated Raman radiation occurs, there is a regular distribution of coherent molecular oscillations generated, with planes of constant phase given by  $(\vec{k}_0 - \vec{k}_-) \cdot \vec{r} = \text{const.}$  Here  $\vec{k}_0$  and  $\vec{k}_-$  are wave vectors for the fields  $E_0$  and  $E_-$ , respectively. Thus there is a time-varying three-dimensional grating which can scatter any other light wave, at the same time producing side-bands at multiples of  $\omega_q$ . Thus a whole series of Stokes (down-shifted) and anti-Stokes (up-shifted) radiation is created from a single original light wave,  $E_0$ . These higher-order Raman shifted waves often extend from  $\omega_0$  into both the infra-red and the ultraviolet regions.

We have noted above that the presence of  $E_-$  produces an oscillation of  $q$  of one phase, and  $E_+$  produces the opposite phase. Thus, while stimulated Raman Stokes light can result in the production of anti-Stokes frequencies, the presence of  $E_+$  quenches the effect of  $E_-$ . When both are present and coupled to the same molecules, the Raman gain is reduced from an exponential form to linear gain only. However, this cannot happen unless there is a phase matching condition (momentum conservation) given by

$$2\vec{k}_0 = \vec{k}_- + \vec{k}_+ \quad (18)$$

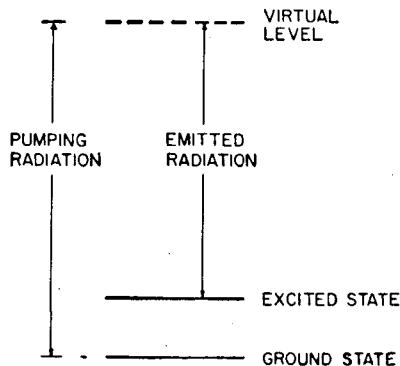


FIG. 3. Representation of energy levels in a Raman maser. This system is closely equivalent to a three-level maser, one of the levels being "virtual" or not characteristic of the molecule when no field is present.

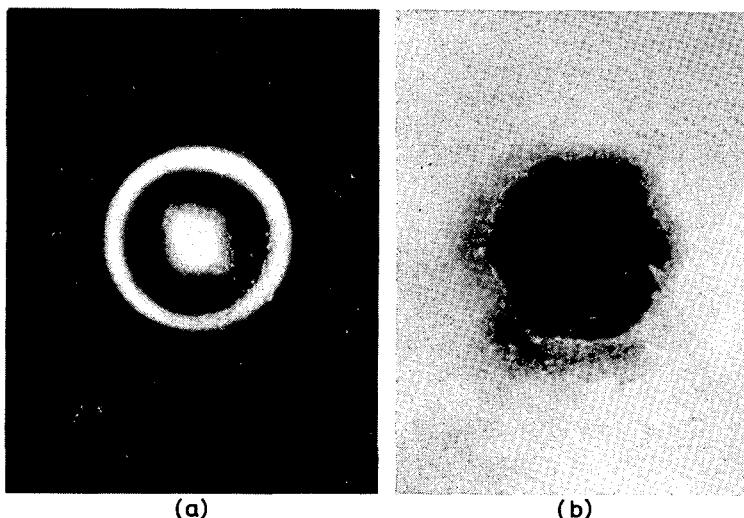


FIG. 4. (a) Cross-section of cone of anti-Stokes light, showing angular distribution emitted from a laser beam (positive photograph).

(b) Angular distribution of Stokes light from an intense laser beam, showing decreased emission in a cone angle which matches Stokes emission ( $2\vec{k}_0 = \vec{k}_- + \vec{k}_+$ ) (negative photograph). [From Chiao, R., Stoicheff, B., Phys. Rev. Lett. 12 (1964) 290.]

Only when this condition is satisfied can  $E_+$  be generated, and then generation of  $E_-$  is somewhat suppressed. If the matching condition (18) is satisfied only for unique directions of  $\vec{k}_-$  and  $\vec{k}_+$ , and if there is dispersion, these vectors are not quite parallel to  $\vec{k}_0$ , as illustrated in Fig.4, where the anti-Stokes light is generated in a cone, and there is a corresponding cone of suppression of Stokes generation. Stokes - anti-Stokes interaction has very important effects on parametric generation. When there is no dispersion, for example, it can almost completely eliminate parametric amplification.

## 7. OTHER VARIATIONS OF NON-LINEARITIES

Molecular vibrations are obviously not the only time-varying phenomena in materials which can be excited by intense light and give parametric frequency conversion. Other motions which have thus been excited, each having its own interesting characteristics, are

- acoustic waves (stimulated Brillouin scattering)
- molecular rotation in liquids (stimulated Rayleigh scattering)
- temperature fluctuations in liquids (stimulated Rayleigh scattering)
- electron orbital polarization (e.g. the parametric amplification discussed as a first example in equations)
- spin waves (in ferrimagnetic materials).

Still other motions which in principle can be stimulated are

- phase transitions (e.g. light scattering at the critical point)
- chemical reactions (chemical association and disassociation with change of polarizability)
- electrical "spin" waves (in ferroelectric materials).

In stimulated Raman scattering, the phase of molecular oscillation at one point is not uniquely determined by the phase at another point, except through the scattered light wave, and hence the scattered wave  $E$  may occur in any direction. However, for an acoustic wave (stimulated Brillouin scattering), the phases at various points are connected through a definite propagation velocity. Hence momentum can be conserved only in a particular direction given by  $\vec{k}_0 = \vec{k} + \vec{k}_{ac}$ , where  $\vec{k}_{ac}$  is the wave vector for the acoustic wave. Any motion for which wavelength varies with frequency shows a similar characteristic. For Brillouin scattering, this requirement results in a relation between the frequency shift  $\Delta z$  of the scattered light and its scattering angle  $\theta$  of

$$\Delta\nu = \frac{zv}{c} \nu_0 \sin^2 \theta / 2 \quad (19)$$

where  $v/c$  is the ratio of acoustic to light velocities, and  $\nu_0$  is the frequency of the original light.

The case of motions which are non-resonant, or lossy, such as molecular orientation or thermal fluctuations, may seem quite different from the above example of molecular vibration. Detailed characteristics are

in fact quite different, but the general scheme of parametric excitation still applies. Consider, for example, a co-ordinate driven by the two fields  $E_0$  and  $E_-$ , but with zero frequency of oscillation (free) and subject to relaxation. It follows an equation of the form  $dq/dt = -q/r + aE_0 E_- \cos \omega t$  where  $\omega$  is the frequency difference between  $E_0$  and  $E_-$ . The steady-state solution is  $q = waE_0 E_- \sin \omega t / (\omega^2 + (1/\tau)^2)$  which gives the strength of the coupling between the motion  $q$  and the field  $E_-$ . The coupling has a maximum value at  $\omega = 1/\tau$ , and hence oscillations can be expected to build up at this frequency, and measurement of the frequency difference between  $E_0$  and the scattered field  $E_-$  gives a good determination of the relaxation time  $\tau$ .

If an absorbing dye is introduced into a liquid or solid, the amount of heating will be proportional to  $E^2$ , which can vary in time if waves of two different frequencies are present. Heating expands the material, changing its susceptibility, and the heated region relaxes at a rate  $1/\tau = Ck^2/\rho C_p$ , where  $k = |\vec{k}_0 - \vec{k}|$ ,  $C$  is the thermal conductivity,  $\rho$  the density, and  $C_p$  the specific heat. This again fulfils requirements for parametric oscillation and one can obtain regular thermal oscillations at a frequency  $\omega = 1/\tau$  and of wavelength  $2\pi/k$ .

Ordinarily, thermodynamic considerations require that  $\chi_3$  be positive, because otherwise it would be possible to steadily feed energy into the electromagnetic waves from the optical material. This in turn implies that the Stokes, or down-converted frequency is amplified steadily but not the anti-Stokes. In the case of thermal fluctuations, the situation is just the opposite, because the presence of a strong field heats the material and decreases  $\chi$ . Hence it is in this case the anti-Stokes radiation which is amplified, and the wave steadily obtains energy from the material by the parametric coupling. However, light absorption is necessarily present which feeds energy steadily back into the material.

Some non-linear processes, particularly those involving multiple processes with very low or zero frequency of modulation, are best handled by integrating Maxwell's equations directly rather than by a perturbation approach, or one which follows individual scattering processes step by step, as has been done above. Several examples include self-focusing and trapping, self-steepening, and the pseudo-Doppler effect.

If a light beam of finite diameter has stronger intensity near its centre than at its edges, it will focus itself in a non-linear medium. The non-linearities produce a converging lens if  $\chi_3$  is positive, as it must be for no flow of energy from the medium. For a given beam intensity and slope, it is not difficult to work out focusing of its rays, and for beams of a few megawatts intensity the focal length is often of the order of a few centimetres. Furthermore, the light tends to trap itself in a light-pipe. The wave equation in this case has the form

$$\nabla^2 E - \frac{n^2}{c^2} \frac{\partial^2 E}{\partial t^2} - \frac{4\pi \chi_3}{c^2} \frac{\partial}{\partial t} (\overline{E^2} \vec{E}) = 0 \quad (20)$$

Here the non-linearity is assumed to be associated with some slow motion such as molecular rotation (Kerr effect) or electrostriction, and  $\overline{E^2}$  is the average over time. This equation has a series of eigensolutions of

cylindrical symmetry, the lowest one varying with the radius only, with an intensity distribution somewhat like a Gaussian curve. The power flow in the beam required for this eigensolution is

$$P = \frac{5.76 \lambda^2 c}{16\pi^4 \chi_3 n_0} \quad (21)$$

where  $\lambda$  is the wavelength, and  $n_0$  the index of refraction. This power for typical liquids and optical wavelengths is a few tens of kilowatts, while for the atmosphere it is a few megawatts. It has the remarkable property of being independent of the beam size, so that this amount of power can travel in a beam of arbitrarily small diameter without any diffraction spreading. This assumes, of course, that use of terms up to  $\chi_3$  only in the expansion of susceptibility is justified.

If the beam power is above the critical trapping power given by Eq.(21), then it tends to be focussed down to a smaller diameter, eventually reaching a diameter where higher-order terms in the expansion of  $\chi$  or parametric instabilities check further decrease in size, and then remaining at this small diameter for some distance. Actual trapped filaments have typical diameters between 1 and 20  $\mu\text{m}$ , and hence as small as a few optical wavelengths. Yet they propagate for many centimetres in a very nearly straight line without spreading.

Any optical beam with a power per spatial mode in excess of the critical trapping value is unstable and likely to contract into one or more such filaments. Inside such a filament, the field intensities are exceedingly high, since the power density is of the order of  $10^{17} \text{ W/cm}^2$ , and many parametric instabilities occur.

If the index of refraction decreases in an intense field, as it can if there is optical absorption and consequent heating, then there is self-defocusing rather than self-focusing. Such an expansion of the beam is known as thermal blooming.

A pulse of light propagating through a non-linear medium undergoes a change of shape and can produce an optical shock, or arbitrarily steep wave-front. The equation of propagation of the envelope of a pulse is

$$-\frac{\partial \rho}{\partial t} = (v_0 - 3v_2\rho) \frac{\partial \rho}{\partial z} \quad (22)$$

where  $\rho$  is the energy density in the pulse ( $\sim E^2$ ),  $v_0$  the velocity and low energy density and  $-3v_2\rho$  a non-linear velocity change. The velocity is always decreased as  $\rho$  is increased under the same general conditions that  $\chi_3$  is positive. The decreased velocity at the higher intensities is just the reverse of the acoustic case, and as a result the shock develops on the trailing, rather than the leading, edge of the pulse. Prior to the shock, the pulse slope develops according to the equations

$$\rho(z, t) = \rho(0, t_0)$$

$$z = \{v_0 t - 3v_2 \rho(0, t_0)\}(t - t_0)$$

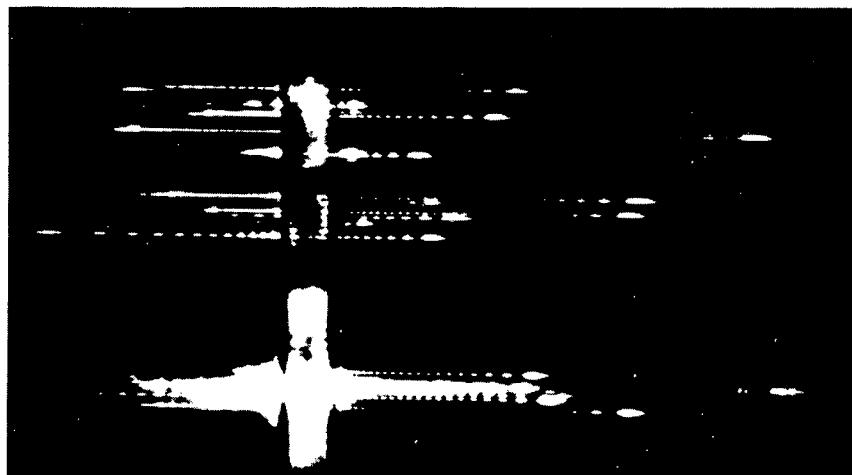


FIG. 5. Spectra of laser beam after travelling in small intense filaments for several centimetres. Each series of dashes represents the spectrum of a single filament. Original laser light was essentially monochromatic, and at the position shown by the arrow. After non-linear propagation in the filament, the spectrum is spread a few hundred wave numbers. [From Shimizu, F., Phys. Rev. Lett. 19 (1967) 1097.]

There is still another and more striking effect in the propagation of a pulse through non-linear material. Because the more intense part of the pulse travels more slowly than the less intense part, the wavelength is stretched out on the side of the pulse where intensity increases with time (negative slope), and lengthened by an amount proportional to the distance travelled and to the slope of the pulse envelope. This results in a frequency shift if the pulse traverses a length  $L$  of non-linear material which is essentially like a Doppler-effect because the optical path length increases as the pulse intensity increases. The frequency shift is given by

$$\delta\omega = \omega_0 \frac{2\pi\chi_3}{n_0} \frac{d(E^2)}{dz} L \quad (23)$$

where  $n_0$  is the index of refraction and  $z$  the distance measured along the pulse of frequency approximately  $\omega_0$  at the time it enters the material. Such shifts can be a substantial fraction of the frequency  $\omega_0$  itself. The shifts are, of course, to lower frequency on one side of the pulse and higher frequencies on the other side, because of the change in sign of the slope. The net resulting spectrum is sometimes surprising. Figure 5 shows the spectra of a number of almost monochromatic beams after they emerge from a few centimetres of path in  $CS_2$ , with frequencies spread over several hundred wave numbers. Theoretically, one should expect much the same general frequency distribution if the light is either initially modulated sinusoidally, or is a very short pulse.

While a number of separate effects have been discussed above, in many cases several may occur together, and there may be many higher-order processes involving iteration of each process in various combinations.

The most typical case where many effects occur together and there is very rapid transfer of energy is in the small collapsed beams described above, usually called small-scale trapping. Many aspects of these filaments are still puzzling, especially the reasons for their particular size, their dynamics, and the generation in them of intensity variations in times of about  $10^{-12}$  s. They are not easy to study because of the extreme conditions. The entire life of such optical filament is usually not much longer than  $10^{-10}$  s. They have optical fields which are at least as large as  $10^7$  V/cm, and frequently are high enough to cause ionization, an effect not within our present discussion. These intensities cannot very well be approached gradually because there is a threshold for formation of the filaments. When the filament is formed, the intensity is usually above the threshold for a large number of instabilities which hence all occur at once. In the more intense filaments, instabilities build up to the point where light is reflected back and forth in the filament, so that much of it is trapped longitudinally as well as being confined to a filament. These filaments, typical of light intensities above a certain threshold for each material, are likely to continue for some time to present a rich variety of puzzles.

#### REFERENCES

- [1] SCULLY, M.O., LAMB, W.E., Phys. Rev. Lett. 16 (1966) 853; Phys. Rev. 159 (1967) 208.
- [2] GLAUBER, R.J., In Physics of Quantum Electronics (KELLEY, P.L., LAX, TANNENWALD, Eds), McGraw-Hill, New York (1966) 788.
- [3] McCALL, S.L., HAHN, E.L., Phys. Rev. Lett. 18 (1967) 908.
- [4] BLOEMBERGEN, N., Nonlinear Optics, W.A. Benjamin, New York (1965).

# QUANTUM STATISTICAL MECHANICS OF SYSTEMS WITH AN INFINITE NUMBER OF DEGREES OF FREEDOM

I. PRIGOGINE

Faculté des Sciences,  
Université Libre de Bruxelles, Belgium,  
and  
Center for Statistical Mechanics,  
University of Texas,  
Austin, Tex., United States of America

## Abstract

QUANTUM STATISTICAL MECHANICS OF SYSTEMS WITH AN INFINITE NUMBER OF DEGREES OF FREEDOM. 1. Introduction; 2. Master equations and kinetic equations; 3. Invariant manifold of the dynamic evolution; 4. Quantization and physical states on the invariant manifold. 5. Conclusions and applications.

## 1. INTRODUCTION

As emphasized by Ruelle in Volume II of these Proceedings, most systems of interest to-day are characterized by a large or even infinite number of degrees of freedom. This is the case for example for the systems studied in chemical physics, hydrodynamics ... as well as for interacting fields.

Now every time we deal with such systems an essential question is the formulation of a "reduced description", of a "simplified language". Even if we could conceive computers big enough to study the molecular dynamics of say  $10^{23}$  molecules in a macroscopic system, the knowledge of their positions and their velocities would be of little interest. On the contrary, a thermodynamic or a hydrodynamic description has proved to be of tremendous value to describe the evolution of such systems.

To some extent, there is a similarity between this situation and the duality which exists between fields and particles. Even if we admit for the moment that the "basic" description of nature is in terms of interacting fields, we still have to take into account that experiments are performed on particles. Is it possible to reduce the field description to a particle theory in which virtual processes would be eliminated and only real processes be retained?

We shall see in this report that both questions are related to the formulation of laws of dynamics specifically appropriated to the study of large systems. In this way we shall be able to give at least sufficient conditions for the validity of the "thermodynamic language" on one hand, and for the reduction of field dynamics to particle dynamics on the other. It is remarkable that these conditions will appear essentially identical.

Let us first stress a few characteristics of large systems to emphasize the deep difference which exists between such systems and "small" systems with only a few degrees of freedom:

- (a) Suppose we know the wave function  $\psi(t)$  at time  $t = 0$ , then

$$\psi(t) = U(t)\psi(0) \quad (1.1)$$

This relation appears as "symmetric" in  $\psi(t)$  and  $\psi(0)$ . We may as well solve it in terms of  $\psi(0)$

$$\psi(0) = U^\dagger(t)\psi(t) \quad (1.2)$$

However, for large systems not only do we not know the initial state but we may also expect that in some sense the relation between times 0 and  $t$  becomes less symmetric. One would like to speak of "loss of information" or of "dissipation".

Indeed, as indicated by the validity of phenomenological thermodynamics, large systems seem to have a mechanism for forgetting initial conditions. For large classes of initial conditions the system approaches in time thermodynamic equilibrium.

What is the relation between this irreversible behaviour and equations (1.1)-(1.2)? Where does the large number of degrees of freedom appear explicitly?

(b) If in small systems we inverse the velocities (or the time) we trace back the evolution: the direct evolution and its time inverse appear as essentially symmetric. For large systems, this is in general no more so.

As an example, let us consider a scattering experiment: a beam of monochromatic particles is scattered by a centre of force. As a result, the velocity distribution becomes more isotropic and simultaneously velocity-dependent space correlations appear between the centre of force and the scattered particles (for more details see Prigogine and Résibois, 1964). If we now use a mirror to inverse the velocities we go from a more symmetric velocity distribution to a less symmetric one, and simultaneously the correlations disappear. While both motions are in agreement with the equations of motion, the type of physical effects associated with one or the other is deeply different. Only the first is associated with the usual idea of "approach to equilibrium" in which the velocity distribution tends to the spherical symmetric Maxwell distribution.

(c) In small systems such as an atom (neglecting the radiation field) we may perform simultaneously all the measurements in the limits permitted by quantum mechanics. In large systems we measure only a very restricted subclass of observables which depend on a small number of particles (usually one or two). Fortunately, we never have to study a quantity such as the product of the x-co-ordinates of all particles present!

It is very important that all thermodynamic properties (such as energy, pressure, etc.) depend only on this class of observables. We shall call them the macro-observables (we are aware that this word is used in a different sense in part of the literature but we have not found a better word).

The irreversible "thermodynamic" evolution refers only to macro-observables.

What we should like to show is that there exists a "macromechanics", a mechanics dealing specifically with large systems and with the evolution of macro-observables. The essential feature will be that well defined classes of trajectories which are distinct in terms of (1.1) can be treated as identical from the point of view of this macromechanics. Inversely, each trajectory in this macromechanics corresponds to the mechanical evolution of systems which differ only in quantities which are unobservable.

The existence of such kind of macromechanics is certainly the most striking result which emerges from the study of dynamics of large systems. As we shall show, concepts such as entropy, "physical particles" or quantum states of finite life time, find their natural place in the frame of this formulation of mechanics.

## 2. MASTER EQUATIONS AND KINETIC EQUATIONS

As a preliminary step, let us summarize briefly the derivation of the master and kinetic equations. Of course in this paper we shall not give any proofs (a list of references is given at the end of this paper). We want only to give a feeling for the physical concepts involved.

We start with von Neumann's equation for the density matrix

$$i \frac{\partial \rho}{\partial t} = [H, \rho] \quad (2.1)$$

In terms of  $\rho$ , the average value of an observable can be written (in occupation number representation)

$$\langle O \rangle = \sum_{nn'} \langle n | \rho | n' \rangle \langle n' | O | n \rangle \quad (2.2)$$

The time  $t$  as it appears in (2.1) or (2.2) is directly the time of "observables" (as in the left-hand side of 2.1) and not the microscopic time of probability amplitudes. Also it should be noticed that as the density matrix satisfies the linear equation (2.1) and is also linearly related to observables through (2.2), asymptotic procedures become specially simple to handle.

It is convenient to perform the change of variables

$$n - n' = \nu, \quad n + n' = 2N \quad (2.3)$$

and to use the notation

$$\langle n | O | n' \rangle \equiv O_{n-n'} \left( \frac{n+n'}{2} \right) \equiv O_\nu(N) \quad (2.4)$$

In this way Eq. (2.2) now becomes

$$\langle O \rangle = \sum_N \left( O_0 \rho_0 + \sum_\nu O_\nu \rho_\nu \right) \quad (2.5)$$

In a model in which a random phase approximation would be valid, all  $\rho_\nu$  would again vanish. It is therefore appropriate to consider  $\rho_\nu$  as expressing the correlations in the system while  $\rho_0$  refers to the "vacuum of correlations".

Now by simple manipulation (use of Fourier-Laplace transforms or equivalently of projection operators, see References at the end of this report), it is possible to derive exact equations for the evolution of both  $\rho_0$  and  $\rho_\nu$ . The equation for  $\rho_0$  is

$$\frac{\partial \rho_0}{\partial t} = \int_0^t d\tau G(t - \tau) \rho_0(\tau) + \mathcal{D}(t; \rho_\nu(0)) \quad (2.6)$$

In this equation,  $G(t)$  is a generalized collision operator acting on the occupation numbers  $N$  and defined formally in terms of all irreducible "vacuum of correlations to vacuum of correlations" transitions. A fundamental role is played by the Laplace transform  $\psi(z)$  of  $G(t)$ .

The finite duration of the collision is expressed through the non-instantaneous character of Eq. (2.6). The second term in the right-hand side of (2.6) expresses the influence of initial correlations  $\rho_\nu(0)$  on the subsequent evolution of the system. Similarly one obtains evolution equations for the correlations. It is convenient to split  $\rho_\nu(t)$  into two parts

$$\rho_\nu(t) = \rho'_\nu(t) + \rho''_\nu(t) \quad (2.7)$$

The evolution of the first part is given by an equation similar to (2.6) which we shall not write here (see e.g. Balescu (b), 1967). The second part corresponds to "creation of correlations"  $\rho'_\nu$  out of  $\rho_\nu$ , with  $\nu' \leq \nu$

$$\rho''_\nu(t) = \sum_{\nu'} \int_0^t d\tau C_{\nu\nu'}(\tau) \rho_{\nu'}(t - \tau) \quad (2.8)$$

In this theory three types of quantities appear:

- (a) diagonal operators such as the collision operator  $\psi(z)$  leading from  $\rho_0$  to  $\rho_0$
- (b) destruction operators such as appears in the second term in (2.6) leading e.g. from  $\rho_\nu$  to  $\rho_0$
- (c) creation operators as in (2.8) leading e.g. from  $\rho_0$  to  $\rho_\nu$ .

In all cases an irreducibility condition is implied: no intermediate state may be the vacuum of correlations. The explicit expressions of these quantities in terms of the matrix elements involving the Hamiltonian have been given elsewhere (see general references).

This description of mechanics as a dynamics of correlations is an alternative exact formulation of quantum mechanics (it is also applicable to classical mechanics). Equation (2.6) is often called a master equation.

The remarkable feature is that the time evolution of  $\rho_0(t)$  appears in Eq. (2.6) as due to the superposition of two effects: a dynamic evolution due to the processes included in  $G(t)$  or  $\psi(z)$  and an "induced evolution which depends essentially through  $\rho_\nu(0)$  on the preparation of the system.

This feature is really unique: if we consider the dynamic evolution as described by the Schrödinger or Heisenberg equations of motion, the problems of initial conditions and of dynamic evolution appear as entirely separate. However, Eq. (2.6) shows that through suitable reformulation of mechanics the initial conditions can be introduced explicitly into the evolution equations. But then we can begin to ask questions which would be otherwise even difficult to formulate. Suppose that we take two different initial preparations, corresponding to different values of  $\rho_\nu(0)$ . How would this influence the subsequent evolution of the system? What is the dynamic memory of the system? The formal exact solution (1.1) of the equations of motion shows that to two different initial conditions at  $t_0$  will correspond two different states at  $t$ .

But here as we are studying a large system we are interested in the mechanical evolution only as far as macro-observables are concerned. Now for large classes of Hamiltonians it can be verified that the influence of  $\mathcal{D}(t; \rho_\nu(0))$  on macro-observables is vanishing for long times (see general references). In this sense

$$\mathcal{D}(t; \rho_\nu(0)) \rightarrow 0 \quad (2.9)$$

In this sense also, different preparations of the system may give rise to the same mechanical evolution. This property is highly non-trivial. It is likely not to apply to systems interacting through long range forces such as gravitation. In our opinion expression (2.9) is the basic mechanical property on which the whole of thermodynamics rests. I believe that it has to replace the condition of metrical indecomposability of classical ergodic theory which even if satisfied would give relaxation times of a completely wrong order of magnitude.

Using (2.9), formula (2.6) gives a closed equation for  $\rho_0(t)$ . Using the Laplace transform  $\psi(\tilde{\gamma})$  of  $G(t)$  it may be written

$$i \frac{\partial \rho_0}{\partial t} = \psi(i \frac{\partial}{\partial t}) \rho_0(t) \quad (2.10)$$

The lowest approximation to Eq. (2.10) is obtained by neglecting the time dependence in the operator  $\psi$ .<sup>1</sup> One obtains then

$$i \frac{\partial \rho_0}{\partial t} = \psi(0) \rho_0(t) \quad (2.11)$$

This is possible when one has widely separated time scales such that the relaxation time is by far the longest characteristic time involved in the time evolution of the system.

Equation (2.11) corresponds to the Boltzmann approximation of statistical physics (or the so-called Pauli equation).

---

<sup>1</sup> The notation  $\psi$  used here should not be confused with the wave function in (1.1).

A more general equation is obtained when  $\psi(i \partial/\partial t)$  is formally developed in a power series of  $i \partial/\partial t$ . After reordering of the series, Eq. (2.10) becomes

$$i \frac{\partial \rho_0}{\partial t} = \Omega \psi(0) \rho_0(t) \quad (2.12)$$

where  $\Omega$  is a functional of  $\psi$  and its derivatives with respect to  $\zeta$ , for  $\zeta \rightarrow +i0$ . The first terms are

$$\Omega = 1 + \psi'(0) + \frac{1}{2} \psi''(0) \psi(0) + [\psi'(0)]^2 + \dots \quad (2.13)$$

with

$$\psi'(0) = \left( \frac{d\psi}{d\zeta} \right)_{\zeta} \rightarrow +i0 \quad (2.14)$$

Equations such as (2.10) - (2.12) are often called kinetic equations. The equations for the correlations may also be simplified. The part  $\rho_v(t)$  vanishes together with (2.9) and (2.8) becomes

$$\rho_v(t) = \int_0^t d\tau C_v(\tau) \rho_0(t-\tau) \quad (2.15)$$

Expanding  $\rho_0(t-\tau)$  around  $\rho_0(t)$  we may obtain  $\rho_v''(t)$  in terms of the distribution function  $\rho_0(t)$  at the same time. We then obtain an expression of the form

$$\rho_v(t) = C_v \rho_0(t) \quad (2.16)$$

This corresponds exactly to (2.12) in which we also have expressed  $\partial \rho_0/\partial t$  in terms of  $\rho_0(t)$  taken at the same time. Starting from the exact equations (2.6) - (2.8) we have obtained the simplified expressions (2.12), (2.16). Equation (2.12) is a closed equation for  $\rho_0$  while Eq. (2.16) expresses the correlation in terms of  $\rho_0$ . Averages of physical quantities involve contributions of both the vacuum of correlations  $\rho_0$ , and of the correlations. However the correlations do not satisfy a separate equation. Therefore one may ask if it would not be possible to combine both  $\rho_0$  and the correlations  $\rho_v$  into a single type of quantity. This is indeed so as we shall show now.

### 3. INVARIANT MANIFOLD OF THE DYNAMIC EVOLUTION

Let us start with the integral representation of the solution of Eq. (2.6)

$$\rho_0(t) = \frac{1}{2\pi i} \int_C dz \sum_{n=0}^{\infty} \frac{1}{z^{n+1}} e^{-izt} \psi^n(z) \left[ \rho_0(0) + \sum_k \mathcal{D}_k(z) \rho_k(0) \right] \quad (3.1)$$

The contour C is taken in the upper half plane above all the singularities of the quantities  $\psi(z)$ ,  $\mathcal{D}(z)$ ; as we have seen  $\psi(z)$  is the Laplace transform of  $G(t)$ . Similarly  $\mathcal{D}_k(z)$  is the Laplace transform of the "destruction fragment", leading from correlation  $\rho_k(0)$  to a diagonal state (for more details see George (a) (b), 1967). The simplified evolution equation (2.12) is obtained by closing from above the contour C in (3.1) around the point  $z = 0$  excluding all other singularities.

Let us compare then the formal solution of (2.12) with the integral representation (3.1) in which only the pole  $z = 0$  is retained. We obtain in this way (see George, loc. cit. as well as general references, and Eq. (2.12))

$$\rho_0(t) = e^{-i\Omega\psi t} A \left[ \rho_0(0) + \sum_k D_k \rho_k(0) \right] \quad (3.2)$$

with

$$A = \sum_{n=0}^{\infty} \frac{1}{n!} \left( \frac{d^n [\psi(z)]^n}{dz^n} \right)_{z \rightarrow +i0} \quad (3.3)$$

and

$$AD_k = \sum_{n=0}^{\infty} \frac{1}{n!} \left( \frac{d^n}{dz^n} \left[ \psi^n \mathcal{D}_k \right] \right)_{z \rightarrow +i0} \quad (3.4)$$

The important point is that the initial conditions associated with the solution of the kinetic equation as given by (3.2) are related to the exact initial conditions through

$$\begin{cases} \bar{\rho}_0(0) = A \left[ \rho_0(0) + \sum_k D_k \rho_k(0) \right] \\ \bar{\rho}_v(0) = C_v \bar{\rho}_0(0) \end{cases} \quad (3.5)$$

The second relation is a consequence of Eq. (2.16). We see that in conjunction with the use of the simplified evolution equations (2.12) and (2.16) we have to use modified initial conditions to obtain the correct long time behaviour for macro-observables. We shall call  $\bar{\rho}_0(0)$ ,  $\bar{\rho}_v(0)$  the "post-initial" conditions.

Equations (3.5) may also be written using an obvious notation ( $\bar{\rho}'(0)$ ) stays for the set of  $\bar{\rho}_v(0)$ )

$$\begin{pmatrix} \bar{\rho}_0(0) \\ \bar{\rho}'(0) \end{pmatrix} = \begin{vmatrix} A & AD \\ CA & CAD \end{vmatrix} \begin{pmatrix} \rho_0(0) \\ \rho'(0) \end{pmatrix} \quad (3.6)$$

or

$$\bar{\rho}(0) = \Pi \rho(0) \quad (3.7)$$

This operator  $\Pi$  introduced by George ((b) 1967) has a number of most remarkable properties:

(a) First of all we have as a consequence of the group property of the equations of motion

$$\Pi^2 = \Pi \quad (3.8)$$

This property insures that to a given solution of the evolution equation (2.12) corresponds the single post-initial condition (3.5). We see that  $\Pi$  is closely related to a projection operator (however it is not Hermitian (see Mandel, 1968), we shall disregard this difference and call  $\Pi$  a projection operator). There are an infinite number of different initial conditions which give the same post-initial condition. Indeed (3.5) shows that two distributions  $\rho'(0)$  and  $\rho''(0)$  satisfying the single condition

$$A \left[ \rho'_0(0) - \rho''_0(0) + \sum_k D_k [\rho'_k(0) - \rho''_k(0)] \right] = 0 \quad (3.9)$$

give the same post-initial condition.

We understand now much better the mechanism through which in large systems the symmetry between the initial and final states is destroyed. The initial conditions appear in conjunction with the projection operator  $\Pi$ . As a result, the study of the time evolution of macro-observables in the Markoffian approximation does not permit us to decide from which state the system started at the initial time. To make such a decision we would have to study the evolution of the system for very short times (of the order of the duration of a collision) or classes of observables which do not belong to macro-observables.

The projection operator  $\Pi$  should of course not be confused with the projection operators which lead from the original von Neumann equation to the master equation (see Zwanzig, 1961; Balescu (a), 1967) and which are introduced to separate  $\rho_0$  from the  $\rho_p$ 's.

(b) The solution (3.2) of the kinetic equation together with Eq. (2.16) may be written in a form similar to Eq. (3.7) (see George (b), 1967)

$$\rho(t) = \Sigma(t) \rho(0) \quad (3.10)$$

with (see Eqs (3.2), (3.6))

$$\Sigma(t) = \begin{vmatrix} e^{-it\Omega\psi} A & e^{-it\Omega\psi} AD \\ Ce^{-it\Omega\psi} A & Ce^{-it\Omega\psi} AD \end{vmatrix} \quad (3.11)$$

This operator may be called the "complete" kinetic operator as it gives both the evolution of  $\rho_0$  and of the correlations. For  $t = 0$ ,  $\Sigma$  reduces to the operator  $\Pi$ .

Now using the semi-group property of  $\Sigma(t)$  it is easy to show that  $\Pi$  commutes with  $\Sigma$  (note that  $\Pi = \Sigma(t) t \rightarrow +0$ ) (3.12)

$$\Pi\Sigma = \Sigma\Pi \quad (3.13)$$

Therefore

$$\Pi\rho(t) = \rho(t) \quad (3.14)$$

where  $\rho(t)$  is an arbitrary solution (3.10) of the kinetic equations. The projection operator  $\Pi$  defines therefore an invariant manifold.

So far as we are interested in the evolution of macro-observables we have only to deal with the projection of the distribution function on the manifold  $\Pi$ . This projection satisfies a closed equation of motion which gives both the diagonal elements and the off-diagonal elements of the density matrix.

(c) It is of course essential to discuss the manifold associated with the projection operator  $\Pi$ . It is defined by

$$\Pi\phi = \phi \quad (3.15)$$

It contains the class of matrices such that (see Eq. (3.5)) their diagonal and off-diagonal elements satisfy the relations

$$\begin{aligned} \varphi_0 &= A \left[ \varphi_0 + \sum_k D_k \varphi_k \right] \\ \varphi_\nu &= C_\nu \varphi_0 \end{aligned} \quad (3.16)$$

The matrix  $\phi$  may represent a density matrix or the matrix associated to another physical quantity such as the energy (see Eq. (3.17)).

It may be verified that the ground state of the system that is the lowest eigenfunction of the Schrödinger equation gives rise to a density matrix which satisfies Eq. (3.16) (see Turner, 1968 and Henin et al., Physica, 1968). More generally the matrix associated to an arbitrary function of the total energy also satisfies Eq. (3.16).

In all these cases we find a very special form of the density matrix such that the off-diagonal elements may be expressed in terms of the diagonal ones through Eq. (3.16).

(d) Instead of considering the effect of  $\Pi$  on density matrices it is of course equivalent to study its effect on operators representing physical quantities (see George (b), 1967; Mandel, 1968).

One may show that

$$\Pi H = H \quad (3.17)$$

or more explicitly if  $H = H_0 + V$  where  $V$  is off-diagonal in the representation in which  $H_0$  is diagonal

$$H_0 = A[H_0 + DV] = H_0$$

$$V = CH_0$$

The relation between the off-diagonal part  $V$  and  $H_0$  is the same as between correlations and the vacuum of correlations (see Eq. (2.16)). As a result we may say that potential energy and correlations play the same dynamic role.

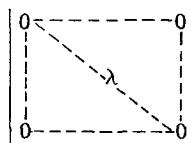
The projection on the manifold  $\Pi$  does not alter the energy. Inversely, the part of the trajectory which is "outside the manifold" does not carry any energy

$$\text{tr } H \rho = \text{tr } H \Pi \rho \quad (3.17')$$

(e) Let us now consider a wave packet  $\psi$  or the "approximate" wave function corresponding to an excited state. The corresponding density matrix  $\rho_\psi$  does not belong to the eigenmanifold of  $\Pi$

$$\Pi \rho_\psi \neq \rho_\psi \quad (3.18)$$

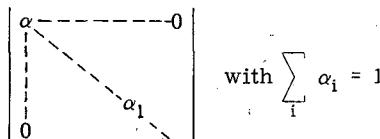
This can be easily verified on simple examples (Henin, 1968). Moreover,  $\Pi \rho_\psi$  cannot be reduced through a canonical transformation to the characteristic form of the density matrix corresponding to pure states



$$\begin{vmatrix} 0 & \lambda \\ \lambda & 0 \end{vmatrix} \quad (3.19)$$

except if  $\psi$  is the eigenfunction of the ground state.

For other wave functions (or wave packets) we can only achieve for  $\Pi \rho_\psi$  the diagonal form



$$\begin{vmatrix} \alpha & \lambda \\ \lambda & \alpha_1 \end{vmatrix} \quad \text{with } \sum_i \alpha_i = 1 \quad (3.20)$$

with more than one element  $\alpha_k$  different from zero. In order to avoid misunderstandings let us notice that the word ground-state is used here in contrast with "excited states". It corresponds to the whole manifold of states with an infinite lifetime (such as an atom in the lowest electronic state with an arbitrary centre of mass motion).

Let us summarize this discussion: the Markoffian evolution of macro-observables is described by density matrices which belong to the invariant and energy carrying manifold  $\Pi$ . The ground-state of the system, as well as the canonical distribution, lie on this manifold. Each trajectory in this manifold is the coalescence of an infinite number of trajectories which differ by the exact initial conditions but correspond all to the same energy. No trajectory can be characterized by a wave function except in the case when the system is in the ground-state. The concept of a wave function loses its meaning as no measure of the macro-observables may determine it. Such measurements lead only to the determination of the state of the system in the

manifold  $\Pi$ . This state is in general represented by a statistical ensemble irreducible to a pure state.

We may also consider our results from a different point of view. As each trajectory in  $\Pi$  is in reality a "tube" corresponding to the confluence of "exact" mechanical trajectories, we may also consider it as a kind of "coarse grained trajectory".

It is important however to stress the difference between the coarse graining in the sense used here and the usual concept of coarse graining as related to the introduction of finite cells in phase space given a priori (see e.g. von Neumann, 1932) or averages over time intervals (according to Kirkwood, see e.g. Rice and Gray, 1965). In our approach, the coarse graining is a consequence of the exact equations of dynamics when applied to the evolution of macro-observables for well defined classes of mechanical systems. As it is given by the projection on the eigenmanifold of  $\Pi$ , its effect depends moreover on the nature of the state. No coarse graining of the ground-state is introduced as this state belongs already to  $\Pi$ .

In a non-dissipative system such as a system with a finite number of degrees of freedom the physical states are the eigenstates of the Hamiltonian. In other words, we have to "decompose" so to speak the von Neumann space into a product of two Hilbert spaces corresponding each to a wave function. The physical states correspond to special choices of the wave function (the eigenfunctions of the Hamiltonian). For each physical state the corresponding density matrix takes the canonical form (3.19).

It is clear that the situation here is deeply different. As far as the macro-observables in the Markoffian approximation are concerned, the physical states lie on the eigenmanifold of  $\Pi$ . The complementary part of this functional space is of no interest to us. Therefore it is in  $\Pi$  that we have to identify the physical states of the system. It is most remarkable that this is at all possible as we shall show now.

#### 4. QUANTIZATION AND PHYSICAL STATES ON THE INVARIANT MANIFOLD

To describe the evolution of a mechanical system in terms of the density matrix through the von Neumann equation (2.1) we need both the diagonal elements  $\rho_0$  and the off-diagonal elements  $\rho_{\nu}$ . However there exists a simple case where  $\rho_0$  alone is sufficient: that is the case of the "Pauli equation" (or the Boltzmann approximation) which corresponds to the form (2.11) of the kinetic equation. Correlations may then be neglected and the equilibrium distribution for long times becomes

$$\rho_0 \rightarrow \exp \left( - \frac{H_0}{kT} \right) \quad (4.1)$$

where  $H_0$  is the unperturbed part of the Hamiltonian. In this case the entropy (or the  $\mathcal{H}$ -quantity) is given both for equilibrium and out of equilibrium by the Boltzmann functional

$$\mathcal{H} = \sum_N \rho_0(N) \log \rho_0(N) \quad (4.2)$$

The evolution is then no more described in terms of mechanical concepts such as interactions or correlations, but in terms of energy conserving collisions.

When we go now to the higher order kinetic equation (2.12) we have also to take into account the correlations through (2.16) (which give contributions to macro-observables of the same order as the correction terms to the Boltzmann approximation in the kinetic equation, see general references). As a result, a description of the evolution in terms of energy conserving collisions becomes impossible. Part of the energy available in collisions is used to build up correlations. The analogy with field theoretical problems is obvious. There also, particles appear in a dual way: both as real entities taking part in collisions and as virtual entities transmitting forces. Here the corresponding problem is the elimination of correlations. This contrasts with our first step which was to disentangle the role of  $\rho_0$  and of the correlations  $\rho_\nu$  in the dynamic evolution. In general the elimination of correlations would be an impossible task. However on the  $\Pi$  manifold correlations appear only in the simple form (2.12). This makes it possible to eliminate them. Let us consider average values

$$\begin{aligned} \langle O \rangle &= \sum_N (O_0 + \sum_\nu O_\nu C_\nu) \rho_0 \\ &= \sum_N (O_0 + \sum_\nu D_\nu O_\nu) \rho_0 \end{aligned} \quad (4.3)$$

where  $D_\nu$  is the Hermitian adjoint operator of  $C_\nu$  (see Mandel, 1968). Moreover if we introduce the new distribution function

$$\tilde{\rho} = \chi^{-1} \rho_0 \quad (4.4)$$

where  $\chi^{-1}$  is a time-independent operator acting on occupation numbers, we may write ( $\chi^\dagger$  is the adjoint of  $\chi$ )

$$\begin{aligned} \langle O \rangle &= \sum_N \chi^\dagger \left( O_0 + \sum_\nu D_\nu O_\nu \right) \tilde{\rho} \\ &= \sum_N O_R \tilde{\rho} \end{aligned} \quad (4.5)$$

In (4.5) the density matrix contains only diagonal elements as well as the matrix  $O_R$  corresponding to the observable  $O$ . All off-diagonal elements have been eliminated. But this elimination can still be achieved in an infinite variety of ways. We have now to introduce rules to make this diagonalization procedure unique. The evolution towards equilibrium has now to proceed only through energy conserving collisions and we have to expect that, exactly as for weakly coupled systems or in the Boltzmann approximation, we shall have (see 4.1)

$$\tilde{\rho} \rightarrow \exp \left[ -\frac{H_R}{kT} \right] \quad (4.6)$$

where  $H_R$  is now a diagonal quantity corresponding to a redefined energy, as in (4.5). Now the kinetic equation (2.12) is itself modified by the transformation (4.4) and takes the form

$$i \frac{\partial \tilde{\rho}}{\partial t} = \varphi \tilde{\rho} \quad (4.7)$$

where  $\varphi$  is the new collision operator. We have therefore to require that

$$\varphi_e - \frac{H_R}{kT} = 0 \quad (4.8)$$

Now both  $\varphi$  and  $H_R$  are functionals of the operator  $X$  appearing in (4.4). Inversely this equation may be used to construct the operator  $X$ . This has been done for various problems in the frame of perturbation theory (see Henin and De Haan, 1968; George, 1968). A single operator satisfying (4.8) has been found. We shall not write its explicit form here but only discuss a few qualitative aspects of this theory.

An equivalent way to present condition (4.8) is to require that there exists a  $\tilde{\rho}$  such that the  $\mathcal{H}$ -quantity of Boltzmann now becomes

$$\mathcal{H} = \sum_N \tilde{\rho}(N) \log \tilde{\rho}(N) \quad (4.9)$$

For this reason we have called our theory an "entropy transformation theory" (Prigogine, Henin and George, 1968).

The transformation  $X$  leads to a kind of canonical representation of all density matrices which lie in the  $\Pi$  manifold and satisfy therefore the conditions (3.16). Instead of using separately  $\rho_0$  and the off-diagonal elements  $\rho_{ij}$  we have now to use  $\tilde{\rho}$ . The relation between  $\tilde{\rho}$  and  $\rho_0$ ,  $\rho_{ij}$  may be written (see (4.5); for the derivation see George (b), 1967)

$$\tilde{\rho} = X^\dagger (\rho_0 + D\rho') \quad (4.10)$$

Therefore  $\tilde{\rho}$  is a kind of weighted average of the diagonal elements  $\rho_{ii}$  and the correlations  $\rho_{ij}$  just as in (4.5). In this sense our description appears as a generalization of the well-known random phase approximation.

Inversely all redefined mechanical quantities may be written exactly as in (4.4)

$$O_R = X^{-1} O_0 \quad (4.11)$$

For example

$$H_R = X^{-1} H_0 \quad (4.12)$$

Our transformation theory leads therefore to a commutative algebra of macro-observables.

We believe that what has been achieved in this way is a consistent particle description of the evolution of many body systems in the Markoffian approximation. The physical states of the system are now described by  $\tilde{\rho}(N_1 \dots N_k \dots t)$  exactly as for weakly coupled systems. The evolution of  $\tilde{\rho}$  is entirely due to energy conserving collisions as they appear in the kinetic equation (4.7). The energy of the system is at any moment the sum of the energies of the particles as

$$\langle H \rangle = \text{tr} \rho (H_0 + V) = \sum_N \tilde{\rho} H_R$$

Let us make a certain number of supplementary comments:

- (a) As the ground state lies in the eigenmanifold of  $\Pi$ , the transformation  $X^{-1}$  should lead to the canonical representation (3.19) of a density matrix for a pure state; the density matrix should then admit the usual factorization

$$\rho = |\psi\rangle\langle\psi| \quad (4.13)$$

where  $\psi$  is the (exact) ground-state function. Similarly  $H_R$  as given by (4.12) should reduce to the corresponding lowest eigenvalue of the Hamiltonian. This can indeed be verified again in the frame of perturbation theory (Henin). Whenever dissipative effects may be neglected (all  $\delta$  functions give vanishing contributions and principal parts can be replaced by their arguments), we have (Henin (b), 1968)

$$H_R = S^{-1}HS \quad (4.14)$$

where  $S$  is an appropriate unitary operator. Our transformation theory leads therefore to a natural generalization of the usual quantum transformation theory applicable to all density matrices in the  $\Pi$  manifold, of which the ground-state is a member. However in general the  $X$  transformation cannot be reduced to (4.14).

It is clear that  $H_R$  cannot be in general the eigenvalue of some Hamiltonian operator. If it would be, the corresponding eigenfunctions would be invariants of motion and the system would not evolve to equilibrium.

- (b) As the equilibrium distribution is also in the  $\Pi$  manifold our method may be applied to the study of equilibrium properties. The partition function of the system

$$Z = \text{tr} e^{-\frac{H}{kT}} \quad (4.15)$$

may be of course evaluated without diagonalizing  $H$ . However using (4.12) we obtain simply

$$Z = \sum_N e^{-\frac{H_R^N}{kT}} \quad (4.16)$$

Again we have verified in the frame of perturbation theory that (4.16) gives results identical to (4.15).

(c) As we have noticed the physical states of the system are now described by the density matrix  $\tilde{\rho}(N_1, N_2 \dots, t)$ . We may transform this state back to the initial representation using (4.4) and (2.16). The corresponding density matrix cannot be written in the form (3.19) and does not correspond to a wave function. As we have already emphasized in Section 3 the wave function concept is lost.

(d) The relation between this method and the Green's formalism has been treated briefly elsewhere (Prigogine, 1968). We want only to mention here that the level shift due to the usual pole on the second Riemann sheet has no simple meaning (beyond lowest order in perturbation theory) when dissipative processes occur. Its real part cannot be used to calculate partition functions as in (4.16). Moreover this method does not lead to any indication of what the meaning of quantum state corresponding to finite life time might be.

## 5. CONCLUSIONS AND APPLICATIONS

The main point in this presentation is the identification of the invariant manifold associated to  $\Pi$  (see Section 3). The class of matrices belonging to this manifold can be simultaneously put into a diagonal form through the transformation (4.4). In this form the evolution may be described in terms of changes of occupation numbers through energy conserving collisions. All virtual effects are eliminated.

From the point of view of thermodynamics we have now a microscopic model of entropy (see 4.9) which satisfies all the necessary requirements ( $\mathcal{H}$ -theorem, as well as correct value of entropy at equilibrium). Moreover we may verify that it leads in the neighbourhood of equilibrium to the basic expressions of thermodynamics of irreversible processes which till now had been postulated by the phenomenologic theory (see Velarde and Wallenborn, 1968).

It is interesting to notice that our theory leads to a synthesis between the points of view of Boltzmann and Gibbs about the statistical formalism of the second law. In agreement with Boltzmann we have now a dynamic derivation of the  $\mathcal{H}$ -theorem but which uses a "coarse grained" concept of mechanical trajectories in the sense discussed in Section 3.

As a consequence of the elimination of all virtual processes, the concept of a particle takes a simple physical meaning. The particle description appears as a non-mechanical (coarse grained) reduction of the initial mechanical description in terms of interacting fields. It corresponds precisely to a simplified language in the sense used in the introduction. Such a reduction is not always possible: if after some time we inverse the velocities we go over to a new ensemble for which property (2.9) no more holds for microscopic times (see Balescu (a), 1967). For the inversed motion no elimination of virtual processes appears possible (at least in the frame of our theory). We see that both the validity of thermodynamic and of a consistent particle description depend on the preparation of the system.

From the point of view of statistical mechanics we may now calculate the excitation spectrum in dissipative systems (see, for example, Henin et al., 1966; Mandel, 1968). Our theory appears from this point of view as a natural generalization of Landau's theory of Fermi liquids.

In atomic physics it gives a new and simple approach to problems such as sequential emission and many-photon transitions (see Henin (b), 1968).

In connection with analytic S-matrix theory, it permits to incorporate unstable particles in the unitarity relations and to calculate the corresponding spectral functions (see Mayné, 1968).

Are the particles as observed in nature related to the particles introduced here? We would like to believe so, precisely because the entities we have introduced carry the full energy of the system.

For all these reasons we are convinced that the theory as it stands is at least consistent. We have now to explore the consequences of our new definition of quantum states (for example, energy splitting due to different life times, symmetry breaking mechanisms) as well as to extend it to strong interactions. Some results have already been obtained and we hope to report on these problems in a near future.

#### ACKNOWLEDGEMENTS

The results presented in this report have been obtained during the last two years by our group in Brussels. A specially important part has been played by Drs. Cl. George, F. Henin and P. Résibois, but I am also indebted to Messrs. Mayné, Mandel, Turner for calculations and discussions.

This work has been sponsored by the following institutions to which I wish to address my appreciation for continuous support: Fonds national belge de la recherche fondamentale collective; Dupont de Nemours International (Geneva); The Air Force Office of Scientific Research through the European Office of Aerospace Research, OAR, United States Air Force under contract AF EOAR 67-25.

#### REFERENCES

- BALESCU, R., (a) *Physica* 36 (1967) 433.
- BALESCU, R., (b) *Physica* 38 (1967) 123.
- GEORGE, CL., (a) *Bull. Acad. r. Belg. Cl. Sci.* 53 (1967) 623.
- GEORGE, CL., (b) *Physica* 37 (1967) 182.
- GEORGE, CL., (1968).
- HENIN, F., DE HAAN, M., *Physica* (1968).
- HENIN, F., (a) *Bull. Acad. r. Belg. Cl. Sci.* (1968).
- HENIN, F., PRIGOGINE, I., GEORGE, CL., MAYNE, F., *Physica* 32 (1966) 1828.
- HENIN, F., *Physica* (to be published).
- MANDEL, P., *Bull. Acad. r. Belg. Cl. Sci.* 54 (1968).
- MAYNE, F., (1968).
- TURNER, J. W., private communication (1968).
- PRIGOGINE, I., RESIBOIS, P., *Estratto degli Atti dell Simposio Lagrangiano, Acad. Delle Scienze di Torino* (1964).
- PRIGOGINE, I., Symposium on Stochastic Processes, La Jolla, March 1968, to be published in *Adv. chem. Phys.*
- PRIGOGINE, I., HENIN, F., GEORGE, CL., *Proc. nat. Acad. Sci. U.S.A.* 59 (1968) 7.
- RICE, S., GRAY, P., *Statistical Mechanics of Simple Liquids*, Wiley-Interscience, New York (1965).
- VELARDE, M. G., WALLENBORN, J., *Physics Lett.* 26A (1968) 584.
- von NEUMANN, J., *Mathematische Grundlagen der Quantenmechanik*, Berlin, Springer (1932).
- ZWANZIG, R. W., in *Lectures in Theoretical Physics III*, Boulder, Summer Institute, Interscience, New York (1961).

In addition, I should like to quote the following general references:

- PRIGOGINE, I., Non Equilibrium Statistical Mechanics, Wiley-Interscience, New York (1962).  
BALESCU, R., Statistical Mechanics of Charged Particles, Wiley-Interscience, New York (1963).  
RESIBOIS, P., Physics of Many-particle Systems, (MEERON, E., Ed.), Gordon and Breach, New York (1966).  
PRIGOGINE, I., Introduction to Non Equilibrium Statistical Physics, in: Nato School on Non Linear Physics  
and Mathematics, Munich, 1966, Springer (1968).

Also various aspects of this work have been presented at the following conferences:

- PRIGOGINE, I., HENIN, F., "Kinetic equations, quasiparticles and entropy", Statistical Mechanics (Proc.  
IUPAP Meeting, Copenhagen, 1966), (BAK, Th. A., Ed.), Benjamin, New York, Amsterdam (1967).  
PRIGOGINE, I., "Quantum theory of dissipative systems and scattering processes", Nobel Symposium V, 1967.  
PRIGOGINE, I., Dissipative processes, quantum states and field theory, XIV<sup>e</sup> Conseil de Physique Solvay,  
Bruxelles, 1967, Wiley-Interscience, New York (in press).



## **ASTROPHYSICS, QUASARS AND PULSARS**



# INTRODUCTION AND TOPICS IN THEORETICAL ASTRONOMY\*

E.E. SALPETER

Laboratory of Nuclear Studies, Physics Department,  
Center for Radiophysics and Space Research,  
Cornell University,  
Ithaca, N.Y., United States of America

## Abstract

INTRODUCTION AND TOPICS IN THEORETICAL ASTRONOMY. Some underlying reasons are discussed for the complexity of phenomena encountered in astronomy. The importance of the very small value of the "gravitational fine-structure constant" (and a few other dimensionless constants) is stressed. A few examples are given of complex problems, e.g. (1) rare phenomena like neutrino pair-emission in stellar interiors; (2) highly non-equilibrium circumstances like the formation and destruction of hydrogen molecules in the interstellar gas; (3) phenomena at very high densities such as (a) the freezing of positive ions into a Coulomb lattice and (b) the formation of higher baryon states in neutron matter at densities beyond nuclear density.

## INTRODUCTION

I should perhaps introduce the most exciting and important topics in astrophysics, but I will not do so; other authors here have chosen these topics. I will discuss a few subjects from a different point of view, namely topics which present a challenge to the techniques of theoretical physics and which occur because of most unusual conditions encountered in astronomy. First I want to summarize what is so unusual about astronomical conditions and why.

Perhaps the main point I want to make is that one often encounters much more complex situations in astronomy than in laboratory physics. In part this is due to the fact that one can only make observations in astronomy instead of controlled experiments, but in part it is an intrinsic feature. A fairly large portion of the astronomy lectures are given here by experimentalists in these Hallowed Halls of Theory. Because of the complexity and interlocking phenomena it is dangerous to abstract a theoretical problem in astronomy at too early a stage and it is important to keep an eye on the observational data continuously.

The unusual conditions in astronomy typically have to do with some extreme numerical values. One of these is the "astronomically large" timescale in astronomy. These long times enable rare phenomena to become important. Another class of extreme values has to do with very high and very low values of both temperature and density. In some kinds of stars temperature and density are certainly higher than can be achieved terrestrially and low densities can be found in interstellar space. The lowest temperatures found in space (3 or 4°K) are not spectacular at all

---

\* Supported in part under NSF Grant GP-6928 and by the Office of Naval Research.

for a low-temperature physicist, but we shall see that the contrast between high and relatively low temperatures leads to complex phenomena.

I will give a few examples which illustrate the effect of these unusual conditions. First, I want to philosophize a little about the reasons for these conditions. More specifically, I want to talk about "magic numbers", the occurrence of a few dimensionless numbers of large magnitude in astronomy.

### MAGIC NUMBERS IN ASTRONOMY [1, 2]

Since most of the audience has a quantum-mechanical background, I would like to express most quantities in the "natural units" of quantum electrodynamics in terms of electron mass  $m$ , velocity of light  $c$  and rationalized Planck's constant  $\hbar$ . Besides the electron's Compton wavelength  $(\hbar/mc)$  and restmass-energy  $mc^2$  for length and energy, our time unit is  $(\hbar/mc^2) \sim 10^{-21}$  s and temperature unit is  $(mc^2/k) \sim 6 \times 10^9$  °K. For density I shall use a slightly mixed unit  $\rho_0 \equiv H/(\hbar mc)^3 \sim 10^6$  gm/cm<sup>3</sup> where  $H = 1837$  m is the mass of a hydrogen atom.

In quantum electrodynamics we are all familiar with the importance of the Sommerfeld fine-structure constant  $\alpha^{-1} \equiv \hbar c/e^2 \sim 137$  as a dimensionless coupling for the Coulomb force. An equivalent "gravitational fine-structure constant" can be defined as

$$\alpha_G^{-1} \equiv \hbar c/GH^2 \sim 10^{38}$$

(some authors prefer a definition involving  $m^2$  instead of  $H^2$  in which case the exponent is about 44 instead of 38). Professor Schwinger [3] derives a small quantum-mechanical correction term to Newtonian gravitation in which  $\alpha_G$  occurs very naturally. I will mainly talk about classical gravitation theory and you may be surprised at the introduction of a quantum-mechanical quantity like  $\alpha_G$ . However, we will have to compare gravitational potential energy with quantum-mechanical Fermi energy (or Heisenberg uncertainty energy) and  $\alpha_G$  enters in this comparison. To anticipate my punch line, many of the peculiarities of astronomy stem from the extreme smallness of  $\alpha_G$ .

It is useful to define a few further quantities involving  $\alpha_G$ , in particular

$$N_0 \equiv \alpha_G^{-3/2}; \quad N_{pl} \equiv \alpha^{3/2} N_0 = \left( \frac{e^2}{GH^2} \right)^{3/2}$$

$$t_0 \equiv \alpha_G^{-1} (\hbar/mc^2) \sim 10^{10} \text{ years}$$

$N_0 H$  is the natural unit of mass for stars. Crudely speaking,  $N_0$  is the number of hydrogen atoms one has to bring together "to make up for the smallness of  $G$ ", i.e. for the gravitational attraction to balance the electron's Fermi pressure (near unit density  $\rho_0$ ). Indeed, common stars have masses within a factor of 100 of  $N_0 H$ . More important, the "Chandrasekhar [4] limiting mass"  $M_{ch}$  is close to  $N_0 H$ . In turn,  $M_{ch}$  is important as the mass below which a star has a state with finite maximum binding energy and maximum density (a final zero-temperature state); for masses exceeding  $M_{ch}$  the temperature and density can both increase indefinitely (or a gravitational

collapse beyond the Schwarzschild singularity may result).  $N_{pl} H$ , on the other hand, is the natural unit of mass for planets (it is close to the mass of Jupiter) or rather the dividing line between planets and stars. We shall restrict ourselves to masses large compared with  $N_{pl} H$ , which ensures that solid-state forces are not a dominant feature.

For masses  $M$  exceeding  $M_{ch}$  we have said that density and temperature can become infinite. For  $M < M_{ch}$  there are finite upper limits but, as long as  $M \gg N_{pl} H$ , the maximum density is large compared with  $\rho_{pl} \equiv \alpha^3 \rho_0$  (the density of zero-pressure solids). It is not surprising then that stars (masses of order  $N_0 H$ ) can exhibit high densities and temperatures.

How about typical timescales in stars? Consider first the time  $t_{ph}$  for the black-body radiation in a star's interior to penetrate. As a crude estimate we write  $t_{ph} \sim (R/c) \tau$  where  $\tau$  is the "optical depth" of the star, the average number of steps between scatterings (or absorption and emission) in the random walk path of a photon from the interior to the surface. To an even cruder order of magnitude we consider unit density  $\rho_0$ , mass  $N_0 H$  and replace photon cross-sections by unity (which involves equating factors like  $\alpha^2$  or  $\alpha^{-1}$  to unity!). In our units thus

$$R \sim N^{1/3} \sim \alpha_G^{-1/2}, \quad \tau \sim \frac{N}{R^2} \sim \alpha_G^{-1/2}, \quad t_{ph} \sim \alpha_G^{-1} \equiv t_0$$

An important feature of stellar evolution is then the enormously large value of the optical depth  $\tau$  of a star and the large value of  $t_{ph}$ . Since most stars lose energy to the outside world only through electromagnetic radiation,  $t_{ph} \sim t_0$  also gives the order of magnitude of the relaxation time for changing the star's total binding energy.

Another timescale of interest for a star is the dynamic timescale  $t_{dyn}$ . As regards order of magnitude, the time taken for a sound-wave to cross the star, the free-fall time and the period of oscillation (for a low normal mode) of the whole star are all the same. This timescale is proportional to  $(G\rho)^{-1/2}$ , so that  $t_{dyn} \sim \alpha_G^{-1/2} (\rho/\rho_0)^{-1/2}$  in our units. For a star  $(\rho/\rho_0)$  is never enormously small, so that  $t_{dyn} \ll t_{ph}$ . To summarize the "definition of a star" in the most elementary terms: it has (i) a mass of order  $N_0 H$  so that densities and temperatures can become high and the time for energy relaxation is long, and (ii) a high enough density for dynamic timescales to be short, so that it is relaxed to an equilibrium shape.

In interstellar space one finds gas clouds which can have masses similar to those of stars, but have a low density, irregular shapes rather than equilibrium shapes and are far from thermal equilibrium. These phenomena are not directly related to cosmology but are made possible by some order of magnitude relations for the "visible universe". We again consider factors like  $10^3$  as "of order unity" and equate the "radius of the visible universe"  $R_u$  to the Hubble distance, a distance at which the cosmological red shift becomes appreciable. With  $N_u$ , the number of nucleons in a sphere of radius  $R_u$ , one finds two observational relations,

$$R_u \sim \alpha_G^{-1} \text{ (i.e. } t_u \equiv R_u/c \sim t_0\text{)}, \quad N_u \sim \alpha_G^{-2}$$

For our present purpose it does not matter whether these relations are purely numerical coincidence or have a deeper theoretical foundation (as many people claim at least for the relation  $GM_u/R_u c^2 = \alpha_G N_u / R_u \sim 1$ );

I merely want to mention some equalities and inequalities which follow. In particular, we have

$$\rho_u \sim \frac{N_u}{R_u^3} \sim \alpha_G \ll 1, \quad t_{dyn} \sim (\alpha_G \rho_u)^{-1/2} \sim \alpha_G^{-1} = t_0$$

$$\tau_u \sim \frac{N_u}{R_u^2} \sim 1, \quad t_{ph} \sim R_u / \tau_u \sim \alpha_G^{-1} = t_0$$

From our present point of view a main feature of the universe is its extremely low average density, which makes it optically thin (the optical depth to photons  $\tau_u$  is actually closer to 0.01 than to unity). Consequently the "age of the universe"  $R_u/c$ , the photon diffusion time  $t_{ph}$  and the dynamic timescale  $t_{dyn}$  are all of the same large order of magnitude, namely  $t_0$ . This feature is one reason why parts of the universe can be so far from equilibrium, but we also need another observational fact, namely that the universe is expanding and its mean temperature very low so that "the night-sky is black". The universe thus acts as a heat-engine with the black night-sky providing the heat-sink and the thermonuclear full of stars feeding the energy source. The low average density of the universe plus the fact that gravitation is an attractive force also makes for a less familiar kind of "entropy engine". We are dealing with a hierarchy of average densities (from universe to clusters of galaxies to galaxies to gas clouds to star clusters to stars) and the gravitational contraction of sub-structure of the universe (contrasted with the cosmological expansion of the universe as a whole) releases energy and makes for complexity.

#### RARE EVENTS IN STARS

To illustrate the possible importance of rare phenomena in the interior of stars, let me start with a purely hypothetical example, but one for which the physics would be simple. Imagine the possibility of longitudinal photons with a finite but very small restmass  $m_1$ . For photon energies much larger than  $m_1 c^2$  these hypothetical photons would have a very weak but non-zero coupling with electrons and nuclei. Let us then ask the question how low an upper limit one can put on  $m_1$  from the fact that no evidence for such particles has appeared from solar astronomy.

One effect such weakly coupled longitudinal photons would have on the sun would be an additional luminosity  $L_{long}$  in the form of such photons, which could not be easily detected on the earth but would nevertheless contribute to the sun's rate of energy loss. Some uncertainties remain in our knowledge of the past evolution of the sun. Nevertheless, the sun could not have an additional invisible energy loss as large as or larger than the visible luminosity or it would already be a red giant star (but a contribution of 10%, say, could not be ruled out). Since the upper limit on the luminosity of longitudinal photons is not much smaller than the luminosity due to ordinary photons, you might expect that the upper limit on their effective coupling strength would also not be much lower than that for ordinary photons. The real situation is drastically different, primarily because of the extremely small value of  $\alpha_G$ .

For ordinary photons we have seen that the optical depth  $\tau$  of the sun is very large ( $\sim \alpha_G^{-1/2}$ ). The bottle-neck for their escape is not their production rate in the sun's interior (they are in thermal equilibrium there anywhere) but their random-walk diffusion path due to the many scatterings. As a consequence, if one decreased the coupling of ordinary photons with matter, the luminosity of the sun would not decrease but increase! As a function of the coupling, the luminosity would have an enormously large maximum when the optical depth was about unity and a further decrease in coupling would start to decrease the luminosity again. In this way one could in principle have a mechanism with extremely weak coupling which nevertheless gives an energy drain from the sun equal to that of ordinary electromagnetic radiation. Since the effective coupling of longitudinal photons with electrons and ions depends on their restmass  $m_1$ , I estimate an upper limit to  $m_1$  of something like  $10^{-18}$  times the electron mass! Note that this corresponds to a lower limit on the Compton wavelength and indirectly on the range of the Coulomb force of more than  $10^6$  km.

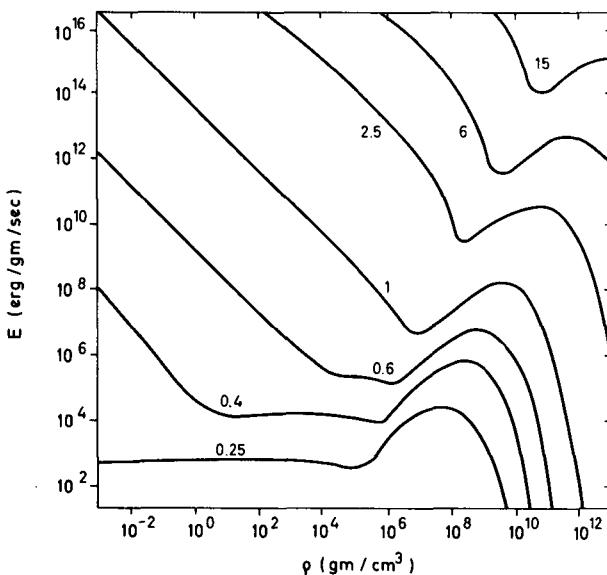


FIG. 1. The total energy loss-rate  $E$  due to the pair, plasma and photo-neutrino process as a function of the density  $\rho$ . The curves are labelled with the value of the parameter  $\lambda = kT/mc^2$ .

Of course, I did not mean to take longitudinal photons too seriously but mentioned them partly to illustrate how even very weak phenomena can add up over the long evolutionary timescales of stars. In part they also serve as an introduction to another weak phenomenon, connected with the universal Fermi interaction, which is not important in the sun but is thought to be important in hotter, more evolved stars - the production of  $\nu\bar{\nu}$  neutrino pairs.

In our units, the effective dimensionless coupling constant of beta-decay is very roughly of the order of  $\alpha_G^{1/2}$ . This particular numerical

coincidence is unimportant for our present purpose but it is important that this coupling is small compared with  $\alpha$  (slow compared with electromagnetic interactions) and large compared with  $\alpha_G$  (fast on an evolutionary timescale). The relevance of ordinary beta-decay is known, but the existence of a direct coupling of the form  $(e\nu) - (e\nu)$  has a different consequence in stellar interiors. Any electromagnetic interaction deep inside a star could be accompanied by an inelastic process in which a neutrino-anti-neutrino pair is emitted [5, 6, 7] (instead of a photon). If this coupling exists with universal strength and there is no other modification of the weak interactions (such as a neutral lepton current [8]), the energy loss-rates as a function of density and temperature are as shown in Fig. 1. Such processes can be quite important in highly evolved stars (e. g. central stars of planetary nebulae, pre-supernovae, etc.). This then is one place where modern particle physics has a direct impact on astrophysics.

#### MOLECULAR HYDROGEN IN INTERSTELLAR SPACE

The major constituent of the interstellar gas is neutral hydrogen, so the question arises what fraction of it is in the form of molecular hydrogen? This is of some (but not overriding) importance in its own right, but I will use it mainly to illustrate how far from equilibrium interstellar space is. If I told you that the temperature of the interstellar gas is about 100°K, you might be inclined to substitute into the well-known formula in statistical mechanics for the equilibrium between atoms and diatomic molecules. If you did that, you would conclude that only about one part in  $10^{100}$  of hydrogen is atomic and the rest molecular. Nothing could be further from the truth!

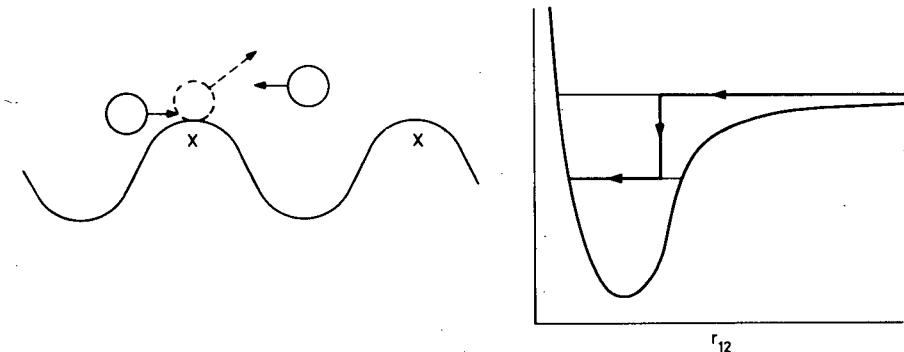


FIG. 2. A schematic picture for two hydrogen atoms approaching each other on the surface of a dust-grain. After they have speeded up due to their mutual attraction, one of the hydrogen atoms collides with a stationary surface molecule; the pair of hydrogen atoms is now vibrationally bound and has some rotational and some centre-of-mass kinetic energy.

The peculiarities of molecular hydrogen in interstellar space is partly due to the fact that the radiative recombination of two hydrogen atoms in their ground-state is very highly forbidden. Since the fraction of H-atoms in excited states is extremely low, radiative recombination is not effective in forming hydrogen molecules in interstellar space. You might think of three-body recombination as the only alternative for making  $H_2$ . The rate

for such three-body recombinations is very density-dependent and there is quite a lot of density variation in the interstellar gas; nevertheless, even the densest gas clouds are too tenuous to give an appreciable rate. However, there is even more variety in interstellar space, including small solid dust-grains, and surface recombination on such grains does work [9-12]. Hydrogen atoms hitting a grain surface are adsorbed and, if the grain temperature is in the right regime, stay long enough to find a partner; the recombination then proceeds as shown schematically in Fig. 2 with a grain-surface atom acting as the third body. The mean time for a hydrogen atom to recombine to  $H_2$  is now estimated to be quite short (by a factor of more than 10) compared with the lifetime of our galaxy. Since the statistical mechanics equilibrium at the gas temperature strongly favours molecules and since the relaxation time is short, you might now really expect the concentration of atomic hydrogen to be negligible. That again is wrong!

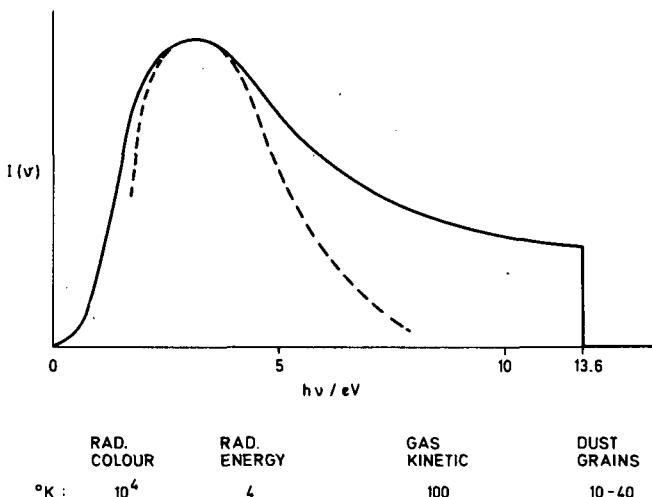


FIG. 3. The intensity distribution (solid curve) versus frequency of diluted starlight in predominantly neutral interstellar gas (HI-region). The dashed curve represents diluted black-body radiation which peaks at the same frequency.

It is wrong because interstellar space is so far from thermal equilibrium that there are many definitions of "temperature" and the gas kinetic temperature of about  $100^{\circ}\text{K}$  is only one of them. The destruction of hydrogen molecules proceeds mainly through photo-dissociation and the relevant parameters are those of the radiation field in interstellar space. The radiation field consists of diluted starlight and is itself a highly non-equilibrium affair with an energy density of only about  $4^{\circ}\text{K}$  but its spectral distribution something like  $10^4^{\circ}\text{K}$  black-body radiation diluted  $10^{14}$ -fold. In the mainly neutral regions of interstellar space (HI-regions) even the shape of the actual spectral distribution (solid curve in Fig. 3) differs somewhat from the nearest black-body shape (dashed curve), notably in the sharp cut-off at the Lyman absorption edge ( $h\nu = 13.6 \text{ eV}$ ). The threshold for photo-dissociation of  $H_2$  had previously been estimated as

about 15 eV but is now thought to be near 12 eV. You might think that a 20% error would hardly be noticed in a controversial field of astrophysics. However, this change of threshold means that H<sub>2</sub> can be dissociated rather easily by photons below the Lyman limit in dust-free HI-regions. Consequently, H<sub>2</sub> is now expected to be abundant only in the denser HI-regions which are rich in dust-grains which absorb u.v. radiation and shield the molecules.

## DENSE STARS

I want to give next two examples of phenomena occurring at high densities in the interior of highly evolved stars. The first example relates to the effect of Coulomb forces on stars slightly less massive than the Chandrasekhar limit  $M_{ch}$ . We are still considering masses  $M \gg M_{pl}$  and from what I said previously you might expect the Coulomb effects to be unimportant throughout. This is true for the pressure but not necessarily for the specific heat and for some other phenomena. Since  $M < M_{ch}$ , the interior temperature T of such a star increases with increasing density  $\rho$  only up to a point when electron-degeneracy sets in and then T decreases towards zero as the electrons become more and more degenerate. The pressure increases monotonically with increasing density (quantum mechanical Fermi pressure of the electrons replacing thermal pressure at low temperature) and Coulomb effects remain at a small percentage.

The positive ions, however, remain non-degenerate because of their large mass, and their thermal energy drops with dropping temperature whereas their Coulomb energy  $E_c$  keeps increasing with increasing density (shown schematically in Fig. 4). For the positive ions, then, the ratio  $E_c/kT$  increases towards infinity as the density approaches its limiting value and the ions freeze [13, 14] into a Coulomb lattice. Estimates now exist for the melting temperature of a Coulomb lattice from numerical Monte Carlo calculations [15]. At this stage the ion contribution to the pressure is unimportant, but the specific heat mainly comes from the ions and not from the degenerate electrons, so that the Coulomb effects are important here. As the ions freeze, their specific heat increases about a factor of two and some latent heat is released; finally, when the ions cool further below the Debye temperature of the lattice their specific heat drops drastically. Analysing statistical data on white dwarf stars is notoriously difficult, but there is even some indication that these properties of the heat content will explain some observational data on the frequency of occurrence of some types of white dwarfs [16, 17].

These ion Coulomb lattices are also important in calculating rates for "pycnonuclear reactions" [18]. In ordinary thermonuclear reactions most reactions are accomplished by pairs of nuclei with relative kinetic energy appreciably larger than the thermal energy  $kT$ . When the ion lattice is only slightly below the freezing point ("strong screening" in Fig. 5), the reacting nuclei are still free, but below a still lower temperature (the "pycnonuclear regime" in Fig. 5) even the reacting nuclei are bound in the lattice (see Fig. 6). To calculate the reaction rate in this regime [19, 20] one needs to evaluate that particular "tail" of the wave function for the lattice's zero-point vibrations which corresponds to a neighbouring pair of nuclei

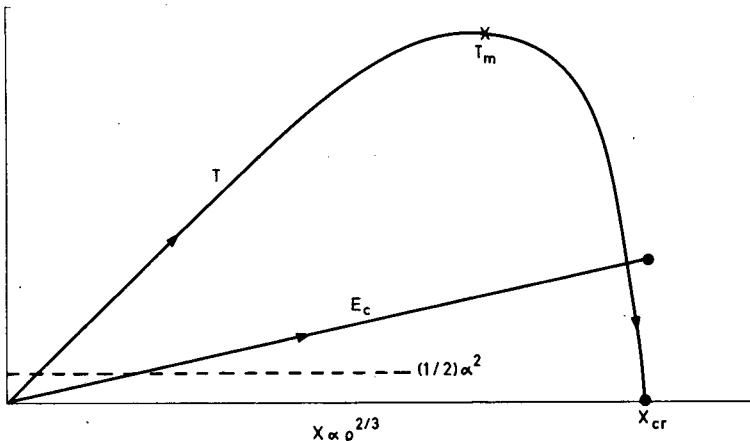


FIG. 4. A schematic plot of Coulomb energy  $E_c$  (per particle) and temperature  $T$  versus density  $\rho$ . The point  $x_{cr}$  corresponds to the final zero-temperature configuration of maximum density.

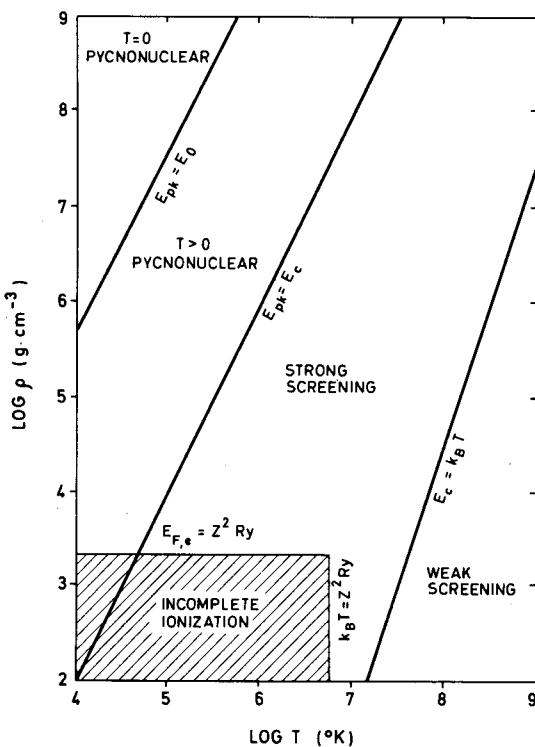


FIG. 5. The various regimes in the density-temperature plane for a  $^{12}\text{C}$  plasma. In the shaded region the Fermi and thermal energies are insufficient for complete ionization. The sloping line on the right is essentially the freezing-curve. Along the middle sloping line the excitation energy  $E_{\text{pk}}$  at which the over-all rate has a peak is comparable with the Coulomb binding energy.

approaching each other within nuclear range - a curious juxtaposition of solid-state and nuclear physics.

At much higher densities still, one also encounters the borderline between particle physics and statistical mechanics. For a range of masses (again near the Chandrasekhar mass) stable neutron stars are a possibility with interior densities so high that most protons plus electrons have undergone inverse beta-decays. The remaining small fraction of ( $p + e$ ) still gives a sufficiently high electron Fermi energy to make the neutrons stable. As the density increases still further, the neutron Fermi energy rises and so does the electron Fermi energy  $E_e$ . When  $E_e$  exceeds the  $\mu$ -meson restmass energy, some  $\mu^+$ 's are also present at equilibrium [21] (the abundance of  $e^-$  and  $\mu^-$  as a function of density is sketched in Fig. 7) and this replacement of some neutrons by ( $p + e$  or  $\mu^-$ ) lowers the pressure slightly below that for pure neutron matter (also shown in Fig. 7).

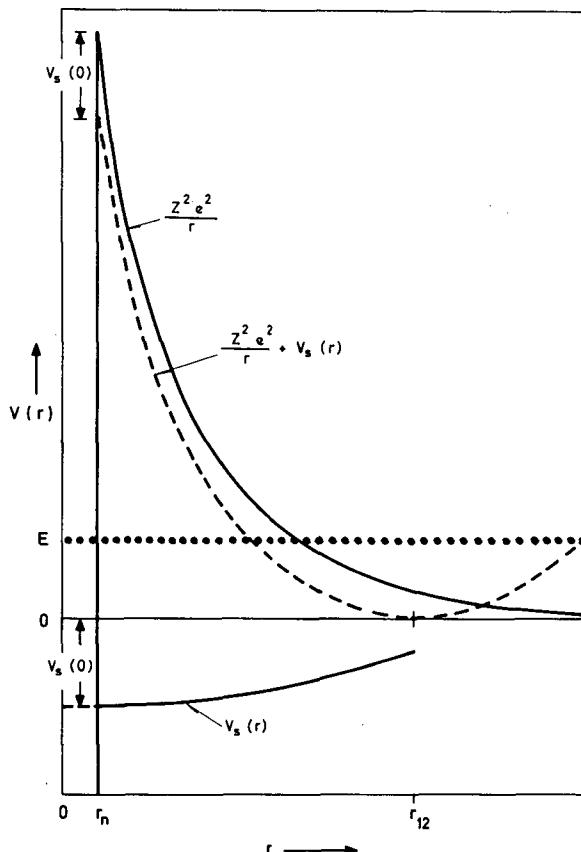


FIG. 6. The total potential  $V(r)$  for a pair of neighbour nuclei in the lattice as a function of their separation  $r$ .  $r_{12}$  is their separation at the equilibrium lattice sites,  $r_n$  the nuclear force range,  $V_s(r)$  is the screening potential due to the rest of the lattice.

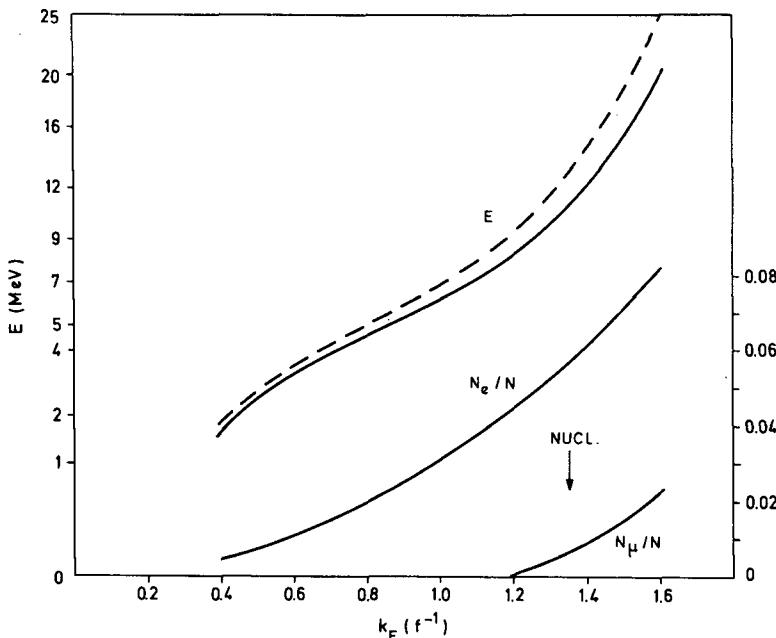


FIG. 7. The average energy  $E$  per nucleon plotted against density (expressed in Fermi momentum units). The dashed curve for pure neutron matter, the solid curve with the effect of ( $p + e$  or  $\mu^-$ ) included. The scale on the right gives the fractional abundance  $N_e/N$  of electrons and  $N_\mu/N$  of  $\mu^-$ -mesons.

As densities increase higher and higher, it is more and more effective if we decrease the fraction of baryons in neutron form in order to decrease the Fermi energy, and the repulsive core of the nuclear forces also becomes more important. At some density [22-24] for instance the change  $2n \rightarrow p + \Sigma^-$  becomes energetically favourable and so presumably do other baryon excitations (unless the repulsive core is worsened). In particle physics one sometimes argues about the difference between fundamental and composite particles. Here the related question of quantum statistics is a very concrete one. In calculations to date one has assumed that other baryon states do not contribute to the neutron Fermi energy at all (and assumed the same repulsive core for all states); is this justified? I will end with this challenge to the particle theorist.

#### REFERENCES

- [1] DIRAC, P.A.M., Proc. R. Soc. A 165 (1938) 199.
- [2] SALPETER, E.E., in Perspectives in Modern Physics (MARSHAK, R.E., Ed.), John Wiley and Sons, New York (1966).
- [3] SCHWINGER, J., these Proceedings, Vol.II.
- [4] CHANDRASEKHAR, S., Astrophys.J. 74 (1931) 81.
- [5] PONTECORVO, B.M., Soviet Phys. JETP 9 (1960) 1148.
- [6] RUDERMAN, M.A., Rep. Prog. Phys. 28 (1965) 411.

- [7] BEAUDET, G., PETROSIAN, V., SALPETER, E.E., *Astrophys. J.* 150 (1967) 979.
- [8] WU, T.T., *Phys. Rev.* 147 (1966) 1033.
- [9] GOULD, R., GOLD, T., SALPETER, E.E., *Astrophys. J.* 138 (1963) 393, 408.
- [10] KNAAP, H.F.P. et al., *Bull. astr. Insts Neth.* 18 (1966) 256.
- [11] STECHER, T., WILLIAMS, D., *Astrophys. J.* 146 (1966) 88.
- [12] STECHER, T., WILLIAMS, D., *Astrophys. J.* 149 (1967) L29.
- [13] ABRIKOSOV, A.A., *Soviet Phys. JETP* 12 (1961) 1254.
- [14] SALPETER, E.E., *Astrophys. J.* 134 (1961) 669.
- [15] BRUSH, S., SAHLIN, H., TELLER, E., *J. chem. Phys.* 45 (1966) 2102.
- [16] MESTEL, L., RUDERMAN, M., *Mon. Not. R. astr. Soc.* 136 (1967) 27.
- [17] VAN HORN, H.M., *Astrophys. J.* 151 (1968) 227.
- [18] CAMERON, A.G.W., *Astrophys. J.* 130 (1959) 916.
- [19] WOLF, R.A., *Phys. Rev.* 137 (1965) B1634.
- [20] SALPETER, E.E., VAN HORN, H.M., *Astrophys. J.* (in press).
- [21] NEMETH, J., SPRUNG, D., *Phys. Rev.* (in press).
- [22] SALPETER, E.E., *Ann. Phys.* 11 (1960) 393.
- [23] SAHAKIAN, G., VARTANIAN, Y., *Nuovo Cim.* 30 (1963) 82.
- [24] TSURUTA, S., CAMERON, A., *Can. J. Phys.* 44 (1966) 1895.

# SURVEY OF CURRENT PROBLEMS IN EXTRAGALACTIC ASTRONOMY

E. M. BURBIDGE

University of California,

Sand Diego, Calif., United States of America

## Abstract

SURVEY OF CURRENT PROBLEMS IN EXTRAGALACTIC ASTRONOMY. 1. Major constituents of the extragalactic universe; 2. Normal galaxies; physical properties; current problems; 2.1. Masses; 2.2. Mass-to-light ratio; 2.3. Spiral arms; angular momentum; barred spirals; 2.4. Evolution; 3. Nuclei - radio galaxies; 4. Compact galaxies; 4.1. The "iron galaxy"; 4.2. The chain of galaxies, VV 172.

## 1. MAJOR CONSTITUENTS OF THE EXTRAGALACTIC UNIVERSE

For decades the only constituent about which we have had any direct knowledge has been the luminous condensed matter in the form of galaxies. Counts of these by various people (e.g. Shane and Wirtanen, 1954, 1967; Zwicky, 1957, 1959), together with reasonable estimates of their masses, have characteristically yielded a matter density in the universe that is one or two orders of magnitude lower than that calculated from the various cosmological models in favour during this period (Oort, 1958). The most recent galaxy counts made at the Lick observatory give  $2.01 \times 10^{-10} L_{\odot}$  per pc<sup>3</sup>, where  $L_{\odot}$  is the solar luminosity, and this yields a density of luminous matter of  $1.36 \times 10^{-32} M/L$  gm cm<sup>-3</sup>, where  $M/L$  is some suitable average ratio of mass to light for the galaxies, in units of solar mass and luminosity,  $M_{\odot}$  and  $L_{\odot}$  (Shane and Wirtanen, 1967). Taking  $M/L = 25.8$ , Shane and Wirtanen derived  $\rho = 3.5 \times 10^{-31}$  gm cm<sup>-3</sup>. Although there are cosmological models that can give this "visible density", both the steady-state universe of Bondi, Gold and Hoyle, and the currently favoured evolving model with cosmological constant  $\Lambda = 0$  and deceleration parameter  $q_0 \approx 1$  give  $\rho \approx 2 \times 10^{-29}$  gm cm<sup>-3</sup> (see my other paper in these Proceedings).

If there is "missing" matter in the universe, it could exist in various forms, e.g.:

- (1) Diffuse cold gas or gas + solid particles (dust) uniformly spread between the galaxies;
- (2) Diffuse uniform hot gas;
- (3) Large numbers of high-density galaxies which might have been underestimated because of being missed in the counts ("compacts");
- (4) Dark compact massive clouds or even solid lumps;
- (5) QSOs if these are massive and not at the distances given by taking their red shifts as Doppler shifts due to the expansion of the universe;
- (6) Single collapsed objects of galactic mass or aggregates of evolved collapsed stars.

A cold diffuse gas consisting mainly of hydrogen used to be the most popular idea. However, the absence of a detectable absorption edge on the short-wavelength side of the Lyman- $\alpha$  line of hydrogen in the spectra of

QSOs with red shifts about 2 (on the assumption that the red shifts are cosmological) has set a very low limit on a uniform density of neutral atomic hydrogen in intergalactic space. This has been discussed by Scheuer (1965) and Gunn and Peterson (1965). It was originally thought that such an absorption edge was actually visible in the first QSO with a red shift  $\sim 2 - 3$  C 9. Later it was found that no absorption is detectable, and this sets a limit of  $\rho \leq 10^{-35}$  gm cm<sup>-3</sup> for a uniform density of neutral atomic hydrogen.

Recent evidence from a quite different point of view has suggested that possibly our galaxy is sweeping up clouds of intergalactic cold hydrogen. The Leiden radio astronomers have made surveys of our galaxy by means of the 21-cm line of neutral atomic hydrogen and they have discovered that, in intermediate and high galactic latitudes, clouds of hydrogen are apparently raining into the equatorial plane from both sides, though mainly from the northern side in an area centred on longitude 120°, latitude +40°. With rotation of our galaxy carrying the sunward galactic longitude 90°, these clouds could indeed be intergalactic matter being swept up (Oort, 1968, and references therein). Alternative suggestions are that the clouds are a part of a circulation of gas in the galaxy, or a falling-back of matter ejected by an explosive event in the nucleus, or even expanding supernova shells.

The masses of these clouds cannot be estimated until their distances are known. If they contain other chemical elements as well as hydrogen, and lie within our galaxy, then the resonance lines of Ca II might be visible in absorption in the spectra of stars lying behind the clouds. These spectral lines are absorbed by interstellar gaseous clouds in our galaxy, and if they could be detected in the spectra of any stars lying in the appropriate areas of the sky, with the same high velocity of approach as shown by the 21-cm line, then one would know that the clouds lie in front of those stars. Spectral classification and photometry of the stars, determining their distances, would thus yield the maximum distances of the clouds. Attempts are being made by Kraft to carry out these observations at Lick observatory. For the present, Oort favours the hypothesis that the clouds are being swept up by our galaxy out of intergalactic space; this would then suggest the rather high local intergalactic density of cold hydrogen of  $\sim 10^{-28}$  gm cm<sup>-3</sup>, but there are great uncertainties in this estimate. The large discrepancy with the upper limit suggested by QSOs at large red shifts is, however, interesting.

Evidence concerning the possible presence of hot intergalactic gas and evidence that might indicate the presence of dark cold clouds giving absorption lines in the spectra of QSOs is discussed elsewhere in these Proceedings.

Diffuse or discrete matter producing non-selective extinction or reddening of the light from distant galaxies has been suggested by various scientists. The absence of a "Stebbins-Whitford effect" - i.e. the absence of a generalized uniform excess reddening of light from distant galaxies - has been definitively established by recent work of Oke and Sandage. (This work is referred to in my other paper in these Proceedings.) Zwicky (1961) has found that his counts of galaxies have suggested clumping of distant clusters of galaxies due to patchy extinction (see a series of papers by Zwicky and Rudnicki, 1963, Zwicky and Berger, 1965, and Zwicky and Karpowicz, 1965, 1966). Dark patches have occasionally been seen

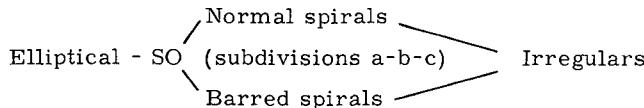
in photographs of galaxies, and the suggestion has been put forward that these are produced by small, fairly dense intergalactic dust clouds, but this evidence is inconclusive (Kowal, 1964). Dark dust clouds outside the main body of a galaxy have been found — a good case is NGC 4438 (Burbidge, Burbidge and Hoyle, 1963). Such dust clouds might have been ejected by a violent explosive event in the galaxy and may eventually leave its vicinity.

An interesting speculation about the presence of dark absorbing matter in intergalactic space is suggested by the absence of visible objects at the positions of some hitherto unidentified radio sources in the revised 3C catalogue. These are the so-called "empty fields". The radio positions are accurately known, and no visible objects can be detected on the Palomar 48-in. Schmidt photographs at these positions. In a search for faint radio galaxies — too faint to be seen with the 48-in. telescope — Sandage has been photographing these fields with the Palomar 200-in. telescope, with a technique designed to reach a very faint limiting magnitude. It is possible that the "empty fields" are empty because of intergalactic obscuration; if so, this could be revealed by a diminished surface density of faint extragalactic images in these areas.

In conclusion, however, it must be said that, with the possible exception of a high-temperature diffuse gas, there is no clear evidence of matter in any of the forms listed previously, in a sufficient amount to give a density even of the same order as the known "visible" density.

## 2. NORMAL GALAXIES; PHYSICAL PROPERTIES; CURRENT PROBLEMS

The past ten years have brought a big increase of observational data available on the physical properties of normal galaxies. A general description of their forms is given in the Hubble Atlas (Sandage, 1961). The morphological classification described there is an extension of the original Hubble scheme



The visible components of galaxies are stars (comprising most of their mass), uncondensed gas (both ionized and neutral), and dust. In their integrated spectra, the continuous radiation and absorption lines come in general from the stellar population, and emission lines from ionized interstellar gas. If no absorption lines are visible, as is sometimes the case both in radio galaxies and galaxies that are not known to be radio sources, then we have no direct information on the stellar content.

### 2.1. Masses

There are individual mass determinations (mostly from rotation curves in which the Doppler shift is measured along the major axis of a galaxy) for some 40 or 50 galaxies. There is a strong selection effect in the types

of galaxy studied, for the following reasons. Most rotation curves have been obtained from Doppler shifts of emission lines coming from the ionized gas. The Hubble morphological sequence

Irregular  $\rightarrow$  Sc  $\rightarrow$  Sb  $\rightarrow$  Sa  $\rightarrow$  SO  $\rightarrow$  Elliptical

shows from left to right a decreasing content of uncondensed matter and of the young, hot stars necessary to ionize the gas. Finally, the spatial orientations, necessary to correct line-of-sight velocities for projection, cannot usually be determined for irregular galaxies. Therefore, our knowledge of individual masses is much better for the Sc galaxies. Some Sb, and about two each of Sa and SO, have mass determinations. The SO galaxies are flattened disk-shaped galaxies without spiral structure and contain very little matter that is not condensed into stars. Therefore, their rotations can be measured only by means of absorption lines arising in the stars of which they are composed; this is a particularly difficult observational problem.

For the elliptical galaxies that are not much flattened - the E0, E1, or E2 galaxies - the mass can be obtained by means of the virial theorem  $2 \text{K.E.} + \Omega = 0$ . The kinetic energy is obtained, in terms of the unknown mass, by measuring the velocity dispersion of the stars in the centre, while the luminosity profile of the galaxy is measured and the assumption is made that this can be used to give the mass distribution and thence the potential energy.

Mass determinations range from  $10^9$ - $10^{12}$  solar masses, but large numbers of lower-mass dwarf galaxies also exist. There is a general increase in mass from irregulars and class Sc along the Hubble sequence to class E, but there is a wide range of masses for each class and great overlap.

For all types of galaxy, statistical masses from orbital motion of pairs (Page, 1962; Page, Dahn and Morrison, 1961) can be obtained. For spirals, there is good agreement with individual determinations; for E and SO, masses from binaries tend to come out too large. There are, however, too few individual determinations to know yet whether this is a significant discrepancy.

The virial theorem has been applied to galaxies in clusters, assuming - as for stars in a galaxy - that a cluster of galaxies is in equilibrium (for the method see Zwicky, 1957; Oort, 1958). This method has always tended to yield masses much larger than the individual determinations. This is a long-standing puzzle<sup>1</sup>. Are the clusters not in fact in equilibrium - i.e. do they have too large a kinetic energy, so that they are dissipating on a time-scale short compared with the Hubble time-scale for expansion of the universe ( $\sim 10^{10}$  yr)? Or do the clusters contain much unseen mass, e.g. collapsed configurations? This brings us to the general question of the mass-to-light ratio in galaxies.

## 2.2. Mass-to-light ratio

The mass-to-light ratio of a galaxy (or the average in a cluster) can be obtained if the luminosity is measured, the distance known, and the mass determined. For E galaxies this ratio is very large and, from what I have

---

<sup>1</sup> The proceedings of a small conference on this topic can be found in Astron. J. 66 (1961) 533.

said about the apparently large masses of clusters of galaxies, even more so for the galaxies they contain. For individual ellipticals, one has to account for values

$$M/L = 50 - 70 \text{ M}_\odot/\text{L}_\odot$$

where obviously an aggregate of stars exactly similar to the sun would give  $M/L = 1$ . One has to synthesize a stellar population that will give these high values and also be consistent with the observed properties of the galaxy, such as the integrated line spectrum coming from its stars and its continuum energy distribution. In the case of stars in our solar neighbourhood, we have

$$L \propto M^3 \text{ or } 4$$

so that a high  $M/L$  can in principle be obtained if one supposes that there is a very large population of low-mass stars (or white dwarfs) which yield very little light for their mass. It is well known that the mass distribution function of stars in our galaxy in the solar neighbourhood yields  $M/L$  around 2, much too low to account for the E galaxies. If we suppose that a typical elliptical galaxy is so old that in a stellar population initially like that of the solar neighbourhood all but the low-mass stars have evolved to white dwarfs, then the time necessary for this to occur is excessively long. Further, there cannot be too large a number of stars of very low mass, because these will be relatively cool and may produce an infra-red magnitude too bright to be compatible with the observations. Spinrad at Berkeley has been doing narrow-band filter photometry of E galaxies – in and out of spectral features over a wide range of wavelengths – and fitting the results with synthetic stellar populations. He can achieve quite high  $M/L$  values, but there is not perfect agreement between theory and observation.

This dilemma has led some of us to favour the possibility that hidden mass is present in the form of evolved stars whose initial masses were greater than the critical mass for a stable white dwarf or neutron star configuration ( $\sim 1 \text{ M}_\odot$ ). Such stars, when they have exhausted their nuclear fuel, will, according to present theory, collapse through the Schwarzschild radius to a singularity unless they can shed sufficient mass. They will then be detectable only by their gravitational field.

### 2.3. Spiral arms; angular momentum; barred spirals

One long-standing problem – that of the spiral arms in spiral galaxies – has recently yielded to theoretical attack, by C. C. Lin and others. The problem has been to explain why spiral arms, which are predominantly a phenomenon exhibited by the uncondensed gas and dust and young stars that have recently formed out of this material, maintain themselves with one or two turns only, under the action of differential rotation in a system whose age is at least 100 times one rotation period. This work by Lin is discussed by Burke in these Proceedings.

The barred spirals provide an interesting hydrodynamical problem. The flow pattern around a prolate spheroid (the bar) rotating end over end has been studied by Prendergast (1962) and Freeman (1965, 1966).

Relatively little has been done yet to connect observations with theory. Velocities have been measured in the gas in a number of barred spirals by Burbidge, Burbidge and Prendergast and the results have been published in a series of papers in the Astrophysical Journal from 1960 to 1964. In the absence of axial symmetry it has proved impossible, however, to derive the three-dimensional velocity field from observations of only the line-of-sight velocity component.

There are, however, two suggestions: first, the difference between barred and the normal spirals may be that the former have greater angular momentum per unit mass; and, second, barred spirals may evolve into normal spirals. The angular momentum per unit mass would be of great interest if it could be determined for the whole range of galactic types, but we have practically no reliable information, even for the Sc galaxies. Most of the angular momentum resides in the outer parts, despite the decrease in density with distance from the centre, and the rotation curve is usually least well determined in the outer parts which are faint, so that the calculations that have been made are most uncertain.

#### 2.4. Evolution

No proper theory of galactic evolution exists. The difficulty is basic and stems ultimately from uncertainty at the cosmological level. The results of mass determinations preclude a simple evolution along the morphological sequence Irr  $\rightarrow$  S  $\rightarrow$  E, which is the sense suggested by the decrease of uncondensed gas along the sequence (Holmberg, 1964). A simple theory of the collapse of a large spherical, initially slowly rotating gas mass will account for some, but not all, of the facts (Eggen, Lynden-Bell and Sandage, 1962; Oort, 1964; Hatanaka et al., 1964). I shall mention just two aspects of the morphology of galaxies which are particularly difficult to fit into this kind of theory.

Very high star densities in the nuclei of some galaxies have been proposed to explain the activity in nuclei of radio galaxies and Seyfert galaxies through star collisions and even agglomeration into a short-lived massive object (Ulam and Walden, 1964; Gold, Axford and Ray, 1965). However, such high densities are hard to achieve from an original low-density gas cloud in times of the order of the Hubble time-scale.

The collapse of a roughly spherical cloud will not account for some objects that look, on the grounds of their gas content and young stellar population, to be relatively early in their evolutionary life-times (photographs and discussions of some of these objects can be found in two papers, by Burbidge, Burbidge and Hoyle, 1963, and by Burbidge, Burbidge and Shelton, 1967). Some of these galaxies have large dimensions and very elongated forms and are of interest because it would seem that they cannot persist long in their present forms.

### 3. NUCLEI - RADIO GALAXIES

The outstanding problem with radio galaxies is to explain the enormous energy release that must have occurred in the form of relativistic particles and magnetic fields; these are what are needed to produce the synchrotron radiation that is observed. There have been so many discussions of this

problem during the last ten years that it would not be profitable to attempt to give a complete list of references. Accounts of optical studies of radio galaxies may be found in papers by Matthews, Morgan and Schmidt (1964), Schmidt (1965), and Burbidge (1965). These studies have shown that a variety of galactic types can become radio sources, but the strongest radio sources are the giant massive extended elliptical or D-type objects, or the N-type objects that consist of a very bright nucleus and a fainter wispy outer part. Optical evidence has now made it clear that the origin of the activity appears to be some kind of violent explosion in the nucleus.

Ambartsumian first suggested that the nuclei of galaxies may be the seat of some very interesting physics (see accounts by Ambartsumian, 1958; 1964). Burbidge, Burbidge and Sandage (1963) collected evidence for the occurrence of violent events in the nuclei of radio and other galaxies. The first good case of such an explosion in the centre of a galaxy was that of the bright, relatively nearby irregular galaxy M 82, studied by Lynds and Sandage (1963).

The Seyfert galaxies, isolated as a small class by Seyfert (1943), have small-diameter intensely bright nuclei; these show very broad emission lines in their spectra corresponding to  $\sim 3000$  km/s in velocity spread, and some show emission lines from high states of ionization. In February 1968, a conference devoted to Seyfert galaxies and related objects was held in Tucson, and the proceedings, prepared by Pacholczyk and Weymann, will be published in the *Astronomical Journal*. Some Seyfert galaxies are radio sources. NGC 1275 is a particularly interesting case, as the ionized gas in it displays two discrete velocities. One velocity is about 5200 km/s, the same as the stars in the galaxy and the same as other galaxies belonging to the same cluster of galaxies. The other velocity is about 8200 km/s and occurs in the gas lying mostly on one side of the galaxy. The obvious explanation is that this gas has been exploded out of the nucleus and now lies behind the galaxy, but the gas with the high velocity runs right up to the nucleus and this suggests that the ejection (if this is what has happened) is still going on. The time taken for the gas to reach the farthest-out point at which it has been seen would be a few times  $10^6$  yr (Burbidge and Burbidge, 1965).

That nuclei of active galaxies can be variable in optical energy output on a time-scale of years or even months has come as a surprise (see Oke, 1967). Strong infra-red emission has been found in Seyfert nuclei (the flux actually peaks in the infra-red in these galaxies), and this shows that most of the energy output in these cases is non-thermal in origin - presumably synchrotron radiation.

The large velocity spread in the excited gas in Seyfert nuclei suggests continuing activity rather than one large outburst. The masses of three Seyfert galaxies are now known, and they are in the normal range for Sc's - about  $3 \times 10^{10}$  solar masses. The escape velocity at the centre is then less than the observed velocity dispersion, and the gas should escape in  $\sim 10^4$  yr. But these galaxies comprise 2% of all spirals. Thus an age of  $\sim 10^8$  yr for the phenomenon is suggested. Therefore, the phenomenon must be either a continuing or a recurrent one.

Gold has linked the Seyfert phenomenon to star collisions in a region of high stellar density. The problem is how to achieve the high density in an originally very diffuse gas with considerable angular momentum. McCrea, in a version of the steady-state cosmology, has proposed that

matter is created in the nuclear regions - i.e. creation of matter occurs preferentially in regions that already have high density.

Non-circular motions occur in the gas in the nuclei of even relatively quiet normal galaxies, such as our own, M 31 (Münch, 1961), or M 51 (Burbidge and Burbidge, 1964a). A variety of non-circular motions occurs in the neutral hydrogen gas in the central region of our own galaxy; these are discussed by Burke in these Proceedings. In M 31 and M 51 the motions are detected in the ionized gas component.

#### 4. COMPACT GALAXIES

Observers have recently turned their attention to what may prove to be a large constituent of the extra-galactic universe - the compact galaxies. Zwicky catalogued objects that appear almost starlike on the Palomar 48-in. Schmidt plates; Haro (Mexico) and Markarian (USSR) have surveyed for small blue galaxies. Zwicky (1964, 1966) has studied some compacts spectroscopically, and recently others have also started working on them - e.g. Khachikian on the Markarian objects and Sargent on the Zwicky objects (papers by these authors can be found in the Proceedings of the Tucson Conference on Seyfert Galaxies and Related Objects, already referred to). Other objects that may be in the same category are the small blue objects discovered by Rubin, Moore and Bertiau (1967) in the Virgo cluster of galaxies, on which Mrs. Rubin has recently started spectroscopic work. Still others, occurring in small groups, have been independently catalogued by Vorontsov-Velyaminov (1959) as interacting galaxies and by Arp (1966) in his atlas of peculiar galaxies.

These objects, on spectroscopic examination, are found to be of many varieties. Not all are high-density systems, but some undoubtedly are. Although it is clear that they are quite numerous, it is too early yet to derive a mean space density. What their place is in the evolutionary scheme is quite unclear at the moment - particularly the small objects of low red shift that have strong emission lines indicating a large content of ionized gas.

Some objects have absorption-line spectra with very broad lines; the stars which comprise these objects evidently have very large velocity dispersions, and these must have (by the virial theorem) large masses and large mass-to-light ratios (Zwicky, 1964).

Some objects fall into the category of Seyfert galaxies; VV 144 is one such (Burbidge and Burbidge, 1964b). It has a jet extending from it on one side. It is not known what is producing the light in the jet, since neither the form of the continuous energy distribution nor the line spectrum has yet been observed. The radiation may be non-thermal in origin, as in the well-known jet in the bright radio galaxy M 87. (In M 87 the radiation is known to be synchrotron radiation, from the high degree of polarization found.)

There are two particularly interesting cases of compact galaxies which have been found by Sargent. I shall discuss these now, as they both indicate phenomena that do not fit the conventional ideas on physical processes in galaxies.

#### 4.1. The "iron galaxy" (Sargent, 1968)

An object in one of Zwicky's lists - I Zw 0051 + 12 - has a most unusual spectrum for a galaxy and may have an anomalous chemical composition. It is of Seyfert type, with a small starlike nucleus and a faint wispy outer part, and its spectrum shows strong broad emission lines of hydrogen and permitted transitions of Fe II. The usual lines found in the ionized gas in galaxies - forbidden lines of singly and doubly ionized oxygen, ionized nitrogen, and ionized sulphur - are absent. Further, no forbidden lines of Fe II are found. This means that the gas must have a much higher electron density than that found in normal galaxies, and also higher than that found in the usual kind of Seyfert galaxies, i.e.

$$N_e \geq 10^6 \text{ cm}^{-3}$$

The spectrum bears some resemblance to that of the quasi-stellar object 3C 273, which also shows Fe II lines, but in 3C 273 these lines are quite weak and can be explained by a normal chemical composition (roughly the same as the sun's) together with a rather high electron density,  $N_e \sim 10^7 \text{ cm}^{-3}$ . No absorption lines are found in the spectrum of the compact galaxy, so nothing is known about its stellar population. The iron-to-hydrogen abundance has not yet been determined.

The red shift,  $z = 0.0605$ , yields an unusually bright intrinsic luminosity for the object - a factor of 5 or 6 brighter than the brightest elliptical galaxies - again suggesting a connection with QSOs and raising the query whether perhaps the red shift is not wholly cosmological.

#### 4.2. The chain of galaxies, VV 172

The question of non-cosmological red shifts is raised in more crucial form by another compact, a member of a chain of galaxies catalogued by Vorontsov-Velyaminov and studied by Burbidge, Burbidge and Hoyle (1963). We were at that time concerned with the time-scale for disruption of such a configuration of galaxies - five strung out in a closely-spaced chain; this time-scale appeared to be only  $\sim 10^9$  yr if the velocity dispersion of the members was of the order 100 km/s. The two brightest members had a mean recession velocity of 15 700 km/s (which set the scale of the system for us, since we had no reason to doubt that this could be used in the normal way with the Hubble expansion constant to derive the distance), and the velocity difference of the two was about 170 km/s.

Now Sargent (private communication) has measured the red shifts of all the members. He has found values similar to the value of 15 700 km/s in two of the remaining three, but the second galaxy in the chain has a velocity more than twice as much - 36 900 km/s; it has strong forbidden emission lines of O II from low-density gas. Sargent believes that the Ca II resonance lines are present in absorption (these are the most usually found absorption lines in low-dispersion spectra of galaxies; they usually come from the stellar population but could come from uncondensed gas). All the other galaxies in the chain have only absorption lines.

It strains coincidence too far to suggest that the galaxy with the discrepant velocity could be a background object, since it falls into place as the second member of this unusual and compact group and has appro-

ximately the same brightness as the others. If the mass is normal – say  $10^{10}$  solar masses – the kinetic energy required to give such a velocity difference is about  $4 \times 10^{61}$  erg, right up in the range of the energies we are so hard put to explain in the radio galaxies.

This is not the first time that large velocity differences have been found in what appear to be physical groups, though it is much the largest difference so far. A group called IC 3481-3, studied by Zwicky (1957), has one of its apparent members with a velocity some 7000 km/s less than the other two; a group called Stephen's Quintet (Burbidge and Burbidge, 1961) is another. In both these cases, the odd member of the group has a lower velocity than the rest and many scientists have assumed that it lies in the foreground. A related case was found by Sersic (1966) – a galaxy with a double nucleus having a velocity difference of some thousands of km/s. We may also recall the radio and Seyfert galaxy NGC 1275, with its velocity difference of 3000 km/s.

Such velocity differences – particularly in the chain of galaxies VV 172 – do not fit into the conventional scheme of extragalactic astrophysics, and I believe the time has now come when it is necessary to pay serious attention to these observational results.

## R E F E R E N C E S

Asterisks indicate articles that either are of a review character or occur in conference proceedings containing many other papers of interest. The rest of the references are of a more specialist character, of less interest to the general reader.

- \* AMBARTSUMIAN, V. (1958) Structure and Evolution of the Universe (STOOPS, R., Ed.), 11th Solvay Conf., 241.
- \* AMBARTSUMIAN, V. (1964) Structure and Evolution of Galaxies, 13th Solvay Conf., Interscience Publishers, New York, 1.
- \* ARP, H. (1966) *Astrophys. J. Suppl. Ser.* 14, 1.
- \* BURBIDGE, E. M. (1965) Proc. 35th Course Fermi School Int. Physics, 43; Structure and Evolution of Galaxies, 13th Solvay Conf., Interscience Publishers, New York, 137.
- BURBIDGE, E. M., BURBIDGE, G. R. (1961) *Astr. J.* 66, 541.
- BURBIDGE, E. M., BURBIDGE, G. R. (1964a) *Astrophys. J.* 140, 144.
- BURBIDGE, E. M., BURBIDGE, G. R. (1964b) *Astrophys. J.* 140, 1307.
- BURBIDGE, E. M., BURBIDGE, G. R. (1965) *Astrophys. J.* 142, 1351.
- BURBIDGE, E. M., BURBIDGE, G. R., HOYLE, F. (1963) *Astrophys. J.* 138, 873.
- BURBIDGE, E. M., BURBIDGE, G. R., PRENDERGAST, K. H. (1960) *Astrophys. J.* 132, 654.
- BURBIDGE, E. M., BURBIDGE, G. R., PRENDERGAST, K. H. (1962) *Astrophys. J.* 136, 119.
- BURBIDGE, E. M., BURBIDGE, G. R., SHELTON, J. W. (1967) *Astrophys. J.* 150, 783.
- \* BURBIDGE, G. R., BURBIDGE, E. M., SANDAGE, A. R. (1963) *Rev. mod. Phys.* 35, 947.
- EGGEN, O. J., LYNDEN-BELL, D., SANDAGE, A. R. (1962) *Astrophys. J.* 136, 748.
- FREEMAN, K. C. (1965) *Mon. Not. R. astron. Soc.* 130, 183.
- FREEMAN, K. C. (1966) *Mon. Not. R. astron. Soc.* 133, 47; 134, 1, 15.
- \* GOLD, T., AXFORD, I., RAY, E. C. (1965) Quasi-Stellar Sources and Gravitational Collapse (ROBINSON, I., SCHILD, A., SCHUCKING, E. L., Eds), Univ. Chicago Press, Chicago, 93.
- GUNN, J. E., PETERSON, B. A. (1965) *Astrophys. J.* 142, 1633.
- HATANAKA, T., HAYAKAWA, S., ISHIDA, K., TAKETANI, M. (1964) *Prog. theor. Phys., Osaka, Suppl.* 31.
- HOLMBERG, E. (1964) *Ark. Astr.* 3 30, 387.
- KOWAL, C. T. (1964) *Astr. J.* 69, 757.
- LYNDS, C. R., SANDAGE, A. R. (1963) *Astrophys. J.* 137, 1005.

- MATTHEWS, T.A., MORGAN, W.W., SCHMIDT, M. (1964) *Astrophys. J.* 140, 35.  
\* MÜNCH, G. (1961) Problems in Extragalactic Research (McVITTIE, G.C., Ed.), I.A.U. Symp. No. 15, MacMillan, New York, 119.  
OKE, J.B. (1967) *Astrophys. J.* 150, 65.  
\* OORT, J.H. (1958) Structure and Evolution of the Universe (STOOPS, R., Ed.), 11th Solvay Conf., 163.  
\* OORT, J.H. (1964) *Trans. int. astr. Un.* XII A, 789.  
\* OORT, J.H. (1968) Galaxies and the Universe (WOLTJER, L., Ed.), Columbia Univ. Press, New York, 1.  
PAGE, T. (1961) *Astrophys. J.* 132, 910.  
PAGE, T. (1962) *Astrophys. J.* 136, 685.  
PAGE, T., DAHN, C.C., MORRISON, F.F. (1961) *Astr. J.* 66, 614.  
\* PRENDERGAST, K.H. (1962) Interstellar Matter in Galaxies (WOLTJER, L., Ed.), Benjamin, New York, 217.  
RUBIN, V.C., MOORE, S., BERTIAU, F.C. (1967) *Astr. J.* 72, 59.  
\* SANDAGE, A.R. (1961) The Hubble Atlas of Galaxies, Carnegie Institution, Washington.  
SARGENT, W.L.W. (1968) *Astrophys. J.* 152, L31.  
SCHEUER, P.A.G. (1965) *Nature* 207, 963.  
\* SCHMIDT, M. (1965) *Astrophys. J.* 141, 1; Structure and Evolution of Galaxies, 13th Solvay Conf., Interscience Publishers, New York, 130.  
SERSIC, J.L. (1966) *Z. Astrophys.* 64, 202.  
SEYFERT, C.K. (1943) *Astrophys. J.* 97, 28.  
SHANE, C.D., WIRTANEN, C.A. (1954) *Astr. J.* 59, 285.  
SHANE, C.D., WIRTANEN, C.A. (1967) *Univ. Calif. Publs Lick Obs.* 22, 1.  
ULAM, S.M., WALDEN, W.E. (1964) *Nature* 201, 1202.  
VORONTSOV-VELYAMINOV, B.A. (1959) Atlas of Interacting Galaxies, Moscow.  
ZWICKY, F. (1957) Morphological Astronomy, Springer, Heidelberg.  
\* ZWICKY, F. (1959) *Handb. Phys.* 53, 390.  
\* ZWICKY, F. (1961) Problems of Extragalactic Research (Mc VITTIE, G.C., Ed.), I.A.U. Symp. No. 15, MacMillan, New York, 347.  
ZWICKY, F. (1964) *Astrophys. J.* 140, 1467.  
ZWICKY, F. (1966) *Astrophys. J.* 143, 192.  
ZWICKY, F., BERGER, J. (1965) *Astrophys. J.* 141, 34.  
ZWICKY, F., KARPOWICZ, M. (1965) *Astrophys. J.* 142, 625.  
ZWICKY, F., KARPOWICZ, M. (1966) *Astrophys. J.* 146, 43.  
ZWICKY, F., RUDNICKI, K. (1963) *Astrophys. J.* 137, 707.  
ZWICKY, F., RUDNICKI, K. (1966) *Z. Astrophys.* 64, 246.



# SOLAR NEUTRINO ASTRONOMY \*

W.A. FOWLER

California Institute of Technology,  
Pasadena, Calif., United States of America,  
and  
Institute of Theoretical Astronomy,  
Cambridge, United Kingdom

## Abstract

SOLAR NEUTRINO ASTRONOMY. The current experimental, observational and theoretical situation in regard to the flux of neutrinos from the sun is reviewed. It is concluded that there are no reasons at the present time to question seriously our fundamental concepts of stellar structure and evolution.

This story will be a sad one especially for those of us who had hoped that things would work out to give a solar neutrino flux even greater than our most sanguine expectations. Alas, this has not been the case and although I will conclude that there are still no reasons to question seriously our fundamental concepts of stellar structure and evolution, yet at the same time it is already clear that the road to definitive results in neutrino astronomy will be a hard and rocky one.

There is perhaps a moral to be learned from my story in regard to the current situation in theoretical astrophysics. On the one hand, since the discovery of the pulsars there has been put forth a spate of "sophisticated" theories of highly evolved objects such as white dwarfs, neutron stars or what have you, all purporting to pulsate or rotate in agreement with the observations. On the other hand, as we shall see, it requires detailed and careful studies of the not too highly evolved sun to bring observation and theory into the even precarious agreement attainable at the present time.

I am reminded of a story. A friend of mine acquired a young dog and soon found that the pup enjoyed nothing more than returning a thrown stick. My friend had occasion to visit the seashore and there, in the game with the dog, threw a stick of wood, which he found on the beach, far out into the water. The pup ran down the beach, ran right over the water, retrieved the floating stick and returned in the same manner. My friend quite naturally decided that perhaps he had had one too many the night before but also being an experimental physicist decided to repeat the experiment. The result was the same. At this junction my friend spotted a colleague several hundred yards down the beach and eventually caught up with him with the remark "I want to show you something remarkable". Thereupon my friend again threw the stick far out into the water, the dog ran out over the water, retrieved the stick, and returned in the same way. My friend turned to his colleague "Don't you think that is remarkable?". His colleague replied "I don't see anything remarkable except that your

---

\* Supported in part by the National Science Foundation and the Office of Naval Research, USA.

dog can't swim". Perhaps it is so in theoretical astrophysics — we are trying to run on the water while we are still unable to swim.

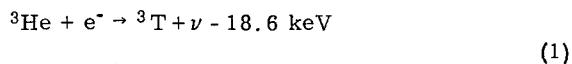
In the sequel I shall discuss (1) how neutrinos are emitted by the sun; (2) how solar neutrinos may be detected; (3) predictions based on solar models using current empirical nuclear and atomic data; (4) observations to date; and (5) afterthoughts. My remarks will be mainly based on my early paper on energetic neutrinos from the decay of  $^8\text{B}$  in the sun and other stars (FOWLER, W.A., *Mém. Soc. r. Liège, Ser. 5, 3* (1960) 207) and on two very recent papers — one observational (DAVIS, R., Jr., HARMER, D.S., HOFFMAN, K.G., *Phys. Rev. Lett.* 20 (1968) 1205) and one theoretical (BAHCALL, J.N., BAHCALL, N.A., SHAVIV, G., *Phys. Rev. Lett.* 20 (1968) 1209). An adequate bibliography will be found in these papers.

## 1. HOW NEUTRINOS ARE EMITTED BY THE SUN

Energy generation and neutrino production are thought to occur in the sun through the proton-proton chain and the CNO bi-cycle. The nuclear reactions involved are shown for the chain in Table I and for the bi-cycle in Table II. In the first case the neutrinos are produced in the weak reaction  $\text{p} + \text{p} \rightarrow ^2\text{D} + \text{e}^+ + \nu$  or in the decay of  $^7\text{Be}$  or  $^8\text{B}$ . In the second case the neutrinos are mainly produced in the decay of  $^{13}\text{N}$  and  $^{15}\text{O}$  with a very small contribution (<0.1%) from the decay of  $^{17}\text{F}$ .

The magnitude of the energy of the various neutrinos produced in the sun is of key importance, since the efficiency of detection by a particular neutrino absorption reaction is proportional to the square of the neutrino energy in excess of the threshold for that reaction. The energy spectra for the various neutrinos listed in Tables I and II are shown in Fig. 1. For beta decay processes the discrete energy of the neutrinos emitted in electron capture is shown as well as the continuous energy spectrum for those accompanying positron emission. Where positron emission is energetically permitted, the continuum neutrinos usually far exceed in number the mono-energetic capture neutrinos, but the energy of the latter exceeds the maximum energy of the former by  $2m_e c^2 = 1.022 \text{ MeV}$ .

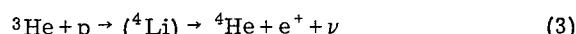
The flux strength for the various solar neutrinos will be discussed below in connection with the predictions from solar models. The question arises as to whether the reactions of Tables I and II are complete and, of course, they are not. For example, there are the following additional processes in principle for the conversion of  $^3\text{He}$  into  $^4\text{He}$ :



and



or



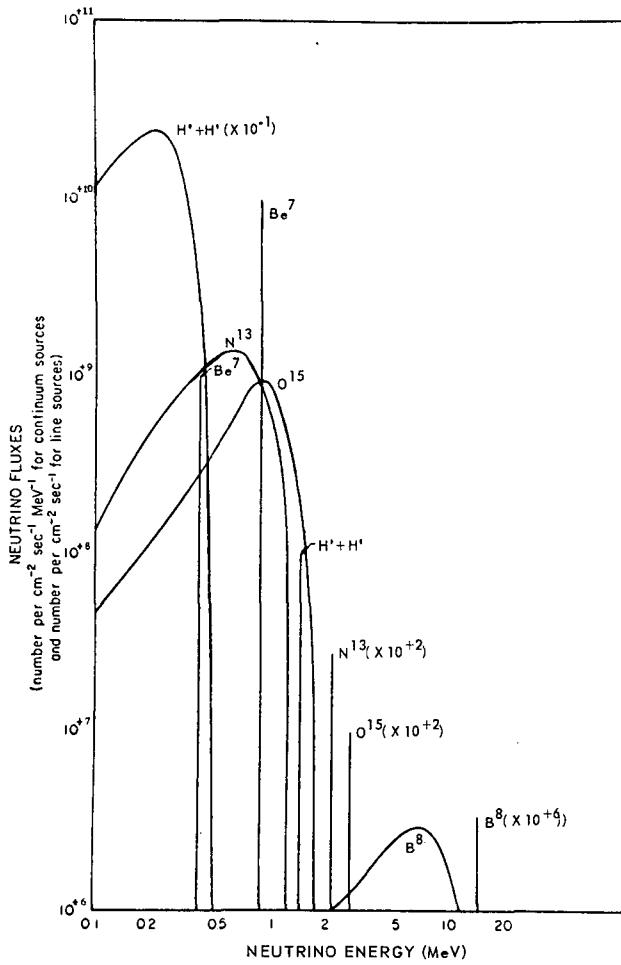


FIG. 1. Predicted neutrino spectrum from the sun. Fluxes given here are evaluated at the earth's surface. The neutrino lines are produced by the capture of free electrons; the small thermal widths ( $\sim 1$  keV) of these lines have been neglected. (Reproduced from BAHCALL, J. N., Science 147 (1965) 117.)

In the first set of reactions it can be shown that the endoergic electron capture rate is very slow since it contains a Boltzmann factor  $\exp(Q/kT)$ , which for  $Q = -18.6$  keV and  $kT \approx 1.3$  keV at the centre of the sun is equal to  $10^{-6}$ . The second set of reactions requires that the ground-state of  ${}^4\text{Li}$  be bound, which has been conjectured by numerous theoreticians. The third process only requires that this state has a low enough energy to serve as a "resonance" in this process at the effective thermal interaction energy for  ${}^3\text{He} + p$ . It has been clear for years from nuclear systematics that this conjecture was nonsense and difficult experiments in our laboratory at Caltech and elsewhere have laid this ghost once and for all by showing that the "ground-state" of  ${}^4\text{Li}$  is unbound relative to  ${}^3\text{He} + p$  by 4.7 MeV. In concluding this particular discussion I can only

TABLE I. REACTIONS OF THE PROTON-PROTON OR PP CHAIN (Aug. 1968)

The pp chain	Energy release	Solar $S_{\text{eff}}$ (keV-barn) or $\bar{\tau}$	Solar $(fS)_{\text{eff}}$
$^1\text{H} + ^1\text{H} \rightarrow ^2\text{D} + e^+ + \nu$	$1.192 \times 2 = 2.38 \text{ MeV}$	$4.2 \pm 0.2 \times 10^{-22}$ (calculated)	$4.4 \times 10^{-22}$
$^2\text{D} + ^1\text{H} \rightarrow ^3\text{He} + \gamma$	$5.494 \times 2 = 10.99$	$3.1 \pm 0.3 \times 10^{-4}$ (direct capture)	$3.3 \times 10^{-4}$
$^3\text{He} + ^3\text{He} \rightarrow ^4\text{He} + ^1\text{H} + ^1\text{H}$ or $^3\text{He} + ^4\text{He} \rightarrow ^7\text{Be} + \gamma$	12.86 26.23 (2% $\nu$ -loss)	$5.0 \pm 1.0 \times 10^3$ ( $n$ -exchange)	$6.0 \times 10^3$
$^7\text{Be} + e^- \rightarrow ^7\text{Li} + \nu + \gamma$ $^7\text{Li} + ^1\text{H} \rightarrow ^4\text{He} + ^4\text{He}$ or $^7\text{Be} + ^1\text{H} \rightarrow ^8\text{B} + \gamma$ $^8\text{B} \rightarrow ^8\text{Be}^* + e^+ + \nu$ $^8\text{Be}^* \rightarrow ^4\text{He} + ^4\text{He}$	1.59 0.05 17.35 25.68 (4% $\nu$ -loss) 0.13 7.7 3.0 19.1 (29% $\nu$ -loss)	$0.45 \pm 0.15$ (direct capture) $\bar{\tau} = 120 \text{ d}$ $125 \pm 15$ (non-resonant) $3.0 \pm 1.0 \times 10^{-2}$ (direct capture) $\bar{\tau} = 1.1 \text{ s}$ $\bar{\tau} = 10^{-21} \text{ s}$	0.55 150 $3.7 \times 10^{-2}$
$^1\text{H} + ^1\text{H} + ^1\text{H} + ^1\text{H} \rightarrow ^4\text{He}$	Total = $\pm 0.0005$	$26,7312 \text{ MeV}$	

TABLE II. REACTIONS OF THE CNO BI-CYCLE (Aug. 1968)

The CNO bi-cycle	Energy release	Solar $s_{\text{eff}}$ (keV-barn) or $\bar{\tau}$	Solar $(fs)_{\text{eff}}$
$\rightarrow ^{12}\text{C} + ^1\text{H} \rightarrow ^{13}\text{N} + \gamma$	1.94	$1.53 \pm 0.15$ ( $E_{\text{r}} = 460$ )	2.2
$^{13}\text{N} \rightarrow ^{13}\text{C} + e^+ + \nu$	1.50	$\bar{\tau} = 870$ s	
$^{13}\text{C} + ^1\text{H} \rightarrow ^{14}\text{N} + \gamma$	7.55	$5.9 \pm 0.8$ ( $E_{\text{r}} = 555$ )	8.4
$\rightarrow ^{14}\text{N} + ^1\text{H} \rightarrow ^{15}\text{O} + \gamma$	7.29	$3.0 \pm 0.6$ (non-resonant)	4.5
$^{15}\text{O} \rightarrow ^{15}\text{N} + e^+ + \nu$	1.73	$\bar{\tau} = 178$ s	
$^{15}\text{N} + ^1\text{H} \rightarrow ^{12}\text{C} + ^4\text{He}$	4.96	$7.5 \times 10^4$ (destructive interference 340, 1050)	$1.1 \times 10^5$
or (1/2200)		24.97 MeV	
(6% $\nu$ -loss)			
$^{15}\text{N} + ^1\text{H} \rightarrow ^{16}\text{O} + \gamma$	12.13	32 (constructive interference 340, 1050)	48
$^{16}\text{O} + ^1\text{H} \rightarrow ^{17}\text{F} + \gamma$	0.60	$9.9 \pm 1$ (non-resonant)	16
$^{17}\text{F} \rightarrow ^{17}\text{O} + e^+ + \nu$	1.76	$\bar{\tau} = 95$ s	
$^{17}\text{O} + ^1\text{H} \rightarrow ^{14}\text{N} + ^4\text{He}$	1.19	$10 \pm 2$ (resonant)	16
(1/2200)		15.68 MeV	
$^1\text{H} + ^1\text{H} + ^1\text{H} + ^1\text{H} \rightarrow ^4\text{He}$	Total	26.7312 MeV $\pm 0.0005$	

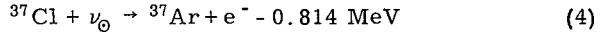
emphasize that a great deal of thought has been given to alternative processes other than those in Tables I and II and that even more effort has been spent in experiments tracking down and eliminating all serious conjectures.

In view of the experimental travails alluded to above and of others associated with the observational technique to be discussed below it might be asked "why bother about neutrinos from the sun?". The answer is of course obvious to anyone who has thought about the problem. Solar neutrinos bring us information from the centre of the sun — only one in  $10^8$  are scattered in travelling from the centre out through the sun. The light, the X-rays and the high-energy particles from the sun tell us about conditions at the surface of the sun. We can infer the interior conditions through the construction of solar models which have the correct mass, radius, luminosity and composition but, on the other hand, the  ${}^8\text{B}$  neutrinos are, for example, a very sensitive "thermometer" for solar interior temperatures since the reaction rate for  ${}^7\text{Be} + \text{p} \rightarrow {}^8\text{B} + \gamma$  varies approximately as  $T^{14}$ . Astronomy without neutrinos reminds me of a proverb which I learned as a graduate student:

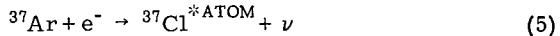
You cannot tell  
The depth of the well  
By the length of the handle on the pump!

## 2. HOW SOLAR NEUTRINOS ARE DETECTED

In this discussion I shall limit attention to the work of Davis and his collaborators at Brookhaven, who use the Pontecorvo technique in which  ${}^{37}\text{Cl}$  is the solar neutrino absorber according to the reaction



The  ${}^{37}\text{Ar}$  produced in this reaction decays with a half-life of 35 d as follows



The emitted neutrino cannot be detected but the  ${}^{37}\text{Cl}$ -atom is left in an excited state which decays to its ground-state with the emission of  $\sim 2.8$  keV Auger electrons via



Davis has designed a remarkable proportional counter with extremely low background but still capable of detecting the low-energy Auger electrons.

Tetrachloroethylene,  $\text{C}_2\text{Cl}_4$ , is used as the target material and the  ${}^{37}\text{Ar}$  is recovered with  ${}^{36}\text{Ar}$  carrier by bubbling helium through the  $\text{C}_2\text{Cl}_4$  and then freezing out the argon over charcoal. Numerous tests using artificially produced  ${}^{37}\text{Ar}$  or  ${}^{37}\text{Ar}$  produced by means of a neutron source immersed in the  $\text{C}_2\text{Cl}_4$  have shown that the recovery technique is more

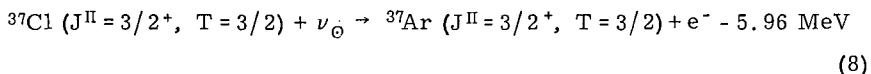
than 95% efficient. The neutron production occurs through  $^{35}\text{Cl}(\text{n}, \text{p})^{35}\text{S}$  and the secondary reaction  $^{37}\text{Cl}(\text{p}, \text{n})^{37}\text{Ar}$ . This secondary reaction can also be stimulated by protons produced in the interactions of cosmic ray muons so that the observations must be carried out at considerable depth to shield out these muons. In the observations under discussion here Davis has used a 100 000-gal tank of  $\text{C}_2\text{Cl}_4$  located at a depth of  $\sim 1$  mile in the famous Homestake Gold Mine at Lead, South Dakota. Davis calls the installation the Brookhaven Solar Neutrino Observatory. Previously he had used 1000 gal of  $\text{C}_2\text{Cl}_4$  under the Savannah River Reactor and found that



and in this way shown that antineutrinos are not identical to neutrinos.

### 3. PREDICTIONS FROM SOLAR MODELS

In my 1960 paper referred to above I pointed out that the neutrinos from  $^8\text{B}$  would greatly enhance the probability of detection in observations such as those envisaged by Davis because of their relatively high energy with a continuum maximum of 14 MeV. This suggestion became even more attractive when, in 1964, Bahcall pointed out that these neutrinos had more than enough energy to produce the state in  $^{37}\text{Ar}$  which is the isobaric spin counterpart of the ground-state of  $^{37}\text{Cl}$  ( $J^{\pi} = 3/2^+$ ,  $T = 3/2$ ). At the time the energy excitation of this state was not known but it was subsequently located at 5.15 MeV. The matrix element for the super-allowed transition



is approximately equal to 3 (unity for each excess neutron in  $^{37}\text{Cl}$ ) and can be accurately calculated. It is thus considerably greater than the known value for reaction (4) derived from the electron capture rate for  $^{37}\text{Ar}$  in its ground-state. Other excited states of  $^{37}\text{Ar}$  also contribute, but in lesser amounts. Thus, although the threshold for Eq.(8) is much greater than that for Eq.(4), Bahcall showed that the over-all detection cross-section was increased by a factor of 17. It was this more than anything else which convinced the powers that be to give Davis the go-ahead for the construction of his neutrino observatory. At this point we must emphasize that the matrix elements and Q-values for reactions (4) and (8) are well established and thus the cross-section for  $^{37}\text{Cl}$  detection of neutrinos of any energy can be calculated with precision.

The importance of the  $^8\text{B}$  neutrino flux is of course dependent on the detailed rates of the nuclear reactions in Tables I and II and in particular whether the sun operates on the proton-proton chain or the CNO bi-cycle. A long series of experimental measurements of the rates of all the reactions involved has indicated that the proton-proton chain is at least ten times as effective in the sun as the CNO bi-cycle. There is one important exception to this statement in that the basic proton-proton reaction rate must be calculated theoretically using empirical data on

the axial vector weak interaction constant, the phase shift for low-energy proton-proton scattering and the properties of the ground-state of the deuteron. It should be sufficient to note at this point that Bethe and Critchfield and Salpeter have put this calculation on a solid theoretical foundation and Bahcall and May have recently reviewed all details of the theory and up-dated the numerical data, with the result that the rate of this reaction is also well established empirically, albeit indirectly.

Until recently there have remained some lingering doubts concerning the rate of  $^{14}\text{N} + \text{p} \rightarrow (^{15}\text{O}) \rightarrow ^{15}\text{O} + \gamma$  which is the slowest and thus the governing reaction in the CN part of the CNO bi-cycle. If an excited state of  $^{15}\text{O}$  fell within the effective thermal range ( $\sim 25$  keV) for  $^{14}\text{N} + \text{p}$ , the stellar rate of the reaction would be markedly enhanced without increasing the observed rates which are performed measured at higher energies ( $\gtrsim 100$  keV). However, in 1967, Hensley laid this ghost once and for all by showing that the last possible candidate state for resonance at thermal energies in  $^{14}\text{N} + \text{p}$  actually was bound by 21.6 keV and in any case could be formed in its high-energy wing only by a d-wave ( $\ell = 2$ ) interaction. The effect of this state in the stellar thermal range is negligible and the extrapolation from the higher-energy continuum is valid. This extrapolation yields a relatively slow rate such that the CNO bi-cycle competes successfully with the proton-proton chain only in stars somewhat more massive than the sun and with central temperatures some  $5 \times 10^6$ °K greater.

Disregarding the above strong conclusions, let us calculate the predictions for the solar neutrino observations by Davis on the assumption that we are all wrong and the sun does operate on the CNO bi-cycle. Such a calculation is instructive in its simplicity. The conversion of protons into helium releases  $4 \times 10^{-5}$  erg and is always accompanied by the emission of two neutrinos whether the two weak interactions involve two positron emissions (the most probable), one positron emission and one electron capture, or two electron captures (rare). The effective energy losses via the neutrinos are only a few per cent with one exception to be discussed later. Thus  $2 \times 10^{-5}$  erg emerges with each neutrino from the sun and the known bolometric light flux at the earth,  $2 \text{ cal cm}^{-2} \text{ min}^{-1} = 1.2 \times 10^6 \text{ erg cm}^{-2} \text{ s}^{-1}$ , can be used to give the total neutrino flux, namely  $1.2 \times 10^6 / 2 \times 10^{-5} = 6 \times 10^{10} \nu \text{ cm}^{-2} \text{ s}^{-1}$ .

When the CNO bi-cycle predominates, then one of the neutrinos is from the decay of  $^{18}\text{N}$  and the other from  $^{15}\text{O}$ , neglecting the rare  $^{17}\text{F}$  cases. For these neutrinos Bahcall has found the average absorption cross-section to be  $5 \times 10^{-46} \text{ cm}^2$  so that  $\Sigma \phi_\nu \sigma_\nu = 3 \times 10^{35}$  events  $\text{s}^{-1}$  per  $^{37}\text{Cl}$ -nucleus for the CNO bi-cycle. The result is essentially model independent as long as the bi-cycle predominates.

It will be abundantly clear that the expected value for  $\Sigma \phi_\nu \sigma_\nu$  from the proton-proton chain is very model dependent. The run of density and temperature in a given solar model will determine the relative rates of the various reactions in the chain and thus the competition between  $^3\text{He}(\tau, 2\text{p})^4\text{He}$  and  $^3\text{He}(\alpha, \gamma)^7\text{Be}$  and between  $^7\text{Be}(e^-, \nu)^7\text{Li}$  and  $^7\text{Be}(p, \gamma)^8\text{B}$ . The relative numbers of neutrinos from the proton-proton reaction, from  $^7\text{Be}$  and from  $^8\text{Be}$  are in turn dependent on this competition. Since the neutrinos from these three sources have quite different energy spectra and thus different detection cross-sections, the predicted  $\Sigma \phi_\nu \sigma_\nu$  is quite model dependent. The proton-proton neutrinos have maximum energy

equal to 420 keV which is 394 keV below the threshold for reaction (4). The  $^7\text{Be}$  neutrinos exceed the threshold by only 48 keV and the absorption cross-section is quite small,  $3 \times 10^{-46} \text{ cm}^2$ . The  $^8\text{B}$  neutrinos can trigger both reaction (4) and reaction (8) and the cross-section is relatively large,  $1.4 \times 10^{-42} \text{ cm}^2$ .

Bahcall and his collaborators have made a continuing theoretical study of hydrogen burning in the sun and have from time to time revised the solar neutrino flux values as improvements in solar models and new atomic and nuclear data became available. It is interesting to follow their predictions before, during, and after Davis performed his first observations. In what follows I shall list the predicted  $^8\text{B}$  neutrino flux which contributes the major part of  $\Sigma \phi_\nu \sigma_\nu$  and also the predicted values for the total  $\Sigma \phi_\nu \sigma_\nu$  in such a way as to show trends and causes over the past few years. In 1966, using the accepted heavy element abundance by mass,  $Z = 2.0\%$ , and the accepted cross-section factor for the  $^3\text{He} + ^3\text{He}$  reaction,  $S_{33} = 1.1 \text{ MeV-barn}$ , it was found that  $\phi_\nu(^8\text{B}) = (2.1 \pm 1) \times 10^7 \nu \text{ cm}^{-2} \text{ s}^{-1}$  and  $\Sigma \phi_\nu \sigma_\nu = (3.0 \pm 1.5) \times 10^{-35} \nu \text{ s}^{-1}$ . The standard deviations are my own estimates. In 1967 it was shown that  $S_{33} = 5.0 \text{ MeV-barn}$  with the result that less production of  $^7\text{Be}$  via  $^3\text{He}(\alpha, \gamma)^7\text{Be}$  occurred and thus less production of  $^8\text{B}$ . The upshot was:  $\phi_\nu(^8\text{B}) = (1.6 \pm 0.8) \times 10^7$  and  $\Sigma \phi_\nu \sigma_\nu = (2.3 \pm 1.2) \times 10^{-35}$ . The year 1968 has brought two important revisions: new Danish results for the lifetime of the neutron have lowered the value to  $10.8 \pm 0.16 \text{ min}$ , well outside the error range of the previously accepted Russian value which was  $11.7 \pm 0.3 \text{ min}$ ; Lambert and Warner of Oxford showed that the heavy element mass fraction was much less than previously thought,  $Z = 1.5\%$ , on the basis of their analysis of their solar spectra. This result has been anticipated by Hoyle in an unpublished analysis of existing data in the summer of 1967. The effects of these changes will now be described.

The decay rate of the neutron is related to the weak interaction coupling constants by  $\lambda_n = \tau_n^{-1} = 0.693 t_n^{-1} \propto 3G_A^2 + G_V^2$  with  $G_A$  the axial vector constant and  $G_V$  the polar vector constant.  $G_V$  is determined by the decay rate of such Fermi transitions as  $\Delta J = 0$ ,  $\Delta \Pi = 0$  to be  $^{14}\text{O}(e^+, \nu) ^{14}\text{N}^*$ , where in the initial and final nuclear states  $J^\Pi = 0^+$ . Thus a decrease in  $t_n$  from  $11.7$  to  $10.8 \text{ min}$  increases  $G_A$ . The basic proton-proton reaction through the s-wave state, where the exclusion principle allows only  $J^\Pi = 0^+$ , to the ground-state of the deuteron,  $J^\Pi = 1^+$ , is a Gamow-Teller transition ( $\Delta J = 0, \pm 1$ ;  $\Delta \Pi = 0$ ). The rate is thus proportional to  $G_A^2$  alone and was accordingly enhanced by the change in  $t_n$ . This means that the central temperature of the sun is lowered to keep the known energy generation rate fixed. Lowering the temperature lowers the production rate for  $^8\text{B}$  markedly and this lowers  $\phi_\nu(^8\text{B})$ . This flux is also decreased when  $Z$  is decreased since lowering  $Z$  decreases the opacity of solar material and decreases the temperature gradient from the centre of the sun to the surface. Since the surface temperature is fixed by observation, the central temperature is decreased (from  $15.7$  to  $14.9 \times 10^6 \text{ }^\circ\text{K}$ ). This small decrease in the temperature in the solar interior lowers the production rate of  $^8\text{B}$  via  $^7\text{Be}(p, \gamma)^8\text{B}$  markedly. Thus at about the same time in 1968 that Davis completed the analysis of his preliminary experiments, Bahcall and collaborators found  $\phi_\nu(^8\text{B}) = (0.5 \pm 0.2) \times 10^7$  and  $\Sigma \phi_\nu \sigma_\nu = (0.75 \pm 0.3) \times 10^{-35}$ . Parenthetically it should be noted that lowering  $Z$  results in lower values of  $Y$ , the helium abundance by mass. The 1968 value is  $Y = 0.22 \pm 0.03$ .

which may be significantly lower than the "universal" values so much in the news of late. Further comments on this point are given in the next section.

#### 4. OBSERVATIONS TO DATE

The Brookhaven Solar Neutrino Observatory has been in operation for only about one year and yet Davis has already found a surprising and significant result: the observational upper limit for the solar neutrino flux is quite low indeed! Details are given in the paper mentioned before. At this point I will dwell only upon the raw data of the observations resulting from the exposure made from June 23rd to October 11th, 1967. Counting was done for 71 d after the exposure and necessary processing. In the first 35 d, 11 counts were recorded in the counter channels in which  $^{37}\text{Ar}$  decays register with reasonable probability, and in the second 36 d, three counts were recorded. The expected background counting rate for 35 or 36 d is 12 counts.

How does one analyse data involving 12, 11, 3 counts! Davis and collaborators argue with considerable justice that during the first 35 d, or the first  $^{37}\text{Ar}$  half-life, an upper limit on the solar neutrino produced counts was at most of the order of one standard deviation in 11 or 12 counts. This can be expressed after proper analysis as

$$\sum \phi_\nu \sigma_\nu \leq 0.3 \times 10^{-35} \nu \text{ s}^{-1} \text{ per } ^{37}\text{Cl-nucleus} \quad (9)$$

A comparison with the predictions previously discussed will explain the keen disappointment caused by this result.

There is another way of looking at the results. In the second  $^{37}\text{Ar}$  half-life only three counts were observed compared to 11 in the first. This difference may just be statistically significant and may represent a decay effect. It may well be that the "standard" background observed by Davis does not apply to the radioactive and carrier argon which is subject to considerable pre- and prior-exposure purification and processing. It is my impression that the upper limit given in Eq. (9) may well represent a real effect. Only time will tell! This is quite literally true since it is now generally agreed that Davis must continue counting on his exposed samples until they decay to a constant background. Only in this way will the applicable background be found. Clearly a number of exposures must be made and it will be several years before reliable statistics can be obtained. Davis must be encouraged and supported financially, but should be left alone to reap his meagre harvest of counts.

#### 5. AFTERTHOUGHTS

The observational results given in Eq. (9) can now be used to place limits on the neutrino fluxes either from the CNO bi-cycle or the proton-

proton chain. Since the mean absorption cross-section for  $^{13}\text{N}$  and  $^{15}\text{O}$  is  $5 \times 10^{-46} \text{ cm}^2$ , one finds

$$\begin{aligned}\phi_\nu(^{13}\text{N} + ^{15}\text{O}) &\leq 0.3 \times 10^{-35} / 5 \times 10^{-46} \\ &\leq 0.6 \times 10^{10} \nu \text{ cm}^{-2} \text{ s}^{-1}\end{aligned}\quad (10)$$

But this flux is then at most one-tenth of that calculated above on the basis that we were all wrong and the CNO bi-cycle predominated over the proton-proton chain in the sun. This is thus not at all the case and the observations have led to a most significant and satisfying result. The arduous nuclear experimentation has not been all in vain!

Since the theoretical predictions attribute most of the proton-proton chain effects to the neutrinos from  $^8\text{B}$  we have

$$\begin{aligned}\phi_\nu(^8\text{B}) &\leq 0.3 \times 10^{-35} / 1.4 \times 10^{-42} \\ &\leq 0.2 \times 10^7 \nu \text{ cm}^{-2} \text{ s}^{-1}\end{aligned}\quad (11)$$

The observational results thus seem definitely lower than the lowest predicted ones:  $\Sigma \phi_\nu \sigma_\nu = (0.75 \pm 0.3) \times 10^{-35}$  and  $\phi_\nu(^8\text{B}) = (0.5 \pm 0.2) \times 10^7$ . Moreover, these values will be reduced to the observed upper limits if Y and Z are reduced by about one third from the 1968 values previously quoted, to  $Y \approx 0.15$  and  $Z \approx 0.01$ . However, there is still considerable uncertainty in the cross-section factor for  $^7\text{Be}(p, \gamma)^8\text{B}$  which Bahcall and collaborators took to be  $S_{17} = 0.043 \text{ keV-barn}$ . One of our former students, Dr. Donald Kohler, has informed us that he and his collaborators would suggest a weighted average,  $S_{17} = 0.029 \text{ keV-barn}$ , based on their still unpublished measurements and the previously published measurements of Kavanagh and of Parker (subsequently revised). With this value,  $\Sigma \phi_\nu \sigma_\nu = (0.5 \pm 0.2) \times 10^{-35}$  and  $\phi_\nu(^8\text{B}) = (0.3 \pm 0.1) \times 10^7$ . There is thus no reason to push the panic button on stellar model calculations at the present time!

Another significant result emerges in that the  $^8\text{B}$  neutrino flux does not equal its maximum possible value,  $4 \times 10^7 \nu \text{ cm}^{-2} \text{ s}^{-1}$ , which could occur if the proton-proton chain always proceeded through  $^8\text{B}$ . In this case the effective energy loss in neutrinos amounts to 29%, a not insignificant figure. Hydrogen burning in stars slightly over one solar mass would effectively yield only 71% of the full energy release and the main sequence lifetimes would be considerably reduced say from ten billion years to seven billion years. Again this is not the case and there is no reason to change the current values of globular cluster lifetimes.

Bahcall and collaborators have pointed out that there is a relatively model independent result which sets a lower limit on the  $\Sigma \phi_\nu \sigma_\nu$  to be expected from the proton-proton chain. The pep-reaction



competes at a known but small calculable rate with the positron emitting reaction listed in Table I. The monoenergetic neutrinos have more than

enough energy to trigger the  $^{37}\text{Cl}$  absorption reaction. On this basis it is found that

$$\sum \phi_\nu \sigma_\nu \geq 0.03 \times 10^{-35} \nu \text{ s}^{-1} \text{ per } ^{37}\text{Cl-nucleus} \quad (13)$$

Thus if Davis eventually pushes his upper limit down by more than a factor of ten it will indeed be necessary to push the panic button!

There is one other possibility which I think merits some consideration. The diffusion time for radiation from the centre of the sun to the surface is  $\sim 10^5$  yr. Thus variations in energy generation over periods of several years may not produce observable effects at the surface. On the other hand, the neutrino flux would accurately monitor such variations. Perhaps Davis has just been unlucky so far.

I would like to point out for the benefit of the high-energy theorists that the main uncertainty in our understanding of the solar structure problem under discussion is the opacity of solar material which is so crucial in determining the interior temperature of the sun. Perhaps they could take some time off from their most commendable labours on high-energy interactions to solve this simple low-energy atomics problem which has such great astrophysical significance.

In conclusion I must pass on a most amusing anecdote related to me by Dr. Ray Davis. The miners in the Homestake Gold Mine have realized his disappointment in not obtaining positive results during the summer of 1967. They consoled him with "Don't worry Dr. Davis, after all it has been a cloudy summer".

#### A C K N O W L E D G E M E N T S

It is a pleasure to thank my cognato e cognata, Remington and Diana Varé-Olmsted for hospitality at their beach club, Aeneas' Landing, near Gaeta, Italy, where this paper was written from notes prepared for my Symposium talk.

# DISCRETE EXTRASOLAR X-RAY SOURCES\*

B. B. ROSSI

Laboratory for Nuclear Science and Physics Department,  
Massachusetts Institute of Technology,  
Cambridge, Mass., United States of America

## Abstract

DISCRETE EXTRASOLAR X-RAY SOURCES. Consideration of atmospheric absorption, cosmic ray background and flux density of radiation from celestial sources shows that the observation of (1) soft X-rays, (2) hard X-rays, and (3)  $\gamma$ -rays requires the use, respectively, of (1) rockets or satellites, (2) balloons or satellites, and (3) satellites. Discrete sources have been observed only in the spectral regions of soft and hard X-rays. The most likely processes responsible for the X-ray emission from these sources are thermal radiation from a hot plasma cloud and synchrotron radiation by relativistic electrons. About 30 discrete X-ray sources are known. They cluster round the galactic equator; hence most of them must be galactic objects. Some of the galactic X-ray sources have been identified with supernova remnants, others with peculiar point-like optical objects. These two groups of X-ray sources exhibit characteristic differences with regard to angular diameter, variability and spectrum. A few sources at high galactic latitude have been detected. These are likely to be extragalactic objects; one of them has been tentatively identified with the radio galaxy M-87.

## I

Astronomical observations with ground-based instruments are restricted to the comparatively small regions of the electromagnetic spectrum that are not absorbed by the terrestrial atmosphere (see Fig. 1). These include the optical window, the radio window and several narrow (but very important) windows in the infra-red. All the remainder of the spectrum is the exclusive domain of space astronomy<sup>1</sup>, i.e. it is accessible only to instruments carried aloft by high flying balloons, rockets or artificial satellites.

After the limitations imposed by atmospheric absorption have been removed, one still has to reckon with the attenuation of electromagnetic waves by interstellar matter, i.e. by the interstellar gas and by the interstellar dust.

Aside from discrete absorption lines, interstellar gas is completely transparent over galactic distances for all wavelengths from the radio region to the far ultra-violet (see Fig. 2). Interstellar dust, however, which is most unevenly distributed in space, blocks out completely the visible and ultra-violet radiations from certain regions of the galaxy; light from distant objects, not located behind particularly dense gas clouds, may still undergo considerable attenuation accompanied by a characteristic "reddening".

At 912 Å, i.e. at a photon energy equal to the ionization energy of hydrogen (13.5 eV), interstellar gas becomes suddenly very opaque over

\* This work was supported in part through funds provided by the National Aeronautics and Space Administration under Grant NsG-386 and in part by Contract At (30-1) 2098 from the USAEC.

<sup>1</sup>  $\gamma$ -rays of exceedingly high energy (above  $10^{13}$  eV) could, in principle at least, be detected from the ground (at mountain altitudes) through the production of air showers.

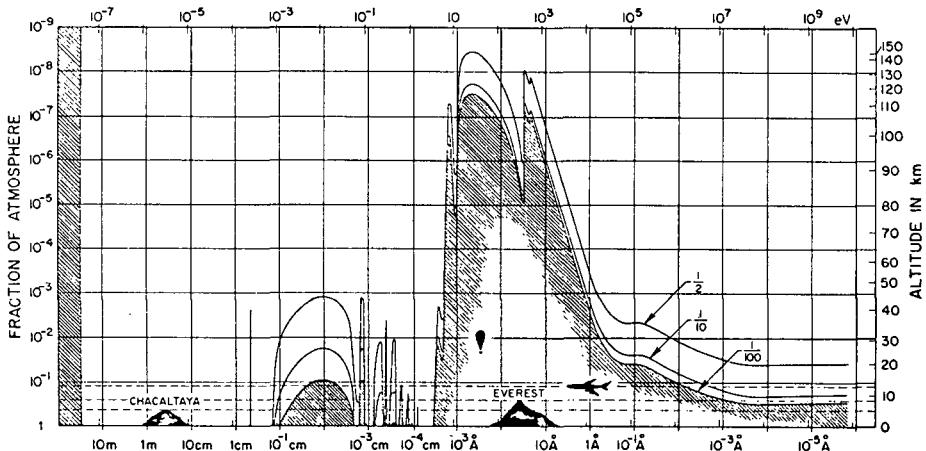


FIG. 1. The absorption of radiations of different wavelengths in the atmosphere; the lines on the graph represent the atmospheric levels where the intensities are reduced to  $1/2$ ,  $1/10$  and  $1/100$  of their initial values (from Rossi, 1965).

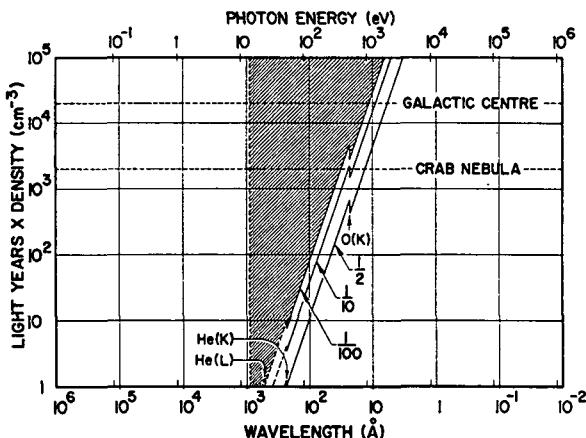


FIG. 2. The absorption of radiations of different wavelengths in interstellar gas; the ordinate is the distance (in light-years) times the average density (in atoms per  $\text{cm}^3$ ); the lines represent attenuations of  $1/2$ ,  $1/10$  and  $1/100$  (from Rossi, 1965; more recent and accurate evaluation of the absorption of interstellar gas will be found in Bell and Kingston, 1967).

galactic distances. Beyond this energy, the absorption decreases again with increasing photon energy, the general drop being interrupted by discontinuities occurring at the ionization energies of helium and of the other less abundant components of the interstellar gas. At all energies greater than 13.5 eV the absorption of dust is negligible compared with that of gas.

From Fig. 2 we see that interstellar space becomes transparent again at photon energies that vary from about one hundred electron volts for the nearby stars to about one or two thousand electron volts for the more dis-

tant galactic objects. Thus interstellar absorption quite naturally divides the electromagnetic spectrum into a low-energy region, extending from the radio waves to a photon energy of 13.5 eV, and a high-energy region, starting from X-rays of about 100 or 1000 eV and extending into the spectral range of  $\gamma$ -rays.

This paper and that of Morrison in these Proceedings will deal with astronomical observations in the high-energy region of the electromagnetic spectrum, as defined above. From the point of view of the experimental techniques, it is convenient to further subdivide this region into the following subregions.

(a) Soft X-rays extending from the low-energy cut-off due to interstellar absorption to about 15 keV

Observations of soft X-rays can only be carried out at altitudes above 100 km and thus require the use of rockets or artificial satellites. Most of the data so far available have been obtained with rockets.

The instruments most commonly used for the detection of soft X-rays have been large-area banks of proportional gas counters provided with thin windows of low-Z material. Photoelectric detectors, potentially capable of reaching photon energies lower than those accessible to gas counters, are also being developed. Collimators are usually placed in front of the detectors to provide a certain amount of angular resolution. These collimators range from relatively simple structures, consisting of thin parallel slabs or honeycomb-like cells (with angular resolutions of the order of degrees), to very sophisticated systems of grids (with angular resolutions of the order of minutes or less; see Oda, 1965).

Soft X-rays hold a special position because they are the only rays in the high-energy region of the electromagnetic spectrum that can be focussed by reflection under grazing incidence. Grazing incidence telescopes have already been built which use a reflection on a parabolic surface and a reflection on a hyperbolic surface (see Figs 3 and 4) and are capable of providing images of X-ray sources with a resolution of the order of seconds of arc. So far, they have been used only for solar observations (see Fig. 5), but it is expected that they will soon be applied to the observation of extrasolar sources.

(b) Hard X-rays extending from about 15 keV to about 0.5 MeV

Since the spectrum of high-energy photons is very steep, rocket flights do not provide sufficiently long observation time in this spectral region. Satellites, so far, have only been available to a very limited extent for X-ray astronomy. On the other hand, hard X-rays penetrate to a sufficient depth in the atmosphere to be detectable by balloon-borne instruments, which provide observation times of many hours. For these reasons, most of the existing information on hard X-rays has come from balloon observations. These observations are hampered, although not in a crucial way, by a diffuse background of X-rays arising from the interactions of cosmic rays with atmospheric gases.

The instruments most commonly used for the detection of hard X-rays have been scintillation counters. Large proportional counters with high-Z

gas filling have also been employed. Angular resolution has been obtained by means of suitable collimators.

(c) Gamma-rays extending from 0.5 MeV to several hundred MeV

The very low intensity of the radiation in this energy range rules out the use of rockets. Moreover, since the intensity of high-energy photons of celestial origin decreases with increasing energy more rapidly than that of the secondary photons from cosmic ray interactions in the atmosphere, the background problem that already plagues balloon observations of hard X-rays becomes here unmanageable. Thus  $\gamma$ -ray astronomy requires the use of satellites. The detectors used for  $\gamma$ -ray astronomy include scintillation counters, Cerenkov counters, spark chambers, solid-state counters, and combinations of such instruments. Devices designed to discriminate between photons and charged particles, which are already useful in X-ray astronomy, become absolutely essential in  $\gamma$ -ray astronomy because, in this energy range, the flux of photons is only a minute fraction of the cosmic ray flux.

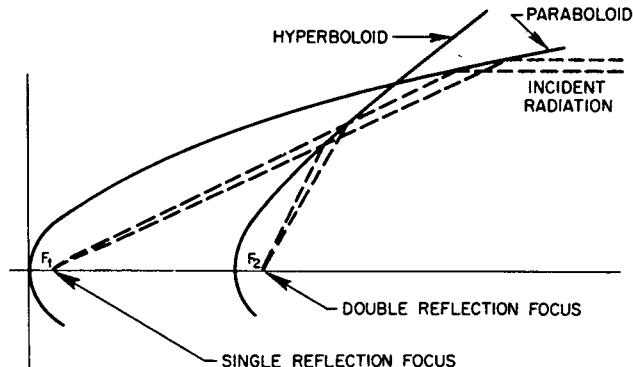


FIG. 3. Schematic diagram of a grazing incidence X-ray telescope. The image of a distant source appears in the focal plane of the hyperboloid, which intersects the axis at  $F_2$  (from Giacconi et al., 1965a).

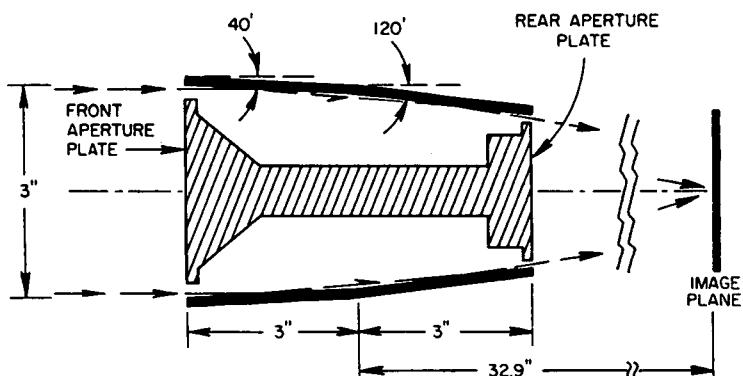


FIG. 4. Cross-section of a grazing-incidence telescope used to take X-ray pictures of the sun (from Giacconi et al., 1965b).

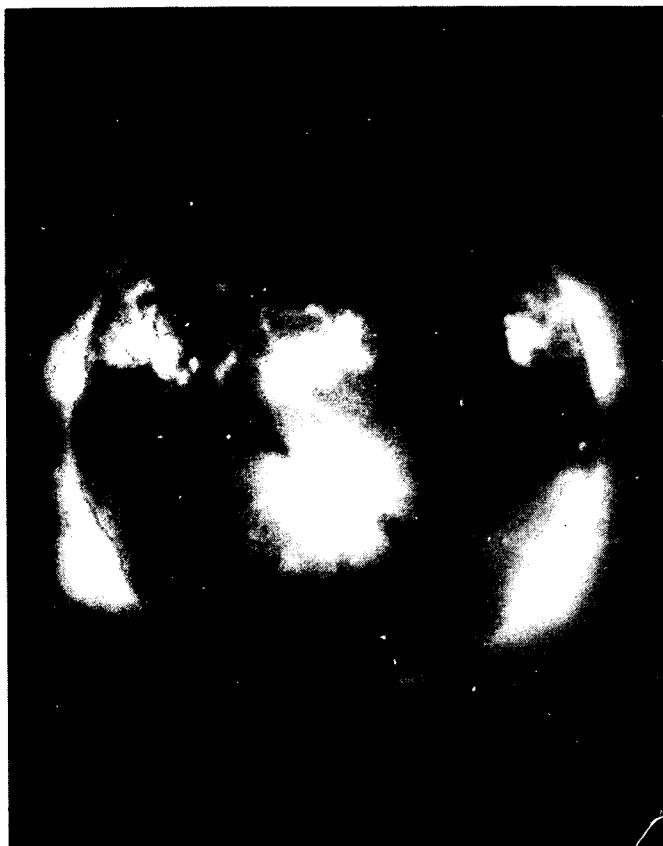


FIG. 5. Solar X-ray photograph at the time of an important 1N flare, obtained on 8 June 1968, by G.S. Vaiana, W.P. Reidy, T. Zehnpfennig, L. Van Speybroeck and R. Giacconi of American Science and Engineering.  $H_{\alpha}$  flare onset 1732 UT, end 1905 UT, accompanied by type III radio burst. Rocket flight 1741 UT to 1746 UT. Exposure time: 30 sec. Filter: 0.00015" mylar + 1100 Å Al (windows 3 - 15 Å and 44 - 60 Å). Rocket pointing accuracy: ± one second of arc (by Sounding Rocket Branch of NASA/GSFC). The rocket, sponsored by NASA, carried aboard a high-resolution telescope, a prototype of the AS&E instrument to be flown as part of the Apollo Telescope Mount mission being prepared at the Marshall Space Flight Center of NASA for the manned post-Apollo program.

Note that most of the detectors used for X-ray and  $\gamma$ -ray astronomy (such as proportional counters, scintillation counters and solid-state counters) provide some information on the spectrum of the radiation, because the average amplitude of the pulses produced by photons of a given energy is a known function of this energy. However the spectral resolution obtained from pulse-height analysis is usually rather poor. In the case of soft X-rays, grazing incidence telescopes, used in conjunction with gratings or crystals, are potentially capable of providing much higher spectral resolution.

## II

One may think of a number of processes, occurring either in the interstellar and intergalactic space, or in localized celestial objects, which are capable of producing high-energy photons. In fact, observations have revealed both a diffuse background and a number of discrete sources of such photons. The diffuse background, which is discussed by Morrison in these Proceedings, extends from soft X-rays to  $\gamma$ -rays of more than 100 MeV. Discrete sources, which form the subject of this paper, have only been detected, so far, in the spectral region of soft and hard X-rays.

Practically all phenomena capable of producing photons in this spectral region involve electrons. Here, as in other fields of astronomy, one might conveniently divide the possible sources into two groups: thermal sources and non-thermal sources.

In a thermal source, the material particles (atoms, molecules, ions, electrons) are in thermal equilibrium with one another. Photons are emitted in free-free transitions of electrons (bremsstrahlung), free-bound transitions (recombination radiation), and bound-bound transitions (giving rise to a line spectrum). If the temperature is sufficiently high (several times  $10^7$  °K at least), a large fraction of the photons belong to the spectral region of X-rays.

The spectrum of the emitted radiation depends in an essential manner on the density of the source. In the extreme case of a source sufficiently dense to be completely opaque to all rays arising within the source itself, the spectrum obeys Planck's law. At the moment there is no evidence that any of the observed X-ray sources are of this type.

The opposite extreme case is that of a source sufficiently thin to be practically transparent to its own radiation. In this case bremsstrahlung (which represents the most effective emission process at the temperatures considered here) gives rise to an exponential spectrum of photons, i.e. to a spectrum of the form

$$j(h\nu) d(h\nu) = \text{const} \times e^{-\frac{h\nu}{kT}} \quad (1)$$

where T is the temperature and  $j(h\nu) d(h\nu)$  represents the energy flux in the range of photon energies between  $h\nu$  and  $h\nu + d(h\nu)$ . As we shall see, some at least of the celestial X-ray sources may well be thermal sources approaching this limiting case. Because of the high temperature and low density, the source material will be in the form of gases, practically one-hundred percent ionized ("hot plasma cloud").

In a non-thermal source, the X-ray emission is due to supra-thermal electrons, i.e. to electrons with energies much greater than the thermal energy of the material particles in the medium. There is evidence that supra-thermal electrons are indeed produced in celestial phenomena, such as solar flares.

Supra-thermal electrons may generate X-rays by bremsstrahlung, by inverse Compton effect (i.e. by collisions with photons of lower energy, such as photons of star light or radio waves) and by synchrotron radiation (or magnetic bremsstrahlung).

It seems that inverse Compton effect is negligible with respect to synchrotron radiation, except, perhaps, as a source of the diffuse background radiation; thus I shall not discuss it here.

Synchrotron radiation is definitely a possible mechanism of X-ray emission for at least some of the discrete sources. I recall that the synchrotron spectrum emitted by an electron of energy  $E$  is sharply peaked at an energy  $(h\nu)_{\max}$  given by

$$(h\nu)_{\max} = 0.44 h\nu_{\text{cycl}} \left( \frac{E}{mc^2} \right)^2 \quad (2)$$

where  $\nu_{\text{cycl}}$  is the cyclotron frequency and  $m$  the electron mass. The energy  $h\nu_{\text{cycl}}$ , measured in electron volts, is related to the component of the magnetic field perpendicular to the electron's velocity,  $H_{\perp}$ , measured in Gauss, by the equation

$$h\nu_{\text{cycl}} = 1.1 \times 10^{-8} H_{\perp} \quad (3)$$

Equation (3) shows that, with the magnetic fields that one might reasonably expect to find in the source region, electrons of exceedingly high energy are needed to generate X-rays by synchrotron emission.

We note that, for reasons at best only partially understood, supra-thermal electrons arising from celestial phenomena appear to have spectra that may be adequately described by a power law with slowly-varying exponent. One then finds that photons produced by such electrons through the synchrotron process have also a spectrum of the same type (although with a different exponent), i.e. a spectrum represented by the equation

$$j(h\nu) d(h\nu) = \frac{\text{const}}{(h\nu)^{\alpha}} d(h\nu) \quad (4)$$

where  $\alpha$  is a slowly-varying function of  $h\nu$ .

Bremsstrahlung by supra-thermal electrons is another possible source of celestial X-rays. However, the following argument indicates that this process cannot contribute more than a small fraction of the X-ray flux originating from the known discrete sources, except the sun. The bremsstrahlung spectrum of electrons of energy  $E$  is essentially flat and extends from very low photon energies to  $h\nu = E$ . On the other hand, the observed X-ray spectra show a rapid decrease with increasing energy. Therefore, if the X-rays are produced via bremsstrahlung, most of the electrons responsible for their emission must have energies below 100 keV. Electrons in this low-energy range lose many times more energy by collision than by radiation. Collision losses will heat the gas, which will then become a thermal source of radiation. If the gas reaches a sufficiently high temperature, it will radiate abundantly in the X-ray region; in this case, however, the supra-thermal electrons will have only served the function of providing the energy to a thermal X-ray source of the kind considered previously. If the temperature does not exceed about 10 million degrees,

then the gas will radiate most of its energy at wavelengths longer than those of X-rays; in this case, the X-ray emission will account for only a small fraction of the total energy radiated by the source. This, in fact, is true for the sun, but not for the other discrete X-ray sources which, as we shall see, have been found to radiate most of their energy in the form of X-rays.

From the above discussion we conclude that, at least in a preliminary analysis of the observational results concerning discrete extrasolar X-ray sources, we are entitled to consider only two emission processes, i.e. thermal emission by a hot gas cloud and synchrotron radiation by electrons of very high energies.

### III

After these introductory remarks, I shall now summarize briefly some of the most significant results obtained to date.

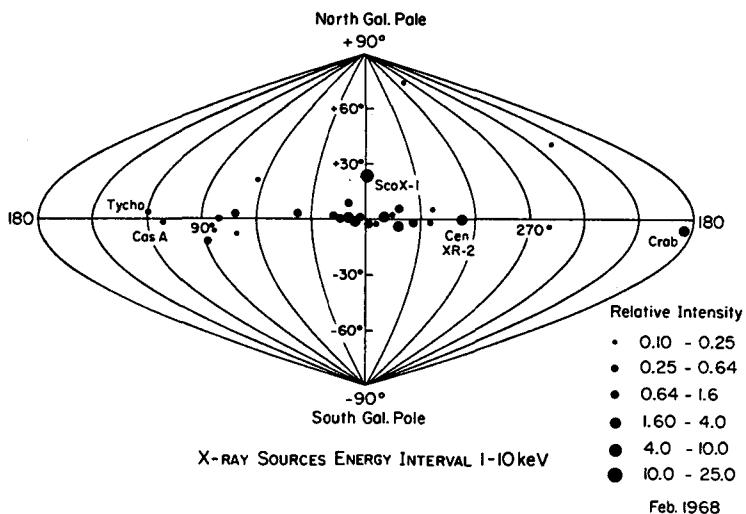


FIG. 6. Distribution of X-ray sources in the sky (galactic co-ordinates).

About 30 discrete X-ray sources are known; Fig. 6 shows their distribution in the sky, in a polar frame of reference with the equatorial plane coincident with the plane of the galaxy. The outstanding feature of this distribution is the concentration of the sources around the galactic equator, which proves that most of the sources are galactic objects. So far, only two sources at high galactic latitude have been unambiguously detected. These may well be extragalactic objects and, in fact, one of them has been tentatively identified with a radio-galaxy (see below).

Among the comparatively few galactic sources about which some measure of detailed information is available, there appear to be objects of at least two different kinds.

The objects of the first kind are remnants of supernovae. The prototype of this class is the X-ray source in the Crab Nebula (the remnant of

the supernova of 1054 AD). The reliability of the identification may be judged from Fig. 7, which shows the location of the X-ray source as reported by two different groups (Bowyer et al., 1964; Oda et al., 1967; the data of the latter group were obtained by means of a modified "modulation collimator"; see Oda, 1965) superposed to a photograph of the nebula in ordinary light. The X-ray source was found to have an angular diameter of the order of 100'', i.e. comparable to, although perhaps somewhat smaller than the angular dimensions of the visible nebula. The centre of the X-ray source and that of the visible nebula were found to be coincident, within the experimental errors.

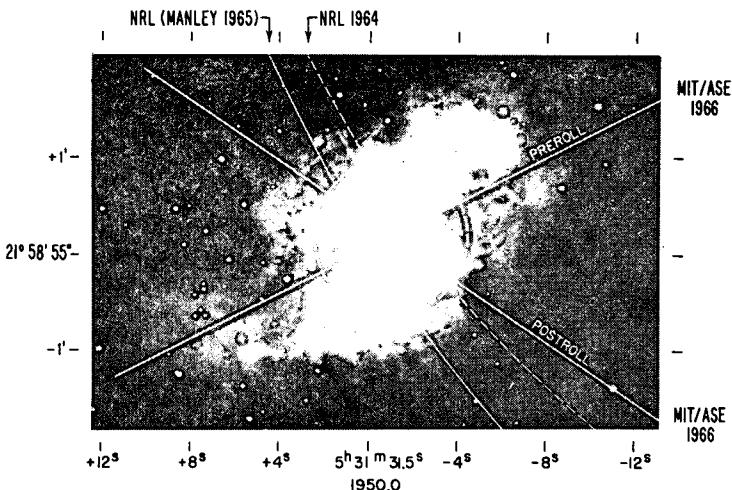


FIG. 7. Observational results on the location and size of the X-ray source in the Crab Nebula, superposed to a photograph of the nebula in ordinary light (from Oda et al., 1967). The data were obtained by Bowyer et al. (1964) who observed the occultation of the Crab by the moon, and by Oda et al. (1967), using a modulation collimator. The arc marked "NRL 1964" shows the position of the moon's limb at the time when it crossed the centre of the X-ray source, as given by Bowyer et al. The arc marked "NRL (Manley 1965)" shows the same data, corrected by Manley and Ouellette. The intersection of the "pre-roll" and "post-roll" lines is the most likely position of the centre of the source, as determined by Oda et al.; the observational errors of this determination are also indicated. The dotted circle represents the approximated dimensions of the X-ray source.

There is no clear evidence for any time-variation of the X-ray flux.

In the interval of photon energies between 2 keV and several hundred keV, the X-ray spectrum of the Crab follows closely a power law (see Fig. 8). If we plot, on the same graph, the X-ray spectrum together with the radio spectrum and the infra-red, visible and ultra-violet spectra of the Crab (exclusive of the line spectrum), and if we use some imagination to fill the rather wide gaps, we find that the whole electromagnetic spectrum of the Crab may be adequately represented by a power law with slowly-varying spectral index (see Fig. 9). Such a spectrum is characteristic of the synchrotron radiation. Moreover, from polarization measurements it is known that the low-energy portion of the spectrum is due, in fact, to synchrotron emission. It is thus natural to suppose that the same mechanism may also be responsible for the X-ray portion of the

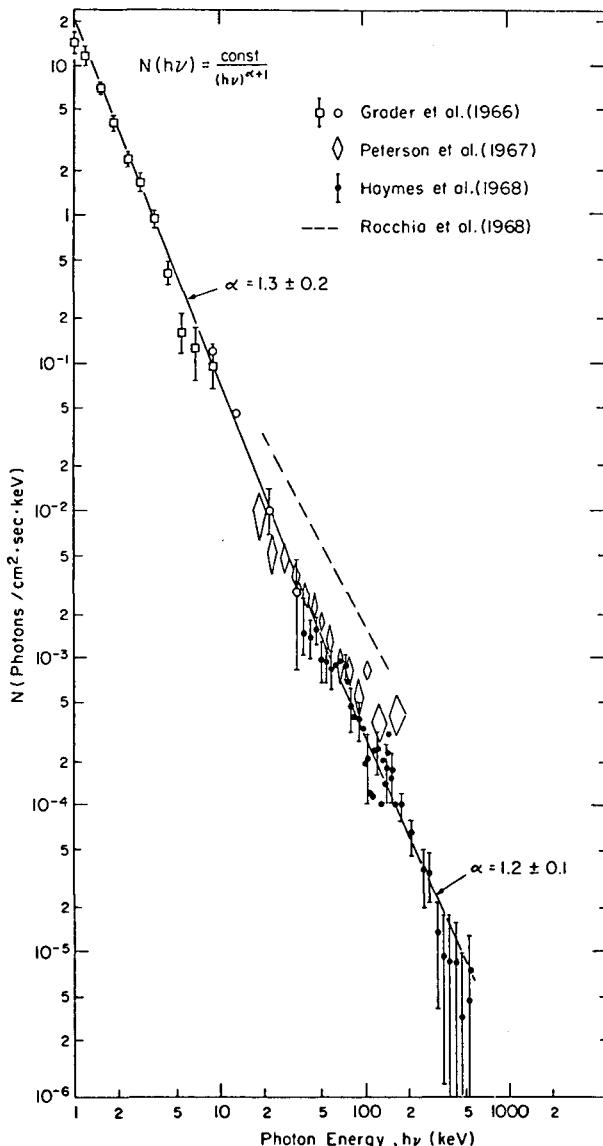


FIG. 8. Log-log plot of the X-ray spectrum of the Crab Nebula. Ordinates represent the quantity  $N = j/h\nu$ , i.e. the number of photons per  $\text{cm}^2$ , per second and per unit energy interval. The energy is measured in keV. If  $j$  obeys a power law with exponent  $\alpha$ ,  $N$  obeys a power law with exponent  $\alpha + 1$ . Experimental data are taken from Grader et al. (1966), Peterson et al. (1967), Haymes et al. (1968a), Rocchia et al. (1968).

spectrum. With a magnetic field of about  $10^{-4}$  G, such as it is usually supposed to exist in the Crab Nebula, synchrotron emission in the u.v. spectrum requires electron energies of the order of  $10^{12}$  eV. With the same field, synchrotron emission in the hard X-ray region would require electron energies of the order of  $10^{14}$  eV. Because of the rapid increase

of synchrotron losses with increasing energy, these electrons dissipate their energy at a much faster rate than those responsible for the synchrotron emission in the low-energy portion of the spectrum. In fact, their lifetime turns out to be about 10 yr, which implies that they would have to be accelerated continuously. On the other hand, it has been pointed out that the magnetic field in the Crab Nebula is probably not uniform, and that X-ray emission may occur only in the regions of enhanced magnetic field. The electrons responsible for the X-ray spectrum would then have a lower energy than that indicated above. They would also have a longer mean life, since they would spend only a fraction of their time in the region where strong synchrotron emission occurs; therefore they could have been produced at the time of the initial supernova explosion (see Apparao, 1967).

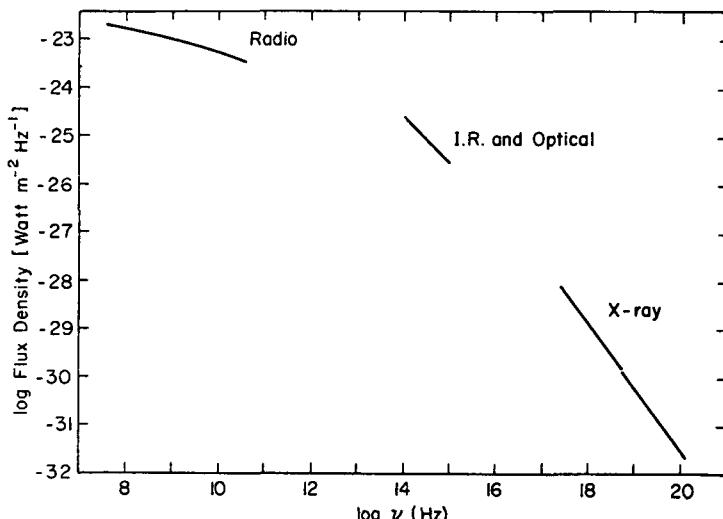


FIG. 9. Electromagnetic spectrum of the Crab Nebula from radio waves to hard X-rays. Abscissa is the logarithm of the frequency  $\nu$  (in hertz); ordinate is the logarithm of  $j$  (in watt per  $\text{m}^2$  and per unit frequency interval).

Although the shape of the spectrum suggests synchrotron radiation, it is by no means certain that this is indeed the mechanism responsible for the X-ray flux. Morrison, for example, prefers the assumption that the X-ray emission is of thermal character, originating from a hot plasma cloud. While an optically thin isothermal cloud produces an exponential spectrum, a non-isothermal cloud with proper temperature distribution may well produce an X-ray spectrum simulating a power law (Sartori and Morrison, 1967).

One may think of two experiments that would help deciding between the two hypotheses outlined above. The first is a search for a polarization of the X-ray emission (a positive result would support the synchrotron hypothesis). The second is a series of X-ray pictures, taken with different filters (if the thermal assumption is correct, the cloud should appear different in different pictures since the radiation from the hotter regions of

the non-isothermal source would be harder than the radiation from the cooler regions). Both of these observations may become possible in the near future.

The electromagnetic spectrum shown in Fig. 9 implies that the X-ray emission of the Crab (in the spectral interval from 1 to 200 keV) is a few times greater than its optical emission, and about 1000 times greater than its radio emission (for  $\lambda \gtrsim 1$  cm). Taking the distance of the Crab as 4000 light-years, its total energy output in X-rays (from 1 to 200 keV) is about  $2 \times 10^{37}$  erg/sec $^{-1}$  (or 5000 times the total energy output of the sun).

Several other galactic X-ray sources have been tentatively identified with remnants of supernovae, among them a source in Cas A (Byram et al., 1966). The X-ray flux from this object is about nine times smaller than that of the Crab. On the other hand, its distance is about three times greater. Thus the X-ray outputs of the Cas A and the Crab appear to be similar (Cas A, however, has a much greater intrinsic luminosity than the Crab in the radio region). Nothing is known as yet concerning the dimensions of the X-ray source in Cas A, nor about its spectrum.

While the identification of the Crab is practically beyond doubt, and the identification of Cas A is quite convincing, all other identifications of supernova remnants with X-ray sources are, individually, very doubtful. However, considerations of a statistical character, based on the density of supernovae, the density of X-ray sources, and the observational uncertainty in their positions, suggest that several additional X-ray sources might be supernova remnants (Poveda and Woltjer, 1968).

Moreover, from what is known today, it appears quite possible that all supernova remnants are X-ray sources of a strength, say, within a factor of 10 of that of the Crab (the absence of an observable X-ray flux from the Kepler supernova, for example, may be explained simply by the large distance of this object - 30 000 light-years).

#### IV

While we are still uncertain as to the fraction of the observed X-ray sources that are supernova remnants, we know for sure, as I already pointed out, that some of the galactic sources are objects of an entirely different nature.

The prototype of these objects is Sco X-1, the brightest source of soft X-rays in the night's sky, whose discovery in 1962 marked the beginning of X-ray astronomy (Giacconi et al., 1962).

Sco X-1 and the Crab Nebula are the two X-ray sources whose positions are presently best known. Observations by means of a modulation collimator have placed Sco X-1 within one or the other of the two rectangles,  $1' \times 2'$  in size, drawn in Fig. 10 over a star map (Gursky et al., 1966b). The brightest star within the combined area of the two rectangles is a point-like variable object of average magnitude 12.5 (arrow), whose peculiar properties, upon which I shall return later, had not been noticed before. If, as is practically certain, this object is the optical counterpart of Sco X-1 (Sandage et al., 1966), one finds that the X-ray flux from Sco X-1 is about 1000 times greater than its energy flux in the visible. The radio flux from Sco X-1 is exceedingly small (about 20 000 times

smaller than that of the Crab at centimetre wavelengths, according to Andrew and Purton, 1968). Thus, in a very real sense, Sco X-1 may be regarded essentially an X-ray star.

Unlike the X-ray source in the Crab, which is definitely extended, the source in Scorpio has angular dimensions below the resolution limit of the instruments used to date ( $20''$ ; Gursky et al., 1966a).

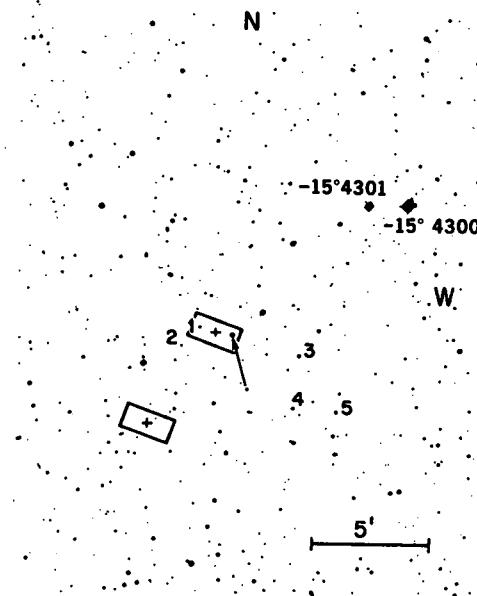


FIG. 10. Photograph of the region containing Sco X-1, reproduced from the Palomar Sky Survey prints. The X-ray source lies within one of the two rectangles. The arrow indicates the star that is regarded as the optical counterpart of the X-ray source (from Sandage et al., 1966).

The X-ray spectrum of Sco X-1 is much softer than that of the Crab. It appears to follow an exponential law and thus suggests that the X-ray source may be a hot thin plasma cloud. In Fig. 11 the spectrum of Sco X-1 is compared with the spectrum of the Crab. The solid curve represents the exponential spectrum due to a thin cloud at a temperature of about  $5 \times 10^7$  °K ( $kT = 4.4$  keV).

Sco X-1 is variable, not only in the visible but also in the X-ray portion of the spectrum (although no clear information is yet available on the correlation between the variations in the two spectral regions). For example, several balloon flights performed with similar detectors, sensitive to X-rays of energy greater than 20 keV, showed that, during the period from March to May 1967, the intensity was about one fourth of that observed in June 1965; during the following month the intensity went up again by a factor of four (Peterson and Jacobson, 1966; Lewin et al., 1967; Overbeck and Tananbaum, 1968).

Not only the intensity but also the spectrum appears to be variable. For example, the soft X-ray spectra observed in two rocket flights per-

formed in May and in September 1967 were characterized by  $kT = 7$  keV and  $kT = 4$  keV respectively (Chodil et al., 1968b).

The most striking variation in the X-ray emission of Sco X-1 was observed during a balloon flight carried out on 15 October 1967 (Lewin et al., 1968b), when, in a period of about 10 min, the intensity of hard X-rays was found to increase by about a factor of four. The increase was immediately followed by a decrease, characterized by a time constant of about 30 min; but, unfortunately, the flight terminated before the intensity had reached again a stable value. There was also evidence that during the "flare" the X-ray spectrum was somewhat softer than before.

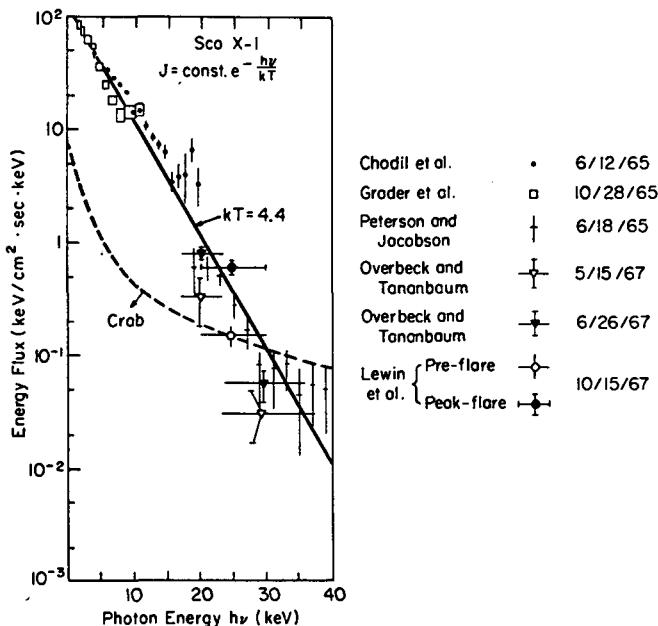


FIG. 11. Semi-logarithmic plot of the X-ray spectrum of Sco X-1 ( $J$  versus  $h\nu$ ). For the sake of comparison the spectrum of the Crab is also shown (dotted curve). The experimental data are taken from Chodil et al. (1965), Grader et al. (1966), Peterson and Jacobson (1966), Overbeck and Tananbaum (1968), Lewin et al. (1968b). The dates of the various flights are indicated.

Two other X-ray sources, i.e. Cyg X-2 and Cen X-2, appear to resemble Sco X-1, at least in some of their properties. It is thus tentatively assumed that they belong to the same category, although, admittedly, the evidence to this effect is not yet compelling.

Cyg X-2 is about 20 times less intense than Sco X-1. Its angular coordinates have been determined with an accuracy of the order of  $10''$  (Giacconi et al., 1967a). On the basis of this determination, Cyg X-2 has been identified with a variable star of average magnitude 15, as shown in Fig. 12 (Giacconi et al., 1967b). Because of the very large number of stars of brightness comparable to or greater than that of this particular object, lying within the area of uncertainty for the position of the X-ray source, one might question the reliability of the identification. The

strongest argument in favour of the proposed identification is the fact that the assumed optical counterpart of Cyg X-2 has properties that greatly resemble those of the optical counterpart of Sco X-1, and that objects of this kind are exceedingly rare.

No significant data are yet available concerning the angular dimensions of Cyg X-2.

We know that the X-ray spectrum of Cyg X-2 is softer than that of the Crab; but the measurements performed so far do not permit to establish the shape of this spectrum. If, however, we assume that the spectrum has an exponential shape like that of Sco X-1, we find for  $kT$  a value of about 3.2 keV (Chodil et al., 1967; Gorenstein et al., 1967).

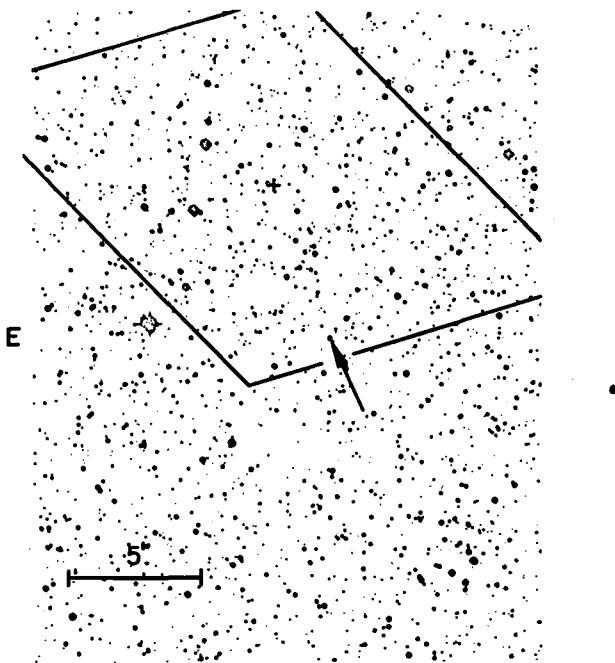


FIG. 12. Photograph of the region containing Cyg X-2, reproduced from the Palomar Sky Survey prints. The X-ray source lies within the marked area. The arrow indicates the star that is regarded as the optical counterpart of the X-ray source (from Giacconi et al., 1967b).

Sufficient data are not yet available to establish whether or not the intensity and the spectrum of Cyg X-2 are variable with time.

Cen X-2 (which lies close to the south geographic pole) has only been observed, so far, with instruments of very modest angular resolution. Therefore no significant data are available concerning its dimensions, and its angular co-ordinates are only known with a precision of the order of one degree. For this reason, any attempt to identify Cen X-2 with an optical object appears premature. On the other hand, many accurate results have been obtained concerning the spectrum and the variability of Cen X-2.

Cen X-2 was first detected in a rocket flight carried out by an Australian group on 4 April 1967 (Harries et al., 1967). Subsequently, it was seen in three additional rocket flights carried out by the same group and by two other groups during the months of April and May (Cooke et al., 1967; Francey et al., 1967; Chodil et al., 1967). On 4 April, the intensity at 2 keV was about the same as that of Sco X-1; on 18 May, the intensity in the same spectral region had dropped to about one third. In all four observations the spectrum was found to follow closely an exponential law; but the corresponding values of  $kT$  were found to decrease from about 3.7 keV on 4 April to about 1.5 keV on 10 May.

The region of the sky where Cen X-2 is located had been explored by a rocket flight carried out in October 1965, and was explored again by another rocket flight carried out in September 1967. Neither of these two rocket flights detected Cen X-2, which means that its intensity (in the soft X-ray region) was at least 10 times smaller than in May 1967 (Chodil et al., 1968a).

The situation concerning the spectrum and the variability of Cen X-2 is further complicated by a balloon observation carried out in October 1967, which revealed a hard component of the spectrum (Lewin et al., 1968a). It is not yet known whether or not this component, which does not appear to be the natural extrapolation of the spectra of the soft component observed in the various rocket flights, is a permanent feature of the spectrum of Cen X-2.

Following the identifications of Sco X-1 and Cyg X-2, many observations of these objects were made in ordinary light and in the near ultraviolet. I do not have either the time nor the specific competence to discuss the results of these observations in detail. Thus I shall only mention a few points, which seem to me particularly significant in connection with the problem concerning the nature of the objects under consideration.

The spectra of both Sco X-1 and Cyg X-2 contain a flat continuum (which makes the objects appear considerably "bluer" than ordinary stars) as well as a line emission spectrum. Various absorption lines are also visible.

In the case of Sco X-1, it seems that the optical continuum may be interpreted as due to the same hot plasma cloud responsible, supposedly, for the X-ray emission. In fact, this assumption has been shown to be capable of explaining quantitatively the results of simultaneous optical and X-ray measurements, when suitable allowance is made for attenuation of the optical spectrum in the interstellar space and in the object itself (see Chodil et al., 1968b). In the case of Cyg X-2, however, it seems that a large fraction of the optical continuum does not arise from the hot cloud (Kristian et al., 1967). While the origin of the optical continuum is still somewhat questionable, it is clear that the emission lines observed in the spectra of both Sco X-1 and Cyg X-2 cannot arise from a hot plasma cloud. In fact their Doppler width indicates temperatures of the order of only  $10^4$  K, i.e. over 1000 times smaller than that of a plasma capable of producing the observed X-rays. Therefore Sco X-1, Cyg X-1 and, presumably, Cen X-2 are complex objects containing regions at temperatures close to the surface temperatures of ordinary stars, and regions at temperatures several orders of magnitude greater.

As I already mentioned, both Sco X-1 and Cyg X-2 are variable optical objects. Intensity variations are observed both in the continuum and in

the line spectrum, but a clear correlation has not yet been established between the variations of the two components of the spectrum. Recent photo-electric measurements by Hiltner and Mook (1967) have shown that Sco X-1 undergoes aperiodic variations of about 0.5 magnitudes during times of the order of hours; moreover, when the average luminosity is high, groups of "flares" occur, each lasting several minutes, and involving variations of about 0.2 magnitudes.

The emission lines and some at least of the absorption lines of Cyg X-2 show large and time-dependent Doppler shifts, different in phase and amplitude for the different lines, which correspond to velocity changes of 200 or 300 km sec<sup>-1</sup>. This is clear evidence that Cyg X-2 is a multiple star, containing at least two, possibly more objects (Burbidge et al., 1967; Kristian et al., 1967; Kraft and Demoulin, 1967). There is still some question as to whether or not the emission lines of Sco X-1 show a time-dependent Doppler shift; in any case this shift is certainly much smaller than that observed in the line spectrum of Cyg X-2. The current view is that Sco X-1 is a double star viewed from a direction nearly perpendicular to the orbital plane.

It has been pointed out that the optical properties of Sco X-1 and Cyg X-2 are strongly reminiscent of those of old novae, which are also known to be binary objects; on the other hand, none of the objects previously classified as old novae that have been looked at with X-ray detectors were found to be X-ray emitters. Pertinent to this question are the observations of Kraft and Demoulin (1967) which show that dominant features of the spectrum of Cyg X-2 coincide with those of late-type stars.

From the absence of detectable proper motion and from the strength of the Ca II interstellar absorption line, the distance of Sco X-1 has been tentatively estimated at about 3000 light-years (Wallerstein, 1967). There is still a large uncertainty in this estimate, but 1000 light-years appears to be a lower limit. For a distance of 1000 light-years the total energy emission from Sco X-1 (which is practically all in the form of X-rays) amounts on the average to about  $6 \times 10^{36}$  erg sec<sup>-1</sup>, i.e. about 1500 times that of the sun.

Various estimates, ranging from about 1500 to about 3000 light-years, have been made for the distance of Cyg X-2 (Kraft and Demoulin, 1967; Prendergast and Burbidge, 1968; Cathey and Hayes, 1968). These estimates are based on different assumptions concerning the absolute optical magnitude of the object (as suggested by the spectral characteristics) and the amount of interstellar absorption.

Various models have been suggested for Scorpio-like objects, and some computations have been carried out to investigate whether these models are capable to explain quantitatively the very large X-ray fluxes that are observed (Shklovsky, 1967; Cameron and Mock, 1967; Prendergast and Burbidge, 1968). The basis of these models is the multiple nature of the objects in question, which, at least in the case of Cyg X-2, may be regarded as experimentally established. The mechanical energy (gravitational and kinetic) of the systems is amply sufficient to maintain the observed fluxes for long periods of time. The transformation of mechanical energy into thermal energy may occur through the gradual accretion by one partner of gases coming from the atmosphere of the other partner, a process particularly effective if the first partner is very dense (neutron

star, white dwarf) so that its surface lies at the bottom of a deep well of gravitational potential.

Clearly many more observations, both on the X-ray and on the optical spectra of the objects in question, are needed before reliable conclusions about their physical nature may be reached. I would like to point out that, at the moment, not even the thermal character of the X-ray emission can be regarded as established. Spectral measurements, with sufficient resolution to detect the emission lines to be expected from a thermal source, would provide crucial evidence on this important question. Very desirable are also simultaneous long-time observations of the X-ray and optical emissions, designed to study the correlation between the intensity and possibly the spectral changes in the two regions.

## V

To this date, a comparatively small fraction of the galactic sources has been classified into Crab-like objects (i.e. supernova remnants) and Scorpio-like objects and in most cases the classification is only tentative. It is still an open question whether or not all galactic X-ray sources belong to one or the other of the two above categories. An indication that this may not be so comes from the observations on a strong source in Cygnus (Cyg X-1) which resembles the Crab as far as the shape of the spectrum is concerned, but differs from the Crab because it is strongly variable and does not coincide with a detectable radio source.

I shall omit the still scattered data that have been obtained on several other galactic sources, the very tentative identifications that have been suggested for some of them, and the speculations concerning the association of X-ray sources with various kinds of optical objects. I would like to note, however, that the X-ray sources appear to be strung along the galactic arms. I also would like to note that, from an estimate of the average distance of the observed sources and of the fractional volume of the galactic disk which they occupy, one arrives at a very tentative figure of  $7 \times 10^{39}$  erg sec<sup>-1</sup> for the total X-ray output of our galaxy, in the spectral range from 1 to 10 Å (Friedman and Byram, 1967).

Finally I would like to point out that important advances in the study of galactic sources may be expected from future observations by means of image-forming telescopes. The great accuracy in the determination of the angular co-ordinates of the sources afforded by this instrument will undoubtedly lead to new reliable identifications. Moreover, measurements of the angular dimensions will help in establishing the character of the sources.

## VI

Very little information is available so far concerning discrete extragalactic sources. In fact, the extragalactic location may be regarded as established with some reliability only for the source in Virgo, near the north galactic pole (see Fig. 6).

The source in question was seen by two different groups in three rocket flights (Byram et al., 1966; Friedman and Byram, 1967; Bradt et al., 1967), and by a third group in a balloon flight (Haymes et al.,



FIG. 13. Photograph of M-87 (radio galaxy NGC4486), from the Mt. Wilson and Palomar collection.

1968b). Thus its existence is well established; however, its angular co-ordinates are only known with a precision of the order of one degree. The X-ray spectrum appears to be similar to that of the Crab, but the observed X-ray intensity is about 50 times smaller. Within the area of uncertainty for the position of the source lies the galaxy M-87, which is the strongest radio source in that region of the sky. For this reason, and because X-ray sources are very rare at high galactic latitudes, it appears likely that the X-ray source in Virgo coincides with M-87. If this identification is correct, from the observed X-ray flux and from the approximately known distance of M-87 (about 10 megaparsec or  $3 \times 10^7$  light-years) one finds that the energy radiated by this object in the form of X-rays is of the order of  $10^{43}$  erg sec $^{-1}$ .

In ordinary light, M-87 has the appearance of a nebulosity of several minutes in diameter, with a jet about 20" long (Fig. 13) whose optical emission is due to synchrotron radiation. It is natural to speculate that perhaps the X-ray emission (whose power is of the same order as that of the optical emission and about 70 times greater than that of the radio emission) comes from the same jet and originates from the same process.

Clearly larger detectors and/or longer times of observations, such as provided by satellites, are essential to extend the range of X-ray astronomy into the field of extragalactic objects. Future observations with improved sensitivity will be capable of detecting the X-ray emission from nearby galaxies similar to our own, such as Andromeda. Moreover, they will undoubtedly reveal many more X-ray sources associated with distant unusual extragalactic objects such as M-87. (In this connection I would like to mention that the quasar 3C273 has been claimed as an X-ray source by Friedman and Byram (1967); however the observed peak in the counting rate on which this claim is based does not appear to be clearly outside the statistical fluctuations of the diffuse X-ray background.) It is even possible that future X-ray observations may bring to light extragalactic objects of a new kind, just as past X-ray observations have revealed previously unknown types of galactic objects.

The image-forming telescope will have a no less important function in extragalactic X-ray astronomy than in galactic X-ray astronomy. Because of its freedom from background problems, the telescope may be competitive with large-area detectors in the search for very weak X-ray sources associated with specific extragalactic objects. Moreover, the telescope can determine the positions of the sources with far greater accuracy. Finally, in the case of extended objects such as M-87, the telescope can determine the spatial distribution of the X-ray emission and thus help in clarifying the emission mechanism.

#### BIBLIOGRAPHY

- ANDREW, B.H., PURTON, C.R., Nature, Lond. 218 (1968) 855.  
APPARAO, M.V.K., Proc. Indian Acad. Sci. 16 1, Sec.A. (1967).  
BELL, K.L., KINGSTON, A.E., Mon. Not. R. astr. Soc. 136 (1967) 241.  
BOWYER, S., BYRAM, E.T., CHUBB, T.A., FRIEDMAN, H., Science 146 (1964) 912.  
BRADT, H., MAYER, W., NARANAN, S., RAPPAPORT, S., SPADA, G., Astrophys. J. 150 (1967) L199.  
BURBIDGE, E.M., LYNDS, C.R., STOCKTON, A.N., Astrophys. J. 150 (1967) L95.  
BYRAM, E.T., CHUBB, T.A., FRIEDMAN, H., Science 152 (1966) 66.  
CAMERON, A.G.W., MOCK, M., Nature, Lond. 215 (1967) 464.  
CATHEY, L.R., HAYES, J.E., Astrophys. J. 151 (1968) L89.  
CHODIL, G., JOPSON, R.C., MARK, H., SEWARD, F.D., SWIFT, C.D., Phys. Rev. Lett. 15 (1965) 605.  
CHODIL, G., MARK, H., RODRIGUES, R., SEWARD, F., SWIFT, C.D., HILTNER, W.A., WALLERSTEIN, G., MANNERY, E.J., Phys. Rev. Lett. 19 (1967) 681.  
CHODIL, G., MARK, H., RODRIGUES, R., SWIFT, C.D., Astrophys. J. 152 (1968a) L45.  
CHODIL, G., MARK, H., RODRIGUES, R., SEWARD, F.D., SWIFT, C.D., TURIEL, I., HILTNER, W.A., WALLERSTEIN, G., MANNERY, E.J. (1968b) to be published.  
COOKE, B.A., POUNDS, K.A., STEWARDSON, E.A., ADAMS, D.J., Astrophys. J. 150 (1967) L189.  
FRANCEY, R.J., FENTON, A.G., HARRIES, J.R., McCracken, K.G., Nature, Lond. 217 (1967) 773.  
FRIEDMAN, H., BYRAM, E.T., Science 158 (1967) 257.

- GIACCONI, R., GURSKY, H., PAOLINI, F.R., ROSSI, B.B., Phys. Rev. Lett. 9 (1962) 439.
- GIACCONI, R., HARMON, N.F., LACEY, R.F., SZILAGYI, Z., J. opt. Soc. Am. 55 (1965a) 345.
- GIACCONI, R., REIDY, W.P., ZEHNPENNIG, T., LINDSAY, J., MUNNEY, W.S., Astrophys. J. 142 (1965b) 1274.
- GIACCONI, R., GORENSTEIN, P., GURSKY, H., WATERS, J.R., Astrophys. J. 148 (1967a) L119.
- GIACCONI, R., GORENSTEIN, P., GURSKY, H., USHER, P.D., WATERS, J.R., SANDAGE, A., OSMER, P., PEACH, J.V., Astrophys. J. 148 (1967b) L129.
- GORENSTEIN, P., GIACCONI, R., GURSKY, H., Astrophys. J. 150 (1967) L85.
- GRADER, R.J., HILL, R.W., SEWARD, F.D., TOOR, A., Science 152 (1966) 1499.
- GURSKY, H., GIACCONI, R., GORENSTEIN, P., WATERS, J.R., ODA, M., BRADT, H., GARMIRE, G., SREEKANTAN, B.V., Astrophys. J. 144 (1966a) 1249.
- GURSKY, H., GIACCONI, R., GORENSTEIN, P., WATERS, J.R., ODA, M., BRADT, H., GARMIRE, G., SREEKANTAN, B.V., Astrophys. J. 146 (1966b) 310.
- HARRIES, J.R., McCACKEN, K.G., FRANCEY, R.J., FENTON, A.G., Nature, Lond. 215 (1967) 38.
- HAYMES, R.C., ELLIS, D.V., FISHMAN, G.J., KURFESS, J.D., TUCKER, W.H., Astrophys. J. 151 (1968a) L9.
- HAYMES, R.C., ELLIS, D.V., FISHMAN, G.J., GLENN, S.W., KURFESS, J.D., Astrophys. J. 151 (1968b) L131.
- HILTNER, W.A., MOOK, D.E., Astrophys. J. 150 (1967) 851.
- KRAFT, R.P., DEMOULIN, M.H., Astrophys. J. 150 (1967) L183.
- KRISTIAN, J., SANDAGE, A., WESTPHAL, J.A., Astrophys. J. 150 (1967) L99.
- LEWIN, W.H.G., CLARK, G.W., SMITH, W.B., Astrophys. J. 150 (1967) L153.
- LEWIN, W.H.G., CLARK, G.W., SMITH, W.B., Astrophys. J. 152 (1968a) L49.
- LEWIN, W.H.G., CLARK, G.W., SMITH, W.B., Astrophys. J. 152 (1968b) L55.
- ODA, M., Appl. Optics 4 (1965) 143.
- ODA, M., BRADT, H., GARMIRE, G., SPADA, G., SREEKANTAN, B.V., GURSKY, H., GIACCONI, R., GORENSTEIN, P., WATERS, J.R., Astrophys. J. 148 (1967) L5.
- OVERBECK, J.W., TANANBAUM, H.D. (1968) to be published.
- PETERSON, L.E., JACOBSON, A.S., Astrophys. J. 145 (1966) 962.
- PETERSON, L.E., JERDE, R.L., JACOBSON, A.S., AIAA Journal 5 (1967) 1921.
- POVEDA, A., WOLTJER, L., Astr. J. 73 (1968) 65.
- PRENDERGAST, K.H., BURBIDGE, G.R., Astrophys. J. 151 (1968) L83.
- ROCHCHIA, R., ROTHENFLUG, R., BOCLET, D., DUROUCHOUX, Ph. (1968) to be published.
- ROSSI, B., Space Res 5, Proc. 5th Int. Symp. Space Science, Florence, 1964, North-Holland Publishing Co., Amsterdam (1965).
- SANDAGE, A.R., OSMER, P., GIACCONI, R., GORENSTEIN, P., GURSKY, H., WATERS, J., BRADT, H., GARMIRE, G., SREEKANTAN, B.V., ODA, M., OSAWA, K., JUGAKU, J., Astrophys. J. 146 (1966) 316.
- SARTORI, L., MORRISON, P., Astrophys. J. 150 (1967) 385.
- SHKLOWSKY, I.S., Astrophys. J. 148 (1967) L1.
- WALLERSTEIN, G., Astrophys. Lett. 1 (1967) L31.



# DIFFUSE RADIATION IN THE HIGH-ENERGY REGION

P. MORRISON

Massachusetts Institute of Technology,  
Cambridge, Mass., United States of America

## Abstract

DIFFUSE RADIATION IN THE HIGH-ENERGY REGION. 1. Introduction; 2. The power-law X-ray background; 3. The thermal X-ray background -(if any); 4. Gamma rays; 5. Higher energy; 6. Other low-energy pools; 7. Other radiations.

## 1. INTRODUCTION

Professor Rossi has split the electromagnetic spectrum into two domains, the division coming at the energy where continuum states are made available in the most abundant of atoms, that is, above one Rydberg, 13.6 eV/photon. I shall aim at the region beyond there. But it turns out that the plausible if uncertain theories for what we observe in fact require that we also know the radiation below that splitting point, for it is likely that such photons are in fact the origin of some we observe above the hydrogen K-edge.

One may begin, therefore, by examining the situation with respect to all the components of radiation which do not appear to arise from some individual astronomical object, the so-called diffuse radiation. (For, so far, no discrete source of photons above the X-ray range is known; Professor Rossi has already treated the discrete X-ray sources.) Naturally, this property may not be genuine; the apparent continuity in direction may one day be resolved, but in general the absence of anisotropy implies that the sources are either very close — within the galactic disc or even in the solar system near the sun — or well beyond the galaxy. In that case they may belong to cosmology, the large scale properties of the world, or at least they represent phenomena of importance in galaxies or their clusters. That is the unifying principle which defines my present topic.

One may tabulate the properties, always speaking rather broadly and approximately of such components of the radiation. Most important is the m. f. p., then the energy density locally seen for a component (if the m. f. p. is large, the total energy must be large or the sources very close), then the spectral shape, then the signs of contribution from non-isotropic sources, if any. These are the parameters with which we must deal.

It is important to note the present uncertainty in the material content of the universe. The observable matter in galaxies, making stars and their associated gas and dust, amounts with reasonable accuracy (the Hubble constant of 100 km/sec/megaparsec —  $3 \times 10^{17}$  sec) to a mean density of 3 to  $6 \times 10^{-31}$  g/cm<sup>3</sup> = 200 to 400 eV/cm<sup>3</sup>. This is the minimal driving energy of it all. If the universe were closed, the critical gravitational requirement would be for  $2 \times 10^4$  eV/cm<sup>3</sup>.

Since neutral hydrogen can hardly be present in the latter density, or it would have been seen in absorption, the possibility of closing the universe

TABLE I. WHAT THERE IS IN SPACE . . . . . so far

Type	Observation or theory	m. f. p. (in $R_{\text{univ}}$ )	Directional dist'n	Energy density (eV/cm <sup>3</sup> )	Photon or particle energy	Origin
<u>Non-electromagnetic</u>						
1. Matter in galaxies	obs	-	isotropic and clumpy	200 to 400	1 Gev - $\frac{1}{2}$ MeV	God?
2. Unseen matter	th'y closure	-	either as 1 or isotropic?	20 000	as 1	as 1
3. Gravitons (thermal)	th'y	big	isotropic	0.1	1 MeV	fireball
4. Neutrinos thermal stellar	th'y obs (?)	big	as 3 - as 1	0.1 - 30 MeV	1 MeV - 1 MeV	fireball-stars
5. Cosmic ray protons	obs?	ca R?	isotropic to 10%	1 MeV	$>10^8$ GeV	explosive galaxies
6. CR electrons	th'y	less than R	perhaps like 1	100 $\mu$ eV	0.1 to 10 GeV	like 5 but in small B fields ( $1/5 \mu G$ )
<u>Photons</u>						
7. Radio noise	obs	big	like 1 plus strong galaxy component	1 $\mu$ eV	1 $\mu$ eV	distant sources + galaxy
8. Microwave	obs	big	isotropic to 1/500	0.4(?)	1 MeV	fireball?
9. Deep i. r.	spec'n	big	like 1	$0.1 < U < 1$ ?	$10^{-2}$ to 0.1	explosive sources, dust?
10. Visible and near visible	obs	big	like 1	0.3 locally, $10^{-2}$ on average	3 eV	galaxies and explosive sources
11. Soft X-ray	obs	stopped by disc	to pole	50-100 $\mu$ eV	100-1000 eV	from 2?
12. Hard X-ray	obs	big	isotropic	100 $\mu$ eV	1 to 500 keV	by 6?
13. Gamma ray	obs	big	disc plus isotropic?	few $\mu$ eV	100 MeV	disc CR + 6?

Space becomes opaque to photons  $10^{15}$  eV or so by photon-photon pair production (m. f. p.  $/R_{\text{univ}} \ll 1$ )

depends either on plasma or on exotic items like pebbles, neutrinos, gravitons, etc. Thus, one has a special interest in soft X-rays, a sign of plasma (see Table I).

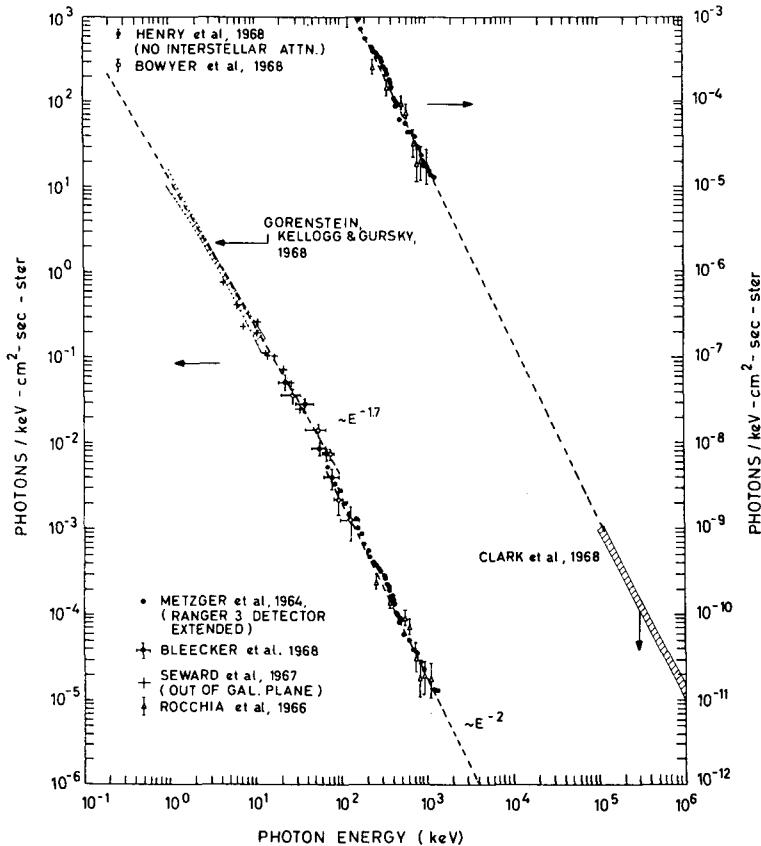


FIG. 1. Measurements made so far.

## 2. THE POWER-LAW X-RAY BACKGROUND

Figure 1 sets forth the measurements made so far in the region from a couple of keV up to about 1 or 2 MeV. The uniform power law, familiar from CR particles and from the normal spectra of radio sources, holds here, with the proviso that these are not such excellent measurements. Reliance on the 'Ranger' data (Arnold et al.) is particularly strong. Let us however assume the broadly correct nature of the case.

The power law spectrum led immediately to the idea that these might be made by Compton recoil on a supply of photons, say in the galactic halo,

where there are plenty of CR electrons — with or without a radio halo. The familiar relations are

$$\frac{dI(\nu)}{d\nu} = K_{n_0} U T^{(m-3)/2} \nu^{(1-m)/2} ; \quad \frac{dn(\gamma)}{d\gamma} = \frac{n_0}{\gamma^m}$$

$$h\nu_T \approx \gamma^2 h\nu_0$$

Note the similarity of Compton recoil to synchrotron emission, especially the energy relationship  $E_r \propto \gamma^2 h\nu_0$ . We take an energy density and a temperature T to characterize the pool of isotropic photons which need not be particularly Planckian in spectrum. Though the cosmological shift effects the radiation temperature T, increasing it in the past as the inverse scale length,  $1/R$  (so that  $U_{rad} \sim 1/R^4$ , while matter density goes like  $1/R^3$ ), the red shift lowers the recoil photon energy as distantly observed. The two effects cancel, so a given electron energy gives the same X-ray energy at every distance.

The slope of the observed X-rays fits an electron spectrum with  $n = 1.2$  or  $1.3$ . This implies electrons of  $m = 3.4$  to  $3.6$ . But the parallel synchrotron radio noise from our halo or from radio sources gives about the electron slope  $2.4$ . This is hopeful and even better if one realizes that there are slope-changing processes.

Consider the steady-state continuity equation in space and in energy space for electrons

$$\vec{\nabla} \cdot \vec{j}(E) + \frac{\partial}{\partial E} \left( \frac{\partial E}{dt} n(E) \right) = q(E)$$

If there is a loss process like  $dE/dt = A + BE + CE^2$  the solution at equilibrium for an injection spectrum of  $1/E^m$  is given by

$$dn/dE \propto K(1/E)^m (A/E + B + CE)^{-1}$$

so the electron spectrum steepens by one power at high energies, flattens by one at low, and retains the source shape  $q(E)$  in between. The recoil photon spectrum follows these breaks, but naturally changes only by half a power (synchrotron also). Losses: (a) ionization, (b) diffusion or expansion leakage, (c) synchrotron or photon recoil.

So the game is to find which populations give what is observed, in intensity, slope and cut-offs. So far there is promise, but no certain solution. The X-ray observed spectrum is up by about 30 to 50 over the result from all external galaxy haloes, and  $10^4$  over our own. These electrons might be metagalactic escapees from radio sources, giving a power like 2.4 for injection, steepened by one unit by collision with the black-body photons. But no one knows how much there should be. We need a few hundred microvolts per  $cm^3$ . Cosmology might help. One first thought it would do everything, for the black-body temperature goes up and the energy density goes up by four powers of  $R$ . But the life of CR electrons goes down just the same way. To keep up the electrons, one needs a strongly evolving CR flux into the past. This is not even implausible, but it is arbitrary; for one assumption, one result follows. One might as well wait for other evidence. It is not at all to be thought of as excluded.

Note that radio is not seen currently in large amounts, so magnetic fields cannot be made to outweigh the photons. A field as large as a micro-gauss would be seen today. The true field between galaxies must be much smaller, at least if electrons are there.

The cut-offs are of interest. If the electrons are Compton-limited by the black-body radiation, the curve should flatten at a gamma near 200, or at some 100 eV of recoil. There is a recent observation which claims an effect at about 1 keV. This might mean a few times greater effect than the black-body, perhaps from the infra-red. Then the rise below that might be from another cause. But the matter is far from secure. It could well be that the ancient and distant evolving sources have a high B, so that the black-body does not drain their energy. Then everything fits, with one assumed evolution and much radio in the past.

### 3. THE THERMAL X-RAY BACKGROUND - (IF ANY)

People who seek to close the universe have a lot of matter to hide. It cannot be neutral, unless molecular, not even a little, from quasar spectra and perhaps also from 21 cm line observations. (The radio limit is about  $10^{-6}$  or a little more atomic H/cm<sup>3</sup>.) So it must be ionized, which means hot compared to 10 eV. But if it is hot compared to 1 keV, it will radiate neatly in the X-ray. Indeed, the lifetime for free-free plasma (neglect helium, which makes a real but not decisive change) is given by the result

$$\tau_{(\text{seconds})} = 10^{11} T^{\frac{1}{2}} / z^2 n_e \quad n_e \text{ in electrons per cm}^3, T \text{ K}$$

Now the flux  $F = q R$  and  $q$  is  $U/\tau$ . Unless the lifetime is of the order of  $10^{17}$  sec, one can hardly heat the material globally. The energy density per cm<sup>3</sup> is only  $10^{-5}$  times the thermal energy per particle. Then the flux is

$$F \sim 10^5 \times (kT/1 \text{ keV})^{3/2} \text{ keV/cm}^2 \text{ sec}$$

All we can allow above the power law at energies beyond a few keV is perhaps 10 keV/cm<sup>2</sup> sec. So we have to look lower, between 10 eV and a few hundred eV. Then the observed flux actually goes down and the allowed density limits up, so we can perhaps see a plasma of somewhat under the critical density and with T at a fraction of 1 keV. This has been claimed lately but there seem to be many alternatives. It is of urgent importance to confirm the newest results.

The spectrum will have an exponential cut-off like  $\exp(-hv/kT)$  at high frequencies and remain flat at low frequencies where other effects occur. This is hard to see except at the upper edge — the low X-ray energies — or perhaps in the deep i.r. But it seems likely to be swamped by the i.r. galaxies.

### 4. GAMMA-RAYS

Beyond the X-rays there will be galactic gamma-rays mainly from the neutral pion decay of proton, gas collisions in the galactic disc. These have been seen; it is too soon to make strong remarks about them. There is no

sign of a very large extragalactic component; it is certainly less than the poleward flux. This would perhaps fit the X-ray slope, though at present it appears there may be a cut-off about 1 to 10 MeV, of uncertain origin. These gamma-ray data, first ever seen beyond the sun, are unpublished and come from G. Clark, MIT, and W. Kraushaar of Wisconsin, with a detector in OSO III and some 400 events up to June of 1968. They clearly see a strong concentration to the galactic plane with a centre-to-antacentre ratio of almost three and a centre-to-pole ratio of about eight. Their energy band begins at about 70 or 80 MeV and amounts to rays about 100 to 200 MeV, probably most of the energy within 200 MeV. It is premature to compare with the expected amount from taking the electron spectrum seen at the sun and the integrated radio spectrum. This will become available soon; there appear to be no strong surprises. Note that the gammas measure the secondary electron source from the cosmic rays; there are also primary electrons, plenty of them (maybe 80% or more of all), as measured here from the positron-electron ratio. This gives extra freedom in reconciling the models.

## 5. HIGHER ENERGY

No data are at hand. It is almost sure that the spectrum must be cut off beyond, say  $10^{15}$  eV photons which cannot travel through a universe made opaque to time by photon-photon collisions, one partner a gamma-ray, the other a millivolt photon of the black-body pool, such that the electron pair threshold is reached. The cross-section is then Thompson-like and the number of black-body photons per  $\text{cm}^3$  is measured not in  $10^{-4}$  or less like the electrons of the assumed universal matter, if it exists, but in hundreds per  $\text{cm}^3$ . So any high-energy photons must start implausibly close at hand, or else be of such high energy that they find the falling part of the pair cross-section well beyond threshold. Even there, infra-red, and eventually even radio photons will do the job. We then expect photons to come to an end well ahead of protons. Recall so far no photons above some 100 MeV have been seen, though much looked for.

## 6. OTHER LOW-ENERGY POOLS

It will be noted that the X-rays seem to come from the presence of a big pool of black-body photons. It is valuable to seek any other pool of photons, perhaps still unseen. The most suspect is the deep i.r., from, say, 10 or so microns up to some hundreds. Here there are certainly many strong emitters among galaxies. A conservative estimate would put this emission at least as much as visible starlight - 10 mV per  $\text{cm}^3$  in all space - and probably up to ten times more. There could even be a surprise here. It is worth noting that the galactic X-rays from recoils, the lifetime of galactic CR electrons, the tail of the high-energy CR protons (cut off by photon collisions which near  $10^{20}$  eV can make photopions) and the over-all X-rays all seem to set an upper limit of about 10 eV/ $\text{cm}^3$  for all the photons in space. This is of course well past the black-body and if achieved would be of radical importance. Probably it is not there, but we need to become

more certain. Perhaps the odd behaviour of the low-energy X-rays, now viewed as free-free radiation from a universal plasma, may be showing recoil effects from other photon pools; in any case, i.r. data are much needed.

## 7. OTHER RADIATIONS

Gravitons and neutrinos are worth mentioning; they occur in our table of energy content in space, in forms held rather likely. There is no direct information, nor is any likely in the near future even from indirect processes.

## R E F E R E N C E S

Up-to-date (1968) experiments on the X-rays at 1 keV and under:

Luna XII: Lebedev group, experiment — Vainshtein, Kurt, Mandelstam, Prsnyakov, Sirovatski, Sunyaev and Tindo. *Cosmicheskie Issledov* n. 2, 1968.

Interpretation: Sunyaev and Vainshtein, preprint 76, Lebedev, 1968.

Rocket: Berkeley group: Bowyer, Field and Mack, *Nature, Lond.* 217 (1968) 32.

NRL group: Henry, Fritz, Meekins, Friedman and Byram, preprint 1968. Washington DC.

Theory papers — with further references:

Free-free: Rees, Sciama and Setti, *Nature, Lond.* 217 (1968) 326.  
Weyman, *Astrophys. J.* 147 (1967) 887.

Compton Recoil: Brecher and Morrison, *Astrophys. J.* 150 (1967) 161. Felten and Morrison, *Astrophys. J.* 146 (1966) 686.

Gamma rays: Fazio, Stecher et al. See *Ann. Reviews* (1967), Fazio review.



# RADIO STUDIES OF GALACTIC STRUCTURE

B.F. BURKE

Department of Physics and Research Laboratory of Electronics,  
Massachusetts Institute of Technology,  
Cambridge, Mass., United States of America

## Abstract

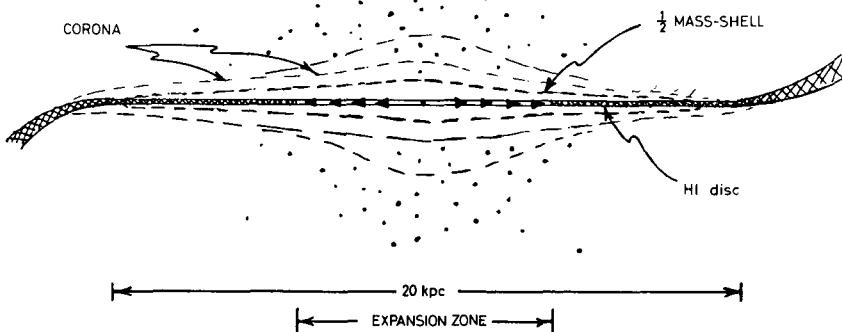
RADIO STUDIES OF GALACTIC STRUCTURE. 1. Introduction, 2. Galactic dynamics, 3. Spiral structure, 4. The anomalous motions.

## 1. INTRODUCTION

The Milky Way Galaxy, the stellar system to which our sun belongs, presents a dual set of problems. The broader, more fundamental questions relate to the origin and formation of the galaxy: of how and when the galaxy condensed from the primeval matter; indeed, whether the process is one of condensation of accretion or whether each galactic nucleus is a source of matter; whether the process is a continuing one; in brief, how do galaxies form? The second set of questions are more descriptive: what are the forms of energy in the galaxy; what are the forces that govern the motions of its components; how large is it, how massive is it, how is the matter and energy distributed and what processes are taking place that we can measure and describe? Our galaxy is only one of many, but if we are to solve the more general problems, detailed knowledge of our own environment can provide, hopefully, the more complete tests of cosmogony that can only be outlined sketchily by our dim view of other galactic systems. I shall concentrate, therefore, on the problems associated with the physical description of our galaxy, particularly with those properties that are determined through radio measurements. Thus the gaseous components of the galaxy will be the principal subject, although the larger questions concerning the overall structure of the galaxy necessarily involve discussion of the stellar component as well.

The general structure of our galaxy as we understand it today is shown in Fig. 1. The oldest population, the globular clusters and high-velocity stars of extreme Population II, form a nearly spherical, centrally condensed cloud whose constituent stars follow highly eccentric orbits with little angular momentum, plunging through the centre of the galaxy. The great mass of stars are in less eccentric orbits, and in Schmidt's (1965) model perhaps half the stellar mass lies within an ellipsoid of axial ratio 20:1, with the sun lying close to this surface. (This is perhaps a danger signal - the median location of the sun betrays anthropocentric tendencies. A larger fraction than we suspect may well lie outside the sun's galactic orbit.) The gas and the youngest stars - the extreme Population I component - lie in an extraordinarily flat disc, less than 200 parsecs thick out to a galactic radius of 10 000 parsecs, then becoming thicker and bending out of the plane in the sense that would be expected naively from the tidal effects of the Large Magellanic Cloud. It is clear that the stars and the gas

exhibit spiral patterns similar to those we see in external galaxies. The problem of spiral structure, however, has plagued the theoretician and the observer - the theoretician because the spiral pattern has been so hard to explain, and the observer because the pattern, in our own galaxy, is so difficult to detect and describe. Finally, the most elusive and surprising of all components can be characterized as the galactic corona, often called the galactic halo, although its actual shape has not been conclusively demonstrated. Figure 1 shows a compromise galactic corona, not much more extended than the major stellar disc, although the caution should be given that this reflects my prejudices, largely based on the observations by Turner and Burke (1967). Mills (1959) would make the corona much more spherical although some of us at times would have eliminated it altogether.



## 2. GALACTIC DYNAMICS

First, let us consider the force fields that affect the large-scale motion of the galactic constituents. Nuclear forces and weak interactions have no direct influence that we know of, although their effects are manifest in spectacular fashion when supernovae explode, and life itself depends on the generation of energy in stars through nuclear processes. Electromagnetic forces, and especially large-scale magnetic fields, must not be ignored, and it was widely accepted a few years ago that spiral structure itself bore witness to the importance of hydromagnetics. Mechanical forces, i.e. gas pressure, must also be considered, but of prime importance, when dealing with dimensions of the order of a kiloparsec or more, is the gravitational field, as one can easily show by computing the order of magnitude of each class of force. At the solar neighbourhood, the gravitational force per unit mass will be  $\sim V_c^2 / R_0$  where  $V_c$  is the circular component of velocity and  $R_0$  is the distance of the sun from the galactic centre.

The effects of gas pressure and magnetic pressure will be of the same order of magnitude, presumably, and so will be  $\nabla p / \rho \sim V_T^2 / L$  for gas density,  $\rho$ , and mean turbulent velocity,  $V_T$ . The characteristic scale-

length,  $L$ , for the density fluctuations in this case are of the order of one kiloparsec. The recent pulsar observations suggest that the effects of magnetic pressure may be smaller still. At any rate, using accepted values for these quantities,

Gravitation:

$$\frac{V_c^2}{R_0} = \frac{(2.5 \times 10^7)^2}{3 \times 10^{22}} = 2 \times 10^{-8} \text{ cm/sec}^2$$

Pressure:

$$\frac{1}{\rho} \nabla p \sim \frac{V_t^2}{L} = \frac{(8 \times 10^5)^2}{3 \times 10^{21}} = 2 \times 10^{-10} \text{ cm/sec}^2$$

Thus, the effects of gravity are overwhelming, not only on the stellar component but on the gaseous component as well. (The cosmic-ray component of the galactic corona is a special case.) Only on a rather small scale, 100 pc or even less, does the importance of the gravitational field diminish to the same order of magnitude as kinetic and magnetic pressure.

The knowledge of the space motions of the stars and gas could, in principle, yield the mass distribution, for we would know the force field, and from that we would solve the Poisson equation to get the galactic density  $\rho(r, \theta, z)$ . We are far from that state of affairs, since galactic distance and time scales are so large that we can determine radial velocities only, except for the closest stars. The effects of interstellar dust further limit observations to the radio wavelengths, where the 21-cm hydrogen line provides information on the motions of the neutral gas, while the high- $n$  recombination lines of excited hydrogen give information on the HII regions, the ionized component of the interstellar medium. The observations are of Doppler shift only, and so a distance scale must be inferred. If one makes the simplifying assumption that the orbits of the hydrogen gas clouds in the galaxy are circular, it is easy to show that the apparent radial velocity,  $V_r$ , is given by

$$V_r = R_0(\omega(R) - \omega(R_0)) \sin \ell$$

where the angular rotation law is  $\omega(R)$ ,  $R_0$  is the distance of the sun from the galactic centre, and  $\ell$  is galactic longitude measured from the centre, as shown in Fig. 2. At this point,  $\omega(R)$  is not known, nor  $R_0$ , and conclusions about the mass of the galaxy and its distribution depend critically upon these inferences.  $R_0$  must be determined by optical methods, and is now taken as 10 kpc, although there is perhaps 20 per cent uncertainty in this number. The form of  $\omega(R)$  is best inferred from the 21-cm radial velocity measurements themselves, by noting that the circle passing through the sun and the galactic centre is the locus of maximum velocity. Using a velocity law determined by these means, the well-known Leiden-Sydney galactic model (Fig. 3) was derived from the observations. The general spiral pattern of the galaxy is clearly displayed in this model, and this impressive work is a good starting point from which to discuss galactic structure. A casual inspection shows that the left and right halves are not

similar, for the spirals from  $\ell = 0^\circ$  to  $\ell = 180^\circ$  are circular, and from  $\ell = 180^\circ$  to  $360^\circ$  the patterns trail markedly. This is not an effect of combining results from two observatories; the effect is real, and gives the first hint that purely circular motion is not a completely accurate assumption. Secondly, the absence of model structures within 4 kpc is not entirely an observational effect, since the motions in this region are certainly non-circular. Thus, we can distinguish several regimes of gas motion: a spiral zone, from 4 kpc outward, characterized by largely (but not entirely) circular motion, an expansion zone to be discussed in Section 4, to which one has to add a third regime, the nuclear region, extending from the centre to about 600 pc, whose nature is still most puzzling.

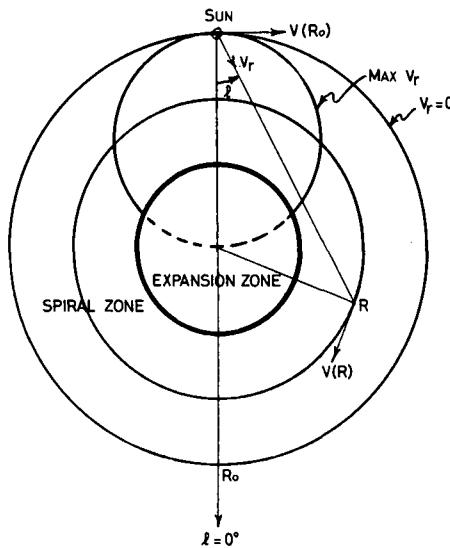


FIG. 2. Geometry for circular motion in the galaxy.

The derivation of a mass distribution then proceeds from the angular velocity law,  $\omega(R)$  by solving for the gravitational potential,  $\phi$ :

$$(\nabla\phi)_R = -R\omega^2$$

This does not give  $\phi$  uniquely, but from this point one can compare the results of mass models, fitting them to our knowledge of  $\phi(R_0)$ , the potential in the solar neighbourhood, which is more completely determined through optical observations. This has been done by several authors, and a simple model has been given by Schmidt (1965), who uses non-homogeneous spheroids, with an axial ratio of 20:1, which gives the system schematized in Fig. 1. Note that the stars are more widely dispersed in  $z$  than the gas by about one order of magnitude. Schmidt's model yields a mass for the galaxy of  $1.8 \times 10^{11}$  solar masses, or about half the mass of the Andromeda nebula. The accuracy with which we know the mass of the galaxy is hard to specify, since it depends (a) upon the ellipsoidal mass-model, (b) upon our

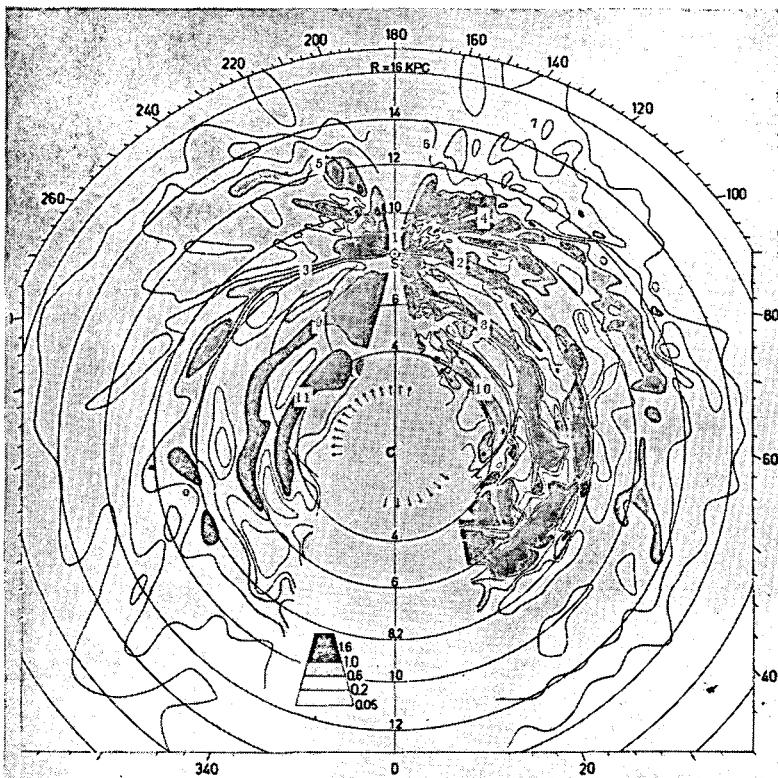


FIG. 3. The Leiden-Sydney model of the distribution of neutral hydrogen (Kerr and Westerhout, 1965).

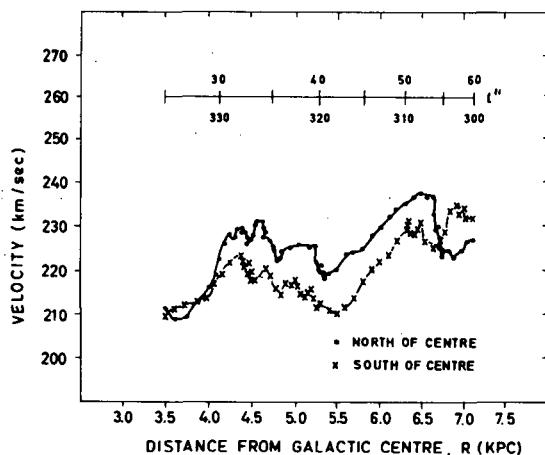


FIG. 4. Comparison of maximum velocity observed for the northern and southern halves of the galaxy (Kerr, 1964).

knowledge of the scale-factor  $R_0$ , and (c) on the assumption of circular orbits for the hydrogen. Assumptions (a) and (b) are hard to test with present measurements, but we have reached a point where the accuracy of (c) can be discussed.

The simplest direct test is given by comparing the velocity law,  $\omega(R)$ , derived from observations along the two halves of the maximum-velocity locus shown in Fig. 2. Kerr (1964) has shown that these are, indeed, different, as shown in Fig. 4. The differences in rotational velocity are typically 10 km/sec, compared to a total rotational velocity of about 230-250 km/sec. This shows that the gross assumption of circular velocity is a very good approximation indeed, although it does raise interesting questions about the detailed (but still large-scale) motions of the gas. It is curious that the curves do not vary monotonically with  $R$ , but show similar wavy deviations, even though the two loci are in different quadrants of the galactic plane.

A related study has now been made by the joint MIT-NRAO group of the motion of the ionized gas, by studying the radial velocity of the hydrogen recombination line  $109\alpha$  ( $n = 110$ ,  $n = 109$ ) in HII regions. The HII regions, since they mark the occurrence of bright, recently formed stars, form part of the gas population, but mark special regions where the condensation of stars is favoured. There is no a priori reason to suppose that the dynamical behaviour of the HII system is the same as the HI as a whole, and a test was made, independent of galactic models, to compare the behaviour of ionized and neutral gas.

The intensity of both HI and HII hydrogen-line radiation is measured as a function of velocity, galactic longitude, and galactic latitude, and one cannot show the complete functional dependence in a two-dimensional contour diagram. By restricting one co-ordinate to a constant value, however, a plot can be made, and for studying motions in the plane it is natural to take a section along the galactic equator, i.e.  $b = 0$ , and to show the sky brightness temperature as a function of velocity and galactic longitude. Figures 5 and 6 show  $T_b(\ell, v_r)$  for both the recent Sydney HI data (Kerr, 1966) and the MIT-NRAO HII data (Reifenstein, Wilson, Burke and Mezger, unpublished). Since only observables are shown, the comparison is independent of galactic model. The HI observations are shown as contours in °K, while the HII observations, since they refer to discrete objects, are shown as black spots.

The HI from galactic longitude  $-20^\circ$  to  $+20^\circ$  is shown in Fig. 5, and the anomalous motion of hydrogen in the expansion zone can be seen clearly (the observations at the centre itself are left out of the diagram). Hydrogen in circular orbits should show low radial velocity at longitudes near the centre, because  $\sin \ell$  is small, and the intense ridge at low velocities can therefore be assigned to hydrogen in the spiral region. The high-velocity hydrogen between  $340^\circ$  and  $20^\circ$  has been shown by Oort and Rougoor (1960) to be associated with features exhibiting radial motions in the galaxy comparable with their circular motions. One, in particular, the '3.5 kpc expanding arm' can be seen as the intense ridge at the left of the diagram, and is certainly an expanding feature, located between the sun and the galactic centre, since it shows as an absorption feature at -53 km/sec against the galactic centre.

The hydrogen features in Fig. 6 do not show such anomalies, and can be fitted by models in which purely circular motions are assumed. The

allowable non-circular velocities in such models may amount to perhaps 10 km/sec, and such motions are indeed probable, as will be shown in Section 3.

The ionized hydrogen observations are more limited, but the strong concentration along the spiral ridge from  $\ell = 348^\circ$  to  $\ell = 18^\circ$  can be seen easily. Three points are sharply anomalous, and are almost certainly associated with the galactic centre ( $\ell = 0^\circ$ ). The expansion zone appears almost devoid of HII regions, with one possible exception, but when one reaches the edge of the expansion zone, at  $\ell = 18^\circ$ , suddenly high-velocity HII regions appear. These persist along the hydrogen ridge to  $\ell = 52^\circ$ , with lower-velocity points also present, corresponding to closer HII regions. The observed maximum radial velocities are very similar to the maximum radial velocities observed for the HI, which means that the rotation laws for the two systems are similar. In order to derive a model for the HII distribution, the same rotation law was used as for the HI in Fig. 3. There is a difficulty in resolving distance ambiguities, since a given velocity in the line of sight corresponds to two different distances if the object lies closer than  $R_0$  to the galactic centre. This is obvious from Fig. 2 and Eq. (2), since a given circle,  $R$ , is intersected twice by the chord to the given point.

The HI and HII models are both plagued by the distance ambiguity inside  $R_0$ , but the HI observations are aided by continuity arguments, starting from the non-ambiguous locus of maximum velocity. The HII map presented in Fig. 7 shows both possible locations for a given observation, as open and filled circles. In most cases, the nearer point, i.e. the filled circle is correct, since the inverse-square law discriminates against distant sources observationally. Note that the incidence of HII regions increases sharply just outside the expansion zone, at  $R = 4$  kpc. Although the relative intensities of the HII regions are not shown, these are also extremely luminous HII regions, and if one were viewing our galaxy from outside, it is this concentration that would be most prominent. The number of HII regions diminishes as one approaches  $R_0$ , and outside the sun, there are few HII regions of importance. We can see, therefore, that the galaxy shows rather different aspects in its neutral and ionized gas distributions. The hydrogen spiral structure extends much farther from the centre than the HII system, and while the HII is usually formed near HI concentrations, the converse does not hold. The fundamental question of the appearance of our galaxy still cannot be answered, although the observations are consistent with the old belief that our galaxy resembles the great nebula in Andromeda. The masses of the two galaxies are the same (within a factor of two, which can be ascribed to observational uncertainty), they both have similar spiral structure, and perhaps the only remarkable feature of this coincidence is that I have not found any occurrence in the sky of such a pair of giant galaxies, only ten galactic diameters apart.

### 3. SPIRAL STRUCTURE

The striking patterns exhibited by external spiral galaxies, and more dimly shown in the spiral models of our own galaxy, have presented an intriguing puzzle to theoreticians, and the problem is not yet completely solved. At one time, it was widely believed that the spiral structure was

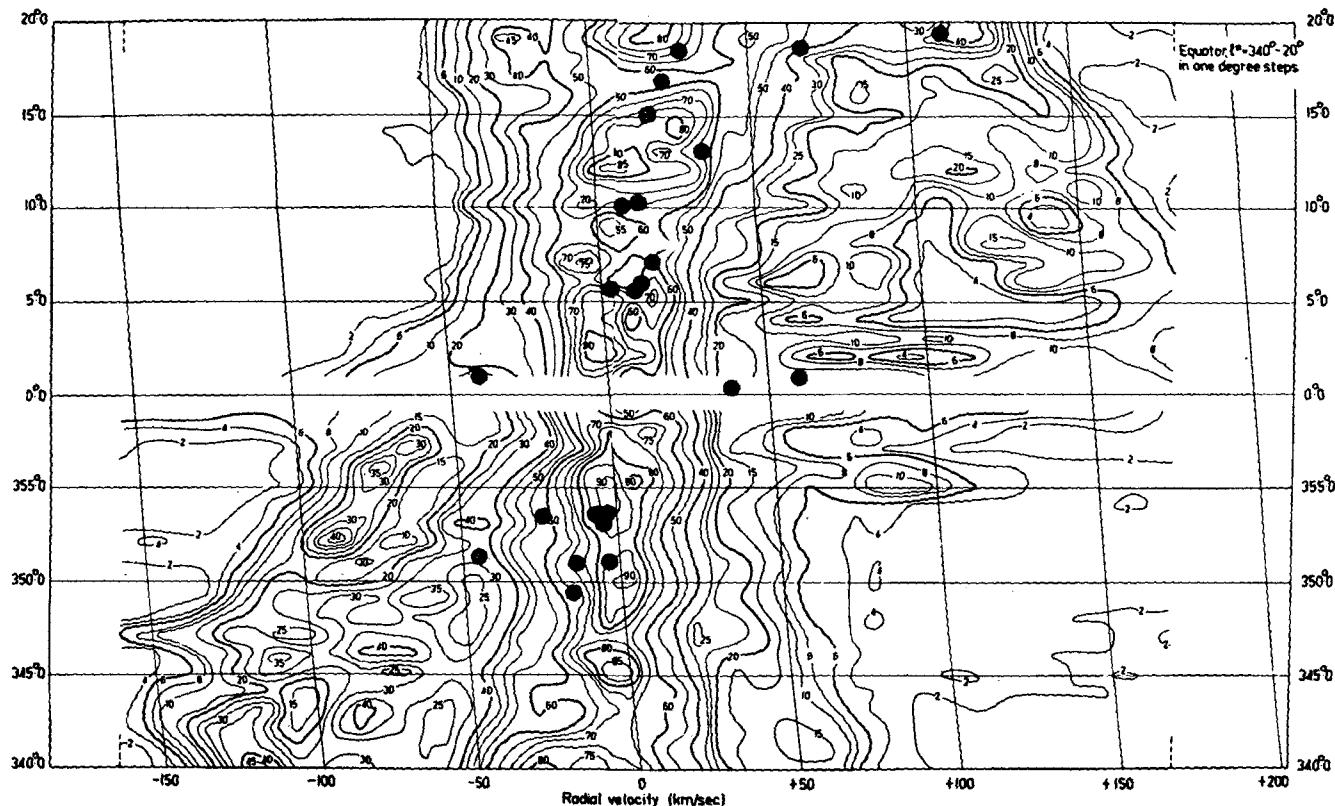


FIG. 5. The observed radial velocities of neutral and ionized hydrogen along the galactic plane from longitude  $340^\circ$  to  $+20^\circ$ . The contours are from the Parkes Survey by Kerr (1967) and the circles are HII region velocities measured by Reifenstein, Wilson, Burke, and Mezger (in preparation).

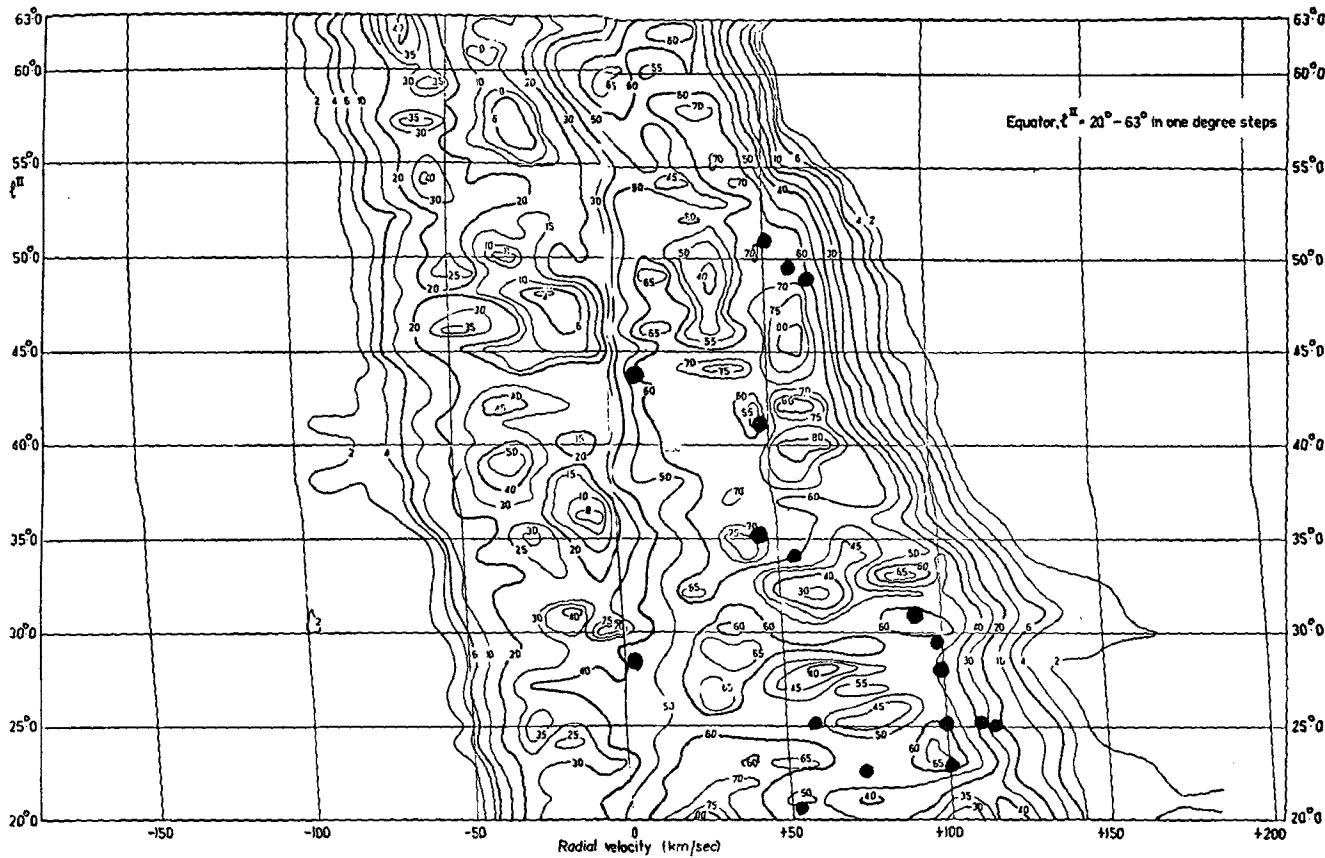


FIG. 6. Comparison of HI and HII from longitude  $20^\circ$  to  $60^\circ$ .

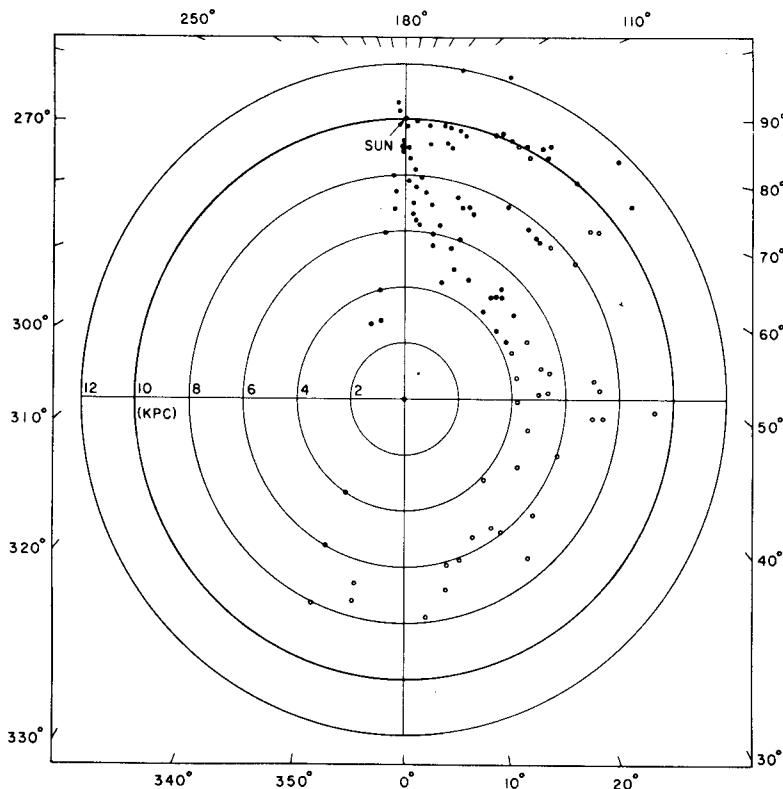


FIG. 7. Model distribution of HII regions in the galaxy. Most positions are ambiguous, the near positions being indicated by filled circles and the far positions by open circles. Most of the HII regions beyond  $R=7\text{ kpc}$  are at the close position. (Reifenstein et al., to be published).

manifest evidence of the general galactic magnetic field. Lack of success in producing an acceptable theory was laid to the fierce difficulties presented by the hydromagnetic problem. Nature provides evidence, however, that magnetic forces are not necessary to produce the spiral patterns observable in smaller-scale phenomena such as hurricanes, or cream in a stirred cup of coffee, and an exciting development in recent years has been the renewal of interest in purely gravitational theories, led by C.C. Lin and his co-workers.

The spiral problem has several aspects, in addition to the basic demonstration. The differential galactic rotation is such that in only a few galactic rotations -  $10^9$  years or so - the spiral arms would be much more tightly wound if they move with the galactic matter and are long-lived structures. They are probably not transitory phenomena, since the appearance of spiral structure is general in galaxies, provided the gas represents 2-5% of the total galactic mass. Curiously, the irregular galaxies, with 10-30% of their mass in gaseous hydrogen, do not show pronounced spiral structure, while the ellipticals, which lack spiral arms completely, are conspicuously devoid of gas. Thus, the 'winding-up problem'

must necessarily be solved by a spiral theory. The general form of the spiral structure can be described in many cases as a pair of logarithmic spirals, and so the predominance of two-armed structures must also be explained. Trailing, rather than opening, spirals must be produced (opening spirals may exist, but only trailing spirals have been found so far). Finally, the spiral structure must also be compatible with approximately circular motions, although 5% deviations are allowed by the observational evidence.

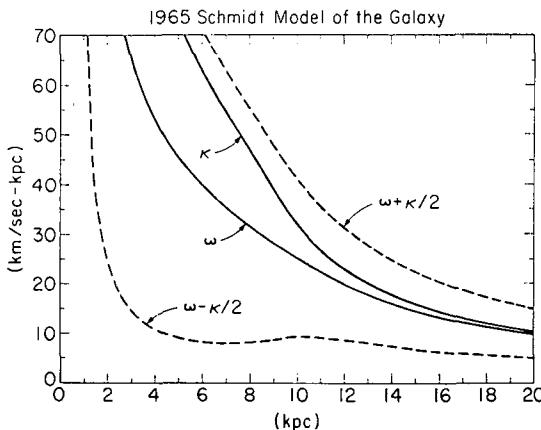


FIG. 8. The rotational frequency  $\omega(R)$  and epicyclic frequency  $\kappa(R)$  for the Schmidt 1965 galactic model. (Lin, 1966).

B. Lindblad (1963) originally suggested that a spiral pattern could be maintained by density waves, within the framework of gravitational forces only. Thus, while the material in the galaxy was rotating according with angular velocity  $\omega(R)$ , Lindblad suggested that the superposition of a density wave would result in a spiral density concentration that maintained its form indefinitely, rotating with a pattern speed  $\omega_p$ . He also noted that smaller-scale motions, at the epicyclic frequency  $\kappa(R)$  could result in a resonance if  $\omega(R)$  and some multiple of  $\kappa(R)$  were commensurate. He and P. O. Lindblad noted that if one takes the observed galactic rotation  $\omega(R)$ , from which one can derive the epicyclic frequency  $\kappa(R)$  (which is just the simple harmonic motion of a test particle when it is displaced by a small amount from its circular orbit), the difference

$$\omega_p = \omega(R) - \frac{1}{m} \kappa(R)$$

is nearly constant over much of the galaxy if  $m = 2$ , as shown in Fig. 8. A resonance condition is satisfied over much of the galactic disc, therefore, implying a long-lived pattern rotating at frequency  $\omega_p$ . Furthermore, the factor 2 dividing  $\kappa(R)$  implies that a two-armed spiral is favoured.

The more rigorous tests of stability (including the winding-up problem), trailing arms, and scale were not met at this stage, but over the last few years Lin and his co-workers (1964, 1966, 1967) have met with remarkable success in demonstrating, for a thin disc, the existence of spiral density

Spiral Pattern for 1965 Schmidt Model

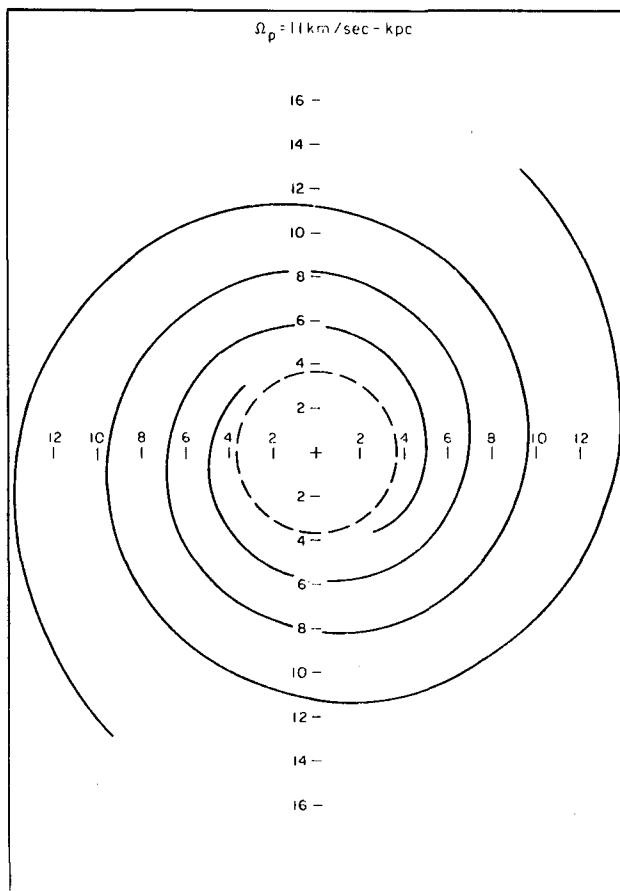


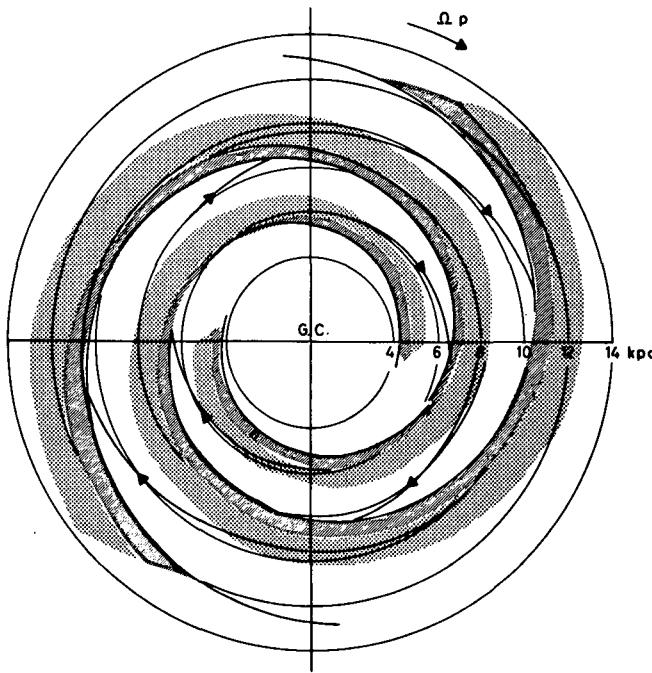
FIG. 9. Spiral structure derived by Lin (1966).

waves that meet many tests. The first demonstrated the stability of such spiral waves, considering the response both of the gaseous disc (which is reasonably hydrodynamic, with negligible viscosity) and of the stellar disc including the observed dispersion in stellar velocities (and thus involves solving the collisionless Boltzmann equation). They found that trailing waves were preferred, and were sustained if the pattern speed,  $\omega_p$ , met the condition

$$\omega - \frac{\kappa}{m} < \omega_p < \omega + \frac{\kappa}{m}$$

The existence of a stable pattern is implied if  $\omega_p$  is constant, and from inspection of Fig. 8 it can be seen that this condition is met over a wide range of  $R$ , if  $m = 2$ . Furthermore, if  $m$  is larger than 2, implying more arms, the condition is only met over a small range of  $R$ . Lin also derived

a dispersion relation relating the radial wavelength,  $\lambda$ , to the frequency  $\nu$  at which stars encounter the pattern, and from this he derived the spiral model shown in Fig. 9. The model was derived entirely on the basis of observables, with no adjustable parameters: the rotation  $\omega(R)$ , epicyclic frequency  $\kappa(R)$ , and surface density  $\sigma$ , were taken from Schmidt's 1965 galactic model;  $m$  was set equal to 2 to give a spiral pattern that extended over a wide range of  $R$ ; the remaining parameter was  $R_L$ , the radius at which the Lindblad resonance occurs. In a sense, this is a free parameter, but  $R = 4 \text{ kpc}$ , the radius that divides the expansion zone from the spiral zone, is a natural distance to choose for  $R_L$ . The general resemblance of this model to the Leiden-Sydney model is striking.



SPIRAL PATTERN IN THE GALAXY

FIG. 10. Gas streamlines in the rotating co-ordinate system of the pattern velocity. (Lin and Roberts, to be published).

The pattern speed for Lin's model in Fig. 9 is  $11 \text{ km/sec/kpc}$  and should be compared to the rotational velocity of the local system, which is  $25 \text{ km/sec/kpc}$ . Thus, the pattern moves rapidly with respect to the stars, raising the kinematic question of what deviations from circular motion are required. It is clear that the velocity now contains a radial component,  $\pi(R, \theta)$ , and the circular component of velocity,  $V_c = R\omega(R)$ , now varies with galactic azimuth:  $V_c = R\omega(R, \theta)$ . Surprisingly, the velocity anomalies are not large; in the neighbourhood of the sun,  $R_0$ ,  $\pi$  is of the order of  $6 \text{ km/sec}$ , and the azimuthal variations in  $V_c$  are of the order of  $\pm 14 \text{ km/sec}$ . The gas motions derived by Lin (1967) are shown in Fig. 10, in a rotating co-ordinate

system in which the pattern stands still. The gas speeds up between the arms, to satisfy the condition that the gas density remain low in the inter-arm regions, and the streamlines close.

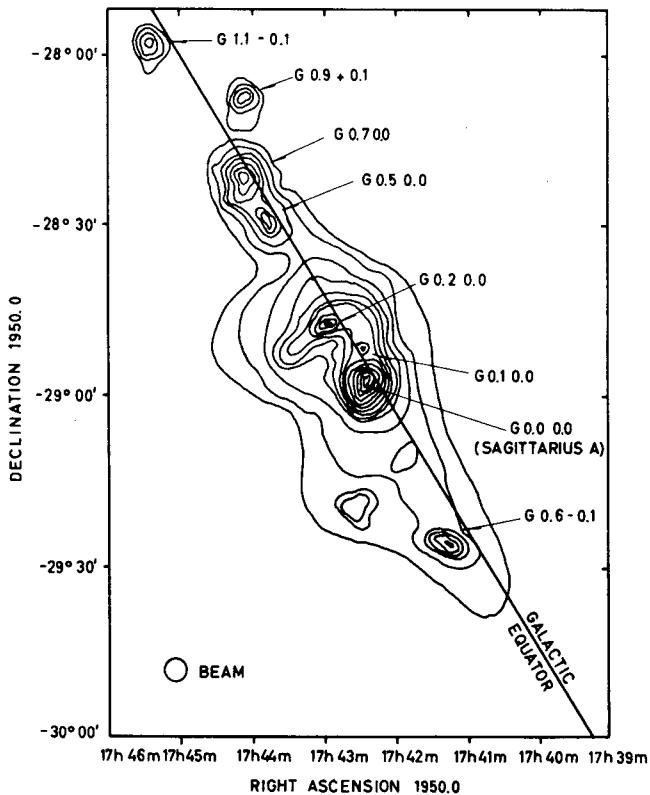


FIG.11. The continuum sources at the position of the galactic centre. (Downes, Maxwell, and Meeks, 1966).

Recently Lin and Roberts (private communication) have derived a further interesting result. The orbits of Fig. 10 were derived from a model in which the magnitude of the fluctuation in gravitational potential was taken from the quasi-sinusoidal variations in Kerr's observed radial velocities in Fig. 4. A numerical integration showed that a shock forms on the inner edge of the spiral arm, as shown by the cusp in the streamlines. Now, it has always been difficult to form stars by condensation in normal inter-stellar clouds, but Lin believes that the factor of 8 compression at this shock may provide a high enough density to encourage the formation of stars. This would mean that stars should form at this shock, and we should therefore find the youngest stars, marked by the HII regions, on the inside edge of the spiral arms, with the older (but not very old) stars spread more into the outside edge. It is possible to persuade oneself that many of the HII regions shown in the comparison of Figs 5 and 6 are consistent with this picture, since they tend to occur at radial velocities whose absolute value is greater than the radial velocity of the hydrogen ridge, i.e.  $\omega(R) - \omega(R_0)$

is greater for the HII than for the HI, so the HII lies closer to the galactic centre than its associated HI ridge. More refined observations will certainly be needed to prove this point, however.

#### 4. THE ANOMALOUS MOTIONS

The nuclei of galaxies are remarkable in many respects. The Seyfert galaxies possess nuclear regions in which mass motions of thousands of km/sec are evident from the Doppler-broadened emission lines that are observed. The nuclei of some galaxies are like quasars in many respects, giving off intense radio emission and exhibiting brightness fluctuations, both radio and optical (e.g. NGC 1275 and 3C120). The nucleus of our own galaxy, while less spectacular than some, has remarkable properties. The continuum radio emission from the cluster of sources near the galactic centre is shown in Fig. 11. A few of these features may be foreground objects, but the symmetry about the dynamical centre of the galaxy is striking enough to justify identification of the complex with the centre. The MIT-NRAO observations of the  $109\alpha$  recombination line show that features G(0.5, 0.0), G(0.7, 0.0) and G(1.1, -0.1) are HII regions, while most of the remaining sources are either non-thermal emitters or are much hotter than normal HII regions. Neutral hydrogen is observed, over this entire region, to be rotating rapidly about the centre with circular velocities of 250 km/sec, and with radial components (inward and outward) of at least 80 km/sec. The rapid rotation suggests that there is a remarkable mass concentration within 500 pc of the centre:  $10^9$  solar masses within a disc 1000 pc in diameter and 200 pc thick, or 100 times the stellar density in the solar neighbourhood. At the very centre, the density may be much higher, for both the Andromeda nebula and several nearby spirals like M33 have quasi-stellar nuclei, with  $10^6$ - $10^7$  solar masses inside a 10 pc sphere.

Beyond the nuclear region in our galaxy, the expansion zone bears testimony to a large-scale process whose relation to the nucleus is not understood. Rougoor and Oort (1960) have shown that at least two major features are segments of expanding arms, and at least one of these, the 3.5 kpc expanding arm, is rotating at an angular velocity of the order of 50 km/sec/kpc, and hence possesses appreciable angular momentum. The major questions that still remain unanswered are:

- (1) Is the process continuous or intermittent?
- (2) Does the expanding material come from the nucleus or the galactic corona?
- (3) Where does the angular momentum come from?
- (4) What sort of transition occurs between the expansion and spiral zone?

It is difficult to judge the existing theoretical work, since there is so little certainty in the observational data. The data strongly suggest an explosive event in the galactic nucleus, not more than 30 million years ago.

The chemical complexity of the nuclear region was demonstrated by Bolton and his co-workers (1965), and by Lilley, Gundersen and Goldstein

(1965), who showed that the entire nuclear region is blanketed by OH absorption lines that imply, from their extraordinary strength that the density of the OH radical, relative to hydrogen, is of the order of 300 times the relative OH abundance in the solar neighbourhood. The entire subject of OH line observations is now too complex to include in a brief review, but it is clear that the OH lambda-doublet 18 cm lines show evidence of complicating physical processes, including maser action on a cosmic scale.

A few comments should also be made about the high-velocity HI clouds in the solar neighbourhood. In the discussion of the motions in the spiral zone, the orderly motion of the greatest fraction of the HI allows one to construct pleasingly regular models. There are exceptions, however, for Hulsbosch and Raimond (1966) and Blaaw et al. (1967) discovered hydrogen clouds at high galactic latitudes, exhibiting peculiar velocities of -30 to -200 km/sec (towards the sun, i.e. falling into the galactic plane). These clouds, whose velocity in some cases is comparable to galactic rotation, come predominantly from one quadrant in the sky, between 60° and 200° galactic longitude and +10° to +80° galactic latitude. The data have been discussed by Oort (1967), who examined the following possibilities:

- (1) Supernova shells
- (2) Super explosions ( $10^6$  times a supernova)
- (3) Condensation from the galactic corona
- (4) Swept-up material from intergalactic space.

There are objections to all four hypotheses, but (2) or (4) appear to be most likely.

# THE MAGNETIC FIELD OF THE GALAXY

F. G. SMITH

Nuffield Radio Astronomy Laboratory,  
Macclesfield, Cheshire, United Kingdom

## Abstract

THE MAGNETIC FIELD OF THE GALAXY. 1. General evidence; 2. Optical evidence; 3. Galactic radio emission; 4. Polarized galactic emission; 5. Faraday rotation from extragalactic sources; 6. Zeeman splitting; 7. Faraday rotation from Pulsar CP1950; 8. A possible reconciliation; 9. Small-scale irregularities.

The existence of a general magnetic field throughout interstellar space has been accepted for 20 years. The configuration of the field is a matter for speculation, except in the vicinity of the Sun where considerable evidence now indicates a field roughly along the direction of the local spiral arm. The strength of the field, both locally and on the larger scale, has not so far been measured, but confident predictions from several different physical arguments agree in values of about  $10^{-5}$  G in the spiral arm, and  $10^{-6}$  G extending into a halo. It is therefore disturbing that the first direct measurement of the field along one particular line of sight has yielded only an upper limit of  $2 \times 10^{-7}$  G.

I shall first review the arguments for the existence of the field, and the estimates of its value, bearing in mind the reconciliation that may become necessary between the new observation and these predictions.

## 1. GENERAL EVIDENCE

A general argument comes from the dynamics of the galaxy. Interstellar gas is a very good conductor, with a very long time constant of decay of any system of fields and currents within it. An equality may therefore be expected between kinetic energy  $\frac{1}{2} \rho v^2$  and magnetic energy  $H^2/8\pi$ . Setting  $\rho = 10^{-24}$  g cm<sup>-3</sup> and  $v = 10$  km sec<sup>-1</sup>, admittedly rough values for interstellar material which has in fact a wide range of density and possibly also of velocity, we find  $H \sim 3 \times 10^{-6}$  G. Higher densities, as in spiral arms, would indicate  $H \sim 10^{-5}$  G, and lower densities, as in the halo, indicate  $H \sim 10^{-6}$  G.

The confinement of cosmic rays in the galaxy, if indeed they are so confined, requires a field of the same order. The radius of curvature of a proton with energy E (electron volts) in a field of H (gauss) is

$$r = \frac{E}{300 H} \text{ cm}$$

A proton with  $E = 10^{16}$  eV in a field  $H = 3 \times 10^{-6}$  G will gyrate within a radius of 30 parsec, and is therefore confined effectively within the galaxy. Evidently when  $E = 10^{20}$  eV the cosmic rays are not confined, so there is no precise borderline to draw and no exact value of H to be deduced. Another value of H can be obtained by equating the energy density of cosmic rays to the

magnetic energy density, so that the field is strong enough to confine most of the cosmic ray flux to the galaxy. The value deduced for a halo field is  $7 \times 10^{-6}$  G, as shown by Van de Hulst [1] in a recent and most valuable review.

Although these arguments are not precise, it should be noted that they may have to be discarded entirely if the value of H is out by a factor of ten. This would mean that an equilibrium value of H has not been built up by kinetic motions, which seems unlikely, and that only cosmic rays with lower energies are confined to the galaxy, which is easier to accept.

## 2. OPTICAL EVIDENCE

The first evidence of the existence and the alignment of the magnetic field came from the observation of optical polarization of starlight (Hiltner, 1949). This arises from the alignment of dust grains across the magnetic field, so that the anisotropic absorption indicates directly the direction of the field, but not its magnitude. The field lies along the spiral arm, in the direction  $\ell = 50\text{--}80^\circ$ . The same direction is indicated by the distribution in the local arm of early-type stars and H II regions.

An extremely interesting new development of these observations is the finding that the plane of polarization varies with optical wavelength [2]. This has been interpreted by Ireland et al. [3] as a misalignment of larger dust particles, as might occur if the field direction was changing in a period of the order of  $10^6$  years.

## 3. GALACTIC RADIO EMISSION

The background radio brightness of the sky is separable into a component originating outside the galaxy, with a spectral index of 0.75 and presumably originating in the discrete radio galaxies and quasars, and a component with smaller spectral index originating within the galaxy, as shown by its angular distribution over the sky. This separation has been discussed by Bridle [4].

The spectral index of the galactic component varies somewhat over the galaxy, but it lies between 0.38 and 0.45 at frequencies between 10 and 100 MHz. The volume emissivity depends on the depth of emission. However, for a region near  $\ell = 140^\circ$ , Bridle [5] estimates the emissivity of the disc to be between  $4.5$  and  $9 \times 10^{-41}$  W m<sup>-3</sup> sr<sup>-1</sup> Hz<sup>-1</sup>.

The galactic radio emission is synchrotron radiation from cosmic ray electrons gyrating in the galactic magnetic field. If the flux of electrons with energy E is I(E), following an exponential population law

$$I(E) = KE^{-\gamma}$$

then the emissivity per unit volume is proportional to  $KH^{(\gamma+1)/2} \nu^{-(\gamma-1)/2}$ . Using appropriate numerical values [6], we can relate the intensity and spectrum of the radio emission to the electron population, which has been measured (see Ref.[1]), as

$$I(E) = 5 \times 10^{-3} E^{-2} \text{ electrons cm}^{-2} \text{ sr}^{-1} \text{ GeV}^{-1}$$

Here the trouble begins. Firstly the spectrum is nearly, but not quite, right. Since  $\alpha = (\gamma - 1)/2$  we expect  $\alpha = 0.5$  instead of about 0.4. Possibly the true value of  $\gamma$  is 1.8 instead of 2, indicating that there are relatively more high energy electrons than have been detected. However, it is more to be expected that low energy electrons have been missed because they are excluded from the solar system by the interplanetary magnetic field. Ignoring this difference the volume emissivities quoted by Bridle lead to  $H = 6 \times 10^{-6}$  G and  $1.6 \times 10^{-5}$  G respectively, for emission spread uniformly or concentrated in clouds. The electrons mainly responsible for this emission then have energies of the order of 1 GeV, and the solar modulation effect should not seriously have affected the measurements of electron flux.

These values for H are uncomfortably large, and we must probably look for a reduction by a factor of two or three. If the Faraday measurement of  $\sim 2 \times 10^{-7}$  G is typical, then we need a factor of 20. A factor of 20 in magnetic flux requires an increase in electron flux by a factor  $20^{3/2}$  to give the same emissivity, entirely contrary to the measurements. Alternatively the volume emissivity might be wrong, possibly because the observed emission occurs over a longer line of sight; this also seems unlikely.

#### 4. POLARIZED GALACTIC EMISSION

Synchrotron radiation is generally linearly polarized. The galactic magnetic field appears to be sufficiently well organized that a high degree of linear polarization is in fact observed over a large part of the sky at short radio wavelengths. At longer wavelengths a smaller degree of polarization is observed because of Faraday rotation, which rotates the plane of polarization by an increasing amount along a line of sight, progressively destroying any alignment. This 'internal depolarization' is least at one point in the sky,  $\ell = 140^\circ$ ,  $b = +5^\circ$ , which appears to be in a direction perpendicular to the local field.

The actual distribution of polarization at  $\lambda = 0.7$  m and shorter has been used by Bingham and Shakeshaft [7] to find the local configuration of the magnetic field.

A surprising recent result [8] is that substantial polarization persists to the long wavelength of  $\lambda = 1.25$  m over much of the sky from  $\ell = 125 - 155^\circ$  and  $b = 0 - 10^\circ$ . Faraday rotation must be particularly small over a distance of at least 50 pc in this region, giving a product  $NH < 3 \times 10^{-8}$  cm<sup>-3</sup> G.

#### 5. FARADAY ROTATION FROM EXTRAGALACTIC SOURCES

Many extragalactic sources have a linearly polarized component of radio emission, and the position angle of this component varies with wavelength according to a Faraday rotation law. The rotation measure R is defined as  $\theta/\lambda^2$ , where  $\theta$  (radians) is the rotation at wavelength  $\lambda$  (cm); R is a measure of  $\int NH_{11} d\ell$ , where  $H_{11}$  is the line of sight component of H.

Values of R for about 200 sources have been found to be distributed over the sky in a simply organized pattern [9]. This indicates a field running along the arm in a direction close to  $\ell = 70^\circ$ , but it also shows that the direction is reversed on the underside of the arm. A simple model suggests that the reversal occurs above the position of the sun.

The value of  $R$  in this model is about 0.2 per parsec, giving a value of  $NH = 0.25 \times 10^{-6}$  averaged locally through the galactic disc. An average value for  $N$  is available from low frequency radio astronomical measurements from satellites [10], which show an absorption in the interstellar medium at frequencies below 3 MHz. This gives an estimate of  $N \sim 0.1 \text{ cm}^{-3}$ . Combining this with rotation measures we have  $H \sim 2.5 \times 10^{-6} \text{ G}$ .

## 6. ZEEMAN SPLITTING

A direct measurement of  $H$  in a cloud of unionized hydrogen may be made by observing the Zeeman splitting of the 21 cm hydrogen line absorption of a discrete radio source. So far only upper limits have been found for the line of sight component  $H_{11}$ . These have been interpreted by Verschuur [11] as giving an upper limit of  $7 \times 10^{-6} \text{ G}$  for any field aligned along  $\ell = 140^\circ$ . There seems to be a possibility of considerable improvement in these measurements, but the results can only apply to the line of sight components of the field in a very few unionized clouds.

## 7. FARADAY ROTATION FROM PULSAR CP 0950

This recent observation [12] gives the average value of  $H_{11}$ , weighted according to the electron density along the line of sight to the pulsar. Linear polarization is often, but not always, observed in the radio pulse from CP 0950, and Faraday rotation is observable if the plane of polarization is measured over a range of frequencies. The rotation gives a measure of  $\int NH_{11} d\ell$  along the line of sight, but a value of  $\int N d\ell$  is independently available from the measurements of group delay of the pulses over a wider range of wavelengths.

Measurements on CP 0950 showed a value of rotation corresponding to a field of only  $4 \times 10^{-7} \text{ G}$ . Even so, the measured rotation could be attributed almost entirely to the terrestrial ionosphere, and the upper limit of  $H_{11}$  was therefore set at  $2 \times 10^{-7} \text{ G}$ .

No measurements are available as yet from the three other known pulsars.

## 8. A POSSIBLE RECONCILIATION

In favour of a field  $10^{-6} - 10^{-5} \text{ G}$  we have some general dynamical arguments, the deduction from observed synchrotron radiation, and the values of rotation measure from outside the galaxy. The Zeeman measurements are inconclusive and refer only to two clouds. Of these arguments the strongest seems to be the deduction from observed synchrotron radiation, which not only gives a good value for the general field but shows also that it is organized simply enough to produce substantial linear polarization.

The very small field given by the pulsar measurement must therefore be explained as a local phenomenon, either as a locally weak field or one in which a chance geometric arrangement gives a small net line-of-sight component. The general field alignment revealed by the polarized galactic emission does not lead us to expect a suitable alignment, but rotation

measures do suggest that this may have happened. The rotation measure of 3C 227, at  $\ell = 228.6^\circ$ ,  $b = 42.3^\circ$ , within  $1^\circ$  of CP 0950, is  $-6 \pm 3$  (Ref.[13]), which is large compared with  $< 0.5$  for CP 0950, but which is small and reversed in sign compared with adjacent sources. Again, it is probable that the line of sight to CP 0950 is long enough to cross the plane where the field reverses, as determined from the distribution of rotation measures.

Although these geometrical situations seem to provide adequate possibilities for explanation, there is also the observation of local regions of small rotation appearing in the background polarization at long wavelengths. This is not in the same region of sky, but it does imply a locally low value of N or H, or both.

The situation is seen to be awkward but not impossible. We must hope for some rather more satisfactory observations on the three other pulsars.

## 9. SMALL-SCALE IRREGULARITIES

So far only a general field has been considered. Although the polarization of the background radio emission shows that this organized component predominates, there is now an indication of smaller scale structure. Bologna et al. [14] and Davies [9] have shown that radio galaxies are less polarized when seen through the plane of our own galaxy. This is interpreted as a Faraday rotation which is different for different parts of the sources, implying that there is structure in electron density or magnetic field with linear dimensions less than 1 parsec across. It is already evident from the distribution of rotation measures across the sky that there is more to the field than a simple alignment along the arms: it may be that we shall soon have at least a statistical picture of the irregularities.

## R E F E R E N C E S

- [1] VAN DE HULST, H. C., A. Rev. Astr. Astrophys. 5 (1967) 167.
- [2] COYNE, G. V., GEHRELS, T., Astr. J. 71 (1966) 355.
- [3] IRELAND, J. G., NANDY, K., REDDISH, V. C., WICKRAMASINGHE, N. C., Nature, Lond. 212 (1966) 990.
- [4] BRIDLE, A. H., Mon. Not. Roy. Ast. Soc. 136 (1967) 219.
- [5] BRIDLE, A. H., Mon. Not. Roy. Ast. Soc. 138 (1968) 251.
- [6] GINZBURG, V. L., SYNOVATSKI, S. I., A. Rev. Astr. Astrophys. 3 (1965) 297.
- [7] BINGHAM, R. C., SHAKESHAFT, J. R., Mon. Not. Roy. Ast. Soc. 136 (1967) 347.
- [8] SMITH, F. G., Nature, Lond. 217 (1968) 831.
- [9] DAVIES, R. D., Nature, Lond. 218 (1968) 435.
- [10] SMITH, F. G., Mon. Not. Roy. Ast. Soc. 131 (1965) 145.
- [11] VERSCHUUR, G. L., IAU Symposium No. 31 (1967) 385.
- [12] SMITH, F. G., Nature, Lond. 218 (1968) 435.
- [13] BERGE, G. L., SEIELSTAD, G. A., Astrophys. J. 148 (1967) 367.
- [14] BOLOGNA, J. M., McCALPIN, E. F., SLOANAKER, R. M., Science 154 (1966) 1656.



# THE PECULIAR A-TYPE STARS

R.P. KRAFT

Lick Observatory, University of California,  
Santa Cruz, Calif., United States of America

## Abstract

THE PECULIAR A-TYPE STARS. 1. Definitions; location in the Hertzsprung-Russell diagram;  
2. Rotational velocities; 3. Magnetic fields; 4. Suggested explanations of the abundance anomalies.

Since a coherent general discussion of the evolution of single and double stars is not possible here, I thought it advisable to select a problem of more limited scope, but one which nevertheless engages the interest of workers in several diverse areas of stellar astronomy. Moreover, I wanted to find a subject in which the connections between theory and critical observation with optical telescopes are both rich and reasonably tractable. And finally, out of pure personal predilection (peevishness of a stellar spectroscopist?), I wanted to describe some astronomy in which a large telescope equipped with a Coudé spectrograph was the sine qua non of its existence.

These conditions are readily met in the study of what are called "peculiar A-type stars". The bewildering variety of observational facts about these objects engages workers in such fields as stellar evolution, nuclear physics, stellar magnetic fields, plasma physics, stellar photometry, stellar spectroscopy, binary stars, and stellar rotation. An entirely satisfactory theory linking the observed facts has yet to be advanced, and the origin and evolution of these stars has not been worked out. The emphasis in this treatment will be on the observations, but details will be put forward only to the extent that they bear on the leading theoretical ideas.

## 1. DEFINITIONS; LOCATION IN THE HERTZSPRUNG-RUSSELL DIAGRAM

I begin by reminding you of the leading properties of the Hertzsprung-Russell (HR) diagram and its interpretation. This is a plot of luminosity  $L$  versus effective temperature  $T_e$  for a large sample of stars; the latter is defined as the temperature of a black body whose rate of emission of energy per unit area per unit time is the same as that of the real star. While the relative frequency of stars per unit luminosity and temperature interval will depend on the way in which the sample is chosen, the leading features of the diagram are clear (Fig.1); the vast majority of stars lie on the main sequence; most of the others are giants. (If the sample had been drawn by volume of space, most of the main sequence stars would lie below the sun in the diagram and many more white dwarfs would have been represented.) Studies of binary stars show that an increase of the main sequence luminosity corresponds to a monotone increasing function of mass, with  $L$  going roughly as  $M^{3.5}$ . Since for a fixed and uniform

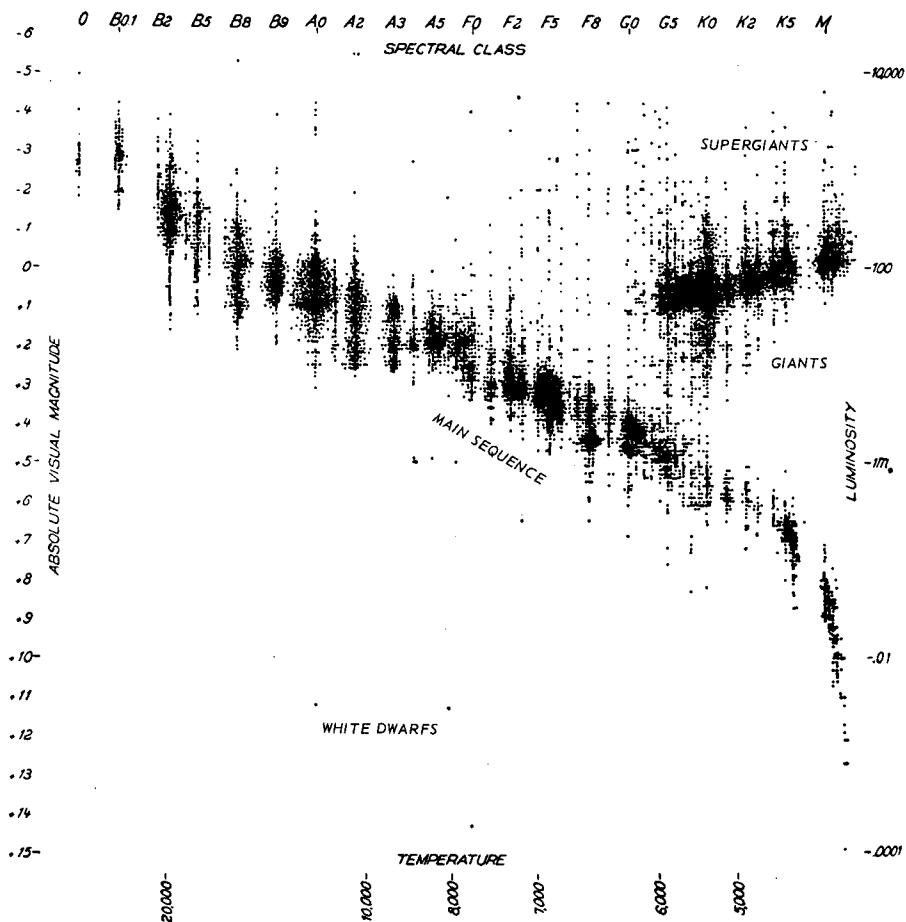


FIG. 1. HR diagram of stars limited by apparent magnitude (after Struve, Stellar Evolution, p. 32).

chemical composition the central temperatures and densities of self-gravitating fluid spheres in hydrostatic equilibrium increase with increasing mass, and since the reaction rates for converting hydrogen into helium are highly temperature and density sensitive, the main sequence is interpreted as a family of stars of different masses and nearly the same chemical composition burning hydrogen in equilibrium. Post-main sequence evolution comes about for any given star when it uses up something like 10% of its critical hydrogen supply; calculations of stellar interior models by Iben and others show that such a star moves into the giant or supergiant region of the diagram, living off gravitational contraction of the core, hydrogen burning in a shell, or if the central temperature gets high enough, He burning. The well-known interpretation of the HR diagram of galactic clusters by Sandage and others, as illustrated in Fig. 2, confirms the picture in a sample of stars of presumed equal age and chemical composition, and, if the hydrogen reaction rate is known

in a star of given mass, one is able to assign a nuclear age-date to clusters based on the turnoff of the brightest main sequence star from the main sequence. It is important to note that the location of the main sequence is, in principle, dependent on the chemical composition, not only through the hydrogen abundance itself, but also because at a given mass the central temperature and density depend on the mean molecular weight of the material. The especial relevance of this point to the interpretation of peculiar A stars will be noted later.

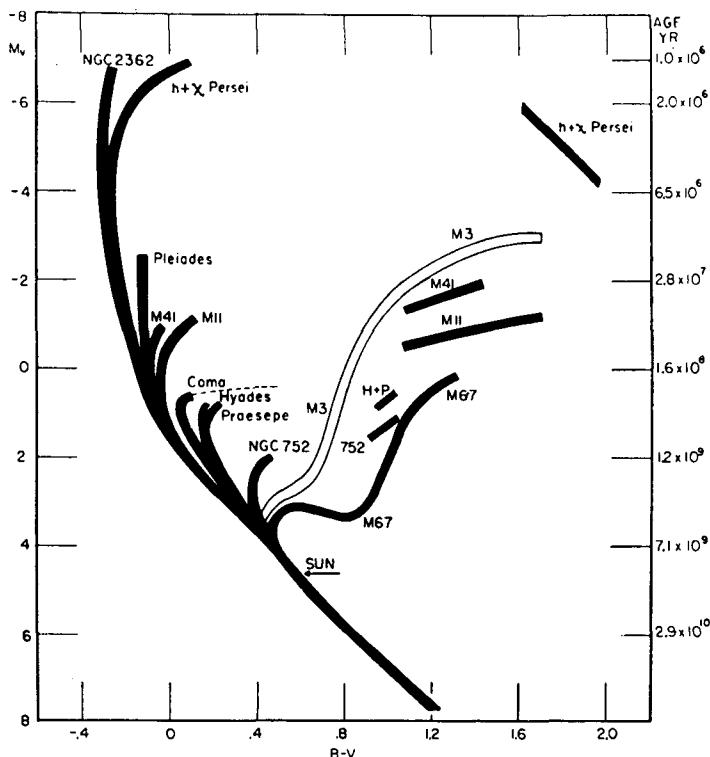


FIG. 2. Composite HR diagram of galactic clusters (after Sandage, *Astrophys. J.* 125, p. 436).

A post-main sequence evolutionary track computed by Iben for a star of mass  $5\mathcal{M}_\odot$  is shown in Fig. 3; the chemical composition is like that of the sun, namely,  $X = 0.71$ ,  $Y = 0.27$ ,  $Z = 0.02$ , the hydrogen, helium, and "metal" abundances by weight, respectively. In the long time-scales near the main sequence, one has hydrogen burning in a convective core which gradually involves a smaller and smaller fraction of the stellar mass; this is followed by overall contraction and the onset of hydrogen burning in a shell source whose thickness diminishes with time. Entering the red giant stage, one finds that the opacity of the envelope drops with decreasing envelope temperature, and the nuclear energy production of the hydrogen shell increases as the core contracts. The end

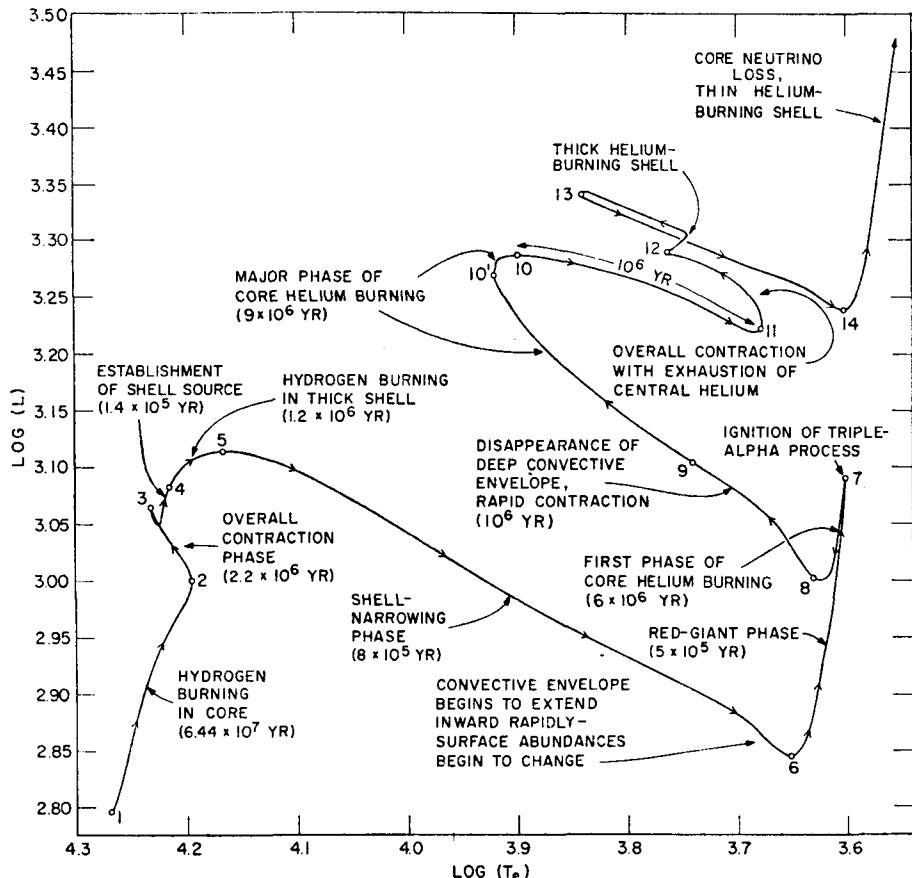


FIG. 3. Iben track for a star of mass  $5 M_{\odot}$  (A. Rev. Astr. Astrophys. 5 (1967) p. 573).

of the red giant branch is reached when He is fired with the triple- $\alpha$  reaction. This ignition is somewhat explosive in the sense that the central temperatures and densities at first rise to such an extent that the energy transport cannot carry the energy away and the central regions therefore expand. But this in turn causes the temperature and density of the hydrogen-burning shell to drop; since the shell still supplies the larger fraction of the total energy generation, the luminosity actually drops. The star now contracts, keeping the leading edge of the hydrogen shell at a temperature high enough to supply most of the energy production.

The evolutionary tracks as a function of mass for a fixed solar-like composition are illustrated also from Iben's work in Fig. 4. The points marked "1" define the theoretical zero age main sequence (ZAMS). As we shall see, a point of particular relevance to the peculiar A-type stars is the qualitative change-over in the shapes of the tracks above mass  $2.25 M_{\odot}$ . (According to Iben this change occurs essentially at mass  $2.25 M_{\odot}$ ) At point 5, the stars enter the red giant phase and move upward in luminosity to the onset of the triple- $\alpha$  reaction at point 6. The stars

of low mass increase their luminosities much more than the stars of high mass between points 5 and 6. According to Iben, this comes about because the low mass stars develop large electron degenerate cores before the onset of He burning. In the more massive stars, the core temperatures rise monotonically as the stars ascend the giant branch, and He burning sets in before electron degeneracy in the core becomes important. In the less massive stars, on the other hand, the central temperatures drop when the electron Fermi energy is of the order of  $kT$ .

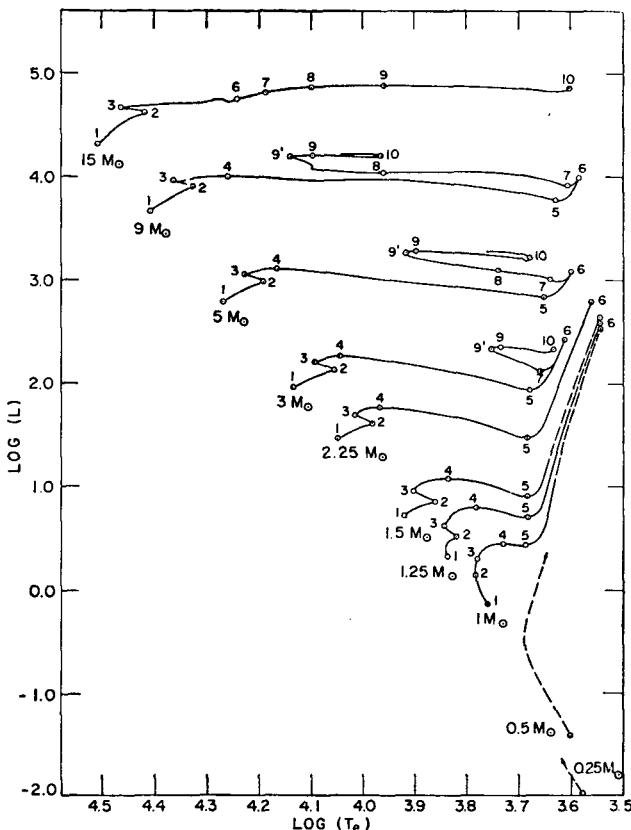


FIG. 4. Paths in the HR diagram for metal-rich stars of various masses (after Iben, *A. Rev. Astr. Astrophys.* 5 (1967) p. 585).

Once the core becomes isothermal owing to electron conduction, the central temperatures begin to rise again but much more slowly than in the large mass case, since electron conduction does not allow the temperature of the core to exceed very much the temperature of the hydrogen burning shell. Mestel suggested, and Schwarzschild and Härm demonstrated, that under the circumstance of core electron degeneracy, the onset of He burning was likely to be an explosive event. The significance of this for peculiar A-type stars will be returned to somewhat later.

A few additional remarks need to be made about the relationship between the theoretical and observational HR diagrams and the transformation of the one to the other. The preceding discussion has all been conducted in the theoretical plane, but astronomers do not observe directly either effective temperatures or luminosities. The latter are obtainable for some of the nearby stars by trigonometric parallaxes; in the case of the one nearest galactic cluster, the Hyades, the distance can be derived from the convergent point of the proper motions of the individual stars. Whether the latter is accurately enough determined has been a matter of hot debate in the last year. Beyond these determinations, stellar distances are unknown, except as they can be obtained by fitting spectra (or equivalent colour measures) into the HR diagram. In place of  $T_e$ , astronomers work with colours or spectral types based on line ratios that are sensitive to excitation and ionization in the stellar atmosphere. Colours are effectively spectral gradients determined by comparing the energy fluxes intercepted by broad-band filters separated by several hundred ångströms. We will have occasion to speak here of the UBV system; U, B, and V are magnitudes, i.e. fluxes expressed on a logarithmic scale with an arbitrary zero-point, centred at  $\lambda\lambda$  3600, 4400, and 5500 Å. Thus (U-B) and (B-V) are colours. The scale is set so that B-V = 0.0 corresponds to an A-type star with a temperature of about 9500°K and B-V = + 1.0 to a K-type star with  $T_e$  near 4500°K.

In any one group of stars with uniform chemical composition, as, for example, the stars of the Hyades cluster, it is possible from model atmosphere calculations to establish the relation between  $T_e$  and (say) (B-V); but this will be valid only for that group of stars. Thus two A-type stars with the same effective temperature might have different (B-V)'s if their metal abundances are different, even though the sources of continuous opacity are, in both cases, due entirely to hydrogen, i.e. the bound-free and free-free continua of both H and H<sup>-</sup>. This is because (B-V) and especially (U-B) are sensitive to the blanketing produced by metal lines. A further complication is that due to reddening by interstellar matter; this begins to be serious for all stars whose distances exceed about 75 to 100 pc.

Having disposed of the preliminaries in an undeservedly cursory fashion, let us turn belatedly to the peculiar A stars themselves. These objects, first isolated by Morgan some 35 yr ago, constitute about 5 to 10% of late B-, A-, and early F-type stars of the main sequence, i.e. they range with approximately main sequence luminosities from  $T_e$  near 8000°K to  $T_e$  near 15 000°K. Their generic defining properties are spectroscopic. When compared with normal stars of equivalent effective temperature, Ap stars exhibit certain lines which are abnormally strong or weak. Though the variety of types of anomalous line intensities is legion, several rough groupings are possible as a function of temperature. In the hottest Ap's, He is weak and Hg and P are sometimes strong; He is present predominantly as <sup>3</sup>He. In order of decreasing temperature, we come to the Mn stars, the Si stars, the Cr-Eu stars, and the Sr stars; in all cases the element named indicates that its lines are greatly enhanced in a star of that variety. Some samples from the original Yerkes Atlas (dispersion near 125 Å/mm) are illustrated in Figs 5 and 6. Two remarks are in order. First, it does not necessarily follow that the apparent

overabundance of some element in one star and the absence of its lines altogether in another type of Ap star exclude its existence in the latter. We do not know, for example, if He is underabundant in Cr-Eu stars because these stars are too cool to excite the lines of He I anyway; conversely, Preston points out that at the temperature of a hot peculiar A-type star, all Eu would cease to be Eu II and would instead be Eu III and Eu IV whose lines lie essentially in the inaccessible ultra-violet.

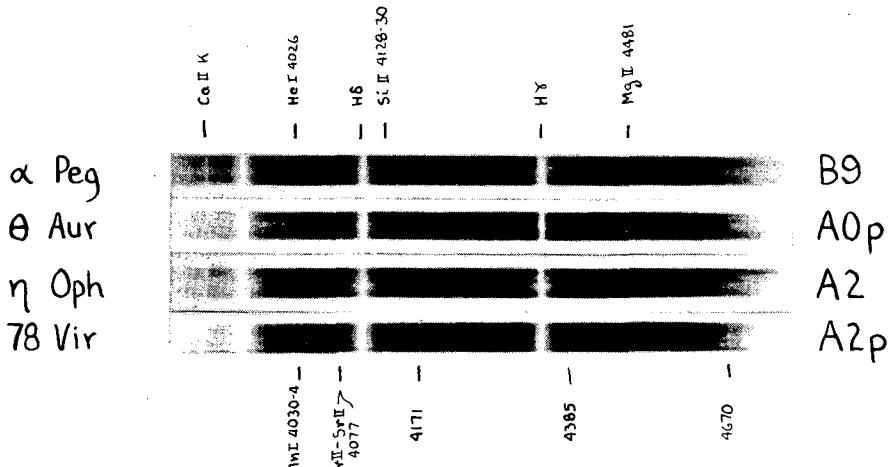


FIG. 5. The Ap stars  $\theta$  Aur (Si) and 78 Vir (Cr-Eu) (after Morgan).

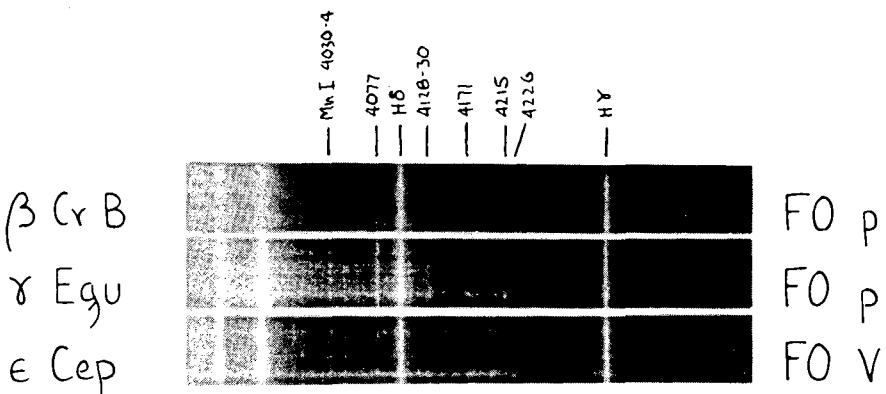


FIG. 6. The Ap stars  $\beta$  CrB (Cr-Eu) and  $\gamma$  Equ (Sr) (after Morgan).

A second point has to do with the dispersion of the spectrograms used to detect Ap stars. Some types of peculiar line intensities, for example those involving Y and some of the rare earths, would, because of blending, become apparent only at high dispersion, say 20  $\text{\AA}/\text{mm}$ . Moreover, because stellar rotation broadens the lines of many stars in this temperature range, some mild peculiar A-type stars have no doubt not been detected.

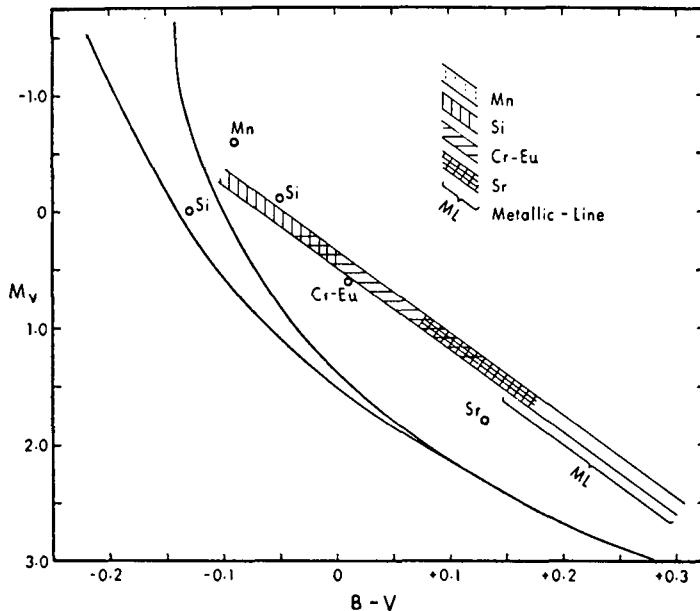


FIG. 7. The position of the Ap and Am stars in an  $M_V$ , ( $B-V$ ) plot (after Eggen and Fowler, the Burbidges, and Hoyle).

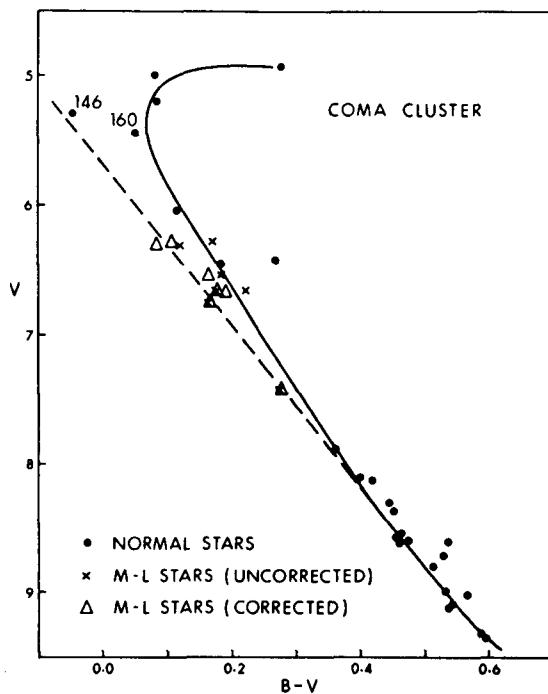


FIG. 8. Strittmatter-Sargent corrections for the positions of Ap and Am stars in Coma (Astrophys. J. 145, p. 135).

The position of the major classes of Ap stars are shown in the observed HR diagram, i.e.  $M_V$  versus  $(B-V)_c$ , in Fig.7, following Eggen and the Burbidges. The so-called metallic-line A's are also shown (Am's), another somewhat larger subgroup of anomalous A-type stars. These stars are characterized principally by the peculiarity that, for a given hydrogen line strength, the K-line of Ca II is too weak and the other metallic lines such as Fe and Ti are too strong. Eggen's photometry, taken at face value, shows a dichotomy in the positions of Ap and Am stars in the HR diagram, but in the theoretical diagram they definitely overlap, the more so with Conti's recognition of an extension of the class of Am's to higher temperature. The Am's bear a certain relation to the Ap's that will be discussed later; at this stage we merely point out that their spectra may have some mild characteristics similar to Ap stars, e.g. a small enhancement of some of the rare earth lines, but not all investigators agree about this.

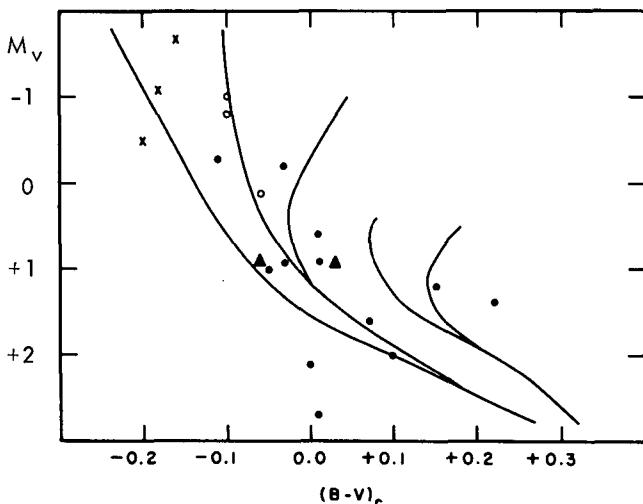


FIG. 9. Corrected positions of Ap stars in the HR diagram (after S. C. Wolff, *Astrophys. J.*, Suppl. 15, p. 34).

Figure 7 gives a misleading picture of the position of the Ap's in the physical HR diagram that plots effective temperature against luminosity because the  $(B-V)$  colours of Ap's are more seriously affected by the blanketing of metallic lines than are the colours of normal A's; this effect is statistically the stronger the cooler and fainter the star. Strittmatter and Sargent showed that, in the case of the Am's in galactic clusters such as Coma, Praesepe, and Hyades, the blanketing corrections move the stars actually somewhat to the left of the main sequence defined by the cluster stars themselves; this is not the same as the ZAMS because of the evolutionary effect among the more luminous stars of the cluster. This is shown for the Coma cluster in Fig.8; also shown are the two Ap stars in Coma, Nos 146 and 160. These are among the hotter Ap's and thus the correction for blanketing is rather small. The blanketing corrections for Ap's were considered also by Mrs. Wolff (see Fig.9), who

concluded that they were sufficient to bring the Ap's essentially into coincidence with that main sequence appropriate to the age of the particular star under consideration; she considered that too few Ap's are found in clusters to be absolutely certain that the corrections bring them to the left of the main sequence. I give high weight to the case of Coma, however, because the cluster is unreddened and unreddened differentially, and I assume in what follows that Ap's actually wind up slightly to the left of the main sequence, in analogy with stars of type Am.

The interpretation of the anomalous line intensities of peculiar A-type stars is rendered still more complex by the fact that in at least 25% of them the lines are variable in strength, in many cases in periodic fashion, and among the elements, *inter alia*, not necessarily in phase. Typical of the variations of line intensity in these so-called "spectrum variables" are those of  $\alpha^2$  CVn; the period is 5.5 d. The lines of Fe I, Fe II, Si II, Cr II, and Mg II vary in phase; the strengths of lines of Ti II, Eu II, Gd II, and Ca II vary together but are about 180° out of phase with respect to the others.

## 2. ROTATIONAL VELOCITIES

A further important, and related, observed property of Ap stars is the fact that they have relatively narrow spectral lines in a portion of the HR diagram characterized by stars with wide lines broadened by rotation. Abt and his associates recently studied the values of  $V \sin i$ , the projected rotational velocities, of the 63 brightest hot Ap stars, and concluded that these velocities averaged only 22% of normal stars in the same temperature range. The latter run from an average value of 180 km/s at the hot end to 150 km/s at the cool end of the Ap range under consideration. There are two ways in which this result can be interpreted. Either the Ap stars represent a group of truly slow rotators or they constitute the subset of A-type stars viewed "pole-on". The former hypothesis appears now to have rather incontrovertible support from the work of Abt and his associates who found that the frequency function of  $V \sin i$  is incompatible with the latter hypothesis, and thus if both normal and peculiar A-type stars have a random orientation of rotational axes, there are no normal stars with  $V < 100$  km/s and no Ap stars with  $V > 150$  km/s.

Further support for this view comes also from Deutsch's study of the frequency function of  $V \sin i$  for A-type stars, based on his assumption that the function is approximately Maxwellian, which seems observationally justified from studies of stars of other spectral types. It is worth noting here that the Am's are also characterized by narrower than average lines. It has been found by Abt, however, that the vast majority, perhaps all, of the Am's are spectroscopic binaries with periods mostly of the order of days. Available information shows, however, that this is not true for Ap's. Small  $V \sin i$  for the Am's must, however, mean that their rotations are truly small because the orientation of the orbital angular momentum vectors is at random; this in turn means that it is not reasonable to suppose that their rotational angular momentum vectors are preferentially aligned.

Now if the rotations of both Am's and Ap's are truly slow, and if, as the statistics suggest, they constitute the entire group of slow rotators in

the domain of the HR diagram characterized by rapid rotation, then two further observed facts about Ap's in general and about the subset of periodic spectrum variables in particular fall rather nicely into place. A rather considerable amount of work has been done in the past few years by Sweet and Roy, by Roxburgh and Strittmatter, by Roxburgh, Griffith, and Sweet, by Strittmatter and Sargent, and by Collins and others on the structure and atmospheric properties of rotating stellar configurations. These authors in general agree that, in the range of temperatures appropriate to Ap's and Am's, rotation moves a star to the right in the HR diagram by an amount that is roughly proportional to  $V^2$ , and one would expect a rotator therefore to have (B-V) a few hundredths of a magnitude redder than a non-rotator. Results of such calculations by Roxburgh and Strittmatter, for example, are shown in Fig.10; the quantity  $\Omega^2$  in these calculations is the ratio of centrifugal force to surface

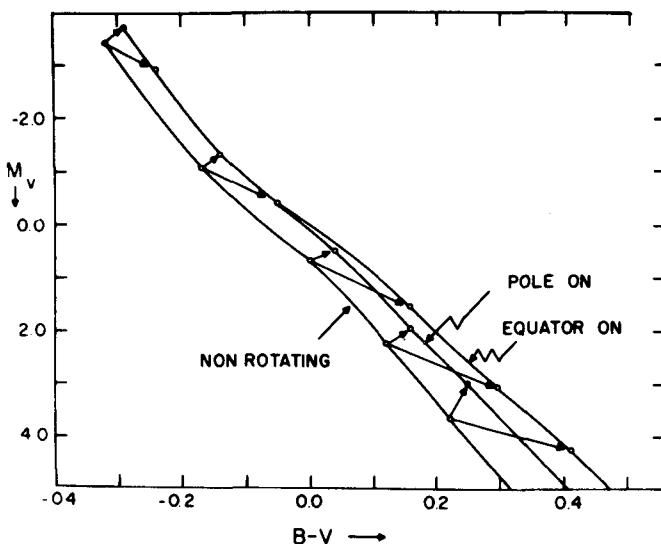


FIG. 10. Stellar rotation vectors in the HR diagram (after Roxburgh and Strittmatter).

gravity at the equator. The deblanketed position of the Am's and Ap's, i.e. the fact that they lie slightly to the left of the main sequence defined by ordinary stars, is completely explicable on the basis of this diagram. The other observed fact explained by slow rotation in Ap's is the period-line width relation among periodic spectrum variables, namely,  $V \sin i \leq 125/P$ , where the period of the spectrum variation is in days. This is illustrated from the original paper by Deutsch in Fig.11. The point here is that the constant in the relationship, 125 km/s, is an upper limit to the rotational velocities, and this is already only about 70% of the average rotation for A-type stars. But more is implied by this result than simply slow rotation for spectrum variables; it is clear that the spectral line strengths are being modulated by rotation. In other words, there is inescapable evidence that the thermodynamic and/or abundance

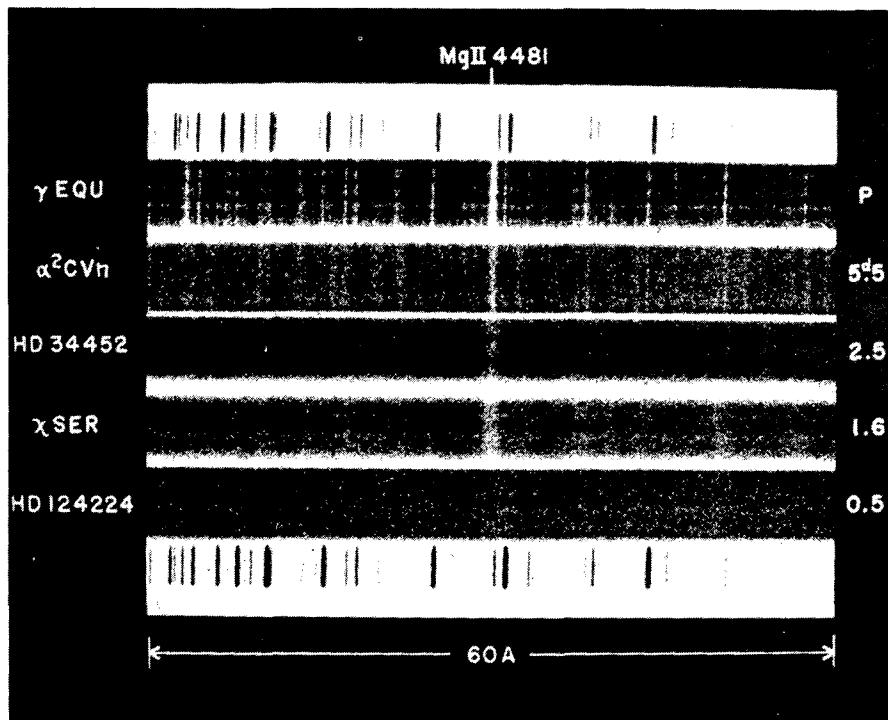


FIG. 11. Period-line width relation for spectrum variables (after Deutsch).

parameters in Ap stars are not isotropic and that spectroscopic "patchiness" is required. Most investigators now agree that the anomalous line strengths found in these stars must be a result of truly unusual abundances of certain nuclear species; there is no universal agreement, however, on whether the patchiness is a result of a non-uniform distribution of elements over the surface or merely reflects a peculiar spatial non-uniformity in the temperature and pressure.

### 3. MAGNETIC FIELDS

We have so far established that when anomalous line blanketing and rotation are taken into account, the positions of the Ap stars in the HR diagram agree with those of normal main sequence stars, that the Ap's and Am's constitute the large majority in the population of truly slow rotators in that part of the HR diagram characterized by rapid rotation, and that the changes in the line intensities in the subgroup of spectrum variables is a result of the existence of surface patches which are modulated by the rotation. Remaining to be discussed are the large-scale coherent magnetic fields discovered by Babcock more than 20 yr ago, and whose existence is virtually a unique property of peculiar A-type stars.

Observations of stellar magnetic fields, first carried on by Babcock but more recently mostly by Preston, depend on the measurement of the

longitudinal Zeeman effect at the Coudé focus of a large reflector. Large spectroscopic scale, excellent resolution, and thermal and structural stability of high order are essential for success in the observations. A differential circular analyser, consisting of a calcite block and a mica quarter wave plate, permits the two  $\sigma$ -components, circularly polarized in opposite senses, to be displayed one above the other, sandwiched between two laboratory sources of Fe I in emission; differential measurements of displacement of magnetically sensitive lines yield directly a Zeeman shift that is proportional to an effective magnetic field strength. The phase shifts arising from the oblique reflections in the Coudé optical system are compensated before the light reaches the spectrograph slit (Fig.12).

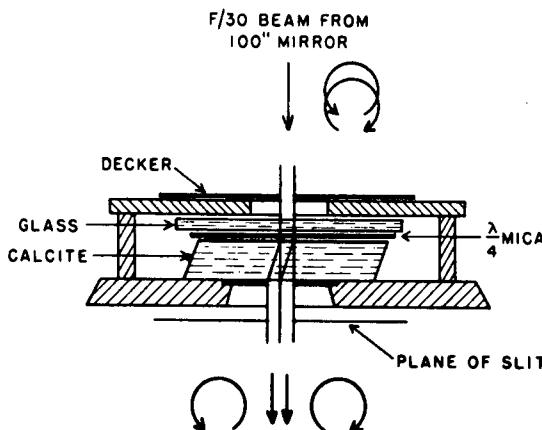


FIG. 12. Differential analyser for circularly polarized light (after Babcock, Stars and Stellar Systems, p.115).

In a purely longitudinal field, the two  $\sigma$ -components of a normal Zeeman triplet are shifted by the same amount on each side of the normal position of the line; thus if  $\Delta s$  represents the shift of the dextrogyrate component in microns,  $F$  the dispersion in  $\text{A/mm}$ ,  $\lambda$  the wavelength, and  $H$  the field strength in gauss, we have

$$H = 1.07 \times 10^9 F (\Delta s / \lambda^2)$$

where the dextrogyrate component has the longer wavelength when the magnetic vector points towards the observer. In a real star, the observed spectrum is much more complex because, first, the field is not purely longitudinal and thus  $\sigma$ - and  $\pi$ -groups of components appear in both spectra with different intensities, second, the field is non-uniform, and third, the spectral lines are broadened thermally and by rotation. Thus only relative displacements of centres of gravity of the lines produced in the two spectra can be measured. Babcock denotes by "z" the mean displacement of the  $\sigma$ -group of components to one side of the centre in units of the displacement of the normal triplet; this quantity can be calculated for each line once the g-factor is known. LS coupling is usually assumed. Since

a large number of lines are measured, an effective field  $H_e$  is defined from all the lines by

$$H_e = 1.07 \times 10^9 F \frac{\sum \Delta s / \lambda^2}{\sum z} \text{ gauss}$$

In Figs 13 and 14 we show respectively a sample spectrum of the Ap star 53 Cam, obtained with an analyser at Palomar, and a plot of displacements against  $z$ -value; the latter relationship shows clearly that a longitudinal magnetic field is in fact being measured. The effective fields that have been measured range from a few hundred gauss to 34 kG for HD 21544; it should be noted, however, that a field of 300 G corresponds on the plate to a shift of only about  $2 \mu\text{m}$  at a dispersion of 5 A/mm. Thus it cannot be overemphasized that detection of stellar magnetic fields is confined to sharp-lined stars. Among A-type stars, this means that fields could reasonably be expected to be detected in fact only among the intrinsically slow rotators, which are the Ap stars themselves, or among the rapid rotators seen pole-on. Yet, to the best of my knowledge, no sharp-lined normal star, which presumably falls in the latter category, shows the existence of a magnetic field. This is strong, but not absolutely conclusive, evidence for believing that the existence of a magnetic field is a necessary and sufficient condition for a star to be a peculiar A-type star.

Magnetic fields have been detected in about 100 bright stars; the vast majority of these are Ap stars. The literature of this subject is highly detailed and cannot be covered in a short review. But a few points are relevant. First, it appears from the work of Preston and Steinitz that the fields of Ap stars are either constant, within the errors, or are periodically variable. Babcock's earlier division of the variability into periodic and irregular resulted from limitations in the observations. Evidently each magnetic cycle, normally involving a reversal of polarity, is not necessarily exactly the same in amplitude and phasing with respect to all prior or subsequent cycles. Thus attempts to assemble magnetic observations from different cycles to one period were frequently unsuccessful. These phase shifts imply long-term secular variations in the position and strengths of the fields, and it is noteworthy that similar variations have been found by Deutsch in the variations of line strengths of periodic spectrum variables.

Second, it appears that in every case of detection of a variable magnetic field in a spectrum variable the periods of the two phenomena are exactly the same. Thus, in view of our earlier conclusions, it is clear that the magnetic field is also modulated by the rotation of the star, and the spectroscopic patches are no doubt related to the inhomogeneities of the field.

Third, variations in the colours and magnitudes of Ap stars have been known since the pioneering work of Provin and later of Jarzebowski, but a recent extensive survey by Stepien shows that periodic variations in light and colour of very small amplitude are characteristic of magnetic variables, and these variations have the same period as the magnetic field. Normally, the amplitudes do not exceed a few hundredths of a magnitude. If the variations in brightness were directly related to a dipole magnetic field distribution over the surface, then from symmetry arguments one might have expected the period of the light variation to be one-half that of the rotational

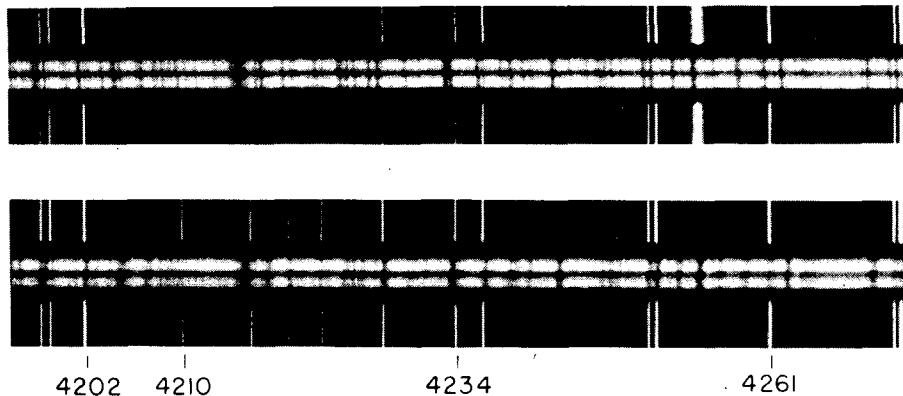


FIG. 13. Magnetic spectra of 53 Cam (after Babcock, Stars and Stellar Systems, p.120).

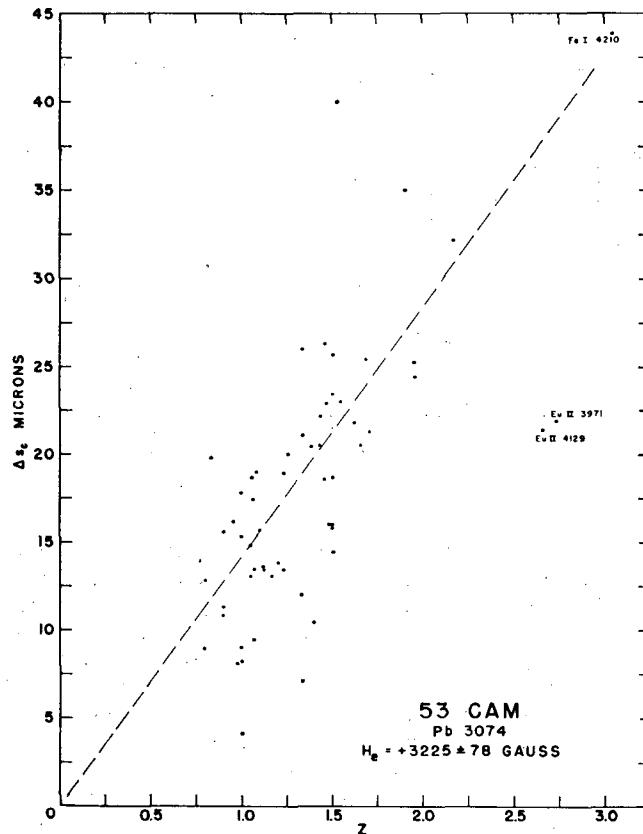


FIG. 14. Shifts in microns versus  $z$ -value for 53 Cam (Stars and Stellar Systems, p.121).

or magnetic period. The observations suggest that, in accordance with the theory of Mestel, the dipole field distorts the star into an oblate spheroid, and the light variations result from the tumbling of such an object with concomitant variations of effective gravity and boundary temperature over the surface. Figure 15 gives an example of the variations found by Stepien in the case of HD 153882; the magnetic variation is that measured by Preston and Pyper.

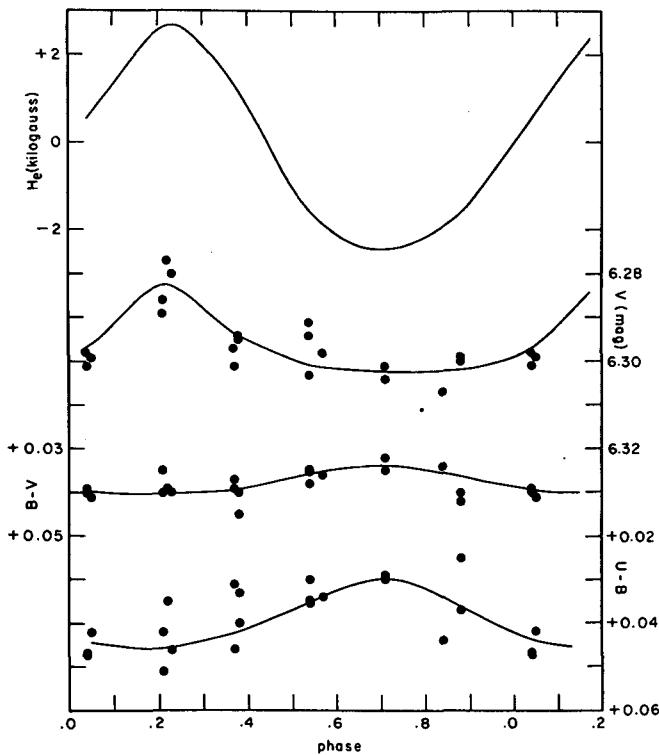


FIG. 15. Light, colour, and magnetic field variations of HD 153882.

Fourth, it appears reasonable from the preceding, and indeed such a view has been advanced for a number of years by Deutsch and more recently by Ledoux and Renson, and others, that the magnetic variations of Ap stars can be understood in terms of a magnetic dipole, rigidly rotating with the star, but with the magnetic axis inclined to the rotation axis, a model no doubt inspired by the terrestrial, Jovian, and solar examples. The dipole is probably rather distorted since the magnetic variations are often rather far from sinusoidal, but most people who subscribed to this picture presumably imagined that the magnetic "axis" was not greatly inclined to the rotation axis. Preston and Sturch showed, however, in the case of  $\beta$  CrB, that the strongest part of the field lay essentially in the plane of the rotational equator. They pointed out that the rotational velocities of very sharp-lined Ap's (of which  $\beta$  CrB is an example) have been seriously

overestimated because most of the lines are distinctly broadened by magnetic intensification. To correctly estimate  $V \sin i$  in such objects, one must seek out the "null" lines, i.e. those lines for which  $z = 0$ . In Fig.16 is illustrated the null line  $\lambda 4065$  of Fe I in  $\beta$  CrB. From this, an upper limit to the projected rotational velocity  $V \sin i \leq 2$  km/s was estimated. On the oblique

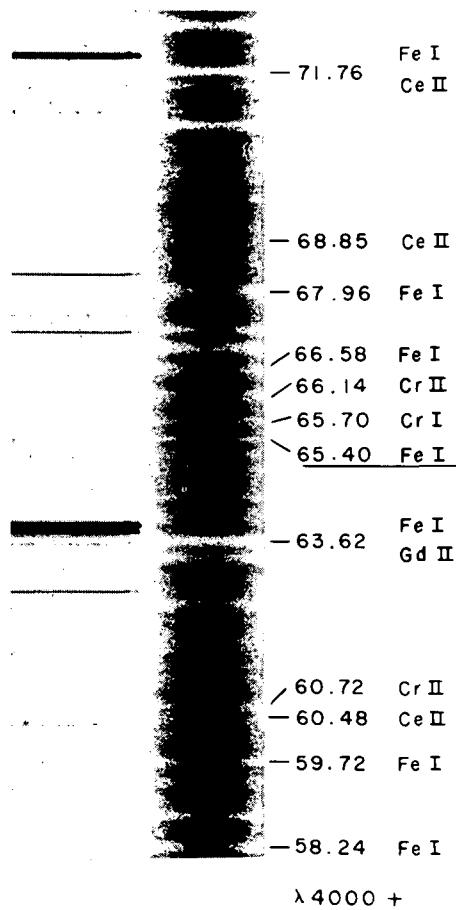


FIG. 16. Null line of Fe I in  $\beta$  CrB (after Preston and Sturch, Magnetic Stars, p.117).

rotator model, the magnetic and spectrum variation period of 18 d leads to  $V = 6$  km/s, so it follows that  $\sin i \leq 1/3$  or  $i \leq 20^\circ$ . In other words, we see the object nearly pole-on, i.e. almost along the rotation axis. Clearly if the magnetic axis were inclined only slightly to the rotation axis, we could have no field reversal, yet the field varies from -800 to +1000 G. If  $\beta$  is the angle between the rotation axis and the magnetic dipole axis, Preston and Sturch estimated  $86^\circ < \beta < 90^\circ$ .

On the assumption of a random orientation of rotational axes for Ap's, Preston later showed that the observed frequency distribution of

$R = H_e(\text{min.})/H_e(\text{max.})$  for 15 stars was compatible with  $\beta = 80^\circ$ . Despite the limited size of the sample, there seems little doubt that the rotational model is compatible with the dipole lying more or less in the equatorial plane, but the axis is conceivably seriously twisted. The observed and computed distribution functions of  $R$  for various values of  $\beta$  are illustrated in Fig. 17.

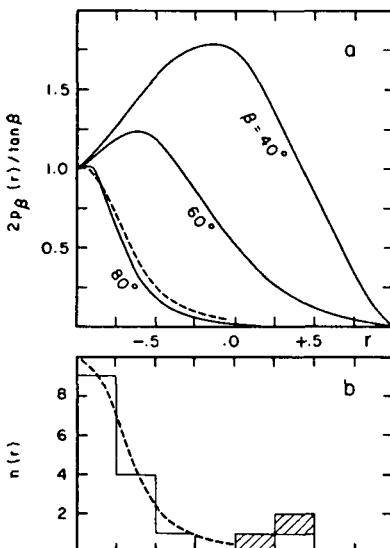


FIG. 17. The distribution functions of  $R$  for various values of  $\beta$  for Ap stars with variable  $H$  (after Preston, Astrophys. J. 150, p. 548).

It must not be supposed that the rigid rotator model does not encounter some observational and especially theoretical objections; these are reviewed in a recent article by Ledoux and Renson. But the work of Preston and his associates and of Steinitz and others has made it increasingly clear that the weight of observational evidence is now strongly on the side of this model, despite some complications related to the small radial velocity changes and the cross-over effect.

#### 4. SUGGESTED EXPLANATIONS OF THE ABUNDANCE ANOMALIES

We have seen that the Ap stars are characterized by anomalous line intensities, slow rotation, and the presence of large-scale coherent magnetic fields lying probably in the equatorial plane. Some show variations of line intensities, magnetic fields, and small variations of light and colours; these are periodic and concomitant with one another. If the light variations in fact result from a distortion of the geometrical shape of the star by the field, then an attempt should be made to detect light variations in broad-lined normal stars; these should not exist if, in fact, the magnetic fields characteristic for Ap's and slow rotation are genetically associated.

We return now to the anomalous line intensities, their interpretation, and the relation of this to stellar evolution and the presence of the magnetic field. There seems to be no doubt that the anomalous line intensities are related to anomalous abundances; these are summarized by Fowler, the Burbidges, and Hoyle in the plot shown in Fig.18. The dashed line shows the normal abundances, and the vertical lines indicate the abundances of the elements in those stars in which, in fact, that particular abundance

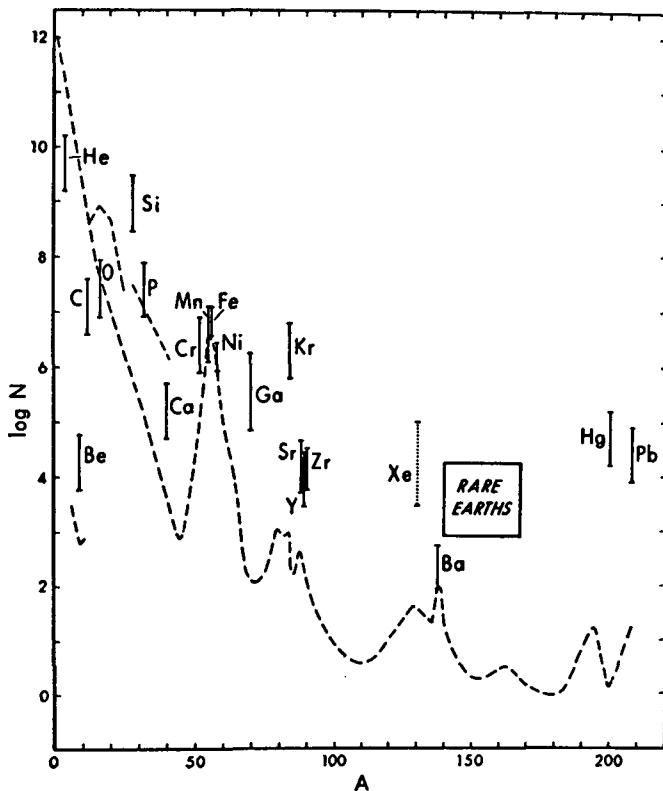


FIG. 18. Schematic abundance curve of Ap stars (after Fowler, the Burbidges, and Hoyle).

anomaly occurs (we note that not all the anomalies occur together in the same star!). Because of magnetic intensification, the possibility of abundance patches, and uncertainties with some of the atomic parameters such as ionization potentials, only the order of magnitude of some of the abundances can be trusted, but the broad outlines cannot be doubted. The leading features of the diagram are the underabundance of such light elements as He, Be, and C, the slight overabundance of Fe-peak elements, the enormous overabundance of such heavy elements as Kr, Sr, Y, Zr, and probably Xe, the overabundance of rare earths, and the apparently normal abundance of Ba. We have already mentioned that He seems mostly to be present as  $^3\text{He}$ , and Wallerstein and Merchant have shown that in at least two stars  $^6\text{Li}$  is unusually abundant relative to  $^7\text{Li}$ .

In an earlier treatment of the subject in what is by now a classical paper, Fowler and the Burbidges proposed that the acceleration of fast protons by surface magnetic fields concentrated in every limited area would give rise to surface nuclear reactions of the (p, n) variety, and that the liberation of neutrons from light elements would permit the building-up of heavy elements from the Fe-peak group. The addition of neutrons in this picture occurs on the slow time-scale (*s*-process), with a relatively low neutron flux liberated in a thin outer layer of the star. A major difficulty with the theory is the problem of the normal abundance of Ba and the overabundance of Eu. The former is very stable against neutron capture, as are Sr, Y, and Zr; therefore if the latter cannot be by-passed in a process of slow neutron addition, and thus build up in abundance, why is not Ba also built up? And the converse situation exists for Eu for which neutron addition is highly probable.

This problem, plus the later recognition of the very small abundance of  $^4\text{He}$  in 3 Cen A led to a reconsideration by Fowler, the Burbidges, and Hoyle. They proposed a combination of highly energetic surface spallation, *s*-processing and interior *r*-processing, that is, neutron addition with large fluxes. The first is designed to "spall the surface clean", so to speak, destroying virtually all the  $^4\text{He}$  by proton bombardment, and leaving a layer containing largely spallation products, namely, protons, neutrons, and deuterium. Various processes such as  $\text{D}(\text{p}, \gamma)^3\text{He}$ ,  $\text{D}(\text{d}, \text{n})^3\text{He}$  and  $\text{D}(\text{d}, \text{p})\text{T}(\beta^-)^3\text{He}$  are considered to produce  $^3\text{He}$ ; some further discussion of the surface mixing processes is required to get the right ratio of the He isotopes. At the same time, it was recognized that such surface processes cannot solve the Ba, Eu problem, and what is needed is a large neutron flux. If one *r*-processes on the  $\beta$ -active parents of Sr and Ba, for example, they will capture another neutron before they have had a chance to  $\beta$ -decay, and thus some stage in which *r*-processing is possible seems essential. The above authors then refer to interior nuclear processes, and in particular to the red-giant stage of stellar evolution mentioned at the outset of this lecture. It is proposed that the sudden removal of degeneracy concurrent with the onset of the He-flash in red giants of mass less than  $2.25 M_{\odot}$  triggers the neutron flux on a rapid time-scale, and thus peculiar A-type stars are the descendants of low mass red giants which have moved back into the vicinity of the main sequence after having mixed the products of *r*-processing into the spalled surface regions. This is admittedly an extremely cursory treatment of an extensive and ingenious theory, but neither time nor the scientific qualifications of the speaker permit a more critical review.

Without criticizing nuclear physics at all, and admitting fully the importance of the need for *r*-processing, I think it must be admitted that this aspect of the theory is incompatible with what we think we know about stellar evolution, even though the so-called "facts" are not by any means incontrovertible. The first difficulty is this. If the Ap stars are really post red giants whose original masses did not exceed  $2.25 M_{\odot}$ , then, even aside from the improbably remarkable coincidence that their blanketing and rotational rectified positions in the HR diagram bring them exactly onto the main sequence, we are forced to the conclusion that they cannot exist presently in galactic clusters with a break-off at a main sequence mass exceeding  $2.25 M_{\odot}$ . Yet Ap stars are present in at least five clusters in which the break-off mass is around  $3 M_{\odot}$  or considerably higher. These

are shown in Figs 19 to 23. For NGC 6633 and NGC 2281 (Figs 19, 20), the break-off masses are near  $3 M_{\odot}$ , in IC 2391 and in the  $\alpha$  Per cluster, the break-off is near  $8 M_{\odot}$ , and in the Sco-Cen Association, the break-off is near 15 to  $20 M_{\odot}$ .

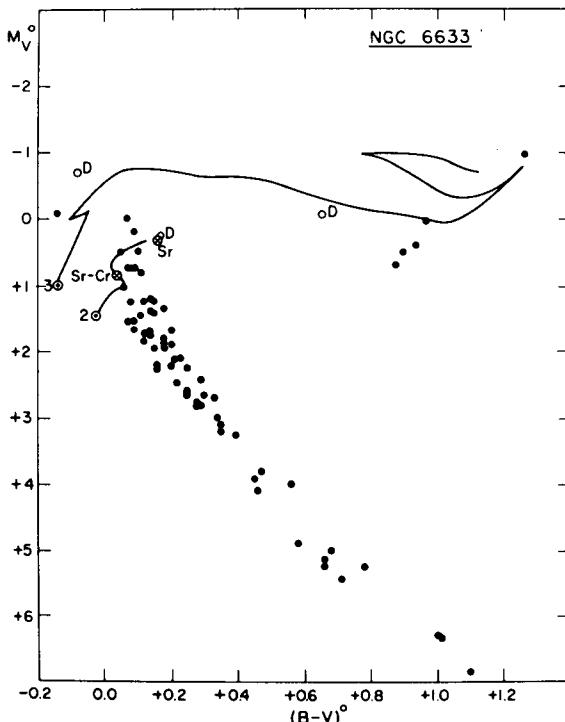


FIG. 19. HR diagram of NGC 6633 (after Johnson, Hiltner, and Iriarte; and Kraft).

Second, if the Ap stars are post red giants that have undergone the He flash with internal mixing, then they must certainly have a mean molecular weight different from stars on the main sequence for the first time, and they would not satisfy the same mass-luminosity relation. To investigate this question, we might look for binary stars with Ap components. Visual binaries are not too suitable for the purpose, since their masses depend on some sort of parallax, usually obtainable not directly but rather from some main sequence fitting procedure which in turn already presupposes some knowledge of the composition. But for spectroscopic binaries this is not true, provided the inclination of the orbit can be found or estimated. Recently, the remarkable magnetic star HD 98088 has been analysed by Abt, Conti, Deutsch, and Wallerstein. The star is a double lined spectroscopic binary with an orbital period of 5.9 d, and this is also the period of the magnetic variation, the spectrum variation, and the rotational period of the Ap primary. The synchronism between these and the orbital period is unique. The system regrettably does not eclipse, and this leads only to lower limits on the masses of the com-

ponents; the minimum masses of the Ap star and the A8 V secondary are  $1.7 M_{\odot}$  and  $1.3 M_{\odot}$ , respectively. If the inclination of the orbit is taken from the condition that the secondary is normal and satisfies the mass-luminosity relation, then  $i = 67^\circ$  and the mass of the primary becomes  $2.2 M_{\odot}$ , which is exactly right for its hydrogen line-based position in the HR diagram. Though there is some evidence that the secondary is a metallic line star, the masses of these objects are known to be normal, and thus the normal mass of the Ap primary would still be maintained.

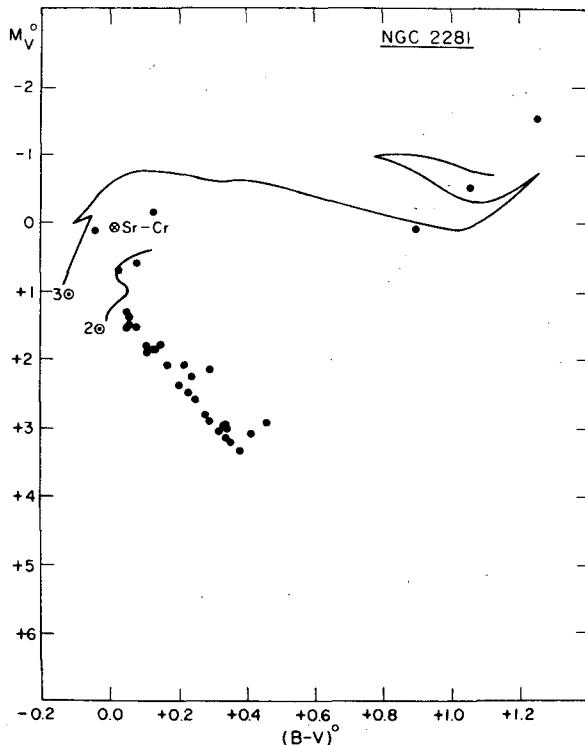


FIG. 20. HR diagram of NGC 2281 (after Pesch, and after Kraft).

If the Ap star had been, for example, an inhomogeneous star with a He core, then from the work of Faulkner and Iben its mass would have been about  $1.2 M_{\odot}$  which is already less than the minimum orbital mass. On the other hand, from the observational side, since the period and surface rotational velocity are known, the inclination of the orbit might be obtainable from the observed width of some line with a null Zeeman width, if such a line can be found in a hot Ap star.

An ingenious suggestion to avoid these difficulties was first offered by Fowler, the Burbidges, and Hoyle and later elaborated by Vanden Heuvel. They proposed that the present Ap stars are the original secondaries of fairly close binary systems in which the original primary, of mass less

than  $2.25 M_{\odot}$  evolved off the main sequence, enlarged its radius, passed through the He flash, either before or after encountering the critical inner Lagrangian surface, and thus spilled the r-processed contaminated material onto the secondary. The latter then evolved up the main sequence into the present Ap and became the new primary of the system. The picture does not unfortunately stand up too well in the face of both theoretical and observational objections. In the first place, Kippenhahn, Kohl, and Weigert have made evolutionary calculations of a system of this kind in which the

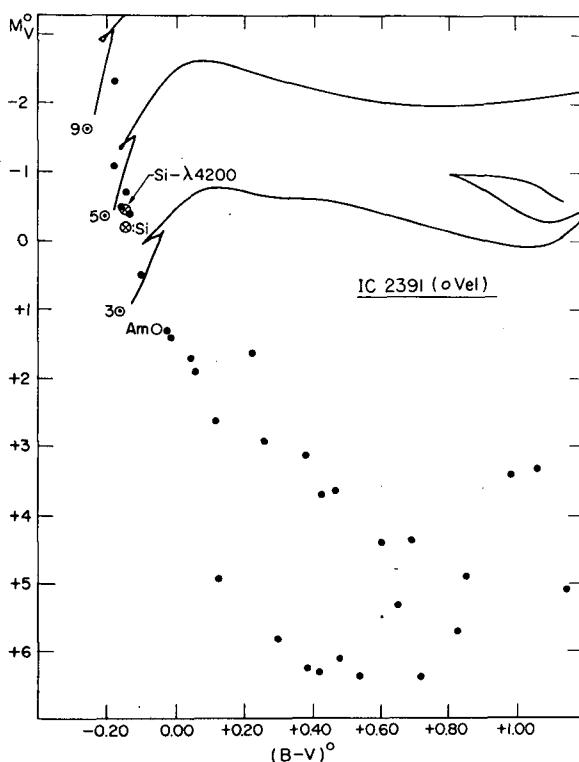


FIG. 21. HR diagram of IC 2391 (after Searle and Hyland, and Kraft).

masses of the original stars are  $2 M_{\odot}$  and  $1 M_{\odot}$ . As soon as the primary encounters the surface, it spills matter very rapidly to the secondary; when the primary returns to the immediate vicinity of the main sequence its mass has already been reduced to  $0.3 M_{\odot}$ , and the mass of the secondary has been increased to  $2.7 M_{\odot}$ . This does not fit well with the observed mass ratio of HD 98088. In the second place, we know of some types of eclipsing binary stars in which the present secondary (former primary) has been caught while it was in the red giant or subgiant stage. These invariably have masses too small for their luminosities and fill the critical inner Lagrangian surface in accordance with the theory of Kippenhahn et al. and with the earlier suggestion of Crawford. Such stars

would quickly evolve into Ap binaries if the above picture is correct. An observed feature of these mass transfer binaries is that their eccentricities are invariably close to zero. But, as Lucy has pointed out, the alleged Ap binary descendants are characterized by eccentricities that, as a class, are far from zero, a property which they share in common with Algol-type binaries recognized as objects on the main sequence for the first time.

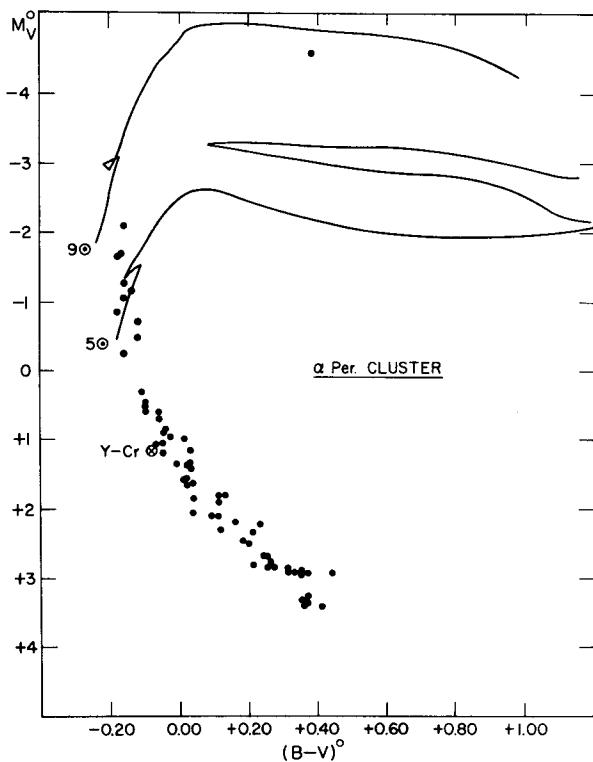


FIG. 22. HR diagram of the  $\alpha$  Per Cluster (after Kraft).

The only remaining possibility seems to be that of postulating white dwarf companions for Ap stars, following Vanden Heuvel. In this picture the present white dwarf has previously spilled its r-processed material over to the present Ap star; this is in general consistent with the Kippenhahn evolutionary process. But this does not explain systems such as HD 98088. The secondary in this case is definitely not a white dwarf. Stability arguments suggest that if this star were in fact a triple system, the white dwarf would be at a considerable distance from both components and might very well not have been large enough at any point in its evolution to have filled the critical surface. Moreover, in this picture both visible components would have to be Ap's, but this does not seem to be the case in HD 98088.

More recent discussions on the nuclear physics of the problem have been advanced by Brancazio and Cameron (1967) who advocate a return of the surface spallation mechanism as the entire source of the anomalous abundances. They demonstrate that in a gas with the normal cosmic abundance distribution, bombardment with  $\alpha$ -particles distributed after an inverse power law spectrum in the 1-40 MeV range can produce the overabundance of rare earths and other elements in Ap stars, including normal abundance for Ba.

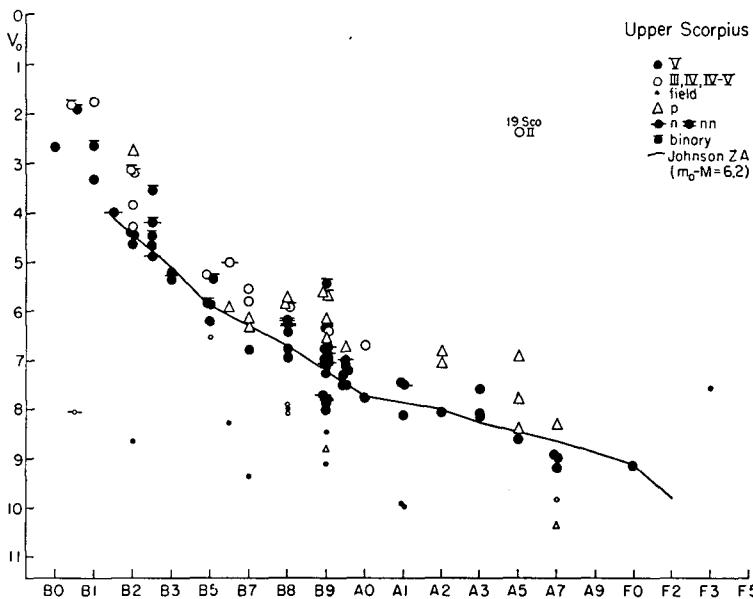


FIG. 23. HR diagram of the Sco-Cen stream (after Garrison).

We have seen that a self-consistent picture of the place of Ap stars in stellar evolution cannot at present be offered. The existence of a large scale ordered magnetic field, the slow rotation, and the evidence of spalled material could perhaps be made internally consistent if we postulated that the magnetic fields of Ap stars somehow drag on the ionized interstellar matter in the vicinity as the star rotates. But this does not seem to be capable of explaining the significance of the r-processed material, and the explanations that have been offered seem to run counter to stellar evolution. The binary hypothesis also runs into fatal objections. In short, the problem may be described as a typical astronomical problem, fascinating, irritating, and mildly perverse.



# STELLAR POPULATIONS AND THE EVOLUTION OF THE GALAXY

R.P. KRAFT

Lick Observatory, University of California,  
Santa Cruz, Calif., United States of America

## Abstract

STELLAR POPULATIONS AND THE EVOLUTION OF THE GALAXY. The author, in discussing the evolution of the galaxy, refers to the work of Baade and the questions of metal and He abundance.

The concept of stellar populations was initiated nearly 25 years ago by Baade with his discovery of the essential dichotomy in the HR diagrams of globular clusters and elliptical galaxies on the one hand and, on the other hand, of the stars in the vicinity of the sun moving slowly with respect to the local standard of rest. This is illustrated in Fig. 1 from Baade's original paper. The shaded area corresponds to the majority of solar vicinity stars. To this group of the common bright stars Baade gave the name Population I. The O and B stars and other young stellar objects belong to this group. To the stars of the hatched area, recognized as extending also down along the main sequence, Baade gave the name Population II. Its giants are characterized as being on the average brighter than the Population I giants, there exists a "horizontal branch" to which the RR Lyr or globular cluster short period cepheids belong, and there are no ordinary O or B type stars. The bright stars of Population II must be judged mostly as of late type.

In the original definition, the distinction was based primarily on the dichotomy in the HR diagrams and on the position within our own galaxy. Many of the globular clusters belonged to a spherical galactic halo but the Population I stars belonged to the galactic disk and especially to the spiral arm regions. Later, it became clear that the distinction was also correlated with kinematic characteristics and with the spectroscopic characteristic of "weak-lined"-ness. In other words, it was found that the stars of Population II exhibited, on the average, weaker metal lines at a given effective temperature than did stars of Population I. This was interpreted by Greenstein, Wallerstein, Kinman, Roman, Schwarzschild, and many others as indicating a decline in metal abundance relative to hydrogen by factors ranging from 10 to 1000 compared with the sun and other Population I stars. At the same time it was recognized that the weak-lined stars showed a strong tendency to have kinematical characteristics different from those of the spiral arm population; the mildly and moderately metal-weak stars seemed to be associated with motions lying more or less in the galactic disk but often with velocities differing from the local circular velocity, as if they were in highly elliptical orbits about the galactic nucleus. Many of the extremely metal-poor stars were found to have large velocity components at right angles to the galactic plane, in what is known as the Z-direction.

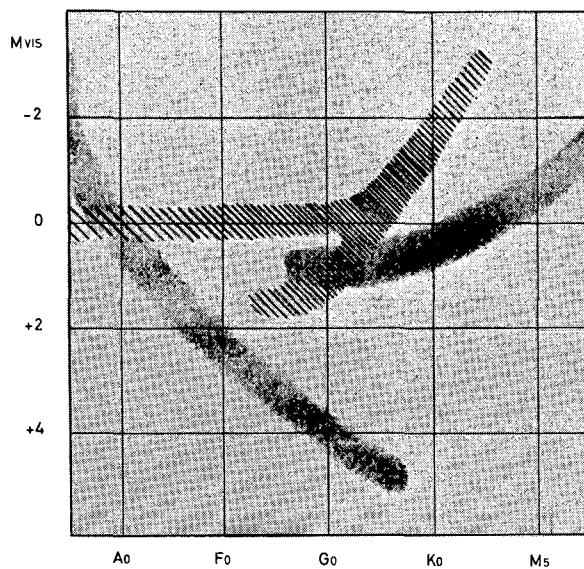


FIG. 1. The HR diagrams of Populations I and II (after Baade, 1944).

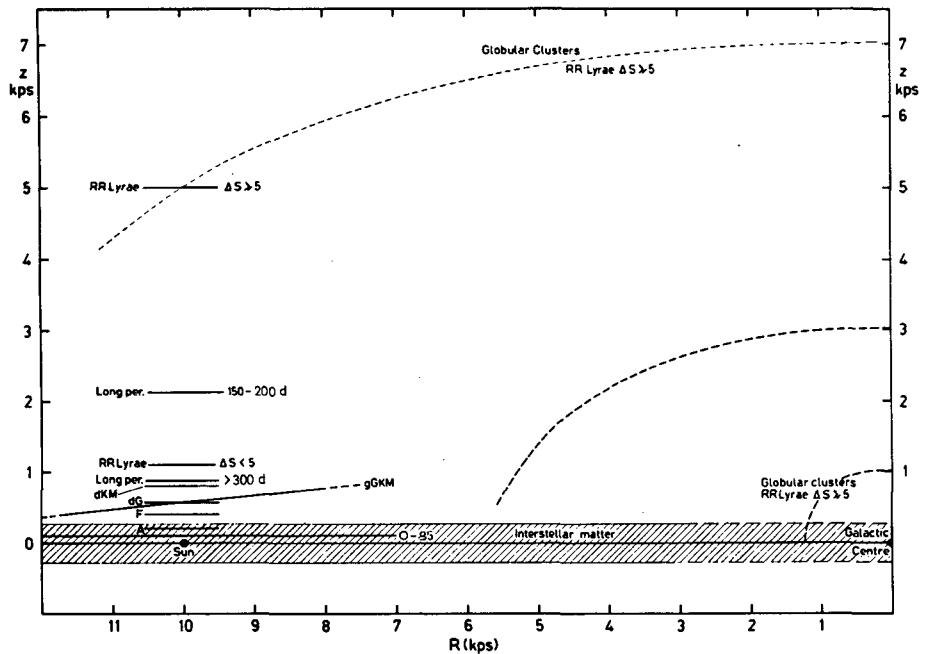


FIG. 2. Distribution of stars in the galaxy by metal line weakness (after Blaauw).

A picture gradually emerged in which the original Baade concept of two stellar populations was replaced with a continuum of population "states", in which metal weakness was correlated with position in the galaxy and with corresponding kinematical characteristics. The stellar content of the galaxy was imagined to consist of a superposition of stellar sub-systems of varying ellipticity. The spherical halo of globular clusters contained the stars representative of extreme metal weakness such as those of M92, and the local RR Lyraes with weak metal lines by Preston's criterion had high Z-components of velocity; with gradual metal enrichment, one encountered sub-systems less and less spherical such as the long-period variables with periods in the 150- to 200-day range, the RR Lyraes with stronger metallic lines, then the long-period variables with periods near 300 days, the ordinary giants, then the globular clusters of the galactic nucleus with moderately strong lines, and finally the disk population, with an older part consisting of somewhat weak-lined main sequence stars with "lagging" circular velocities, followed by main sequence stars with ordinary circular velocities, and finally the spiral arm population of O, B stars, interstellar matter and cepheid variables. These are illustrated schematically in Fig. 2, following Blaauw.

An analysis of this sort permits a reconstruction of the early history of the galaxy along lines originally put forth by Oort. One supposes that the protogalaxy was an inhomogeneity in the expanding universe that contracted towards higher density, conserving its angular momentum in rotation around some fixed axis. As the Z-motion of the gas gradually dissipated and formed a flattened disk, stars were formed in condensations of the gas. Nucleosynthesis of heavy elements is thought to have gone on as the stars evolved. The stars returned metals to the interstellar medium in the red giant stage and in supernova outbursts and the products of such synthesis gradually enriched the gas from which subsequent generations of stars were formed. This simplified picture, while probably correct in broad outline, must be modified to allow for the possibility, supported by the observational data of Arp, Pagel, Spinrad, and others, that there exists considerable spatial variation in metal abundance of the interstellar medium at a given time.

For faint stars such as those in globular clusters and in other distant regions of the galaxy, it is not practicable to make abundance analyses based on spectrograms. One uses photometric indices based on some kind of broad-band photometry. This is provided in the UBV system by the colours U-B, B-V, and a quantity  $\delta(U-B)$  at a fixed value of (B-V). Fixing attention on main sequence stars, we find that (B-V) is sufficient to determine the temperature of a star in a given stellar system with a known chemical composition, such as the stars of a galactic cluster. In particular, one can set up the relationship between (B-V) and  $T_e$  for the main sequence of the Hyades from model atmosphere considerations. In general, in a star of weak metal lines, such as that in a globular cluster, or in a star of the solar vicinity with a large Z-velocity, (U-B) will be too negative for the observed value of (B-V). This comes about because of the deblanketing of metal lines which are more heavily crowded into the region of the spectrum occupied by the U-filter; the B-filter is less affected and the V-filter still less. A quantity  $\delta(U-B)$  which measures this departure is a good index of metal deficiency, as has been shown by Wallerstein and Carlson from a comparison of metal abundance based on high dispersion spectrograms

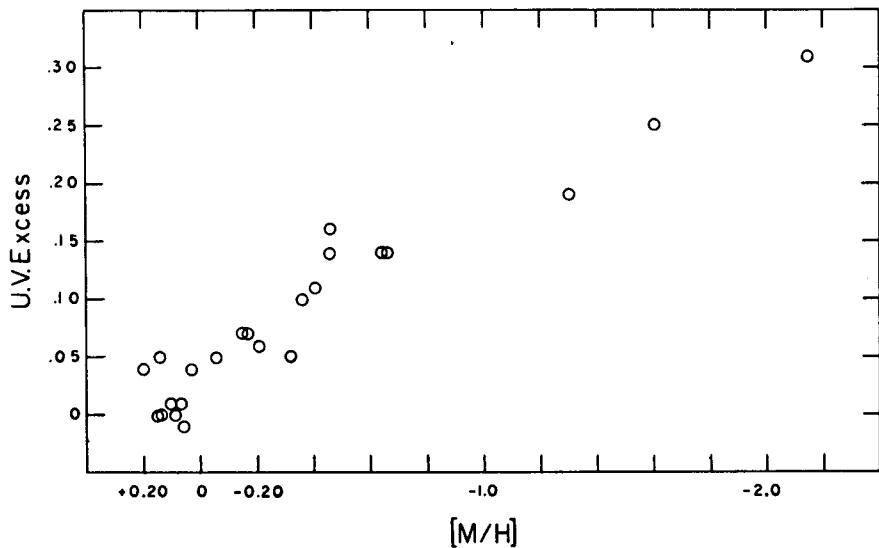


FIG. 3.  $\delta(U-B)$  as a function of  $[M/H]$  from high dispersion spectrograms (after Wallerstein and Carlson).

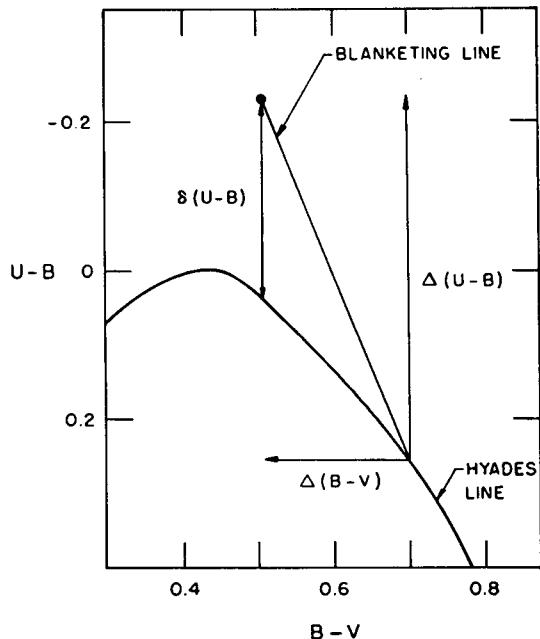


FIG. 4. Definition of  $\delta(U-B)$ . The lower curve corresponds to the Hyades main sequence (after Wildey, Sandage, and Burbidge).

and the quantity  $\delta(U-B)$  (Fig. 3). The latter is defined in Fig. 4 in a  $(U-B)$  versus  $(B-V)$  plot. The lower curve corresponds to the Hyades main sequence, and a typical globular cluster main sequence star would be found at the top of the vertical vector. The slant line gives the slope of the trajectory along which the sun would move if the lines were removed from its spectrum. The lower end of this line gives the corrected position of such stars if they had the metal abundances of Hyades stars; in this way their "proper" effective temperatures can be determined. The existence of such stars in the solar vicinity is revealed by Fig. 5; these objects are known as "sub-dwarfs", and most have kinematical properties that put them on highly elliptical orbits around the galactic nucleus, or that indicate large  $Z$ -values for the motion. They become sub-dwarfs, of course, because their observed  $(B-V)$ 's are too blue for their luminosities based on trig parallaxes, and they thus wind up lying below the Hyades main sequence in the observed HR diagram.

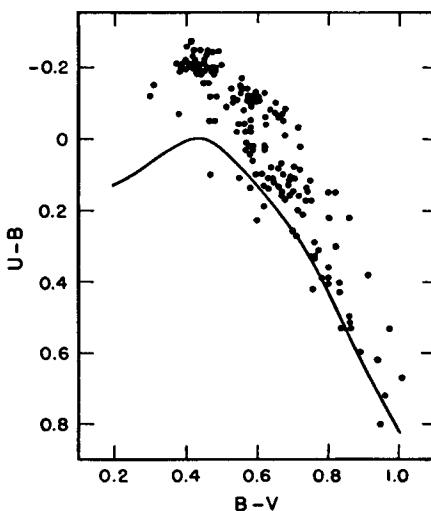


FIG. 5. Position of "sub-dwarfs" in the HR diagram (after Sandage).

Eggen and Sandage concluded that when these sub-dwarfs were corrected for blanketing as indicated by Fig. 4, they moved essentially onto the Hyades main sequence. More recently, however, the problem has been reconsidered by Cayrel who based his temperatures for Hyades and other Population I stars and for sub-dwarfs on the virtually line-free infra-red colours of Johnson and his associates and on the photometric system of Strömgren which measures hydrogen-line strengths. He concluded that the slope of the blanketing line of the UBV system depended on temperature more than Eggen and Sandage had supposed, and that for the sub-dwarfs of greatest weight the corrected positions in the HR diagram actually lay below the Hyades main sequence by about 0.7 mag. The significance of this result for the He abundance will be discussed later.

Using the  $\delta(U-B)$  index, Eggen, Lynden-Bell and Sandage showed that it was well correlated with velocity in the  $Z$ -direction, and therefore

with the maximum distance that could be achieved by a star responding to the force-field of the galaxy (Fig. 6). They then argued that the eccentric and highly inclined orbits of these stars are explained as a result of a large inward motion of the gas at the time they were formed, whereas the circular motion of the Population I stars in the vicinity of the sun reflects the original motion of the much younger gas out of which they were formed in comparatively recent times. Such inward radial velocities for the old stars imply a collapse of the gas in virtual free-fall in a time of only a few times  $10^8$  yr.

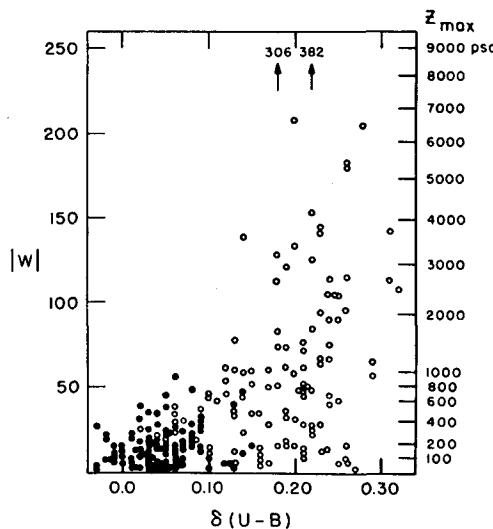


FIG. 6.  $\delta(U-B)$  as a function of  $Z$ -velocity (after Eggen et al.).

However, as noted by Field, the interpretation of Eggen and co-workers of their observational results might not be unique if one were willing to admit that the present high velocity stars in the vicinity of the sun were actually ejected with these speeds from a gaseous sub-stratum distributed in the disk, much as it is now. While it is true that the present random motions of interstellar gas clouds are too small, it is possible that these motions were larger in the past, and one must certainly keep in mind that even at present there exists in the galaxy the large outward motion associated with the 3 kpc arm.

We have discussed so far the dynamical evolution of the galaxy as it is correlated with metal abundance in stars and the character of their motions. It has been possible to do this more or less independently of stellar evolutionary considerations because stars of the main sequence could be compared for metal content directly from observation. It is true, of course, that the topology of colour-magnitude diagrams of globular and galactic clusters is sensitive to metal abundance. For example, it is well known that with increasing  $Z/X$  ( $Z$  here refers to the metal abundance not to the velocity component at right angles to the galactic plane!) the red giant branch rises less high above the horizontal branch in globular clusters, and this is confirmed by evolutionary calculations. Similarly,

the amount of the blueward extension of the horizontal branch and the number and mean period of RR Lyrae stars in globular clusters are both correlated with  $Z/X$ . But these topological considerations are not in themselves necessary to determine  $Z/X$ , and thus stellar evolutionary models need not be called upon. It is different with the important question of the He abundance, i.e. the ratio of  $Y/X$ . This is because the He lines cannot be excited in the low temperature stars of the main sequence.

The primordial He abundance is obviously a matter of great interest, and one would like to know if the He was all produced in the interiors of stars, or if it was synthesized in the early stages of the "big-bang". Most recently, Dicke has called attention to the importance of determining the He abundance in first generation stars as a test of the validity of general relativity versus the scalar-tensor theory of gravitation. He notes that, under general relativity, the universe would be about  $1.2 \times 10^{10}$  yr old, its present mass density would be about  $7 \times 10^{-31}$  g/cm<sup>3</sup>, a low value, and the initial He content of the first generation stars, presumably those of globular clusters of low  $Z/X$ , would be high, i.e. close to 30%. On the other hand, scalar-tensor theory predicts  $7 \times 10^9$  yr,  $2 \times 10^{-29}$  g/cm<sup>3</sup>, and 0% for these quantities, respectively. We may therefore ask if there is any direct observational determination of the He abundance in extreme Population II, i.e. in globular cluster stars of small  $Z/X$ .

The horizontal branch supplies the only types of objects capable of exciting the He lines. In M15, a globular cluster with extremely weak metal lines, there exists a planetary nebula in which lines of H and He are excited in emission. The abundance ratio He/H in such objects is not very sensitive to excitation conditions and depends mostly on the radiative recombination rates which are well known. An analysis by O'Dell, Peimbert, and Kinman leads to a near normal value of Y of about 40%. At the same time, calculations by Christy of the position in the HR diagram of the instability strip for RR Lyraes shows that the position of the high temperature edge is sensitive to the He abundance. Comparison with observation, while dependent on the interstellar reddening for a given globular cluster and on the adopted relationship between  $T_e$  and (B-V), seems to require again Y near 30%. On the other hand, Greenstein, Searle and Rodgers, and Sargent and Searle have all pointed out the considerable weakness of He lines in horizontal branch stars in comparison with main sequence stars of equivalent temperature, and have expressed doubt that this effect can be explained as a result of a change in gravity alone. They suggest a decline in the He/H ratio (by number) of one or two orders of magnitude.

Regardless of the precise interpretation of these effects, it must be admitted that, in every case, we are dealing with evolved stars. Even if the original He abundance in these objects were zero, they must have manufactured He in the deep interior, and at the time of the He flash, this could well have been mixed into the envelope. It is appropriate to remember here also that, even in extreme Population I, among the stars of the Orion association, there are stars with apparent underabundances of He. Is this primordial, resulting from some fluctuation of He/H in the interstellar medium, or have these stars been "spalled-clean" of  $^4\text{He}$  by protons accelerated in intense magnetic fields? And if the latter were true, could such processes operate in horizontal branch stars as well?

A different kind of approach to the He problem is being made by Faulkner and Iben who compute the evolutionary paths for horizontal-branch

stars and study the topology of the globular cluster HR diagrams as a function of Y. We need not discuss the details of these calculations, but we can note that high He abundance seems consistent with (1) the existence of a gap between the giant and horizontal branches, (2) an explanation for the horizontal branch that requires no mass loss between the giant and horizontal branch stages, and (3) the "correct" luminosity difference between the cluster turn-off point and the RR Lyraes.

Returning to the main sequence F and G type stars of globular clusters, we ask if a direct determination of Y among these might still be possible, thus permitting us to be independent of evolutionary considerations. In the UBV system, this cannot be done independent of assumptions we do not wish to make. We have

$$U - B = f_1(T_e, Z/X, \mathcal{M}/R^2, E)$$

$$B - V = f_2(T_e, Z/X, \mathcal{M}/R^2, E)$$

and from the Vogt-Russell theorem,  $L = L(\mathcal{M}, Z/X, Y/X)$ . The quantity  $\delta(U-B)$  can be calculated from these and determines a functional relation between  $\mathcal{M}/R^2$  and  $Z/X$ , but is not independent from them. There are simply not enough observed quantities to determine independently  $T_e, \mathcal{M}$ , E,  $Z/X$ , and  $Y/X$ . If we fit to the sub-dwarfs as indicated by Cayrel, for example, then we are forcing agreement with whatever Y versus  $\mathcal{M}$  relation is appropriate for these stars; this may not be the same for globular stars or even for globular clusters *inter alia*. On the other hand, Strömgren has defined a colour system in which the filters are carefully chosen to measure astrophysical quantities based on model atmosphere considerations. We can write

$$b - y = f(T_e, \mathcal{M}/R^2, Z/X, E) \text{ temperature and reddening}$$

$$c_1 = c_1(T_e, \mathcal{M}/R^2, Z/X) \text{ surface gravity}$$

$$m_1 = m_1(T_e, \mathcal{M}/R^2, Z/X) \text{ metal-line blanketing}$$

$$\beta = \beta(T_e) \text{ temperature}$$

$$\text{with } L = L(\mathcal{M}, Z/X, Y/X) \text{ Vogt-Russell theorem}$$

Since  $R^2 = L/T_e^4$ , each main sequence star of a globular cluster gives  $T_e, \mathcal{M}/L, Z/X$ , and E, with  $Z/X$  and E in principle the same for all. Combining this with the Vogt-Russell theorem enables the  $Y/X$  and  $\mathcal{M}$  to be determined independently. In practice, however, the procedure is beset with two difficulties, one observational and one theoretical. Observationally, the technique would depend on measuring  $\beta$  with a precision of 2 or 3% in a number of main sequence stars of apparent magnitude fainter than 18 or 19. This filter is one order of magnitude narrower than the conventional broad-band filters of UBV; several hours of photon counting with a large telescope would be needed to determine  $\beta$  with the required precision for one star, even with the 200-in. telescope. The theoretical difficulty has to do with the models of main sequence stars with external convection

layers; for an assumed mass and composition, the radii depend on the ratio of mixing length to scale height, a freely assignable parameter at the present state of knowledge. This could for example vary from one stellar sub-system to another as a function of e.g. rotation or magnetic field strength. It therefore appears that the only hope for immediate progress in the question of the He abundance for Population II is along the lines of evolutionary calculation in the manner of Iben and Faulkner, but it must be recognized that this may be far removed from the direct "observations".



# RADIO GALAXIES AND QUASARS

F. G. SMITH

Nuffield Radio Astronomy Laboratory,  
Macclesfield, Cheshire,  
United Kingdom

## Abstract

RADIO GALAXIES AND QUASARS. 1. The log N-log S relation; 2. Identifications; 3. The empty fields; 4. Special populations of radio sources; 5. Common properties of quasars and radio galaxies; 6. Angular sizes; 7. Cosmological puzzles.

It is in some ways surprising that radio astronomy has proved so apt a tool in cosmological studies. The most advanced techniques in radio have brought radio telescopes only a little way towards the capabilities of the simplest optical instruments. Only some thousands of individual radio sources have been individually observed, and most radio telescopes cannot detect more because of their large beam-width. Radio maps of the whole sky have an angular resolution only of the order of one degree. The spectrum of cosmic radio waves is almost featureless, in contrast to the wealth of physical information available in the optical spectrum. If the progress of astronomy had been from radio to optical, instead of vice versa, the advance must surely have been a revolution in which the old was completely discarded. Nevertheless, almost the only substantial evidence in cosmology comes from radio astronomy. The reason lies in the comparative rarity and the great power of the quasars and radio galaxies, which can be detected as individuals at distances beyond the reach of optical telescopes.

At the same time it must be emphasized that the radio studies depend heavily on the optical, particularly because the basic measurement of distance is not otherwise available. Distance is available optically from a chain of argument which builds up into the relation between red shift and magnitude for a recognizable class of objects, but which becomes more controversial when this class is extended to include the quasars, which are less common and less familiar objects. Distance is not available at all for those radio sources which do not have an optical counterpart; here the radio astronomer uses the correlated information of the identified objects to interpret the unidentified ones, and if he can relate the properties of identified and unidentified radio sources, he may deduce that he is exploring to greater depths of the universe than can the optical astronomer.

The difficult arguments centre on this relation between identified and unidentified radio sources. If the sample of radio sources which are well understood from many studies of their radio and optical properties is indeed representative, then we can use without inhibition the relation between number and flux density, the log N - log S relation, to test cosmological theory.

## 1. THE LOG N - LOG S RELATION

At various stages of the extension of the number counts to smaller flux densities the counts were made both with pencil beam telescopes and with interferometers. An advantage of the interferometer has been that it allowed an extension of the analysis to a number of sources approaching one per primary telescope beamwidth, by using Scheuer's statistical analysis. Hewish [1] was able to delineate from this analysis the main features of the number counts, namely the rise in numbers with a slope of -1.8, and a subsequent fall beyond about  $10^3$  sources per steradian. It was, however, necessary to examine carefully the possible effects of angular resolution on the analysis, especially as this might reduce the flux densities of the nearer and therefore the most intense sources.

The counts of individual sources require a coverage of the whole sky at large flux densities where numbers are small, and coverage of successively smaller areas for smaller flux densities where the numbers would otherwise become impossibly large. The whole sky is covered to a flux density  $S_{400} > 10$ , while the counts from Pooley and Ryle [2] extending to  $S_{400} = 10^{-2}$  cover only  $3 \times 10^{-4}$  sr. The small areas must be shown to be representative by comparing the results for different areas of the sky. Here the interferometer has again been used to provide a comparison in statistics (Hughes and Longair [3]), but so far this has only been achieved for  $S_{178} > 0.2$ .

The results of assembling all these counts, covering almost four decades of flux density and extending to  $10^5$  sources per steradian, are given by Pooley and Ryle both as a log N - log S curve and as a relative count  $N/N_0$  versus S, where  $N_0$  is the number of sources expected in a static Euclidean universe. The slope of the log N - log S curve is initially -1.85, extending to  $N = 10^2$  /sr.  $N/N_0$  correspondingly increases, up to a maximum of  $2\frac{1}{2}$  where  $N = 10^3$  /sr. Then comes a fall of  $N/N_0$  to about 0.2 where  $N = 10^5$  /sr; at this point the slope of the log N - log S curve is -0.8.

The static Euclidean universe could only be accepted if the whole phenomenon were very local indeed. Another more suitable model, the Einstein-de Sitter model, may be fitted on to the more intense sources and used to predict a form of the  $N/N_0$  curve. The excess of observed faint sources over the number expected on this model reaches a factor of almost ten.

So far, the only explanation of this curve has been in terms of evolutionary models. Longair [4] has shown how to combine the source counts with the observed value of integrated sky brightness so as to test any hypothetical distribution of sources, specified as a luminosity function  $\rho(P)$  which may vary with both distance and time. He shows that either a time variation of mean luminosity  $P \propto t^{-2.2}$  or a density evolution  $\rho \propto t^{-3.8}$  would explain the increase in numbers, and that a cut-off is required in either case for a value of red shift  $Z \sim 3$  or 4. Only the more powerful sources can evolve in this way, otherwise the reduction of  $N/N_0$  at low flux densities will not be sharp enough to fit the observations.

A further most important result is that the individual sources now seen by the one-mile telescope account for as much as one half of the total extragalactic radio radiation. Not only is the population of radio sources isotropic, it is also unique. If it is unique, it must have the scale of the universe as a whole, since we cannot reasonably consider any other unique and isotropic population centred on the solar system.

## 2. IDENTIFICATIONS

Those catalogues of radio sources which are complete down to a definite limiting flux density, as for example the Parkes catalogue for the southern hemisphere, and 3C, followed by 4C, for the northern hemisphere, provide material for a thorough investigation of the nature of the observed sources, and in particular their luminosity distribution  $n(P)$ . This is not, of course, the same as investigating the nature of a sample of radio sources representing all that exist in a given volume of space; the luminosities of a sample of this kind are distributed according to a luminosity function  $\rho(P)$ .

The results of these identifications are of obvious interest in the nature of the visible sources, but they are also vitally important in establishing the function  $\rho(P)$  which is needed in interpreting the source counts. It will be appreciated that the slope of a straight line relation between  $\log N$  and  $\log S$  is not affected by the form  $\rho(P)$ , but that any features in the  $\log N - \log S$  curve tend to be smoothed out by a widely distributed  $\rho(P)$ . The sharpness of the cut-off at large values of  $Z$  can only be discussed when  $\rho(P)$  is known; of course,  $\rho(P)$  may itself be a variable, and the identifications only lead to a local determination.

At present, the most interesting part of the identification story lies in the failure to identify a considerable proportion of sources even when accurate positions are available and the sky in these positions is not obscured by dust clouds.

## 3. THE EMPTY FIELDS

In 1966, largely through the efforts of Wyndham and of Véron, a systematic search for identifications of the 3C source catalogue was completed. Of the 328 sources Véron [5] selected 254 by omitting 32 which were clearly galactic, such as the supernova remnants, and 42 which were in absorbing regions of the sky. The 254 sources in clear fields were identified as

100 definite 44 possible	}	radio galaxies
39 definite 21 possible	}	quasars
36 empty fields, $m_v > 21$		
14 not identified (inaccurate positions)		

The "empty fields" possibly represented quasars at great distances, since the slopes of the  $\log N - \log S$  curves were

Radio galaxies	$-1.55 \pm 0.05$
Quasars	$-2.2 \pm 0.1$
All except radio galaxies	$-2.2$

Bolton [6] performed a similar analysis on a uniform sample of 383 sources in the southern hemisphere. He found 143 radio galaxies and 100 quasars - rather a smaller proportion of radio galaxies than Véron

found for the 3C sources. Bolton suggested that his "empty fields" represented radio galaxies rather than quasars, on the basis of a difference between the spectra at frequencies between 1410 and 2650 MHz. There seemed to be no clear distinction between quasars and galaxies on the basis of the log N - log S plots. The radio galaxies had a somewhat smaller slope, but it was suggested that this could be accounted for by the effects of plate limits on the optical identifications.

There was not a strong conflict here, since Véron's counts for the empty fields were not very accurate, allowing at least some to be radio galaxies, and Bolton had apparently identified a smaller proportion of radio galaxies so that any quasars in the empty fields might be lost among the radio galaxies on the limit of identification.

Little [7] has attempted to resolve this question by a careful study of interplanetary scintillation, which sorts out the sources with significant components below about 1" arc in angular size. He examined those sources in 3C which are close enough to the ecliptic to show scintillation. Of 22 quasars (or possible quasars) 13 scintillated; of 56 radio galaxies 5 scintillated; of 19 "empty fields" 10 scintillated.

This seems to show that the "empty fields" are in fact quasars. However, Little remarks that it is better to say that the sources which scintillate are also those which are the more powerful, whether or not they are quasars. The radio galaxies have luminosity  $P_{178}$  between  $10^{23}$  and  $10^{26} \text{ W sr}^{-1} \text{ Hz}^{-1}$ , but the only scintillators have  $P_{178}$  close to  $10^{26} \text{ W sr}^{-1} \text{ Hz}^{-1}$ . The quasars all have  $P_{178}$  above  $10^{26} \text{ W sr}^{-1} \text{ Hz}^{-1}$ , and most of them scintillate.

The conclusion so far is that the empty fields contain either quasars or the most powerful radio galaxies.

#### 4. SPECIAL POPULATIONS OF RADIO SOURCES

The evidence has been presented so far as though any given sample of space contained a selection of radio sources which is closely related to the selection found anywhere else. Only a progressive evolution is envisaged, over distance scales of the order of the Hubble distance. Is it possible to accommodate a substantial proportion of sources in an entirely different configuration, say in the galaxy, or in the local group of galaxies?

A local population of sources might be postulated to explain the excess of sources, or to place quasars at distances smaller than their red shifts indicate. The excess of sources cannot be explained in this way, since it is such a large effect that 80% of the sources must take part in it at a flux density  $S_{408} > 0.1$ . This is not compatible with the observation that the population of sources is isotropic and unique. Some quasars could, however, be accommodated, since the total numbers are fairly small and the source counts would not be much changed by removing the identified quasars.

The argument about the distance of the quasars relies on quite other grounds, namely the close relation between their properties and those of radio galaxies. The small numbers do not allow much statistical analysis to be made, but it should nevertheless be noted that the slope of the log N - log S curve for identified quasars is not far from that of all sources taken together.

## 5. COMMON PROPERTIES OF QUASARS AND RADIO GALAXIES

The distinction between quasars and radio galaxies seems clearer for their optical than for their radio properties. The radio properties for which sufficient data are available are the luminosity distribution, the spectra, and the angular diameters. Some data are becoming available on variability of luminosity.

The luminosity distribution depends, of course, on the interpretation of red shift as distance, and for large red shifts it depends on the cosmological model which is used in this interpretation. The distribution function has no break between quasars and radio galaxies, which overlap in the region of  $P_{178} = 10^{26} \text{ W sr}^{-1} \text{ Hz}^{-1}$ .

The spectra are practically indistinguishable at the lower frequencies. A complete sample made at 178 MHz, i.e. the 4C catalogue, has been extensively studied by combining observations made at Cambridge, Jodrell Bank and Parkes. Up to 1410 MHz the mean spectral index of quasars is  $0.72 \pm 0.04$ , and of radio galaxies  $0.71 \pm 0.03$  (Williams et al. [8]). From 1410 to 2695 MHz the spectrum shows a steepening, and there is some doubt about whether this is less marked for quasars than for radio galaxies. The Parkes survey tends to show that it is, while the Jodrell Bank observations combined with Cambridge suggest that it is not. There is certainly no large effect in the known spectra, although individual quasars and radio galaxies both have complicated spectra at the higher frequencies.

Any difference in curvature of spectrum between the two classes, quasar and radio galaxy, may be better interpreted as a difference between spectra according to brightness temperature. This depends of course on a measurement of angular size, either by scintillation studies or by interferometry. We shall return to this later.

Finally, any attempt to distinguish the radio properties of quasars and radio galaxies must take account of radio galaxies which are prominent visible galaxies but with a quasar-like radio object contained in them. These quasar-like radio galaxies are amongst the known variable radio sources and provide very strong evidence of the close connection, amounting probably to a continuum of properties between classes of objects only distinguishable when they show extremes of intrinsic luminosity and size.

## 6. ANGULAR SIZES

There are many maps of the brightness distributions across sources whose angular sizes are greater than about  $\frac{1}{2}'$ arc. These are usually, but not always, identified with radio galaxies. Some, such as 3C 47, are identified with quasars, even though the radio source is double separated by  $1'$  arc. Some are hard to identify because they cover a large area of sky.

Greater angular resolution can be achieved, at the expense of actual map-making, by interferometry with various lengths of baseline. At Jodrell Bank we have reduced the baseline for our major series of observations to 20 km, so that the angular resolution for the longer wavelengths is of the order of  $5''$  arc. This followed the well-known series of experiments in which fewer and fewer sources were found to have appreciable

components with angular sizes eventually reaching down to 0.04" arc; at this point trans-Canadian and trans-U.S.A. interferometry stepped in and reduced the limits on a few sources to 0.003" arc.

These very long baselines have so far only been used on a small number of intense sources, and only at short wavelengths. We need a systematic study, and we need it at a range of wavelengths. Some of the best work is still that from interplanetary scintillation, which can be used to distinguish sources greater or less than 1/3" angular size. Recent interferometer observations at Jodrell Bank have shown that the majority of radio sources observed at 408 MHz are easily visible with an interferometer spacing of 32 000 wavelengths. Furthermore, the small angular structure which this implies is found in radio galaxies as well as in quasars.

When quasars are examined with sufficient angular resolution, many of them are found to be double radio sources, like many radio galaxies. A new example is 3C 147, which has recently been resolved by Donaldson (private communication) into a "double double" structure, with components separated by 0''.14 arc and 1''.1 arc. The essential difference between the quasars and the radio galaxies lies in the angular diameter and hence the brightness of the individual components. It is still reasonable to suppose that there is a continual gradation of angular diameters of components, from some minutes of arc for a "typical" radio galaxy to some hundredths of a second of arc for a "typical" quasar. When making the comparison it must be realized that the spectra of the small components of a source tend to be flatter than those of the large components, so that short wavelength measurements emphasize the small concentrations. It would be interesting to know the spectra of various angular sizes of components, but these are not yet available.

The angular structure of the sources in the empty fields is not obviously different from that of the quasars.

## 7. COSMOLOGICAL PUZZLES

The difficulties in the optical properties of quasars include the interpretation of the absorption lines, the absence of Lyman- $\alpha$  absorption in the continuum, and irregularities in the distribution of numbers of quasars both against red shift and across the sky. There are no additional serious problems of this nature in the radio observations, possibly because there are no observable radio lines and because the optical quasars are nearly all found from the radio positions. At this observational level the most interesting problem is that of the empty fields, which contain radio sources very like the identified sources, with similar spectra and angular diameters, but without a visible counterpart. If these are indeed quasars, their optical luminosity must be rather low.

The interpretation of the radio counts according to cosmological theory, once having accepted the demise of the steady state theory, is a large subject of its own. It would be useful to add some relations between red shift, flux density, spectrum, and angular diameter, but it has been distressing to find no clear relations between these quantities. Presumably there is a large dispersion in the intrinsic properties of the radio sources.

The only recent light on this subject has come from Miley [9], who found a definite relation between red shift and angular diameter for the

quasars with steep spectra. This observation has not yet been fitted into cosmological theory; so far it only helps to establish that the ordering of sources in terms of red shift does have a physical significance.

The puzzles about the origin of the energy in the quasars and radio galaxies are beyond the scope of this lecture.

#### R E F E R E N C E S

- [1] HEWISH, A., Mon. Not. R. astr. Soc. 123 (1961) 167.
- [2] POOLEY, C.G., RYLE, M., Mon. Not. R. astr. Soc. 139 (1968) 515.
- [3] HUGHES, R.G., LONGAIR, M.S., Mon. Not. R. astr. Soc. 135 (1967) 131.
- [4] LONGAIR, M.S., Mon. Not. R. astr. Soc. 133 (1966) 421.
- [5] VERON, P., Astrophys. J. 144 (1966) 861.
- [6] BOLTON, J.G., Nature 211 (1966) 917.
- [7] LITTLE, L.T., Observatory 88 (1968) 52.
- [8] WILLIAMS, P.J.S., COLLINS, R.A., CASWELL, J.L., HOLDEN, D.J., Mon. Not. R. astr. Soc. 139 (1968) 289.
- [9] MILEY, G.K., Nature 218 (1968) 933.



# OBSERVED PROPERTIES OF QUASI-STELLAR OBJECTS

M. SCHMIDT

Mount Wilson and Palomar Observatories,  
Pasadena, Calif., United States of America

## Abstract

OBSERVED PROPERTIES OF QUASI-STELLAR OBJECTS. 1. Radio properties; 2. Optical continuum; search for QSO; 3. Distribution; 4. Isolated, star-like objects; 5. Optical variations; 6. Emission line spectra; 7. Red shifts; 8. Absorption line spectra; 9. Red shift - magnitude relation.

Quasi-stellar objects (QSO) may be defined as star-like objects with emission-line spectra showing large red shifts. We shall limit this review to a summary of their observed properties.

### 1. Radio properties

Quasi-stellar radio sources (QSS) are quasi-stellar objects observed as radio sources. Among the 300 extragalactic sources in the 3C revised catalogue, 44 have been identified as QSS. In total, several hundred radio sources in other radio catalogues are identified as QSS. Many QSS have radio components of angular size less than  $1''$ , but rather few radio galaxies contain so small a component. Intercontinental interferometry shows that the brightest QSS, 3C 273, has a radio component of less than  $0''.001$ . However, some QSS have radio components as large as a minute of arc. Many QSS have a relatively flat radio spectrum, in which the radio flux at frequencies above 1000 MHz is usually variable with a typical time-scale of a year. This has also been observed for a few radio galaxies. It appears that there is considerable overlap of radio properties between the radio galaxies and the QSS, the differences being of a statistical character.

### 2. Optical continuum; search for QSO

The relation between optical and radio continuum is not simple. In some cases, like 3C 47, there can be none, as the radio source is double and well away from the star-like object. The optical energy distribution of QSS is much flatter than that of typical faint galactic stars. Hence the optical QSS show a relative excess in the ultra-violet and in the infra-red. Sandage and Luyten have used the ultra-violet excess to find QSO without reference to radio. Braccesi uses both the ultra-violet and the infra-red excess to isolate relatively pure samples of QSO among faint stars. Thus all QSO found till now have an ultra-violet excess by their selection. If QSO without ultra-violet excess exist, they would be difficult to find. It may be doubted that they exist, since all QSS identified from radio positions do show an ultra-violet excess. Most of the energy in the quasi-stellar sources is radiated in the visual region of the spectrum; in a few

cases, like 3C 273, most is radiated in the infra-red. The optical properties of QSS and QSO have not revealed any significant differences.

Sandage and Luyten find 0.3 QSO per square degree down to visual magnitude 18. The corresponding surface density of QSS in the 3C revised catalogue is only 0.0014 per square degree.

### 3. Distribution

The distribution over the sky of 3C QSS does not deviate from a random distribution. Ryle has emphasized that counts of radio sources at all levels of flux density show isotropy so that the QSS, which probably make out some 30% of the 3C revised catalogue sources, must show a considerable degree of isotropy, too. Surveys of QSO are too few to allow the study of their distribution over the sky.

### 4. Isolated, star-like objects

All objects identified are star-like, i.e. their optical diameters are less than 1''. Some of the quasi-stellar sources, such as 3C 273, 3C 48 and 3C 196, show faint nebulosity near the object. Except for this, all the objects seem to be isolated and not associated with galaxies or clusters of galaxies. It is this isolation which is to a large degree responsible for the continuing uncertainty about the nature of the quasi-stellar objects.

### 5. Optical variations

Variations are seen in the optical continuum of some QSO with characteristic time-scales of a few days up to ten years or more. No general secular variation is seen. A few sources show large variations: 3C 446 has varied over a factor of 20 in the past few years. Kinman has found that its polarization properties vary over an interval of months.

### 6. Emission line spectra

The strongest emission lines seen are Ly- $\alpha$   $\lambda$  1216, C IV  $\lambda$  1549, C III  $\lambda$  1909, and Mg II  $\lambda$  2798; weaker lines are due to Si IV, He II, Fe II, etc. Above 3000 Å the spectra show the Balmer lines and forbidden lines of O II, O III, etc., as usually observed in the spectra of gaseous nebulae around hot stars. The emission lines are usually about 40 or 50 Å wide. In a few cases the forbidden lines are much narrower, in exceptional cases the Balmer lines too.

### 7. Red shifts

The emission lines show red shifts  $z = \Delta\lambda/\lambda_0$  ranging from 6 to 236% (the latter for the quasi-stellar source 4C 25.5). The red shift does not vary with time; in particular, the red shift of 3C 446 did not change when the optical output increased by a factor of 20.

### 8. Absorption line spectra

Some QSO exhibit both emission and absorption lines in their spectra. The main absorption lines in order of strength are Ly- $\alpha$ , C IV, Si IV, Si III, Si II, C II, Mg II and others. Most often the absorption line red shift is slightly less than the emission line red shift  $z_{\text{abs}} \lesssim z_{\text{em}}$ . In a very few cases  $z_{\text{abs}} \gtrsim z_{\text{em}}$ , e.g. 4C 25.5 with  $z_{\text{em}} = 2.358$ ,  $z_{\text{abs}} = 2.368$ . Several objects exhibit one or more absorption red shifts that are considerably smaller than  $z_{\text{em}}$ . Dramatic cases are:

- (a) PHL 5200 ( $z_{\text{em}} = 1.98$ ) which according to Lynds shows absorption bands with  $z_{\text{abs}}$  varying from 1.9 to 1.98. Margaret Burbidge claims that the structure in the absorption lines changed recently.
- (b) PKS 0237-23 ( $z_{\text{em}} = 2.22$ ) which exhibits a multitude of narrow absorption lines. Burbidge, Lynds and Stockton find three to eight absorption red shifts, while Bahcall, Greenstein and Sargent find five absorption red shifts, all ranging from  $z_{\text{abs}} = 1.36$  to 2.20. The widths of the absorption lines correspond to only 100 km/s.
- (c) PHL 938 ( $z_{\text{em}} = 1.95$ ) in which Burbidge, Lynds and Stockton identified two absorption red shifts,  $z_{\text{abs}} = 1.91$  and 0.61.

### 9. Red shift - magnitude relation

It has sometimes been claimed that the observed red shifts show no correlation with the magnitude. The only complete material available is for the sources of the 3C revised catalogue. We find for decreasing optical brightness:

<u>Visual magnitude</u>	<u><math>\langle \log z \rangle</math></u>
16	-0.28
17	-0.09
18	-0.02

Hence the mean red shift does increase as we go to optically fainter QSS. The observed rate of increase of  $\log z$  with magnitude is affected by radio selection effects, the discussion of which is outside the scope of this review.



# INTRODUCTION

## Opening talk on pulsars

J. G. BOLTON

Commonwealth Scientific and Industrial Research Organization,  
Division of Radiophysics,  
Chippendale, N.S.W., Australia

### Abstract

INTRODUCTION: OPENING TALK ON PULSARS. The discovery of the four known pulsars up to June 1968 is described, the characteristics are given, and the components of the pulse shapes are discussed.

From a theoretical point of view the discovery of the pulsars is particularly interesting as they may offer us the opportunity to study highly condensed matter on a small scale in just the same way as the quasars might on a large scale.

In a way history was repeating itself in their initial discovery. In 1946 Hey deduced the existence of the first radio source from fluctuations in the radio emission from the galaxy in a small region of sky - by analogy with the known variations in solar radio emission. His surmise of discrete sources proved to be correct, but the fluctuations were shown to be scintillations in the ionosphere due to irregularities in electron density. These scintillations have a period of  $\sim 1$  min and are rarely seen above 100 MHz. Some years later a shorter period scintillation was discovered at Cambridge during radio source position observations where a short output time constant was employed. These scintillations have a period of 0.1 to 1 s, are seen only for sources  $\sim 1''$  arc or less and are due to the interplanetary medium. Close to the sun they occur at frequencies of  $\sim 1000$  MHz and can occur as far as  $70^\circ$  from the earth-sun line at much lower frequencies.

The characteristics of the pulsars are given in Table I. They were first discovered at Cambridge with an aerial system designed to monitor the interplanetary scintillations of many thousands of radio sources. Each source is only seen for a few minutes each day at transit. A very short output time constant was used in order to faithfully follow the scintillations. After some weeks of observation the Cambridge team, led by Hewish, realized that certain sporadic pulse type interference occurred at the same sidereal time - hence the pulsars.

Their radiation occurs in the form of short pulses of the order of 30 ms duration and extends over most of the observable radio spectrum. The pulses repeat at intervals of the order of a second and the pulse interval remains constant to one part in  $10^8$  over six months.

All four known pulsars were discovered at Cambridge and the Cambridge observers have provided positions for the objects, accurate to  $10''$  arc in two cases and to  $1'$  arc in the other two cases. Following the announcement of their existence, work immediately began on them with the large steerable reflectors at Jodrell Bank, Arecibo and Parkes.

TABLE I. CHARACTERISTICS OF THE FOUR KNOWN PULSARS  
(TO JUNE 1968)

Pulsar	CP 0834	CP 0950	CP 1133	CP 1919
Approximate period (s)	1.27	0.25	1.19	1.38
Mean flux density at 80 MHz	0.3	0.8	0.3	0.4
Galactic latitude	26°	44°	70°	4°
Estimated distance (pc) (for electron density = 0.2 cm <sup>-3</sup> )	60	15	30	60
Pulse length (ms)	~35	~20	~35	~35
Spectral index	-2 ± 0.5	-1 ± 0.5	-1.5 ± 0.5	-1.5 ± 0.5

These instruments have the advantage of being able to observe at many frequencies simultaneously and at different polarizations. The fact that they can track the pulsars for some hours means that many successive pulses can be integrated to give an average pulse envelope. One of the difficulties of investigating these objects is that there are severe variations in pulse-to-pulse intensities and in the longer term averages. However, the pulse envelope as an average of ~100 pulses is remarkably constant and permits the measurement of the onset to ~3 ms and thus the average period over a day to three parts in 10<sup>8</sup>.

This period changes during the year due to the orbital motion of the earth round the sun and a smaller change takes place during the day due to the earth's rotation. The former effect was shown to exist in the initial Cambridge observations. The magnitude of this effect depends on the celestial position of the source and, conversely, the position of the source can be determined from the orbital variation in the pulse period. The heliocentric period of one pulsar (CP 1919) has been shown to vary by not more than a few parts in 10<sup>8</sup> over six months and this sets an upper limit to the change in relative motion of the pulsar and the sun. From it we can say that, if the pulsar is one member of a binary system, then its period must be ≫ 1 yr or the same as the pulse period, i.e. 1.38 s.

The pulsed nature of the radiation permits an approximate distance to be determined, for the pulses are delayed in the interstellar medium by an amount dependent on the frequency. The frequency dispersion is given by

$$\left[ \frac{d\nu}{\nu^3} \right] \rightarrow \frac{dO}{\nu^3} = \frac{-c}{L} \frac{dt}{\nu_p^2} \quad (1)$$

where  $\nu \gg \nu_p$  is the plasma frequency. It was first determined by the Cambridge observers from measurements over a very small frequency range. The measurements now extend over the range from 40 to 5000 MHz and the relationship given by Eq.(1) has been shown to be accurate to one part in  $10^4$ . This ensures that  $\nu \gg \nu_p$  for most of the intervening medium and means that only a very small fraction of the integrated electron density in the line of sight could be in a stellar corona surrounding the radiation source.  $\int n_e dL$  for the pulsar CP 1919 =  $12.55 \text{ ne cm}^{-3}$ . For  $n_e = 0.2 \text{ cm}^{-3}$  the corresponding distance is 60 pc. The closest object, CP 0950, must be of the order of 15 pc from the sun.

The average pulse spectrum shows considerable variation from day to day. Average values over a five-minute period may vary by a factor of ten at the low frequencies and a factor of three at the high frequencies. Spectral indices are within the range  $-1.5 \pm 0.5$  for CP 1133 and CP 1919; for CP 0955 the index is flatter at  $-1.0 \pm 0.5$  and steeper for CP 0834.

Individual pulses exhibit up to 100% linear polarization. For CP 0950 Graham Smith has determined the interstellar Faraday rotation. Combining this with the integrated electron density, he obtains a value of  $\leq 0.2 \mu\text{G}$  for the longitudinal component of the galactic magnetic field in the direction of this object. However, it is known from observations of the Faraday rotation of nearby extragalactic objects that the galactic field must be almost transverse to this direction.

Substructure is seen in both the individual and average pulse shapes. Three pulsars have pulses which show a moderately distinct double structure with both components of approximately equal amplitude. On occasions a third component may appear. For CP 0950 there is a single main pulse, but a minor sub-pulse occurs 100 ms before the main pulse. The sub-structure changes markedly from pulse to pulse. On occasions this sub-structure persists over a quite wide range of frequencies, but on others the correlation may vanish within 100 kHz. While some of the observed phenomena may be due to ionospheric or interplanetary scintillation, most are believed to be inherent in the source. It is difficult to account for a wide range of intensity variations at both high and low frequencies and the very strong focussing involved by a scintillation mechanism.

The question arises as to whether any of the more normal radio sources could be pulsars, as about 25% of the first thousand sources remain unidentified. Objects which are intrinsically more luminous than the four known pulsars and are more distant, so that the frequency dispersion is greater, or have a much shorter period, would appear as continuous sources. They might also have steeper than normal spectra. There are two unidentified radio sources, 3c 318.1 and one southern source, which might be of this nature from the form of their radio spectra.

Radial pulsations or rotation phenomena in white dwarfs or neutron stars have been proposed as the origin of the radio pulses. Attempts at optical identification have not been definitive so far. In the field of CP 1919 the brightest object is a yellow star of 18th magnitude. This is at a low latitude where the obscuration is very high. For a distance of 60 pc its absolute magnitude must be about 12. For CP 0950 the brightest object is 19th magnitude and the likely absolute magnitude about 18. This is outside the normal range of white dwarfs whose magnitudes range from 10 to 15 and one might wonder why none of the known white dwarfs, which

are on the average closer to the sun, exhibit the pulsing phenomenon - if white dwarfs are the responsible objects. If neutron stars are involved, it is unlikely that they would be visible at the estimated distances of the pulsars.

## ATTEMPTS AT OPTICAL DETECTION OF PULSARS

A. G. W. CAMERON

Belfer Graduate School of Science,  
Yeshiva University,  
New York, N. Y.,  
United States of America

### Short contribution

Some classes of pulsar hypotheses require that the radio pulses be produced in the atmosphere of a compact star, presumably a white dwarf. In view of the energy involved, one might expect that a progressive wave propagates upwards through such an atmosphere, becoming a shock either in the photosphere or in the lower corona, and then some of its energy goes into the radio radiation mechanism, whatever that is. Hence it is natural to look for light flashes or sinusoidal variations arising from the pulsars.

The optical searches carried out so far have concentrated on the strongest pulsar, CP 1919. Some eleven groups have reported on their attempts. Most of these groups used a photomultiplier with either a magnetic tape recorder or a multichannel scaler. The majority of the groups have attempted to see if there is optical variation synchronized with the pulsar period of 1.337 s. All such attempts have produced negative results.

The error box round the radio position of CP 1919 contains a "blue star". This is so designated because it is bluer than a nearby red star on the Palomar sky survey plates. Actually, the "blue" star is quite red. A recent 200" plate taken by H. C. Arp shows two fainter red stars, outside the error box but close to it, and a very faint (23rd magnitude) red star near the centre of the error box. The "blue star" was suggested as an identification of CP 1919 by Ryle and Bailey [1].

A positive detection of optical signals from the region of the "blue" star has been claimed by Lynds, Maran and Trumbo [2]. They used a 50" telescope, an unfiltered 1P21 photomultiplier tube and a 400-channel multi-scaler. They found a sinusoidal variation with double the radio period, a very surprising result, which if correct would account for the failure to observe any variation with the radio period itself. In recent work at Kitt Peak they have found this variation with a circular diaphragm as small as 3 arcseconds radius centred on the "blue" star. The amplitude of the sinusoidal variation was some 4% of the light of the "blue" star in their early work, but in their later work they found that the amplitude diminished close to zero during many of the hours of observation. They have not obtained this signal with the telescope pointed away from the "blue" star. However, in view of the lack of signal at times when the telescope is pointed at the "blue" star, the amount of time spent looking away from it may well be inadequate to check the reality of the signal. Hence confirmation with other types of equipment, preferably without a simulated pulsar radio period fed into the apparatus for timing purposes, is certainly required.

C. R. Lynds has obtained a good spectrum of the "blue" star. It shows H and K calcium absorption lines and the Balmer series of hydrogen lines. These lines have widths of about 15 Å. Lynds has classified the spectrum as A7 to F0, which indicates that the star is greatly reddened, presumably by interstellar absorption.

If the "blue" star should be a white dwarf correctly identified with CP 1919, then it is unlikely that any rotating model for the pulsar emission can be correct. Such a model, rotating with a period of 2.7 s or shorter, would have much broader lines unless viewed nearly pole-on. In the latter case, it is difficult to imagine a suitable radio "searchlight" narrow enough to operate in the polar regions and still to be detected for such a short portion of the pulsar period.

If the double radio period of the optical variation is correct, then it is difficult to associate it with a fundamental or low overtone pulsation of a star. So much energy would be associated with such a vibration that it is unlikely that the amplitude of the vibration could be variable in intervals of a few hours.

#### A C K N O W L E D G E M E N T S

I am indebted to S. P. Maran for information on the Kitt Peak measurements. This work is supported in part by the National Science Foundation and by the National Aeronautics and Space Administration.

#### R E F E R E N C E S

- [1] RYLE, M., BAILEY, J.A., Nature, Lond. 217 (1968) 907.
- [2] LYNDS, C.R., MARAN, S.P., TRUMBO, D.E., Science 161 (1968) 42.

# THE NATURE OF PULSARS

## Survey of present views\*

T. GOLD

Cornell-Sydney University Astronomy Center,  
Cornell University,  
Ithaca, N.Y.,  
United States of America

### Abstract

THE NATURE OF PULSARS: SURVEY OF PRESENT VIEWS. The constraints upon any theory which attempts to explain pulsars are discussed. Pulsar models must invoke emitting regions that are remarkably small and intense, and only a coherent motion of charges can reasonably provide such radiation densities. Secondly, high density matter has to be invoked in order that such a concentration of energy can be held together for long periods of time and in order to provide dynamical time constants that are short enough and that allow one to account for the extreme regularity. Current models are discussed based on pulsating white dwarfs, pulsating neutron stars, rotating white dwarfs, neutron stars with orbiting planet, neutron star binaries, or neutron stars with rotating magnetospheres. The last seems best capable of giving some of the observed features although there, as in the other cases, the radiation mechanism is not understood.

There is no theory yet that satisfactorily describes the observations, nor does anyone make a claim to understand the remarkable phenomena that have been observed. Nevertheless, one may discuss the basic types of points of view that have been developed.

The outstanding feature that has to be explained is the extreme regularity of the timing of the pulses. The constancy of the period is now known to an accuracy of better than one part in  $10^8$ . This is a precision that seems to belong to the realm of celestial mechanics or the motion of large masses rather than to that of low density plasma physics. For this reason alone the presence of dense matter has been surmised, so as to have mechanical periods as short as those observed.

Another reason for suggesting dense matter to be associated with pulsars is the need to invoke very small but intense regions for the emission, and hence the need to confine the energy required for at least a year's emission to a small volume. Intense gravitation seems the only way in which the necessary fields and particles can be ultimately held together, since all other forms of energy are on average repulsive.

The intensity of the radiation at the source must be extremely high. The dimensions of the emitting region cannot be larger than a few light-milliseconds, assuming that the earth is not in a particularly favoured position with respect to them. At the distances as they are approximately known, this implies an electron temperature of the order of  $10^{17}$  eV if it were independent electrons that cause the radiation, whatever the mechanism. But such high-energy electrons would surely lose their energy into high-energy quanta rather than the low radio frequencies, below 50 MHz, where the bulk

---

\* Theoretical work at Cornell on this subject is sponsored under a contract with the Office of Naval Research, No. N00014-67-A-0077-0007.

of the energy appears. This argument makes effectively certain that the emission mechanism is a highly coherent one, in which very large numbers of elementary charges perform a co-ordinated motion. If, for example, one guesses that the energy per charge has to be brought down to  $10^5$  eV to cause such low frequency emission, then  $10^{12}$  charges have to move in unison. Perhaps the correct number is higher or lower, but it is in any case a formidable number, and this makes clear that one cannot invoke any of the conventional cosmical radio emission mechanisms. An ordinary shock wave moving through a tenuous plasma could not be held responsible, nor could any form of synchrotron or Cherenkov radiation, except with such very large units of charge.

Without any detailed mechanism being mentioned, one may still discuss the various sets of circumstances that have been invoked.

### 1. The pulsating white dwarf

The fundamental pulsational mode of a white dwarf is thought to be always at a lower frequency than the observed range from 1.3 to 0.25 s. It is necessary to suppose that the timing is derived from a higher mode which is excited and for which an energy source is available, without exciting or causing a modulation by the lower modes.

The constancy of the period would require the structure of the star to undergo no sensible change in a year, since the period of oscillation is critically dependent upon the structure. Whether such a star may have an energy supply to feed the appropriate oscillatory mode without the diminution in this supply showing up in a slight change of structure seems questionable.

The fine structure of polarization that has been observed to repeat in successive repetition periods would require the phenomenon to be other than a simple radial pulsation. A different medium must be associated with each of the sub-pulses, so that the polarization effects may repeat in the observed manner (without invoking complex and artificial assumptions about cyclical changes in the medium within each repetition cycle).

The mechanism that would need to be responsible for generating the observed very high intensity of radiation from the mechanical motion of a large amplitude wave is not understood.

### 2. Pulsating neutron star

The periods of pulsation even of the lowest mode of neutron stars are estimated as much shorter than the observed ones. A small number of milliseconds would be possible for this, but 1.3 is not.

The decay of pulsations in a neutron star is estimated to be much faster than is compatible with the data.

### 3. Rotation of a white dwarf

It is not thought that a white dwarf star can rotate as fast as once in 0.25 s. The free orbital period around the densest model of white dwarfs would be as long as 0.5 s ( $3 \times 10^{33}$  g,  $10^8$  cm) and any spin period of the flattened star would presumably be substantially longer.

For the constancy of period a model of this kind seems favourable, since the period would be associated with the rotation and independent of all physical detail including the amplitude of the actual radio frequency generating mechanism. The sub-pulses can be associated with different parts of the surface of the star, and therefore the polarization sub-structure can be accounted for.

In all models where rotation provides the repetition period, the lengths of the individual observed pulses are determined by the beam-widths of the various rotating beams, just as in the case of a light-house beacon. The transverse dimensions available for generating the beams have thus to be large enough to provide for the shortest rise-times, and those are of the order of 3 ms. If the model consisted of something like a sunspot beaming the radiation normal to the surface there, only 1/400 of the circumference could be concerned to give the narrowest beams that occupy only 1/400 of the repetition period. For a star of radius 1000 km this would imply 15 km. A 15-km antenna can provide a beam as narrow as 1/1000 rad at a wavelength of 15 m, and such a model is thus adequate in this respect (the much smaller neutron star would be inadequate).

The mechanism of radiation giving rise to the enormously high intensities required to generate all the power within dimensions of the order of 15 km is of course still completely obscure. This would be a much higher intensity than the one demanded by the criterion mentioned earlier.

#### 4. Neutron star with orbiting planet

The possibility of an orbiting solid mass close to a neutron star has been mentioned. Such an orbiting mass could well have a period in the observed range. The lesson learnt in the case of Jupiter, where the radio emission appears to be affected by the satellite Io, is clearly at the root of this suggestion, even if the mechanism of radio emission is not clear.

However, for any orbit of such short period the differential gravitational field is immense. For a tensile strength of steel one calculates a maximum size of the order of one metre, and of course less for the case of rock. A single particle of such small size cannot be held responsible, and a swarm could not remain clustered to a small fraction of the orbit. We see no way in which orbiting solid objects could be held responsible for the observed effects.

#### 5. Rotation of neutron star binary

A pair of neutron stars could, it seems, orbit each other with the pulse repetition periods quoted. It is not clear what other details of such a model should be proposed to account for the sub-pulses, polarization, intensity, etc.

The general theory of relativity would predict a rapid decrease of angular momentum of such a binary, due to the large quadrupole moment and the consequent gravitational radiation. If this is correct, the constancy of the period could not be accounted for.

#### 6. Rotation of a neutron star and its magnetosphere

For this case the periods are in a permissible range, though of course much longer than the shortest possible. (The maximum rotation frequency

would be, like the lowest pulsational mode, in the range of milliseconds.) Many stars rotate slower than their bursting speeds, and there is no incongruity there. On the other hand, if this were the correct point of view, one might expect many more such objects to possess higher pulse repetition frequencies.

The constancy of the period can be accounted for in this case, since no gravitational radiation is expected in the case of axial symmetry. The spin energy of a neutron star is probably its largest remaining energy source, and if there are no other sources of dissipation of it, it could supply the radio emission with a decrease of less than one part in  $10^9$  per year.

As in all rotating models, the sub-structure of the pulses and the polarization properties could be understood in terms of different parts being responsible for the different features.

The mechanism of radiation proposed here is connected with a rapidly rotating magnetosphere. Co-rotation of any plasma contained in it, up to velocities close to the speed of light must then be expected, if the magnetic field is strong enough. Field-strengths at the surface of a neutron star may be as high as  $10^{12}$  G, and the electrical conductivity is expected to be very high. Whether this can be so high as to lead to really long magnetic decay time constants, despite the very small size of the object, estimated as of the order of 10 km, is not yet known. The observed structure of the pulses could only be understood in such a model if either field or plasma was very far from axial symmetry. A skew field would radiate energy principally at the basic frequency, and possibly some conversion to higher frequencies might take place in the surrounding plasma. Any plasma that is available in a particular tube of force will radiate where co-rotation enforces a high value of the relativistic factor  $\gamma$ , that is, very near to the circle at which co-rotation would imply motion at the velocity of light. This radiation will be responsible for preventing the occurrence of higher rotation speeds, through the radiation reaction. It would seem that the less plasma there is, the closer the speed of light will be approached; but in each case the radiation has to build up to just what is necessary to overcome the magnetic forces that tend towards co-rotation.

A confinement of the plasma content to a narrow sector of the field would be required, and this could be maintained only if there was a supply of plasma to such a tube of force, possibly from the stellar surface and some kind of spot on it. The longitude of that spot would then define the phase of the pulses. The beaming would be into the forward direction of the motion, for relativistic reasons, just as in the case of synchrotron radiation. Since the plasma is essentially stationary in the co-rotating frame, it is the unevenness of its charge distribution that defines the radio frequency radiation, essentially emitting the frequencies obtained by transforming the spatial Fourier components of the charge distribution in the co-rotating system to fixed co-ordinates. The observed spectrum would thus have to be understood in terms of that charge distribution, and the high intensities observed would imply that this distribution must be very uneven.

Having given a survey of the various points of view, I must say that I favour the last one. The problems of understanding a relativistically rotating magnetosphere are of course very great, but if magnetized neutron stars or white dwarfs exist, these problems must arise in the real world. For larger stars the radius at which  $rw = c$  is so large that the magneto-

sphere does not extend to it, being cut short by interaction with the external gas.

Another type of object has to be mentioned in this context, and that is a Schwarzschild singularity. If neutron stars exist, then we cannot ignore these objects, since the range of initial conditions that will set them up is larger. From the point of view of an external observer the shrinkage into the singularity takes an infinite time. For what length of time such an observer would continue to see a rotating magnetosphere is not clear, nor do we understand the rate of change of rotation speed in the external system that should be seen resulting from the collapse. Our lack of understanding of these problems does not make them less likely as candidates for the explanation of observed events. Magnetospheres rotating at relativistic velocities anchored in objects of high density, possibly close to the Schwarzschild singularity limit, are clearly a subject for further detailed study. They might have been a subject for study even if pulsars had not been discovered; now after their discovery our lack of understanding of such configurations of matter seems particularly disturbing.

There is a great variety of further radio observations that can be made and that will allow one to narrow down to the correct model. The pulse nature allows many types of radio observations that cannot be done on any other type of source, and we have here a magnificent field for the combined ingenuity of theoreticians and observers.



## **GRAVITATIONAL THEORY AND COSMOLOGY**



# THEORETICAL IMPLICATIONS OF THE KNOWN FACTS ABOUT GRAVITATION

W. THIRRING

University of Vienna, Austria

## Abstract

THEORETICAL IMPLICATIONS OF THE KNOWN FACTS ABOUT GRAVITATION. The empirical facts known about gravitation are classified and the properties of gravitational theory proved by them are discussed.

The following analysis of the theoretical implications of the empirical facts known about gravitation will be based on what is regarded as reasonable theories. These are theories constructed after the pattern of classical electrodynamics where the field equations and the equations of motion of the particle result from the same action functional  $W$ . This is the sum of the action functional for free particles  $W_p$  and that for gravitational field  $W_f$  and an interaction term  $W_i$ . If we describe the motion of the particle by the four co-ordinates  $Z^i$ ,  $i = 0, 1, 2, 3$  as function of proper time  $s$  we have

$$W_p = \frac{m}{2} \int ds \dot{Z}^i(s) Z^k(s) g_{ik}$$

with the metric tensor

$$g_{ik} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}$$

The interaction term is, e.g. for electrodynamics,

$$W_i = e \int ds \dot{Z}^i(s) A_i(Z(s))$$

where  $A_i$  is the vector potential. The structure of  $W_f$  will depend on the kind of field used and is more complicated for tensors of higher rank.

To classify the empirical evidence we shall introduce the following concepts:

I. Strong universality: A theory is said to be strongly universal when the trajectory of particles of identical internal structure in a gravitational field depends only on their initial conditions and not on their mass.

II. Weak equivalence: This requires that the effect of a constant external field can be transformed away by going into a suitably accelerated frame provided the gravitational interaction with the system is neglected.

III. Strong equivalence: The same conditions without the proviso.

These conditions gradually sharpen our feeling about the universal nature of gravitation, i.e. the logical implications are III → II → I. I is called strong universality because electromagnetism which we usually call universal would satisfy it only if  $e/m$  were the same for all particles. Generally, if  $W_i$  is proportional to  $m$  we have strong universality. This does not yet imply II because different initial conditions may require a differently accelerated frame to transform the field away. This happens, e.g. in electrodynamics, even if  $e/m$  is universal, i.e. if I holds. A large class of theories which satisfy weak equivalence are obtained if one sets

$$W_i = m \int ds \dot{Z}^i(s) \dot{Z}^k(s) \psi_{ik}(Z(s))$$

where  $\psi_{ik}$  specifies the gravitational potential. Of course, this coupling has to be universal, that is to say for each particle  $\psi_{ik}$  has to be coupled to its energy-momentum tensor. In this case these theories have the remarkable property that the observable metric is not the original pseudo-euclidian  $g_{ik}$  but the Riemannian  $g_{ik} + \psi_{ik}$ . This conclusion is reached by studying the effect of  $\psi_{ik}$  on the size and frequency of atoms. Thus, the Riemannian structure of space-time is automatically generated by these theories and need not be postulated separately. Whether III is also satisfied depends on  $W_f$ . Generally the symmetric tensor  $\psi_{ik}$  contains as irreducible spin components  $0 \oplus 0 \oplus 1 \oplus 2$ . If one uses a conserved energy-momentum tensor only a spin 0 and the spin 2 couple. Thus, these theories are characterized by one parameter  $\alpha$  which describes the admixture of a scalar field. It turns out that only for the extreme cases,  $\alpha = 0$  and  $\alpha = 1$ ,  $W_f$  can be arranged so that the theories contain strong equivalence.  $\alpha = 0$ , i.e. only spin 2 corresponds to Einstein's theory.

In terms of these concepts the empirical facts test the properties shown in Table I.

TABLE I. PROPERTIES TESTED BY CONCEPTS CONSIDERED

Strong Universality	Weak Equivalence	Strong Equivalence	Effect
$W_i = m$ (independent of $m$ )	$W_i = \int T^{mn} \psi_{mn}$	$\alpha = 0, 1$	
$10^{-14} - 10^{-20}$	-	-	$K^0 \rightarrow 2\pi$
$10^{-11}$	$10^{-9}$	-	Eötvos, Dicke
no	yes	yes	determined metric
no	yes	yes	determined redshift
no	no	no, only if	determined right light deflection
no	no	no, only if	determined right motion

The most precise evidence for strong universality comes from the  $K^0 \rightarrow 2\pi$  decay. Any difference in the gravitational mass between  $K^0 - K^0$  would result in a  $K_1 - K_2$  mixing and therefore a fast  $K_L \rightarrow 2\pi$  decay. Present experiments require this mass difference relation to  $m_K$  to be less than  $10^{-14}$  or  $10^{-20}$  depending on whether one believes that we feel here the gravitational potential of our galaxy only or the whole universe. However, this tests only strong equivalence since  $K^0$  and  $\bar{K}^0$  presumably have the same internal structure. By this we mean that its constituents move with the same relative velocity, unlike proton and nuclei or proton and electron. The ratio of their gravitational coupling to their mass is measured in the Eötvös experiment where Dicke has now achieved an accuracy of  $10^{-11}$ . However, electrons or the kinetic energy of nucleons contribute only about 1% to the weight of matter such that it tests weak equivalence to  $10^{-9}$ . Of course, protons and neutrons contribute in the same amounts to the weight of matter but since they should have essentially the same internal structure one measures only strong universality to  $10^{-11}$ .

The redshift tests only weak equivalence and less accurately, and there is no direct experimental evidence on strong universality. The observed light deflection rules out a pure scalar theory,  $\alpha = 1$ , but is compatible with a small admixture,  $\alpha < 1/10$ . The perihelion motion is a very subtle effect which comes out just right for  $\alpha = 0$ . In view of the many corrections which have to be applied to the data it is not clear how stringent this test is.

Summarizing, one may say that Einstein's theory is, at least, very close to the truth.



# RECENT DEVELOPMENTS IN OBSERVATIONAL COSMOLOGY

D.W. SCIAMA

Department of Applied Mathematics and Theoretical Physics,  
University of Cambridge,  
Cambridge, United Kingdom

## Abstract

RECENT DEVELOPMENTS IN OBSERVATIONAL COSMOLOGY. 1. Radio sources; 2. Helium; 3. The X-ray background; 4. The excess microwave background; 4.1. Existence; 4.2. Origin; 4.3. Effects; 4.4. Isotropy.

The last few years have seen dramatic developments in observational cosmology thanks to radio astronomy (together with related optical discoveries) and to X-ray astronomy. In fact it is no longer possible to say where astrophysics ends and cosmology begins. This has been amply demonstrated at this symposium where many of the astrophysical lecturers were forced into cosmological considerations. Let me first briefly summarize the developments discussed in these Proceedings.

## 1. RADIO SOURCES

(a) Quasi-stellar objects. These are discussed by Schmidt. If their red shifts are cosmological in origin then, since the record red shift  $z (= \delta\lambda/\lambda)$  is 2.36, we are able to see out to distances comparable with the radius of the universe and back over times comparable with the age of the universe.

(b) Radio source counts. F.G. Smith has discussed the most extensive counts which have recently been published by Pooley and Ryle. Their chief characteristic (as has been known for some years) is an excess of faint sources. The most immediate interpretation of this is in terms of an intrinsic evolution in the sources with cosmic epoch.

## 2. HELIUM

The abundance of helium in, for instance, the sun is about 22% that of hydrogen by mass. It is difficult to account for this great quantity of helium in terms of processes occurring since the formation of our galaxy. On the other hand, as has been explained by Hoyle, the  $\alpha$ - $\beta$ - $\gamma$  theory accounts nicely for the helium as a result of thermonuclear processes occurring about 100 s after the "hot big bang", especially if one uses in the calculation a value of about 3°K for the present black-body temperature of the universal radiation field (see below).

### 3. THE X-RAY BACKGROUND

This is discussed by Morrison. Its origin is uncertain, but is likely to be cosmological, that is, most of the flux probably comes from regions of substantial red shift. A possible mechanism seems to be (inverse) Compton collisions between relativistic electrons and the black-body radiation. Whatever the mechanism, the observations provide a useful upper limit for any universal process which leads to the production of X-rays.

### 4. THE EXCESS MICROWAVE BACKGROUND

This is discussed in detail below. The present data suggest that this background has a black-body spectrum over the wavelength range, so far studied with a temperature of about 3°K. The most likely explanation is that it is radiation left over from the hot big bang.

These results, especially the radio source counts and the excess microwave radiation, have tended to swing opinion away from the steady-state theory. However, the situation is complicated and an assessment is given by Gold.

I now discuss the excess microwave background in more detail. It is convenient to do this under four headings: existence; origin; effects; and isotropy.

#### 4.1. Existence

Either by direct measurement or by indirect argument from excited inter-stellar CN absorption lines, the intensity of the excess background is known from 70 cm to 2.5 mm. I am not competent to give a critical discussion of these observations, but for what it is worth my impression is that the case for a black-body spectrum is now rather strong (despite the reservations that have been expressed to me privately by several astronomers). The critical region of the spectrum is around the peak which for 3°K would be near 1 mm. Several groups are planning to observe this region, which has to be done from above the atmosphere. I might add that small deviations ( $\sim 25\%$ ) from an exact black-body spectrum could be explained by reasonable cosmological processes. For the purpose of this paper I shall assume that the radiation has a black-body character.

#### 4.2. Origin

##### 4.2.1. Hot big bang

Dicke and his colleagues have revived the old (1948) theory of Gamow and Alpher that the universe had an infinitely hot dense beginning at  $t = 0$ . At early times the radiation would rapidly have reached thermal equilibrium with matter and so would have had a black-body spectrum. As the universe expands, the radiation cools but, apart perhaps from small deviations, retains its black-body character.

I do not think that this theory should be taken too literally. It fails to explain why the present value of the black-body temperature is 3°K

rather than any other value. We can express the problem more invariantly in terms of the entropy of the radiation per baryon of matter. If localized entropy production may be neglected, the entropy per baryon is independent of time during the expansion of the universe; and we may almost think of it as a "constant of nature". Its observed value is about  $10^8 k$  ( $k$  is Boltzmann's constant), and this value requires explanation. According to the conventional view, it is one of the initial conditions holding at  $t = 0$ . It may be necessary to assume this, but we should at least attempt to find an explanation.

#### 4.2.2. Previous contraction

It has often been conjectured that the present expansion of the universe was preceded by a contracting phase, despite our inability to follow this through in general relativity (see 4.4 (f), singularities). In such a contracting phase galaxies and stars may have formed, and the resulting release of radiation would, because of the blue shift, lead to a net entropy in the form of radiation at the start of the expansion phase (binding energy would be restored to galaxies and stars near the end of the collapse without using up all the blue-shifted radiation). If this point of view is correct, it would be possible in principle to compute the entropy per baryon at the start of the expansion phase.

#### 4.2.3. Dissipation processes during expansion

The importance of dissipation processes in cosmology has recently been realized. I shall mention only that neutrino-electron interactions at  $10^{10} \text{ }^\circ\text{K}$  and photon-electron interactions at  $10^4 \text{ }^\circ\text{K}$  give rise to appreciable viscosity. Inhomogeneities and anisotropies would tend to be smoothed out and in the process heat would be generated. It seems possible that most of the  $10^8 k$  entropy per baryon could be produced in this way (as suggested by Misner and by Rees), but this remains to be seen.

#### 4.2.4. Discrete sources

Several people have proposed that the excess microwave background may arise from the superposed radiation of many discrete radio sources whose spectra rise with increasing frequency in the manner required to simulate the present observations. At the moment this seems less plausible than the hot big bang explanation, although of course it will always be necessary to estimate the integrated radiation from discrete sources at all wavelengths of interest.

### 4.3. Effects

#### 4.3.1. Primeval element formation

From what we have already said, it is clear that the radiation provides the heat necessary to promote the thermonuclear reactions envisaged by the  $\alpha$ - $\beta$ - $\gamma$  theory. In particular it seems quite likely that most of the helium in our galaxy was formed in this way, although there are still difficulties to be overcome.

#### 4.3.2. Galaxy formation

In the early stages of the expansion the radiation dominated dynamically over the matter. It seems unlikely that galaxies could begin to form until this domination ceased.

#### 4.3.3. Interaction with high energy particles

The following is a summary of a part of Morrison's lecture in these Proceedings.

(a) Electrons. Relativistic electrons in radio sources (including our own galaxy) suffer (inverse) Compton collisions with black-body photons. These collisions constitute a serious energy drain on the electrons and lead to the production of an appreciable flux of X-rays.

(b) Protons and heavy nuclei. In the energy region around  $10^{20}$  eV, particles have a mean free path through the radiation which is much less than the radius of the universe (as a result of pion formation). This is important because it is likely that such high energy particles in the cosmic ray flux have an extragalactic origin.

(c)  $\gamma$ -rays. As a result of pair production,  $\gamma$ -rays of energy exceeding  $10^{15}$  eV have a mean free path through the radiation which is less than the radius of our galaxy. We may soon know whether  $\gamma$ -rays of such energy are in fact incident on the earth.

#### 4.4. Isotropy

The original measurements by Penzias and Wilson on the excess microwave background showed that the radiation was isotropic to within a precision of 5%. More recently, the precision has been improved to about 0.1% over the limited regions of sky which have been specially studied (by Partridge and Wilkinson and by Conklin and Bracewell). If the background is cosmological in origin, this is by far the most accurate measurement in cosmology. It has many profound implications. In order to understand them it is helpful to consider the immediate rather than the ultimate source of the background radiation. A similar distinction arises when we look at a hot body like the sun. The ultimate source of its radiation is its very hot central regions, but the radiation from the centre is scattered many times before it reaches us. The immediate source, what we actually see, is the surface of last scattering, that is, the photosphere of the sun. This is not a strict surface, of course, but for many purposes we may suppose that we can see a distance into the sun corresponding to unit optical depth.

In the same way the ultimate source of the black-body background is (presumably) the hot big bang which is at essentially infinite red shift from us. However, on its way to us this radiation is scattered many times by matter (mainly free electrons), and its immediate source are the electrons which last scattered it. The red shift  $z_0$  of these electrons depends on the cosmological model adopted, but in no case is less than seven and may be very much greater. We are thus "seeing" out to very great distances and

the black-body radiation carries information to us about physical conditions at these great distances.

The main information is precisely the high isotropy we are now discussing. Its implications can be inferred from the simple but general result that, if the present temperature of the radiation in any given direction is  $T$ , then its temperature on the last scattering surface is  $T(1 + z_0)$ . If  $T$  is independent of direction to a high precision then it can be assumed to a corresponding precision that:

- (a) The radiation temperature was the same everywhere in the universe at a red shift  $z_0$  from us;
- (b) The red shift  $z_0$  itself is the same in all directions, that is, the universe is isotropic with respect both to optical depth and to expansion rate;
- (c) Peculiar velocities of the electrons on the last scattering surface (with respect to the substratum) are low;
- (d) Large-scale density irregularities along the lines of sight to the last scattering surface are severely limited (since the gravitational effects of such irregularities would modify the red shift of the last scattering surface).

These results taken together show that on a large scale the universe is highly homogeneous and isotropic. These symmetry properties underlie the standard Friedmann-Robertson-Walker models of the universe and they now appear to hold with greater precision than any cosmologist (except perhaps Milne!) had ever dared to hope. The urgent question then arises: why is the universe so homogeneous and isotropic? Some people believe that it is a matter of initial conditions at  $t = 0$  (as in the case of the value of the radiation entropy per baryon). However, recent work by Misner, by Doroshkevich, Novikov and Zeldovich, and by Stewart has opened up the possibility that some at least of this uniformity may be due to dissipative processes smoothing out non-uniformities after  $t = 0$  (the same dissipative processes indeed as may produce most of the observed entropy per baryon).

We now consider two further consequences of the high isotropy of the background radiation:

(e) Our own peculiar velocity must be low, since otherwise, because of the Doppler effect, there would be a dipole distribution of intensity with direction in any plane. Partridge and Wilkinson have looked explicitly for such a dipole distribution, and their failure to find it limits our peculiar velocity (at a declination of  $-8^\circ$ ) to about 300 km/s. This result is of great significance because our estimated peculiar velocity is of the same order. A slight improvement of observational precision is thus likely to lead to a positive measurement.

Such a positive measurement would be important for two types of problem, namely, the validity of Mach's principle and the determination of the largest local structure to which we are dynamically related. Mach's principle requires that a dynamically non-rotating inertial frame should be non-rotating relative to the bulk of the matter in the universe. We may ask: with what precision can the identity of these two types of non-rotating frame be established experimentally? Now the sun is believed to rotate around the centre of our galaxy with a velocity of about 250 km/s. This rotation is with respect to an inertial frame (the galaxy bulges at its equator and is flattened at its poles) but we do not yet know whether this rotation is also with respect to external galaxies. There is a program at Lick observatory, and I believe in the Soviet Union, to use external galaxies as

reference points (and probably quasi-stellar objects will also be used). If these programs are successful, the agreement of the two non-rotating frames would be tested to a precision somewhat exceeding 0.5 seconds of arc per century. We can make a similar test using the background radiation. The reason is that a measured velocity relative to this radiation represents a velocity relative to the last scattering surface, that is, relative to the bulk of the matter in the universe. Thus if, because of the rotation of our galaxy, the sun has a velocity of 250 km/s through the background radiation, this fact should show up in the next round of measurements, leading again to an agreement of the two non-rotating frames to better than 0.5 seconds of arc per century.

However, this situation is complicated by the second of our problems, namely, that we may belong to a dynamical structure larger than the galaxy which confers on us a further peculiar velocity. For instance, our galaxy probably belongs to a local group of 20 or so galaxies, and it has been estimated to move relative to this group at a velocity of about 100 km/s. Some astronomers also believe that the local group may belong to a flattened supercluster of galaxies which may be rotating about a centre in the Virgo cluster of galaxies. This rotational motion of the supercluster is very controversial, but since estimates of the resulting velocity of the local group are of the order of several hundred kilometres per second it is clear that this question will be much clarified by a reliable measurement of our velocity through the background radiation. We may add that if the effect is found, the identity of the two non-rotating frames would be tested to a precision of about  $10^{-3}$  seconds of arc per century.

(f) In conjunction with certain other reasonable assumptions, the high isotropy of the background radiation implies that there was a physical singularity of the universe somewhere within or on our past light-cone. This is perhaps the most unexpected consequence of the high isotropy. Penrose explains the significance of theorems about singularities in general relativity. Such theorems are of special importance in cosmology, partly because we lie within the system which contains the singularity (unlike the collapsing star problem where the singularity cannot communicate with us) and partly because we would like to investigate whether a collapsing universe can turn into an expanding universe.

The Friedmann-Robertson-Walker models with zero cosmical constant have a point singularity which effectively prevents any continuation of the solution. The question has been much discussed whether the singularity is an artefact resulting from the exact symmetry assumptions underlying the models. Hawking in particular has proved a number of powerful theorems which show that this is not so. Under reasonable assumptions any somewhat irregular universe contains a singularity (although not, in general, of a point-like character). The most powerful theorem, due to Hawking, Ellis and Penrose states that Einstein's field equations together with the energy and causality conditions described in Penrose's lecture, and one more condition, guarantee the presence of a singularity. This final condition is that there exists a point whose past light-cone begins to reconverge at some distance from the point. The idea of such reconvergence is well known to astrophysical cosmologists for whom it corresponds to a minimum in the angular diameter of an object as a function of its distance from us. In the Einstein-de Sitter model, for instance, this minimum occurs at a

red shift of less than two, that is, at a red shift less than that of several quasi-stellar objects.

The point of all this is that the isotropy of the black-body radiation ensures that this final condition is satisfied in the actual universe. The reason is that, if the radiation was last scattered at the minimum possible red shift of seven, then there is enough intergalactic matter to produce the reconvergence. If, however, the scattering occurred at a red shift very much greater than seven, then the isotropy of the radiation shows that the universe has been close to a Friedmann-Robertson-Walker model out to this large red shift. Calculation shows that this also guarantees the re-convergence of our past light-cone.

Arguments of this sort do not, however, tell us much about the nature of the singularity, nor how to avoid it if this is thought desirable. These are hard problems whose solution lies in the future.



# OBSERVATIONAL COSMOLOGY: OPTICAL WAVELENGTHS

E. M. BURBIDGE

University of California,

San Diego, Calif., United States of America

## Abstract

OBSERVATIONAL COSMOLOGY: OPTICAL WAVELENGTHS. 1. Hubble relation; 1.1. Cosmological models; 1.2. Results; 1.3. Observational problems; 2. Impact of quasi-stellar objects on cosmology.

In the first part of this paper I want to give the material that Sandage would have presented, if he had been able to come to this symposium. This will be the present situation regarding the determination of the form of the relation between red shift  $z = (\lambda_{\text{meas}} - \lambda_0)/\lambda_0$  and the apparent luminosity of galaxies — the Hubble relation. Much of it was contained in his Halley Lecture (Sandage, 1968a; for earlier work see Humason, Mayall, and Sandage, 1956, and Sandage, 1961).

In the second part, I shall say a few words about the quasi-stellar objects and their place in the cosmological scheme.

## 1. HUBBLE RELATION

That a systematic relation exists between the luminosity of galaxies and their red shifts was discovered by Hubble in 1929 and followed up by Hubble and Humason. At this point I have to define the usual term used by astrophysicists to measure luminosity — the magnitude — because discussions of the Hubble relation always use this unit. If two objects have luminosities  $\ell_1, \ell_2$ , then their magnitudes are connected by the relation

$$\frac{\ell_1}{\ell_2} = 10^{0.4} (m_2 - m_1) \quad (1)$$

Thus magnitude is on a logarithmic scale, and the fainter the object, the larger is its magnitude.

For nearby objects, where effects of space curvature are negligible and the first order Doppler formula applies,

$$z = \frac{\Delta\lambda}{\lambda_0} = \frac{v}{c} \quad (2)$$

Hubble's relation implied a proportionality between the magnitude  $m$  and  $\log z$  for galaxies. He interpreted this as being due to the expansion of the universe. We have from the inverse square law for luminosity

$$\ell \propto D^{-2} \quad (3)$$

if D is the "luminosity distance" of any galaxy. If we have a set of galaxies of equal and known intrinsic luminosity, the measured luminosity can be used to give D by Eq. (3). In an expansion starting at a point, the distance is proportional to the velocity; combination of this with Eqs (1), (2), and (3) for relatively nearby galaxies gives

$$m = \text{const.} + 5 \log z \quad (4)$$

This is the relation Hubble found, and we shall see presently how well it is obeyed by Sandage's most recent data.

### 1.1. Cosmological models

Of all the possible cosmological models, very few have been investigated in detail. What has been done amounts to a study of just two kinds of model; exploding or big-bang models, first suggested by Friedmann in 1922, and the steady-state model in its original form as suggested by Bondi, Gold, and Hoyle in 1948. In both of these, a uniform isotropic universe is assumed. Friedmann models were investigated mathematically by Robertson and Walker; they have a zero "cosmological constant"  $\Lambda$ . Starting with a Friedmann singularity, a non-empty universe expands and the subsequent expansion velocity will be decelerated by self-gravitation of the matter. In the steady-state model, the expansion accelerates as new matter is created and fills the void left by the expansion, so that a density constant in space and time is maintained. The existence of large-scale inhomogeneities is possible. In a version of the steady-state theory suggested by McCrea, matter forms preferentially in regions of higher-than-average density, i.e. in nuclear regions of galaxies.

The task that Sandage has been tackling for the past ten years has been the determination of the present-day values of the Hubble constant,  $H_0$ , and the so-called deceleration parameter,  $q_0$ . These are given by

$$H_0 D = c \frac{\Delta \lambda}{\lambda_0} = c z \quad (5)$$

and

$$q_0 = - \frac{\ddot{R}_0}{H_0^2} R_0 \quad (6)$$

where  $R_0$  is the present value of the radius of the universe, in the "big-bang" theories. The present mean matter-density of the universe will then be

$$\rho_0 = \frac{3 H_0^2 q_0}{4 \pi G} \quad (7)$$

and the cycle time T for Friedmann models with positive curvature ( $q > \frac{1}{2}$ ) is

$$T = \frac{2 \pi q_0}{H_0 (2 q_0 - 1)^{3/2}} \quad (8)$$

In the steady-state model,  $R \propto \exp Ht$  and  $q_0 = -1$ .

### 1.2. Results

Sandage has used the work of Mattig (1958), in which explicit relations are derived connecting the observable quantities  $m$  and  $z$ , for the Friedmann models. His latest results are shown in Fig. 1 (Sandage, 1968b). The abscissa represents what is essentially a bolometric magnitude (I shall return to this later, as it has been a source of uncertainty in the past), and the ordinate is the directly observable  $\log cz$ . The points extend to  $z = 0.46$ , but there are few points beyond  $z = 0.2$ . From Eq. (4) the nearby points should be fitted by a line of slope 5, and the line that has been drawn has exactly this slope. The small black box in the lower left corner is the range over which Hubble first formulated the velocity-distance relation in 1929. A least squares fit to the solid dots gives

$$\text{abscissa} = \text{const.} + (4.998 \pm 0.003) \log cz$$

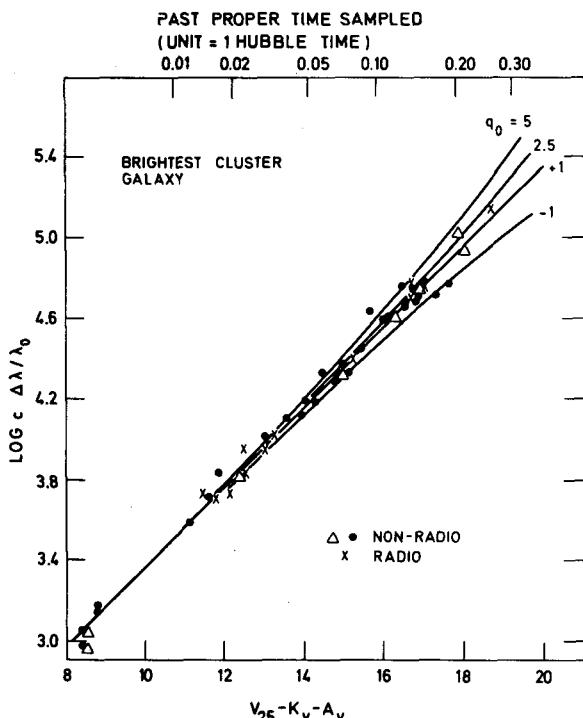


FIG. 1. Latest results by Sandage.

Only the non-radio galaxies have been used in this fit, though Sandage believes that the radio galaxies fit a parallel sequence 0.3 magnitudes below this. The slope of 5 is given by all theories as  $z \rightarrow 0$ . It is the exact relation for Friedmann models with  $q_0 = +1$ ,  $\Lambda = 0$ .

Figure 2 shows the same points, and curves for several values of  $q_0$ , from Sandage (1968a). Sandage believes that the data are consistent with  $q_0 = +1$ , although his formal solution from the points gives  $q_0 = 1.6 \pm 0.5$

as a preliminary result. The universe, according to these results, has a closed form (oscillating model).

The determination of  $H_0$  depends on the calibration of the plot — i.e. on determining the actual linear distance for one of the points at the lower left of the diagram. The current calibration gives

$$H_0 = 75.3^{+19}_{-15} \text{ km/s per Mpc}$$

or

$$H_0^{-1} = 12.9^{+3.7}_{-2.7} \times 10^9 \text{ yr}$$

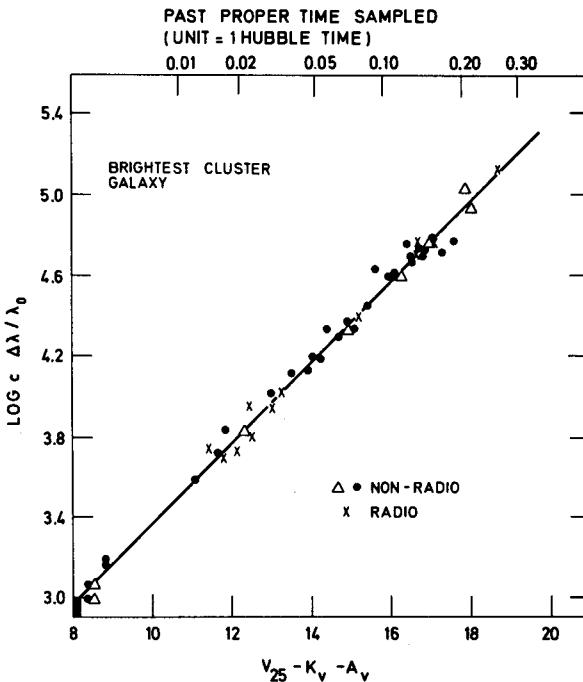


FIG. 2. Same points as in Fig. 1, but with different values of  $q_0$

The formal dispersion, considered to be only in the abscissa in the diagram, is quite small,  $\pm 0.3$  mag. Various random uncertainties in the calibration have been included in giving the errors on  $H_0$ . Additional possible systematic errors in the calibration could decrease  $H_0$  to

$$H_0 = 55 \text{ km/s per Mpc}$$

The time  $T$  for the whole cycle, for  $H_0 = 75.3$ ,  $q_0 = +1$ , is

$$T = 8.2 \times 10^{10} \text{ yr}$$

and the time since the last Friedmann singularity is

$$t = 7.4 \times 10^9 \text{ yr}$$

The mean mass density  $\rho_0$  is

$$\rho_0 = 2 \times 10^{-29} \text{ gm/cm}^3$$

### 1.3. Observational problems

There are four particular problems involved in the work just described. These are:

- (1) Galaxies of the same intrinsic luminosity (standard candles) must be selected.
- (2) Magnitudes measured through a filter of limited band-pass must be converted to something that is equivalent to a bolometric or total magnitude.
- (3) Evolutionary effects due to the aging of stars in galaxies must be evaluated - i.e. it is necessary to show that  $d\ell/dt = 0$  or to determine its value.
- (4) The calibration for nearby galaxies must be determined.

Point (1) seems on a better footing now. Sandage has found that the brightest elliptical galaxies in clusters have less dispersion in magnitude than the average of the brightest several members (as used by Humason, Mayall and Sandage, 1956). The former appears to be independent of the number of galaxies in a cluster, while the latter is not. There is no correlation between residuals from the fit and the number of galaxies in a cluster, and this eliminates the possibility of a selection effect in that brightest galaxies in distant clusters might be systematically chosen from richer clusters and might therefore be brighter than nearby brightest galaxies (the so-called "Scott effect").

The second point has given an inordinate amount of trouble. In principle there is no problem, if one could measure the total radiation at all wavelengths coming from all galaxies that one uses. In practice this is an impossible task; it would require far more telescope time than is available, since so few telescopes in the world are large enough and well-enough equipped to do this work, and in addition, wavelengths shorter than the atmospheric cut-off in the rest wavelength system become detectable as the red shift increases while the long wavelengths disappear likewise into the infra-red where the atmosphere again blocks radiation.

In practice one measures the radiation received through a broad-band filter centred on about 5500 Å, on half-width about 900 Å. A typical elliptical galaxy, in which all the light is coming from a stellar population of known characteristics, has a distribution that resembles a Planck function for a temperature in the region  $T = 4000^\circ\text{K}$  but that is actually not well represented by such a function. For galaxies with different red shifts, the filter band-pass is located at different places in the intrinsic  $I(\lambda)$  curve.

Oke and Sandage (1968) now have an accurate determination of  $I(\lambda)$  for nearby elliptical galaxies, and they have, by measuring the radiation through different band-pass filters, shown that, over the whole range of the figures, there appears to be no change in the intrinsic  $I(\lambda)$  curve - i.e. out to  $z = 0.46$ ,  $I(\lambda)$  is the same. Earlier work by Stebbins and Whitford had suggested there was a change in  $I(\lambda)$  with  $z$ , in the sense that the more distant galaxies were redder than the nearby ones.

This result, in conjunction with the known theory of stellar evolution for the stars in the appropriate mass range, has led Oke and Sandage to

conclude that  $d\ell/dt \approx 0$  also. The large mass-to-light ratios of elliptical galaxies mean that a large fraction of the light must be coming from un-evolved low-mass dwarfs, and the time-scales we are talking about out to  $z = 0.46$  are only  $\sim 3 \times 10^9$  yr, which is short compared to the evolutionary lifetime of even a solar-mass star.

The calibration of the bright end of the Hubble diagram is another point that has taken a great deal of effort. The steps are as follows.

The pulsating variable stars, in which intrinsic luminosity is closely correlated with the period of variation — the Cepheids — can be calibrated in our own galaxy, in the solar neighbourhood. These can then be seen in the Andromeda Nebula and other nearby galaxies, so that the distances of these can be determined. But galaxies in the local group are too near to be partaking of the general expansion of the universe, and for galaxies a little further away, the random motions are not small relative to the general outward motion. So new distance indicators, brighter than the Cepheids, must be found.

Sandage has recently used two — the clouds of ionized gas that surround complexes of hot young stars, which have a remarkably constant distribution of diameters in a given class of spiral galaxy, and the luminosities of the dense star clusters that lie in the outer parts of the massive galaxies of both spiral and elliptical classes (the "globular clusters"). These are calibrated in nearby galaxies, by means of the Cepheids, and then one can measure them in the nearest of the large galaxy clusters in which a reliable recession velocity can be determined — the Virgo cluster.

This looks now to be on a rather firm footing. But this concerns only the horizontal axis in the Hubble diagram; what about the vertical axis? Since red shifts can be measured with arbitrary precision, no one has bothered about errors here until recently. But we must really worry about whether the mean red shift of the Virgo cluster is wholly due to the expansion of the universe, because a local anisotropy in the velocities out to  $v = 4000$  km/s has been found by de Vaucouleurs. This is still far from being adequately studied (mainly because of lack of a good large telescope in the southern hemisphere). Sandage has, however, estimated its amount, and has derived

$$H_\infty/H_{vc} = 1.17 \pm 0.09$$

where  $H_{vc}$  is the Hubble constant derived from the Virgo cluster data alone and  $H_\infty$  is the value derived above. This ratio is considerably less than an earlier estimate by de Vaucouleurs. The effect dies out, Sandage believes, beyond  $z = 0.05$ , because the scatter in the Hubble diagram is small beyond this. If the corrections for this effect were to move the clusters appreciably upwards in the diagram, the whole grid of lines would be displaced to the left in the figures, and this would change the best fit in the direction towards  $q_0 = 0$ .

Investigation of this anisotropy, and research into its physical meaning, seems to me to be the most urgent and interesting new task awaiting us in this field. It is obviously necessary to take account of the large velocity differences that can occur in groups of galaxies to which I refer in my paper on extragalactic astronomy.

## 2. IMPACT OF QUASI-STELLAR OBJECTS ON COSMOLOGY

When the large red shifts of the QSOs were discovered, it was hoped that they would be a most powerful tool for cosmology. As Schmidt shows in these Proceedings, the enormous scatter in the plot of  $\log z$  versus magnitude for QSOs has clearly ruled out the possibility of using this. The plots of  $\log N - \log S$  for radio sources (number counts to given radio flux limits) are discussed elsewhere in these Proceedings. I might mention that this approach was attempted by Hubble in the 1930s using optical galaxies, but he abandoned it because of lack of knowledge of the intrinsic luminosity distribution of galaxies. Ryle and Longair have made analyses of the data on  $\log N - \log S$ , and Schmidt has studied the number density as a function of  $z$ ; both have concluded that evolutionary effects would be necessary to explain the data.

These studies, however, assume that the red shifts of QSOs are cosmological. As has been indicated, there are grave difficulties in the way of accepting this interpretation. As a result of these difficulties, Terrell suggested that the red shifts are Doppler shifts due to relativistic outward velocities of comparatively local objects, while Burbidge and Hoyle, and Hoyle and Fowler have suggested that the red shifts might be gravitational and intrinsic to the objects themselves.

One set of data that have led to serious consideration of this latter possibility has been the distribution of the red shifts among the QSOs with  $z$  near 2. Originally, it was found that absorption lines were preferentially in QSOs of large  $z$ , and Schmidt outlined the remarkable recent result that absorption-line red shifts much lower than the emission-line value are found in certain QSOs. One has either to interpret these as being due to intergalactic absorbing clouds or galaxies lying along the path length to a cosmologically distant QSO, or relativistically moving shells or puffs of gas around the objects, or shells lying in regions of different gravitational potential in the gravitational red shift theory. There are difficulties in the way of all three interpretations.

The original situation with the absorption lines was that several QSOs were found to have these at  $z = 1.95$ . Now there seems to be significance in this number for emission-line red shifts also. The situation is shown in Table I. We have found also some clumping of red shifts at other values of  $z$ , but this work is in a preliminary stage.

Mushotzky (1968) has plotted  $N(z)$  against  $z$  for QSOs and has found two distinct distributions. There is one large peak, containing the greatest number of QSOs, centred at about  $z = 0.5$  or  $0.6$ . The curve then drops to a low minimum, and there is a second quite narrow peak at  $z = 1.9$  or  $2.0$ . This suggests two kinds of QSOs, with quite different properties. It is in the second peak that the queer objects with multiple absorption-line red shifts occur.

Shklovsky and Khardashev have suggested that the numbers of absorption-line red shifts around  $z = 1.95$  might be produced in a universe of Lemaitre type, with a long halt in the expansion at an epoch corresponding to this red shift. Such a universe would have a non-zero cosmological constant. This has led them to a large age of  $\sim 7 \times 10^{10}$  yr for the time since the expansion started. There are spectroscopic difficulties in interpreting the absorption lines in this way. Petrosian and Salpeter have made an alternative suggestion to explain the general piling-up of red shifts around  $z = 2$ .

TABLE I. QUASI-STELLAR OBJECTS WITH EMISSION-LINE RED SHIFTS GREATER THAN 1.9

Object	$z_{\text{em}}$	$z_{\text{abs}}$	Object	$z_{\text{em}}$	$z_{\text{abs}}$
4C 29.50	1.927	-	3C 9	2.012	-
PHL 61	1.93	-	Ton 1530	2.046	1.9362, 1.9798, 2.0553
3C 191	<u>1.956</u>	<u>1.947</u>	PHL 1305	2.064	-
PKS 0119-04	<u>1.955</u>	<u>1.965</u>	PKS 0229 + 13	2.07	-
PHL 938	<u>1.955</u>	1.9064, 0.6128	B 189	2.075	-
BSO 6	<u>1.956</u>	-	BSO 11	2.084	-
PHL 5200	1.98	1.90-1.98, <u>1.9502</u>	PKS 0106 + 01	2.107	-
PKS 1148-00	1.982	-	PKS 1116 + 12	2.118	<u>1.947</u>
PHL 1127	1.990	<u>1.95</u> (one line)	PKS 0237 - 23	2.223	2.2017, <u>1.9556</u> , 1.6744, 1.6715, 1.6564, 1.5958, 1.5132, 1.3646
LB 8755	2.010	-			

In conclusion, I think that we have two quite different situations in observational cosmology. Sandage's presentation of his data on galaxies suggests that we have a smooth, orderly, expanding universe out to  $z = 0.46$ , or, at any rate, out to  $z = 0.2$ . (But even here we have to understand and take account of the local velocity anisotropy.)

The QSOs, on the other hand, suggest either that the cosmology is far more complicated than anything that has been considered yet, or that we are far from understanding the behaviour of matter in strong gravitational fields.

#### R E F E R E N C E S

- HUMASON, M.L., MAYALL, N.U., SANDAGE, A. (1956) *Astr. J.* 61, 97.  
MATTIG, W. (1958) *Astr. Nachr.* 284, 109.  
MUSHOTZKY, R. (1968) *Nature, Lond.*, in preparation.  
OKE, J.B., SANDAGE, A. (1968) *Astrophys. J.*, in press.  
SANDAGE, A. (1961) *Astrophys. J.* 133, 355.  
SANDAGE, A. (1968a) *Observatory* 88, 91.  
SANDAGE, A. (1968b) *Astrophys. J.* 152, L149.



# REMARKS ON GRAVITATION AND COSMOLOGY

R. H. DICKE  
Palmer Physical Laboratory,  
Princeton, N. J.,  
United States of America

## Abstract

REMARKS ON GRAVITATION AND COSMOLOGY. In discussing a relation between stellar evolutionary ages and gravitational theory, the author makes reference to the scalar-tensor theory and the technique of dating by the radioactive decay of uranium.

I wish to discuss here a relation between stellar evolutionary ages and gravitational theory. A star of about one solar mass has a luminosity varying roughly as  $G^7$ . Under the scalar-tensor theory  $G$  should have been larger in the past and the star should have been brighter. The brighter star would exhaust its fuel more quickly and appear to be older than it actually is (Dicke, 1966). For the extreme population II stars, the oldest in the galaxy, this age correction should be by a factor of two to three, provided that the stars were formed in the first  $2 \times 10^8$  yr. For young stars the age correction is minor. Population II stars  $7 \times 10^9$  yr old would appear to be  $(15-20) \times 10^9$  yr old. Evolutionary ages for the oldest stars in our galaxy fall in the range  $(9-20) \times 10^9$  yr, with 33% helium being required for the low apparent age and 0% leading to  $20 \times 10^9$  yr. (See Iben and Faulkner, 1968).

An initial helium content in the sun of  $\sim 21\%$  provides part of the explanation for the low flux of solar neutrinos. But less initial helium would be expected in the older population II stars. The evolutionary age of these stars would be expected to be in excess of  $12 \times 10^{10}$  yr.

Under the scalar-tensor theory, with a hot cosmology in a closed space, no helium would have been formed in the primeval fire-ball (Dicke, 1968). Extreme population II stars should then be very low in helium. A closed space is favoured by the curvature in the magnitude - red shift relation. A closed space would also be consistent with the average mass density of the universe, assuming that there exists enough non-galactic matter in clusters of galaxies to gravitationally bind them and that the same ratio of galactic to non-galactic matter exists throughout the universe.

While the evidence for a closed space is not strong, if space is closed under the scalar-tensor theory (with  $\omega = 5$ ), the oldest stars in the galaxy should have been formed with little helium and their apparent evolutionary ages should be  $\sim 20 \times 10^9$  yr, their real ages being  $(7-10) \times 10^9$  yr.

Assuming a closed space under general relativity, using Peebles' (1966) calculation (see also Waggoner et al., 1967) of helium formation in the primeval fire-ball, the oldest stars should have an initial helium content of  $\sim 27\%$ .

If space is open, the matter in the universe being only that seen in galaxies, there is little difference between general relativity and the scalar-tensor theory, in expectations regarding stellar ages and helium abundance.

From Sandage's recent work on the distance of the Virgo cluster of galaxies, as reported here by Mrs. Burbidge, the Hubble age of the universe is  $13.7 \pm 3.4 \times 10^9$  yr (Sandage, 1968). The corresponding range of ages for an Einstein-de Sitter space is  $(7.3 - 11.4) \times 10^9$  yr. For closed spaces the age is less and it has the above limits when the curvature is sufficiently small.

Apparently the ages of the oldest stars are reasonably consistent with the age of the universe under either general relativity or the scalar-tensor theory, but the explanations are quite different, involving different initial helium content in the oldest stars.

As a final method for establishing an age of our galaxy I shall discuss the technique of dating by the radioactive decay of uranium (Burbidge et al., 1957). The method is based on the decay of  $^{235}\text{U}$  and  $^{238}\text{U}$ . The present abundance ratio  $[^{235}\text{U}/^{238}\text{U}]$  and the theoretical ratio at the time of formation determine the age of the galaxy. Burbidge et al. (1957) obtained  $6.6 \times 10^9$  yr and  $(11.5 - 18) \times 10^9$  yr for "prompt" and "continuous" formation, respectively.

Fowler and Hoyle (1960), concluding that "prompt" formation was incompatible with the present abundance ratio of  $[^{232}\text{Th}/^{238}\text{U}]$  in the solar system, investigated with care a model based on continuous production. To introduce the abundance ratios of chemically different materials, such as thorium and uranium, causes difficulties, particularly when the mean lives of  $^{238}\text{U}$  and  $^{232}\text{Th}$  are so great. Fowler and Hoyle deduced from an analysis of meteorites a present abundance ratio  $[^{232}\text{Th}/^{238}\text{U}] = 3.8 \pm 0.3$ . For "prompt" production I find that the ratio should be  $3.3 \pm 0.3$ . I would conclude that the two ratios agree sufficiently well. The chondritic meteorites of various types vary in this abundance ratio over the range 3.1 to 5.0. A very large number of recent carbonaceous meteorites of types I and II have given mean values of  $3.2 \pm 0.7$  and  $3.5 \pm 0.5$ , respectively.

Not being persuaded by the "discordance" between the thorium-uranium and uranium-uranium age determinations and not being convinced that the inclusion of thorium would improve the significance of the age determinations, I based a calculation of galactic age on only the  $[^{235}\text{U}/^{238}\text{U}]$  ratio, but on more realistic assumptions than the prompt synthesis of the elements (Dicke, 1962). The following assumptions were made (here modified slightly to accommodate new knowledge):

- (1) The uranium incorporated in the solar system was representative of the interstellar medium at this galactic radius at the time of formation.
- (2) The theoretical value of the abundance ratio at formation,  $[^{235}\text{U}/^{238}\text{U}]$ , falls in the range of the calculations of Fowler and Hoyle and Cameron, i.e. 1.3 - 1.8.
- (3) There exists no specific delay in the formation of r-process elements; these elements are made along with other heavy elements.
- (4) At least 60% of the present heavy element content of the interstellar medium was formed promptly during the collapse of the galaxy, in a time of about one hundred million years.
- (5) The remainder was formed afterwards at a uniform rate (per unit mass of the interstellar medium).

A detailed discussion of these assumptions is out of place here. Observational support for them was present in 1961 and it has increased, in some cases dramatically. While it certainly cannot be claimed that these

assumptions are above question, if they are correct, the present observed abundance ratio

$$[{}^{235}\text{U} / {}^{238}\text{U}] = 0.00723$$

and the mean lives of  ${}^{235}\text{U}$  and  ${}^{238}\text{U}$  permit a determination of the age of the galaxy. This age is  $6.7 \pm 0.9 \times 10^9$  yr, essentially the same as found for prompt synthesis by Burbidge et al. in 1957. It should be noted that this value is in reasonable agreement with Sandage's new Hubble age if the universe is closed.

Figure 1 shows the dependence of the age of the galaxy, and the present ratio of abundances  $[{}^{232}\text{Th} / {}^{238}\text{U}]$  expected in the solar system, for different values of the initial abundance formation ratio  $[{}^{235}\text{U} / {}^{238}\text{U}]$ ,  $\Delta T$ , the age of the solar system, and  $p$ , the fraction of the present heavy element content of the interstellar medium formed promptly (during the collapse of the proto-galaxy).

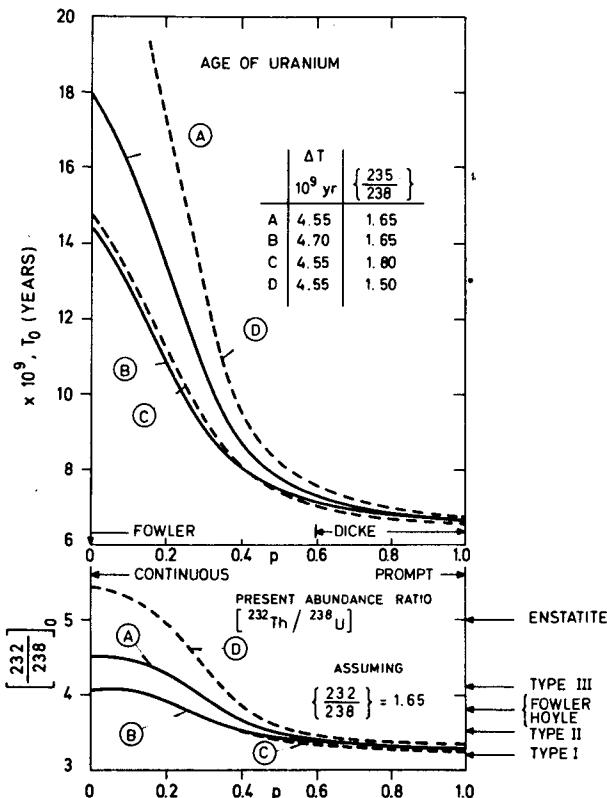


FIG. 1. The age of our galaxy  $T_0$  and the calculated present abundance ratio (in the solar system),  $[{}^{232}\text{Th} / {}^{238}\text{U}]$ , as a function of  $p$ , the fraction of the present heavy element content of the interstellar medium formed promptly.  $\Delta T$  is the time of formation of the solar system and  $\left\{ \frac{235}{238} \right\}$  is the (theoretical) abundance ratio  $[{}^{235}\text{U} / {}^{238}\text{U}]$  at the time of formation. The value ( $p = 0$ ) marked "Fowler" refers to an article by W. A. Fowler (Proc. Rutherford Jubilee Int. Conf., Manchester (BIRKS, J. B., Ed.), Heywood Co. Ltd., London (1961)). This analysis differs in several details from that of Fowler and Hoyle (1960).

## R E F E R E N C E S

- BURBIDGE, E.M., BURBIDGE, G.R., FOWLER, W.A., HOYLE, F. (1957) Rev. mod. Phys. 29, 547.
- DICKE, R.H. (1962) Nature 194, 329.
- DICKE, R.H. (1966) in Stellar Evolution(STEIN, R.F. , CAMERON, A.G.W., Eds), Plenum Press, N.Y.
- DICKE, R.H. (1968) Astrophys. J. 152, 1.
- FOWLER, W.A., HOYLE, F. (1960) Ann. Phys. 10, 280.
- IBEN, Icko, Jr., FAULKNER, J. (1968) Astrophys. J. 153, 101.
- PEEBLES, P.J.E. (1966) Astrophys. J. 146, 542.
- SANDAGE, A. (1968) Astrophys. J. 152, 149.
- WAGONER, R.V., FOWLER, W.A., HOYLE, F. (1967) Astrophys. J. 148, 3.

# REMARKS ON THE GENERAL PRINCIPLES OF EINSTEIN'S GRAVITATION THEORY

V. FOCK

Physical Institute of the University of Leningrad,  
Leningrad, USSR

## Abstract

REMARKS ON THE GENERAL PRINCIPLES OF EINSTEIN'S GRAVITATION THEORY. The author discusses some of the ambiguities that arise, and comments on some of the terms.

When the general principles of Einstein's gravitation theory are discussed, confusion very often arises from the lack of definition of some important notions like the notion of relativity or of the relativity principle or of the equivalence principle and of other notions.

To make a notion unambiguous one has to specify the conditions under which it applies. Einstein's theory is a theory of space and time. Therefore one must first of all specify the space region to which the problem refers. This region can never be the complete physical universe: what a human mind can study is always only some limited part of it. The term "universe" so often used in cosmological considerations is merely a technical term introduced to specify some kind of boundary conditions (Friedmann's universe, de Sitter's universe, etc.). The boundary conditions are expressed in the form of a hypothesis concerning the asymptotic behaviour of the gravitational potentials.

The simplest case of a universe (in the technical sense) is a system of massive bodies, like the solar system, immersed in an infinite space which is Euclidean at infinity and such that the four-dimensional metric at infinity is pseudo-Euclidean. In this case the space-time is uniform at infinity and a physical principle of relativity holds which is essentially the same as the principle of relativity of Galilei. In a suitable system of co-ordinates this principle is expressible with the help of a Lorentz transformation. It holds in spite of the fact that space-time is non-uniform near the masses but only uniform at infinity.

The suitable system of co-ordinates mentioned above are the harmonic co-ordinates. The fact that the principle of relativity is expressible in terms of a linear transformation of the harmonic co-ordinates makes their position exclusive.

To be correctly understood I must explain in a more exact way what I mean by physical principle of relativity. I use this term in the sense defined by Galilei who considered identical physical processes in two ships in uniform relative motion (or in a ship first lying at the shore and then being in uniform motion with respect to the shore).

In mathematical language we can say that two processes (or two phenomena) are identical if they are described by functions of the same mathematical form. In order that two processes in two frames of reference should be identical in the above sense, the mathematical transition from one

frame of reference to another must be accompanied by a corresponding adjustment of physical conditions. If and only if such an adaptation of the physical conditions to the transformation of co-ordinates is possible, one can say that the principle of relativity holds with respect to this transformation.

We see that the concept of physical relativity is connected with the mathematical notion of invariance rather than with that of covariance.

Physical relativity implies specific physical conditions (like uniformity of space-time at infinity) and is not possible in the general case of Riemannian space-time.

If one uses the word "relativity" in its physical meaning, then general relativity cannot exist. When Einstein speaks of "General Relativity" he changes the original (Galilean) meaning of the word "relativity". He means by it a mere covariance of equations which is a purely formal requirement and not an expression of any physical fact or physical law. (This was pointed out by Kretschmann as early as 1917.)

We tried to analyse the notion of relativity and came to the conclusion that physical relativity exists in a space-time which is pseudo-Euclidean at infinity. This is an idealization of the real space-time which can be applied to such objects as the solar system or perhaps a galaxy. In a certain sense this model of a space-time is also a local one.

Now let us consider Einstein's Principle of Equivalence. It is clear that this principle is purely local. The possibility of compensating the gravitational field by a field of acceleration exists only for a small region of space (inside a rocket, say). On the contrary, it is impossible to remove the gravitational field round the terrestrial globe. Thus the answer to the question whether the principle of equivalence is true, depends on the position of the problem. It is true locally (and is connected with the equality of inertial and gravitational mass) but cannot be considered as a general principle permitting the construction of some purely kinematical theory of gravitation (Einstein's theory is by no means a kinematical one).

The principles of relativity and of equivalence, in spite of their heuristic value and of the great role they played in Einstein's reasoning are, from the logical point of view, not the true foundations of Einstein's Gravitation Theory.

What are the true principles of this theory? Since we know its complete formulation, there is no difficulty in answering this question.

The first basic idea of the theory is the unification of space and time into a unique four-dimensional manifold with indefinite metric. This unification is connected with the law of propagation of any action transmitted with the limiting velocity (and consequently with the cause-effect relationship of events in space and time). A unique space-time manifold is considered in the usual (so-called "special") relativity as well, but in the gravitation theory a more general metric is introduced (the Riemannian metric).

The second basic idea of the theory is the assumption that physical processes can influence the metric, and the establishment, on that basis, of the unity between metric and gravitation. Formally, this unity manifests itself in the fact that the components of the metric tensor are at the same time gravitation potentials.

The idea that space is not necessarily rigid was for the first time expressed by Riemann but it was Einstein who built a physical theory on that basis.

Einstein himself understood perfectly well the importance of the two ideas stated above, but at the same time he never ceased to consider his theory as an expression of what he called the "general relativity principle" and of the equivalence principle. He always attributed a physical meaning to the general covariance of equations and never agreed with the fact that the general principle of relativity is either empty (if relativity is understood as mere covariance) or untrue (if relativity is understood as physical relativity).

I would like to conclude by stressing my great admiration for Einstein's contribution to physics and especially for his wonderful gravitation theory. This admiration is not incompatible with the observation that Einstein also made mistakes when trying something new.



# GENERAL RELATIVITY: SURVEY AND EXPERIMENTAL TESTS

R.H. DICKE

Palmer Physical Laboratory,  
Princeton, N.J., United States of America

## Abstract

GENERAL RELATIVITY: SURVEY AND EXPERIMENTAL TESTS. 1. Field theories of gravitation; 2. The null experiments; 3. The famous three tests; 4. The solar gravitational quadrupole moment.

Before discussing the structure of gravitational theory, a few words on Principles and philosophy are needed. Principles spelled with a capital P have played an important role in the development of General Relativity. In his autobiographical notes Einstein (1949) tells us how he was influenced in his thinking by philosophical matters and how his considerations of Mach's (1877) ideas concerning inertia guided him to General Relativity. He elevated these somewhat unsharp ideas to the status of a Principle, calling them "Mach's Principle". While Mach's Principle now may not seem very important, Einstein's own "Equivalence Principle" has continued to be a central pillar on which our ideas rest.

Perhaps the lack of observations drives physicists to philosophy. Certainly no other central part of physics has had less direct support from observations than General Relativity.

The "Equivalence Principle" has been taken to mean two different things by various workers. The differences may seem to be slight but they are important, sufficiently important to warrant making a careful distinction between them. The difference in meaning has been brought into focus in two different statements, a "weak" and a "strong" form of the Principle (Dicke, 1962a).

The "weak" form states that the gravitational acceleration of all small, localized bodies is the same, independent of composition. The restriction to small bodies is to avoid "tidal" accelerations due to gravitational gradients acting on the quadrupole moments of extended bodies.

The "strong" form states that the laws of physics observed in a small, freely-falling, non-rotating laboratory have a standard form and a standard numerical content, independent of the location in space and time. In particular, the form and numerical content of these laws is that of gravity-free space. By "numerical content" one means the total of dimensionless numbers such as particle mass ratios and coupling constants.

It is evident that only the "weak" form of the law is supported directly by the Eötvös experiment (Eötvös, 1922, Roll et al., 1964). Parts of the "strong" principle draw indirect observational support from the experiment, but other parts are without visible support.

## 1. FIELD THEORIES OF GRAVITATION

The principal local field theories of gravitation are indicated in Fig.1 with connections meant to show their interrelations.

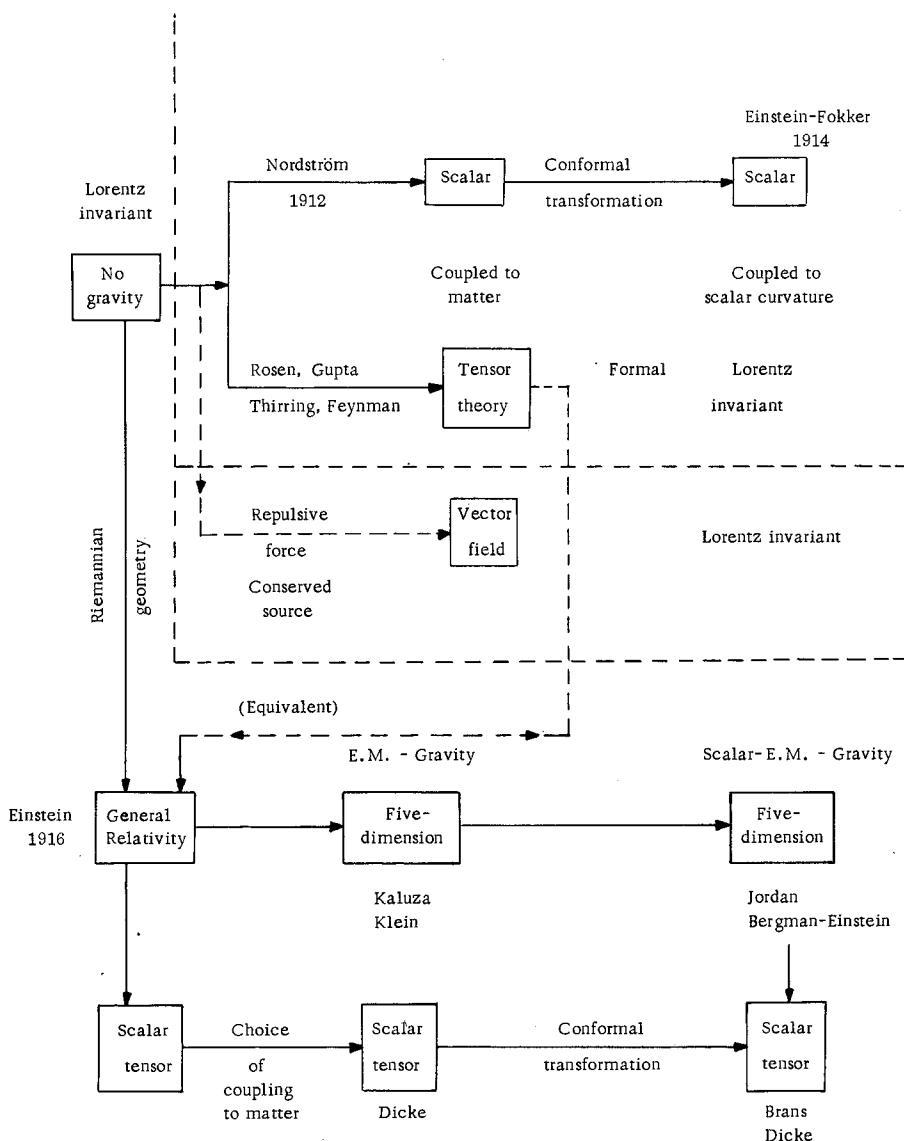


FIG. 1. Diagrammatic representation of the interrelations between local field theories of gravitation.

At the upper left there is a box meant to represent the body of Lorentz-invariant field theory associated with gravity-free space. The first formally Lorentz-invariant relativistic theory of gravitation to appear was the scalar theory of Nordström (1912), derived in a straightforward way from special relativity and based on a universal coupling of a scalar field to matter. The theory is only "formally Lorentz-invariant" because the scalar coupling to particle mass introduces a variation in the mass.

The masses of elementary particles become functions of the scalar and the sizes of "rods" and the periods of "clocks" become functions of the scalar. Thus the Minkowski metric tensor of the Lorentz-invariant geometry is not measured by real rods and clocks. These measure the metric of a curved but conformably flat space. But this should not be taken to mean that the Minkowski metric is completely non-physical and beyond measurement. Units of length and time can be defined operationally as  $(G\hbar/c^3)^{\frac{1}{2}}$  and  $(G\hbar/c^5)^{\frac{1}{2}}$ , respectively. These units do not involve elementary particles, and the geometry measured is that of flat space (Dicke, 1965, 1967).

A conformal transformation of the Nordström theory can be used to transfer the dependence of mass on the scalar to the locally measured gravitational constant. This transformation which is geometrically conformal also transforms mass. The conformal transformation is conveniently interpreted as a scaling of the units of length, time, and reciprocal mass by a common co-ordinate dependent factor. The Einstein-Fokker (1914) theory is a conformably flat theory of this type.

A final example of a formally Lorentz-invariant theory of gravitation is the tensor theory as developed by Rosen (1940), Gupta (1957), Feynman (1957) and Thirring (1961). This is a tensor theory of gravitation which employs the standard machinery of Lorentz-invariant field theory. The Lorentz invariance of this theory is completely formal, for the Minkowski metric is not subject to measurement by any naturally defined units of measure. This theory can be transformed into Einstein's (1916) General Relativity and is physically equivalent to it.

Einstein's (1916) generally covariant theory of gravitation, General Relativity, represents a generalization of the geometry of his special relativity from a four-dimensional flat space to a four-dimensional Riemannian space. The metric tensor of this Riemannian space is treated as a dynamic variable, and the curvature tensor is connected to the matter content expressed in terms of the stress-energy tensor of matter. The Einstein field equations are obtained from the variational principle

$$\delta \int (R + GL) \sqrt{-g} d^4x = 0 \quad (1)$$

where  $R$  is the scalar curvature and  $L$  is the Lagrangian density of matter. Varying Eq.(1) with respect to  $g_{ij}$  gives Einstein's field equations. Varying with respect to the other variables in  $L$  gives all the remaining equations of motion of physics.

A further generalization of the geometry was carried out by Kaluza (1921) and Klein (1926). They found that a purely formal extension of the four-dimensional space-time by one dimension to a five-dimensional space permitted the incorporation of the laws of charge-free electromagnetism into the geometry of space. The so-called Einstein-Maxwell field equations appear in the formalism as the equations of motion of the metric components of the five-dimensional space. The 15 elements of the five-dimensional metric tensor contain the ten elements of the metric tensor of a four-dimensional projected space, the four elements of the electromagnetic vector potential, and a single constant.

Jordan (1948, 1955) extended the Kaluza-Klein theory by freeing the constraint on the 15th component of the metric tensor, permitting it to be

a new dynamic variable, a scalar field variable. This theory is equivalent to a dynamic system coupling the scalar field and the electromagnetic field to the curvature of the four-dimensional projected subspace. Jordan introduced the theory to provide a formal basis for Dirac's (1938) cosmology.

The scalar-tensor theory (Brans and Dicke, 1961) is closely related to Jordan's theory, though it was independently developed. This theory cannot be put in a five-dimensional form, but the field equations are similar to the four-dimensional equations contained in a special case of Jordan's theory. To show the similarity, one of Jordan's parameters must be fixed, and the "matter Lagrangian" must be extended to include non-electromagnetic effects. The scalar-tensor was developed to avoid difficulties vis-à-vis Mach's Principle that we saw in General Relativity. It is one of the paradoxes of physics that the ideas of Mach which led Einstein to General Relativity seemed to be imperfectly incorporated in the theory.

Another possible path to the scalar-tensor theory starts with ordinary General Relativity but admits a long-range scalar field as part of the ordinary field content of space. Such a scalar field (like electromagnetism) affects the curvature of space through its stress-energy tensor. The gravitational phenomenon is partially geometrical, through Einstein's field equation, and partially the effect of the scalar force-field interacting with matter. The gravitational effect has two separate origins, but the weak equivalence principle is satisfied. The reason that all small bodies fall with the same acceleration is traced to the equivalence for a localized, isolated and stationary body of the integral of the time-averaged contracted stress-energy over the body to the total energy of the body (Dicke, 1964a).

As in the case of Nordström's theory, the metric with a scalar force field acting is only formal, for the scalar interaction modifies ordinary rods and clocks. But as with Nordström's theory, this metric is measured by rods and clocks based on gravitational units  $(G\hbar/c^3)^{\frac{1}{2}}$  and  $(G\hbar/c^5)^{\frac{1}{2}}$ .

By choosing the appropriate form of coupling to matter, a special case is obtained (Dicke, 1962b) for which a conformal transformation leads directly to the physically equivalent Brans-Dicke form of the scalar-tensor theory. Thus there are two physically equivalent forms of the scalar-tensor theory. For the first form (Brans and Dicke, 1961) there is no direct interaction of the scalar field with matter and a modification of Einstein's field equations is introduced. In the second form (Dicke, 1962b), Einstein's field equations are satisfied, but the scalar field couples to matter as a force field.

The scalar-tensor theory is based on two gravitational fields, but there is only a single coupling constant,  $\omega$ , a dimensionless number of the order of magnitude of 5, the coupling constant for the scalar field. The scalar field itself sets the magnitude of the tensor coupling. In General Relativity the coupling constant  $Gm_p^2/\hbar c^4 \sim 10^{-40}$  is given by nature and is unrelated to other dimensionless numbers. The equivalent "constant" under the scalar-tensor theory is

$$\frac{(4 + 2\omega)m_p^2}{(3 + 2\omega)\hbar c\varphi}$$

It is found by solving a field equation. The value of this "constant" is so small because the mass of the seen part of the universe is so large relative to the mass of a proton. The scalar is presently (Brans and Dicke, 1961)

$$\varphi = \frac{8\pi}{2\omega + 3} \langle \rho t \rangle T_0$$

where  $\rho$  refers to mass density and  $t$  is time.  $T_0$  is the age of the universe, and the average of the product of matter density and time is to be taken over the history of the universe. For an Einstein-de Sitter universe

$$\varphi = \frac{4\pi(4+3\omega)}{3+2\omega} \rho_0 T_0^2$$

and the ordinary gravitational coupling "constant" becomes

$$\frac{(4+2\omega)m_p^2}{(3+2\omega)\hbar c \varphi} = \frac{(4+2\omega)}{4\pi(4+3\omega)} \left( \frac{m_p}{\rho_0 T_0^3 c^3} \right) \left( \frac{T_0 c}{\hbar/m_p c} \right)$$

The ratio of the radius of the universe  $T_0 c$  to the Compton wavelength of the proton is a large number ( $\sim 10^{40}$ ) but the second term, the ratio of the proton mass to the "mass of the universe" ( $\sim 10^{-80}$ ) is very small.

The formal differences between General Relativity and the scalar-tensor theory are seen in the two variational equations

$$\begin{aligned} \delta \int (R + GL) \sqrt{-g} d^4x &= 0 \\ \delta \int \left( \varphi R - \omega \frac{\varphi_{,i}\varphi^{,i}}{\varphi} + L \right) \sqrt{-g} d^4x &= 0 \end{aligned} \tag{2}$$

It is noticed that the gravitational constant  $G$  has disappeared and is replaced by  $\varphi^{-1}$ , carrying the same dimensions. The matter Lagrangian  $L$  is the same for both equations.

The results expected from observations under the scalar-tensor theory differ from expectations under ordinary General Relativity in several ways. The Schwarzschild solution differs somewhat, leading to a relativistic perihelion rotation ( $1 - (4/3)s$ ) of the Einstein value and a light deflection ( $1 - s$ ) of the Einstein value. Here  $s = 1/(4+2\omega)$  is the fraction of the gravitational acceleration due to the scalar force field (under the Dicke 1962b formulation). Present observations indicate that if the scalar interaction exists,  $\omega \sim 5$  and  $s \sim 0.07$ .

The other difference which may be presently susceptible to observation is the slow weakening of gravitation expected under the scalar-tensor theory. This is due to the slow change in  $\varphi$  induced by the change of the gravitational coupling constant  $-(\psi/\varphi)_0$ . The steady but slow weakening of gravitation with time carries a number of implications for astrophysics and geophysics which may some day be capable of yielding a sound basis for accepting or rejecting the theory.

## 2. THE NULL EXPERIMENTS

Owing to the small size of a laboratory relative to astronomical bodies and the short life span of a physicist relative to the age of the universe, the co-ordinate patch occupied by a typical laboratory experiment is too small to permit the exhibition of space-curvature effects. Thus significant laboratory experiments tend to be null experiments. While the laboratory experiment on the gravitational red shift was not a null experiment (Pound and Snider, 1965), it did not exhibit a space-curvature effect.

Of the various null experiments probably the most important is the Eötvös experiments on the compositional independence of the gravitational acceleration (Eötvös et al., 1922, Roll et al., 1964). It has been shown with a precision of  $1 \times 10^{-11}$  that gold and aluminium fall towards the sun with the same acceleration.

Figure 2 shows a diagram of the torsion balance used for the Princeton experiment. The quartz triangle carried an optical flat polished on one edge. This was used automatically to monitor rotation of the torsion balance in a high vacuum, the apparatus being held at constant temperature in a deep instrument well. The approximate three-fold symmetry axis of the torsion balance was to eliminate a gravitational quadrupole moment which would cause the balance to respond to gradients in the gravitational field.

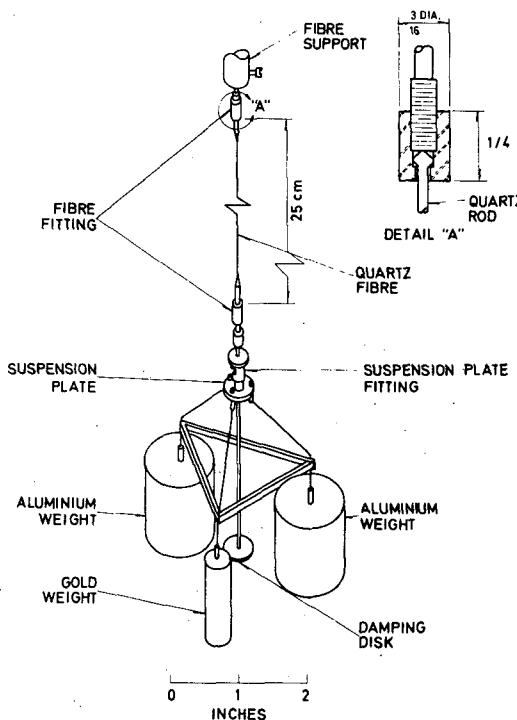


FIG. 2. The Princeton version of the Eötvös experiment, torsion balance.

The balance was part of an optical-electric-electronic feed-back loop that damped the free oscillations and shortened the pendulum period. At 6.00 a.m. and 6.00 p.m. the weights were in positions relative to the sun for which an anomalous gravitational acceleration of gold relative to aluminium would cause a rotation of the balance (in the absence of the feed-back loop). After averaging data for weeks at a time (leaving the balance undisturbed), it was found that the diurnal variation in the torque acting on the torsional balance was negligibly small, corresponding to a gravitational anomaly of  $10^{-11}$  or less.

While this experiment represents a direct test of only the "weak principle of equivalence", it also provides an indirect test for much of the "strong principle" (Wapstra and Nijgh, 1955).

To illustrate how this comes about, consider the question of a possible variability of the fine structure constant, i.e. part of the numerical content of physical laws. This question was raised by Wilkinson (1958) and was considered by Peebles and me (see Dicke, 1960; Peebles and Dicke, 1962a, 1962b; Peebles, 1962). The question has been raised again by Gamow (1967) without his being aware of the earlier work. Arguments based on radioactive decay have been used by Peebles and Dicke (1962b), Dyson (1967), Peres (1967), Gold (1968), and Chitre and Pal (1968) to show that little variation of  $\alpha$  with time could have occurred.

The argument based on the Wapstra and Nijgh (1955) paper would proceed along the following lines: The electrostatic energy associated with the nucleus varies as  $Z(Z - 1) A^{-1/3} \alpha$ . A variation of  $\alpha = e^2/\hbar c$  with time associated with the coupling of the electromagnetic Lagrangian to a scalar field would imply also a variation with position, in particular a variation with height. If an atom were lifted and  $\alpha$  were to vary, work would be required to change the internal energy and this part of the gravitational work would be proportional to  $Z(Z - 1) A^{-1/3}$ . An anomalous gravitational acceleration would result. The observed constancy of the gravitational acceleration sets severe limits to such a variation of  $\alpha$  with position.

The remote possibility that this argument would be invalidated by the addition of a compensatory coupling to a new tensor field was considered (Dicke, 1960, 1962a). Peebles (1962), Peebles and Dicke (1962a) showed that this was incompatible with another null experiment, that of Hughes et al. (1960) and Drever (1961).

The experiments of Hughes et al. and Drever are null experiments on spatial isotropy. The coupling of matter to a long-range tensor field (additional to the usual metric tensor) would give space anisotropic properties. A co-ordinate transformation generally can be used to reduce a tensor (normally only one tensor) locally to the isotropic Minkowski form. Thus a second long-range tensor would introduce anisotropy into physical laws.

The Hughes-Drever experiment is one of the two most precise null experiments known and it sets extremely stringent limits on a coupling of matter to a second tensor field. As a result, the above argument involving the Eötvös experiment appears to be valid, and very little variation of the fine structure constant with position is permitted. The Eötvös experiment appears to set a much more stringent limit to the coupling of a scalar field to

the electromagnetic Lagrangian (and the resulting variation of  $\alpha$  with position) than does argumentation based on the history of  $\alpha$  and  $\beta$  decay (Peebles and Dicke, 1962b; Dyson, 1967).

The Hughes-Drever experiment involved observations of the separation in very weak magnetic fields of the magnetic sub-states of nuclei having ground-state spins greater than  $\frac{1}{2}$ . The anisotropy present as a result of coupling to another long-range tensor field would be expected to lead to inequalities in the separation of the magnetic sub-states. (See Dicke, 1964a, for an interpretation.)

While there are many other null experiments which could be discussed, the Eötvös experiment and the Hughes-Drever experiment appear to be the most significant for gravitational theory.

### 3. THE FAMOUS THREE TESTS

The three well-known positive effects associated with General Relativity are the gravitational red shift, the relativistic part of the perihelion rotation of the orbit of Mercury, and the gravitational deflection of light. It is a measure of Einstein's genius that his first publications suggested all three tests and that in the intervening half century no other distinctly different effect has been observed. In making this statement I am mindful of the remarkable results reported recently by Shapiro et al. (1968) on the retardation of radar waves bounced from Venus and Mercury and passing close to the sun. While from the standpoint of technique this observation is far removed from the classical observations of the deflection of light, they are closely related. In optics, the close connection between the deflection of a light wave by a wave-retarding medium and wave retardation are well known. The connection is essentially a wave propagation phenomenon and it holds also for the vacuum surrounding the sun.

Probably, the three famous tests should be reduced to two, for the gravitational red shift experiment (Pound and Snider, 1965) does not exhibit any space curvature effect and its results are easily derived by assuming only the weak form of the equivalence principle and the assumption of mass-energy equivalence.

The gravitational deflection of light was an effect predicted by Einstein and later discovered. Unfortunately the unknown systematic errors associated with the observations of star positions during an eclipse of the sun cast doubt on the precision of the results. The suspicion that there may be serious errors in these results is reinforced by examination of the discordance in the data from different eclipses. (See Bertotti et al., 1962, for a modern discussion.)

The recent related data from planetary radar (Shapiro et al., 1968) already represent an improvement over the classical data. Also, a new type of instrument, designed by H. Hill, is capable of measuring star positions near the sun without the necessity for an eclipse; ultimately, it should be capable of yielding star positions of considerable precision.

The results of Shapiro et al. (1968) are equivalent to a light deflection  $0.9 \pm 0.2$  times Einstein's predictions. The error is twice the standard deviation of the data. Under the scalar-tensor theory I would expect

a factor of  $\sim 0.93$ . Thus, these new data are in good agreement with both General Relativity and the scalar-tensor theory.

The final relativistic effect for which observations exist is the relativistic rotation of the perihelion of a planet's orbit. This effect was known as an observed anomaly in Mercury's motion from the middle of the 19th century. Many attempts were made to account for the anomaly along classical lines, including the introduction of a perturbation by an unobserved planet Vulcan and the effect of a not-yet-observed oblateness of the sun.

The planet Vulcan has never been found, but we now believe that the sun has a quadrupole moment (Dicke, 1964b) associated with an oblateness which we have observed (Dicke and Goldenberg, 1967a). I shall return to this observation and its significance, but first I shall summarize what is known about the observed excess motion of the line of apsides of the various planets and one asteroid.

Clemence's (1943) classic discussion of the motion of Mercury provides most of the basis for our knowledge of the relativistic motion of planetary orbits. There are three primary sources of error in determining the relativistic motion of perihelia: (1) the errors in the observation, (2) errors in adopted masses of planets, and (3) error in the adopted value of the general precession.

In 1943 there was considerable doubt about the mass of Venus. This doubt is now dispelled, as a result of the program in space research. Duncombe (1958) has used improved values of planetary masses and a correction of  $0.^{\circ}79$  to Newcomb's value of the general precession to obtain excess centennial motions of  $43.^{\circ}0 \pm 0.44$  and  $5.^{\circ}01 \pm 1.79$  for Mercury and the Earth respectively. These are to be compared with the values  $43.^{\circ}03$  and  $3.^{\circ}84$ , computed under General Relativity.

More recently, Wayman (1966) has recomputed these excess motions using both Duncombe's (1958) adopted planetary masses and the newly determined masses of Marsden, in part based on planetary radar measurements. These results are given in Table I. They are based on Duncombe's correction to Newcomb's value for the general precession ( $\Delta p = 0.^{\circ}79/\text{century}$ ) and on the new value ( $\Delta p = 1.^{\circ}63 \pm 0.22/\text{century}$ ) obtained by Wayman from a study of the proper motions of 34 distant OB stars. Also included in the 6th and 7th columns of Table I are corrections for the observed solar oblateness,  $\Delta r/r = (5.0 + 0.7) \times 10^{-5}$  (Dicke and Goldenberg, 1967a). This correction is  $-3.^{\circ}4/\text{century}$  and  $-0.^{\circ}1/\text{century}$  for the "observed" rotation of the perihelion of Mercury and the Earth respectively. The last two columns of the table contain the theoretical value under General Relativity and scalar-tensor theory.

From Table I it is evident that the interpretation of the observations depends critically upon the solar oblateness, whether this correction should be included or not. Except for this uncertainty, the "observed" relativistic rotation of Mercury's perihelion appears to have a precision of the order of 2%. This precision seems to be adequate to rule strongly in favour of, or against, General Relativity, depending upon the disposition made of the solar oblateness data.

Shapiro et al. (1968b) have analysed the few available optical observations of the asteroid Icarus. While this asteroid has been known less than two decades, it is so favourably situated for the observation of the rela-

TABLE I. THE "OBSERVED" RELATIVISTIC ROTATION OF THE PERIHELION OF MERCURY AND THE EARTH AS COMPUTED BY WAYMAN (1966)

Results are based on Duncombe's (1958) and Marsden's (1965) planetary masses and two values for the correction  $\Delta p$  to Newcomb's general precession

Solar oblateness correction	No	No	No	No	Yes	Yes	General Relativity	Scalar-tensor $\omega = 5$
$\Delta p$	0°79	0°79	1°63	1°63	1°63	0°79		
Masses	D	M	D	M	M	D		
Mercury	43°1	43°2	42°3	42°3	38°9	39°7	43°03	38°93
Earth	5°3	2°3	4°5	1°4	1°3	5°2	3°84	3°47

D = Duncombe's planetary masses and M = Marsden's planetary masses.

tivistic perihelion rotation that a meaningful value has already been obtained. Shapiro et al. fit the observations to an equation containing the relativistic rotation of the perihelion (computed under General Relativity) multiplied by a parameter  $\lambda$ . The least-square fit gives a value for  $\lambda$  of  $0.97 \pm 0.20$ . Under the scalar-tensor theory, with the correction due to the solar quadrupole moment included,  $\lambda$  is expected to have the value of  $\sim 1.04$ .

#### 4. THE SOLAR GRAVITATIONAL QUADRUPOLE MOMENT

Little will be said here about the actual measurements of the oblateness of the sun (Dicke and Goldenberg, 1967a). A proper discussion would take far too long. The observations are usually not questioned.

The optical system of the instrument used for the observations is shown schematically in Fig.3. The instrument determines photoelectrically the oblateness of the solar image defined by the limb of the sun. The edge (or limb) is remarkably sharp, the scale height for change in brightness being only a few tens of kilometres. The location of the limb is determined almost completely by a surface of constant density. The position of the limb is not observed directly but rather is inferred from the integrated light flux from the outer  $6°5 - 19°1$  arc of the photosphere passed by an occulting disk placed over the sun's image. This technique frees the measurement from most of the systematic errors that could be introduced by the atmosphere. It also provides a measure of the distribution of brightness about the limb, a factor important to the interpretation.

The oblateness observed in the summer of 1966 was  $\Delta r/r = (5 \pm 0.7) \times 10^{-5}$  and the sun was remarkably uniformly bright about the limb. In the mean the brightness difference between the pole and the equator could not be reliably determined. Expressed as the temperature of an equivalent black body, the average temperature difference between the pole and the equator was under 3 deg.

While most of the observations were made at the limb, a substantial number of brightness distribution measurements were made much further in. The general conclusion is that (in 1966) there was no noticeable latitude dependence in the brightness of the solar disk.

Owing to bad weather and perhaps also to a more active sun, the 1967 data scattered more than those of 1966. However, the observations extend over a longer time period and the data are almost as meaningful as those of 1966. The over-all results are consistent with the conclusions based on 1966 data.

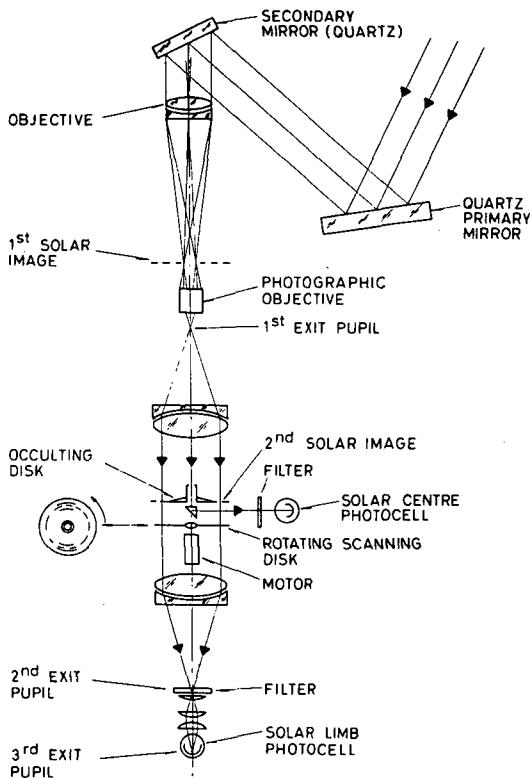


FIG. 3. Optical system of the instrument used to measure the solar oblateness.

In my opinion the most compelling reason for believing the observations is provided by the measurement of the orientation of the minor axis of the elliptical image of the sun. In 1966 this minor axis coincided with the known rotation axis to within  $2^\circ$  throughout the summer. In this period of time the orientation of the sun's axis changed by  $40^\circ$ , the effect of changing aspect as the earth moved about the sun.

I can find nothing wrong with the observations, and I conclude that, in 1966 and 1967, the sun had an oblateness of  $\Delta r/r \sim 5 \times 10^{-5}$ . There was no substantial latitude dependence in the brightness of the solar disk in either 1966 or 1967.

The interpretation of part of the observed solar oblateness as the effect of a solar mass quadrupole moment is based on the following theorem (von Zeipel): A homogeneous static fluid subject only to gravitational and pressure stresses has common surfaces of constant density, pressure, temperature and gravitational potential. The proof is elementary. The equation of static equilibrium

$$\nabla P + \rho \nabla \varphi = 0 \quad (3)$$

implies that the normals to surfaces of constant  $P$  and  $\varphi$  coincide, implying that the surfaces themselves coincide. From the curl of the above equation one finds that surfaces of constant  $\rho$  and  $\varphi$  coincide. The theorem then follows from the assumption of homogeneous composition and the connection between  $P$ ,  $T$  and  $\rho$ .

The solar oblateness measurement gives the oblateness of an outer surface of constant density. If the above theorem were applicable, the shape of the external surface of constant gravitational potential would be known and hence the mass quadrupole moment directly could be computed.

The theorem is certainly not rigorously satisfied in the sun, but the above statement is also much stronger than needed. First it should be noted that the statement is needed only for the outer "seen layers" of the sun. Complex stress patterns deep in the interior of the sun do not affect the validity of conclusions based on observations of the surface. Second, the rotation of the sun violates the condition of static equilibrium. To the extent that the outer "seen layers" rotate rigidly, the rotation is easily included by adding to  $\varphi$  the centrifugal potential  $-\frac{1}{2} \omega^2 r^2 \sin^2 \theta$ . In a rotating co-ordinate system the fluid is static.

While the surface of the sun is not rotating uniformly, the contribution from surface rotation is found to be the same as would be computed assuming rigid rotation at a rate observed for the sun at a latitude of  $\sim 49^\circ$ .  $0.8 \times 10^{-5}$  of the observed oblateness of  $5.0 \times 10^{-5}$  is due to the surface rotation.

It is believed that the velocity stress of surface rotation is the only stress, additional to pressure and gravitation, that contributes appreciably to the oblateness of the sun.

I have developed a general theory of the effect of additional stresses on the oblateness of the sun. The added stresses are believed to be due only to magnetic and velocity fields. The form of the stress tensor assuming the presence of these added fields is indicated in the upper part of Table II. Assuming a stationary state (not static since the fluid is moving), I obtain the  $(2, 0)$  spherical harmonic expansion coefficients indicated at the bottom. The coefficient for  $\rho$  directly yields this contribution to the oblateness. A combination of the coefficients for  $\rho$  and  $P$  yields the latitude dependence of the brightness at the extreme solar limb.

Inasmuch as surface distributions of both magnetic and velocity fields are known for the sun, it is possible to place reasonable limits on their contributions to solar oblateness. In this connection, the explicit expression for the brightness variation induced at the extreme limb is very useful. Any distribution capable of distorting the surface will normally

TABLE II. SURFACE STRESSES GENERATED BY MAGNETIC AND VELOCITY FIELDS  
Effects on surfaces of constant density and pressure

Effect of surface stress on shape and polar brightening of sun  
(in spherical co-ordinates)

Stress tensor  $\varphi$ , gravitational potential

$$T_i^j = P \delta_i^j + \frac{1}{4\pi G} [\varphi_{,i} \varphi^{,j} - \frac{1}{2} \delta_i^j \varphi_{,k} \varphi^{,k}] + M_i^j$$

$$M_i^j = \rho v_i v^j - \frac{1}{4\pi} (B_i B^j - \frac{1}{2} \delta_i^j B_k B^k)$$

Stationary state

$$T_{i;j}^j = 0$$

If  $M_i^j = 0$ , surfaces of constant  $\varphi$ ,  $P$ ,  $\rho$ ,  $T$  coincide

$$\text{Let } M_i^j = R_i S^j + \dots \delta_i^j Q \quad S_{ij}^j = 0$$

$$P_s = Q + \frac{1}{3} R_i S^i$$

$$A = 2 R_1 S^1 - R_2 S^2 - R_3 S^3$$

$$B = R_2 S^2 - R_3 S^3$$

$$C = r^{-1} R_2 S^1$$

Spherical harmonic expansion coefficients  $[ ]_{\ell, m}$

$$[P]_{2,0} = -[P_s]_{2,0} + \frac{1}{6} [A - B]_{2,0} + \sqrt{\frac{5}{36}} [B]_{0,0}$$

$$+ \frac{1}{\sqrt{6}} \frac{1}{r^2} \frac{\partial}{\partial r} r^3 [C e^{-i\varphi}]_{2,1}$$

$$- g[\rho]_{2,0} = \frac{1}{2} \frac{1}{r^2} \frac{\partial}{\partial r} r^2 [A]_{2,0} - \frac{1}{6} \frac{\partial}{\partial r} [B]_{2,0} + \sqrt{\frac{5}{36}} \frac{\partial}{\partial r} [B]_{0,0}$$

$$+ \frac{1}{\sqrt{6}} \frac{1}{r} [8 + r^2 \frac{\partial^2}{\partial r^2}] [C e^{-i\varphi}]_{2,1}$$

also affect the brightness distribution. Only for very special sets of stress distributions (e.g. uniform surface rotation) can the shape of the sun be affected by these surface stresses without large companion changes being induced in surface brightness.

A number of interesting suggestions have been made in the past  $1\frac{1}{2}$  years for producing the solar oblateness using surface stresses. I have found that they all fail the brightness test, some miserably. (See Dicke and Goldenberg, 1967b.)

TABLE III. STRESS DISTRIBUTIONS OVER THE SURFACE FOR A NON-GRAVITATIONAL ACCOUNT OF THE SOLAR OBLATENESS  
The absence of a latitude dependence of the solar brightness has been used to provide a constraint on otherwise possible field distributions

Implications of uniformly bright sun

Surfaces of constant  $\rho$ , T, P coincide

$$\rho_{,i} + \rho \varphi_{,i} + \rho W_{,i} = 0$$

$$\rho W_{,i} = M_1^j_{,j} \quad M_1^j = \rho v_i v^j - \frac{1}{4\pi} (B_1 B^j - \frac{1}{2} S_1^j B_k B^k)$$

Spherical co-ordinates

$$ds^2 = g_{ij} dx^i dx^j = dr^2 + r^2 d\theta^2 + r^2 \sin^2 \theta d\phi^2$$

Two solutions of interest

$$I. \quad M_1^j = \rho W \delta_1^j - \delta_1^j \frac{1}{r^2} \int r^2 \frac{d\rho}{dr} W dr$$

Randomized fields

$$\text{Turbulence} \quad \langle v_\theta^2 - v_r^2 \rangle^{\frac{1}{2}} = \langle v_\phi^2 - v_r^2 \rangle^{\frac{1}{2}} \sim 3 \times 10^5 \text{ km/s}$$

Excess in polar region.

$$\text{Magnetic field (randomized)} \quad \langle B_r^2 \rangle = 2 \langle B_\theta^2 \rangle = 2 \langle B_\phi^2 \rangle$$

Excess in polar region

Dependence on depth

$$\langle B^2 \rangle^{\frac{1}{2}} \sim 200 \text{ G at } \lambda = 0.004$$

$$\sim 700 \text{ G at } \lambda = 0.5$$

$$II. \quad M_1^j = M_3^3 (r, \psi) = -2 \rho r^2 \sin^2 \psi \frac{d}{d(g_{33})} W$$

$$= -2 \rho g_{33} \frac{d}{d(g_{33})} W$$

Rotation on cylinders

$$M_3^3 = \rho r^2 [\omega(r \sin \psi)]^2$$

Of even more help in interpreting the observations is the knowledge that there is no substantial variation in the latitude of brightness over the whole photosphere (not just the limb). If von Zeipel's theorem were rigorously satisfied at the surface, a small variation in brightness (viewed normally) would be expected because of a non-uniform spacing of surfaces of constant temperature. The variation of layer spacing expected from

the solar oblateness is small and the corresponding variation in brightness, with the surface viewed normally, is equally small. These variations will be neglected in the following. The observations will be taken to mean that the intensity and angular distribution of radiation flux from a patch of the solar surface is the same everywhere (in the mean, i.e. without a systematic latitude dependence). But this implies that surfaces of constant density and temperature coincide. Because of the uniform composition, surfaces of constant pressure and density coincide. Then the added terms to the stress tensor have the form shown at the top of Table III, namely the covariant divergence of this part of the stress tensor is in the form of a gradient multiplied by the density.

While I have not proved that there are no more, I have found only two solutions to the equation

$$\rho W_{,i} = M_{i;j}^j \quad (4)$$

The first requires an anisotropic randomized magnetic or velocity field. The required velocity and magnetic field distributions are such that the mean square averages of the  $\theta$  and  $\varphi$  components are equal and the mean square vertical component is generally different. The connection between radial and transverse components is slightly complicated. To account for the solar oblateness with turbulence fields requires a polar excess of  $v_\theta^2 - v_r^2$  of  $\sim 10^{11}$  (km/s)<sup>2</sup>. For magnetic fields the required excess squared field strength in the polar region is approximately  $(200)^2 G^2$  in the upper photosphere and  $(700)^2 G^2$  in the lower. The exact differences depend upon details of the radial distribution.

Root mean square magnetic fields that strong can be eliminated on observational grounds (Howard, 1967). Information concerning velocity fields is obtained from the widths and shapes of spectroscopic lines, also from spectroscopic observations of Doppler shifts.

The Doppler shift measurements show anisotropic velocities of the order of 1 km/s, probably the result of disturbances propagated from the underlying convective region. Low velocities (transverse) are seen in large-scale convective motion in the photosphere (Leighton, 1963). It has been inferred from observations of spectroscopic line shapes and widths that small-scale turbulence velocities may also be anisotropic with velocities of the order of 3 km/s (Schmalberger, 1963).

While anisotropy is found in the velocity field, a dependence on latitude is missing. There seems to be no indication from either line widths or Doppler shifts of a dependence on latitude of the turbulence field.

The second type of solution of Eq.(5) represents a rotation of the sun with the angular velocity in the observed part of the sun as a function only of the distance from the axis of rotation. This solution was used to obtain the correction  $0.8 \times 10^{-5}$  to the oblateness of the sun from the observed latitude dependence of angular velocity.

To summarize, I believe that a solar oblateness of  $\Delta r/r \sim 5 \times 10^{-5}$  has been observed and that only  $0.8 \times 10^{-5}$  of this can be associated with extraneous effects. If this is the correct interpretation of the observations, the 6th and 7th columns of Table I may reasonably be taken to represent the present state of knowledge of the relativistic motion of the perihelion.

What of the future? Clearly, other independent tests of General Relativity are needed. In a sense other tests already exist, for the value  $\omega \sim 5$  for the coupling constant of the scalar-tensor theory was deduced prior to the solar oblateness measurement from an assessment of the significance of a number of astrophysical observations. (See Dicke, 1967, for a résumé and references.) But astrophysics has not yet progressed to a point where a single unambiguous interpretation of a complex set of observations always exists, and these "tests" of General Relativity are not nearly as clear and unambiguous as a typical laboratory experiment, or Shapiro's radar retardation experiment.

The interpretation of an observation of the gravitational deflection of light, or the radar retardation effect, is free of ambiguity resulting from the solar oblateness. Future observations might be improved by the factor of 5 needed for a really definitive test of General Relativity.

Other aspects of the program of planetary radar observations also could be important. Accurate observations of the perihelia motions of two planets permit the separate determination of a solar quadrupole moment and the relativistic effect. Observations over a sufficiently long period of time would permit a determination of the secular slowing of planetary motion associated with a weakening of gravitation.

Finally, the attempt by W.M. Fairbank and his colleagues to observe the Thirring-lens precession and the geodesic precession, using spinning tops in a satellite (Schiff, 1959), could provide a test of General Relativity distinctly different from those originally suggested by Einstein.

#### REFERENCES

- BERTOTTI, B., BRILL, D., KROTKOV, R. (1962) in Gravitation (WITTEN, L., Ed.) Wiley and Sons, New York.
- BRANS, C., DICKE, R.H. (1961) Phys. Rev. 124, 925.
- CHITRE, S.M., PAL, Y. (1968) Phys. Rev. Lett. 20, 219.
- CLEMENCE, G.M. (1943) Astr. Pap., Wash. 11, 1.
- DICKE, R.H. (1960) Am. J. Phys. 28, 344.
- DICKE, R.H. (1962a) in Evidence for Gravitational Theories (MOELLER, C., Ed.) Academic Press, New York.
- DICKE, R.H. (1962b) Phys. Rev. 125, 2163.
- DICKE, R.H. (1964a) in The Theoretical Significance of Experimental Relativity, Gordon and Breach, New York.
- DICKE, R.H. (1964b) Nature 202, 432.
- DICKE, R.H. (1965) Ann. Phys. 31, 235.
- DICKE, R.H. (1967) Physics to-day 20, 1, 1.
- DICKE, R.H., GOLDENBERG, H. Mark (1967a) Phys. Rev. Lett. 18, 313.
- DICKE, R.H., GOLDENBERG, H. Mark (1967b) Nature 214, 1294.
- DIRAC, P.A.M. (1938) Proc. R. Soc. A165, 199.
- DREVER, R.W.P. (1961) Phil. Mag. 6, 683.
- DUNCOMBE, R.L. (1958) Astr. Pap., Wash. 16, Pt.1.
- DYSON, F.J. (1967) Phys. Rev. Lett. 19, 1291.
- EINSTEIN, A. (1916) Annln Phys. 49, 769.
- EINSTEIN, A. (1949) in Albert Einstein, Philosopher-Scientist (SCHILPP, P.A., Ed.) Tudor Publ. Co., New York.
- EINSTEIN, A., FOKKER, A.D. (1914) Annln Phys. 44, 321.
- FEYNMAN, R. (1957) Chapel Hill Conf. on Relativity.
- GAMOW, G. (1967) Phys. Rev. Lett. 19, 759.

- GOLD, R. (1968) Phys. Rev. Lett. 20, 219.  
GUPTA, S. (1957) Rev. mod. Phys. 29, 334.  
HOWARD, R. (1967) Ann. Rev. Astr. Astrophys. 5, 1.  
HUGHES, V.W., ROBINSON, H.G., BELTRAN-LOPEZ, V. (1960) Phys. Rev. Lett. 4, 342.  
JORDAN, P. (1948) Astr. Nachr. 276.  
JORDAN, P. (1955) Schwerkraft und Weltall, Wieweg, Braunschweig.  
KALUZA, Th. (1921) S.B. preuss Akad. Wiss. 1921, 966.  
KLEIN, O. (1926) Nature 118, 516.  
LEIGHTON, R.B. (1963) Ann. Rev. Astr. Astrophys. 1, 19.  
MACH, E. (1877) Conservation of Energy, reprint Open Court Publ. Co., Chicago (1911).  
MARDEN, B.G. (1965) Bull. astr. phil. Soc. 93, 532.  
MORGAN, H.R., OORT, J.H. (1951) Bull. astr. Insts Neth. 2, 379.  
NORDSTRÖM, G. (1912) Phys. Z. 13, 1126.  
PEEBLES, P.J.E., DICKE, R.H. (1962a) Phys. Rev. 127, 629.  
PEEBLES, P.J.E., DICKE, R.H. (1962b) Phys. Rev. 128, 2006.  
PEEBLES, P.J.E. (1962) Ann. Phys. 20, 240.  
PERES, A. (1967) Phys. Rev. Lett. 19, 1293.  
POUND, R.V., SNIDER, R.L. (1965) Phys. Rev. 140, B788.  
ROSEN, N. (1940) Phys. Rev. 57, 147.  
SCHIFF, L.I. (1959) Proc. natn. Acad. Sci. U.S.A. 45, 69.  
SCHMALBERGER, D.C. (1963) Astrophys. J. 138, 693.  
SHAPIRO, I.I., PETTINGILL, G.H., ASH, M.E., STONE, M.L., SMITH, W.B., INGALLS, R.P.,  
BROCKELMAN, R.A. (1968a) Phys. Rev. Lett. 20, 1265.  
SHAPIRO, I.I., ASH, M.E., SMITH, W.B. (1968b) Phys. Rev. Lett. 20, 1517.  
THIRRING, W.E. (1961) Ann. Phys. 16, 96.  
WAPSTRA, A.H., NIJGH, G.J. (1955) Physica 21, 796.  
WAYMAN, P.A. (1966) Quart. J. R.A.S. 7, 138.



# GENERAL RELATIVITY THEORY AND EXPERIMENTS\*

J. WEBER

Department of Physics and Astronomy,  
University of Maryland,  
College Park, Md.,  
United States of America

## Abstract

GENERAL RELATIVITY THEORY AND EXPERIMENTS. 1. Introduction; 2. Dicke's recent experiments on solar oblateness; 2.1. Challenges to Dicke's interpretation; 2.2. Spin down problem; 2.3. Possible time dependence of the mass quadrupole moment; 2.4. Solar neutrino problem; 2.5. Summary of comments on Dicke's presentation; 3. University of Maryland experimental program.

## 1. INTRODUCTION

The first part of my discussion will attempt to defend Einstein's General Theory of Relativity in the light of Dicke's recent criticisms<sup>1</sup>. The second part will deal with my own program of development of apparatus and search for gravitational radiation.

I believe that Dicke's recent beautiful experiments teach us something new about the sun, but his results cannot be accepted as a disagreement with General Relativity.

There are many reasons for the wide acceptance of Einstein's General Relativity and time doesn't permit discussing even a few. I wish to stress the very nice connection with present-day particle physics. A particle point of view must start with bosons for a classically describable field and the infinite range of the force requires zero rest mass. A spin-zero theory (in vacuum) is

$$\square^2 \varphi = 0 \quad (1)$$

Equation (1) gives no light deflection and the incorrect perihelion advance. A spin-one theory would be (in vacuum)

$$\square^2 \varphi_\mu = 0 \quad (2)$$

Equation (2) is like electromagnetism, leading to repulsion between like masses. Therefore it is unacceptable.

A spin-two theory is, for the vacuum

$$\square^2 \varphi_{\mu\nu} = 0 \quad (3)$$

\* Supported in part by the National Science Foundation.

<sup>1</sup> DICKE, R. H., "General relativity: survey and experimental tests", these Proceedings.

But we must include self-interactions. The gravitational field is a source of energy and therefore a source for itself. To discuss this we note that Eq. (3) may be obtained from the Lagrangian density

$$L = -\frac{1}{2} \varphi^{\alpha\beta,\mu} \varphi_{\alpha\beta,\mu} \quad (4)$$

and from Eq.(4) a stress energy tensor for the gravitational field may be constructed according to

$$t_{\mu\nu} = \delta_{\mu\nu} L - \varphi^{\alpha\beta,\mu} \frac{\partial L}{\partial \varphi^{\alpha\beta,\nu}} \quad (5)$$

To include self-interactions, Eq.(3) needs to be replaced by

$$\square^2 \varphi_{\mu\nu} = \kappa t_{\mu\nu} \quad (6)$$

with  $\kappa$  as a suitable coupling constant, but the Lagrangian (4) gives Eq.(3) rather than Eq.(6). To obtain Eq.(6) we must modify Eq.(4) to obtain

$$L' = -\frac{1}{2} \varphi^{\alpha\beta,\mu} \varphi_{\alpha\beta,\mu} + L_1 \quad (7)$$

Now Eqs (7) and (5) lead to

$$t'_{\mu\nu} = \delta_{\mu\nu} L' - \varphi^{\alpha\beta,\mu} \frac{\partial L'}{\partial \varphi^{\alpha\beta,\nu}} \quad (8)$$

But Eq.(6) should be

$$\square^2 \varphi_{\mu\nu} = \kappa t'_{\mu\nu} \quad (9)$$

In order to obtain Eq.(9) we must again add a term to  $L$  getting

$$L'' = -\frac{1}{2} \varphi^{\alpha\beta,\mu} \varphi_{\alpha\beta,\mu} + L_1 + L_2 \quad (10)$$

Proceeding in this way leads to a Lagrangian density with an infinite number of terms, a non-linear theory which is identical with that written by Einstein. We see therefore that Einstein's General Relativity beautifully fits the particle (by spin) classification schemes of present-day physics.

Describing an apparently simple field like gravitation by mixed spin-zero and spin-two particle interactions is a step most of us are very reluctant to take. Of course this is not a good argument. The above discussion of gravitation from the standpoint of particle physics is due to Gupta [1] and similar treatments have been given by Thirring and Feynman.

## 2. DICKE'S RECENT EXPERIMENTS ON SOLAR OBLATENESS

The history of the perihelion advance is that Leverrier discovered it more than a century ago. A solar mass quadrupole moment was suspected as the cause before General Relativity. German astronomers looked for it and found nothing. Dicke's great ingenuity and skill were required to discover this exceptionally small 35 km oblateness in the solar surface.

In order to deduce a mass quadrupole moment from the shape of the surface, Dicke requires that the surface of the sun be an equipotential. All of us recall the theorem that for suitable boundary conditions the solution of Laplace's equation

$$\nabla^2 \phi = 0 \quad (11)$$

for the gravitational potential  $\phi$  is unique if one equipotential (and suitable boundary conditions at infinity) are given. Therefore, if the shape of the surface is known, the quadrupole moment follows – if the surface is truly an equipotential. It is not difficult to show that this requires neglect of shear viscosity forces.

### 2.1. Challenges to Dicke's interpretation

Roxburgh [2], Cocke [3] and others have questioned whether or not the solar surface is truly an equipotential. Cocke [3] notes that in the photosphere there is turbulent viscosity high enough to cause a great deal of distortion when acting on large-scale meridional currents. He could not, however, determine the sign of the resultant effect.

### 2.2. Spin down problem

To accept the interpretation of oblateness as a mass quadrupole effect demands a rapidly spinning interior (roughly once per day). This is not consistent with a slow spinning exterior (roughly once per month), unless the relaxation time for the angular momentum is extremely long and approximately comparable with the age of the sun. If the relaxation time is short, the sun cannot have the assumed quadrupole moment.

No-one really knows the answer, so let us resort to arguments about a cup of tea. Spin the cup. How long does it take for the fluid spin to relax? If the relaxation mechanism is simple viscosity, the relaxation time is  $\tau$  with

$$\tau_1 = L^2/\nu \quad (12)$$

In Eq.(12)  $L$  is the radius,  $\nu$  is the kinematic viscosity coefficient given in terms of the ordinary coefficient of viscosity  $\eta$  and the density  $\rho$  by

$$\nu = \frac{\eta}{\rho} \quad (13)$$

For a cup of tea,  $\nu \approx 10^{-2} \text{ cm}^2/\text{s}$ ,  $L \approx 3 \text{ cm}$

$$\tau = 900 \text{ s} \quad (14)$$

Clearly Eq.(14) is incorrect, yet this is the kind of mechanism required for a long solar relaxation time.

What happens in a cup of tea is that there are boundary layer phenomena and secondary flows [4] (Fig. 1). These secondary flows yield a much shorter spin down time

$$\tau \approx \left( \frac{L^2}{\Omega \nu} \right)^{\frac{1}{2}} \quad (15)$$

with  $\Omega$  the angular velocity.

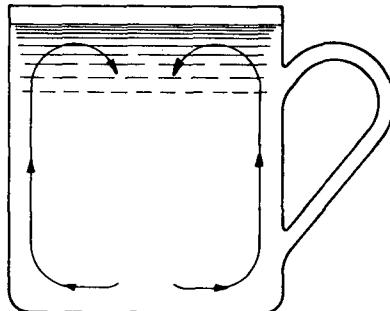


FIG.1. Secondary flows in a cup of tea.

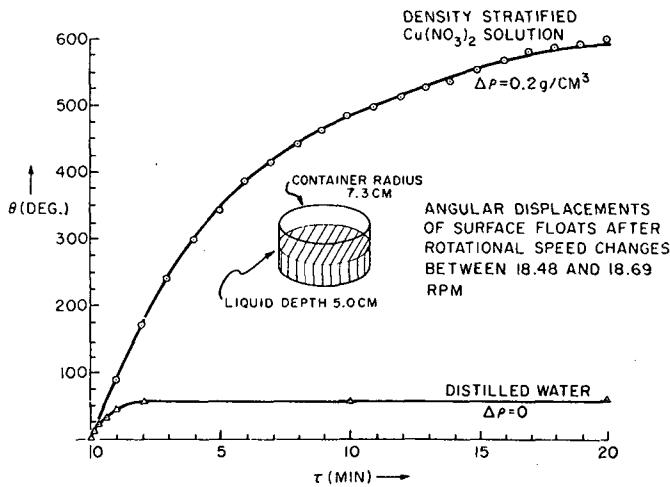


FIG.2. Comparison of spin down times for stratified and unstratified liquids.

Dicke argues that the sun behaves like a stratified medium in which buoyant forces oppose secondary flows. Figure 2 shows data on an experiment done by Dicke to support this view; clearly the stratified liquid has a longer relaxation time than the unstratified one. Spiegel has challenged Dicke's estimates of the spin down time (see Ref. [5]).

The applicability of the cup of tea arguments is perhaps best summarized by Dicke's eloquent statement that "the sun is no cup of tea" and Spiegel's equally eloquent reply that "the spin down problem is no piece of cake".

### 2.3. Possible time dependence of the mass quadrupole moment

If there is indeed a mass quadrupole moment, it may be related to the solar cycle or other phenomena. It would have to be established that the magnitude and direction are sufficiently stable in time to contribute to the perihelion advance.

### 2.4. Solar neutrino problem

The recent observations of Davis et al. [6] set a limit on solar neutrinos of  $\approx 0.3 \times 10^{-35}$  neutrinos per Cl atom per second for his apparatus. Reasonable solar models [7] give  $\approx 0.4 \times 10^{-35}$ . If Professor Dicke's theory is correct, G will have been stronger in the past, the sun's luminosity would therefore have been greater, and the sun will have evolved along the main sequence faster than we think. The helium abundance in the sun's core would be greater than expected and we should expect roughly  $0.8 \times 10^{-35}$  neutrinos per Cl atom in Davis' experiment. Thus the apparent disagreement here is evidence against the scalar-tensor theory.

### 2.5. Summary of comments on Dicke's presentation

It is difficult to argue with Professor Dicke because he has done an experiment. His interpretation rests on a number of incomplete arguments which are, taken one at a time, not entirely convincing.

My rebuttal consists of a series of also incomplete arguments, each of which is not convincing either. As Dirac has remarked, the sum of n incomplete arguments does not add up to a complete argument. All of us can only admire the brilliance of Dicke in devising this experiment and defending his position against a number of physicists and astrophysicists.

## 3. UNIVERSITY OF MARYLAND EXPERIMENTAL PROGRAM

The problem of the experimental confirmation of the existence of gravitational radiation is one of the challenging problems of our generation. I have discovered a method for the measurement of the Riemann tensor and search for gravitational radiation.

The idea is simple. Consider an elastic body (see, for example, Ref. [8]). In a curved space it will become deformed. In the presence of a time-dependent curvature its normal modes of oscillation may be excited. We employ a laboratory mass for search at the relatively high frequency  $\omega = 10^4$  and the normal modes of oscillation of the earth for the much lower frequencies upwards of one cycle per hour [8].

To test the detector we have gravitationally induced relative displacements [8] of the end faces of a two-metre cylinder of roughly 0.01 nuclear diameter, corresponding to strains of a few parts in  $10^{16}$ . These have been observed and measured. We have a second detector several kilometres away. Roughly once a month these detectors see a coincidence which cannot be statistical [9]. Analyses indicate, for example, that one of these events has a probability of occurring randomly about once in 8000 yr. These events are coincident with a time resolution much shorter

than the time for any kind of sound to propagate through the earth. An elaborate seismic array essentially guarantees that these events are not seismic in character.

A smaller number of events have been observed which do not appear to be coincident to less than a second. However, one of the detectors has two long period degrees of freedom and different kinds of excitation might conceivably result in a delayed output response.

I should emphasize that the counting rate is exceedingly small. We observe the noise output  $3 \times 10^7$  times per month and see roughly one count.

Have we detected gravitational waves? It is hard to be sure. If the signals are not gravitational, they are either some sort of unsuspected electromagnetic or seismic disturbance which enters the system in a way which elaborate precautions may have overlooked. To verify these results, we have installed a third detector at the Argonne National Laboratory near Chicago. It will be very interesting to see if the coincidences persist over the 1000-km baseline.

We are also designing a detector to search for gravitational radiation from the pulsars. The earlier developed techniques appear capable of extension using large masses and lock in detection techniques made possible by the very precise frequencies.

#### ACKNOWLEDGEMENTS

I wish to acknowledge the valuable assistance of Professor H. Zapsolsky in preparing these remarks, and many helpful and enlightening discussions with Professor R.H. Dicke. The task of preparing a rebuttal to his point of view was assigned to me by Professor Marshak. I deeply appreciate Professor Dicke's kindness in assisting me in the task.

#### REFERENCES

- [1] GUPTA, S.N., Rev. mod. Phys. 29 (1957) 334.
- [2] ROXBURGH, I.W., Nature 213 (1967) 1077.
- [3] COCKE, W.J., Phys. Rev. Lett. 19 (1967) 609.
- [4] GREENSPAN, H.P., HOWARD, L.N., J. Fluid Mech. 17 (1963) 385.
- [5] HOWARD, L.N., MOORE, D.W., SPIEGEL, E.A., Nature 214 (1967) 1997.
- [6] DAVIS, R., HARMER, D., HOFFMANN, K., Phys. Rev. Lett. 20 (1968) 1205.
- [7] BAHCALL, J., BAHCALL, N., SHAVIV, G., Phys. Rev. Lett. 20 (1968) 1209.
- [8] WEBER, J., General Relativity and Gravitational Waves, Chapter 8, Interscience Publishers, New York, London (1961); Physics to-day (April 1968).
- [9] WEBER, J., Phys. Rev. Lett. 20 (1968) 1307.

# THE QUANTIZATION OF THE GRAVITATIONAL FIELD

P.A.M. DIRAC

Department of Applied Mathematics and Theoretical Physics,  
University of Cambridge,  
Cambridge, United Kingdom

## Abstract

THE QUANTIZATION OF THE GRAVITATIONAL FIELD. 1. The classical rules for constraints; 2. The quantum rules for constraints; 3. The constraints for the classical gravitational field; 4. The quantization problem.

There is no experimental evidence for the quantization of the gravitational field, but we believe quantization should apply to all the fields of physics. They all interact with one another, and it is difficult to see how some could be quantized and others not. I shall here deal with the quantization of the gravitational field of Einstein. I shall take the gravitational field to be all alone, so our starting point is the Einstein equations for empty space.

As a basis for a quantum theory we need the corresponding classical theory in the Hamiltonian form. We can get the Einstein theory in the Hamiltonian form by working from the action principle, which was discovered by Einstein, and applying standard methods. The result is a Hamiltonian theory with constraints.

## 1. THE CLASSICAL RULES FOR CONSTRAINTS

I must first discuss the general rules governing constraints in classical Hamiltonian dynamics. (The underlying theory is contained in my paper, Proc. R. Soc. A 246 (1958) 326.) The constraints are equations connecting the dynamical co-ordinates and momenta, the  $q$ 's and the  $p$ 's, say

$$\phi_m \approx 0 \quad m = 1, 2, 3 \dots \quad (1)$$

Only those physical states are allowed for which the constraints are satisfied. We must not use constraints in Poisson brackets, i.e.  $\phi_m \approx 0$  does not lead to  $[A, \phi_m] \approx 0$ . Owing to this limitation on the use of constraint equations, I call them weak equations and use the special sign  $\approx$  for them.

The constraints must be real

$$\bar{\phi}_m = \phi_m \quad (2)$$

Constraints can of course be added and can be multiplied by factors which may be any real functions of the  $q$ 's and  $p$ 's. According to the

general theory, constraints can be added to the Hamiltonian to give another Hamiltonian on the same footing as the original one

$$H^* = H + \sum_m \lambda_m \phi_m \quad (3)$$

where the  $\lambda$ 's are any real functions of the  $q$ 's and  $p$ 's.

The initial  $q$ 's and  $p$ 's may be chosen arbitrarily, provided they satisfy the constraints (1). But then the later  $q$ 's and  $p$ 's are not determined, owing to the arbitrariness in the  $\lambda$ 's. The physical state at all times is determined by the initial  $q$ 's and  $p$ 's. It follows that the arbitrariness in the later  $q$ 's and  $p$ 's corresponds to making mathematical transformations that are not of physical importance, such as the gauge transformations of electrodynamics or the co-ordinate transformations of general relativity. The existence of constraints in the theory corresponds to the existence of such mathematical transformations.

If a dynamical variable  $\alpha$  satisfies the conditions

$$[\alpha, \phi_m] \approx 0 \quad \text{all } m \quad (4)$$

then it is invariant under the mathematical transformations.

The constraints have to satisfy certain consistency requirements for the theory to be consistent, namely

$$[\phi_m, H] \approx 0 \quad (5)$$

$$[\phi_m, \phi_n] \approx 0 \quad (6)$$

These weak equations are equivalent to the strong equations

$$[\phi_m, H] = \sum_n a_{mn} \phi_n \quad (7)$$

$$[\phi_m, \phi_n] = \sum_k b_{mnk} \phi_k \quad (8)$$

with suitable coefficients  $a$ ,  $b$ .

## 2. THE QUANTUM RULES FOR CONSTRAINTS

We must now consider how to use the constraint equations (1) in the quantum theory. We cannot use them in Poisson brackets, so we cannot multiply them by general factors, functions of the  $q$ 's and  $p$ 's, both on the left and on the right. Let us make the rule that we can multiply them only by factors on the left, so that  $\phi_m \approx 0$  leads to  $\lambda \phi_m \approx 0$ , but not in general to  $\phi_m \lambda \approx 0$ .

We keep the consistency requirements (5) and (6). When they are written as strong equations (7), (8), it is now necessary that all the coefficients  $a$ ,  $b$  shall be on the left of their respective  $\phi$ 's.

The conditions (8) with the coefficients  $b$  all on the left mean that there exist kets  $|P\rangle$  satisfying

$$\phi_m |P\rangle = 0 \quad \text{all } m \quad (9)$$

Such kets will play a special role in the theory. We assume that these kets, and only these kets, correspond to physical states. The equations (9) provide supplementary conditions on the wave function.

The classical  $\phi_m$  are real. We should expect the quantum  $\phi_m$  to be also real, i.e. Hermitian operators. In the past, people (including myself) have not paid sufficient attention to this reality requirement (2). With  $\phi_m$  real, in general  $\lambda\phi_m$  is not also real. We have to impose the restriction that, from  $\phi_m \approx 0$ , we can infer  $\lambda\phi_m \approx 0$  only in the special case when  $\lambda\phi_m$  is real as well as  $\phi_m$ , which usually requires that  $\lambda$  shall be real and shall commute with  $\phi_m$ . The symmetry between left and right multiplication is now restored.

The restriction on the coefficients of  $\phi_m$  must apply to the  $a$ 's and  $b$ 's in Eqs (7) and (8). The consistency requirements in the quantum theory are then much more rigid than in the classical theory.

With the  $\lambda$ 's satisfying the restriction, we can introduce  $H^*$  by Eq. (3) and use it on the same footing as  $H$  to provide Heisenberg equations of motion or a Schrödinger wave equation.

### 3. THE CONSTRAINTS FOR THE CLASSICAL GRAVITATIONAL FIELD

The classical Hamiltonian formulation for the gravitational field has been worked out in my paper in Proc. R. Soc. A 246 (1958) 333. We choose co-ordinates such that the surfaces  $x^0 = \text{constant}$  are space-like. Let  $e^{rs}$  be the reciprocal matrix to  $g_{rs}$  ( $r, s = 1, 2, 3$ ) and let the determinant of the  $g_{rs}$  be  $-K^2$  and the determinant of the  $g_{\mu\nu}$  be  $-J^2$ .

The dynamical co-ordinates, the  $q$ 's, are the  $g_{rs}$  at all points of the space-like surface  $x^0 = \text{constant}$ . The conjugate momenta  $p^{rs}$  are given by

$$p^{rs} = (e^{rs} e^{ab} - e^{ra} e^{sb}) J \Gamma_{ab}^0 \quad (10)$$

where  $\Gamma$  denotes a Christoffel symbol.

There are four constraints for each point of the surface  $x^0 = \text{constant}$ . Three of them form a three-vector density in the surface; they are ( $u = 1, 2, 3$ )

$$\phi_u = p^{rs} g_{rs,u} - 2(p^{rs} g_{ru})_s \approx 0 \quad (11)$$

The fourth, which we shall write as  $\phi_L$ , is a three-dimensional scalar density in the surface and is

$$\phi_L = K^{-1} (g_{rs} g_{ab} - g_{ra} g_{sb}) p^{rs} p^{ab} + K R_3 \approx 0 \quad (12)$$

where  $R_3$  is the three-dimensional scalar curvature of the surface, and is a function of the  $g_{rs}$  and their first and second spatial derivatives, but does not involve the  $p^{rs}$ .

A possible Hamiltonian is  $H=0$ . The general Hamiltonian, according to Eq.(3), is just a linear function of the  $\phi$ 's with arbitrary real coefficients. The consistency conditions (5) or (7) are now trivial.

One finds that the  $\phi$ 's satisfy the Poisson bracket relations (the prime denotes a variable taken at the point  $x_1^!, x_2^!, x_3^!$ )

$$[\phi_u, \phi_v^!] = \phi_v \delta_{,u} (x - x^!) + \phi_u^! \delta_{,v} (x - x^!) \quad (13)$$

$$[\phi_L, \phi_u^!] = \phi_L^! \delta_{,u} (x - x^!) \quad (14)$$

$$[\phi_L, \phi_L^!] = -\{e^{rs} \phi_r + e^{irs} \phi_i^!\} \delta_{,s} (x - x^!) \quad (15)$$

Since the right-hand sides here are all linear combinations of the  $\phi$ 's, the consistency conditions (6) or (8) are confirmed.

#### 4. THE QUANTIZATION PROBLEM

When we quantize the theory we make the  $g_{rs}$ ,  $p^{rs}$  into operators satisfying the commutation relations

$$[g_{rs}, p'^{ab}] = \frac{1}{2} (\delta_r^a \delta_s^b + \delta_s^a \delta_r^b) \delta(x - x^!) \quad (16)$$

The problem is to choose suitably the order of the non-commuting factors in the  $\phi$ 's, given classically by Eqs (11) and (12), so as to get the appropriate consistency conditions for the quantum theory.

The classical  $\phi_u$  contains the term  $p^{rs} g_{rs,u}$ , in which the two factors are both taken at the same field point  $x^1, x^2, x^3$ . Thus the difference between  $p^{rs} g_{rs,u}$  and  $g_{rs,u} p^{rs}$  involves the derivative of  $\delta(0)$ , which is an undetermined number. The same remarks apply to the second term in the classical  $\phi_u$ , namely  $(p^{rs} g_{ru})_s$ . However, there is no difficulty here, because the quantum  $\phi_u$  has to be Hermitian, so we must take it to be

$$\phi_u = \frac{1}{2} (p^{rs} g_{rs,u} + g_{rs,u} p^{rs}) - (p^{rs} g_{ru} + g_{ru} p^{rs})_s \quad (17)$$

The Poisson bracket  $[\phi_u, \phi_v^!]$  must then also be Hermitian and must just equal the right-hand side of Eq.(13) with Hermitian  $\phi_u$ 's. The quantum consistency condition (13) is thus fulfilled.

We must now consider  $\phi_L$ , given classically by Eq.(12). The term  $KR_3$  here can be taken over directly into the quantum theory, as it involves only commuting factors. However, the first term of  $\phi_L$  contains non-commuting factors, and different ways of ordering the factors give expressions which differ by functions of the  $g_{rs}$ ,  $p^{rs}$  multiplied by  $\delta(0)$ , which is infinite. The Hermitian condition is no longer adequate to fix the quantum expression.

$\phi_L$  must satisfy the two consistency conditions (14) and (15). The first of these presents no difficulty, because it is a general formula applying to any three-dimensional scalar density  $\phi_L$ , and merely expresses how such a scalar density gets altered by a transformation of the coordinates  $x^1, x^2, x^3$  on the surface. The second condition is the difficult one, the difficulty showing itself by the appearance of coefficients on the right-hand side of Eq.(15) which are not numbers.

The problem has been investigated by J. Schwinger (Phys. Rev. 132 (1962) 1317). Schwinger uses a different notation, based on the dynamical co-ordinates  $q^{rs} = K^2 e^{rs}$ , with appropriate momenta conjugate to them, but his equations are equivalent to the ones given here.

Schwinger makes his attack more powerful by considering the quantum analogue, not of  $\phi_L$ , but of  $K^n \phi_L$ , where  $n$  is some number at our disposal. Classically,  $K^n \phi_L$  satisfies, as the conditions replacing Eqs (14) and (15)

$$[K^n \phi_L, \phi'_u] = (K^n \phi_L)_{,u} \delta(x-x') + (n+1) K^n \phi_L \delta_{,u}(x-x') \quad (18)$$

$$[K^n \phi_L, K'^n \phi'_L] = -K^n K'^n \{e^{rs} \phi_r + e^{is} \phi'_r\} \delta_{,s}(x-x') \quad (19)$$

Schwinger assumes for the quantum  $K^n \phi_L$  the Hermitian expression

$$(K^n \phi_L)_Q = p^{rs} K^{n-1} (g_{rs} g_{ab} - g_{ra} g_{sb}) p^{ab} + K^{n+1} R_3 \quad (20)$$

It satisfies Eq. (18) for any  $n$ , since this equation holds for any quantity that transforms suitably under transformations of the co-ordinates  $x^1, x^2, x^3$  of the surface.

Schwinger proceeds to examine what value of  $n$ , if any, would make  $(K^n \phi_L)_Q$  satisfy Eq. (19) with the quantum expressions (17) for the  $\phi_u$ 's on the right-hand side, with their coefficients all on the left. The calculation involves quantities of the form  $\delta(x-x') \delta_{,r}(x-x')$ . There does not exist any general method for handling such quadratic quantities in the  $\delta$ -function, free from inconsistencies. However, by using simple and plausible but non-rigorous methods, Schwinger is able to show that the condition is satisfied with  $n=3$ . Since the right-hand side of Eq. (19) must be Hermitian, it must then be equally possible to have all the coefficients of the  $\phi_u$ 's on the right.

The problem of the quantization of the gravitational field is thus left in a rather uncertain state. If one accepts Schwinger's plausible methods, the problem is solved. But one cannot be happy with such methods without having a reliable procedure for handling quadratic expressions in the  $\delta$ -function.

The sort of inconsistency one must prevent is illustrated by the following calculation with  $\delta(y)$  for one variable  $y$ . We have

$$y \delta(y) = 0 \quad (21)$$

Differentiating, we get

$$y \delta'(y) = -\delta(y) \quad (22)$$

Multiplying Eq. (21) by  $\delta'(y)$ , we get

$$y \delta(y) \delta'(y) = 0$$

and multiplying Eq. (22) by  $\delta(y)$ , we get

$$y \delta(y) \delta'(y) = -\{\delta(y)\}^2$$



# ON GRAVITATIONAL COLLAPSE

R. PENROSE

Department of Mathematics,  
Birkbeck College,  
London, United Kingdom

## Abstract

ON GRAVITATIONAL COLLAPSE. The gravitational collapse of a spherically symmetrical body, to within its Schwarzschild radius, is discussed in qualitative terms. The concept of a "trapped surface" is presented, in order to deal with asymmetrical cases. The effect of rotation is illustrated by reference to the Kerr solution. Two cases, namely of "small" and "large" angular momentum, are contrasted. In the latter case no trapped surface (and no event horizon) arises.

A new theorem, due jointly to S.W. Hawking and the present author, is described, which effectively incorporates and generalizes the previous theorems predicting space-time singularities. In particular, if a physically acceptable space-time contains a trapped surface (or, alternatively, if it is spatially closed), then on the basis of Einstein's equations (with  $\lambda = 0$ ), it must possess a singularity.

I have taken it that I am to present a "point of view" on relativity. In fact, I think I shall present not one but two (more or less orthogonal) points of view - neither of which really adequately represents my own viewpoint.

The first point of view is that of the pure relativist. The pure relativist treats general relativity as a branch of differential geometry. He is really more a pure mathematician than a physicist and is not normally concerned with the relation of his theory to observation or experiment. He strives for elegance in his arguments and regards this as an end in itself. He has a strong distaste for co-ordinates in general, although on occasion he reluctantly finds he has the need to use them. He is an extreme sceptic as regards physical concepts, even those which have historically been regarded as lying at the root of his subject. He cannot, for example, comprehend the meaning of such terms as "general covariance", "the strong equivalence principle" (although the "weak equivalence principle" he can probably manage) or "Mach's principle". He cannot, of course, understand the concept of "force" or of "potential energy" but a "total energy-momentum" concept he can, with great effort, just about accept.

The second point of view I wish to present can be referred to as that of the gravitational optimist (although some people might feel that "pessimist" might be a more appropriate term). The gravitational optimist believes that general relativity has fundamental importance to the physics of our world. He believes, in particular, that the observational aspects of the theory will not always be restricted to a few very minor effects like the perihelion advance of Mercury; that cosmological interest in the theory will not ultimately be restricted to a few simple smoothed-out symmetrical models. Moreover, he would like to believe that the effects of space-time curvature ought somehow to be involved in determining the particles we observe about us. It would seem then, that because of the extreme smallness of numbers, like  $2.4 \times 10^{-43}$ , which govern the size

of the gravitational coupling, the gravitational optimist should require that "fantastically strong" gravitational fields must exist (or have existed) somewhere in our universe. By a "fantastically strong" field, I mean one involving radii of curvature (say) as small as the order of an elementary particle size, so that one must expect that the local physics would then be very drastically affected by the presence of gravitation - the local Lorentz covariance having been completely destroyed.

Although the aims of the pure relativist and of the gravitational optimist are so different, there is perhaps a thread of logic connecting them. For if the pure relativist is really a pure mathematician at heart, why does he spend his time with Riemannian manifolds only of dimension four, with that peculiarly arbitrary (+, -, -, -) signature? It is as though some physicist friend of his once told him that such manifolds had to do with the structure of our world - but omitted to mention  $10^{-43}$ ! Our pure relativist, having devoted so much of his time to his subject may eventually (in his weaker moments!) feel the need to interest physicists in his results. In this way he may be led to have some sympathy with the views of the gravitational optimist.

The topic I wish to discuss, namely gravitational collapse, illustrates aspects of both the points of view I have mentioned. The role of the pure relativist, here, is in making precise the nature of the situation which arises and in proving exact mathematical theorems applicable to this situation - theorems which one would be hard put to obtain using any other viewpoint. In effect, the pure relativist supplies techniques, whereas it is the gravitational optimist who supplies motivation. The object of the exercise will be, in fact, to establish, on the basis of present physical understanding, the actual existence of regions of enormous curvature in our universe - although the precise results obtained fall perhaps marginally short of this aim.

Let us first establish our conventions. The metric signature (+, -, -, -) will be used, with "gravitational" units chosen, so that

$$G = c = 1 \quad (1)$$

Einstein's equations are then

$$R_{ab} - \frac{1}{2}Rg_{ab} + \lambda g_{ab} = -8\pi T_{ab} \quad (2)$$

allowing for the possibility of a cosmological constant  $\lambda$ . I shall not concern myself here with possible modifications of the Einstein theory, except to mention that essentially everything I shall say will apply equally well whether one adopts the standard Einstein theory (2) or the "scalar-tensor" modification of it mentioned by Prof. Dicke in his lecture (Brans-Dicke theory [3, 6]). In the latter case, one must use for the metric ds in what follows, not the most directly physical metric, but the conformally related one in which the main field equations closely resemble Eq.(2) (see Ref. [5]).

I shall consider, first, the case of an exactly spherically-symmetrical gravitational collapse. We know from the classical result of Chandrasekhar [4] (and Landau) that a cold (non-rotating) body which has reached the end-point of nuclear activity and whose mass is larger than about one and

one third solar masses cannot hold itself apart against the effects of its own gravitational field. Such a body will therefore collapse and there will be nothing to halt this collapse. Owing to the exact spherical symmetry, the metric exterior to the body must be the Schwarzschild solution [1], which in the usual co-ordinates (taking  $\lambda=0$ ) is

$$ds^2 = \left(1 - \frac{2m}{r}\right) dt^2 - \left(1 - \frac{2m}{r}\right)^{-1} dr^2 - r^2 (d\theta^2 + \sin^2 \theta d\phi^2) \quad (3)$$

$m$  being the mass of the body in the gravitational units (1). A drawback of the familiar form (3) is that it does not allow us to discuss collapse to within the critical radius given by  $r = 2m$ . The form clearly becomes singular at  $r = 2m$  but this is really a co-ordinate effect arising from the fact that the  $t$ -co-ordinate behaves badly at  $r = 2m$ . We can see this most easily if we replace  $t$  by the advanced time parameter  $v$  given by

$$V = t + r + 2m \log(r - 2m) \quad (4)$$

The metric (3) then transforms to the Eddington-Finkelstein form [7, 8]

$$ds^2 = \left(1 - \frac{2m}{r}\right) dv^2 - 2 dv dr - r^2 (d\theta^2 + \sin^2 \theta d\phi^2) \quad (5)$$

The form (5) is non-singular over the entire range  $0 < r < \infty$ . The vanishing of the coefficient of  $dv^2$  at  $r = 2m$  implies merely that the particular co-ordinate hypersurface  $r = 2m$  is null. (For finite  $v$ , we have  $t = \infty$  on  $r = 2m$ ; see Eq.(4).) The space-time Eq.(5) in the neighbourhood of a point at  $r = 2m$  is locally Minkowskian, as it also is everywhere else ( $r > 0$ ).

The space-time in the matter region does not have the metric given by Eq.(3) or Eq.(5), but I shall be concerned here only with the empty exterior field. The situation is as depicted in Fig. 1, with one space dimension suppressed. The matter collapses to within  $r = 2m$ . An observer sitting on the surface of the body would notice nothing peculiar occurring when he crosses  $r = 2m$ . The curvature (i.e. gravitational tidal forces) would mount steadily, however, becoming finally infinite at  $r = 0$ . Unlike the hypersurface  $r = 2m$  (often misleadingly referred to as the "Schwarzschild singularity" owing to the failure of the  $t$ -co-ordinate used in Eq.(3)), the region  $r = 0$  is a true physical singularity of the metric and not merely a co-ordinate effect.

Although  $r = 2m$  is not a singularity, this region does have some peculiarities from the global point of view. Let us consider an external observer situated some distance from the collapsing body. To ascertain what he sees of the body, we must trace backwards the light rays entering his eye from the direction of the body. From the way the light cones are tipping over near  $r = 2m$ , it should be evident that, no matter how long he waits, our observer "sees" the body - albeit exceedingly dimly - as it was only just before it plunged into the  $r < 2m$  region. In fact, owing to similarities between this and other situations arising in cosmology, we may refer to the  $r = 2m$  hypersurface as an event horizon. No signal emitted in the empty  $r < 2m$  region can ever escape to the outside, where-

as signals emitted in  $r > 2$  m can cross inwards to be trapped inside  $r = 2$  m. This "one-way traffic" nature of  $r = 2$  m results from the fact that it is a null hypersurface, i.e. everywhere tangent to the light cone, as is depicted in Fig. 1.

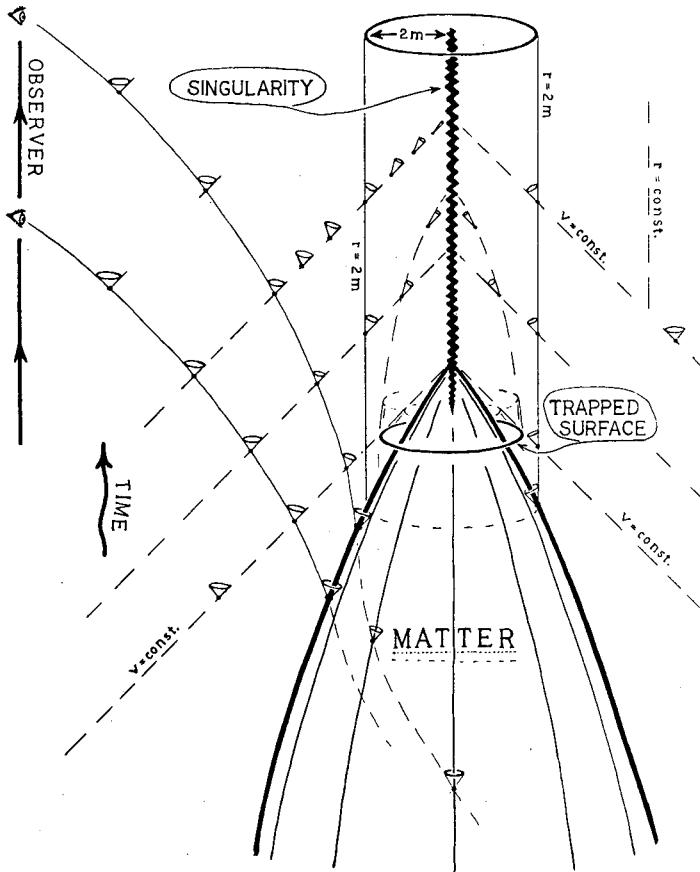


FIG. 1. Space-time picture of a symmetrical collapse to within the Schwarzschild radius  $r = 2$  m.

The radii of space-time curvature in the neighbourhood of  $r = 2$  m (though easily small enough to kill a man in free-fall across  $r = 2$  m in the case of a collapsed star) are not nearly of an order which would impress our gravitational optimist. His attention is drawn instead to the neighbourhood of  $r = 0$ , at which the radii of curvature become arbitrarily small. But here we must worry particularly about the implications of our initial assumption of exact spherical symmetry. For it is to be expected that even the slightest departure from spherical symmetry in the external field could result in very gross departures from the metric (5) near  $r = 0$ . Thus it is not at all clear that the enormous curvatures desired by our gravitational optimist would result at all in a "generic" physical situation. Indeed, we must even ask whether any analogue of  $r = 2$  m is to be

expected in a realistic gravitational collapse. Might not the existence of an event horizon of this type be a very exceptional situation characteristic only of the exact symmetry?

It is in order to answer such questions that the concept of a trapped surface is introduced. A trapped surface is a closed (i.e. compact without boundary) spacelike 2-surface  $T$  which has the property that the null geodesics which meet  $T$  orthogonally are all converging in the neighbourhood of  $T$ . This convergence is measured in the sense of decreasing area of cross-section as we proceed into the future following the null geodesics. In more physical terms, we may envisage a flash of light originating "instantaneously" all over  $T$ . This flash spreads away from  $T$  in two directions orthogonal to  $T$  at each point. The characteristic property of a trapped surface is that the area of the flash decreases in both directions at every point - that is to say, the intensity of the flash everywhere increases. An ordinary spacelike 2-sphere in Minkowski space is clearly not trapped. Here there is an ingoing flash for which the area indeed decreases, but for the outgoing flash the area increases. However, surfaces do exist in Minkowski space which would be trapped except for the fact that they are not closed (e.g. the intersection of the past light cones of two spatially separated points). This illustrates the fact that the existence of a trapped surface in a space-time is not merely a local property of the space-time. For the space-time given by our metric (5) the surfaces  $r = \text{const.}$ ,  $v = \text{const.}$  are trapped for  $0 < r < 2m$ . Thus, the event horizon  $r = 2m$  in fact represents the boundary between the region which contains trapped surfaces and that which does not.

We may picture the situation of Fig. 1 in another way. Let us take a section of Fig. 1 by a spacelike "hyper-plane", that is,  $v-r = A (= \text{const.})$ . We can think of this as representing the situation at one "instant of time". Superimpose on this picture another section of Fig. 1 taken "a moment later", that is to say by  $v-r = A + \epsilon$  ( $0 < \epsilon = \text{const.}$ ). Corresponding to each point  $P$  of the "earlier" section, we draw in the intersection of the light cone of  $P$  with the "later" section. For example, if, instead of Fig. 1, we had chosen to represent Minkowski space in this way (sections by  $t = \text{const.}$ ), then our resulting picture would consist of a three-dimensional Euclidean space, where each point  $P$  would have drawn about it a sphere of radius  $\epsilon$  and centre  $P$ . This sphere and its interior would represent the region that could be causally influenced by an occurrence at  $P$ , a time  $\epsilon$  later; the surface of this sphere would represent the location, after time  $\epsilon$ , of a light flash emitted at  $P$ . In the same way, Fig. 2 illustrates the progress of light flashes in the collapse situation of Fig. 1 (taken after the matter has collapsed down to  $r = 0$ ). In Fig. 3 the same situation is depicted on a larger scale, so as to illustrate a trapped surface  $T$  and the locally decreasing surface area of a light flash emitted at  $T$ .

Let us now drop the assumption of spherical symmetry - and we may also drop the requirement that our collapsing body be surrounded only by empty space. Can we be sure that trapped surfaces can result in a realistic gravitational collapse? I think it is clear that (on the basis of general relativity) trapped surfaces must at least occasionally result from a collapse. The fact that there can be nothing in principle against trapped surfaces arising follows from a variety of reasons. The argument I like best is the following "Gedanken-experiment": A mad dictator, who rules a galaxy containing (say)  $10^{11}$  stars, decides that he wishes to

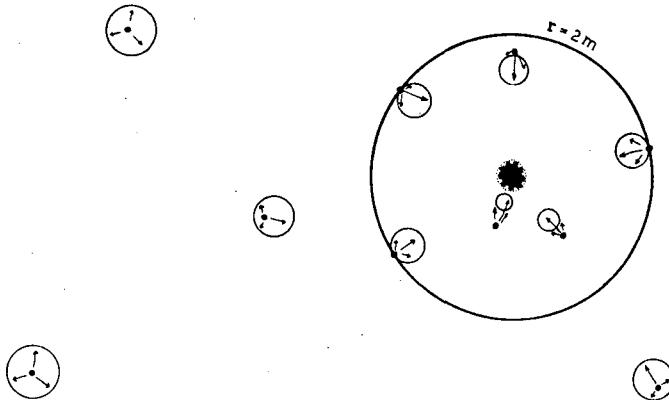


FIG. 2. Space picture of a spherically symmetrical collapsed object.

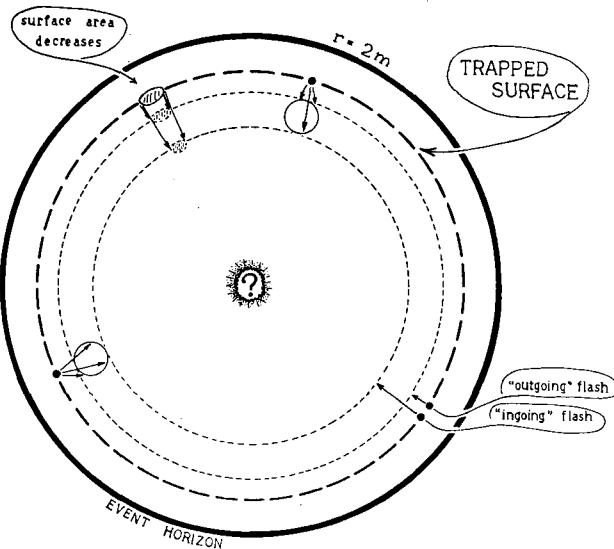


FIG. 3. The locally decreasing surface areas associated with a trapped surface.

build a trapped surface. This he can do (whether or not his galaxy already contains a trapped surface at its centre anyway!) by adjusting the velocities of the stars so that they all fall towards the centre so as to arrive at about the same time. It can then be shown that merely because of the focusing effect on the (idealized) light rays which pass near to (or through) the stars, a trapped surface will result before any serious problem of stellar collisions can arise.

But what are the consequences of a trapped surface arising? In fact, at least two things can be said. In the first instance, it follows from the weak energy assumption (6), which will be discussed shortly, that anything enclosed by a trapped surface (i.e. whose world line threads through

it, roughly speaking) cannot be observed by any external observer. Secondly, as a consequence of the theorem to be stated shortly, singularities in space-time must be expected to result.

Before considering the theorem, it is worthwhile to examine, in addition to the Schwarzschild solution, another exact vacuum solution, namely that of Kerr [15] (see also Ref. [2]), since it enables us to incorporate rotation (angular momentum) into the collapse discussion. The Kerr solution depends on two free parameters  $m$  and  $a$ . Here  $m$  describes the total mass (as in the Schwarzschild case) and  $a$  the angular momentum. The solution gives the field of only a restricted class of spinning bodies, since the quadrupole and higher moments must all be certain explicit functions of the two parameters  $a, m$ . Nevertheless, there are some reasons for believing that whenever a trapped surface develops, the final asymptotic field which results may always be effectively a Kerr solution (or Kerr-Newman solution [16] if there is a non-zero net charge), all the excess multipole moments being ultimately radiated away by gravitational (plus electromagnetic) radiation, as the field itself finally settles down after the collapse has taken place. This is largely conjecture at present, but appears to be borne out in the non-rotating case at least, by a result due to Israel [14].

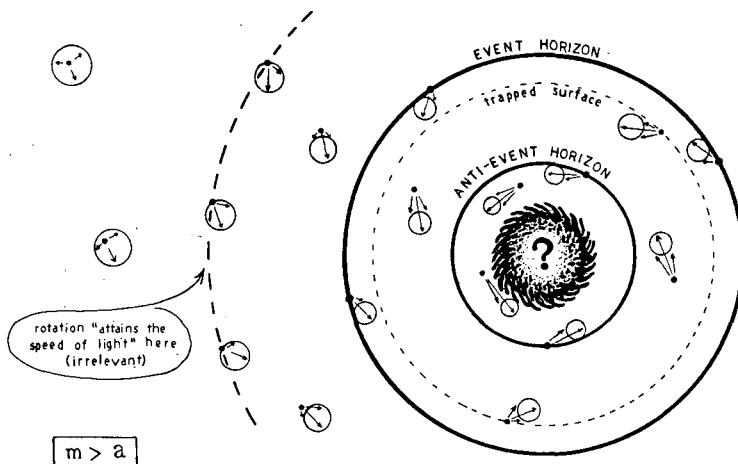


FIG. 4. Space picture of a Kerr solution for "small" angular momentum, representing a rotating collapsed object.

There are two cases to consider in the Kerr solution. First there is the case  $m > a$  of small angular momentum. The situation is depicted in Fig. 4 (viewed along the rotation axis), which is drawn according to the same conventions as Figs 2 and 3. Note that external to the event horizon there is a hypersurface (irrelevant for our purposes) inside which the system is "rotating faster than light", that is to say, a particle would have to exceed the local light velocity in order to appear to be stationary to an observer at infinity. (The picture unduly emphasizes the importance of this region - our pure relativist would not much care for Figs 2 - 5!) The event horizon here is a stationary null hypersurface. Signals can cross

it from the outside to the inside but not the other way about. There is also another stationary null hypersurface inside the event horizon which we may refer to as the anti-event horizon. There is an inward directed one-way traffic of signals between the event horizon and the anti-event horizon. But inside the anti-event horizon signals can pass freely both inwards and outwards provided they spiral around sufficiently. Trapped surfaces reside in the region between the event horizon and the anti-event horizon. This region acts as a kind of shield for whatever chaos resides inside.

Now consider a sequence of Kerr solutions for which  $m$  is kept fixed but  $a$  is steadily increased. The anti-event horizon gets nearer and nearer to the event horizon until finally they annihilate one another leaving the internal chaos visible from the outside. However, it should be emphasized that for any particular collapse situation in which trapped surfaces do form, the event horizon can never be annihilated in this way. The internal chaos must be forever hidden from the outside world. If a collapse situation gives rise to a metric similar to a Kerr solution with  $a > m$ , then it cannot do so via metrics containing trapped surfaces. Furthermore, since we have no trapped surface for  $a > m$ , we also have no theorem predicting enormous curvature regions - at least, this is the situation at present.

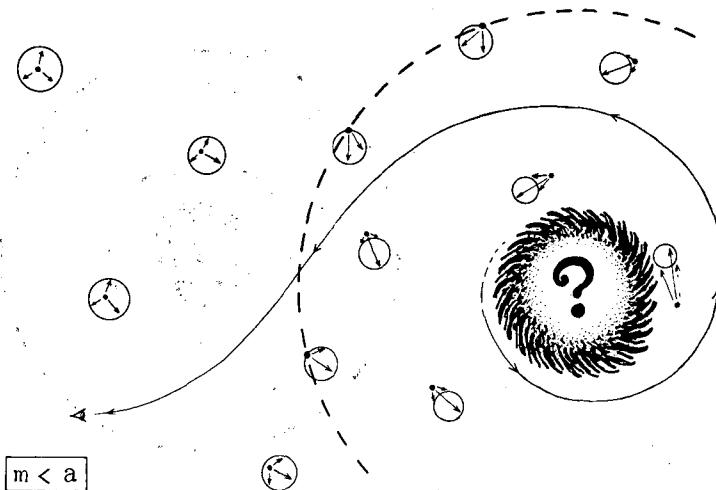


FIG.5. Space picture of a Kerr solution for "large" angular momentum. Signals can escape from the internal region to be observed by an external observer.

Thus, the question mark at the centre of Fig.5 is really a different kind of question mark from that in Fig.4. In a generic (physically realistic) situation like Fig.4, we know that there still remains a region that we cannot understand on the basis of present space-time theory (a "singularity"), while for a generic Fig.5 we do not know whether there is such a region. A point of some interest, but whose significance is not clear, is that the singularity region of Fig.4 in the exact Kerr solution resides some way away from the event horizons and is qualitatively not different from the singularity region of Fig.5. Thus it might seem that the nature of the

true singularity (e.g. enormous curvature region) is not sensitive to the presence or absence of an event horizon<sup>1</sup>. This may be misleading, however, as the existing singularity theorems give no indication as to the location of the singularities in generic situations and very little indication as to their nature. I think that the best description of a generic collapse with  $a > m$  (if such occurs) is, at present, simply a big question mark.

I now wish to describe a precise mathematical theorem which forms the basis of many of the above remarks concerning the existence of singularity regions. Since we have no desire to restrict attention just to vacuum solutions nor just to matter with some particular equation of state, the content of Einstein's equations (2) will manifest itself only as inequalities on the Ricci curvature. Such inequalities we can refer to as energy conditions. The two energy conditions of interest to us here are the following:

The weak energy condition:

$$\ell^a \ell_a = 0 \text{ implies } R_{ab} \ell^a \ell^b \leq 0 \quad (6)$$

The strong energy condition:

$$t^a t_a = 1 \text{ implies } R_{ab} t^a t^b \leq 0 \quad (7)$$

(That condition (6) is a consequence of condition (7) follows by a limiting argument.)

The interpretation of the "weak" condition (6) is that with Einstein's equations (2) it becomes

$$E + p_i \geq 0 \quad (i = 1, 2, 3) \quad (8)$$

where in an eigentetrad of  $T_{ab}$ ,  $E$  is the energy density and  $p_1, p_2, p_3$  are the three principal pressures (unequal only for anisotropic media). In fact Eq.(6) (or Eq.(8)) is a consequence of the non-negative definiteness of the (local) energy density (that is:  $t^a t_a = 1$  implies  $T_{ab} t^a t^b \geq 0$ ). Thus, we expect any reasonable matter to satisfy Eq.(8). From this will follow all the essential properties of trapped surfaces mentioned earlier, namely that space-time singularities must be expected to result, in a region which is not visible from the outside. However, a loophole remains in the application of the precise theorem [17, 18] which uses only Eq.(6). (This is the possibility of the non-existence of an open global Cauchy hypersurface in our universe). This loophole can be removed if we adopt Eq.(7) rather than Eq.(6).

---

<sup>1</sup> I have been somewhat vague about the meaning of the term "event horizon" for a generic situation. More accurately, we could call this the "common event horizon for all external inertial observers". It represents the boundary of the set of all points which cannot be connected to external infinity by time-like future-directed curves (assuming asymptotic flatness). This set contains all the trapped surfaces, but will in general extend out somewhat beyond the trapped surface region.

The interpretation of the "strong" condition (7), provided  $\lambda \leq 0$  in Einstein's equations (2), is

$$E + \sum p_i \geq 0 \quad \text{and} \quad E + p_i \geq 0 \quad (i = 1, 2, 3) \quad (9)$$

Although this does not quite follow from the non-negative definiteness of energy, it is still a very reasonable assumption for "normal" matter. (Equivalent to Eq.(9) is:  $t^a t_a = 1$  implies  $T_{ab} t^a t_b > \frac{1}{2} T_c^c$ .) Much use has been made of Eq.(7) by Hawking, who has obtained several important theorems [9 - 11] predicting singularities, mainly in cosmological contexts. He has referred, descriptively, to Eq.(7) as stating that "gravitation is always attractive". A slight drawback to Eq.(7) is that its strict applicability requires (unlike Eq.(6)) that the cosmological constant  $\lambda$  be zero or negative. However, in regions where curvature is already large compared with  $\lambda$ , it is difficult to believe that the presence of a cosmological constant should seriously affect the subsequent discussion.

The theorem I wish to describe was obtained jointly with Hawking[13]. In its main implications it incorporates and somewhat generalizes the previous results. Ideally, one would like to be able to show that on the basis of Einstein's equations and reasonable physical assumptions, the space-time curvature becomes arbitrarily large (in an appropriate sense) in some region. Unfortunately none of the theorems to date directly shows this. Instead, we must content ourselves with a deduction that the space-time in question is not geodesically complete and, consequently, not extendible to a geodesically complete space-time. By a space-time, I shall here mean a pseudo-Riemannian four-manifold with a  $(+, -, -, -)$  signature, which is time-orientable. Such a manifold is timelike- and null-geodesically complete if every timelike or null geodesic can be extended to arbitrarily large values of its affine parameter. Physically, this means (crudely) that freely moving particles or photons cannot "just disappear" or "just appear" in the space-time.

Theorem: Space-time cannot be timelike- or null-geodesically complete if

- (i) the strong energy condition holds
- (ii) there are no closed timelike curves
- (iii) every timelike or null geodesic (with tangent vector  $t^a$ ) encounters a region where

$$t_{[a} R_{b]cd[e} t_f] t^c t^d \neq 0$$

(iv) there exists a compact achronal set with a trapped edge.

Assumption (ii) is the physically very reasonable requirement that it should be impossible for any observer to travel (or to send signals) into his own past. Assumption (iii) would be expected to hold in any physically realistic or "generic" space-time. Randomly oriented curvature (e.g. gravitational waves, no matter how weak) would be sufficient to ensure that assumption (iii) holds.

The physically important assumption of the theorem is (iv). An achronal set (sometimes referred to as a semi-spacelike set [18]) is a

set of points no two of which can be connected by a strictly timelike curve (e.g. a single point or, in normal circumstances, a spacelike or null hypersurface). The "edge" of such a set  $S$  is "trapped" if the null geodesics, lying on the boundary of the future of  $S$ , leave  $S$  at points at which they locally converge into the future (as with a trapped surface). It is not necessary here to appreciate the meaning of assumption (iv) in its full generality. Only the following three special cases will concern us, any one of which can be inserted in the theorem in place of (iv):

- (iva) there exists a trapped surface
- (ivb) there exists a compact spacelike hypersurface
- (ivc) all the past-directed null geodesics through some point start converging again.

If we use (iva), we effectively obtain the required implication that, on the basis of Einstein's equations, trapped surfaces result in singularities in space-time. Alternatively we can use (ivb), which is essentially the statement that the universe should be spatially closed. Again the implication is that space-time singularities should occur. The final alternative (ivc) has relevance to observational questions. The point referred to can be taken to represent the earth at the present moment. The past-directed null geodesics represent the portion of the universe that we can "see". Roughly speaking, (ivc) states that the solid angle subtended by an object of given size will have a minimum value. Hawking and Ellis [12] have suggested that the present observations of the 3°K "black-body" radiation indicate that (ivc) is satisfied in our universe. I believe Sciama discusses this in greater detail in his lecture [19]. Again, space-time singularities (this time in the past) are to be inferred.

I cannot go into the proof of the theorem here. The methods used involve differential topology; also certain arguments rather specific to the study of space-times have had to be developed. These are very much the realm of the pure relativist.

It is one of the ironies of the subject<sup>2</sup> that by obtaining results of this kind, our pure relativist pulls the rug from beneath him! Now he can no longer regard his space-time as something remote from the rest of physics. And if he is to have any hope at all of appreciating the nature of one of his "singularities", or of the neighbouring space-time, he will presumably have to descend his ivory tower into the murky depths of quantum theory and particle physics - and even this may well not be sufficient!

## REFERENCES

- [1] BIRKHOFF, G.D., Relativity and Modern Physics, Harvard Univ. Press, Cambridge, Mass. (1923).
- [2] BOYER, R.H., LINDQUIST, R.W., J. math. Phys. 8 2 (1967) 265.
- [3] BRANS, C., DICKE, R.H., Phys. Rev. 124 (1961) 925.
- [4] CHANDRASEKHAR, S., R. astr. Soc. monthly Notices 95 (1935) 207.
- [5] DICKE, R.H., Phys. Rev. 125 (1962) 2163.

---

<sup>2</sup> There is, perhaps, a twin irony that our gravitational optimist cannot "see" (and therefore cannot "use") the singularity that he had commissioned the pure relativist to find for him inside the collapsing star. But I do not feel that this is quite valid, since the gravitational optimist has also been presented with a "visible" singularity in the remote past.

- [6] DICKE, R.H., these Proceedings.
- [7] EDDINGTON, A.S., Nature 113 (1924) 192.
- [8] FINKELSTEIN, D., Phys. Rev. 110 (1958) 965.
- [9] HAWKING, S.W., Proc. R. Soc. A294 (1966) 511.
- [10] HAWKING, S.W., Proc. R. Soc. A295 (1966) 490.
- [11] HAWKING, S.W., Proc. R. Soc. A300 (1967) 187.
- [12] HAWKING, S.W., ELLIS, G.F.R., Astrophys. J. 152 (1968) 25.
- [13] HAWKING, S.W., PENROSE, R., to be published.
- [14] ISRAEL, W., Phys. Rev. 164 (1967) 1776.
- [15] KERR, R.P., Phys. Rev. Lett. 11 (1963) 237.
- [16] NEWMAN, E.T., COUCH, E., CHINNAPARED, K., EXTON, A., PRAKASH, A., TORRENCE, R., J. math. Phys. 6 (1965) 918.
- [17] PENROSE, R., Phys. Rev. Lett. 14 (1965) 57.
- [18] PENROSE, R., "Structure of space-time", in Battelle Rencontres, Seattle, 1967 (De WITT, C., WHEELER, J.A., Eds), Benjamin, New York (1968).
- [19] SCIAMA, D.W., these Proceedings.

# SINGULARITY IN THE GENERAL SOLUTION OF THE GRAVITATIONAL EQUATIONS

E. M. LIFSHITZ

Institute for Theoretical Physics,  
Moscow, USSR

## Short contribution

The problem of the existence of a singularity in the general solution of the gravitational equations is of great importance for relativistic cosmology, the more it can be stated in a precise form in the framework of general relativity.

For several years, this problem was considered by Khalatnikov and myself in the following analytic statement:

The general solution is the one which allows completely arbitrary conditions (distribution of matter and gravitational field) at an initial moment of time. The criterion for the degree of generality of a solution is the number of arbitrary space functions contained in it ("physically arbitrary" functions are those the number of which cannot be diminished by any choice of the reference system). For a solution to be general it must contain eight such functions. To obtain the general solution in an exact form is, of course, an insoluble problem. Instead, we can state the problem as follows: Assuming the existence of a singularity one must find, in its neighbourhood, the widest class of solutions and then judge by the number of arbitrary functions whether it is the general solution.

An extensive search led to a general solution containing eight arbitrary functions. However, the singularity which the solution contained turned out not to be of a physical nature; this is due merely to the intersection of coordinate lines in the synchronous reference system we have used. The widest class of solutions with a physical singularity (density of the matter tending to infinity) turned out to contain only seven arbitrary functions (details can be found in our review paper: *Adv. Phys.* **12** (1963) 185).

We have thus been led to the conclusion that there is no general singularity in the cosmological solutions of the gravitational equations. Since this result was negative, the proof could of course not be of a decisive character. At that time we did not yet know of Penrose's remarkable theorem which indicates the opposite.

The mathematical reasoning of Penrose is of such a kind that, I must confess, I am unable to follow it satisfactorily enough for me to be completely convinced of its validity. At the same time, of course, I have no arguments against it.

Now what does this theorem imply in its physical aspects? It states that (if a trapped surface exists in an initial moment of time) in a future moment at least a part of the space must become singular. So the theorem would be satisfied already if a single star collapsed - which is evidently not what is meant by a cosmological singularity. Moreover, there are as yet no complete results as to the actual occurrence of trapped surfaces in the general case and the theorem also does not state what kind of singularity it predicts; does the density of matter become infinite or not?

So perhaps there is still no real contradiction between our results. Anyway, I should say that the question could be finally settled (in the positive sense) only by finding out the form of the general solution in the neighbourhood of the singularity. If this really exists, the problem must have a definite solution.

# REMARKS ON GRAVITATIONAL RADIATION

S. WEINBERG

Massachusetts Institute of Technology,

Cambridge, Mass.,

and

University of California,

Berkeley, Calif., United States of America

## Abstract

REMARKS ON GRAVITATIONAL RADIATION. The author discusses gravitational radiation and makes some observations about soft gravitons.

The chairman has said that this session should emphasize aspects of general relativity that are relevant to possible experiments, and has asked me to talk about cosmic black-body gravitational radiation. I can therefore give my talk in one sentence: There are no experimental aspects of cosmic black-body gravitational radiation. To fill up the rest of the time, I will explain why not, and will make a few general remarks about soft gravitons.

The possibility that the universe contains a good deal of gravitational radiation has been considered by Wheeler, Zeldovich, Weber, Zipoy, Winterberg, and probably many others. It is obvious that if the universe has expanded from a singularity then it must indeed contain thermal gravitational radiation described by the Planck distribution law. When the Robertson-Walker "radius"  $R(t)$  was very small then the matter of the universe was necessarily opaque to gravitons, and hence in thermal equilibrium with black-body gravitational radiation. If the universe became transparent to gravitons at a radius  $R_1$ , then the left-over gravitons still obey the Planck law, but with temperature red-shifted to  $T_g = T_1 (R_1/R_0)$ , where  $R_0$  is the present radius and  $T_1$  the temperature of matter at radius  $R_1$ . The universe subsequently became transparent to photons, say at a radius  $R_2$ , so we now see electromagnetic radiation with temperature  $T_\gamma = T_2 (R_2/R_0) \approx 3^\circ\text{K}$ . Hence the ratio of the present graviton and photon temperatures is simply given by

$$T_g/T_\gamma = R_1 T_1 / R_2 T_2$$

We therefore have to ask how  $RT$  varied when  $R$  increased from  $R_1$  to  $R_2$ . If during this period the specific heat of the universe was dominated by a fixed number of relativistic particles then  $RT$  was constant, and the graviton temperature should now be about  $3^\circ\text{K}$ . In the opposite extreme case, we might suppose that the number of particle types with mass less than  $m$  grows like  $m^A$ , in which case  $RT$  decreases with  $R$  like  $R^{-A/(A+3)}$ , and  $T_g$  should be much less than  $T_\gamma$ . As a reasonable compromise we may assume that the specific heat was dominated by a variable number  $N$  of

types of highly relativistic particles, in which case  $RT$  goes like  $N^{-1/4}$  and

$$T_g/T_\gamma \approx (N_2/N_1)^{1/4}$$

If  $N_1/N_2$  is about 100 then  $T_g$  is about  $1^\circ\text{K}$ .

It is slightly less obvious that the same conclusions hold even if the universe were not ever condensed enough to be opaque to gravitational radiation, provided we assume that it has oscillated between expansion and contraction for a very long time. (The "bounce" might be caused by a strong short-range attractive force among hadrons, which could violate the energy conditions discussed by Penrose.) In such cosmologies the probability of a given graviton's being absorbed in any one cycle might be small, but it will be absorbed if we wait for enough cycles to pass, and therefore the graviton population is the same as if the universe were opaque. This argument can be made more precise by computing the present graviton distribution function from the previous emission and absorption rates. Since the induced emission rate depends upon the number of gravitons present, we obtain in this way an integral equation, for which we can fortunately find a general solution. If  $RT$  is constant near the minimum of  $R$ , this solution is essentially just the Planck law, with temperature similar to that discussed above.

Unfortunately there does not seem to be any way to detect  $1^\circ\text{K}$  black-body gravitational radiation. Through inverse inner bremsstrahlung the electrons in a metal can be heated at a rate at most  $10^{-45}^\circ\text{K}/\text{s}$ . Through excitation of rotational levels a free molecule will absorb about  $10^{-57} \text{ eV/s}$ . The lowest vibrational mode of the big aluminium cylinder described by Weber picks up about  $10^{-51}$  watts, whereas  $10^{-23}$  watts is needed to give a discernible signal. The planet Mercury picks up kinetic energy at a rate which increases the radius of its orbit by  $10^{-75} \text{ cm/s}$ . And so on.

One other imaginable detection scheme, which also doesn't work, but is rather interesting in its own right, makes use of the effect due to Bose statistics of a thermal radiation background on the virtual gravitons which accompany all physical processes. You will recall that infra-red divergences make the radiative corrections due to soft virtual photons important if we try to measure reaction rates with energy resolution  $\Delta E/E$  small compared with  $e^{-1/B}$ , where  $B$  is an energy-dependent quantity generally of the order  $1/137$ . The same is true for gravitons, except that  $B$  is of the order  $10^{-17}$  at the highest cosmic-ray energies, and much less at lower energies. In both cases the effect of soft virtual quanta is very much enhanced if the process takes place in the presence of black-body radiation; we then find that radiative corrections become substantial when  $\Delta E$  is less than  $KTB$ . However, it is unlikely that we will be able to observe reactions at very high energy with an energy resolution of  $10^{-21} \text{ eV}$ .

This brings me to the remarks I wished to make about gravitational radiation in general. Professional general relativists have doubted the existence of exact wave solutions of Einstein's equations. How then can I have the effrontery not only to discuss the cosmic distribution of gravitons, but also to compute gravitational radiative corrections? It is because I have tacitly taken as my starting point not classical general relativity but special relativity and quantum mechanics. On this basis one can prove that long-range (i.e. inverse square) forces can only be

associated with massless particles of spin 0, 1 or 2. We know that gravitation is not transmitted solely by massless particles with  $j = 0$  and / or  $j = 1$ , so special relativity and quantum mechanics allow no alternative to the existence of a graviton with mass zero and spin two. Also, the same argument that leads to the theorem excluding long-range forces from particles with spin higher than two also leads to low-energy theorems (including the principle of equivalence itself) that underlie the calculations discussed above. Thus it seems rather beside the point to worry whether Einstein's equations yield exact wave solutions or whether classical general relativity can be quantized. The important open questions are whether we can construct a consistent dynamical theory of particles with mass zero and spin two, and whether this theory will then agree with general relativity in the classical limit. There is no reason to doubt that both questions will be answered affirmatively, and in the meanwhile there are available enough low-energy theorems to handle problems of the sort I have been describing.



# HAMILTONIAN DYNAMICS AND POSITIVE ENERGY IN GENERAL RELATIVITY\*

S. DESER

Physics Department, Brandeis University,  
Waltham, Mass., United States of America

## Abstract

HAMILTONIAN DYNAMICS AND POSITIVE ENERGY IN GENERAL RELATIVITY. A review is first given of the Hamiltonian formulation of general relativity; the gravitational field is a self-interacting massless spin-two system within the framework of ordinary Lorentz covariant field theory. The recently solved problem of positive-definiteness of the field energy is then discussed. The latter, a conserved functional of the dynamical variables, is shown to have only one extremum, a local minimum, which is the vacuum state (flat space). This implies positive energy for the field, with the vacuum as ground-state. Similar results hold when minimally coupled matter is present.

The new results I shall report on are concerned with the physical properties of the Hamiltonian in general relativity, namely with the recent demonstrations [1, 2] (in collaboration with D.R. Brill and L.D. Faddeev) that the energy of the gravitational field is positive. It follows also that the ground-state is the vacuum (flat space), and that the four-momentum  $P^\mu$  of a gravitating system is a positive time-like vector; all these conclusions are required physically for normal systems in Lorentz covariant theories.

In view of the general aims of this symposium, however, it seems worthwhile to first establish why these questions, and indeed the very notion of energy, are important in what is at first sight a purely geometrical system, devoid of such familiar dynamical concepts. I shall therefore begin with a brief analysis of general relativity in the spirit of usual Lorentz covariant field theory, in order to indicate how well, in fact, gravitation fits into the conceptual framework of the rest of physics. This "point of view", which is about ten years old, seems to me to have been a particularly fruitful approach to general relativity. It has allowed us to use experience and intuition accumulated elsewhere, particularly in electrodynamics, and provides, especially for non-relativists, a detailed key to the basic concepts of the theory, by translating the (sometimes mysterious) geometry into more familiar dynamics.

To review the Hamiltonian analysis of general relativity, I shall use an approach developed by Arnowitt, Misner and myself [3] independently and at about the same period as that which Professor Dirac [4] has sketched in his lecture. The analysis deals entirely with bounded, asymptotically flat systems, and is therefore entirely non-cosmological; it exhibits the gravitational field as a normal massless spin-two field, which happens to be highly self-interacting. Let us set up a brief electrodynamics-gravitation dictionary to provide a rapid orientation. Both theories are based on:

---

\* Work supported by U.S.A.F. OAR under OSR Grant AF 368-67.

1. A universal minimal coupling:  $\partial \rightarrow \partial - ie\Lambda \leftrightarrow \partial \rightarrow \nabla$ . The covariant derivative ( $\nabla$ ) prescription may even, as for the Yang-Mills field, be used to generate the gravitational self-interaction, namely the Einstein equations themselves from their weak-field, linear, form. Associated with universality is:

2. Gauge invariance

$$\delta A = \partial \Lambda, \quad \partial \psi = ie \Lambda \delta \psi \leftrightarrow \text{co-ordinate transformations of metric and matter variables}$$

which necessarily implies that not all variables are dynamical, but that some are:

3. Gauge variables

$$A_0 \leftrightarrow g_{0\mu}$$

A further consequence of gauge invariance is the existence of:

4. Constraint equations

$$(\vec{\nabla} \cdot \vec{E} - \rho) = 0 \leftrightarrow (G_\mu^0 - \kappa T_\mu^0) = 0$$

which express certain field components as functions of the others, rather than being normal time evolution equations involving second time derivatives. Consistency of the constraints with the time development equations is guaranteed by the:

5. Bianchi identities

$$\partial_\mu (\partial_\nu F^{\mu\nu}) \equiv 0 \leftrightarrow \nabla_\mu G^{\mu\nu} \equiv 0$$

which also imply corresponding conservation requirements on the sources. The constraints in both theories have the form of the divergence of a variable equal to a source, and naturally lead to associated flux integrals over the:

6. Constraint variables

$$\vec{E}_{i,i} \leftrightarrow (g_{ij,ij} - g_{ii,jj})$$

where the right side (corresponding to the longitudinal electric field) is schematic for the four metric constraint or Newtonian components involved in the four constraint equations; we shall return to them. These flux integrals, defining the generators of associated gauge transformations, are guaranteed to be conserved by the identities. They define the

7. Charge  $Q \leftrightarrow$  energy-momentum  $P^\mu$

$$Q \equiv \int d^3r \rho = \oint \vec{E} \cdot d\vec{S} \leftrightarrow 16\pi m \equiv \int d^3r \mathcal{H} = \oint (g_{ij,j} - g_{jj,i}) dS_i$$

using the traditional notation  $m$  for the energy of the field. Here we have exhibited both the surface integral forms and the volume integrals over the source (charge density  $\rho \leftrightarrow$  energy density  $\mathcal{H}$ ). Similar forms exist for the spatial translations  $\vec{P}$ .

Finally, in addition to the gauge and constraint variables, there are the  
8. Unconstrained dynamical (radiation field) variables

$$(\vec{A}^T, \vec{E}^T) \leftrightarrow (g_{ij}^{TT}, \pi^{ijTT})$$

In electrodynamics they are the two transverse ( $\vec{\nabla} \cdot \vec{A}^T \equiv 0 \equiv \vec{\nabla} \cdot \vec{E}^T$ ) pairs of conjugate variables; likewise in gravitation where the six pairs ( $g_{ij}^{TT}, \pi^{ijTT}$ ) satisfy the four requirements of transversality and tracelessness. This is a consequence of masslessness in both cases.

This highly schematic dictionary is actually a rather accurate representation, in each case, of the Hamiltonian form of the theory in appropriate (radiation) gauges. We may perhaps insert a parenthetic comment on quantization here. If one is less ambitious than Professor Dirac, and does not seek a closed form quantization prescription, but more optimistic than Professor Schwinger in allowing operators, it is possible to discuss quantization within this framework, at least in a perturbation sense, though problems common to any non-linear field theory (plus some new ones) are certainly present. Returning to our electrodynamics-gravitation parallelism, it is clear that there will also be differences, owing to the different nature of the invariances involved, particularly through the non-abelian and space-time aspects of the co-ordinate group. This is most strikingly brought out by the following simple form of the Einstein action

$$I_E \equiv \int d^4x \mathcal{R} = \int d^4x \{ \pi^{ij} \partial_0 g_{ij} - N_\mu R^\mu(g_{ij}, \pi^{ij}) \}$$

when written in terms of the spatial metric components  $g_{ij}$ , their time derivatives  $\pi^{ij}$ , the gauge components  $N_\mu$  which are just the  $g_{0\mu}$ , and the  $R^\mu$  which are essentially the  $G_\mu^0$  and depend only on the Cauchy data ( $g_{ij}, \pi^{ij}$ ). Here is an apparently Hamiltonian,  $L = p\dot{q} - H(p, q)$ , form with the one difference that the "Hamiltonian"  $N_\mu R^\mu$  vanishes by virtue of the constraints  $R^\mu(g, \pi) = 0$  obtained by varying the Lagrange multipliers  $N_\mu$  (the remaining variations yield the rest of the Einstein equations  $G_{\mu\nu} = 0$ ). Physically, the "Hamiltonian", not having any well-defined time with respect to which it generates translations - because of co-ordinate invariance - vanishes. Precisely the same thing occurs, and for the same reason, in the "parameterized" form of classical mechanics when the time is made a dynamical variable and an arbitrary new parameter is introduced as the independent variable. (A partial analogue of this situation also exists in electrodynamics; where the constraint term  $A_0(\vec{\nabla} \cdot \vec{E} - \rho)$  in the Hamiltonian ( $A_0$  is a Lagrange multiplier) also vanishes.) How then is the notion of Hamiltonian and energy to be reinstated? Clearly, one must make a choice of time co-ordinate, but for every such choice the time translation generator will have a different form. On the other hand, its value, the energy, must be the same for any given physical system or geometry if it is to be of any use. Further, the energy must hopefully have all the "good" properties associated with it in usual theory, particularly positive definiteness. This requirement is essential classically, even aside from quantization. For there is no freedom of translating away the zero point of energy in general relativity on the one hand, and on the other,

existence of solutions with any finite negative energy can be shown to imply an unbounded negative spectrum.

The properties and various definitions of gravitational field energy-momentum in general relativity would take too long to develop here [5]. It is possible to show that (almost) all definitions proposed for energy of asymptotically flat systems (a) agree with each other, and (b) have the same value in all co-ordinate frames which approach the same cartesian co-ordinates at infinity. More precisely, the energy, being conserved, may be given as a flux integral over a closed spatial 2-surface at infinity at any instant provided the metric and its first derivatives behave respectively as  $\eta_{\mu\nu} + O(1/r)$  and  $O(1/r^2)$  at spacelike infinity (when there is radiation, one may still have  $\partial_\alpha g_{\mu\nu} \sim O(1/r)$  on the light cone). This requirement is partly one on the dynamical modes (they must fall off sufficiently rapidly that the energy is finite) and partly one on the choice of co-ordinates: they must be asymptotically cartesian, as is also required of co-ordinates used in flat space to discuss energy. Any change of interior co-ordinates clearly leaves a surface integral unchanged, while asymptotic Lorentz rotations may be shown to transform the various  $P^\mu$  components like a four-vector. We list two forms for  $m$  here, as given in Refs [4, 3] respectively:

$$16\pi m = \oint (gg^{ij})_{,j} dS_i = \oint (g_{jj,i} - g_{ij,j}) dS_i$$

in units in which  $16\pi\gamma=1$  ( $\gamma$  the Newtonian constant). Each form is invariant under both linear gauge transformations ( $\delta g_{\mu\nu} = \partial_\mu \xi_\nu + \partial_\nu \xi_\mu$ ) and interior co-ordinate changes respecting the asymptotic conditions above.

So much for the definition and form of the energy; to show that it is always positive is an altogether different story. For, just as the charge is given by the flux form  $\oint \vec{E} \cdot d\vec{S}$ , its value must be gotten the hard way by summing over the charge density in all space,  $\int d^3r \rho$ . It is here that non-linearity raises its ugly head; even in the absence of matter, the energy is obtained in principle by solving the highly non-linear constraints

$G_\mu^0(g, \pi) = 0$  for the variables  $(g_{ij,ij} - g_{ii,jj})$  or  $(gg^{ij})_{,ij}$  which are the linear terms in  $G_\mu^0$ . This can at best be done by an iteration process, yielding a series expansion, and for this reason direct attempts to prove positiveness have not succeeded except for some particular classes of solutions. Before describing the recent solutions to the positiveness problem, which are based on a variational approach, it may be useful to give an intuitive argument for positiveness, despite the attractive character of the long-range Newtonian potential. Schematically, Newtonian theory gives for the total energy,  $m_N$ , of a system,  $m_N = m_0 - \gamma m_E^2/2r$  where  $m_0$  is the mass in the absence of gravitation and  $r$  a characteristic dimension of the system. This can indeed turn negative for sufficient density, owing to the negative Newtonian potential energy. In Einstein theory, this formula is replaced by  $m_E = m_0 - \gamma m_E^2/2r$  (since the equivalence principle says all energy has gravitational effect);  $m_E$  clearly always remains positive if  $m_0$  is.

What has been demonstrated by the variational method is that the energy  $m$  of the gravitational field, as a functional of the field configuration, i.e. of the geometry, has the following properties:

(A). The only extremum of  $m$ , namely the "point" at which  $\delta m = 0$ , is vacuum, i.e. flat space.

(B) The extremum is a minimum; near flat space the second variation  $\delta^2 m$  is positive-definite.

From this we conclude that  $m$  is positive and vanishes only at flat space (the converse, that  $m = 0$  for flat space, is of course a trivial consequence of any reasonable definition of energy). A point of rigour should perhaps be mentioned. Our conclusions follow, strictly speaking, for a function of a finite number of variables varying over an appropriately complete domain. There, rigorous theorems exist for any finite number of variables (it is of course trivial for a function of one variable). No corresponding rigorous results exist as yet for functionals.

Let me indicate briefly the idea of the derivations. The first, developed in collaboration with Brill [1], simply varies the energy functional with respect to all the Cauchy variables ( $g_{ij}$ ,  $\pi^{ij}$ ) and then proceeds to eliminate the variations of the constraint variables in terms of those of the unconstrained ones. That is, we write

$$\delta_{TOT} m = \int d^3 r \{ \delta m / \delta g (\delta_U g + \delta_C g) + \delta m / \delta \pi (\delta_U T + \delta_C \pi) \}$$

where  $\delta_U$ ,  $\delta_C$  are the free and constrained variations, related by the requirement that  $(\delta_U + \delta_C) G_\mu^0 = 0$ , since the constraints must be respected under variation, so that only systems obeying the Einstein equations are compared. Elimination of  $\delta_C$  in favour of  $\delta_U$  then yields the Euler-Lagrange equations for the extremum as the coefficients of  $\delta_U g_{ij}$  and  $\delta_U \pi^{ij}$ . The latter then turn out to characterize the extremum as flat space. The above procedure can be carried through because we need not solve the constraints but deal only with the - more linear - set of varied relations  $\delta G_\mu^0 = 0$ .

A more intuitive approach, which makes more direct use of the Hamiltonian form of the Einstein equations is that due to Brill, Faddeev and myself [2]. Here we ensure that the constraints are obeyed under variation simply by adding them to the mass functional with Lagrange multipliers

$$16\pi m = \oint (gg^{ij})_{,j} dS_i + \int d^3 r N_\mu R^\mu$$

Now everything may be varied freely; this means first of all that we may drop the surface term itself since all variations can now be taken to drop off sufficiently rapidly at infinity. (In the previous method, these Newtonian components, along with their variations, were the whole story and necessarily dropped only as  $1/r^2$ .) This  $\delta m$  is simply given by  $\delta / \delta d^3 r N_\mu R^\mu$ . But the latter variation is already known to us in a different context: recall the form (1) for the Einstein action. It differs from our "action" only in that the kinetic  $\pi^{ij} \partial_0 g_{ij}$  terms are missing. Thus we know that our Euler-Lagrange equations are just the Einstein equations, at any instant, with all time derivatives set to zero. But one may show that the only solution of these "stationary" (in the sense of time-independent) equations, namely

$$\delta "H" / \delta g = 0 = \delta "H" / \delta \pi, \quad "H" \equiv \int d^3 r N_\mu R^\mu.$$

is flat space. We may also include matter sources in our problem in the same way. The latter involve an addition to the Einstein action of the form

$$I_M = \int d^4x [\Sigma_A \pi^A \partial_0 \phi_A - N_\mu \mathcal{T}_\mu^0(\pi^A, \phi_A, g_{ij})]$$

where  $(\pi^A, \phi_A)$  are the matter variables,  $\mathcal{T}_\mu^0$  the (covariant) energy-momentum density. While the definition of the mass remains unchanged in the presence of matter, the constraints clearly become  $(G_\mu^0 - \kappa T_\mu^0) = 0$  and our "mass functional" is now  $\int N_\mu (G_\mu^0 - \kappa T_\mu^0) d^3r$ . The Euler-Lagrange equations are thus the combined Einstein-matter field equations with all time derivatives set to zero. Now in the absence of gravitation, the equations  $\delta H_M / \delta \pi^A = 0 = \delta H_M / \delta \phi_A$  are satisfied only by the vacuum state for normal systems - there are no other time-independent solutions of, say, the Maxwell or Klein-Gordon equations. This is basically due to the elliptic nature of these equations when time derivatives are dropped. With minimal coupling to gravitation, this remains unchanged, and the elliptic equations in an external gravitational field still have no non-trivial solutions ( $\nabla^2 \phi = 0$  and  $\nabla^\mu \nabla_\mu \phi = 0$  are alike in having no solutions). Once the conclusion that  $\pi^A = 0 = \phi_A$  is drawn, the problem reduces to the source-free one, which has been solved.

Finally, we must show that small matter-gravitation excitations about the vacuum state are positive-definite. This is not very difficult; for this second variation,  $\delta^2 m$ , breaks up into a sum of uncoupled gravitational and matter contributions. The former is essentially that of a source-free linear, spin-two field and may be shown to be positive. The latter is the energy of weak matter excitations in flat space. If the matter is normal, i.e. has positive flat-space energy, then we clearly have

$$\delta_{TOT}^2 m|_{VAC} = \delta_{GRAV}^2 m + \delta_{MATT}^2 m > 0$$

as there are no cross-terms at vacuum. (Of course, if we perversely bring in "negative-energy" matter then vacuum is a saddle point due to the negative character attributed to the sources.)

We have thus shown, subject to as yet unknown questions of rigour regarding variational properties of functionals, that the energy of a gravitational system is positive and vanishes only for the vacuum which is thus also the ground-state. It follows also, since the derivation holds in any Lorentz frame, that the invariant mass has the property that

$$M^2 \equiv -P_\mu P^\mu > 0$$

and so all solutions have positive time-like four-momentum. This may also be demonstrated directly by similar variational techniques [6]. Thus the gravitational field has all the required energy properties of a legitimate special relativistic field, a fact which is reassuring also for quantization considerations.

## REF E R E N C E S

- [1] BRILL, D.R., DESER, S., Phys. Rev. Lett. 20 (1968) 75.
- [2] BRILL, D.R., DESER, S., FADDEEV, L.D., Physics Lett. 26A (1968) 538.
- [3] ARNOWITT, R., DESER, S., MISNER, C.W., "The dynamics of general relativity", Gravitation: An Introduction to Current Research (WITTEN, L., Ed.), Wiley, New York (1962).
- [4] DIRAC, P.A.M., Proc. R. Soc. A246 (1958) 333; Phys. Rev. 114 (1959) 924.
- [5] See for example ARNOWITT, R., DESER, S., MISNER, C.W., Phys. Rev. 122 (1961) 997.
- [6] DESER, S., Nuovo Cim. 55B (1968) 593.



INTERNATIONAL SYMPOSIUM  
ON CONTEMPORARY PHYSICS,  
TRIESTE, 7-28 JUNE 1968

Director of the Symposium: ABDUS SALAM  
Scientific Secretary: L. FONDA

MONITORS

G. BURBIDGE: Dept. of Physics, University of California,  
La Jolla, Calif., USA

D. PINES: Centre for Advanced Study,  
University of Illinois,  
Urbana, Ill., USA

M. ROSENBLUTH: Institute for Advanced Study,  
Princeton, N.J., USA

Chairman of Local Committee: P. BUDINI, ICTP

Scientific Information Officer: A. M. HAMENDE, ICTP

Administrative Officer: P. RENDI, ICTP

*A list of participants appears in Vol. II.*

## AUTHOR INDEX

Numerals refer to the first page of a paper.

- |                           |                        |
|---------------------------|------------------------|
| Abrikosov, A.: 97         | Lifshitz, E. M.: 557   |
| Anderson, P. W.: 47       | Martin, P. C.: 123     |
| Bolton, J. G.: 471        | Montroll, E.: 177, 273 |
| Burbidge, E. M.: 347, 497 | Morrison, P.: 393      |
| Burke, B. F.: 401         | Penrose, R.: 545       |
| Cameron, A. G. W.: 475    | Pethick, C.: 93        |
| Coppi, B.: 249            | Pines, D.: 3           |
| de Gennes, P. G.: 195     | Prigogine, I.: 315     |
| Deser, S.: 563            | Rosenbluth, M. N.: 205 |
| Dicke, R. H.: 507, 515    | Rossi, B.: 371         |
| Dirac, P. A. M.: 539      | Salpeter, E. E.: 335   |
| Doniach, S.: 87           | Schmidt, M.: 467       |
| Dupree, T.: 221           | Schrieffer, J. R.: 55  |
| Ferrell, R. A.: 129       | Sciama, D. W.: 489     |
| Fisher, Michael E.: 19    | Smith, F. G.: 417, 459 |
| Fock, V. A.: 511          | Suhl, H.: 157          |
| Fowler, W. A.: 359        | Thirring, W.: 485      |
| Gold, T.: 477             | Thompson, W. B.: 237   |
| Keller, J. B.: 257        | Townes, C. H.: 295     |
| Khalatnikov, I. M.: 71    | Weber, J.: 533         |
| Kraft, R. P.: 449, 459    | Weinberg, S.: 559      |
| Lieb, E. H.: 163          |                        |



## IAEA SALES AGENTS AND BOOKSELLERS

Orders for Agency publications can be placed with your bookseller or any of the addresses listed below:

### ARGENTINA

Comisión Nacional de  
Energía Atómica  
Avenida del Libertador 8250  
Buenos Aires

### GERMANY, Fed. Rep.

R. Oldenbourg Verlag  
Rosenheimer Strasse 145  
D-8 Munich 80

### AUSTRALIA

Hunter Publications  
23 McKillop Street  
Melbourne, C.1

### HUNGARY

Kultura  
Hungarian Trading Company  
for Books and Newspapers  
P.O. Box 149  
Budapest 62

### AUSTRIA

Publishing Section  
International Atomic Energy Agency  
Kärntner Ring 11  
P.O. Box 590  
A-1011 Vienna

### ISRAEL

Heiliger & Co.  
3, Nathan Strauss Str.  
Jerusalem

### BELGIUM

Office International  
de Librairie  
30, Avenue Marnix  
Brussels 5

### ITALY

Agenzia Editoriale Commissionaria  
A.E.I.O.U.  
Via Meravigli 16  
I-20123 Milan

### CANADA

Canadian Government Printing Bureau,  
International Publications Main Stores,  
Room 2738, Second Floor,  
Sacred Heart Boulevard,  
Hull, Québec

### JAPAN

Maruzen Company, Ltd.  
P.O. Box 5050,  
100-31 Tokyo International

### C.S.S.R.

S.N.T.L.  
Spolena 51  
Nové Mesto  
Prague 1

### MEXICO

Librerfa Internacional, S.A.  
Av. Sonora 206  
México 11, D.F.

### DENMARK

Ejnar Munksgaard Ltd.  
6 Norregade  
DK-1165 Copenhagen K

### NETHERLANDS

Martinus Nijhoff N. V.  
Lange Voorhout 9  
P.O. Box 269  
The Hague

### FRANCE

Office International de  
Documentation et Librairie  
48, Rue Gay-Lussac  
F-75 Paris 5e

### PAKISTAN

Mirza Book Agency  
65, Shahrah Quaid-E-Azam  
P.O. Box 729  
Lahore - 3

**POLAND**

Ars Polona  
Centrala Handlu Zagranicznego  
Krakowskie Przedmiescie 7  
Warsaw

**ROMANIA**

Cartimex  
3-5 13 Decembrie Street  
P.O. Box 134-135  
Bucarest

**SPAIN**

Librería Bosch  
Ronda Universidad 11  
Barcelona - 7

**S W E D E N**

C. E. Fritzes Kungl. Hovbokhandel  
Fredsgatan 2  
Stockholm 16

**S W I T Z E R L A N D**

Librairie Payot  
Rue Grenus 6  
CH-1211 Geneva 11

**U. S. S. R.**

Mezhdunarodnaya Kniga  
Smolenskaya-Sennaya 32-34  
Moscow G-200

**U. K.**

Her Majesty's Stationery Office  
P.O. Box 569  
London S.E.1

**U. S. A.**

UNIPUB, Inc.  
P.O. Box 433  
New York, N.Y. 10016

**Y U G O S L A V I A**

Jugoslovenska Knjiga  
Terazije 27  
Belgrade

IAEA Publications can also be purchased retail at the United Nations Bookshop at United Nations Headquarters, New York, from the news-stand at the Agency's Headquarters, Vienna, and at most conferences, symposia and seminars organized by the Agency.

In order to facilitate the distribution of its publications, the Agency is prepared to accept payment in UNESCO coupons or in local currencies.

Orders and inquiries from countries not listed above may be sent to:

Publishing Section  
International Atomic Energy Agency  
Kärntner Ring 11  
P.O. Box 590  
A-1011 Vienna, Austria





**INTERNATIONAL  
ATOMIC ENERGY AGENCY  
VIENNA, 1969**

**PRICE: US \$13.00  
Austrian Schillings 336,-  
(\$5.8.4; F.Fr. 63,70; DM 52,-)**

**SUBJECT GROUP: III  
Physics, Plasma Physics  
and Electronics**