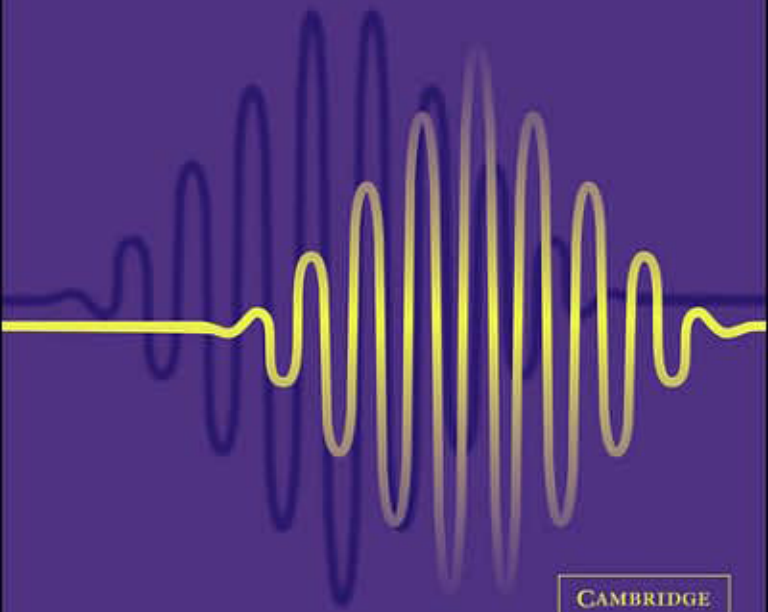
J. F. James

Second Edition

A Student's Guide to Fourier Transforms

With Applications in Physics and Engineering



CAMBRIDGE

This page intentionally left blank

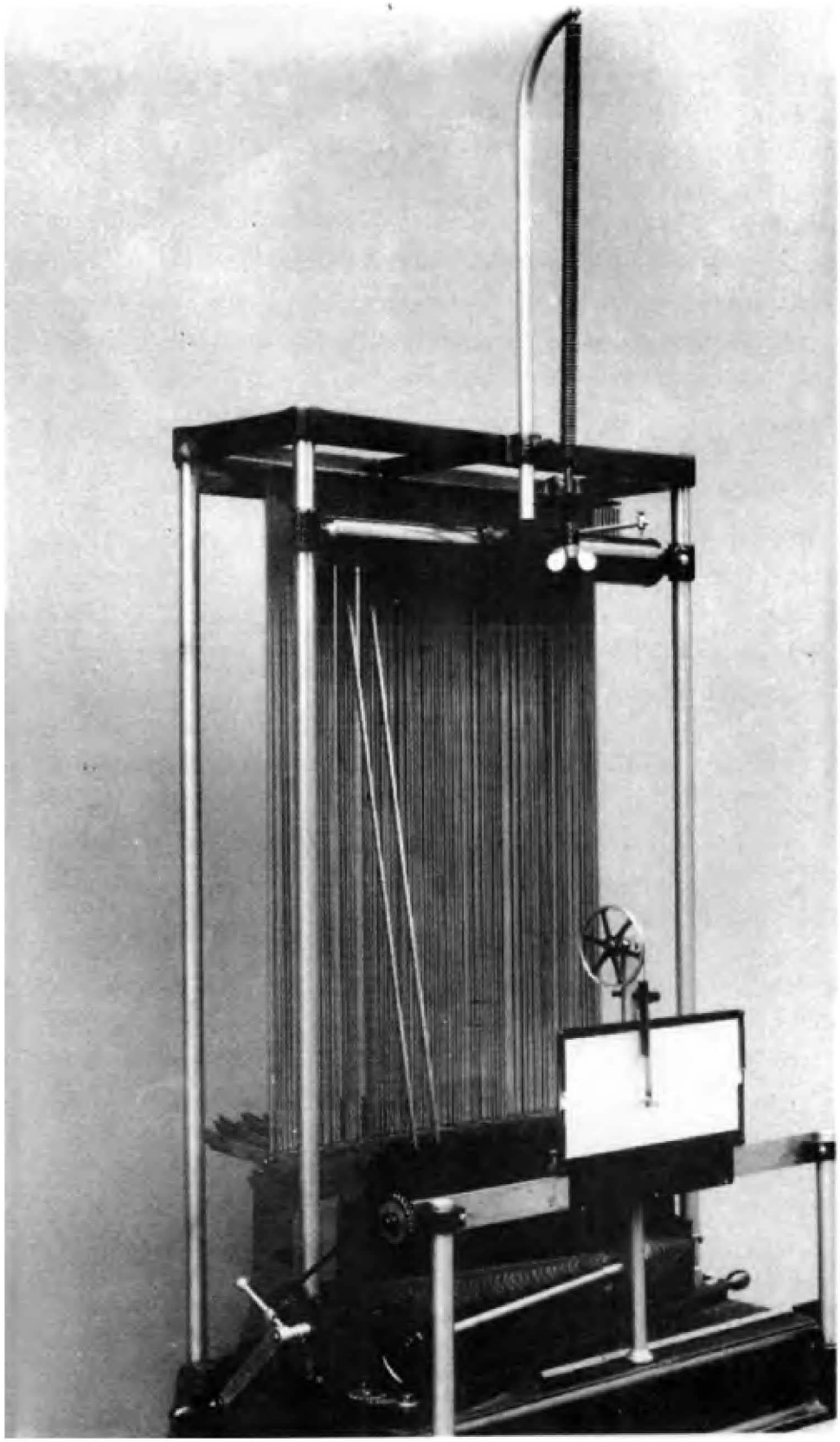
A Student's Guide to Fourier Transforms

Fourier transform theory is of central importance in a vast range of applications in physical science, engineering, and applied mathematics. This new edition of a successful undergraduate text provides a concise introduction to the theory and practice of Fourier transforms, using qualitative arguments wherever possible and avoiding unnecessary mathematics.

After a brief description of the basic ideas and theorems, the power of the technique is then illustrated by referring to particular applications in optics, spectroscopy, electronics and telecommunications. The rarely discussed but important field of multi-dimensional Fourier theory is covered, including a description of computer-aided tomography (CAT-scanning). The final chapter discusses digital methods, with particular attention to the fast Fourier transform. Throughout, discussion of these applications is reinforced by the inclusion of worked examples.

The book assumes no previous knowledge of the subject, and will be invaluable to students of physics, electrical and electronic engineering, and computer science.

JOHN JAMES has held teaching positions at the University of Minnesota, the Queen's University Belfast and the University of Manchester, retiring as Senior Lecturer in 1996. He is currently an Honorary Research Fellow at the University of Glasgow, a Fellow of the Royal Astronomical Society and Member of the Optical Society of America. His research interests include the invention, design and construction of astronomical instruments and their use in astronomy, cosmology and upper-atmosphere. Dr James has led eclipse expeditions to Central America, the Central Sahara, Java and the South Pacific islands. He is the author of about 40 academic papers and co-author with R. S. Sternberg of *The Design of Optical Spectrometers* (Chapman & Hall, 1969).



The Harmonic integrator, designed by Michelson and Stratton (see p. 72). This was the earliest mechanical Fourier transformer, built by Gaertner & Co. of Chicago in 1898. (Reproduced by permission of The Science Museum/Science & Society Picture Library.)

A Student's Guide to Fourier Transforms

with applications in physics and engineering

Second Edition

J. F. JAMES

Honorary Research Fellow,
The University of Glasgow

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 2RU, United Kingdom

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org

Information on this title: www.cambridge.org/9780521808262

First published in print format 2002

ISBN-13 978-0-511-07802-6 eBook (NetLibrary)
ISBN-10 0-511-07802-1 eBook (NetLibrary)

ISBN-13 978-0-521-80826-2 hardback
ISBN-10 0-521-80826-X hardback

ISBN-13 978-0-521-00428-2 paperback
ISBN-10 0-521-00428-4 paperback

Contents

Preface to the first edition

Showing a Fourier transform to a physics student generally produces the same reaction as showing a crucifix to Count Dracula. This may be because the subject tends to be taught by theorists who themselves use Fourier methods to solve otherwise intractable differential equations. The result is often a heavy load of mathematical analysis.

This need not be so. Engineers and practical physicists use Fourier theory in quite another way: to treat experimental data, to extract information from noisy signals, to design electrical filters, to 'clean' TV pictures and for many similar practical tasks. The transforms are done digitally and there is a minimum of mathematics involved.

The chief tools of the trade are the theorems in Chapter 2, and an easy familiarity with these is the way to mastery of the subject. In spite of the forest of integration signs throughout the book there is in fact very little integration done and most of that is at high-school level. There are one or two excursions in places to show the breadth of power that the method can give. These are not pursued to any length but are intended to whet the appetite of those who want to follow more theoretical paths.

The book is deliberately incomplete. Many topics are missing and there is no attempt to explain everything: but I have left, here and there, what I hope are tempting clues to stimulate the reader into looking further; and of course, there is a bibliography at the end.

Practical scientists sometimes treat mathematics in general and Fourier theory in particular, in ways quite different from those for which it was invented[1]. The late E. T. Bell, mathematician and writer on mathematics, once described mathematics in a famous book title as 'The Queen and Servant of Science'.

[1] It is a matter of philosophical disputation whether mathematics is invented or discovered. Let us compromise by saying that theorems are discovered; proofs are invented.

The queen appears here in her role as servant and is sometimes treated quite roughly in that role, and furthermore, without apology. We are fairly safe in the knowledge that mathematical functions which describe phenomena in the real world are 'well-behaved' in the mathematical sense. Nature abhors singularities as much as she does a vacuum.

When an equation has several solutions, some are discarded in a most cavalier fashion as 'unphysical'. This is usally quite right[2]. Mathematics is after all only a concise shorthand description of the world and if a position-finding calculation based, say, on trigonometry and stellar observations, gives two results, equally valid, that you are either in Greenland or Barbados, you are entitled to discard one of the solutions if it is snowing outside. So we use Fourier transforms as a guide to what is happening or what to do next, but we remember that for solving practical problems the blackboard-and-chalk diagram, the computer screen and the simple theorems described here are to be preferred to the precise tedious calculations of integrals.

Manchester, January 1994 J. F. James

[2] But Dirac's Equation, with its positive and negative roots, predicted the positron.

Preface to the second edition

This edition follows much advice and constructive criticism which the author has received from all quarters of globe, in consequence of which various typos and misprints have been corrected and some ambiguous statements and anfractuosities have been replaced by more clear and direct derivations. Chapter 7 has been largely rewritten to demonstrate the way in which Fourier transforms are used in CAT-scanning, an application of more than usual ingenuity and importance: but overall this edition represents a renewed effort to rescue Fourier transforms from the clutches of the pure mathematicians and present them as a working tool to the horny-handed toilers who strive in the fields of electronic engineering and experimental physics.

Glasgow, January 2001 J. F. James

Chapter 1

Physics and Fourier transforms

1.1 The qualitative approach

Ninety percent of all physics is concerned with vibrations and waves of one sort or another. The same basic thread runs through most branches of physical science, from accoustics through engineering, fluid mechanics, optics, electromagnetic theory and X-rays to quantum mechanics and information theory. It is closely bound to the idea of a *signal* and its *spectrum*. To take a simple example: imagine an experiment in which a musician plays a steady note on a trumpet or a violin, and a microphone produces a voltage proportional to the the instantaneous air pressure. An oscilloscope will display a graph of pressure against time, $F(t)$, which is periodic. The reciprocal of the period is the frequency of the note, 256 Hz, say, for a well-tempered middle C.

The waveform is not a pure sinusoid, and it would be boring and colourless if it were. It contains 'harmonics' or 'overtones': multiples of the fundamental frequency, with various amplitudes and in various phases[1], depending on the timbre of the note, the type of instrument being played and on the player. The waveform can be *analysed* to find the amplitudes of the overtones, and a list can be made of the amplitudes and phases of the sinusoids which it comprises. Alternatively a graph, $A(\nu)$, can be plotted (the sound-spectrum) of the amplitudes against frequency.

$A(\nu)$ is the Fourier transform of $F(t)$.

Actually it is the *modular* transform, but at this stage that is a detail.

Suppose that the sound is not periodic – a squawk, a drumbeat or a crash instead of a pure note. Then to describe it requires not just a set of overtones

[1] 'phase' here is an angle, used to define the 'retardation' of one wave or vibration with respect to another. One wavelength retardation for example, is equivalent to a phase difference of 2π. Each harmonic will have its own phase, ϕ_m, indicating its position within the period.

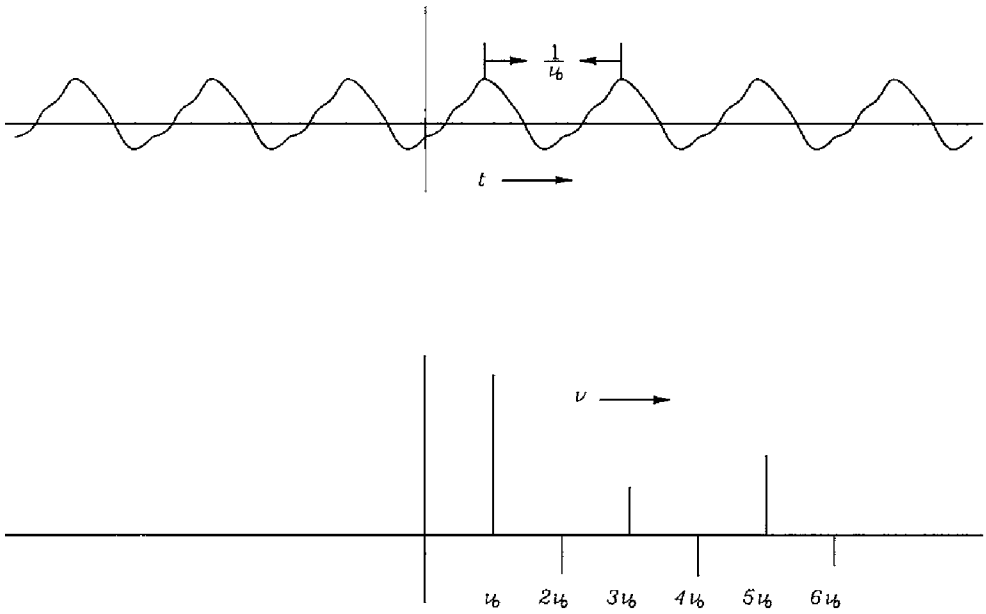


Fig. 1.1. The spectrum of a steady note: fundamental and overtones.

with their amplitudes, but a continuous range of frequencies, each present in an infinitesimal amount. The two curves would then look like Fig. 1.2.

The uses of a Fourier transform can be imagined: the identification of a valuable violin; the analysis of the sound of an aero-engine to detect a faulty gear-wheel; of an electrocardiogram to detect a heart defect; of the light curve of a periodic variable star to determine the underlying physical causes of the variation: all these are current applications of Fourier transforms.

1.2 Fourier series

For a steady note the description requires only the fundamental frequency, its amplitude and the amplitudes of its harmonics. A discrete sum is sufficient. We could write:

$$F(t) = a_0 + a_1 \cos 2\pi \nu_0 t + b_1 \sin 2\pi \nu_0 t + a_2 \cos 4\pi \nu_0 t + b_2 \sin 4\pi \nu_0 t + a_3 \cos 6\pi \nu_0 t + \cdots$$

where ν_0 is the fundamental frequency of the note. Sines as well as cosines are required because the harmonics are not necessarily 'in step' (i.e. 'in phase') with the fundamental or with each other.

More formally:

$$F(t) = \sum_{n=-\infty}^{\infty} a_n \cos(2\pi n \nu_0 t) + b_n \sin(2\pi n \nu_0 t) \tag{1.1}$$

and the sum is taken from $-\infty$ to ∞ for the sake of mathematical symmetry.

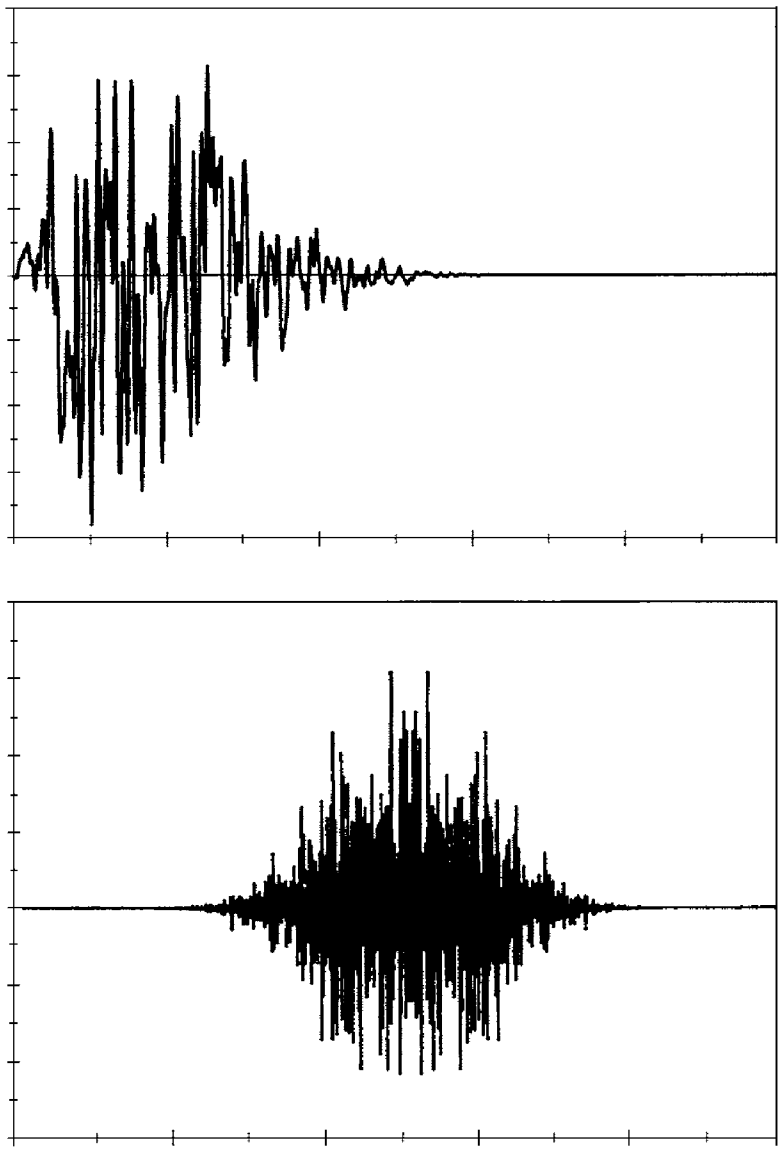


Fig. 1.2. The spectrum of a crash: all frequencies are present.

This process of constructing a waveform by adding together a fundamental frequency and overtones or harmonics of various amplitudes, is called Fourier synthesis.

There are alternative ways of writing this expression:
since $\cos x = \cos(-x)$ and $\sin x = -\sin(-x)$ we can write:

$$F(t) = A_0/2 + \sum_{n=1}^{\infty} A_n \cos(2\pi n \nu_0 t) + B_n \sin(2\pi n \nu_0 t) \tag{1.2}$$

and the two expressions are identical provided that we set $A_n = a_{-n} + a_n$ and $B_n = b_n - b_{-n}$. A_0 is divided by two to avoid counting it twice: as it is, A_0 can be found by the same formula that will be used to find all the A_n's.

Mathematicians and some theoretical physicists write the expression as:

$$F(t) = A_0/2 + \sum_{n=1}^{\infty} A_n \cos(n\omega_0 t) + B_n \sin(n\omega_0 t)$$

and there are entirely practical reasons, which are discussed later, for *not* writing it this way.

1.3 The amplitudes of the harmonics

The alternative process – of extracting from the signal the various frequencies and amplitudes that are present – is called *Fourier analysis* and is much more important in its practical physical applications. In physics, we usually find the curve $F(t)$ experimentally and we want to know the values of the amplitudes A_m and B_m for as many values of m as necessary. To find the values of these amplitudes, we use the *orthogonality* property of sines and cosines. This property is that if you take a sine and a cosine, or two sines or two cosines, each a multiple of some fundamental frequency, multiply them together and integrate the product over one period of that frequency, the result is always zero except in special cases.

If P, $=1/\nu_0$, is one period, then:

$$\int_{t=0}^{P} \cos(2\pi n\nu_0 t).\cos(2\pi m\nu_0 t)\,dt = 0$$

and

$$\int_{t=0}^{P} \sin(2\pi n\nu_0 t).\sin(2\pi m\nu_0 t)\,dt = 0$$

unless $m = \pm n$, and:

$$\int_{t=0}^{P} \sin(2\pi n\nu_0 t).\cos(2\pi m\nu_0 t)\,dt = 0$$

always. The first two integrals are both equal to $1/2\nu_0$ if $m = n$.

We multiply the expression (1.2) for $F(t)$ by $\sin(2\pi m\nu_0 t)$ and the product is integrated over one period, P:

$$\int_{t=0}^{P} F(t)\sin(2\pi m\nu_0 t)\,dt = \int_{t=0}^{P} \sum_{n=1}^{\infty} \{A_n \cos(2\pi n\nu_0 t) + B_n \sin(2\pi n\nu_0 t)\} \times \sin(2\pi m\nu_0 t)\,dt + \frac{A_0}{2}\int_{t=0}^{P} \sin(2\pi m\nu_0 t)\,dt \tag{1.3}$$

and all the terms of the sum vanish on integration except

$$\int_0^P B_m \sin^2(2\pi m\nu_0 t)\,dt = B_m \int_0^P \sin^2(2\pi m\nu_0 t)\,dt$$
$$= B_m/2\nu_0 = B_m P/2$$

so that

$$B_m = (2/P)\int_0^P F(t)\sin(2\pi m\nu_0 t)\,dt \tag{1.4}$$

and provided that $F(t)$ is known in the interval $0 \to P$ the coefficient B_m can be found. If an analytic expression for $F(t)$ is known, the integral can often be done. On the other hand, if $F(t)$ has been found experimentally, a computer is needed to do the integrations.

The corresponding formula for A_m is:

$$A_m = (2/P)\int_0^P F(t)\cos(2\pi m\nu_0 t)\,dt \tag{1.5}$$

The integral can start anywhere, not necessarily at $t = 0$, so long as it extends over one period.

Example: Suppose that $F(t)$ is a square-wave of period $1/\nu_0$, so that $F(t) = h$ for $t = -b/2 \to b/2$ and 0 during the rest of the period, as in the diagram:

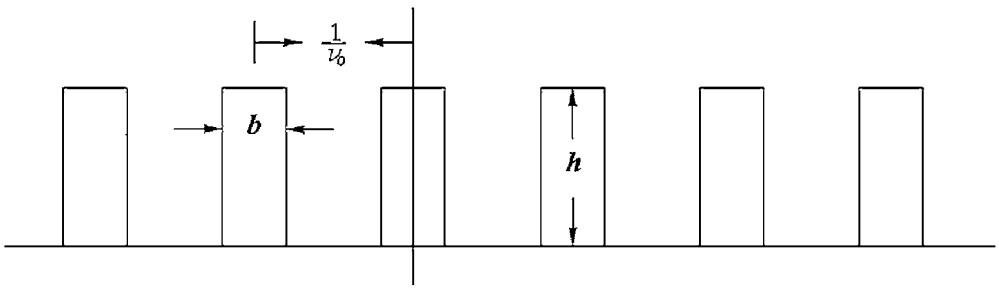


Fig. 1.3. A rectangular wave of period $1/\nu_0$ and pulse-width b.

then:

$$A_m = 2\nu_0 \int_{-1/2\nu_0}^{1/2\nu_0} F(t)\cos(2\pi m\nu_0 t)\,dt$$
$$= 2h\nu_0 \int_{-b/2}^{b/2} \cos(2\pi m\nu_0 t)\,dt$$

and the new limits cover only that part of the cycle where $F(t)$ is different from zero.

If we integrate and put in the limits:

$$\begin{aligned} A_m &= \frac{2h\nu_0}{2\pi m\nu_0}\{\sin(\pi m\nu_0 b) - \sin(-\pi m\nu_0 b)\} \\ &= \frac{2h}{\pi m}\sin(\pi m\nu_0 b) \\ &= 2h\nu_0 b\,\{\sin(\pi\nu_0 mb)/\pi\nu_0 mb\} \end{aligned}$$

All the B_n's are zero because of the symmetry of the function – we took the origin to be at the centre of one of the pulses.
The original function of time can be written:

$$F(t) = h\nu_0 b + 2h\nu_0 b\sum_{m=1}^{\infty}\{\sin(\pi\nu_0 mb)/\pi\nu_0 mb\}\cos(2\pi m\nu_0 t) \quad (1.6)$$

or alternatively:

$$F(t) = \frac{hb}{P} + \frac{2hb}{P}\sum_{m=1}^{\infty}\{\sin(\pi\nu_0 mb)/\pi\nu_0 mb\}\cos(2\pi m\nu_0 t) \quad (1.7)$$

Notice that the first term, $A_0/2$ is the *average* height of the function – the area under the top-hat divided by the period: and that the function $\sin(x)/x$, called 'sinc(x)', which will be described in detail later, has the value unity at $x = 0$, as can be shown using De l'Hôpital's rule[2].

There are other ways of writing the Fourier series. It is convenient occasionally, though less often, to write $A_m = R_m\cos\phi_m$ and $B_m = R_m\sin\phi_m$, so that equation (1.2) becomes:

$$F(t) = \frac{A_0}{2} + \sum_{m=1}^{\infty} R_m\cos(2\pi m\nu_0 t + \phi_m) \quad (1.8)$$

and R_m and ϕ_m are the amplitude and phase of the mth harmonic. A single sinusoid then replaces each sine and cosine, and the two quantities needed to define each harmonic are these amplitudes and phases in place of the previous A_m and B_m coefficients. In practice it is usually the amplitude, R_m which is important, since the energy in an oscillator is proportional to the square of the amplitude of oscillation, and $|\ R_m\ |^2$ gives a measure of the power contained in each harmonic of a wave. 'Phase' is a simple and important idea. Two wave trains are 'in phase' if wave crests arrive at a certain point together. They are 'out of phase' if a trough from one arrives at the same time as the crest of the

[2] De l'Hôpital's rule is that if $f(x) \to 0$ as $x \to 0$ and $\phi(x) \to 0$ as $x \to 0$, the ratio $f(x)/\phi(x)$ is indeterminate, but is equal to the ratio $(df/dx)/(d\phi/dx)$ as $x \to 0$.

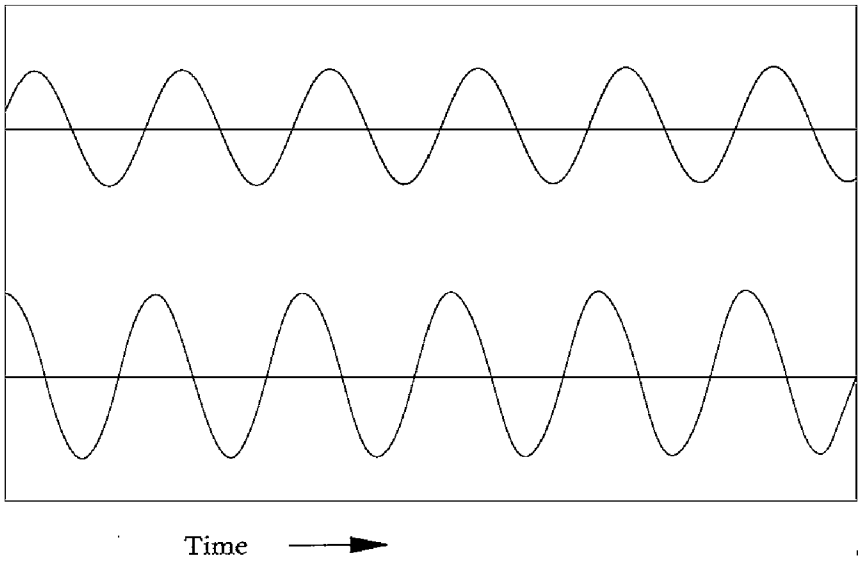


Fig. 1.4. Two wave trains with the same period but different amplitudes and phases. The upper has 0.7× the amplitude of the lower and there is a phase-difference of 70°.

other. (Alternatively they have 180° phase difference.) In Fig. 1.4 there are two wave trains. The upper has 0.7× the amplitude of the other and it *lags* (not *leads*, as it appears to do) the lower by 70°. This is because the horizontal axis of the graph is time, and the vertical axis measures the amplitude at a fixed point as it varies with time. Wave crests from the lower wave train arrive earlier than those from the upper. The important thing is that the 'phase-difference' between the two is 70°.

The most common way of writing the series expansion is with complex exponentials instead of trigonometrical functions. This is because the algebra of complex exponentials is easier to manipulate. The two ways are linked of course by De Moivre's theorem. We can write:

$$F(t) = \sum_{-\infty}^{\infty} C_m e^{2\pi i m \nu_0 t}$$

where the coefficients C_m are now complex numbers in general and $C_m = C^*_{-m}$. (The exact relationship is given in detail in Appendix 1.4). The coefficients A_m, B_m and C_m are obtained from the *Inversion Formulae:*

$$A_m = 2\nu_0 \int_0^{1/\nu_0} F(t)\cos(2\pi m \nu_0 t)\, dt$$

$$B_m = 2\nu_0 \int_0^{1/\nu_0} F(t)\sin(2\pi m \nu_0 t)\, dt$$

$$C_m = 2\nu_0 \int_0^{1/\nu_0} F(t) e^{-2\pi m \nu_0 t}\, dt$$

(The minus sign in the exponent is important) or, if ω_0 has been used instead of ν_0 $(=\omega_0/2\pi)$ then:

$$A_m = \omega_0/\pi \int_0^{2\pi/\omega_0} F(t)\cos(m\omega_0 t)\,dt$$

$$B_m = \omega_0/\pi \int_0^{2\pi/\omega_0} F(t)\sin(m\omega_0 t)\,dt$$

$$C_m = 2\omega_0/\pi \int_0^{2\pi/\omega_0} F(t)e^{-im\omega_0 t}\,dt$$

The useful mnemonic form to remember for finding the coefficients in a Fourier series is:

$$A_m = \frac{2}{period}\int_{one\ period} F(t)\cos\left\{\frac{2\pi mt}{period}\right\}dt \tag{1.9}$$

$$B_m = \frac{2}{period}\int_{one\ period} F(t)\sin\left\{\frac{2\pi mt}{period}\right\}dt \tag{1.10}$$

and remember that the integral can be taken from any starting point, a, provided it extends over one period to an upper limit $a + P$. The integral can be split into as many subdivisions as needed if, for example, $F(t)$ has different analytic forms in different parts of the period.

1.4 Fourier transforms

Whether $F(t)$ is periodic or not, a complete description of $F(t)$ can be given using sines and cosines. If $F(t)$ is not periodic it requires all frequencies to be present if it is to be synthesized. A non-periodic function may be thought of as a limiting case of a periodic one, where the period tends to infinity, and consequently the fundamental frequency tends to zero. The harmonics are more and more closely spaced and in the limit there is a continuum of harmonics, each one of infinitesimal amplitude, $a(\nu)d\nu$, for example. The summation sign is replaced by an integral sign and we find that:

$$F(t) = \int_{-\infty}^{\infty} a(\nu)d\nu\cos(2\pi\nu t) + \int_{-\infty}^{\infty} b(\nu)d\nu\sin(2\pi\nu t) \tag{1.11}$$

or, equivalently:

$$F(t) = \int_{-\infty}^{\infty} r(\nu)\cos(2\pi\nu t + \phi(\nu))\,d\nu \tag{1.12}$$

or, again:

$$F(t) = \int_{-\infty}^{\infty} \Phi(\nu)e^{2\pi i\nu t}\,d\nu \tag{1.13}$$

If $F(t)$ is real, that is to say, if the insertion of any value of t into $F(t)$ yields a real number, then $a(\nu)$ and $b(\nu)$ are real too. However, $\Phi(\nu)$ may be complex and indeed will be if $F(t)$ is asymmetrical so that $F(t) \neq F(-t)$. This can sometimes cause complications, and these are dealt with in Chapter 8: but $F(t)$ is often symmetrical and then $\Phi(\nu)$ is real and $F(t)$ comprises only cosines. We *could* then write:

$$F(t) = \int_{-\infty}^{\infty} \Phi(\nu) \cos(2\pi \nu t)\, d\nu$$

but because complex exponentials are easier to manipulate, we take as a standard form the equation (1.13) above. Nevertheless, for many practical purposes only real and symmetrical functions $F(t)$ and $\Phi(\nu)$ need be considered.

Just as with Fourier series, the function $\Phi(\nu)$ can be recovered from $F(t)$ by inversion. This is the cornerstone of Fourier theory because, astonishingly, the inversion has exactly the same form as the synthesis, and we can write, if $\Phi(\nu)$ is real and $F(t)$ is symmetric:

$$\Phi(\nu) = \int_{-\infty}^{\infty} F(t) \cos(2\pi \nu t)\, dt \tag{1.14}$$

so that not only is $\Phi(\nu)$ the Fourier transform of $F(t)$, but $F(t)$ is the Fourier transform of $\Phi(\nu)$. The two together are called a 'Fourier Pair'.

The complete and rigorous proof of this is long and tedious[3] and it is not necessary here; but the formal definition can be given and this is a suitable place to abandon, for the moment, the physical variables time and frequency and to change to the pair of abstract variables, x and p, which are usually used. The formal statement of a Fourier transform is then:

$$\Phi(p) = \int_{-\infty}^{\infty} F(x) e^{2\pi i p x}\, dx \tag{1.15}$$

$$F(x) = \int_{-\infty}^{\infty} \Phi(p) e^{-2\pi i p x}\, dp \tag{1.16}$$

and this pair of formulae[4] will be used from here on.

[3] It is to be found, for example, in E. C. Titchmarsh, *Introduction to the Theory of Fourier Integrals*, Clarendon Press, Oxford, 1962 or in R. R. Goldberg, *Fourier Transforms*, Cambridge University Press, Cambridge, 1965.

[4] Sometimes one finds:

$$\Phi(p) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(x) e^{ipx}\, dx; \quad F(x) = \int_{-\infty}^{\infty} \Phi(p) e^{-ipx}\, dp$$

as the defining equations, and again symmetry is preserved by some people by defining the transform by:

$$\Phi(p) = \left\{\frac{1}{2\pi}\right\}^{\frac{1}{2}} \int_{-\infty}^{\infty} F(x) e^{ipx}\, dx; \quad F(x) = \left\{\frac{1}{2\pi}\right\}^{\frac{1}{2}} \int_{-\infty}^{\infty} \Phi(p) e^{-ipx}\, dp$$

Symbolically we write:

$$\Phi(p) \rightleftharpoons F(x)$$

One and only one of the integrals must have a minus sign in the exponent. Which of the two you choose does not matter, so long as you keep to the rule. If the rule is broken half way through a long calculation the result is chaos; but if someone else has used the opposite choice, the Fourier pair calculated of a given function will be the complex conjugate of that given by your choice.

When time and frequency are the conjugate variables we shall use:

$$\Phi(\nu) = \int_{-\infty}^{\infty} F(t)e^{-2\pi i\nu t}dt \tag{1.17}$$

$$F(t) = \int_{-\infty}^{\infty} \Phi(\nu)^{2\pi i\nu t}d\nu \tag{1.18}$$

and again, symbolically:

$$\Phi(\nu) \rightleftharpoons F(t)$$

There are two good reasons for incorporating the 2π into the exponent. Firstly the defining equations are easily remembered without worrying where the 2π's go, but more importantly, quantities like t and ν are actually physically measured quantities – time and frequency – rather than time and *angular* frequency, ω. Angular measure is for mathematicians. For example, when one has to integrate a function wrapped around a cylinder it is convenient to use the angle as the independent variable. Physicists will generally find it more convenient to use t and ν, for example, with the 2π in the exponent.

1.5 Conjugate variables

Traditionally x and p are used when abstract transforms are considered and they are called 'conjugate variables'. Different fields of physics and engineering use different pairs, such as frequency, ν and time, t in accoustics, telecommunications and radio; position, x and momentum divided by Planck's constant, $p/\hbar$ in quantum mechanics, and aperture x, and diffraction angle divided by wavelength $p = \sin\theta/\lambda$ in diffraction theory.

In general we will use x and p as abstract entities and give them a physical meaning when an illustration seems called-for. It is worth remembering that x and p have inverse dimensionality, as in time t and frequency, t^{-1}. The product px, like any exponent, is always a dimensionless number.

One further definition is needed: the 'power spectrum' of a function[5]. This notion is important in electrical engineering as well as in physics. If power is

[5] Actually the *energy* spectrum. 'Power spectrum' is just the conventional term used in most books. This is discussed in more detail in Chapter 4.

transmitted by electromagnetic radiation (radio waves or light) or by wires or waveguides, the voltage at a point varies with time as $V(t)$. $\Phi(\nu)$, the Fourier transform of $V(t)$, may very well be – indeed usually is – complex. however the power per unit frequency interval being transmitted is proportional to $\Phi(\nu)\Phi^*(\nu)$, where the constant of proportionality depends on the load impedance. The function $S(\nu) = \Phi(\nu)\Phi^*(\nu) =\mid \Phi(\nu) \mid^2$ is called the power spectrum or the spectral power density (SPD) of $F(t)$. This what an optical spectrometer measures, for example.

1.6 Graphical representations

It frequently happens that greater insight into the physical processes which are described by a Fourier transform can be achieved by a diagram rather than a formula. When a real function $F(x)$ is transformed it generally produces a complex function $\Phi(p)$, which needs an Argand diagram to demonstrate it. Three dimensions are required: $\mathbf{Re}\Phi(p)$;$\mathbf{Im}\Phi(p)$ and p. A perspective drawing will display the function, which appears as a more or less sinuous line. If $F(x)$ is symmetrical, the line lies in the **Re**-p plane, and if antisymmetrical, in the **Im**-p plane. The Figures 8.1 and 8.2 in Chapter 8 illustrate this point.

Electrical engineering students in particular, will recognize the end-on view along the p-axis as the 'Nyquist diagram' of feedback theory. There will be examples of this graphical representation in later chapters.

1.7 Useful functions

There are some functions which occur again and again in physics, and whose properties should be learned. They are extremely useful in the manipulation and general taming of other functions which would otherwise be almost unmanageable. Chief among these are:

1.7.1 The 'top-hat' function[6]

This has the property that:

$$\begin{aligned}\Pi_a(x) &= 0, -\infty < x < -a/2\\ &= 1, -a/2 < x < a/2\\ &= 0, a/2 < x < \infty\end{aligned}$$

and the symbol Π is chosen as an obvious aid to memory.

[6] In the USA this is called a 'box-car' or 'rect' function.

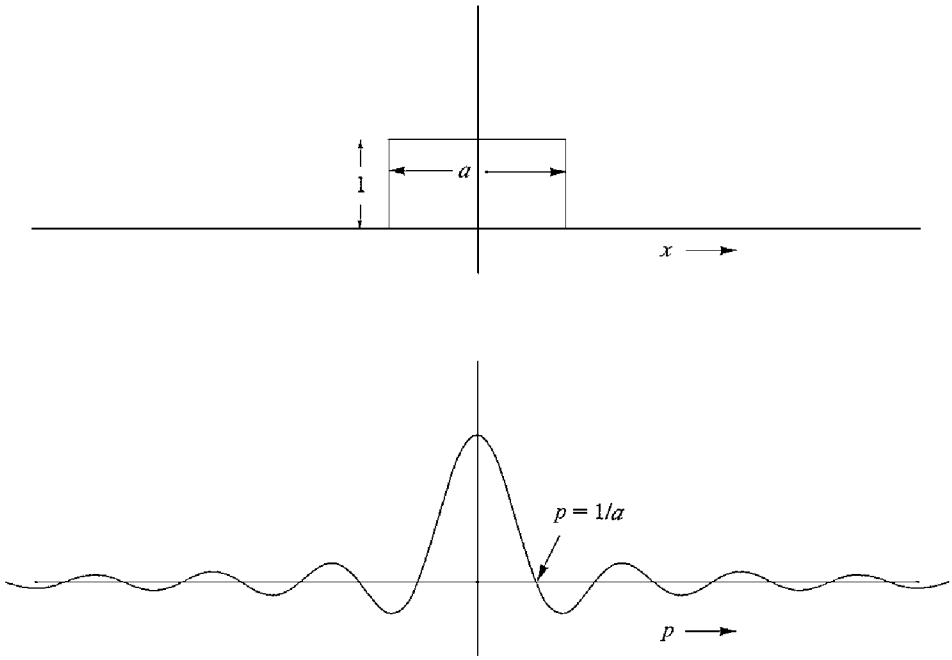


Fig. 1.5. The top-hat function and its transform, the sinc-function.

Its Fourier pair is obtained by integration:

$$\begin{aligned}\Phi(p) &= \int_{-\infty}^{\infty} \Pi_a(x) e^{2\pi i p x} dx \\ &= \int_{-a/2}^{a/2} e^{2\pi i p x} dx \\ &= \frac{1}{2\pi i p}[e^{\pi i p a} - e^{-\pi i p a}] \\ &= a\left\{\frac{\sin \pi p a}{\pi p a}\right\} \\ &= a.\mathrm{sinc}(\pi p a)\end{aligned}$$

and the 'sinc-function', defined[7] by $\mathrm{sinc}(x) = \sin x/x$ is one which recurs throughout physics. As before, we write symbolically:

$$\Pi_a(x) \rightleftharpoons a.\mathrm{sinc}(\pi p a)$$

[7] Caution: some people define $\mathrm{sinc}(x)$ as $\sin(\pi x)/(\pi x)$.

1.7.2 The sinc-function

$$\mathrm{sinc}(x) = \sin x/x$$

Has the value unity at $x = 0$, and has zeros whenever $x = n\pi$. The function $\mathrm{sinc}(\pi pa)$ above, the most common form, has zeros when $p = 1/a, 2/a, 3/a, \ldots$

1.7.3 The Gaussian function

Suppose $G(x) = e^{-x^2/a^2}$

a is the 'width parameter' of the function, and the full width at half maximum (FWHM) is $1.386a$.

and (what every scientist should know!): $\int_{-\infty}^{\infty} e^{-x^2/a^2} dx = a\sqrt{\pi}$

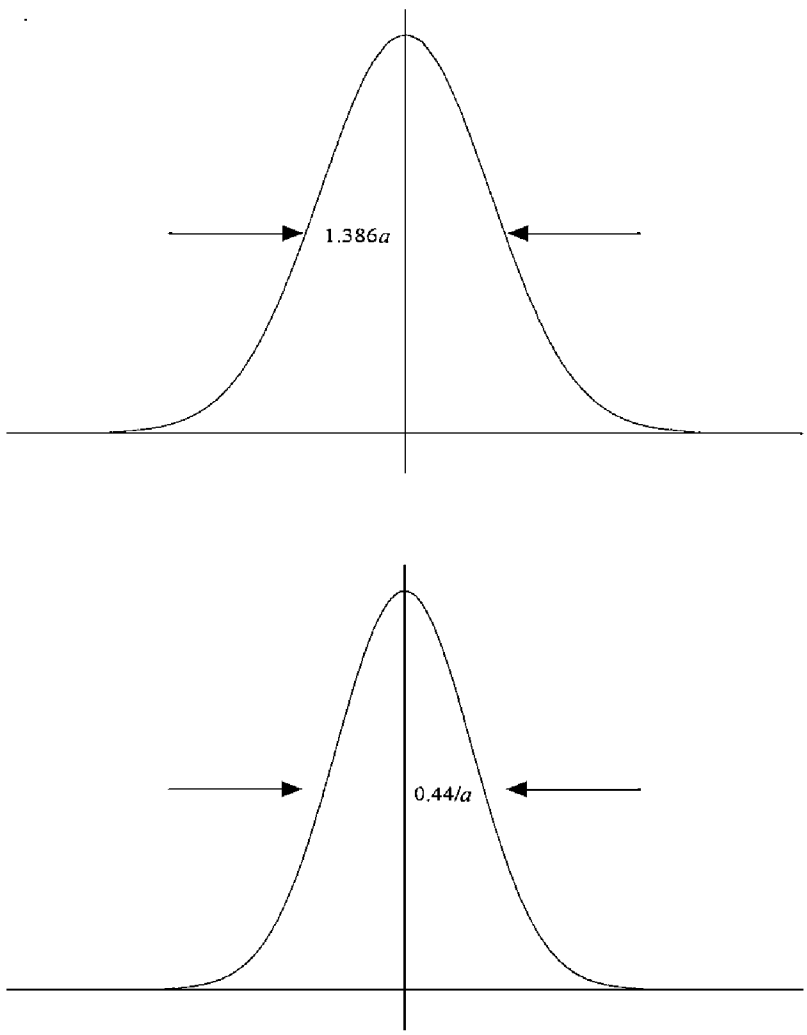


Fig. 1.6. The Gaussian function and its transform, another Gaussian with full width at half maximum inversely proportional to that of its Fourier pair.

Its Fourier transform is $g(p)$, given by:

$$g(p) = \int_{-\infty}^{\infty} e^{-x^2/a^2} e^{2\pi i p x} dx$$

The exponent can be rewritten (by 'completing the square') as

$$-(x/a - \pi i p a)^2 - \pi^2 p^2 a^2$$

and then

$$g(p) = e^{-\pi^2 p^2 a^2} \int_{-\infty}^{\infty} e^{-(x/a - \pi i p a)^2} dx$$

put $x/a - \pi i p a = z$, so that $dx = a dz$. Then:

$$g(p) = a e^{-\pi^2 p^2 a^2} \int_{-\infty}^{\infty} e^{-z^2} dz$$

$$= a\sqrt{\pi} e^{-\pi^2 a^2 p^2}$$

so that $g(p)$ is another Gaussian function, with width parameter $1/\pi a$.

Notice that, the wider the original Gaussian, the narrower will be its Fourier pair.

Notice too, that the value at $p = 0$ of the Fourier pair is equal to the area under the original Gaussian.

1.7.4 The exponential decay

This, in physics is generally the positive part of the function $e^{-x/a}$. It is asymmetric, so its Fourier transform is complex:

$$\Phi(p) = \int_0^{\infty} e^{-x/a} e^{2\pi i p x} dx$$

$$= \left[\frac{e^{2\pi i p x - x/a}}{2\pi i p - 1/a} \right]_0^{\infty} = \frac{-1}{2\pi i p - 1/a}$$

Usually, with this function, the power spectrum is the most interesting:

$$| \Phi(p) |^2 = \frac{a^2}{4\pi^2 p^2 a^2 + 1}$$

This is a bell-shaped curve, similar in appearance to a Gaussian curve, and is known as a Lorentz profile. It has a FWHM $= 1/\pi a$.

It is the shape found in spectrum lines when they are observed at very low pressure, when collisions between emitting particles are infrequent compared with the transition probability. If the line profile is taken as a function of frequency,

$I(\nu)$, the FWHM, $\Delta\nu$ is related to the 'Lifetime of the Excited State', the reciprocal of the transition probability in the atom which undergoes the transition. In this example, a and x obviously have dimensions of time. Looked at classically, the emitting particle is behaving like a damped harmonic oscillator radiating power at an exponentially decreasing rate. Quantum mechanics yields the same equation through perturbation theory.

There is more discussion of this profile in Chapter 5.

1.7.5 The Dirac 'delta-function'

This has the following properties:

$$\delta(x) = 0 \text{ unless } x = 0$$
$$\delta(0) = \infty$$
$$\int_{-\infty}^{\infty} \delta(x)dx = 1$$

It is an example of a function which disobeys one of Dirichlet's conditions,

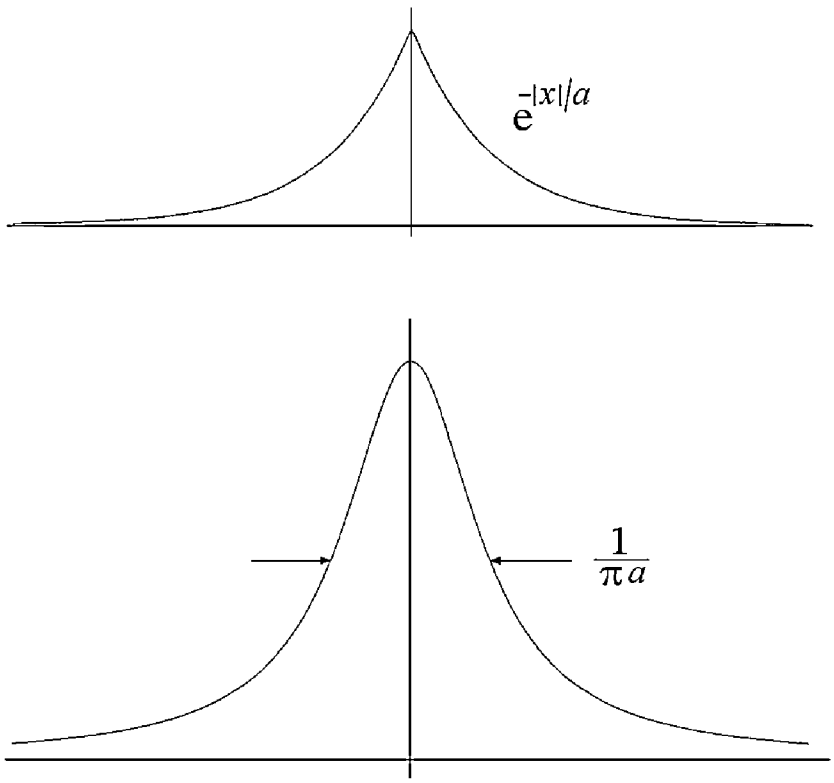


Fig. 1.7. The exponential decay $e^{-|x|/a}$ and its Fourier transform.

since it is unbounded at $x = 0$. It can be regarded crudely as the limiting case of a top-hat function $(1/a)\Pi_a(x)$ as $a \to 0$. It becomes narrower and higher, and its area, which we shall refer to as its *amplitude* is always equal to unity. Its Fourier transform is $\text{sinc}(\pi pa)$ and as $a \to 0$, $\text{sinc}(\pi pa)$ stretches and in the limit is a straight line at unit height above the $x-$axis. In other words,

The Fourier transform of a δ-function is unity

and we write:

$$\delta(x) \rightleftharpoons 1$$

Alternatively, and more accurately, it is the limiting case of a Gaussian function of unit area as it gets narrower and higher. Its Fourier transform then is another Gaussian of unit height, getting broader and broader until in the limit it is a straight line at unit height above the axis.

The following useful properties of the δ-function should be memorized. They are:

$$\delta(x - a) = 0 \text{ unless } x = a$$

The so-called 'shift theorem':

$$\int_{-\infty}^{\infty} f(x)\delta(x - a)dx = f(a)$$

where the product under the integral sign is zero except at $x = a$ where, on integration, the δ-function has the amplitude $f(a)$.

It is then easy to show, using this shift theorem, that for positive[8] values of a, b, c and d:

$$\begin{aligned} \delta(x/a - 1) &= a\delta(x - a) \\ \delta(a/b - c/d) &= ac\delta(ad - bc) \\ &= bd\delta(ad - bc) \\ \delta(ax) &= (1/a)\delta(x) \end{aligned}$$

[8] for negative values of these quantities a minus sign may be needed, bearing in mind that the integral of a δ-function is always positive, even though a, for example may be negative. Alternatively we may write, for example, $\delta(x/a - 1) = \mid a \mid \delta(x - a)$.

And another important consequence of the shift theorem is:

$$\int_{-\infty}^{\infty} e^{2\pi ipx}\delta(x-a)dx = e^{2\pi ipa}$$

so that we can write:

$$\delta(x-a) \rightleftharpoons e^{2\pi ipa}$$
$$\delta(mx-a) \rightleftharpoons (1/m)e^{2\pi ipa/m}$$

and a formula which we shall need in Chapter 7:

$$\frac{1}{n}\delta\left(\frac{p}{l}-\frac{r}{n}\right) = \delta\left(\frac{pn}{l}-r\right) \rightleftharpoons e^{-2\pi i(\frac{pn}{l}-r)}$$

1.7.5.1 A pair of δ-functions

If two δ-functions are equally disposed on either side of the origin, the Fourier transform is a cosine wave:

$$\begin{aligned}\delta(x-a)+\delta(x+a) &\rightleftharpoons e^{2\pi ipa}+e^{-2\pi ipa}\\ &= 2\cos(2\pi pa)\end{aligned} \tag{1.19}$$

1.7.5.2 The Dirac comb

This is an infinite set of equally-spaced δ-functions, usually denoted by the Cyrillic letter *Ш* (Shah). Formally, we write:

$$Ш_a(x) = \sum_{n=-\infty}^{\infty} \delta(x-na)$$

It is useful because it allows us to include Fourier series in the general theory of Fourier transforms. For example, the *convolution* (to be described later) of $Ш_a(x)$ and $(1/b)\Pi_b(x)$ (where $b < a$) is a square wave similar to that in the earlier example, of period a and width b, and with unit area in each rectangle. The Fourier transform is then a Dirac comb, with 'teeth' of height a_m spaced at intervals $1/a$. The a_m are of course the coefficients in the series.

If the square wave is allowed to become infinitesimally wide and infinitely high so that the area under each rectangle remains unity, then the coefficients a_m will all become the same height, $1/a$. In other words, the Fourier transform of a Dirac comb is another Dirac comb:

$$Ш_a(x) \rightleftharpoons \frac{1}{a}Ш_{\frac{1}{a}}(p)$$

and again notice that the period in p-space is the reciprocal of the period in x-space.

This is not a formal demonstration of the Fourier transform of a Dirac comb. A rigorous proof is much more elaborate, but is unnecessary here.

1.8 Worked examples

1. A train of rectangular pulses, as in Fig. 1.8, has a pulse width equal to 1/4 of the pulse period. Show that the 4th, 8th 12th etc. harmonics are missing.

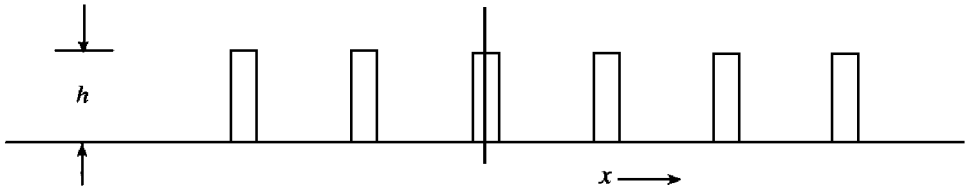


Fig. 1.8. A rectangular pulse-train with a 4 :1 'mark-space' ratio.

Taking zero at the centre of one pulse, the function is clearly symmetrical so that there are only cosine amplitudes.

$$A_n = \frac{2}{P}\int_{-P/8}^{P/8} h\cos\left(\frac{2\pi nx}{P}\right)dx$$

$$= \left(\frac{h}{\pi n}\right)2\sin\left(\frac{2\pi n}{P}.\frac{P}{8}\right)$$

$$= \left(\frac{h}{2}\right)\mathrm{sinc}\left(\frac{\pi n}{4}\right)$$

so that $A_n = 0$ if $n = 4, 8, 12, \ldots$

2. Find the sine-amplitude of a saw-tooth waveform as in Fig. 1.9:

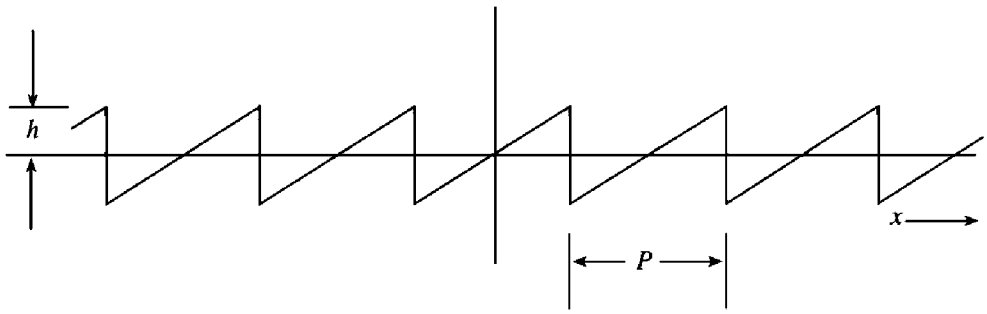


Fig. 1.9. A saw-tooth waveform, antisymmetrical about the origin.

By choosing the origin half way up one of the teeth, the function is clearly made antisymmetrical, so that there are no cosine amplitudes.

$$\begin{aligned} B_n &= \frac{2}{P}\int_{-P/2}^{P/2} 2\frac{xh}{P}\sin\left(\frac{2\pi nx}{P}\right)dx \\ &= 4\frac{h}{P^2}\left[-x\cos\left(\frac{2\pi nx}{P}\right)\frac{P}{2\pi n} + \frac{P^2}{4\pi^2 n^2}\sin\left(\frac{2\pi nx}{P}\right)\right]_{-P/2}^{P/2} \\ &= (-2h/\pi n)\cos\pi n \quad \text{since } \sin\pi n = 0 \end{aligned}$$

so that

$$\begin{aligned} B_0 &= 0 \\ B_n &= (-1)^{n+1}(2h/\pi n),\ n \neq 0 \end{aligned}$$

As a matter of interest, it is worth while calculating the sine-amplitudes when the origin is taken at the tip of a tooth, to see how changing the position of the origin changes the amplitudes. It is also worth while doing the calculation for a similar wave, with negative-going slopes instead of positive.

Chapter 2

Useful properties and theorems

2.1 The Dirichlet conditions

Not all functions can be Fourier-transformed. They are transformable if they fulfil certain conditions, known as the Dirichlet conditions.

The integrals which formally define the Fourier transform in Chapter 1 will exist if the integrands fulfil the following conditions:

The functions $F(x)$ and $\Phi(p)$ are square-integrable, i.e. $\int_{-\infty}^{\infty} \mid F(x) \mid^2 dx$ is finite, which implies that $F(x) \rightarrow 0$ as $\mid x \mid \rightarrow \infty$

$F(x)$ and $\Phi(p)$ are single-valued. For example a function such as that in Fig. 2.1 is not Fourier-transformable:

$F(x)$ and $\Phi(p)$ are 'piece-wise continuous'. The function can be broken up into separate pieces, so that there can be isolated discontinuities, as many as you like, at the junctions, but the functions must be *continuous* in the mathematical sense, between these discontinuities[1].

The functions $F(x)$ and $\Phi(p)$ have upper and lower bounds.

This is a condition which is *sufficient* but has not been proved *necessary*. In fact we shall assume that it is not. The Dirac δ-function, for instance, disobeys this condition. No engineer or physicist has yet lost sleep over this one.

In Nature, all the phenomena that can be described mathematically seem to require only well-behaved functions which obey the Dirichlet conditions. For example, we can describe the electric field of a wave-packet[2] by a function which is continuous, finite and single-valued everywhere, and as the wave-packet contains only a finite amount of energy, the electric field is square-integrable.

[1] The classical nonconformist example is Weierstrass's function, $W(x)$, which has the property that $W(x) = 1$ if x is rational and $W(x) = 0$ if x is irrational. It looks like a straight line but it is not transformable, since it can be shown that between any two rational numbers, however close, there is at least one irrational number, and between any two irrational numbers there is at least one rational number, so that the function is everywhere discontinuous.

[2] I have deliberately avoided the word 'photon', for fear of causing apoplexy among strict quantum theory purists.

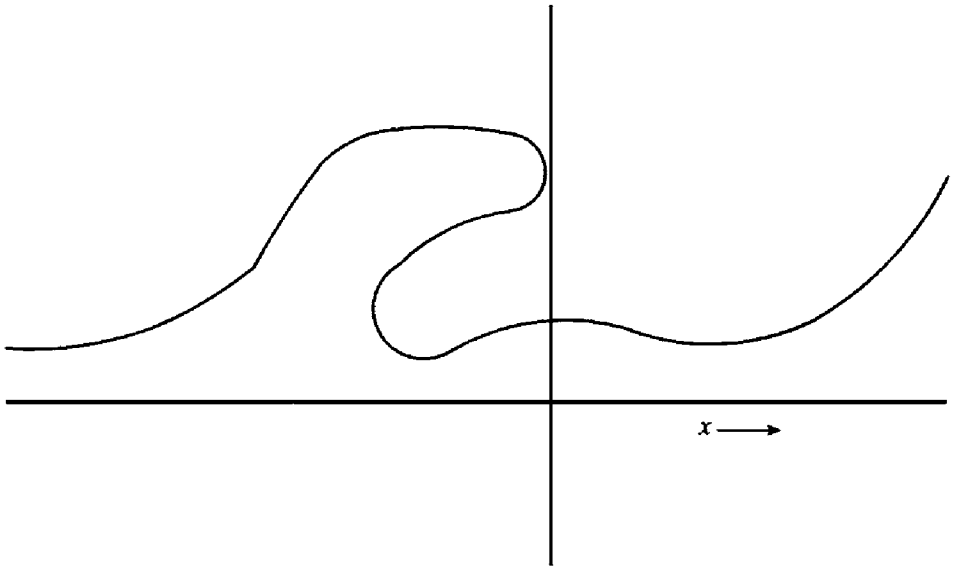


Fig. 2.1. A triple-valued function like this can not be Fourier-transformed.

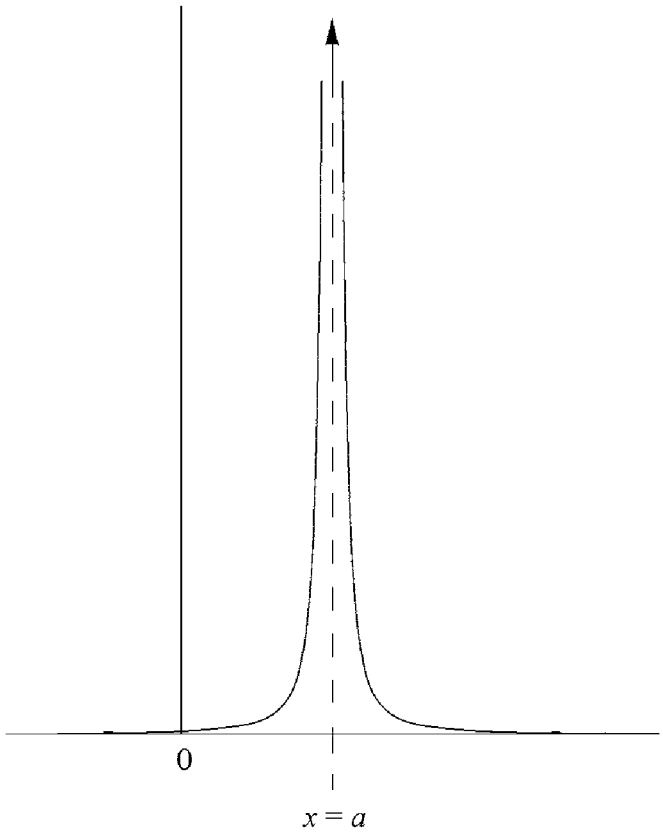


Fig. 2.2. $F(x) = 1/(x - a)^2$, an unbounded non-transformable function of x.

2.2 Theorems

There are several theorems which are of great use in manipulating Fourier-pairs, and they should be memorized. For the most part the proofs are elementary. The art of practical Fourier-transforming is in the manipulation of functions

using these theorems, rather than in doing extensive and tiresome elementary integrations. It is this, as much as anything, which makes Fourier theory such a powerful tool for the practical working scientist.

In what follows, we assume:

$$F_1(x) \rightleftharpoons \Phi_1(p); F_2(x) \rightleftharpoons \Phi_2(p)$$

where '$\rightleftharpoons$' implies that F_1 and Φ_1 are a Fourier pair.

The addition theorem:

$$F_1(x) + F_2(x) \rightleftharpoons \Phi_1(p) + \Phi_2(p) \tag{2.1}$$

The shift theorem already mentioned in Chapter 1 has the following lemmas:

$$\begin{aligned} F_1(x+a) &\rightleftharpoons \Phi_1(p)e^{2\pi ipa} \\ F_1(x-a) &\rightleftharpoons \Phi_1(p)e^{-2\pi ipa} \\ F_1(x-a) + F_1(x+a) &\rightleftharpoons 2\Phi_1(p)\cos 2\pi pa \end{aligned} \tag{2.2}$$

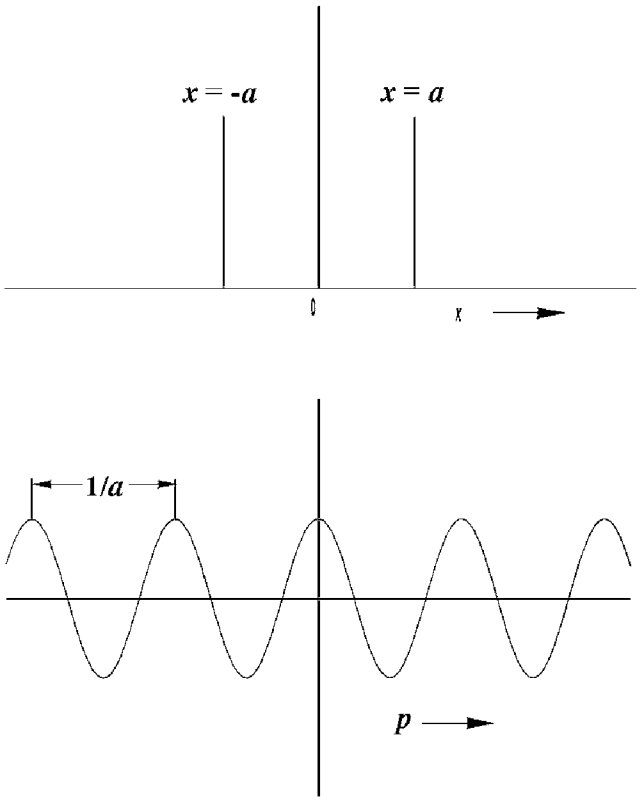


Fig. 2.3. A pair of δ-functions and its transform.

In particular, notice that if $F_1(x)$ is a δ-function, the lemmas are:

$$\begin{aligned} \delta(x+a) &\rightleftharpoons e^{-2\pi ipa} \\ \delta(x-a) &\rightleftharpoons e^{2\pi ipa} \\ \delta(x-a)+\delta(x+a) &\rightleftharpoons 2\cos 2\pi pa \end{aligned} \tag{2.3}$$

The third of these is illustrated in Fig. 2.3:

2.3 Convolutions and the convolution theorem

Convolutions are an important concept, especially in practical physics, and the idea of a convolution can be illustrated simply by an example.

Imagine a 'perfect' spectrometer, plotting a graph of intensity against wavelength, of a monochromatic source of light of intensity S and wavelength λ_0. Represent the power spectral density ('the spectrum') of the source by $S\delta(\lambda-\lambda_0)$. The spectrometer will plot the graph as $kS\delta(\lambda-\lambda_0)$, where k is a factor which depends on the throughput of the spectrometer, its geometry and its detector sensitivity.

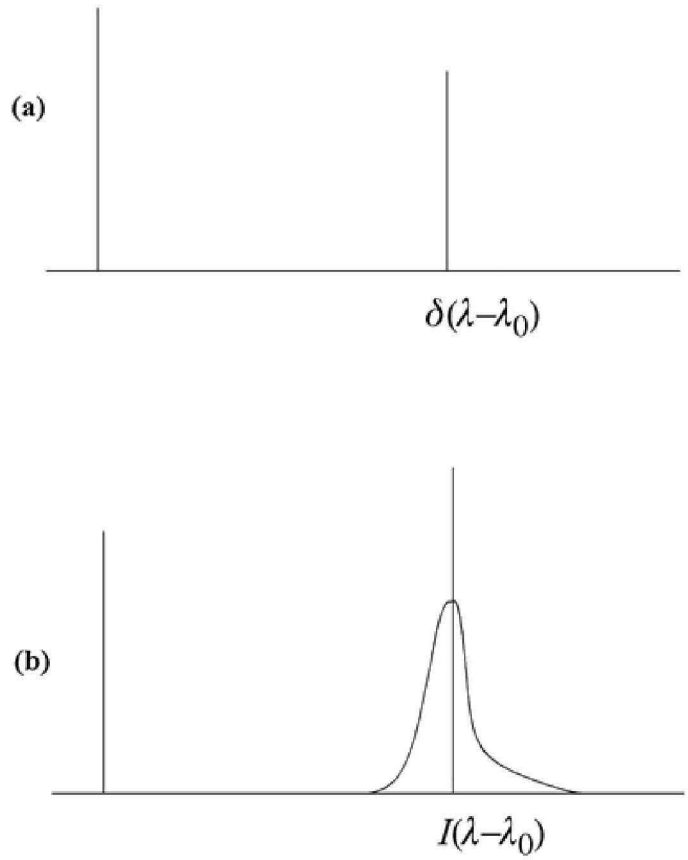


Fig. 2.4. The spectrum of a monochromatic wave (a) entering and (b) leaving a spectrometer. The area under curve (b) must be unity – the same as the 'area' under the δ-function, to preserve the idea of an 'instrumental function'.

No spectrometer is perfect in practice, and what a real instrument will plot in response to a monochromatic input is a continuous curve $kSI(\lambda - \lambda_0)$, where $I(\lambda)$ is called the 'instrumental function' and $\int_{-\infty}^{\infty} I(\lambda)d\lambda = 1$.

Now we inquire what the instrument will plot in response to a continuous spectrum input. Suppose that the intensity of the source as a function of wavelength is $S(\lambda)$. We assume that a monochromatic line at *any* wavelength λ_1 will be plotted as a similarly shaped function $kI(\lambda - \lambda_1)$. Then an infinitesimal interval of the spectrum can be considered as a monochromatic line, at λ_1, say, and of intensity $S(\lambda_1)d\lambda_1$ and it is plotted by the spectrometer as a function of λ:

$$dO(\lambda) = kS(\lambda_1)d\lambda_1 I(\lambda - \lambda_1)$$

and the intensity *apparently* at another wavelength: λ_2 is:

$$dO(\lambda_2) = kS(\lambda_1)I(\lambda_2 - \lambda_1)d\lambda_1$$

The total power apparently at λ_2 is got by integrating this over all wavelengths:

$$O(\lambda_2) = k\int_{-\infty}^{\infty} S(\lambda_1)I(\lambda_2 - \lambda_1)d\lambda_1$$

or, dropping unnecessary subscripts:

$$O(\lambda) = k\int_{-\infty}^{\infty} S(\lambda_1)I(\lambda - \lambda_1)d\lambda_1$$

and the output curve, $O(\lambda)$ is said to be the convolution of the spectrum $S(\lambda)$ with the instrumental function $I(\lambda)$.

It is the idea of an instrumental function, $I(\lambda)$, which is important here. We assume that the same shape $I(\lambda)$ is given to any monochromatic line input. The idea extends to all sorts of measuring instruments and has various names, such as 'impulse response', 'point-spread function', 'Green's function' and so on, depending on which branch of physics or electrical engineering is being discussed. In an electronic circuit, for example, it answers the question 'if you put in a sharp pulse, what comes out?' Most instruments have no fixed unique 'instrumental function', but the function often changes slowly enough (with wavelength, in the spectrometer example) that the idea can be used for practical calculations.

The same idea can be envisaged in two dimensions: a point object – a star for instance – is imaged by a camera lens as a small smear of light, the 'point-spread function' of the lens. Even a 'perfect' lens has a diffraction pattern, so that the best that can be done is to convert a point object into an 'Airy-disc', a spot, $1.22f\lambda/d$ in diameter, where f is the focal length and d the diameter of the

lens. The lens in general, when taking a photograph, gives an image which is the convolution, in two dimensions, of its point-spread function with the object.

The formal definition of a convolution of two functions is then:

$$C(x) = \int_{-\infty}^{\infty} F_1(x')F_2(x - x')\,dx' \tag{2.4}$$

and we write this symbolically as:

$$C(x) = F_1(x) * F_2(x)$$

Convolutions obey various rules of arithmetic, and can be manipulated using them:

The commutative rule:

$$C(x) = F_1(x) * F_2(x) = F_2(x) * F_1(x)$$

or:

$$C(x) = \int_{-\infty}^{\infty} F_2(x')F_1(x - x')\,dx'$$

as can be shown by a simple substitution.

The distributive rule:

$$F_1(x) * [F_2(x) + F_3(x)] = F_1(x) * F_2(x) + F_1(x) * F_3(x)$$

The associative rule: the idea of a convolution can be extended to three or more functions, and the *order* in which the convolutions are done does not matter:

$$F_1(x) * [F_2(x) * F_3(x)] = [F_1(x) * F_2(x)] * F_3(x)$$

and usually the convolution of three functions is written without the square bracket:

$$C(x) = F_1(x) * F_2(x) * F_3(x) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F_1(x - x')F_2(x' - x'') \times F_3(x'')\,dx'dx''$$

In fact a whole algebra of convolutions exists and is very useful in taming some of the more fearsome-looking functions that are found in physics. For example:

$$[F_1(x) + F_2(x)] * [F_3(x) + F_4(x)] = F_1(x) * F_3(x) + F_1(x) * F_4(x) + F_2(x) * F_3(x) + F_2(x) * F_4(x)$$

There is a way of visualizing a convolution. Draw the graph of $F_1(x)$. Draw,

on a piece of transparent paper, the graph of $F_2(x)$. Turn the transparent graph over about a vertical axis and lay this mirror-image of F_2 on top of the graph of F_1. When the two y-axes are displaced by a distance x', integrate the product of the two functions. The result is one point on the graph of $C(x')$.

2.3.1 The convolution theorem

With the exception of Fourier's Inversion Theorem, the convolution theorem is the most astonishing result in Fourier theory. It is as follows:

If $C(x)$ is the convolution of $F_1(x)$ with $F_2(x)$ then its Fourier pair, $\Gamma(p)$ is the *product* of $\Phi_1(p)$ and $\Phi_2(p)$, the Fourier pairs of $F_1(x)$ and $F_2(x)$. Symbolically:

$$F_1(x) * F_2(x) \rightleftharpoons \Phi_1(p).\Phi_2(p) \tag{2.5}$$

The applications of this theorem are manifold and profound. Its proof is elementary:

$$C(x) = \int_{-\infty}^{\infty} F_1(x')F_2(x - x')\,dx'$$

by definition.

Fourier transform both sides (and note that, because the limits are $\pm\infty$, x' is a dummy variable and can be replaced by any other symbol not already in use):

$$\Gamma(p) = \int_{-\infty}^{\infty} C(x)e^{2\pi ipx}dx = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F_1(x')F_2(x - x')e^{2\pi ipx}dx'dx \tag{2.6}$$

Introduce a new variable $y = x - x'$. Then during the x-integration x' is held constant and $dx = dy$

$$\Gamma(p) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F_1(x')F_2(y)e^{2\pi ip(x'+y)}dx'dy$$

which can be separated to give:

$$\begin{aligned}\Gamma(p) &= \int_{-\infty}^{\infty} F_1(x')e^{2\pi ipx'}dx'.\int_{-\infty}^{\infty} F_2(y)e^{2\pi ipy}dy \\ &= \Phi_1(p).\Phi_2(p)\end{aligned}$$

2.3.2 Examples of convolutions

One of the chief uses of convolutions is to generate new functions which are easy to transform using the convolution theorem.

2.3.2.1 Convolution of a function with a δ-function, $\delta(x-a)$

$$C(x) = \int_{-\infty}^{\infty} F(x-x')\delta(x'-a)dx' = F(x-a)$$

by the properties of δ-functions. This can be written symbolically as:

$$F(x) * \delta(x-a) = F(x-a)$$

Applying the convolution theorem to this is instructive as it yields the shift theorem:

$$F(x) \rightleftharpoons \Phi(p); \qquad \delta(x-a) \rightleftharpoons e^{-2\pi ipa}$$

so that $F(x-a) = F(x) * \delta(x-a) \rightleftharpoons \Phi(p)e^{-2\pi ipa}$

More interesting is the convolution of a pair of δ-functions with another function:

$$[\delta(x-a) + \delta(x+a)] \rightleftharpoons 2\cos 2\pi pa$$

hence:

$$[\delta(x-a) + \delta(x+a)] * F(x) \rightleftharpoons 2\cos 2\pi pa.\Phi(p) \tag{2.7}$$

and this is illustrated in Fig. 2.5. The Fourier transform of a Gaussian $g(x) = e^{-x^2/a^2}$ is, from Chapter 1, $a\sqrt{\pi}e^{-\pi^2p^2a^2}$. The convolution of two unequal Gaussian curves, $e^{-x^2/a^2} * e^{-x^2/b^2}$ can then be done, either as a tiresome exercise in elementary calculus, or by the convolution theorem:

$$e^{-x^2/a^2} * e^{-x^2/b^2} \rightleftharpoons ab\pi e^{-\pi^2p^2(a^2+b^2)}$$

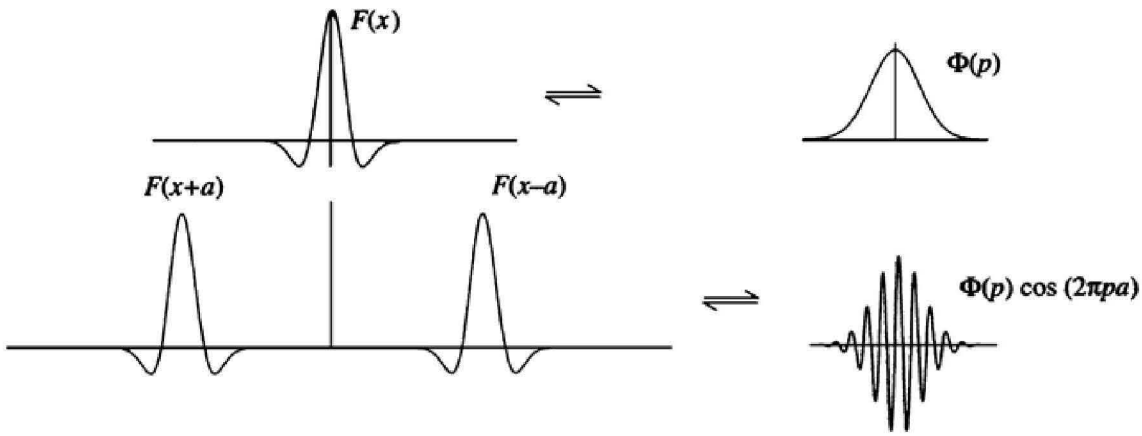


Fig. 2.5. Convolution of a pair of δ-functions with $F(x)$, and its transform.

Fig. 2.6. The triangle function, $\Lambda_a(x)$, as the convolution of two top-hat functions.

and the Fourier transform of the right-hand side is

$$\frac{ab\sqrt{\pi}}{\sqrt{a^2+b^2}}e^{-x^2/(a^2+b^2)} \tag{2.8}$$

so that we arrive at a useful practical result:

The convolution of two Gaussians of width parameters a and b is another Gaussian of width parameter $\sqrt{a^2+b^2}$

or, to put it another way, the resulting half-width is the Pythagorean sum of the two component half-widths.

The convolution of two equal top-hat functions is a good example of the power of the convolution theorem. It can be seen by inspection that the convolution of two top-hat functions, each of height h and width a is going to be a triangle, usually called the 'triangle-function' and denoted by $\Lambda_a(x)$, with height h^2a and base length $2a$.

The Fourier transform of this triangle function can be done by elementary integration, splitting the integral into two parts: $x = -a \to 0$ and $x = 0 \to a$. This too, is tiresome. On the other hand, it is trivial to see that if $h\Pi_a(x) \rightleftharpoons ah.\text{sinc}(\pi pa)$ then $h^2a\Lambda_a(x) \rightleftharpoons a^2h^2\text{sinc}^2\pi pa$

2.3.2.2 *The autocorrelation theorem*

This is superficially similar to the convolution theorem but it has a different physical interpretation. This will be mentioned later in connection with the Wiener–Khinchine theorem. The autocorrelation function of a function $F(x)$ is defined as:

$$A(x) = \int_{-\infty}^{\infty} F(x')F(x+x')\,dx'$$

The process of autocorrelation can be thought of as a multiplication of every point of a function by another point at distance x' further on, and then summing all the products: or like a convolution as described earlier, but with identical functions and without taking the mirror-image of one of the two.

There is a theorem similar to the convolution theorem.

Beginning with the definition:

$$A(x) = \int_{-\infty}^{\infty} F(x')F(x+x')\,dx'$$

Fourier transform both sides:

$$\Gamma(p) = \int_{-\infty}^{\infty} A(x)e^{2\pi ipx}dx = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F(x')F(x+x')e^{2\pi ipx}dx'dx$$

let $x + x' = y$. Then if x' is held constant, $dx = dy$

$$\Gamma(p) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F(x')F(y)e^{2\pi ip(y-x')}dx'dy$$

which can be separated to

$$\begin{aligned}\Gamma(p) &= \int_{-\infty}^{\infty} F(x)e^{-2\pi ipx'}dx'.\int_{-\infty}^{\infty} F(y)e^{2\pi ipy}dy \\ &= \Phi^{\star}(p).\Phi(p)\end{aligned}$$

so that

$$A(x) \rightleftharpoons |\ \Phi(p)\ |^2$$

The Wiener–Khinchine theorem, to be described in Chapter 4, may be thought of as a physical version of this theorem. It says that if $F(t)$ represents a signal, then its autocorrelation is (apart from a constant of proportionality) the Fourier transform of its power spectrum, $|\ \Phi(\nu)\ |^2$.

2.4 The algebra of convolutions

You can think of convolution as a mathematical operation analogous to addition, subtraction, multiplication, division, integration and differentiation. There are rules for combining convolution with the other operations. It cannot be associated with multiplication for example, and in general:

$$[A(x) * B(x)].C(x) \neq A(x) * [B(x).C(x)]$$

But convolution signs and multiplication signs can be exchanged across a Fourier transform symbol, and this is very useful in practice. For example:

$$[A(x) * B(x)].[C(x) * D(x)] \rightleftharpoons [a(p).b(p)] * [c(p).d(p)]$$

(Obviously upper case and lower case letters have been used to associate Fourier pairs) and as further examples:

$$A(x) * [B(x).C(x)] \rightleftharpoons a(p).[b(p) * c(p)]$$
$$[A(x) + B(x)] * [C(x) + D(x)] \rightleftharpoons [a(p) + b(p)].[c(p) + d(p)]$$
$$[A(x) * B(x) + C(x).D(x)].E(x) \rightleftharpoons [a(p).b(p) + c(p) * d(p)] * e(p)$$

So far as we use Fourier transforms in physics and engineering, we are concerned mostly with functions and manipulations like this to solve problems, and fluency in this relatively easy algebra is the key to success. Computation, rather than calculation is involved, and there is much software available to compute Fourier transforms digitally. However, most computation is done using complex exponentials and these involve the full complex transform. A later chapter deals with this subject.

2.5 Other theorems

2.5.1 The derivative theorem

If $\Phi(p)$ and $F(x)$ are a Fourier pair:

$$F(x) \rightleftharpoons \Phi(p), \text{ then } \quad dF/dx \rightleftharpoons -2\pi ip\Phi(p)$$

Proofs are elementary. You can integrate dF/dx by parts or you can differentiate $F(x)$:

$$F(x) = \int_{-\infty}^{\infty} \Phi(p)e^{-2\pi ipx}dp$$

differentiate with repect to x:

$$dF/dx = \int_{-\infty}^{\infty} -2\pi ip\Phi(p)e^{-2\pi ipx}dp$$
$$= -2\pi i \int_{-\infty}^{\infty} p\Phi(p)e^{-2\pi ipx}dp \tag{2.9}$$

and the right-hand side is $-2\pi i$ times the Fourier transform of $p\Phi(p)$.

Example 1: the top-hat function $\Pi_a(x) \rightleftharpoons a\,\text{sinc}\pi pa$. If the top-hat function is differentiated with respect to x, the result is a pair of δ-functions at the points where the slope was infinite:

$$\frac{d\Pi_a(x)}{dx} = \delta(x + a/2) - \delta(x - a/2)$$

Transforming both sides:

$$\delta(x + a/2) - \delta(x - a/2) \rightleftharpoons e^{-\pi ipa} - e^{\pi ipa} = -2i \sin \pi pa$$
$$= -2\pi ip[a \operatorname{sinc}(\pi pa)]$$

The theorem extends to further derivatives:

$$d^n F(x)/dx^n \rightleftharpoons (-2\pi ip)^n \Phi(p)$$

and much use is made of this in mathematics.

Example 2: if the moment of inertia about the y-axis of a symmetrical curve is infinite, its Fourier transform has a cusp at the origin.
Because:

$$\int_{\infty}^{\infty} f(x)dx = \phi(0)$$

and then if

$$\left(\frac{\partial^2 f}{\partial x^2}\right)_{x=0} = -4\pi^2 \int_{-\infty}^{\infty} p^2 \phi(p)\, dp = \infty$$

there is a discontinuity in $(\partial f/\partial x)$ at the origin.

Example 3: the differential equation of simple harmonic motion is:

$$md^2F(t)/dt^2 + kF(t) = 0$$

where $F(t)$ is the displacement of the oscillator from equilibrium at time t. If we Fourier-transform this equation, $F(t)$ becomes $\Phi(\nu)$ and d^2F/dt^2 becomes $-4\pi^2\nu^2\Phi(\nu)$. The equation then becomes:

$$\Phi(\nu)(k/m - 4\pi^2\nu^2) = 0$$

which, apart from the trivial solution $\Phi(\nu) = 0$ requires $\nu = \pm 2\pi\sqrt{k/m}$ and this is just a small taste of the power which is available for the solution of differential equations using Fourier transforms.

2.5.2 The convolution derivative theorem

$$\frac{d}{dx}[F_1(x) * F_2(x)] = F_1(x) * \frac{dF_2(x)}{dx} = \frac{dF_1(x)}{dx} * F_2(x) \qquad (2.10)$$

The derivative of the convolution of two functions is the convolution of either of the two with the derivative of the other. The proof is simple and is left as an exercise.

2.5.3 Parseval's theorem

This is met under various guises. It is sometimes called 'Rayleigh's theorem' or simply the 'Power theorem'. In general it states:

$$\int_{-\infty}^{\infty} F_1(x)F_2^*(x)\,dx = \int_{-\infty}^{\infty} \Phi_1(p)\Phi_2^*(p)\,dp \tag{2.11}$$

where * denotes a complex conjugate.
The proof of the theorem is in the Appendix.

Two special cases of particular interest are:

$$\frac{1}{P}\int_0^P |\,F(x)\,|^2\,dx = \sum_{-\infty}^{\infty}\left(a_n^2 + b_n^2\right) = \frac{A_0^2}{4} + \frac{1}{2}\sum_{1}^{\infty}\left[A_n^2 + B_n^2\right] \tag{2.12}$$

which is used for finding the power in a periodic waveform, and

$$\int_{-\infty}^{\infty} |\,F(x)\,|^2\,dx = \int_{-\infty}^{\infty} |\,\Phi(p)\,|^2\,dp \tag{2.13}$$

for non-periodic Fourier pairs.

2.5.4 The sampling theorem

This is also known as the 'cardinal theorem' of interpolary function theory, and originated with Whittaker[3], who asked and answered the question: how often must a signal be measured (sampled) in order that all the frequencies present should be detected? The answer is: the sampling interval must be the reciprocal of twice the highest frequency present.

The theorem is best illustrated with a diagram (Fig. 2.7). The highest frequency is sometimes called the 'folding frequency', or alternatively the 'Nyquist' frequency, and is given the symbol ν_f.

Suppose that the frequency spectrum, $\Phi(\nu)$, of the signal is symmetrical about the origin and stretches from $-\nu_f$ to ν_f. The convolution of this with a Dirac comb of period $2\nu_0$ provides a periodic function and the Fourier transform of this periodic function is the *product* of a Dirac comb with the original signal (and, to be strict, its reflection in the origin): in other words it is the set of Fourier coefficents in the series representing the periodic function. The periodic function is known provided the coefficients are known, and the coefficients are the values of the original signal $F(t)$, at intervals $1/2\nu_f$, multiplied by a suitable constant. The more coeffcients are known, the more harmonics can be added to make the spectrum, and more detail can be seen in the function when it is

[3] J. M. Whittaker, *Interpolary Function Theory* Cambridge University Press, Cambridge, 1935.

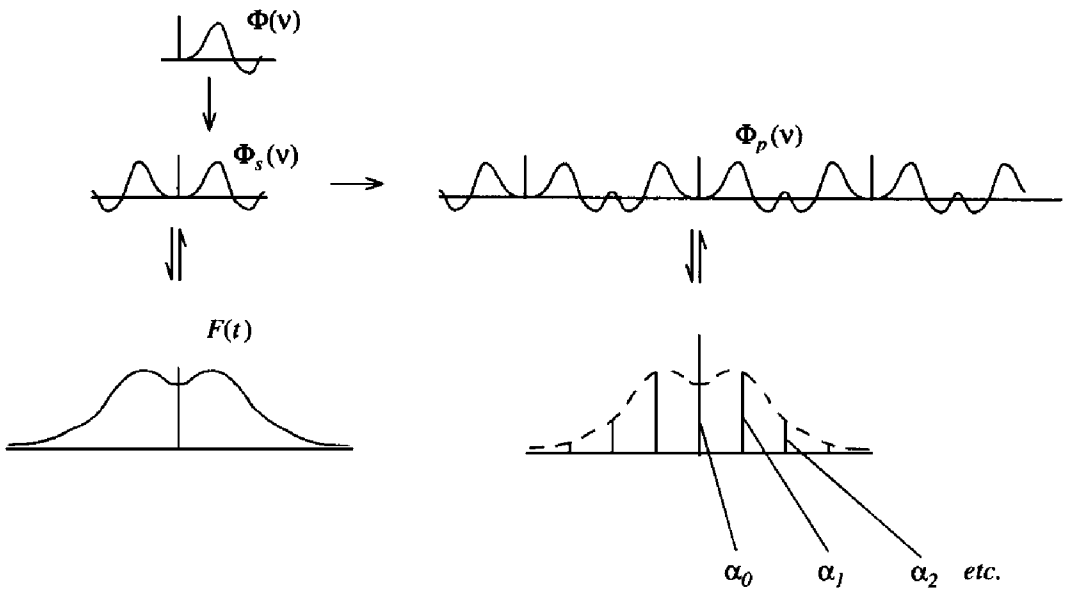


Fig. 2.7. The sampling theorem.

reconstructed. With the help of the interpolation theorem (below) all the points between the sample points can be filled in.

Formally, the process can be written, with $F(t)$ and $\Phi(\nu)$ a Fourier pair as usual. The Fourier transform of $F(t)Ш_a(t)$ is:

$$\int_{-\infty}^{\infty} F(t)Ш_a(t)e^{-2\pi i\nu t}dt = \Phi(\nu) * Ш_{1/a}(\nu)$$

rewrite the left-hand side as:

$$\int_{-\infty}^{\infty} F(t)\sum_{n=-\infty}^{\infty}\delta(t-na)e^{-2\pi i\nu t}dt = \sum_{n=-\infty}^{\infty}\int_{-\infty}^{\infty} F(t)\delta(t-na)e^{-2\pi i\nu t}dt$$
$$= \sum_{n=-\infty}^{\infty} F(na)e^{-2\pi i\nu na} = \Phi'(\nu)$$

The left-hand side is now a Fourier series, so that $\Phi'(\nu)$ is a periodic function, the convolution of $\Phi(\nu)$ with a Dirac comb of period $1/a$. The constraint is that $\Phi(\nu)$ must occupy the interval $-1/2a$ to $1/2a$ only; in other words, $1/a$ is twice the highest frequency in the function $F(t)$, in accordance with the sampling theorem.

2.6 Aliasing

In the sampling theorem it is strictly necessary that the signal should contain no power at frequencies above the folding frequency. If it does, this power will be 'folded' back into the spectrum and will appear to be at a lower frequency. If

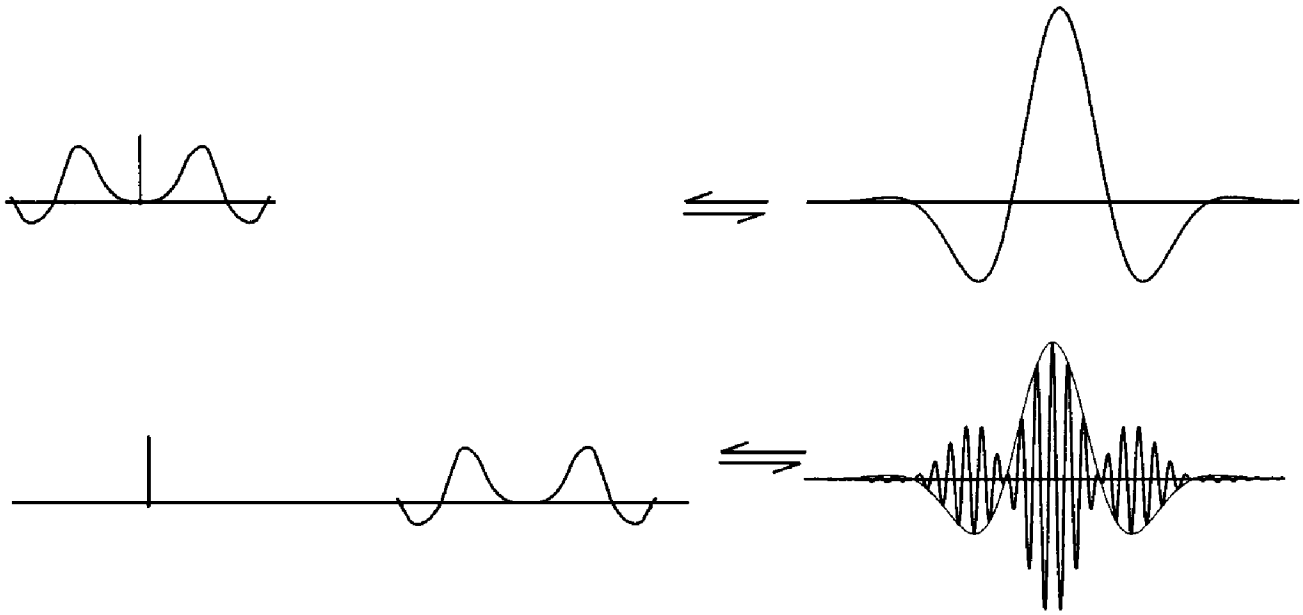


Fig. 2.8. A signal occupying a high alias of a fundamental in frequency space, and its recovery by deliberate undersampling or 'demodulating'.

the frequency is $\nu_f + \nu_a$ it will appear to be at $\nu_f - \nu_a$ in the spectrum. If it is at twice the folding frequency it will appear to be at zero frequency. For example, a sine-wave sampled at intervals $a, 2\pi + a, 4\pi + a, \ldots$ will give a set of samples which are identical. There are, in effect, 'beats' between the frequency and the sampling rate. It is always necessary to take precautions when examining a signal to be sure that a given 'spike' corresponds to the apparent frequency. This can be done either by deliberate filtering of the incoming signal, or by making several measurements at different sampling frequencies. The former is the obvious method but not necessarily the best: if the signal is in the form of a pulse and is in a noisy environment, a lot of the power can be lost by filtering.

Aliasing can be put to good use. If the frequency band stretches from ν_0 to ν_1 the empty frequency band between ν_0 and 0 can be divided into a number of equal frequency intervals each less than $2(\nu_1 - \nu_0)$ The sampling interval then need be only $1/2(\nu_1 - \nu_0)$ instead of $1/2\nu_1$. This is a way of demodulating the signal, and the spectrum that is recovered appears to occupy the first alias although the original occupied a possibly much higher one. The process is illustrated in Fig. 2.8.

2.6.1 The interpolation theorem

This too comes from Whittaker's interpolary function theory. If the signal samples are recorded, the values of the signal in between the sample points can be

calculated. The spectrum of the signal can be regarded as the product of the periodic function with a top-hat function of width $2vf$. In the signal, each sample is replaced by the convolution of the sinc-function with the corresponding δ-function. Each sample, $a_n\delta(t - t_n)$ is replaced by the sinc-function, a_n sinc$\pi\nu_f$ and each sinc-function conveniently has zeros at the positions of all the other samples (this is hardly a coincidence, of course) so that the signal can be reconstructed from a knowledge of its samples which are the coefficients of the Fourier series which form its spectrum.

This is much used in practical physics, when digital recording of data is common, and generally the signal at a point can be well enough recovered by a sum of sinc-functions over twenty or thirty samples on either side. The reason for this is that unless there is a very large amplitude to a sample at some distant point, the sinc-function at a distance of 30π from the sample has fallen to such a low value that it is lost in the noise. It depends obviously on practical details such as the signal/noise ratio in the original data: and more importantly, on the absence of any power at frequencies higher than the folding frequency.

Stated formally, the signal $F(t)$ sampled at times $0, t_0, 2t_0, 3t_0, 4t_0, 5t_0, \ldots$ can be computed at any intermediate point t as the sum

$$F(nt_0 + t) = \sum_{m=-N}^{N} F\{(n+m)t_0\}\,\mathrm{sinc}\,[\pi(m - t/t_0)]$$

where N, infinite in theory, is about $20 \rightarrow 30$ in practice. The sum can not be computed accurately near the ends of the data stream and there is a loss of N samples at each end unless fewer samples are taken there.

2.6.2 The similarity theorem

This is fairly obvious: if you stretch $F(x)$ so that it is twice as wide, then $\Phi(p)$ will be only half as wide, but twice as high as it was. Formally:

$$\text{if } F(x) \rightleftharpoons \Phi(p) \text{ then } F(ax) \rightleftharpoons |\,(1/a)\,|\,\Phi(p/a)$$

The proof is trivial, and done by substituting $x = ay$, $dx = ady$; $p = z/a$, $dp = (1/a)dz$. Because the integrals are between $-\infty$ and ∞, the variables for integration are 'dummy' and can be replaced by any other symbol not already in use.

2.7 Worked examples

The saw-tooth used in Chapter 1 shows an interesting result using Parseval's theorem.

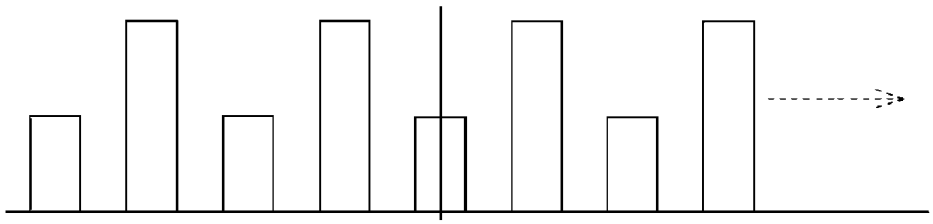


Fig. 2.9.

The nth sine-coefficient as we saw, is $(-1)^{n+1}2h/n\pi$. The sum to infinity of the squares is:

$$\sum_{n=1}^{\infty}\frac{4h^2}{\pi^2 n^2} = \frac{2}{P}\int_{-P/2}^{P/2}\left[\frac{2hx}{P}\right]^2 dx$$

$$= \frac{8h^2}{P^3}\left[\frac{x^3}{3}\right]_{-P/2}^{P/2}$$

$$= 2\frac{h^2}{3} = \frac{4h^2}{\pi^2}\sum_{n=1}^{\infty}\frac{1}{n^2}$$

so that finally:

$$\sum_{n=1}^{\infty}\frac{1}{n^2} = \frac{\pi^2}{6}$$

This is an example of an arithmetic result coming from a purely analytical calculation. As a way of computing π it is not very efficient: it is accurate to only six significant figures (3.14159) after one million terms. Using the fact that $\pi = 6\sin^{-1}(1/2)$, with $\sin^{-1}$ obtained by integrating $1/\sqrt{1-x^2}$ term-by-term, is much more efficient.

In a rectangular waveform with pulses of length $a/4$ separated by spaces of length $a/4$ and with alternate rectangles twice the height of their neighbours, the amplitude of the second harmonic is greater than the fundamental amplitude.

The waveform can be represented by

$$F(t) = h\Pi_{\frac{a}{4}}(t) * [Ш_a(t) + Ш_{\frac{a}{2}}(t)]$$

The Fourier transform is:

$$\Phi(\nu) = (ah/4)\mathrm{sinc}(\pi\nu a/4)\,.\left[\frac{1}{a}Ш_{\frac{1}{a}}(\nu) + \frac{2}{a}Ш_{\frac{2}{a}}(\nu)\right]$$

and the teeth of this Dirac comb are at $\nu = 1/a, 2/a\ldots$, with heights

$$h/4\mathrm{sinc}(\pi/4), 3h/4\mathrm{sinc}(\pi/2), h/4\mathrm{sinc}(3\pi/4), \ldots,$$

and the ratio of heights of the first and second harmonics is $3/\sqrt{2}$.

This effect can be seen in astronomy or radioastronomy when searching for pulsars: the 'interpulses', between the main pulses generate extra power in the second harmonic and can make it larger than the fundamental.

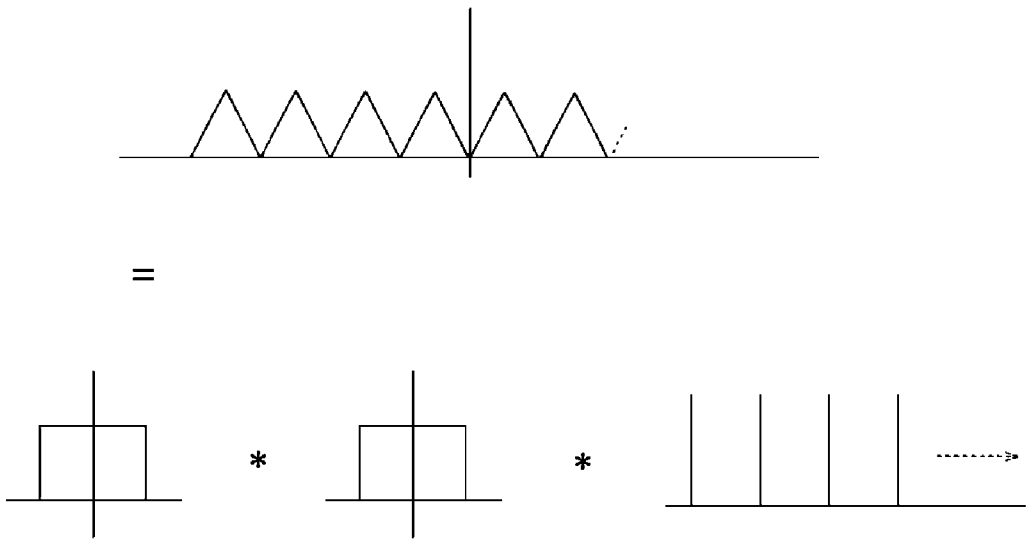


Fig. 2.10. The double-sawtooth waveform.

The double-sawtooth waveform: This can not be regarded as the convolution of two rectangular waveforms of equal mark-space[4] ratio, since the effect of integration is to give an embarrassing infinity. Instead it is the convolution of a top-hat of width a with another identical top-hat and with a Dirac comb of period $2a$. Thus:

$$\Pi_a(t) * \Pi_a(t) * \text{Ш}_{2a}(t) \rightleftharpoons (a/2)\,\text{sinc}^2 \pi \nu a \;.\; \text{Ш}_{\frac{1}{2a}}(\nu)$$

So that the amplitudes, which occur at $\nu = 1/2a, 1/a, 3/2a, \ldots$ are: $2a/\pi^2$, 0, $2a/9\pi^2$, 0, $2a/25\pi^2, \ldots$

[4] The term 'equal mark-space ratio' comes from radio jargon, and implies that the signal is zero for the same interval that it is not.

Chapter 3

Applications 1: Fraunhofer diffraction

3.1 Fraunhofer diffraction

The application of Fourier theory to Fraunhofer diffraction problems and to interference phenomena generally, was hardly recognized before the late 1950s. Consequently, only textbooks written since then mention the technique. Diffraction theory, of which interference is only a special case, derives from Huygens' principle: that every point on a wavefront which has come from a source can be regarded as a secondary source: and that all the wavefronts from all these secondary sources combine and interfere to form a new wavefront.

Some precision can be added by using calculus. In the diagram (Fig. 3.1), suppose that at O there is a source of 'strength' q, defined by the fact that at A, a distance r from O there is s 'field', E of strength $E = q/r$. Huygens' principle is now as follows:

If we consider an area dS on the surface S we can regard it as a source of strength EdS giving at B, a distance r' from A, a field $E' = qdS/rr'$. All these elementary fields at B, summed over the transparent part of the surface S, each with its proper phase[1], give the resultant field at B. This is quite general – and vague.

In Fraunhofer diffraction we simplify. We assume:

- that only two dimensions need be considered. All apertures bounding the transparent part of the surface S are rectangular and of length unity perpendicular to the plane of the diagram.
- that the dimensions of the aperture are small compared with r'.
- that r is very large so that the field E has the same magnitude at all points on the transparent part of S, and a slowly varying or constant phase. (Another

[1] Remember: phase change $= (2\pi/\lambda) \times$ path change and the paths from different points on the surface S (which, being a wavefront, is a surface of constant phase) to B are all different.

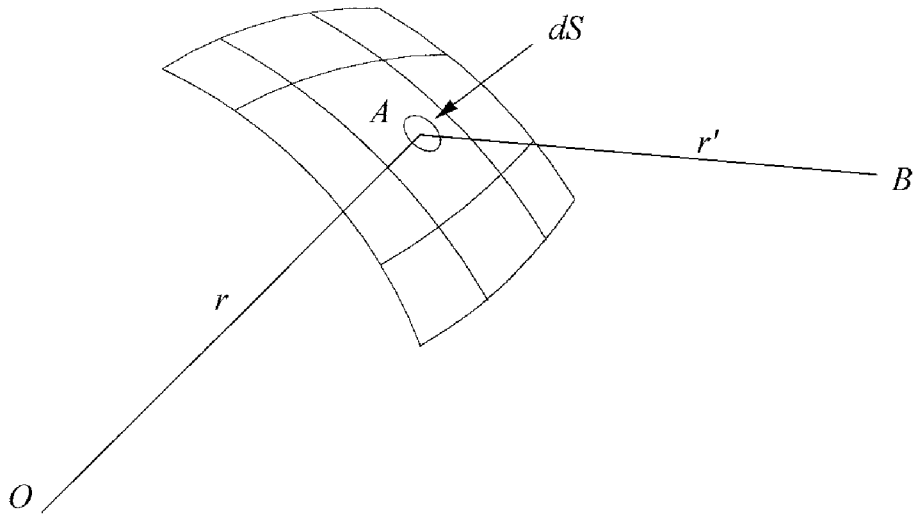


Fig. 3.1. Secondary sources in Fraunhofer diffraction.

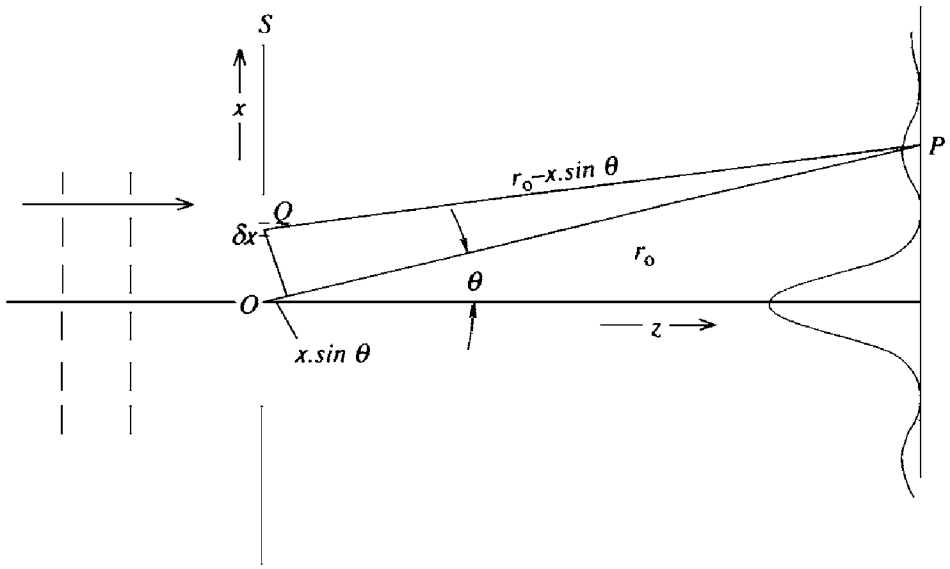


Fig. 3.2. Fraunhofer diffraction by a plane aperture.

way of putting it is to say that plane wavefronts arrive at the surface S from a source at $-\infty$).

- that the aperture S lies in a plane.

To begin, suppose that the source, O lies on a line perpendicular to the surface S, the diffracting aperture. Use Cartesian coordinates, x in the plane of S, and z perpendicular to this (x and z are traditional here). Then the magnitude of the field E at P can be calculated.

Consider an infinitesimal strip at Q, of unit length perpendicular to the $x-z$ plane, of width dx and distance x above the z-axis. Let the field strength[2] there be $E = E_0 e^{2\pi i\nu t}$. Then the field strength at P from this source will be:

$$d\overline{E}(P) = E_0 dx e^{2\pi i\nu t} e^{-2\pi i r'/\lambda}$$

where r' is the distance QP. The exponent in this last factor is the *phase difference* between Q and P.

For convenience, choose a time t so that the phase of the wavefront is zero at the plane S, i.e. $t = 0$. Then at P:

$$\overline{E}(P) = \int_{aperture,S} E_0 dx e^{-2\pi i r'/\lambda}$$

and the aperture S may have opaque spots or partially transmitting spots, so that E_0 is generally a function of x.

This is not yet a useable expression.

Now, because $r' \gg x$ (the condition for Fraunhofer diffraction) we can write:

$$r' \approx r_0 - x \sin\theta$$

and then the field $\overline{E}$ at P is obtained by summing all the infinitesimal contributions from the secondary sources like that at Q, and remembering to include the phase-factor for each. The result is:

$$\overline{E} = E_0 e^{-2\pi i r_0/\lambda} \int_{aperture} e^{2\pi i x \sin\theta/\lambda}\, dx$$

and if we write $\sin\theta/\lambda = p$ we have, finally:

$$\overline{E} = E_0 e^{-2\pi i r_0/\lambda} \int_{-\infty}^{\infty} A(x) e^{2\pi i p x}\, dx$$

where $A(x)$ is the 'aperture function' which describes the transparent and opaque parts of the screen S. The result of the Fourier transform is to give the *amplitude* diffracted through an angle θ. Where it appears on a screen depends on the distance to the screen, and on whether the screen is perpendicular to the z-direction and other geometrical factors[3].

The important thing to remember is: that diffraction of a certain wavelength at a certain aperture is always *through an angle*: the variable p conjugate to x

[2] As usual, we use complex variables to represent *real* quantities – in this case the electric field strength. This complex variable is called the 'analytic' signal and the real part of it represents the actual physical quantity at any time at any place.

[3] This is all an approximation: in fact the field *outside* the diffracting aperture is not exactly zero and depends in practice on whether the opaque part of the screen is conducting or insulating. This is a subtlety which can safely be left to post-graduate students.

is $\sin\theta/\lambda$ and it is θ which matters. Diffraction theory alone says nothing about the size of the pattern: that depends on geometry.

Very often, in practice, the diffracting aperture is followed by a lens, and the pattern is observed at the focal plane of this lens. The approximation, that $r' = r_0 - x\sin\theta$ is now exact, since the image of the focal plane, seen from the diffracting aperture, is at infinity.

Problems in Fraunhofer diffraction can thus be reduced to writing down the aperture function, $A(x)$, and taking its Fourier transform. The result gives the amplitude in the diffraction pattern on a screen at a large distance from the aperture. For example, for a simple parallel-sided slit of width a, the aperture function, $A(x)$ is $\Pi_a(x)$. For two parallel-sided slits of width a separated by a distance b between their centres, $A(x) = \Pi_a(x) * [\delta(x - b/2) + \delta(x + b/2)]$, and so on. Apertures of various sizes are now encompassed by the same formula and the amplitude of the light (or sound, or radio waves or water waves) diffracted by the aperture through an angle θ can be calculated. The *intensity* of the wave is given by the r.m.s. value of the amplitude $\times$ (complex conjugate) and the factor $e^{2\pi i r_0/\lambda}$ disappears when this is done.

If the original source is not on the z-axis, then the amplitude of E at $z = 0$ contains a phase factor, as in Fig. 3.3.

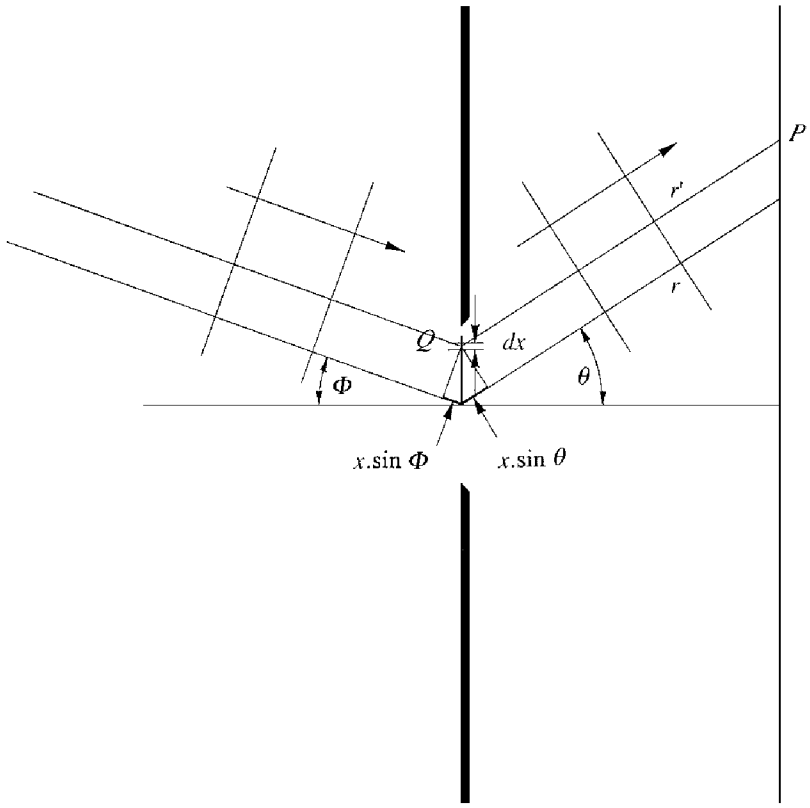


Fig. 3.3. Oblique incidence from a source not on the z-axis.

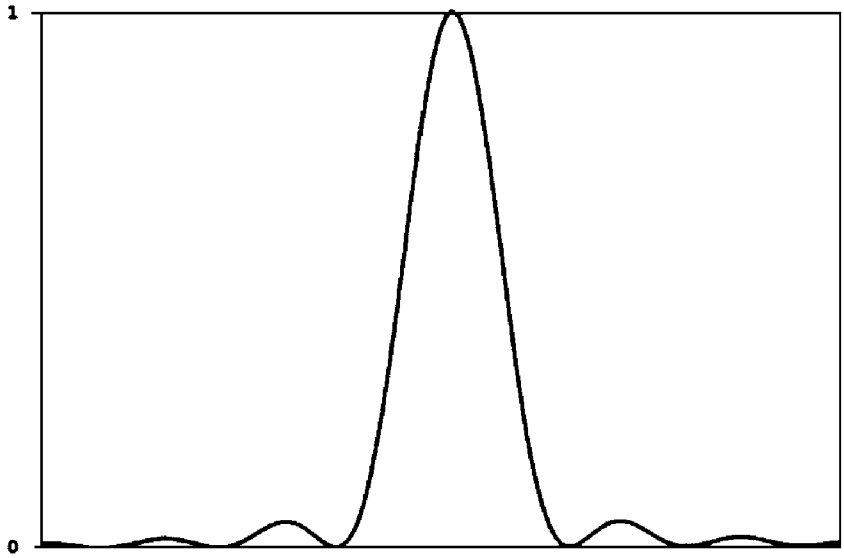


Fig. 3.4. The intensity pattern, $\text{sinc}^2(\pi a \sin\theta/\lambda)$, from diffraction at a single slit.

$W - W'$ is a wavefront (a surface of constant phase) and if we choose a moment when the phase is zero at the origin, the phase at x at that moment is given by $(2\pi/\lambda)x.\sin\phi$, and the phase factor that must multiply E_0 is $e^{(-2\pi i/\lambda)x\sin\phi}$.

The magnitude at P is then

$$\overline{E} = E_0 e^{2\pi i r_0/\lambda} \int_{-\infty}^{\infty} A(x) e^{(-2\pi i/\lambda)x(\sin\theta+\sin\phi)}\, dx$$

and when the Fourier transform is done, the oblique incidence is accounted for by remembering that $p = (\sin\theta + \sin\phi)/\lambda$.

3.2 Examples

3.2.1 Single-slit diffraction, normal incidence

For a single slit with parallel sides, of width a, the aperture function is $A(x) = \Pi_a(x)$. Then:

$$\overline{E} = k.\text{sinc}(\pi a p) = k.\text{sinc}(\pi a \sin\theta/\lambda)$$

(where k is the constant[4] $E_0 a e^{-2\pi i r_0/\lambda}$), and the intensity is this multiplied by its complex conjugate:

$$\overline{E E^*} = I(\theta) = |k|^2.\,\text{sinc}^2(\pi a \sin\theta/\lambda) \qquad (3.1)$$

[4] For most practical purposes, the *unimportant* consant.

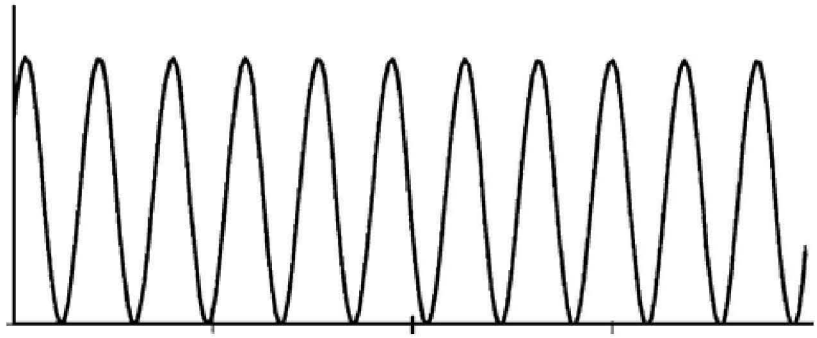


Fig. 3.5. Intensity pattern from interference between two point sources.

3.2.2 Two point sources at $\pm b/2$ (for example, two antennae, transmitting in phase from the same oscillator)

Then:

$$A(x) = \delta(x - b/2) + \delta(x + b/2)$$

and the Fourier transform of this is [Chapter 1, equation (1.19)]:

$$\overline{E} = 2k.\ \cos(\pi b \sin\theta/\lambda)$$

and the intensity is this amplitude multiplied by its complex conjugate:

$$I(\theta) = 4\ |k|^2\ \cos^2(\pi b \sin\theta/\lambda)$$

3.2.3 Two slits, each of width a,with centres separated by a distance b (Young's slits, Fresnel's biprism, Lloyd's mirror, Rayleigh's refractometer, Billet's split-lens)

$$A(x) = \Pi_a(x) * [\delta(x - b/2) + \delta(x + b/2)]$$

Then, applying the convolution theorem:

$$I(\theta) = 4k^2 \mathrm{sinc}^2(\pi a \sin\theta/\lambda) cos^2(\pi b \sin\theta/\lambda)$$

3.2.4 Three parallel slits, each of width a. centres separated by a distance b

To simplify the algebra, put $\sin\theta/\lambda = p$

$$A(x) = \Pi_a(x) * [\delta(x - b) + \delta(x) + \delta(x + b)]$$

$$\overline{A}(p) = k\ \mathrm{sinc}(\pi pa)[e^{2\pi ibp} + 1 + e^{-2\pi ipb}]$$

$$= k\ \mathrm{sinc}(\pi pa)[2\cos(2\pi pb) + 1]$$

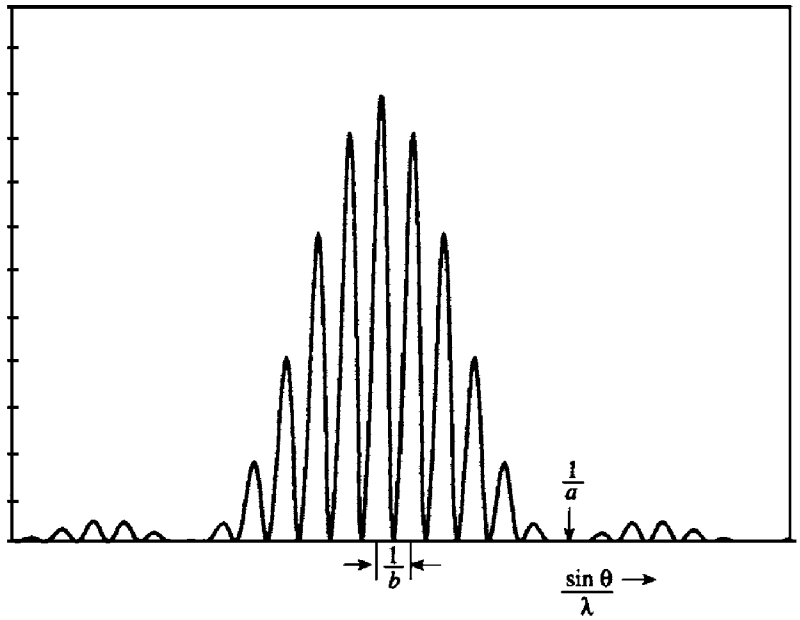


Fig. 3.6. Intensity pattern from interference between two slits of width a separated by a distance b.

and the intensity diffracted at angle θ is:

$$\begin{aligned} I(p) &= k^2 \operatorname{sinc}^2(\pi pa)[2\cos(4\pi pb) + 4\cos(2\pi pb) + 3] \\ &= k^2 \operatorname{sinc}^2(\pi a \sin\theta/\lambda)[2\cos(4\pi b \sin\theta/\lambda) + 4\cos(2\pi b \sin\theta/\lambda) + 3] \end{aligned}$$

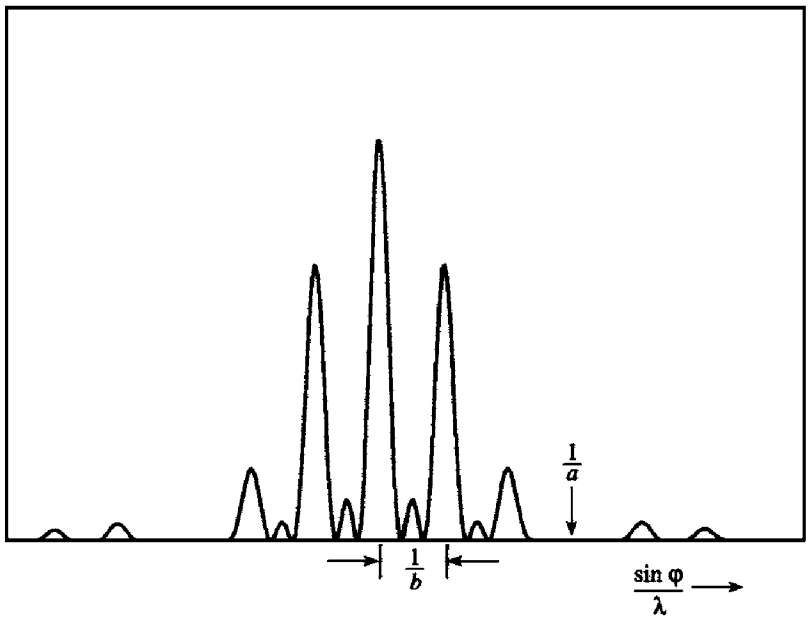


Fig. 3.7. Intensity pattern from interference between three slits of width a, separated by b.

3.2.5 The transmission diffraction grating

There are two obvious ways of representing the aperture function. In either case we assume that there are N slits, each of width w, each separated from its neighbours by a, the grating constant, and that N is a large ($10^4 \rightarrow 10^5$) number.

Then, since $A(x) = \Pi_w(x) * \text{Ш}_a(x)$ represents an infinitely wide grating, its width can be restricted by multiplying it by $\Pi_{Na}(x)$, so that the aperture function is:

$$A(x) = \Pi_{Na}(x).[\Pi_w(x) * \text{Ш}_a(x)]$$

Then the diffraction amplitude is:

$$\begin{aligned}\overline{E}(\theta) &= Na.\text{sinc}(\pi Na \sin\theta/\lambda) * [w.\text{sinc}(\pi w \sin\theta/\lambda).(1/a)\text{Ш}_{(1/a)}(\sin\theta/\lambda)] \\ &= Nw.\text{sinc}(\pi Na \sin\theta/\lambda) * [\text{sinc}(\pi w \sin\theta/\lambda).\text{Ш}_{(1/a)}(\sin\theta/\lambda)]\end{aligned}$$

(N.B. the convolution is with respect to $\sin\theta/\lambda$.)

A diagram here is helpful: the second factor (in the square brackets) is the product of a Dirac comb and a very broad (because w is very small) sinc-function; and the convolution of this with the first factor, a very narrow sinc-function, represents the diffraction produced by the whole aperture of the grating. Since the narrow sinc-function is reduced to insignificance by the time it has reached as far as the next tooth in the Dirac comb, the intensity distribution is this very narrow line profile $\text{sinc}^2(\pi Na \sin\theta/\lambda)$, reproduced at each tooth position with its intensity reduced by the factor $\text{sinc}^2(\pi wa \sin\theta/\lambda)$.

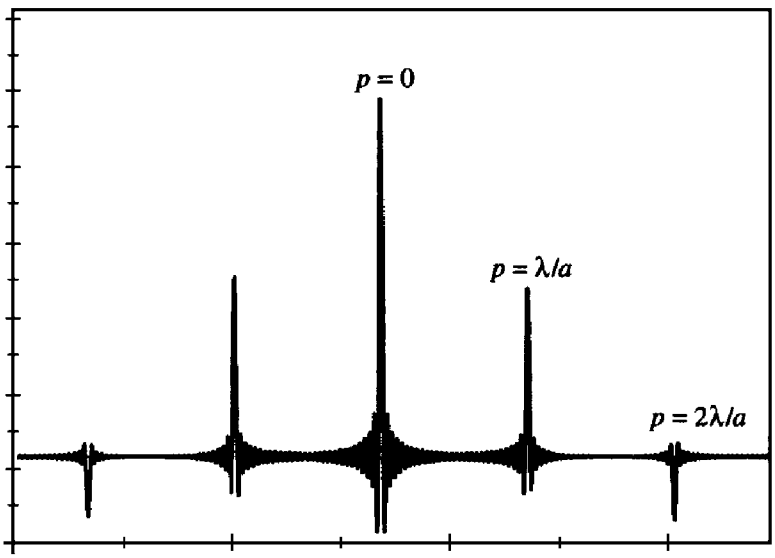


Fig. 3.8. Amplitude transmitted by a diffraction grating.

This is not precise, but is close enough for all practical purposes. To be precise, fastidious and pedantic, the aperture function, as described in the older optics textbooks, is:

$$A(x) = \sum_{n=0}^{N-1} \delta(x - na) * \Pi_w(x)$$

and since $\delta(x - na) \rightleftharpoons e^{2\pi inpa}$ the diffracted amplitude is:

$$\overline{E}(\theta) = k \ \mathrm{sinc}(\pi wp) \sum_{n=0}^{N-1} e^{2\pi inpa}$$

where $k = w.E_0 e^{-2\pi i r_0/\lambda}$. The third factor in the equation is the sum of a geometrical progression of common ratio $e^{2\pi ipa}$ and after a few lines of algebra the equation becomes:

$$\overline{E}(\theta) = k \ \mathrm{sinc}(\pi wp) e^{\pi i(N-1)pa)} \sin(\pi Npa)/\sin(\pi pa)$$

with $p = \sin\theta/\lambda$ as usual. The intensity is given by $\overline{E}(\theta)\overline{E}(\theta)^*$. The exponential factors disappear and if we write I_0 for E_0^2 the intensity distribution is:

$$I(\theta) = I_0. \left(\frac{\sin(\pi Npa)}{\sin(\pi pa)} \right)^2 \mathrm{sinc}^2(\pi wp) \tag{3.2}$$

If N is large, the first factor is very similar to a sinc^2-function, especially near the origin, where $\sin\pi pa \simeq \pi pa$, and although it is exact it yields no more information about the diffraction pattern details than the previous approximate derivation. Either way, the factor in the first bracket gives details about the line shape and the resolution to be obtained, and the third factor, the broad sinc^2-function, gives information about the intensities of the diffraction maxima in the pattern.

In particular, if a maximum for one wavelength λ falls at the same diffraction angle θ as the first zero of an adjacent wavelength $\lambda + \delta\lambda$ (the usual criterion for resolution in a grating spectrometer), the two values of p can be compared:

$$\text{for } \lambda \text{ at maximum, } \sin\theta \sim \theta = m\lambda/a$$

$$\text{for } \lambda \text{ at first zero, } \theta = m\lambda/a + \lambda/Na$$

which is the same angle as for $\lambda + \delta\lambda$ at maximum, i.e. $m(\lambda + \delta\lambda)/a$

$$\text{whence } \delta\lambda = \lambda/mN$$

which gives the theoretical resolution of the grating.

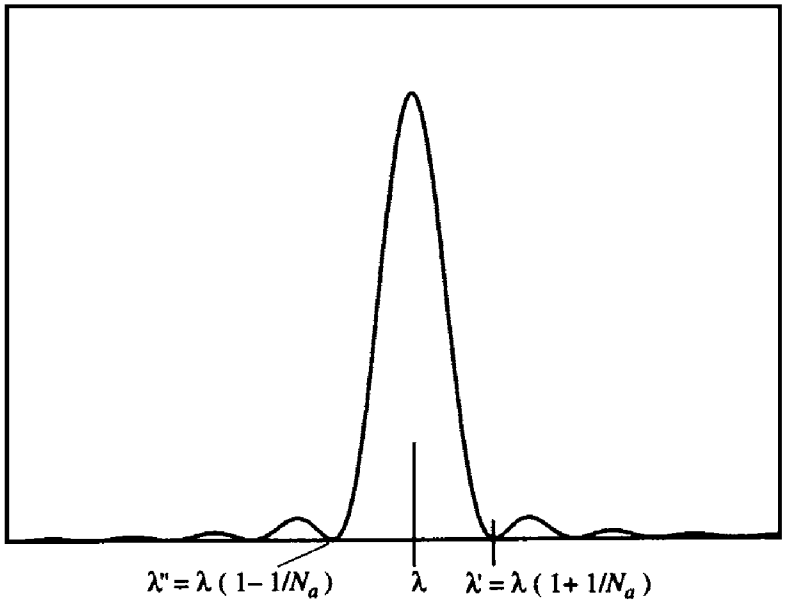


Fig. 3.9. The shape of a spectrum line from a grating. The profile is a sinc^2 function of the form $\mathrm{sinc}^2(\pi Nap)$.

Two points are worth noting.

(1) No one expects to get the full theoretical resolution from a grating. Manufacturing imperfections may reduce it in practice to $\sim 70\%$ of the theoretical value.
(2) Although this is the closest that two wavelengths can still produce separate images, more closely spaced wavelengths can be disentangled if the combined shape is known. The process of *deconvolution* can be used to enhance resolution if need be, although the improvement can be disappointing.

The sinc-function in Fig. 3.9 represents the amplitude near the diffraction image of a monochromatic spectrum line. Although the diffraction amplitude defines a direction, θ, in practice a lens or a mirror will focus all the radiation that comes from the grating at angle θ to a point on its focal surface.

The intensity distribution in the image will be the square modulus of the amplitude distribution, in this case a sinc^2-function, which has its width[5] determined by the width Na of the grating.

The minima at λ' and λ'' are at a wavelength difference $\pm 1/Na$, from the properties of the sinc^2-function.

Interesting things can be done to the amplitude of the radiation transmitted (or reflected) by the grating by covering the grating with a mask. A diamond-shaped

[5] By 'width' we mean here the Full Width at Half Maximum Intensity of the spectrum line, usually denoted by 'FWHM'.

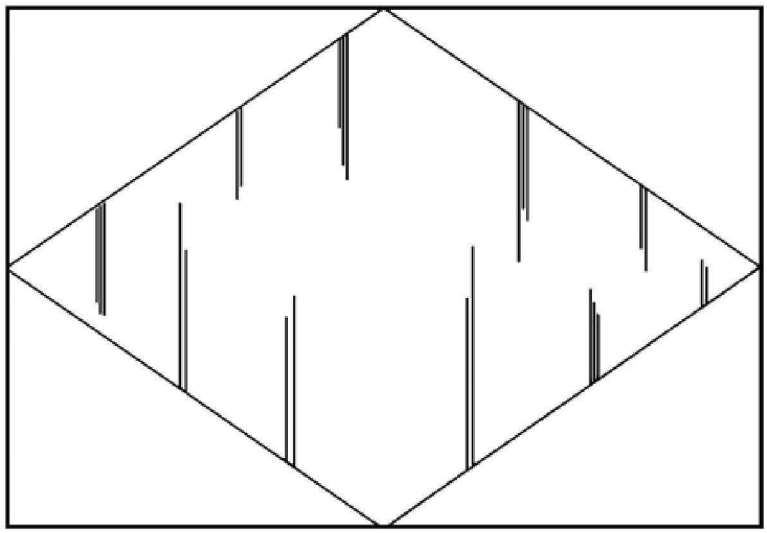


Fig. 3.10. Diffraction grating with a diamond-shaped apodising mask.

mask for example (Fig. 3.10) will change the aperture function from $\Pi_a(x)$ to $\Lambda_a(x)$ and the Fourier transform of the aperture function is then:

$$\overline{E}(\theta) = k\ \mathrm{sinc}^2(\pi(aN/2)\sin\theta/\lambda) * [\mathrm{sinc}(\pi w\ \sin\theta/\lambda).(1/a)\text{Ш}_{(1/a)}(\sin\theta/\lambda)]$$

The shape of the image of a monochromatic line is changed. Instead of $\mathrm{sinc}^2[\pi Na(\sin\theta/\lambda)]$, it becomes $\mathrm{sinc}^4[(\pi Na/2)(\sin\theta/\lambda)]$. The sinc^4 function is nearly twice as wide as the sinc^2 and the intensity of the light is reduced by a factor of 4, but the intensities of the 'side lobes' are reduced from 1.6×10^{-3} to 2.56×10^{-6} of the main peak intensity. This reduction is important if faint satellite lines are to be identified – for example in studies of fine structure or Raman-scattered lines – where the the satellite intensities are 10^{-6} of the parent or less. The process, which is widely used in optics and radioastronomy, is called *apodising*[6].

There are more subtle ways of reducing the side-lobe intensities by masking the grating. For example, a mask as in Fig. 3.11 allows the amplitude transmitted to vary sinusoidally across the aperture according to

$$\Pi_{Na}(x)[A + B\cos(2\pi x/Na)].$$

The Fourier transform of this is

$$\overline{E}(\theta) = Na\ \mathrm{sinc}(\pi pNa) * \{A\delta(p) + B/2[\delta(p - 1/Na) + \delta(p + 1/Na)]\}$$

and this is the sum of three sinc-functions, suitably displaced. Figure 3.13 illustrates the effect.

[6] From the Greek 'without feet', implying that the side-lobes are reduced or removed.

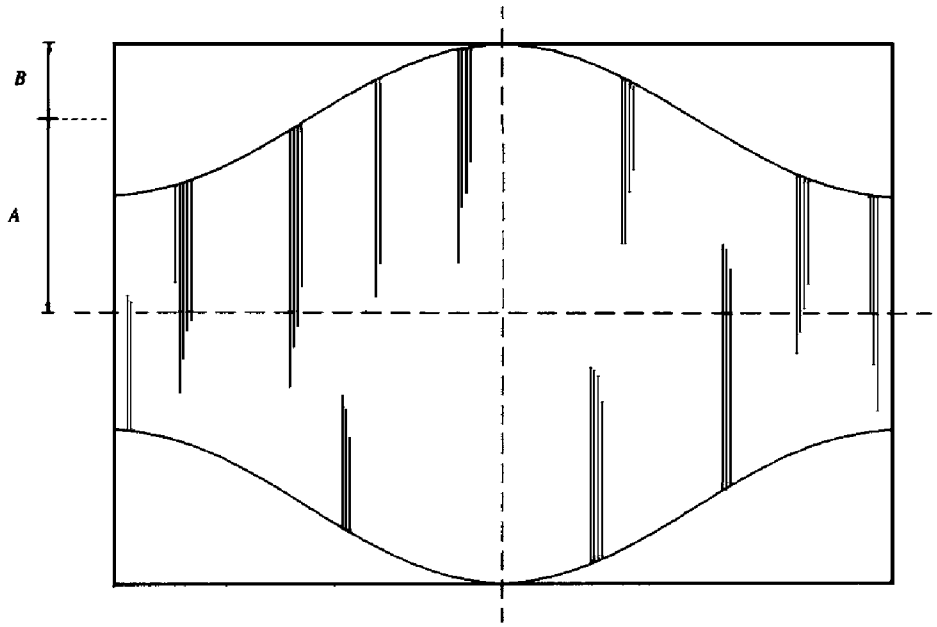


Fig. 3.11. An $A + B\cos(2\pi x/Na)$ apodising mask for a grating.

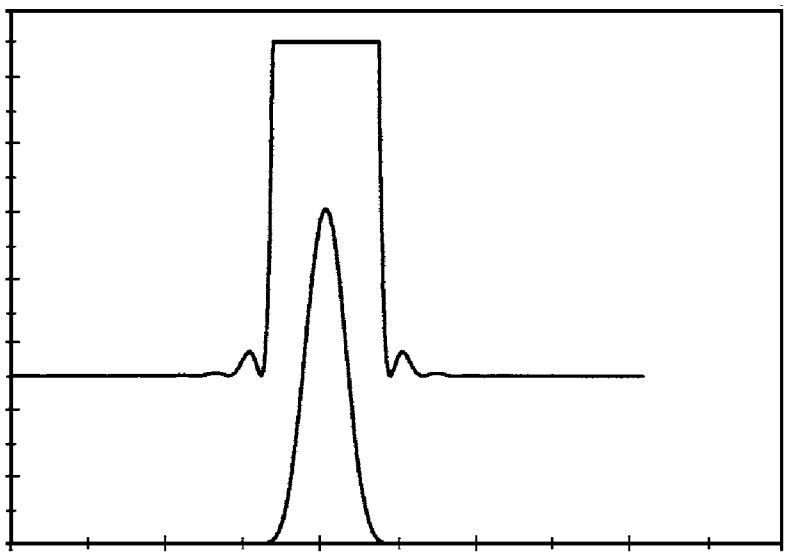


Fig. 3.12. The intensity-profile of a spectrum line from a grating with a sinusoidal apodising mask. The upper curve is the lower curve multiplied by $\times 1000$ to show the low level of the secondary maxima.

Even more complicated masking is possible and in general what happens is that the power in the side-lobes is redistributed according to the particular problem that is faced. The nearer side-lobes can be suppressed almost completely, for example and the power absorbed into the main peak or pushed out into the 'wings' of the line. Favourite values for A and B are $A = B = 0.5H$ and $A = 0.685H$, $B = 0.315H$ where H is the length of the grating rulings. (*Not* the ruled width of the grating.)

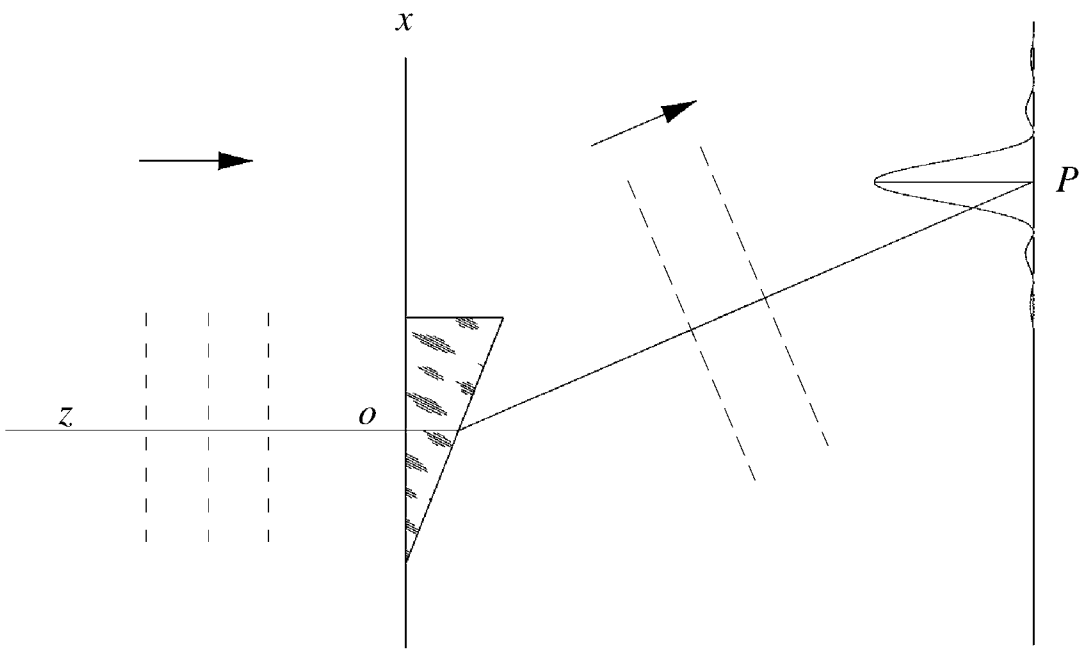


Fig. 3.13. A single-slit aperture with a prism and its displaced diffraction pattern.

3.2.6 Apertures with phase-changes instead of amplitude changes

The aperture function may be (indeed *must* be) bounded by a mask edge of finite size and it is possible – for example by introducing refracting elements – to change the phase as a function of x. A prism or lens would do this.

3.2.7 Diffraction at an aperture with a prism

Because 'optical' path is $n\times$ geometrical path, the passage of light through a distance x in a medium of refractive index n introduces an *extra* 'path' $(n-1)x$ compared with the same length of path in air or vacuum. Consequently there is a phase change $(2\pi/\lambda)(n-1)x$.

There is thus (see Fig. 3.14) a variation of *phase* instead of transmission across the aperture, so that the aperture function is complex. If the prism angle is ϕ and the aperture width is a, the thickness of the prism at its base is $a\tan\phi$ and when parallel wavefronts coming from $-\infty$ have passed through the prism, the phases at the apex and the base of the prism are 0 and $(2\pi/\lambda)(n-1)a\tan\phi$.

However, we can choose the phase to be zero at the centre of the aperture, and this is usually a good idea because it saves unnecessary algebra later on.

Then the phase at any point x in the aperture is $\zeta(x)=(2\pi/\lambda)x(n-1)\tan\phi$ and the aperture function describing the Huygens wavelets is:

$$A(x)=\Pi_a(x)e^{(2\pi i/\lambda)x(n-1)\tan\phi}$$

The Fourier transform of this, with $p = \sin\theta/\lambda$ as usual, is:

$$\overline{E}(\theta) = A \int_{-a/2}^{a/2} e^{(2\pi i/\lambda)x(n-1)\tan\phi} e^{2\pi i p x}\, dx$$

so that, after integrating and multiplying the amplitude distribution by its complex conjugate we get:

$$I(\theta) = A^2 a^2 \mathrm{sinc}^2\{a\pi[p + (n-1)\tan\phi/\lambda]\}$$

Notice that if $n = 1$ we have the same expression as in equation (3.1). Here we see that the shape of the diffraction function is identical, but that the principal maximum is shifted to the direction $p = \sin\theta/\lambda = -(n-1)\tan\phi/\lambda$ or to the diffraction angle $\theta = \sin^{-1}[(n-1)\tan\phi]$. This is what would be expected from elementary geometrical optics when θ and ϕ are small.

3.2.8 The blazed diffraction grating

It is only a small step to the description of the diffraction produced by a grating which comprises, instead of alternating opaque and transparent strips, a grid of parallel prisms. There are two advantages in such a construction. Firstly the aperture is completely transparent and no light is lost, and secondly the prism arrangement means that, for one wavelength at least, all the incident light is diffracted into one order of the spectrum.

The aperture function is, as before, the convolution of the function for a single slit with a Dirac comb, the whole being multiplied by a broad $\Pi_{Na}(x)$ representing the whole width of the grating.

The diffracted intensity is then the same shifted sinc^2 function as above, but multiplied by the convolution of a Dirac comb with a narrow sinc-function, the Fourier pair of $\Pi_{Na}(x)$, which represents the shape of a single spectrum line. Now, there is a difference, because the broad sinc-function produced by a single slit has the same width as the spacing of the teeth in the Dirac comb. The zeros of this broad sinc-function are adjusted accordingly, and for one wavelength, the first order of diffraction falls on its maximum, while all the other orders fall on its zeros. For this wavelength, *all* the transmitted light is diffracted into first order. For adjacent wavelengths the efficiency is similarly high, and in general the efficiency remains usefully high for wavelengths between 2/3 and 3/2 of this wavelength.

This is the 'blaze wavelength' of the grating and the corresponding angle θ is the 'blaze-angle'.

Reflection gratings are made by ruling lines on an aluminium surface with a diamond scribing tip, held at an angle to the surface so as to produce a series

of long thin mirrors, one for each ruling. The angle is the 'blaze-angle' that the grating will have, and a similar analysis will show easily that the phase change across one slit is $(2\pi/\lambda)2a\tan\beta$ where β is the 'blaze-angle' and a the width of one ruling (and the separation of adjacent rulings). In practice, gratings are usually used with light incident normally or near-normally on the ruling facets, that is at an incidence angle β to the surface of the grating. There is then a phase change zero across one ruling, but a delay $(2\pi/\lambda)2a\sin\theta$ between reflections from adjacent rulings. If this phase change $= 2\pi$ then there is a principal maximum in the diffraction pattern.

Transmission gratings, generally found in undergraduate teaching laboratories, are usually blazed, and the effect can be easily be seen by holding one up to the eye and looking at a fluorescent lamp through it. The diffracted images in various colours are much brighter on one side than on the other.

3.3 Polar diagrams

Since the important feature of Fraunhofer theory is the angle of diffraction, it is sometimes more useful, especially in antenna theory, to draw the intensity pattern with a polar diagram, with intensity as r the length of the radius vector and θ as the azimuth angle. The sinc^2-function then appears as in Figure 3.14. Sometimes the logarithm of the intensity is plotted instead, to give the *gain* of the antenna as a function of angle.

A word of caution is appropriate here: although the basic idea of Fraunhofer diffraction may guide antenna design, and indeed allows proper calculation for so-called 'broadside arrays', there are considerable complications when

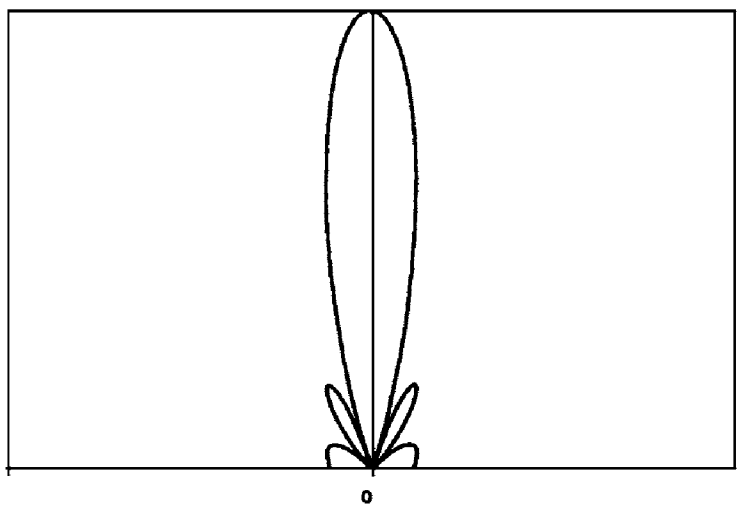


Fig. 3.14. The polar diagram of a sinc-function.

describing 'end-fire' arrays, or 'Yagi' aerials (the sort used for television reception). The broadside array, which comprises a number of dipoles (each dipole consisting of two rods, lying along the same line, each $\lambda/4$ long and with an alternating voltage applied in the middle) behaves like a row of point sources of radiation, and the amplitude at distances large compared with a wavelength can be calculated. Both the amplitude and the relative phase radiated by each dipole can be controlled[7] so that the shape of the radiation pattern and the strengths of the side-lobes are under control. End-fire antennae, on the other hand, have one dipole driven by an oscillator and rely on resonant oscillation of the other 'passive' dipoles to interfere with the radiation pattern and direct the output power in one direction. The phase re-radiated by a passive dipole depends on whether it is really half a wavelength long, on its conductivity, which is not perfect and on the dielectric constant of any sheath which may surround it. Consequently, aerial design tends to be based on experience, experiment and computation, rather than on strict Fraunhofer theory. The passive elements may be $\lambda/3$ apart, for example and their lengths will taper along the direction of the aerial, being slightly shorter on the transmission side and longer on the opposite side to the excited dipole. Such modifications allow a broader band of radiation to be transmitted or received along a narrow cone possibly only a few degrees wide. The nearest optical analogue is probably the Fabry–Perot étalon or, practically the same thing, the interference filter.

3.4 Phase and coherence

Coherence is an important concept, not only in optics, but whenever oscillators are compared.

No natural light source is exactly monochromatic, and there are small variations in period and hence wavelength from time to time. Two sources are said to be coherent when any small variation in one is matched by a similar variation in the other, so that, for example, if a crest of a wave from one arrives at a given point at the same instant as the trough of a wave from the other, then at all subsequent times troughs and crests will arrive together and there is always destructive interference between the two.

In general two separate sources, two laser beams for example, although nominally of the same wavelength, will not be coherent and no interference pattern will be seen when they both shine on to a screen[8]. This is why, to generate

[7] Equivalent to apodising in optics, but with more flexibility.

[8] This is not strictly true: a very fast detector can 'see' the fringes, which are shifting very rapidly on the surface where they are formed. Exposures in nanoseconds or less are required, and the technology involved is fairly expensive.

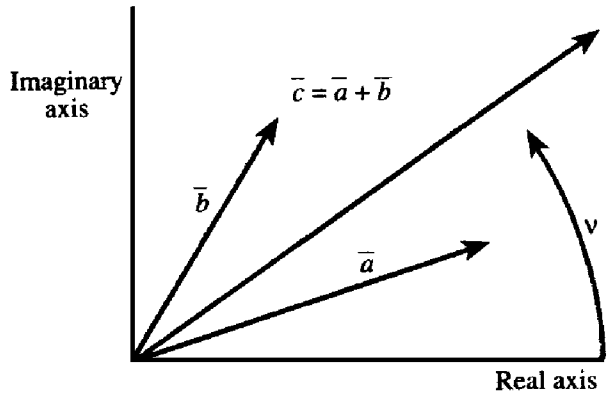


Fig. 3.15. The vector addition of two wave-vectors representing two coherent sources. All three vectors are rotating at the same frequency, ν. The vectors are described by the complex numbers $Ae^{2\pi i\nu t}$, the 'analytic signal', but it is the real part of each – the horizontal component in the graph – which represents the instantaneous value of the electric field of the light wave.

an interference pattern it is necesary to use two images of the same source as with a Fresnel biprism for example, or two sources fed from the same primary source, as in Young's slits.

The idea can be visualized by thinking of the *analytic* wave-vector – the vector in the complex plane whose real component represents the electric field – rotating at frequency ν. If the source is monochomatic the rotation is exactly at this frequency. Now imagine a rotating coordinate system, rotating at frequency ν. The wave-vector will be stationary. In practice the wave-vector will wander about an average direction, and if there are two sources, both vectors will wander independently. If they wander by angles greater than 2π, the vector sum of the two, which represents the resultant amplitude, will vary randomly between $\overline{E_1} + \overline{E_2}$ and $\overline{E_1} - \overline{E_2}$ and the intensity will take an average value $I = I_1 + I_2$, where $I_1 = < E_1 E_1^* >$ and the diagonal brackets denote time-averages.

On the other hand, if the two sources are coherent the phase angle ϕ between the two vectors will stay constant and the resultant amplitude will be $\overline{E_1} + \overline{E_2}e^{i\phi}$ The intensity of the combined sources will then be

$$I = I_1 + I_2 + 2\sqrt{I_1 I_2}\cos\phi$$

Now a new and useful concept can be introduced: suppose that the two vectors are not completely independent but that they are loosely coupled together so that the phase-difference ϕ varies about a mean value, but this variation, although random, is less than 2π. The time average of the vector sum is then not simply $I_1 + I_2$, but is less than the vector sum above. Then there will be an interference pattern, but the minima will not be so deep nor the maxima so high as in the

fully coherent case. We write:

$$I = I_1 + I_2 + 2\Gamma_{12}\sqrt{I_1 I_2}\cos\phi$$

where ϕ is the average phase-difference. The factor Γ_{12} is always less than unity and is called the *degree of coherence* or the *coherence factor*. The condition is called 'partial coherence'.

It can be measured in the laboratory by measuring the maximum and minimum intensities in an interference fringe system. In one case $\phi = 0$ and in the other $\phi = \pi/2$ so that $I_{max} = I_1 + I_2 + 2\Gamma_{12}\sqrt{I_1 I_2}$ and $I_{min} = I_1 + I_2 - 2\Gamma_{12}\sqrt{I_1 I_2}$ The *visibility* of the fringes, which is defined by

$$V = (I_{max} - I_{min})/(I_{max} + I_{min}) = 2\Gamma\sqrt{I_1 I_2}/(I_1 + I_2)$$

is closely related to the spatial degree of coherence. In particular, if the two wave-trains emerging from two slits in a plane are of equal intensity (and they should be in a well-conducted experiment), then:

$$V = \Gamma_{12}$$

This measures in fact the degree of spatial coherence in the two parts of the wave-front arriving at the two slits from the original monochromatic source, and this sort of coherence depends on the size of the source. The purpose of a *stellar* interferometer[9] is to measure the coherence size of the emitter, i.e. the angular diameter of the star being observed. The Van Cittert–Zernike (q.v.) theorem covers the question by showing that the fringe-visibility in a Young's slits-type interferometer, measured as a function of the separation of the two slits, is proportional to the Fourier transform of the angular intensity distribution across the source. The accuracy of measurement obtainable in principle is comparable with that of the telescope with total theoretical resolution $1.22f\lambda/d$.

Radio astronomers, being able to measure phase and amplitude directly, can detect and measure partial coherence with separations d, ('base-lines') comparable with the diameter of the earth and, since the angular resolution obtainable is about the ratio of the wavelength divided by the baseline, can consequently measure angular diameters of radio sources with a resolution of about 2×10^{-8} radians. (An optical telescope would need an aperture of 50 m to compete with this.)

We can also conceive of *temporal* coherence, which is the coherence between one part of a wave-train and a later part. This is seen for example in a two-beam interferometer where the wavefronts are divided and one part travels along a

[9] Such as the Michelson or the Hanbury–Brown stellar interferometers.

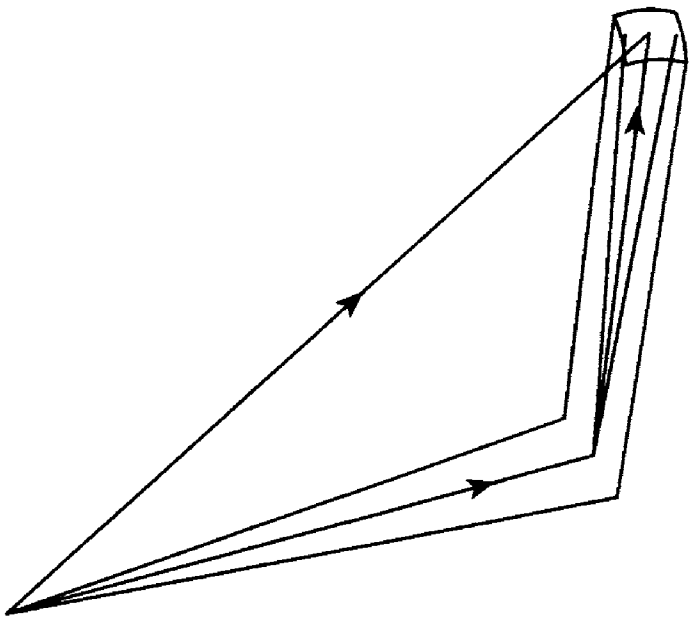


Fig. 3.16. The vector addition of two wave-vectors representing two partially coherent sources. All the vectors are rotating at the same average frequency, but the phase difference ϕ varies randomly over a small range of angles.

different and longer path than the other, so that one part is delayed when they recombine at the beam-splitter. The interference fringes are then not as sharp as at zero path-difference becuse of the reduced coherence between one section of the wave train and another. This gives us the idea of *coherence length* and answers the question often asked of students: 'how long is a photon'. The answer is for 'allowed' atomic transitions, about 1 m. This corresponds to the time taken for the transition – about 10^{-8} s and the distance travelled by light in that time. Light from an atomic beam, unaffected by random motions of the emitting atoms, will show fringes gradually losing coherence over path-differences of up to 0.5 m or so[10].

Since all electromagnetic radiation is quantized (if quantum theorists are to be believed), there may be a conceptual difficulty in reconciling the energy of a photon, $h\nu$, with the very long coherence lengths (light-years in the case of atomic clocks) of radio waves. This can be circumvented by considering that photons are bosons and that, unlike fermions, it consequently is possible to superimpose them, namely to put two or more of them in the same place at the same time and to allow them all to be coherent with each other. The philosophical aspects of this should be left to the disciples of the uncertainty principle.

[10] Keeping an interferometer aligned with ths sort of path-difference is one of the more heroic aspects of optical technique.

3.5 Exercises

(Note: These examples are not just dry academic solutions of artificial apertures: the results which they provide may well form the physical bases for new types of measuring instrument and servo-control devices.)

Find the angular intensity distribution of the diffracted radiation in the following examples:

1. An aperture of width A, of which one half has been covered with a transparent strip which delays the wavefront by $\lambda/2$. What happens if the transparent strip slips so as to cover more or less than one half of the aperture? (Method of monitoring and hence controlling the position of the transparent strip.)
2. Two apertures, each of width a, with their centres separated by b. One of them is covered by a moving transparent ribbon of varying thickness, causing a varying delay in the wavefront of the order of a few wavelengths on average, with a few tenths of a wavelength variation. (Method of monitoring and hence controlling the thickness during manufacture.)
3. Four equi-spaced apertures, the end one and its neighbour covered with a transparent strip of varying thickness. Is there any advantage in using four instead of two? Are there optimum values for a and b?
4. Two identical half-wave dipole antennae are fed from the same transmitter and one feed incorporates a lossless phase-shifting network. How will the polar diagram of the radiation pattern change as the relative phases of the antennae are changed?
5. Work out from first principles the theory of the blazed reflection grating. Find the blaze-angle necessary in a reflection grating with 6000 rulings/cm if it is to be perfectly efficient in first order for light of wavelength 500 nm (or 5000 Å or 0.5 μm).

Chapter 4

Applications 2: signal analysis and communication theory

4.1 Communication channels

Although the concepts involved in communication theory are general enough to include bush-telegraph drums, alpine yodelling or a ship's semaphore flags, by 'communication channel' is usually meant a single electrical conductor, a waveguide, a fibre-optic cable or a radio-frequency carrier wave. Communication theory covers the same general ground as information theory, which discusses the 'coding' of messages (such as Morse code, not to be confused with encryption, which is what spies do) so that they can be transmitted efficiently. Here we are concerned with the physical transmission by electric currents or radio waves, of the signal or message that has already been encoded. The distinction is that communication is essentially an analogue process, whereas information coding is essentially digital.

For the sake of argument, consider an electrical conductor along which is sent a varying current, sufficient to produce a potential difference $V(t)$ across a terminating impedance of one ohm.

The mean-level or time-average of this potential is denoted by the symbol $\langle V(t)\rangle$ defined by the equation:

$$\langle V(t)\rangle = \frac{1}{2T}\int_{-T}^{T} V(t)\,dt$$

The power delivered by the signal varies from moment to moment, and it too has a mean value:

$$\langle V^2(t)\rangle = \frac{1}{2T}\int_{-T}^{T} V^2(t)\,dt$$

For convenience, signals are represented by functions like sinusoids which, in general, disobey one of the Dirichlet conditions described at the beginning of

Chapter 2: they are not square-integrable:

$$\lim_{t\to\infty}\int_{-T}^{T} V^2(t)\,dt \to \infty$$

but in practice, the signal begins and ends at finite times and we regard the signal as the product of $V(t)$ with a very broad top-hat function. Its Fourier transform – which tells us about its frequency content – is then the convolution of the true frequency content with a sinc-function so narrow that it can for most purposes be ignored. We thus assume that $V(t) \to 0$ at $\mid t \mid > T$ and that

$$\int_{-\infty}^{\infty} V^2(t)\,dt = \int_{-T}^{T} V^2(t)\,dt$$

We now define a function $C(\nu)$ such that $C(\nu) \rightleftharpoons V(t)$, and Rayleigh's theorem gives:

$$\int_{-\infty}^{\infty} \mid C(\nu) \mid^2 d\nu = \int_{-\infty}^{\infty} V^2(t)\,dt = \int_{-T}^{T} V^2(t)\,dt$$

The mean power level in the signal is then:

$$(1/2T)\int_{-T}^{T} \mid V \mid^2 (t)\,dt$$

since $V^2(t)$ is the power delivered into unit impedance; and then:

$$(1/2T)\int_{-T}^{T} \mid V \mid^2 (t)\,dt = \int_{-\infty}^{\infty} \frac{\mid C(\nu) \mid^2}{2T}\,d\nu$$

and we *define* $\mid C(\nu) \mid^2 /2T = G(\nu)$ to be the spectral power density (SPD) of the signal.

4.1.1 The Wiener–Khinchine theorem

The autocorrelation function of $V(t)$ is defined to be:

$$\lim_{T\to\infty}(1/2T)\int_{-T}^{T} V(t)V(t+\tau)\,dt = \langle V(t)V(t+\tau)\rangle$$

again the integral on the left-hand side diverges and we use the shift theorem and Parseval's theorem to give:

$$\int_{-T}^{T} V(t)V(t+\tau)\,dt = \int_{-\infty}^{\infty} C^*(\nu)C(\nu)e^{2\pi i\nu\tau}d\nu$$

Then:

$$(1/2T)\int_{-T}^{T} V(t)V(t+\tau)\,dt = \int_{-\infty}^{\infty} \frac{|\,C(\nu)\,|^2}{2T} e^{2\pi i\nu\tau}\,d\nu = R(\tau)$$

so that with the definition of $G(\nu)$ above:

$$R(\tau) = \int_{-\infty}^{\infty} G(\nu)e^{2\pi i\nu\tau}\,d\nu$$

and finally:

$$R(\tau) \rightleftharpoons G(\nu)$$

In other words,

the spectral power density is the Fourier transform of the autocorrelation function of the signal.

This is the Wiener–Khinchine theorem.

4.2 Noise

The term originally meant the random fluctuation of signal voltage which was heard as a hissing sound in early telephone receivers, and which is still heard in radio receivers that are not tuned to a transmitting frequency. Now it is taken to mean any randomly fluctuating signal which carries no message or 'information'. If it has equal power density at all frequencies it is called 'white' noise[1]. Its autocorrelation function is always zero since at any time the signal $n(t)$, being random, is as likely to be negative as positive. The only exception is at zero delay, $\tau = 0$ where the integral diverges. The autocorrelation function is therefore a δ-function and its Fourier transform is unity, in accordance with the Wiener–Khinchine theorem and with this definition of 'white'.

In practice the band of frequencies which is received is always finite, so that the noise power is always finite. There are other types of noise. For example:

- Electron shot noise, or 'Johnson noise', in a resistor, giving a random fluctuation of voltage across it: $\langle V^2(t)\rangle = 4\pi RkT\Delta\nu$, where $\Delta\nu$ is the bandwidth, R the resistance, k Boltzmann's constant and T the absolute temperature[2].

[1] This is a rebarbative use of 'white', which really defines a rough surface which reflects all the radiation incident upon it. It is used, less compellingly, to describe the colour of the light emitted by the Sun or even less compellingly, to describe light of constant spectral power density in which all wavelengths (or frequencies: take your choice) contribute equal power.

[2] $\langle V^2\rangle = 1.3 \times 10^{-10}(R\Delta\nu)^{1/2}$ volts in practice.

- Photon shot noise, which has a normal (Gaussian) distribution of count-rate[3] at frequencies low compared with the average photon arrival rate and, more accurately, a Poisson distribution when equal time samples are taken. This is met chiefly in optical beams used for communication, and only then when they are weak. Typically, a laser beam delivers 10^{18} photons/s, so that even at 100 MHz there are 10^{10} photons/sample, or an S/N ratio of $10^5:1$.
- Semi-conductor noise, which gives a time-varying voltage with a SPD which varies as $1/\nu$ – which is why many semiconductor detectors of radiation are best operated at high frequency with a 'chopper' to switch the radiation on and off. There is usually an optimum frequency, since the number of photons in a short sample may be small enough to increase photon shot-noise to the level of the semi-conductor noise.

4.3 Filters

By 'filter' we mean an electrical impedance which depends on the frequency of the signal current trying to pass. The exact structure of the filter, the arrangement of resistors, capacitors and inductances, is immaterial. What matters is the effect that the filter has on a signal of fixed frequency and unit amplitude. The filter does two things: it attenuates the amplitude and it shifts the phase. This is all that it does[4]. The frequency-dependence of its impedance is described by its filter function $Z(\nu)$. This is defined to be the ratio of the output voltage divided by the input voltage, as a function of frequency:

$$Z(\nu) = V_o/V_i = A(\nu)e^{i\phi(\nu)}$$

where V_i and V_o are 'analytic' representations of the input and output voltages; i.e. they include the phase as well as the amplitude. The impedance is complex since both the amplitude and the phase of V_o may be different from V_i. The filter impedance, Z is usually shown graphically by plotting a polar diagram of the attenuation, A, radially against the angle of phase-shift, eliminating ν as a variable. The result is called a *Nyquist diagram* (Fig 4.1). This is the same figure that is used to describe a feedback loop in servo-mechanism theory, with the difference that the amplitude A is always less than unity in a passive filter, so that there is no fear of the curve encompassing the point $(-1, 0)$, the criterion for oscillation in a servomechanism.

[3] Which may be converted into a time-varying voltage by a rate-meter.

[4] Unless it is 'active.' Active filters can do other things such as doubling the frequency of the input signal.

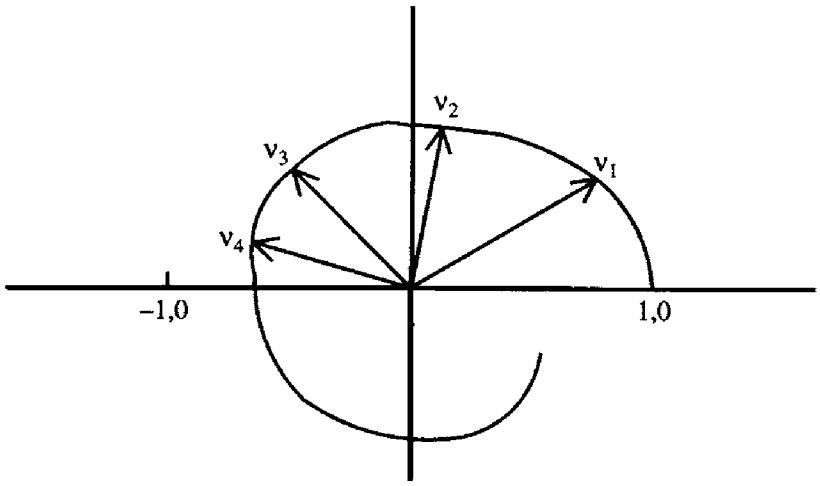


Fig. 4.1. The Nyquist diagram of a typical filter.

4.4 The matched filter theorem

Suppose that a signal $V(t)$ has a frequency spectrum $C(\nu)$ and spectral power density $S(\nu) = | C(\nu) |^2/2T$. The signal emerging from the filter then has a frequency spectrum $C(\nu)Z(\nu)$ and the SPD is $G(\nu)$, given by:

$$G(\nu) = \frac{| C(\nu)Z(\nu) |^2}{2T}$$

If there is white noise passing through the system, with spectral power density $| N(\nu) |^2/2T$ the total signal power and noise power are:

$$\frac{1}{2T}\int_{-\infty}^{\infty} | C(\nu)Z(\nu) |^2 \, d\nu$$

and

$$\frac{1}{2T}\int_{-\infty}^{\infty} | N(\nu)Z(\nu) |^2 \, d\nu$$

For white noise $| N(\nu) |^2$ is a constant, $= A$, say, so that the transmitted noise power is:

$$\frac{A}{2T}\int_{-\infty}^{\infty} | Z(\nu) |^2 \, d\nu$$

and the ratio of signal power to noise power (S/N) is the ratio:

$$(S/N)_{power} = \int_{-\infty}^{\infty} | C(\nu)Z(\nu) |^2 \, d\nu \Big/ A\int_{-\infty}^{\infty} | Z(\nu) |^2 \, d\nu$$

Here we use Schwartz's inequality[5]

$$\left[\int_{-\infty}^{\infty} | C(\nu)Z(\nu) |^2 \, d\nu\right]^2 \leq \int_{-\infty}^{\infty} | C(\nu) |^2 \, d\nu \int_{-\infty}^{\infty} | Z(\nu) |^2 \, d\nu$$

so that the S/N power ratio is always $\leq A \int_{-\infty}^{\infty} | C(\nu) |^2 \, d\nu$ and the equality sign holds if and only if $C(\nu)$ is a multiple of $Z(\nu)$. Hence:

The S/N power ratio will always be greatest if the filter characteristic function $Z(\nu)$ has the same shape as the frequency content of the signal to be received.

This is the matched filter theorem. In words, it means that the best signal/noise ratio is obtained if the filter transmission function has the same shape as the signal power spectrum.

It has a surprisingly wide application, in spatial as well as temporal data transmission. The tuned circuit of a radio receiver is an obvious example of a matched filter: it passes only those frequencies containing the information in the programme, and rejects the rest of electromagnetic spectrum. The tone-control knob does the same for the accoustic output. A monochromator does the same thing with light. The 'radial velocity spectrometer' used by astronomers[6] is an example of a spatial matched filter. The negative of a stellar spectrum is placed in the focal plane of a spectrograph, and its position is adjusted sideways – perpendicular to the slit-images – until there is a minimum of total transmitted light. The movement of the mask necessary for this measures the Doppler-effect produced by the line-of-sight velocity on the spectrum of a star.

4.5 Modulations

When a communication channel is a wireless telegraphy channel (a term which comprises everything from a modulated laser beam to an extremely low frequency (ELF) transmitter used to communicate with submerged submarines) it is usual for it to consist of a 'carrier' frequency on which is superimposed a 'modulation'. If there is no modulating signal, the voltage at the receiver varies with time according to:

$$V(t) = V_c e^{2\pi i(\nu_c t + \phi)}$$

[5] See, for example D. C. Champeney: *Fourier Transforms and their Physical Applications* Appendix F, Academic Press, 1973.

[6] Particularly by R. F. Griffin. See *Astrophys. J.* **148**, 465 (1967).

where ν_c is the carrier frequency; and the modulation may be carried out by making V_c, ν_c or ϕ a function of time.

4.5.1 Amplitude modulation

If V varies with a modulating frequency ν_{mod}, then $V = A + B\cos 2\pi\nu_{mod}t$ and the resulting frequency distribution will be as in Fig. 4.2 and as various modulating fequencies from $0 \rightarrow \nu_{max}$ are transmitted, the frequency spectrum will occupy a band of the spectrum from $\nu_c - \nu_{max}$ to $\nu_c + \nu_{max}$. If low modulating frequencies predominate in the signal, the band of frequencies occupied by the channel will have appearance of Fig. 4.3 and the filter in the receiver should have this profile too.

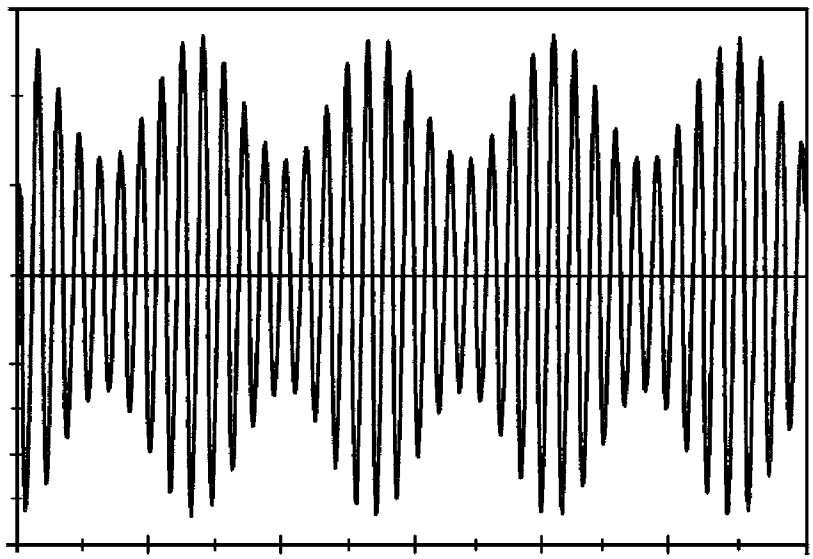


Fig. 4.2. A carrier wave with amplitude modulation.

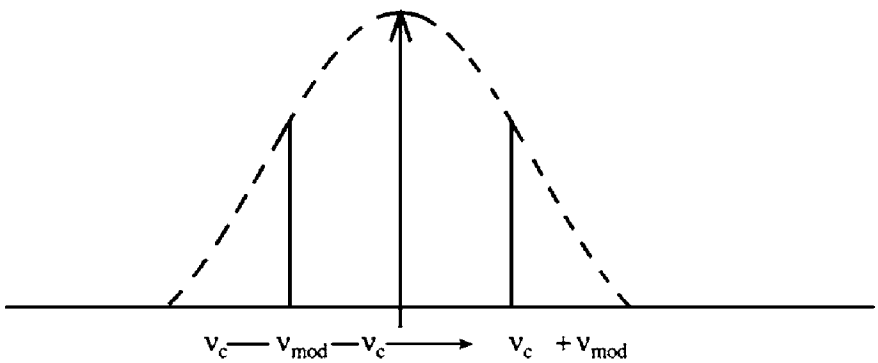


Fig. 4.3. Various modulating frequencies occupy a band of the spectrum. The time function is $A + B\cos(2\pi\nu_{mod}t)$ and in frequency space the spectrum becomes the convolution of $\delta(\nu - \nu_c)$ with $A\delta(\nu) + B(\delta(\nu - \nu_0) + \delta(\nu + \nu_0)/2$.

The power transmitted by the carrier is wasted unless very low frequencies are present in the signal. The power required from the transmitter can be reduced by filtering its output so that only the range from ν_c to ν_{max} is transmitted. The receiver is doctored in like fashion. The result is single sideband transmission.

4.5.2 Frequency modulation

This is important because it is possible to increase the bandwidth used by the channel. (By 'channel' is meant here perhaps the radio frequency link used by a spacecraft approaching Neptune and its receiver on Earth, some 4×10^9 km away.) The signal now is

$$V(t) = A \cos 2\pi \nu(t) t$$

and $\nu(t)$ itself is varying according to $\nu(t) = \nu_{carrier} + \mu \cos 2\pi \nu_{mod}(t) t$. μ can be made very large so that for example a voice telephone signal, normally requiring about 3×10^3 Hz bandwidth can be made to occupy several MHz if necessary. The advantage in doing this is found in the Hartley–Shannon theorem of information theory, which states that the 'channel capacity', the rate at which a noisy channel can transmit information in bits s^{-1} ('bauds') is given by:

$$dB/dt \leq 2\Omega \log_e(1 + S/N)$$

Where Ω is the channel bandwidth, S/N is the *power* signal/noise ratio and dB/dt is the 'baud-rate' or bit- transmission rate.

So, to get a high data transmission rate, you need not slave to improve the S/N ratio because only the logarithm of that is involved: instead you increase the bandwidth of the transmission. In this way the low power available to the spacecraft transmitter near Neptune is used more effectively than would be possible in an amplitude-modulated transmitter. Theorems in information theory, like those in thermodynamics, tend to tell you what is possible, without telling you how to do it.

To see how the power is distributed in a frequency-modulated carrier, the message-signal, $a(t)$, can be written in terms of the phase of the carrier signal, bearing in mind that frequency can be defined as rate of change of phase. If the phase is taken to be zero at time $t = 0$, then the phase at time t can be written as:

$$\phi = \int_0^t \frac{\partial \phi}{\partial t}\, dt$$

and $\partial\phi/\partial t = \nu_c + \int_0^t a(t)\, dt$ and the transmitted signal is:

$$V(t) = a e^{2\pi i [\nu_c + \int_0^t a(t)\, dt] t}$$

Consider a single modulating frequency ν_{mod}, such that $a(t) = k\cos(2\pi\nu_{mod}t)$. Then

$$2\pi i\int_0^t a(t)\,dt = \frac{2\pi ik}{2\pi\nu_{mod}}\sin 2\pi\nu_{mod}t$$

k is the depth of modulation, and k/ν_{mod} is called the *modulation index*, m. Then:

$$V(t) = Ae^{2\pi i\nu_c t}e^{im\sin(2\pi\nu_{mod}t)}$$

It is a cardinal rule in applied mathematics, that when you see an exponential function with a sine or cosine in the exponent, there is a Bessel function lurking somewhere. This is no exception. The second factor in the expression for $V(t)$ can be expanded in a series of Bessel functions by the Jacobi expansion[7]:

$$e^{im\sin(2\pi\nu_{mod}t)} = \sum_{n=-\infty}^{\infty} J_n(m)e^{2\pi in\nu_{mod}t}$$

and this is easily Fourier transformable to:

$$\chi(\nu) = \sum_{n=-\infty}^{\infty} J_n(m)\delta(\nu - n\nu_{mod})$$

The spectrum of the transmitted signal is the convolution of $\chi(\nu)$ with $\delta(\nu - \nu_c)$. In other words, $\chi(\nu)$ is shifted sideways so that the $n = 0$ tooth of the Dirac comb is at $\nu = \nu_c$.

The amplitudes of the Bessel functions must be computed or looked up in a table[8] and for small values of the argument m are: $J_0(m) = 1$; $J_1(m) = m/2$; $J_2(m) = m^2/4$ etc. Each of these Bessel functions multiplies a corresponding tooth in the Dirac comb of period ν_{mod} to give the spectrum of the modulated carrier. Bearing in mind that $m = k/\nu_{mod}$ we see that the channel is not uniformly filled and there is less power in higher frequencies.

As an example of the cross-fertilizing effect of Fourier transforms, the theory above can equally be applied to the diffraction produced by a grating in which there is a periodic error in the rulings. In Chapter 3 there was an expression for the 'aperture function' of a grating which was

$$A(x) = \Pi_{Na}(x)[\Pi_a(x) * \text{Ш}_a(x)]$$

and if there is a periodic error in the ruling, it is $\text{Ш}_a(x)$ that must be replaced. The rulings, which should have been at $x = 0, a, 2a, 3a, \ldots$ will be

[7] See, for example, Jeffreys & Jeffreys, *Mathematical Physics*, Cambridge University Press, p. 589.

[8] e.g. Jancke & Emde or Abramowitz & Stegun.

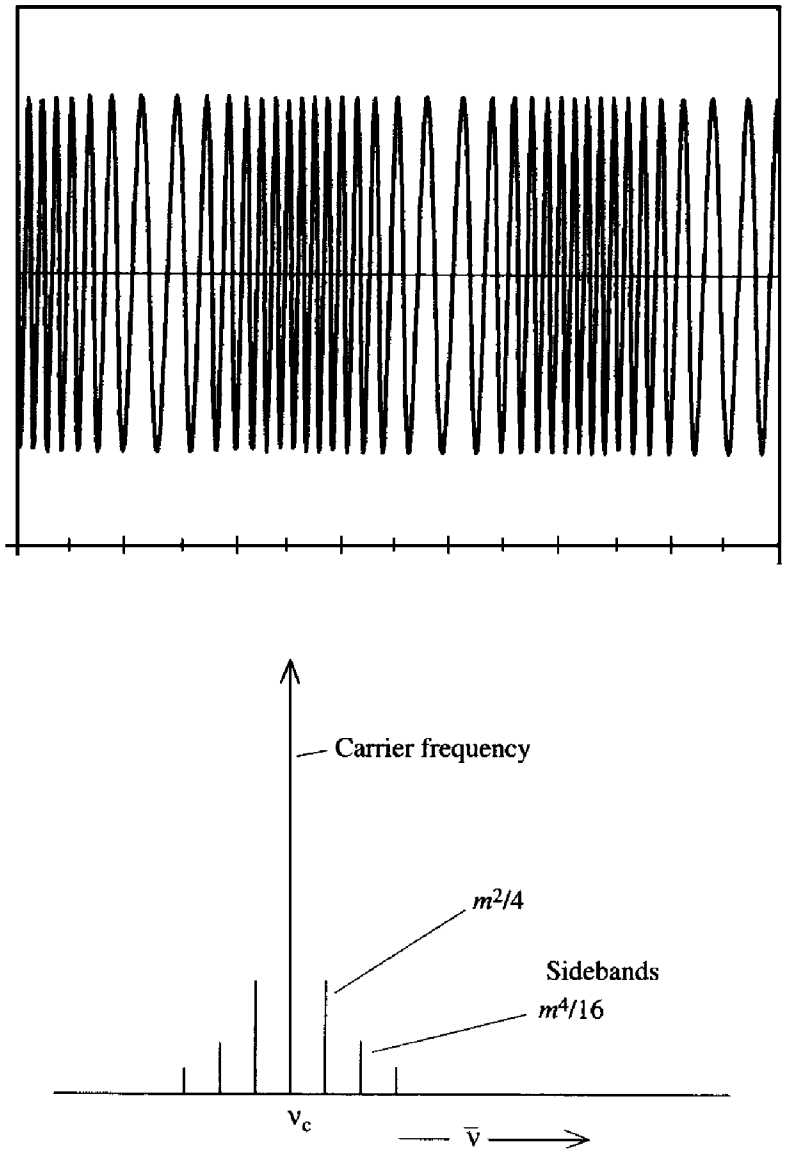


Fig. 4.4. Frequency modulation of the carrier. Many side-bands are present with amplitudes given by the Jacobi expansion.

at $0, a + \alpha \sin(2\pi\beta.a), 2a + \alpha \sin(2\pi\beta.2a), \ldots$ etc. and the *III*-function is replaced by

$$G(x) = \sum_{-\infty}^{\infty} \delta\left[x - na - \alpha \sin(2\pi\beta na)\right]$$

where α is the amplitude of the periodic error. and $1/\beta$ is its 'pitch'. This has a Fourier transform

$$\overline{G}(p) = \sum_{-\infty}^{\infty} e^{2\pi i[na + \alpha \sin(2\pi\beta na)]}$$

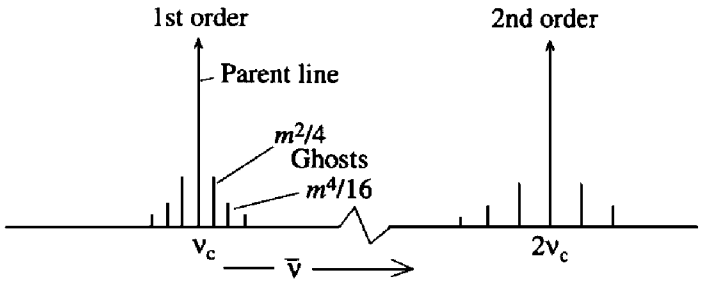


Fig. 4.5. Rowland ghosts in the spectrum produced by a diffraction grating with a period error in its rulings. The spacing of the ghost from its parent line depends on the period of the error, and the intensity on the square of the amplitude of the error.

with $p = \sin\theta/\lambda$ as in Chapter 3. There is a clear analogy with $V(t)$ above. The diffraction pattern then contains what are called 'ghost' lines[9] around each genuine spectrum line as in Fig. 4.5.

The analysis is not quite as simple as in the case of a frequency-modulated radio wave because the simple sinusoids are replaced by δ-functions. What happens is that the infinite sum $\overline{G}(p)$ can be analysed into a whole set of Dirac combs, of periods slightly above and below the true error-free period, and with amplitudes decreasing rapidly according to the amplitude of the Bessel function which multiplies them. The Rowland ghosts are then separated from the parent line by distances which depend on the pitch $1/\beta$ of the lead-screw of the grating ruling engine and have amplitudes which depend on the square[10] of the amplitude α of that periodic error.

These satellites lie on either side of a spectrum line with intensity $\pi^2 p^2 \alpha^2$ times the height of the parent and separated from it by $\Delta\lambda = \pm a\beta\lambda$ are the first-order Rowland ghosts. The next ones, of height $\pi^4 p^4 \alpha^4$ of the parent intensity are the second-order ghosts, and so on. The analogy with the channel occupation of a frequency-modulated carrier is exact.

There are of course many other ways of modulating a carrier, such as *phase* modulation, *pulse-width* modulation, *pulse-position* modulation, *pulse-height* modulation and so on, quite apart from digital encoding which is quite a separate way of conveying information. Several different kinds of modulation can be applied simultaneously to the same carrier, each requiring a different type of demodulating circuit at the receiver. The design of communications channels includes the art of combining and separating these modulators and ensuring that they do not influence each other with various kinds of 'cross-talk'.

[9] Rowland ghosts, after H. A. Rowland, the inventor of the first effective grating-ruling engine.
[10] Because $\overline{G}(p)$ gives the diffraction *amplitude*.

4.6 Multiplex transmission along a channel

There are two ways of sending a number of independent signals along the same communication channel. They are known as *time multiplexing* and *frequency multiplexing*. Frequency multiplexing is the more commonly used. The signals to be sent are used to modulate[11] a *sub-carrier* which then modulates the main carrier. A filter at the receiving end demodulates the main carrier and transmits only the sub-carrier and its side-bands (which contain the message). Different sub-carriers require different filters and it is usual to leave a small gap in the frequency spectrum between each sub-carrier, to guard against 'cross-talk', that is one signal spreading into the pass-band of another signal.

Time-multiplexing involves the 'sampling' of the carrier at regular time intervals. If, for example, there are ten separate signals to be sent, the sampling rate must be twenty times the highest frequency present in each band. The samples are sent in sequence and switched to ten different channels for decoding, and there must be some way of collating each message channel at the transmitting end with its counterpart at the receiving end so that the right message goes to the right recipient. The 'serial link' between a computer and a peripheral, which uses only one wire, is an example of this, with about eight channels[12], one for each bit-position in each byte of data.

4.7 The passage of some signals through simple filters

This is not a comprehensive treatment of the subject, but illustrates the methods used to solve problems. Firstly we need to know about the Heaviside step function.

4.7.1 The Heaviside step function

When a switch is closed in an electric circuit there is a virtually instantaneous change of voltage on one side. This can be represented by a 'Heaviside step' function, $H(t)$. It has the property that $H(t) = 0$ for $t < 0$ and $H(t) = 1$ for $t > 0$[13]. If you differentiate it you get a delta-function $\delta(t)$ and this fact can be

[11] 'Modulate' here means that the main carrier signal is multiplied by the message-bearing sub-carrier. Demodulation is the reverse process, in which the sub-carrier and its message are extracted from the transmitted signal by one of various electronic tricks.

[12] Anywhere between 5 and 11 channels in practice, so long as the transmitter and receiver have agreed beforehand about the number.

[13] Its value *at* $x = 0$ is the subject of debate, but usually taken as $H(0) = 1/2$.

used to find its Fourier transform. We use the differential theorem:

$$H(t) = \int_{-\infty}^{\infty} \phi(\nu) e^{2\pi i \nu t} d\nu$$

$$\delta(\nu) = dH(t)/dt \rightleftharpoons 2\pi i \phi(\nu)$$

so that:

$$\phi(\nu) = 1/2\pi i \nu$$

4.7.2 The passage of a voltage step through a 'perfect' low-pass filter

Suppose that the filter is a 'low-pass' filter with no attenuation or phase-shift up to a critical frequency ν_c and zero transmission thereafter. If the height of the step is V volts, the voltage as a function of time is a Heaviside step-function, $VH(t)$. Its frequency content is then $V/2\pi i \nu$ and the output frequency spectrum is the product of this with the filter profile: that is, $\overline{V}(\nu) = V/(2\pi i \nu).\Pi_{2\nu_c}(\nu)$. The output signal, as a function of time, is the Fourier transform of this, which is

$$f_0(t) = V \int_{-\nu_c}^{\nu_c} \frac{e^{2\pi i \nu t}}{2\pi i \nu} d\nu$$

where the top-hat has been replaced by finite limits on the integral.

The function to be transformed is antisymmetric and so there is only a sine transform:

$$f_o(t) = iV \int_{-\nu_c}^{\nu_c} \frac{\sin 2\pi \nu t}{2\pi i \nu} d\nu = Vt \int_{-\nu_c}^{\nu_c} \operatorname{sinc}(2\pi \nu t)\, d\nu$$

$$= 2Vt \int_{0}^{\nu_c} \operatorname{sinc}(2\pi \nu t)\, d\nu = \frac{1}{\pi} \int_{0}^{2\pi \nu_c t} \operatorname{sinc}(x)\, dx$$

with the obvious substitution $x = 2\pi \nu t$.

The integral is a function of t obviously, and must be computed since sinc-functions are not directly integrable. The result is shown graphically in Fig. 4.6.

The rise-time depends on the filter bandwidth. People who use oscilloscopes on the fastest time-base settings to look at edges will recognize this curve.

4.8 The Gibbs phenomenon

When you display a square-wave on an oscilloscope, the edges are never quite sharp (unless they are made so by some subtle and deliberate electronic trick) but

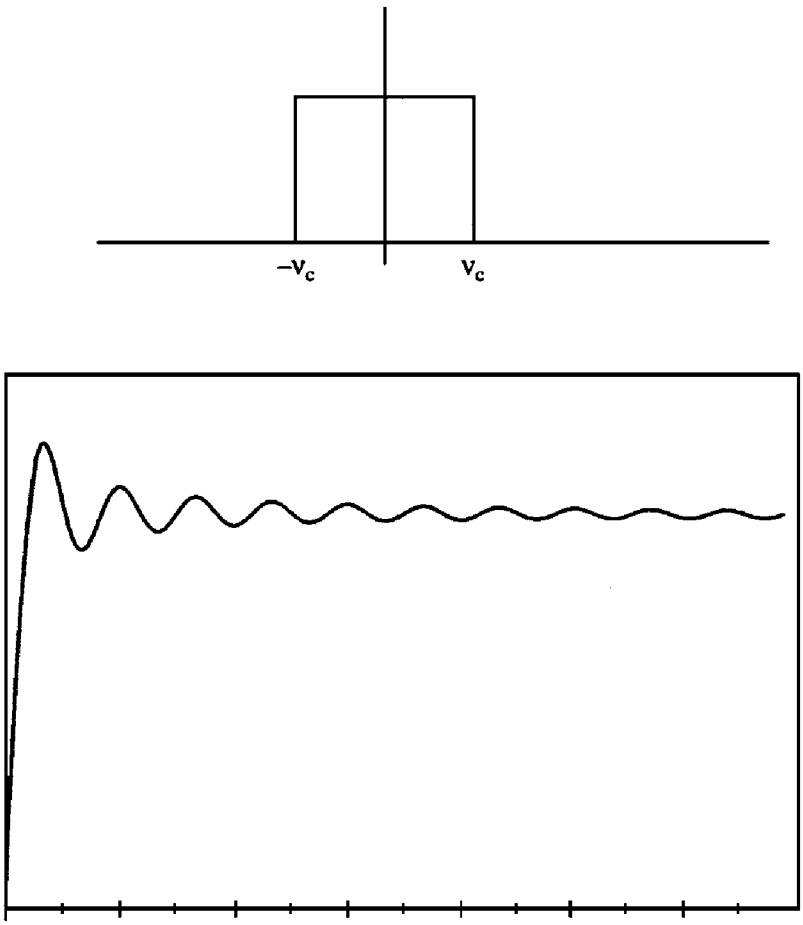


Fig. 4.6. Passage of a Heaviside step-function through a perfect low-pass filter. The pass band is a top-hat function in frequency space, and this sets the limits on the integral of the Heaviside step's transform.

show small oscillations which increase in amplitude as the corner is approached. They may be quite small in a high-bandwidth oscilloscope.

The reason is found in the finite bandwidth of the oscilloscope. The square-wave is synthesized from an infinite Dirac comb of frequencies, with teeth of heights which depend on the mark–space ratio of the square-wave. To give a *perfect* square-wave, an infinite number of teeth are required, that is to say, the series expansion for $F(t)$ must have an infinite number of terms: sharp corners need high frequencies. Since there is an upper limit to the available frequencies, only a finite number of terms are, in practice, included. This is equivalent to multiplying the Dirac comb in frequency-space by a top-hat function of width $2\nu_{max}$, and in t-space, which is what the oscilloscope diplays, you see the convolution of the square-wave with a sinc-function $\mathrm{sinc}(2\pi\,\nu_{max})$. Convolution with an edge (effectively with a Heaviside step-function) replaces the edge with the integral of the sinc-function between $-\infty$ and t, and the result is shown in Fig. 4.6.

The phenomenon was discovered experimentally by A. A. Michelson and Stratton. They designed a mechanical Fourier synthesizer, in which a pen position was controlled by 80 springs pulling together against a master-spring, each controlled by 80 gear-wheels which turned at relative rates of 1/80, 2/80, 3/80 . . . 79/80 and 80/80 turns per turn of a crank-handle. The synthesizer could have the spring tensions set to represent the 80 amplitudes of the Fourier coefficients and the pen position gave the the sum of the series. As the operator turned the crank-handle a strip of paper moved uniformly beneath the pen and the pen drew the graph on it, reproducing, to Michelson's mystification, a square-wave as planned, but showing the Gibbs phenomenon. Michelson assumed, wrongly, that mechanical shortcomings were the cause: Gibbs gave the true explanation in a letter to *Nature*[14].

The machine itself, a marvel of its period, was constructed by Gaertner & Co. of Chicago in 1898. It now languishes in the archives of the South Kensington Science Museum.

4.8.1 The passage of a train of pulses through a low-pass filter

Suppose that we represent the pulse train by a *Ш*-function. If the pulse repetition frequency is ν_0 the train is described by $Ш_a(t)$, where $a = 1/\nu_0$. Suppose that the filter as before, transmits perfectly all frequencies below a certain limit and nothing above that limit. In other words the filter frequency profile or 'filter function' is the same top-hat function Π_{ν_f}. The Fourier transforms of the signal and the filter function are $(1/a)Ш_{\nu_0}(\nu)$ and $\Pi_{\nu_f}(\nu)$ respectively. The frequency spectrum of the output signal is then the product of the input spectrum and the filter function, $(1/a)Ш_{\nu_0}(\nu).\Pi_{\nu_f}(\nu)$ and the output signal is the Fourier transform of this, namely the convolution of the original train of pulses with $\mathrm{sinc}2\pi\nu_f t$. If the filter bandwidth is wide compared with the pulse repetition frequency, $1/a$, the sinc-function is narrow compared with the separation of individual pulses, and each pulse is replaced, in effect, with this narrow sinc-function. On the other hand if the filter bandwidth is small and contains only a few harmonics of this fundamental frequency, the pulse-train will resemble a sinusoidal wave. An interesting sidelight is that if the transmission function of the filter is a decaying exponential[15], $Z(\nu) = e^{-k|\nu|}$, then the wavetrain is the convolution of $Ш_a(t)$ with $(k/2\pi^2)/[t^2 + (k/2\pi)^2]$. The square of the resulting function may be familiar to students of the Fabry–Perot étalon as the 'Airy' profile.

[14] *Nature* **59**, 606 (1899).

[15] Do the Fourier transform of this in two parts: $-\infty \rightarrow 0$ and $0 \rightarrow \infty$.

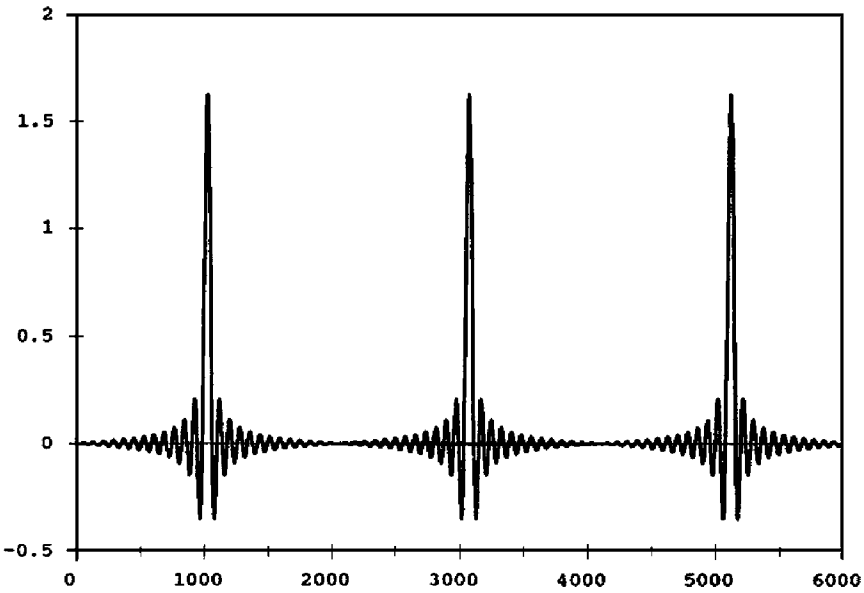


Fig. 4.7. Attenuation of a pulse train by a narrow band low-pass filter.

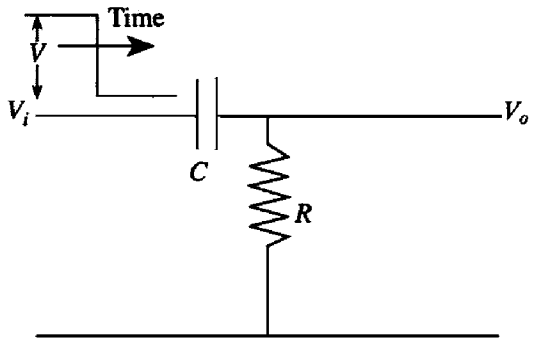


Fig. 4.8. A simple high-pass filter passing a voltage step.

4.8.2 Passage of a voltage step through a simple high-pass filter

This is an example which shows that contour integration has simple practical uses occasionally:

by Ohm's law (Fig. 4.8):

$$V_o = V_i \frac{R}{R + 1/2\pi i\nu C} = V_i \frac{2\pi i\nu RC}{2\pi i\nu RC + 1} = V_i \frac{2\pi i\nu}{2\pi i\nu + \alpha}$$

where R is the resistance, C the capacity in the circuit and $\alpha = 1/RC$

Let the input step have height V so that it is described by the Heaviside step function $\overline{V}_i(t) = VH(t)$. Its frequency content is then $V/2\pi i\nu = V_i(\nu)$ and

$$V_o(\nu) = \frac{V}{2\pi i\nu} \cdot \frac{2\pi i\nu}{2\pi i\nu + \alpha} = \frac{V}{2\pi i\nu + \alpha}$$

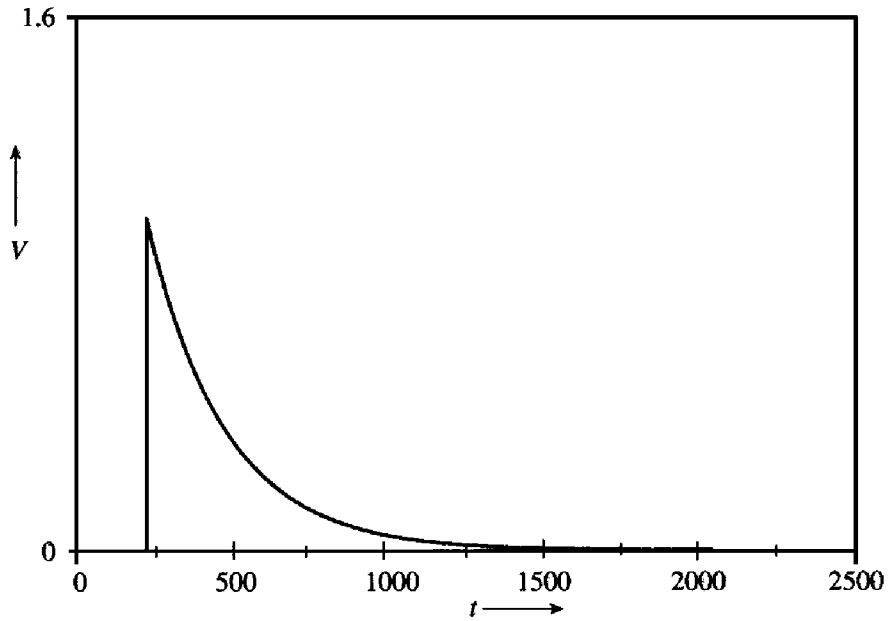


Fig. 4.9. V_0 as a function of time simple high-pass filter when the input is a Heaviside step function.

The time-variation of the output voltage is the Fourier transform of this:

$$\overline{V}_o(t) = V \int_{-\infty}^{\infty} \frac{e^{2\pi i \nu t}}{2\pi i \nu + \alpha} \, d\nu$$

replace $2\pi\nu$ by z:

$$\overline{V}_o(t) = \frac{V}{2\pi} \int_{-\infty}^{\infty} \frac{e^{izt}}{iz + \alpha} \, dz$$

and multiply top and bottom by $-i$ to clear z of any coefficient:

$$\overline{V}_o(t) = \frac{-iV}{2\pi} \int_{-\infty}^{\infty} \frac{e^{izt}}{z - i\alpha} \, dz$$

This integral will not yield to elementary methods ('quadrature'). So we use Cauchy's integral formula[16]: if z is complex, the integral of $f(z)/(z-a)$ anticlockwise round a closed loop in the Argand plane containing the point a is equal to $2\pi i f(a)$. The quantity $f(a)$ is the *residue* of $f(z)/(z-a)$ at the 'pole', a. Written formally it is:

$$\oint \frac{f(z)}{z-a} \, dx = 2\pi i f(a)$$

[16] Of fundamental importance and to be found in any book dealing with the functions of a complex variable.

Here the pole is at $z = i\alpha$, so $e^{izt} = e^{-\alpha t}$ and

$$\frac{-iV}{2\pi}\int_C \frac{e^{izt}}{z - i\alpha}\,dz = -2\pi i\frac{iV}{2\pi}e^{-\alpha t} = Ve^{-\alpha t}$$

and the loop ('contour') comprises (a) the real axis, to give the desired integral with $dz = dx$, and (b) the positive semi-circle at infinite radius where the integrand vanishes. Along the real axis the integral is:

$$\lim_{r\to\infty}\frac{-iV}{2\pi}\int_{-r}^{r}\frac{e^{ixt}}{x - i\alpha}dx$$

which is the integral we want. Along the semicircle at large r, z is complex and so can be written $z = e^{i\theta}$ or as $r(\cos\theta + i\sin\theta)$ so that e^{izt} becomes $e^{ir(\cos\theta + i\sin\theta)t}$. The real part of this is $e^{-rt\sin\theta}$ which, for positive values of t, vanishes as r tends to infinity (this is why we choose the positive semicircle – $\sin\theta$ is positive). The integral around the positive semicircle then contributes nothing to the total.

Thus, for $t > 0$, the time variation $V_o(t)$ of the voltage out, is:

$$V_o(t) = Ve^{-\alpha t}$$

For negative values of t, the negative semicircle must be used for integration in order to make the integral vanish. The negative semicircle contains no pole, so the real axis integral is also zero. So the complete picture of the response is shown in Fig. 4.9.

Chapter 5

Applications 3: spectroscopy and spectral line shapes

5.1 Interference spectrometry

One of the fundamental formulae of interferometry is the equation giving the condition for maxima and minima in an optical interference pattern:

$$2\mu d \cos\theta = m\lambda$$

and m must be integer for a maximum and half-integer for a minimum.

There are five possible variables in this equation, and by holding three constant, allowing one to be the independent variable and calculating the other, many different types of fringe can be described, sufficient for nearly all interferometers; and nearly all the types of interference fringe referred to in optics texbooks[1], such as 'localized' fringes, fringes of constant inclination, Tolansky fringes, Edser–Butler fringes etc. etc., are included.

5.1.1 The Michelson multiplex spectrometer

Consider the fringes produced by a Michelson interferometer.

If monochromatic light of wavenumber[2] $\overline{\nu}$ $(= 1/\lambda)$ and amplitude A is incident, the beam splitter, if perfect, will send light of amplitude $A/\sqrt{2}$ along each arm. It will be reflected at the two mirrors, and on return to the beam-splitter will recombine with different phases, the result of different path lengths travelled in the two arms. If, for convenience, we choose a moment when the phase is zero at the point of division, the two phases will be $2\pi\overline{\nu}2d_1$ and $2\pi\overline{\nu}2d_2$, where the two paths have lengths d_1 and d_2.

[1] e.g. M. Born and E. Wolf *Principles of Optics*, Cambridge University Press, Cambridge.

[2] $\overline{\nu}$ is used here to denote wavenumber rather than k, since k is sometimes used to mean $2\pi/\lambda$.

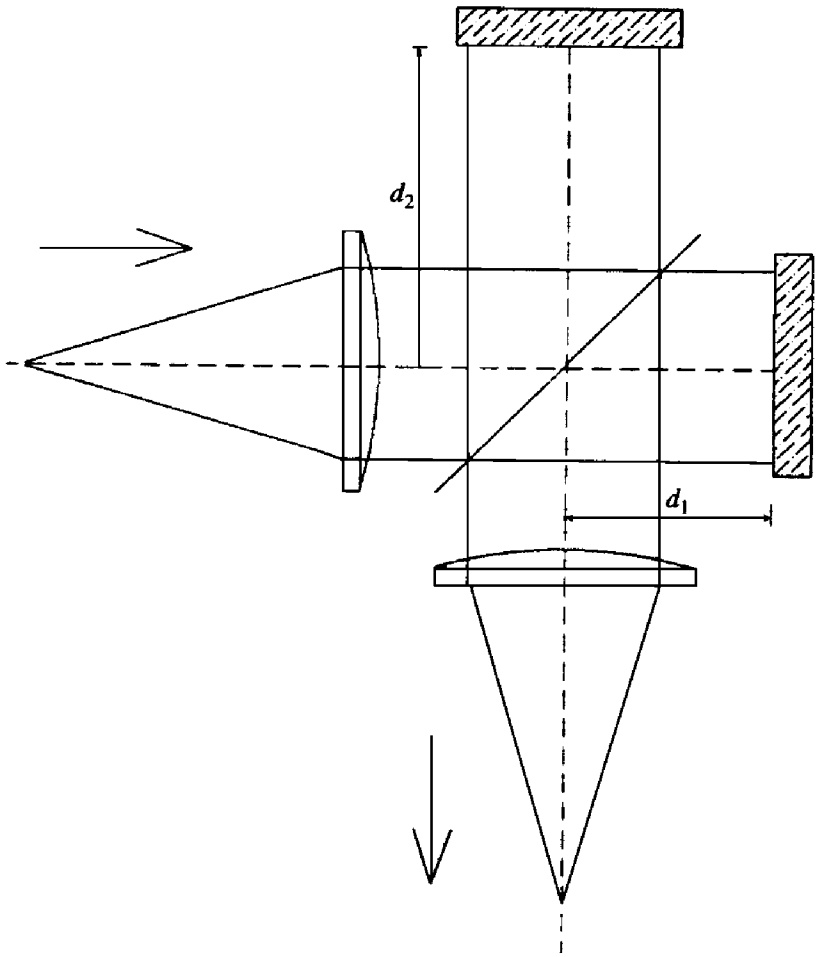


Fig. 5.1. The optical path in a Michelson interferometer.

The two amplitudes, both complex when the phases are included, are added and the transmitted amplitude is $(A/2)[e^{2\pi i\bar{\nu}d_1} + e^{2\pi i\bar{\nu}d_2}]$. The transmitted intensity is then:

$$I = A^2/4[e^{2\pi i\bar{\nu}d_1} + e^{2\pi i\bar{\nu}d_2}][e^{-2\pi i\bar{\nu}d_1} + e^{-2\pi i\bar{\nu}d_2}]$$

The path-difference $2(d_1 - d_2)$ is usually written as Δ so that, on completing the multiplication:

$$I = I_0/2[1 + \cos(2\pi\bar{\nu}\Delta)]$$

where $I_0 = AA^*$ is the input intensity at zero path-difference.

This describes the fringes which appear in succession when the path-difference Δ is steadily changed. If, instead of monochromatic radiation of wavenumber $\bar{\nu}$, a whole spectrum is used, the intensity at wavenumber $\bar{\nu}$ will be $I(\bar{\nu})$. That is to say, the power entering the interferometer between wavenumbers $\bar{\nu}$ and $\bar{\nu} + d\bar{\nu}$ is $I(\bar{\nu})d\bar{\nu}$. The intensity emerging when the path-difference is Δ will be

$$dJ(\Delta) = \frac{I(\bar{\nu})}{2}[1 + \cos(2\pi\bar{\nu}\Delta)]d\bar{\nu}$$

and the integral of this over the whole spectrum gives the total intensity at path difference Δ:

$$J(\Delta) = \int_{\overline{\nu}=0}^{\infty} \frac{I(\overline{\nu})}{2} d\overline{\nu} + \int_{\overline{\nu}=0}^{\infty} \frac{I(\overline{\nu})}{2} \cos(2\pi\overline{\nu}\Delta) d\overline{\nu}$$

The first integral is half the total intensity entering the interferometer, $= I_0/2$; so if we put: $2J(\Delta) - I_0 = K(\Delta)$, then:

$$K(\Delta) \rightleftharpoons I(\overline{\nu})$$

The cosine transform is justified since the interferogram in principle is symmetric, and negative path-differences give the same intensity as positive path differences.

The basic idea, then, is that if the interferogram $J(\Delta)$ is measured at suitable intervals of path-difference, the spectral power density $I(\overline{\nu})$ can be recovered by a Fourier transform.

An alternative way of looking at the method is to consider that half of the incoming waveform has been delayed in the longer arm by $c\Delta$ and that the intensity at the detector therefore is the autocorrelation of the incoming signal. The spectrum is the 'spectral power density', i.e. the Fourier transform of the autocorrelation function, as required by the Wiener–Khinchine theorem.

There are some practical difficultes. For example, the path-difference should be increased in *exactly* equal steps, and the intensity emerging from the interferometer should be measured for exactly the same time interval, that is to say, the same total exposure must be made at each station. As the path difference changes there must be no misalignment of the interferometer mirrors, else the fringe contrast is destroyed.

In practice the 'sampling' of the output (the 'interferogram') is never exactly regular. There should be a sample at zero path-difference, and this too is difficult to achieve precisely. The interferogram should be symmetrical about zero path-difference, so that negative path-differences produce the same intensities as the corresponding positive path-differences: usually they do not. However, these are practical details and they have been overcome, so that Fourier spectroscopy has become a routine technique[3].

There are two powerful reasons for doing infra-red spectroscopy this way.

- The radiation passing through the interferometer can be received from a large solid angle – hundreds of times larger than in a corresponding grating spectrometer, so that spectra are obtained far more quickly.

[3] See, for example, *Fundamentals of Fourier Transform Infra-red Spectroscopy.* B. C. Smith, CRC Press, Boca Raton, FL (ISBN 0849324610).

- There is a so-called 'multiplex advantage'[4], which arises from the fact that, in contrast to a monochromator where one wavelength is selected and nearly all the power is discarded inside the instrument, radiation from the whole spectral band is received simultaneously by one detector. If the spectrum is rich in emission lines and bands, the signal/noise ratio is increased by a factor in the region of $\sqrt{N}$ where N is the number of resolved elements in the spectrum.

The net result is that, provided the detector is the principal source of noise in the system (which it is in the infra-red, though not in the visible or UV), there is a substantial gain in efficiency: much fainter sources of radiation can be examined, or spectra can be obtained in a much shorter time. For example, the combination of a Fourier multiplex absorption spectrometer with a chromatograph column can be used for on-line analysis of crude oil, where thousands of organic chemical compounds, each with its own characteristic spectrum, pass in sequence through an absorption cell in the spectrometer and can be identified in turn.

The sampling theorem, described in Chapter 2, holds: samples of the interferogram must be taken at intervals Δ_0 of path-difference not greater than[5] the reciprocal of twice the highest wave-number in the spectrum. If necessary the spectrum can be filtered optically to ensure that there is no 'leakage' of higher frequencies into the spectral band. If the spectral band is narrow, the sampling can be at a multiple of the proper interval, so that aliasing can be allowed. A stabilized HeNe laser beam can be used produce fringes to ensure that the samples are taken at equal intervals. Sometimes, instead of moving a mirror in steps of equal length, stopping, taking a sample, then moving on one step, the path-difference is increased uniformly and smoothly, using the passage of a fringe of laser light (which has a wavelength much shorter than the infra-red spectrum under analysis) to initiate each sample. Then each sample is the *integral* over one step-length of the intensity in the interferogram. What is recorded is the convolution of the interferogram with a top-hat one step-length wide. The spectrum is then the product of the true spectrum with a sinc-function with zeros at $\pm 2\overline{\nu}_f$ and the computed spectrum must be divided by this sinc-function. The process works so long as points near a zero of the sinc-function are not involved.

The other Fourier-related processes discussed earlier also can be applied. A monochromatic line passed through the instrument will yield a sinc-function shape (note: not a sinc^2 function) the result of a finite range of path differences having been used. This has enormous side-lobes in the modular spectrum with

[4] Sometimes called the Fellgett Advantage, after its discoverer.

[5] In practice, usually substantially less than, to leave a gap between the computed spectrum and its mirror image in wavenumber space.

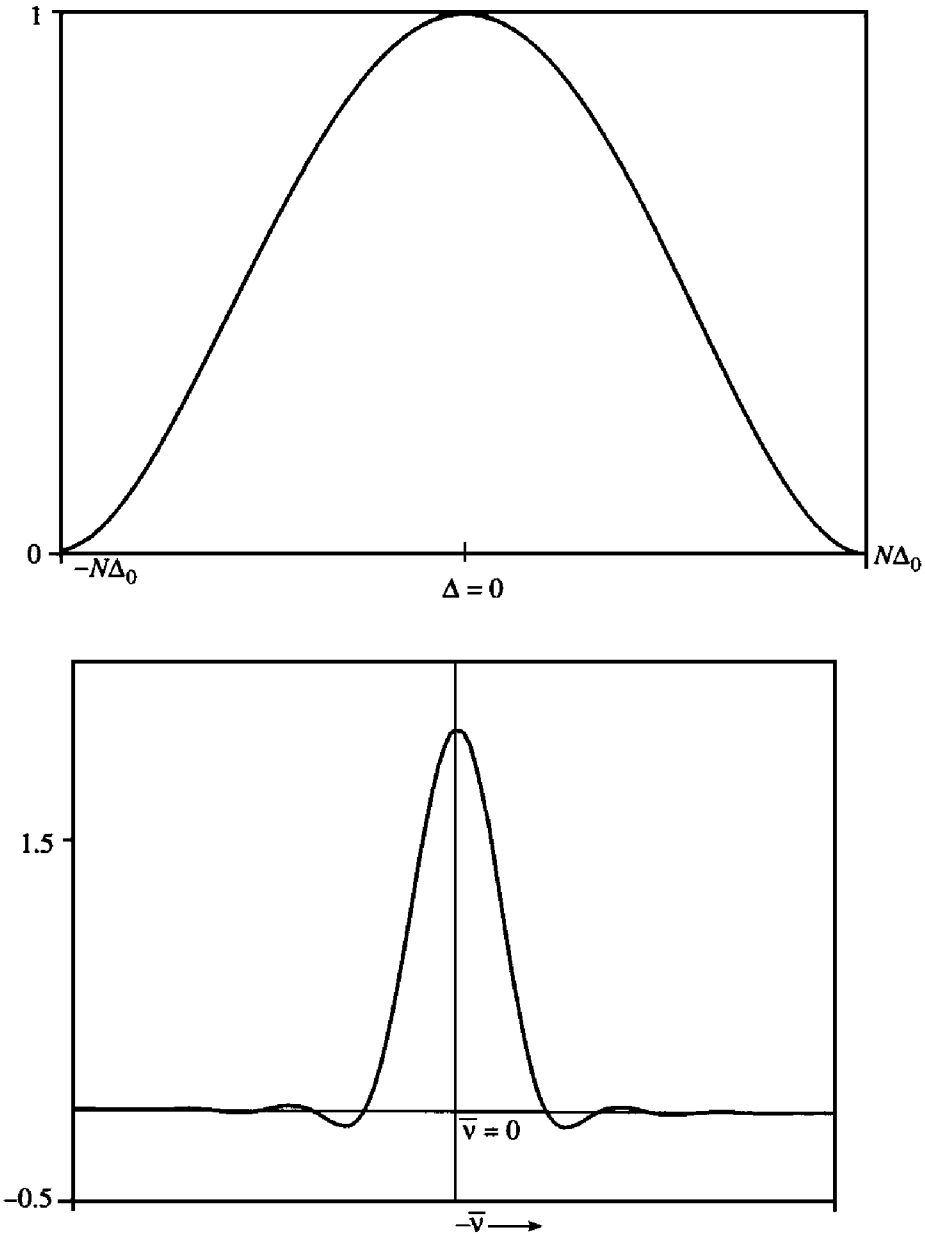


Fig. 5.2. The Connes apodising function for infra-red Fourier multiplex spectroscopy and its ensuing line profile. Without it the line profile would be a sinc-function with secondary peaks *below* zero and −22% of the principal maximum in amplitude.

the amplitude of the first side-lobe being 22% of the principal maximum, and apodisation (see Chapter 3, page 48) is needed to reduce them. There has been much experimentation with apodising functions – which multiply the interferogram before doing the transform – and a function which multiplies the nth sample of the N-sample interferogram $K(\Delta)$ by $[1 - (n/N)^2]^2$ due to Janine Connes[6] has found much favour. It is illustrated in Fig. 5.2.

[6] J. Connes, *Aspen Conference on Multiplex Fourier Spectroscopy*, G. A. Vanasse, A. T. Stair, D. J. Baker (eds). AFCRL-71-0019. 1971, p. 83.

5.2 The shapes of spectrum lines

When an electrical charge is accelerated it loses energy to the radiation field around it. In uniform motion it produces a magnetic field proportional to the current, that is, to $\mathbf{e}\partial x/\partial t$; and if the charge is accelerated the changing magnetic field produces an electric field proportional to $\mathbf{e}\partial^2 x/\partial t^2$. This in turn induces a magnetic field (via Maxwell's equations) also proportional to $\mathbf{e}\partial^2 x/\partial t^2$.

If the charge is oscillating, so are the fields induced around it and these are seen as electromagnetic radiation – in other words, light or radio waves. The power radiated is proportional to the squares of the field strengths $\frac{1}{2}(\epsilon_0\mathbf{E}^2 + \mu_0\mathbf{H}^2)$, which are proportional to $\mathbf{e}(\partial^2 x/\partial t^2)^2$. The total power radiated is $2/(3c^2)\ |\ddot{\mathbf{X}}|^2$, where $\mathbf{X}$ is the maximum value of the dipole moment $\mathbf{ex}$ generated by the oscillating charge. A dipole losing energy in this way is a damped oscillator, and one of Planck's early successes[7] was to show that the damping constant γ is given by:

$$\gamma = \frac{8\pi^2}{3}\frac{\mathbf{e}^2}{mc}\frac{1}{\lambda^2}$$

The equation of motion for an oscillating dipole is then the usual damped harmonic oscillator equation:

$$\ddot{\mathrm{x}} + \gamma\dot{\mathrm{x}} + C\mathrm{x} = 0$$

where C is the 'elastic' coefficient, which depends on the particular dipole, and which describes its stiffness and the frequency of the oscillation. γ is of course the damping coeffient which determines the rate of loss of energy.

The solution of the equation is well known, and is:

$$f(t) = e^{-\frac{\gamma}{2}t}(Ae^{2\pi i\overline{\nu}_0 t} + Be^{-2\pi i\overline{\nu}_0 t})$$

and it is convenient to put $A = 0$ here so that the amplitude, as a function of time is:

$$f(t) = e^{-\frac{\gamma}{2}t}Be^{-2\pi i\overline{\nu}_0 t}$$

The Fourier transform of this gives the spectral distribution of amplitude and when multiplied by its complex conjugate gives the spectral power density:

$$\phi(\overline{\nu}) = \int_0^\infty e^{-\frac{\gamma}{2}t}Be^{2\pi i\overline{\nu}_0 t}e^{-2\pi i\overline{\nu}t}dt$$

(the lower limit of integration is 0 because the oscillation is deemed to begin

[7] M. Planck, *Ann. Physik* **60**, 577 (1897).

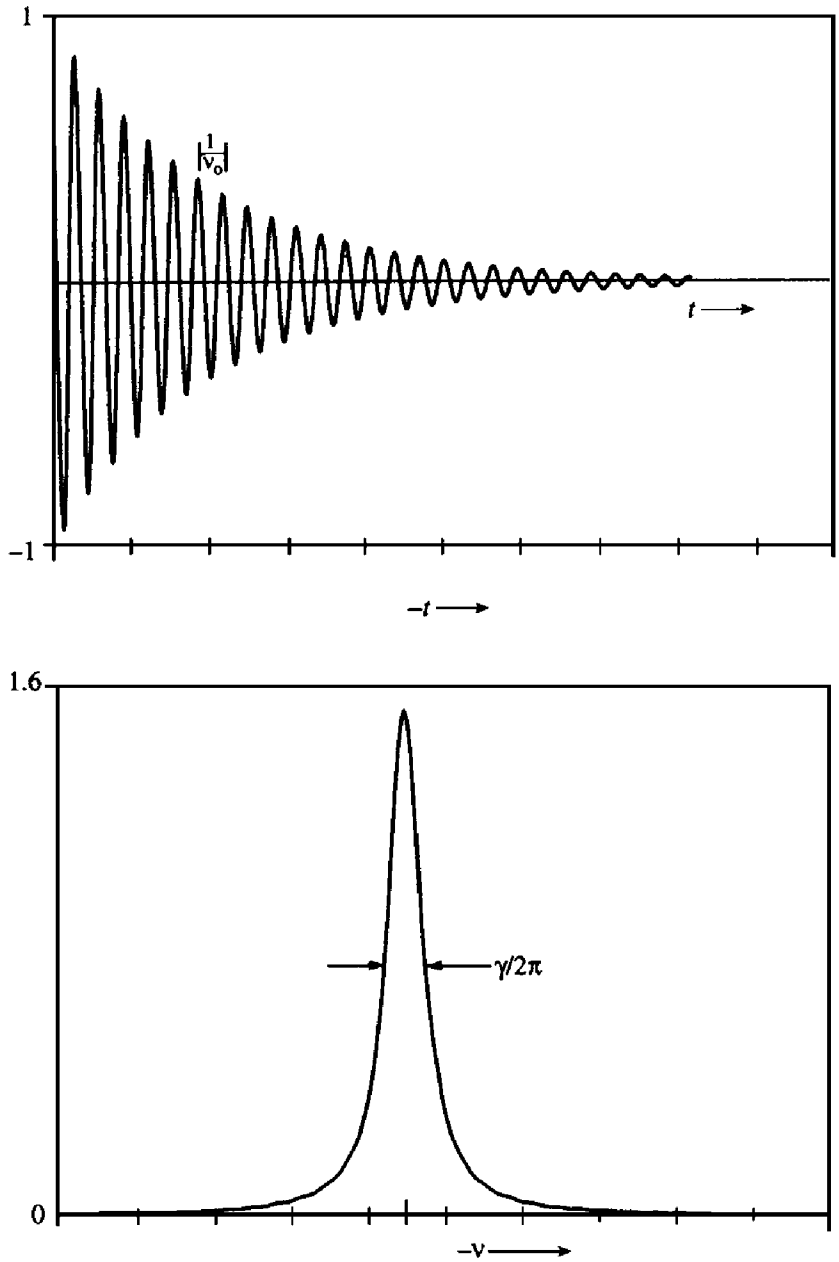


Fig. 5.3. The amplitude of a damped harmonic oscillator and the corresponding spectrum line profile: a Lorentz-function with FWHM $= \gamma/2\pi$. This would be the shape of a spectrum line emitted by an atomic transition if the atoms were held perfectly still during their emission.

then). On integrating we get:

$$\phi(\overline{\nu}) = e^{-\frac{\gamma}{2}t}\left[\frac{e^{-2\pi i(\overline{\nu}_0-\overline{\nu})t}}{2\pi i(\overline{\nu}_0-\overline{\nu})-\gamma/2}\right]_0^{\infty} = \frac{1}{2\pi i(\overline{\nu}_0-\overline{\nu})-\gamma/2}$$

and the spectral power density is then:

$$I(\overline{\nu}) = \frac{1}{4\pi^2(\overline{\nu}_0-\overline{\nu})^2+(\gamma/2)^2}$$

and the line profile is the Lorentz profile discussed in Chapter 1.

The same equation can be derived quantum mechanically[8] for the radiation of an excited atom. The constant $\gamma/2$ is now the 'transition probability', the reciprocal of the 'lifetime of the excited state' if only one downward transition is possible. The FWHM of a spectrum line emitted by an 'allowed' or 'dipole' atomic transition of this sort is usually called the 'natural' width of the line. The shape occurs yet again in nuclear physics, this time called the 'Breit–Wigner formula', and describing in the same way the energy spread in radioactive decay energy spectra. The underlying physics is obviously the same as in the other cases.

There is thus a direct link between the transition probability and the breadth of a spectrum line, and in principle it is possible to measure transition probabilities by measuring this breadth. With typical 'allowed' or 'dipole' transitions – the sort usually seen in spectral discharge lamps – the transition probabilities are in the region of 10^8/s and the breadth of a spectrum line at 5000 Å – in the green – is about 0.003 Å. This requires high resolution, a Fabry–Perot étalon for instance, to resolve it. The measurement is quite difficult since atoms in a gas are in violent motion, and a collimated beam of excited atoms is required in order to see the natural decay by this means.

The violent motion of atoms or molecules in a gas is described by the Maxwellian distribution of velocities. The kinetic energy has a Boltzmann distribution, and the fraction of atoms with velocity v in the observer's line-of-sight has a Gaussian distribution:

$$n(v) = n_0 e^{-mv^2/2KT}$$

with a proportionate Doppler shift, giving a Gaussian profile to what otherwise would be a monochromatic line:

$$I(\lambda) = I_0 e^{-(\lambda-\lambda_0)^2/a^2}$$

The width parameter, a, comes from the Maxwell velocity distribution and $a^2 = 2\lambda^2 kT/mc^2$ where k is Boltzmann's constant, T the temperature, m the mass of the emitting species and c the speed of light.

When we substitute numbers in this formula we find that the intensity profile is a Gaussian with a FWHM proportional to wavelength, and with $\Delta\lambda/\lambda = 7.16 \times 10^{-7}\sqrt{T/M}$ where M is the molecular weight of the emitting species.

This Doppler broadening, or temperature broadening, by itself would give a different line shape from that caused by radiation damping: a Gaussian profile rather than a Lorentz profile. Unless the emitter has a fairly high atomic weight or the temperature is low, the Doppler width is much greater than the natural

[8] See, for example, N. F. Mott and I. N. Sneddon, *Wave Mechanics and its Applications* Oxford. 1948. Ch10, §48.

width. However the line shape that is really observed, after making allowance for the instrumental function, is the convolution of the two into what is called a 'Voigt' profile.

$$V(\lambda) = G(\lambda) * L(\lambda)$$

The Fourier transform will be the product of another Gaussian shape and the Fourier transform of the Lorentz shape. This Lorentz shape is a power spectral density and its Fourier transform is, by the Wiener–Khinchine theorem, the autocorrelation of the truncated exponential function representing the decay of the damped oscillator. This autocorrelation is easily calculated. Let s be the variable paired with λ. Then $L(\lambda) \rightleftharpoons l(s)$ where

$$l(s) = \int_{s'}^{\infty} e^{-\frac{\gamma}{2}s} e^{-\frac{\gamma}{2}(s+s')} ds$$

$$= \frac{1}{\gamma} e^{-\frac{\gamma}{2}s} \quad \gamma > 0 \quad ; \quad = \frac{1}{\gamma} e^{\frac{\gamma}{2}s} \quad \gamma < 0$$

Autocorrelations are necessarily symmetrical and so we can write:

$$l(s) = \frac{2}{\gamma} e^{\frac{\gamma}{2}|s|}$$

So long as s is positive, the Fourier transform of the Voigt line profile is the product

$$v(s) = e^{-\pi^2 s^2 a^2} . e^{-\frac{\gamma}{2}s}$$

and a graph of $\log_e v(s)$ versus s is a parabola. From this parabola the two quantities γ and a can be extracted by elementary methods, and the two components of the convolution are separated.

Voigt profiles occur fairly frequently in spectroscopy. Not only is the line-profile of a damped oscillator a Lorentz curve but, the instrumental profile of a Fabry–Perot étalon is the convolution of a Lorentz profile with a Dirac comb. Fabry–Perot fringes, when used to measure the temperature of a gas or a plasma, therefore show Voigt profiles and if the instrument is used properly – that is with the proper spacing between the plates – the Lorentz half-width will be similar to the Gaussian half-width.

Other causes of spectral lines shapes can easily be imagined. If the pressure is high, atoms will collide with each other before they have had time to finish

their transition. The decaying exponential is then cut short, and the resulting line shape is the convolution of the Lorentz profile with a sinc-function. The width of the sinc-function will be different for every decay, with a Poisson distribution about some average value. The resulting spectrum line then shows 'pressure-broadening', which increases as the intercollision time diminishes, that is, as the pressure increases.

Chapter 6

Two-dimensional Fourier transforms

6.1 Cartesian coordinates

The extension of the basic ideas to two dimensions is simple and direct. As before, we assume that the function $F(x, y)$ obeys the Dirichlet conditions and we can write:

$$A(p, q) = \int_{y=-\infty}^{\infty} \int_{x=-\infty}^{\infty} F(x, y) e^{2\pi i(px+qy)} dx dy$$

$$F(x, y) = \int_{q=-\infty}^{\infty} \int_{p=-\infty}^{\infty} A(p, q) e^{-2\pi i(px+qy)} dp dq$$

The space of the transformed function is of course two-dimensional, like the original space. The extension to three or more dimensions is obvious.

It sometimes happens that the function $F(x, y)$ is separable into a product $f_1(x) f_2(y)$. In this case the Fourier pair, $A(p, q)$ is separable into $\phi_1(p)\phi_2(q)$ and we find separately that:

$$f_1(x) \rightleftharpoons \phi_1(q); \qquad f_2(x) \rightleftharpoons \phi_2(q)$$

If $F(x, y)$ is not separable in this way then the transform must be done in two stages:

$$A(p, q) = \int_{-\infty}^{\infty} e^{2\pi iqy} \left\{ \int_{-\infty}^{\infty} F(x, y) e^{2\pi ipx} dx \right\} dy$$

and whether the x-integral or the y-integral is done first may depend on the particular function, F.

6.2 Polar coordinates

Sometimes – often – there is circular symmetry and polar coordinates can be used. The transform space is also defined by polar coordinates, ρ, ϕ and the substitutions are:

$$x = r\cos\theta; \qquad y = r\sin\theta$$
$$p = \rho\cos\phi; \qquad q = \rho\sin\phi$$

Then

$$A(\rho,\phi) = \int_{r=0}^{\infty}\int_{\theta=0}^{2\pi} F(r,\theta)e^{2\pi i(\rho\cos\phi.r\cos\theta+\rho\sin\phi.r\sin\theta)}rdrd\theta$$

where $rdrd\theta$ is now the element of area in the integration, as can be seen directly or from the Jacobean $\partial(x, y)/\partial(r, \theta)$).

This shortens to:

$$A(\rho,\phi) = \int_{r=0}^{\infty}\int_{\theta=0}^{2\pi} F(r,\theta)e^{2\pi i\rho r\cos(\theta-\phi)}rdrd\theta$$

And if the function A is separable into $P(r)\Theta(\theta)$ the integrals separate into:

$$\int_{r=0}^{\infty} P(r)\left\{\int_{\theta=0}^{2\pi}\Theta(\theta)e^{2\pi i\rho r\cos(\theta-\phi)}d\theta\right\}rdr$$

If there is circular symmetry A is a function of r only, and $\Theta(\theta) = 1$. We can write:

$$A(\rho,\phi) = \int_{r=0}^{\infty} P(r)\left[\int_{\theta=0}^{2\pi} e^{2\pi i\rho r\cos(\theta-\phi)}d\theta\right]rdr$$

We now put $\theta - \phi = \alpha$, a new independent variable, with $d\alpha = d\phi$ (the integral, being taken around 2π, does not depend on the value of θ).

Then the θ-integral becomes

$$\int_0^{2\pi} e^{2\pi i\rho r\cos\alpha}d\alpha$$

and this (see Appendix) is equal to $2\pi J_0(2\pi\rho r)$ where J_0 denotes the zero-order Bessel-function.

Then:

$$A(\rho) = 2\pi\int_0^{\infty} P(r)rJ_0(2\pi\rho r)\,dr$$

which is known as a Hankel transform. It is a close relative of the Fourier transform.

The Bessel functions of any order n, $J_n(x)$, have the property that when they are multiplied by $x^{\frac{1}{2}}$ they form an orthogonal set[1] like the trigonometric functions:

$$\int_0^\infty x J_n(x) J_m(x)\, dx = \delta_m^n$$

where δ_m^n is the usual Kronecker-delta ($\delta_m^n = 0$ if $m \neq n$ and $\delta_m^m = 1$.)

Consequently there is an inversion formula as in the Fourier transform, so that $P(r)$ can be recovered from:

$$P(r) = 2\pi \int_0^\infty A(\rho)\rho J_0(2\pi\rho r)\, d\rho$$

and the two functions are symbolically linked by:

$$P(r) \Leftrightarrow A(\rho)$$

6.3 Theorems

Some, but not all, of the theorems derived in Chapter 2 carry over into two dimensions. As above, assume that $P(r) \Leftrightarrow A(\rho)$

The similarity theorem: $P(kr) \Leftrightarrow (1/k^2)A(\rho/k)$

The addition theorem: $P_1(r) + P_2(r) \Leftrightarrow A_1(\rho) + A_2(\rho)$

Rayleigh's theorem:

$$\int_0^\infty |\, P(r)\,|^2\, r.dr = \int_0^\infty |\, \Phi(\rho))\,|^2\, \rho.d\rho$$

There is a convolution theorem like that in one dimension but one of the functions has to explore the whole plane in two dimensions instead of just sliding over the other. The product integral is done at each point in the plane to obtain the convolution:

$$C(r') = P_1(r) * *P_2(r) = \int_{r=0}^\infty \int_{\theta=0}^{2\pi} P_1(r)P_2(R)\, r dr d\theta$$

where $R^2 = r^2 + r'^2 - 2rr'\cos\theta$ and the symbol ** is used to denote a two-dimensional convolution.

There is a corresponding convolution theorem:

$$C(r) \Leftrightarrow A_1(\rho)A_2(\rho)$$

[1] A proof of the orthogonality is given in Bracewell, *The Fourier Transform and its Applications*, (see bibliography).

6.4 Examples of two-dimensional Fourier transforms with circular symmetry

6.4.1 The top-hat function, also known as 'circ' or 'disk'.

$$P(r) = h, 0 < r < a$$
$$= 0, a < r < \infty$$
$$A(\rho) = 2\pi h \int_0^a r.J_0(2\pi\rho r)\, dr$$

We use the property (See Appendix):

$$\frac{d}{dx}(xJ_1(x)) = xJ_0(x)$$

let $2\pi\rho r = x$; $\quad 2\pi\rho dr = dx$

Then:

$$A(\rho) = 2\pi h \int_0^{2\pi a\rho} \frac{x}{2\pi\rho} J_0(x) \frac{dx}{2\pi\rho}$$

$$= \frac{h}{2\pi\rho^2} \int_0^{2\pi a\rho} xJ_0(x)dx = \frac{h}{2\pi\rho^2} [xJ_1(x)]_0^{2\pi a\rho}$$

$$= \frac{ah}{\rho} J_1(2\pi a\rho) = \pi a^2 h \left\{ \frac{2J_1(2\pi a\rho)}{2\pi a\rho} \right\}$$

and finally:

$$A(\rho) = \pi a^2 h \,\text{Jinc}(2\pi a\rho) \text{ where } \text{Jinc}(x) = \frac{2J_1(x)}{x}$$

Jinc contains the factor of 2 in order that Jinc(0) = 1.

This, with a as the aperture radius and ρ as $\sin\theta/\lambda$, gives the amplitude of diffraction of light or radio waves at a circular aperture. The intensity distribution, which is the square modulus of this, is the famous 'Airy disc' familiar to students of the telescope and other optical imaging instruments.

6.4.2 The thin annulus

$P(r)$ is a circle or radius a. In optics, a very thin ring transmitting light:

$$P(r) = h\delta(r - a)$$

then:

$$A(\rho) = 2\pi h \int_0^\infty r\delta(r - a)J_0(2\pi\rho r)\, dr$$

$$= 2\pi ahJ_0(2\pi a\rho)$$

6.5 Applications

6.5.1 Fraunhofer diffraction by a rectangular slot

The simple two-dimensional Fraunhofer theory of Chapter 3 can now be elaborated. There, we assumed that the element dS on the surface S was equal *in area* to dx, the width of a slit $\times$ unit length perpendicular to the diagram.

Now we can use $dS = dxdy$, a small rectangle in the diffracting aperture, perpendicular to the direction of propagation, and we can calculate the diffracted amplitude in a direction specified by direction cosines l, m, n.

From this we can calculate the intensity at a point on a plane at a distance z from the aperture. If the amplitude at the element of area $dxdy$ at $Q(x, y)$ is $Kdxdy$, then at P, on the distant screen, it will be $Kdxdye^{\frac{2\pi i}{\lambda}R'}$ and from elementary coordinate geometry, $R' = R - lx - my$ where l and m are the direction cosines of the line OP and R is the distance from the origin to the point P on the distant screen.

The total disturbance at P is then the sum of all the elementary disturbances from the $z = 0$ plane, so that we can write:

$$A(p, q) = \int\int_{aperture} Kdxdye^{2\pi i(\frac{R}{\lambda} - \frac{lx}{\lambda} - \frac{my}{\lambda})}$$

$$= C\int\int_{aperture} e^{-2\pi i(px+qy)}dxdy$$

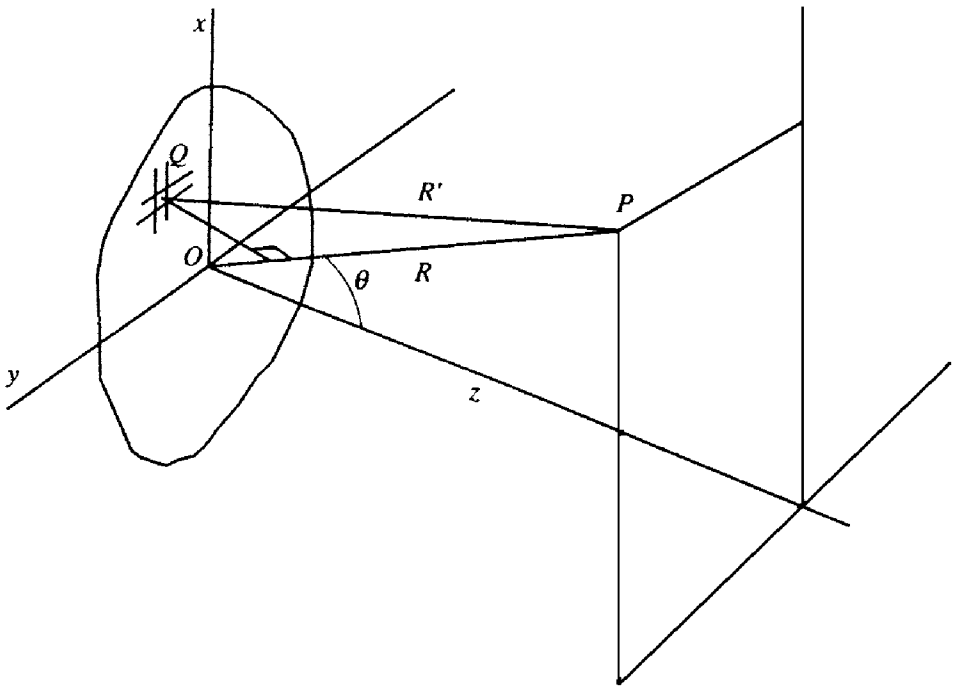


Fig. 6.1. The two-dimensional diffracting aperture, in cartesian coordinates.

where $p = l/\lambda$, $q = m/\lambda$ and C is a constant which depends on the area of the aperture, and contains the constant phase factors and any other things which do not affect the relative intensity in the diffraction pattern.

If the aperture is a rectangle of side $2a$, $2b$ the integrals separate:

$$A(p, q) = C \int_{-a}^{a} e^{-2\pi i p x} dx \int_{-b}^{b} e^{-2\pi i q y} dy$$

and the intensity diffracted in the direction whose direction cosines are $p\lambda$, $q\lambda$ is the square-modulus of this.

$$I(p, q) = I_0 \operatorname{sinc}^2(2\pi a p) \operatorname{sinc}^2(2\pi b q)$$

Notice that, perhaps surprisingly, the intensity at the central peak is proportional to the *square* of the area of the aperture.

6.5.2 *Fraunhofer diffraction by a circular aperture*

If the aperture is circular and of radius a, the Hankel transform is used, with $x = r\cos\theta$, $y = r\sin\theta$ as before and with $p = l/\lambda = \rho\cos\phi$; $q = m/\lambda = \rho\sin\phi$ and $\rho^2 = p^2 + q^2$.

The third direction cosine, n is given by

$$n^2 = 1 - l^2 - m^2 = 1 - (p\lambda)^2 - (q\lambda)^2$$

so that

$$\rho^2 = \frac{1}{\lambda^2}(l^2 + m^2) = \frac{1 - n^2}{\lambda^2}$$

or $\rho = \sin\theta/\lambda$, where θ is the angle between OP and the z-axis. Then, immediately:

$$A(\theta) = A(0)\frac{J_1(2\pi a \sin\theta/\lambda)}{2\pi a \sin\theta/\lambda}$$

$$I(\theta) = I(0)\left[\frac{J_1(2\pi a \sin\theta/\lambda)}{2\pi a \sin\theta/\lambda}\right]^2$$

Which is the formal equation for the intensity in the Airy disc. Again notice that $I(0)$ is proportional to the square of the area of the aperture. The total power in the pattern is of course proportional to the area of the aperture, but as the radius of the diffracting aperture doubles, for example, the pattern on a distant screen has half the radius and one quarter the area, out to the first zero-intensity ring.

As an exercise, the calculation of the intensity distribution in the diffraction pattern made by an annular aperture can be done. If the inner and outer radii of

the annulus are a and b, the amplitude function is:

$$A(\theta) = K\left[a^2\frac{J_1(2\pi a\sin\theta/\lambda)}{2\pi a\sin\theta/\lambda} - b^2\frac{J_1(2\pi b\sin\theta/\lambda}{2\pi b\sin\theta/\lambda}\right]$$

and the intensity distribution is the square of this.

A graph of this function shows that the central maximum is narrower than that of the Airy disc for the same outer radius. A telescope with an annular aperture apparently beats the 'Rayleigh criterion' for spatial resolution. However, it does so at the expense of putting a lot of intensity into the ring around the central maximum, and the gain is usually more illusory than real.

6.6 Solutions without circular symmetry

In general, provided that the aperture function can be separated into $P(r)$ and $\Theta(\theta)$. then as we saw earlier:

$$A(\rho,\phi) = \int_{r=0}^{\infty} P(r)\left\{\int_{\theta=0}^{2\pi}\Theta(\theta)e^{2\pi i\rho r\cos(\theta-\phi)}d\theta\right\}rdr$$

Consider the interference pattern of a set of apertures – or antennae – equally spaced around the circumference of a circle. If there are N of them the θ-dependent function is:

$$\Theta(\theta) = \sum_{0}^{N-1}\delta(\theta - 2\pi n/N)$$

and the r-dependent part is

$$P(r) = \delta(r-a)$$

In other words, the sources are equally spaced at angles $2\pi/N$ around the circle of radius a

Then

$$A(\rho,\phi) = \int_0^{\infty} r\delta(r-a)\sum_{0}^{N-1}e^{2\pi i\rho r\cos(2\pi n/N-\phi)}dr$$

$$= a\sum_{0}^{N-1}e^{2\pi i\rho a\cos(2\pi n/N-\phi)}$$

This is as far as the analysis can be taken. The pattern, $I(\rho,\phi)$ can be computed without difficulty from this expression, and is a typical example of a problem solved by computer after analysis fails. The particular case of $N = 2$ yields the familiar pattern of two-beam interference, including the hyperbolic shapes of

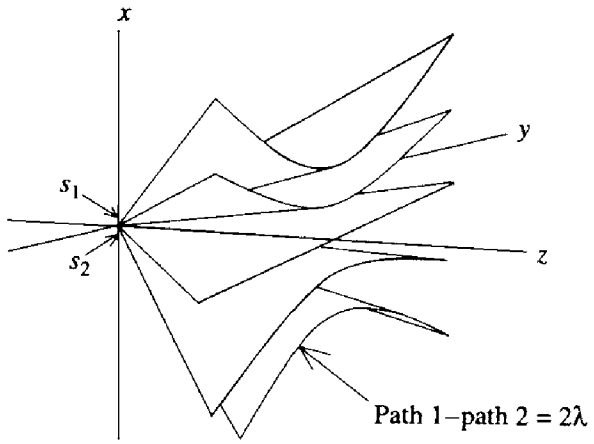


Fig. 6.2. The cones of maximum intensity in a two-beam interference pattern. The two interfering sources are on the x-axis, above and below the origin.

the fringes on a distant plane surface:

$$\begin{aligned} A(\rho, \phi) &= a\left[e^{2\pi i a\rho\cos\phi} + e^{2\pi i a\rho\cos(\pi-\phi)}\right] \\ &= 2a\cos(2\pi a\rho\cos\phi) \end{aligned}$$

and the intensity pattern is given by

$$I(\rho, \phi) = 4a^2\cos^2(2\pi a\rho\cos\phi)$$

which has maxima when $2a\rho\cos\phi$ is integer. Since $\rho = \sin\alpha/\lambda$ the maxima occur when $\phi = n\lambda/2a\sin\alpha$. α is the angle between the z-axis and the direction of diffraction, ϕ is the *azimuth* (angle in the p, q-plane), so that interference fringes, the maxima of $I(\rho, \phi)$, emerge along directions defined by the condition $(2a/\lambda)\sin\alpha\cos\phi = \text{constant}$, that is to say, on cones of semi-angle ϕ about the $\phi = 0$-axis. If they are received on a plane perpendicular to the z-axis they show hyperbolic shapes, but on a plane perpendicular to the x-axis (the $\phi = 0$ axis: the axis containing the two sources) the shapes would be concentric circles.

Other cases, like $N = 4$ yield to analysis as well. But in general, to parody Clausewitz, it is best to regard computation as the continuation of analysis by other means.

Chapter 7

Multi-dimensional Fourier transforms

The physical world seems to comprise four dimensions of space and time, and other dimensions, like electrical potential or temperature are used occasionally for drawing graphs. For this reason Fourier transforms in three or more dimensions can be useful sometimes. The extension is not difficult and can sometimes give greater insight to what is happening in Nature than mere geometry. This chapter describes some of the functions and ideas which are helpful in manipulating multi-dimensional Fourier transforms.

7.1 The Dirac wall

This is described by

$$f(x, y) = \delta(x - a)$$

and is zero everywhere except on the line $x = a$, where it is infinite. Despite this infinity, it may be envisaged as a wall, parallel to the y-axis, of unit height, as in Fig. 7.1.

Its two-dimensional Fourier transform is given by

$$\begin{aligned}\phi(p, q) &= \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} \delta(x - a) e^{2\pi i p x} e^{2\pi i q y} dx dy \\ &= \int_{y=-\infty}^{\infty} e^{2\pi i p a} e^{2\pi i q y} dy \\ &= e^{2\pi i p a} \delta(q)\end{aligned}$$

which has a complex amplitude[1] and is zero except on the line $q = 0$.

[1] In the sense mentioned in Chapter 1, that a δ-function is infinite at each point but its integral, which we consider to be its 'amplitude', is unity.

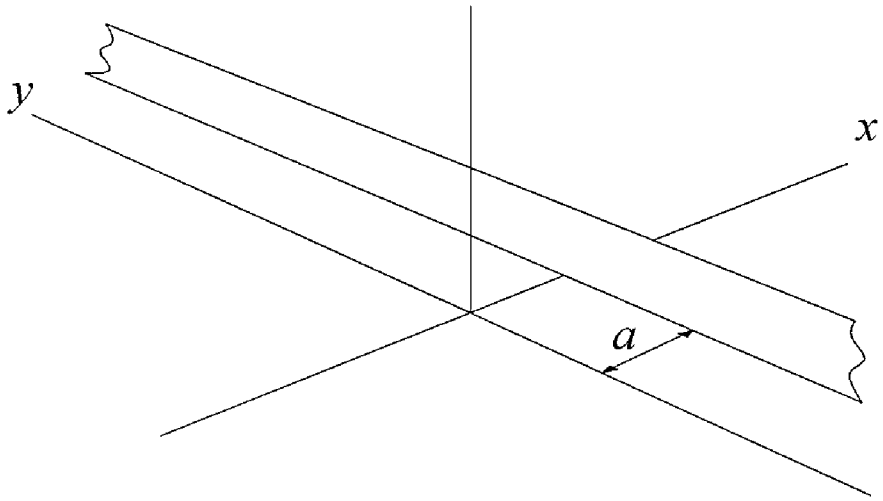


Fig. 7.1. A simple Dirac wall, $f(x, y) = \delta(x - a)$.

A pair of these Dirac walls, equally disposed about the y-axis has a Fourier transform given by:

$$\phi(p, q) = 2\delta(q) \cos 2\pi pa$$

A wall standing on a line inclined to the y-axis at an angle θ is described by $f(x, y) = \delta(lx + my - c)$, where $l = \cos\theta$, $m = \sin\theta$ and c is the length of the perpendicular from the origin to the line. The δ-function is zero everywhere on the x, y-plane except on the line and its two-dimensional Fourier transform is:

$$\phi(p, q) = \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} \delta(lx + my - c) e^{2\pi ipx} e^{2\pi iqy} dxdy$$

Do the y-integration first[2]

$$\phi(p, q) = \frac{1}{m} \int_{x=-\infty}^{\infty} e^{2\pi ipx} e^{2\pi iq(c-lx)/m} dx$$

and notice that 'integration' here is a simple replacement of the variable in the exponential by the argument of the δ-function.

Then, rearranging the exponents:

$$\begin{aligned} \phi(p, q) &= \frac{1}{m} e^{2\pi iqc/m} \int_{x=-\infty}^{\infty} e^{2\pi ix(p-lq/m)} dx \\ &= e^{2\pi iqc/m} \delta(mp - lq) \end{aligned}$$

which is zero except on the line $mp - lq = 0$ in the p, q-plane.

[2] Bearing in mind from Chapter 1 that $\delta(lx + my - c) = \frac{1}{m}\delta(y - (c - lx)/m)$.

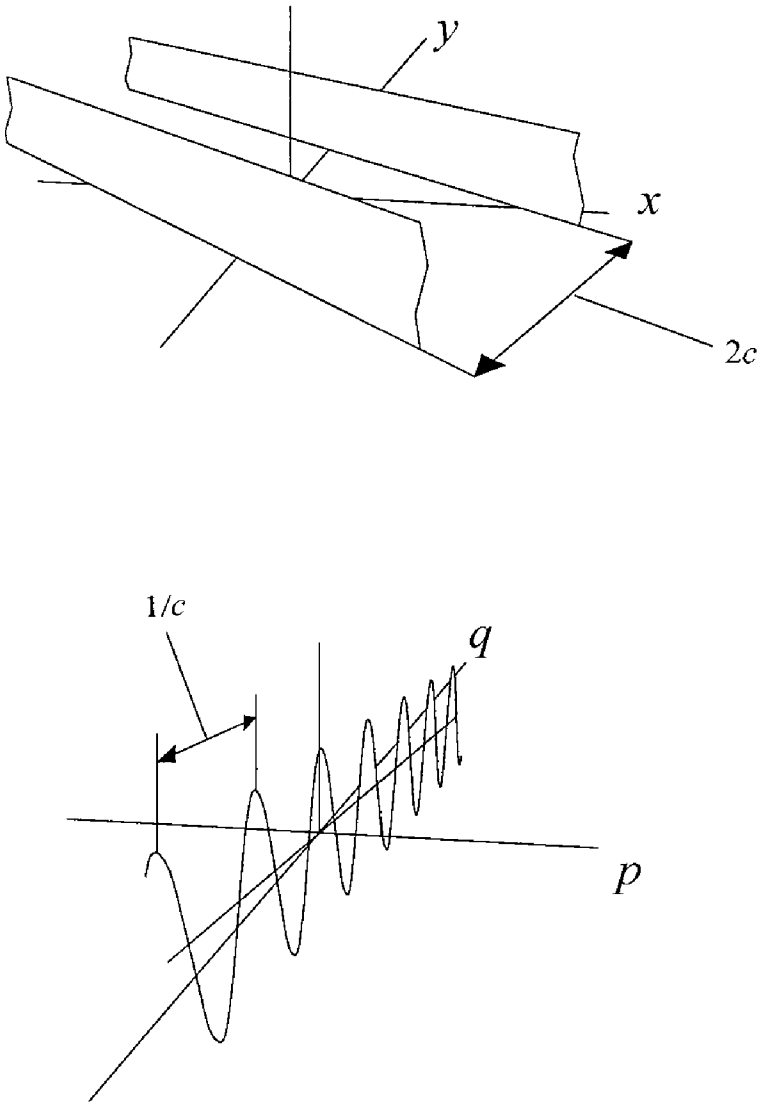


Fig. 7.2. The Fourier transform of a pair of Dirac walls.

Equally, the y-integration could have been done first, in which case the Fourier transform would have been

$$e^{2\pi ipc/l}\delta(mp - lq)$$

The period in the phase-factor is $1/c$ and it is measured along the direction of the line $mp - lq = 0$ in p, q-space. As we shall see later it is possible to envisage a one-dimensional variable $u = p/l$ or q/m, conjugate to c, along that line, and then the above function describes a complex sinusoid of period $1/c$ along the Dirac wall. Its one-dimensional Fourier transform *along that line* would then be a delta function at a distance c from the origin, situated at the point lc, mc in x, y-space. This δ-function would lie on the line $mx - ly = 0$.

This is the place to mention that much insight can be gained by superposing the two planes so that the p- and q-axes of one coincide with the x- and y-axes of the other. In this example, the Fourier transform of the Dirac wall lies on a line in the p, q-plane perpendicular to the wall in the x, y-plane.

A pair of Dirac walls, equally disposed on either side of the origin, has a two-dimensional Fourier transform given by:

$$\delta(lx + my - c) + \delta(lx + my + c) \rightleftharpoons \delta(mp - lq) \cdot 2\cos 2\pi qc/m$$

that is to say, a Dirac wall with a sinusoidally-varying amplitude and lying on the line $mp - lq = 0$.

Notice particularly that, with this superposition of the two planes, the function and its transform are related in spatial position, *irrespective of the orientation*[3] *coordinate systems chosen.* In this example they lie on perpendicular Dirac walls.

7.2 Computerized axial tomography

A particularly useful application of these ideas is to be found in computerized transverse axial scanning tomography, vulgarly known as CAT-scanning or C-T scanning. Imagine a Dirac wall taking a vertical slice through a two-dimensional function $F(x, y)$ lying on the x, y-plane (Fig. 7.3a). If the wall stands on the line $lx + my - c = 0$ the product is zero everywhere except on the line. On the line stands a Dirac wall (Fig. 7.3b) with amplitude varying as $F(x, (c - lx)/m)$. The line-integral (Fig. 7.3c):

$$P_l(c) = \int_{-\infty}^{\infty} F(x, y)\delta(lx + my - c)\, ds \tag{7.1}$$

(where ds is the line element along the direction defined by l), depends only on l and c. Incidentally, $P_l(c)$ is known as the *Radon transform*[4] of $F(x, y)$. It can be imagined, as in the previous section, as a δ-function of amplitude $P_l(c)$ standing on the line $mx - ly = 0$ at a distance c from the origin. With c as variable it becomes a function of c along the line and this function is called the *projection* of $F(x, y)$ in the direction θ where $\cos\theta = l$.

Now as the direction θ rotates from 0 to π, the various functions $P_l(c)$ sweep out a two-dimensional function $Q(x, y)$ on the x, y-plane.

[3] But not of the position of the origin.

[4] vide, e.g. S. R. Deans: *The Radon Transform and Some of its Applications.* J. Wiley, New York, 1983.

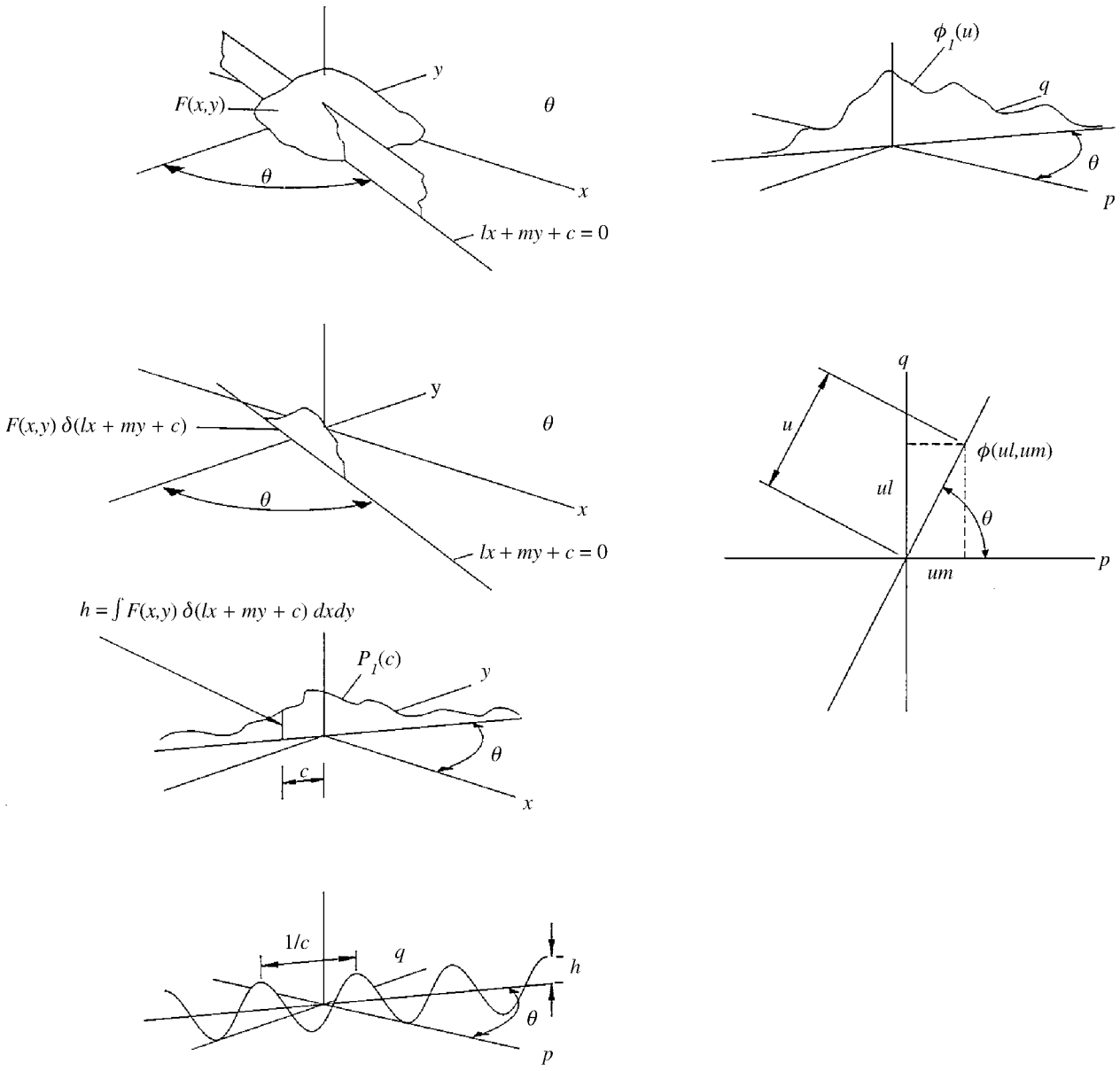


Fig. 7.3. Illustrating the steps in computerized axial tomography.

What is more interesting, however, is the function which results from first taking the *one-dimensional* Fourier transform of $P_l(c)$ along the line $mx - ly = 0$, with c as variable on the x, y-plane and u as its conjugate on the p, q-plane. In the p, q-plane this Fourier transform, $\phi_l(u)$, will lie on the line $mp - lq = 0$ which is superimposed on $mx - ly = 0$ on the x, y-plane. This set of Fourier transforms too, sweeps out a two-dimensional function $\Phi(p, q) = \phi_l(lu, mu)$ in Fig. 7.3e as the direction of the projection changes from 0 to π.

We now demonstrate the remarkable fact that $\Phi(p, q)$ is the two-dimensional Fourier transform of $F(x, y)$.

To do this we first of all write $\delta(lx + my - c)$ as a one-dimensional Fourier integral, using c as the variable and u as its conjugate:

$$\delta(lx + my - c) = \int_{u=-\infty}^{\infty} e^{2\pi i(lx+my-c)u} du$$

$$= \int_{u=-\infty}^{\infty} e^{2\pi iu(lx+my)} e^{-2\pi icu} du$$

and if we insert this into the equation for $P_l(c)$ [equation 7.1] we find:

$$P_l(c) = \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} F(x, y) \int_{u=-\infty}^{\infty} e^{2\pi iu(lx+my)} e^{-2\pi icu} du dx dy$$

and changing the order of integration:

$$P_l(c) = \int_{u=-\infty}^{\infty} \left(\int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} F(x, y) e^{2\pi iu(lx+my)} dx dy \right) e^{-2\pi icu} du$$

Within the brackets is $\Phi(ul, um)$ the two-dimensional Fourier transform of $F(x, y)$, and notice (Fig. 7.3e) that on the p, q-plane $ul = p, um = q$.

Thus:

$$P_l(c) = \int_{-\infty}^{\infty} \Phi(ul, um) e^{-2\pi icu} du$$

Whence, for a fixed direction θ,

$$\Phi(ul, um) = \Phi(p, q) = \int_{-\infty}^{\infty} P_l(c) e^{2\pi icu} dc$$

This is still a one-dimensional transform and it defines $\Phi(p, q)$ along the line $mp - lq = 0$. In Radon transform theory it is called the *projection slice* theorem.

Thus if we know $P_l(c)$ for all azimuths θ, from 0 to π, and do the complete set of one-dimensional transforms, the two-dimensional function $\Phi(p, q)$ is known. The original function $F(x, y)$ is then obtained from:

$$F(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \Phi(p, q) e^{-2\pi i(px+qy)} dp dq$$

If a three-dimensional object is partially transparent to radiation such as X-rays, visible light or a particle beam, it is possible to make a two-dimensional map of the absorption coefficient (α) on a plane section through it. When monochromatic radiation is transmitted through the object it is attenuated

according to *Beer's law*, which states that the intensity of the radiation transmitted in the x-direction falls according to:

$$\frac{\partial I}{\partial x} = -I\alpha$$

where I is the intensity at the point x along the direction of transmission, and α is the absorption coefficient for that wavelength or frequency. The coefficient depends on the nature of the absorbing material and if the absorption is constant along the path, then Beer's law in one dimension takes the form:

$$I(x) = I_o e^{-\alpha x}$$

If α varies from point to point, then the integral along the transmission path (the 'line-integral') must be taken and:

$$I(x) = I_o e^{-\int_0^x \alpha(x)\cdot dx}$$

From this the following useful equation emerges:

$$\int_0^x \alpha(x)\cdot dx = ln(I_o/I(x))$$

The function of computer axial tomography is to make a two-dimensional plot – the map – of α in a plane slice through the object. Notice that if the source and the detector are both outside the object then the line-integral of α is identical with

$$\int_{-\infty}^{\infty} \alpha(x)\cdot dx = ln(I_o/I(x))$$

Consider an absorbing object – a skull for example – through which a narrow beam of X-rays can be transmitted from a source to a detector. The line-integral of the absorption coefficient α follows from the logarithm of the ratio of the intensity at the source to the intensity at the detector.

We now use this narrow beam as a saw-blade to 'cut' a plane section through the object.

In Fig. 7.3 the z axis is used to depict the absorption coefficient, $\alpha(x, y)$, in the section and the radiation beam is directed along the line $lx + my - c = 0$. We replace the $F(x, y)$ by $\alpha(x, y)$ in equation (7.1):

$$P_l(c) = \int_{-\infty}^{\infty} \alpha(x, y)\cdot \delta(lx + my - c)\, ds$$

As the the source and detector move together in the x, y-plane in a direction perpendicular to the transmission direction, c is changing while l is constant. The beam, as it moves, takes a slice through the absorbing object (hence the

word 'tomography') and there will be a measurement of the line-integral $P_l(c)$ as a function of c (Fig. 7.3c).

The one-dimensional Fourier transform, $\phi_l(u)$, of $P_l(c)$ maps out, as the direction θ of the projection changes, the two-dimensional function $\Phi(p, q)$, the aggregate of these ϕ-functions over all azimuths and this, as we have seen, is the two-dimensional Fourier transform of $\alpha(x, y)$.

This is the central idea in computer axial tomography

The inevitable conclusion is that, provided $P_l(c)$ is measured for every azimuth θ $(= cos^{-1}l)$ from $\theta = 0$ to $\theta = 180°$, and the one-dimensional Fourier transforms are taken, then the function $\Phi(ul, um)$ is known over the whole p, q-plane[5] and the inverse transform of $\Phi(ul, um)$ is $\alpha(x, y)$, the original desired function:

$$\alpha(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \Phi(q, p) \cdot e^{-2\pi iqx} e^{-2\pi ipy} dq \cdot dp$$

The function $\alpha(x, y)$ then represents the two-dimensional distribution of density, or absorption cross-section of X-rays or other exploring radiation beams by the material through which the radiation passes.

The practical implementations[6] of the idea have been manifold and to the universal public good. This brief description ignores the extraordinary extensions of the idea[7] in areas as diverse as cosmology and geophysics, and it must also be mentioned that other methods than Fourier transforms may be used to recover the required data. There can have been few inventions more deserving of a Nobel prize than this one.

7.3 A 'spike' or 'nail'

This is described by a two-dimensional δ-function, $\delta(x - a)\delta(y - b)$ and is zero everywhere in the x, y-plane except at the point (a, b). As the product of a function of x and a function of y it is separable and its Fourier transform is $e^{2\pi ipa} e^{2\pi iqb}$.

[5] Or as much of it as the resolution of the source and detector will permit. Instrumental considerations limit the spatial frequencies accessible to a CAT-scanner and only a limited area – about 2 mm^2 – of frequency-space (the p, q-plane) is useable in practice with X-ray tomography.

[6] The 1979 Nobel prize for physiology and medicine was awarded to G. N. Hounsfield and A. Cormack for the invention of CAT-scanning. The prototype CAT-scanner, constructed by EMI, went into service at the Atkinson Morley's Hospital, Wimbledon, in 1971.

[7] Described for example, in G. T. Herman's *Image Reconstruction From Projections* – see bibliography.

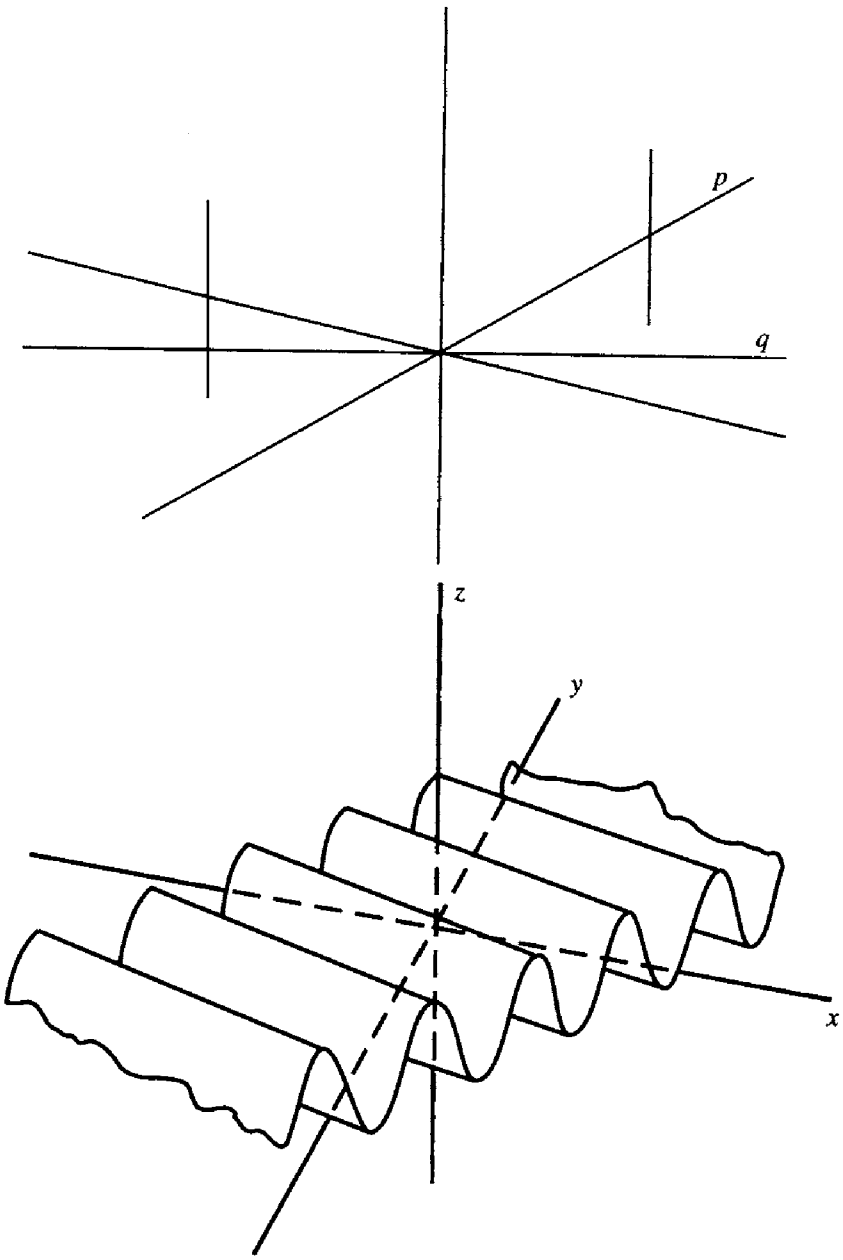


Fig. 7.4. The Fourier transform of a pair of nails at $\pm(x, y)$.

A pair of such nails equally disposed about the origin is described by:

$$f(x, y) = \delta(x - a)\delta(y - b) + \delta(x + a)\delta(y + b)$$

and its Fourier transform is:

$$\phi(p, q) = 2\cos 2\pi(pa + qb)$$

This is a corrugated sheet. Lines of constant phase (wavecrests) lie on the lines $pa + qb =$ integer, and are illustrated in Fig. 7.4 and again, on superposition, the line joining the nails on the (x, y)-plane is perpendicular to the wavecrests on the (p, q)-plane.

7.4 The Dirac fence

This is an infinite row of equally-spaced δ-functions (the fence posts) along a line. When it runs along the x-axis and the spacing of the posts is a, the fence is described by:

$$f(x, y) = \left[\sum_{n=-\infty}^{\infty} \delta(x - na)\right] \delta(y) = \text{Ш}_a(x)\delta(y)$$

Its Fourier transform follows from the Fourier transform of a *Ш*-function mentioned in Chapter 1 and is $\frac{1}{a}\text{Ш}_{\frac{1}{a}}(p)$, a parallel set of walls, all parallel to the q-axis with spacing $1/a$.

If the fence is inclined to the x-axis at an angle θ, then $l(= \sin\theta)$ and $m(= \cos\theta)$ define the direction of the line of the fence, and the fence is described by:

$$f(x, y) = \left[\sum_{n=-\infty}^{\infty} \delta(lx + my - na)\right] \delta(mx - ly)$$

The first factor requires the function to be zero except when $lx + my = na$ (thus defining a set of parallel walls) and the second requires that it be zero except on a line perpendicular to the first set, passing through the origin. This can also be written as:

$$f(x, y) = \text{Ш}_a(lx + my)\delta(mx - ly)$$

The Fourier transform can be seen graphically as the convolution of the two separate transforms. The transform of the first factor is:

$$\phi_1(p, q) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\sum_{n=-\infty}^{\infty} \delta(lx + my - na)e^{2\pi ipx}e^{2\pi iqy}dxdy$$

and once more the simple rule for integrating a product which includes a δ-function applies:

$$\begin{aligned}\phi_1(p, q) &= \frac{1}{l}\int_{-\infty}^{\infty}\sum_{n=-\infty}^{\infty} e^{2\pi ip(na-my)/l}e^{2\pi iqy}dy \\ &= \frac{1}{l}\sum_{n=-\infty}^{\infty} e^{2\pi ipna/l}\int_{-\infty}^{\infty} e^{2\pi iy(q-pm/l)}dy \\ &= \delta(ql - pm)\sum_{n=-\infty}^{\infty} e^{2\pi ipna/l}\end{aligned}$$

which is a row of fence-posts spaced $1/a$ apart[8] lying on the line $lq = mp$.

[8] Actually the product of a wall lying on the line $ql = pm$, and an infinite set of walls of spacing a/l lying perpendicular to the p-axis.

The second factor transforms similarly:

$$\begin{aligned}\phi_2(p,q) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \delta(mx-ly)e^{2\pi ipx}e^{2\pi iqy}dxdy \\ &= \frac{1}{m}\int_{-\infty}^{\infty} e^{2\pi ip(ly/m)}e^{2\pi iqy}dy \\ &= \delta(lp+mq)\end{aligned}$$

which is a wall passing through the origin, lying on the line $lp = -mq$; that is, perpendicular to the fence post of the first factor when the p, q-plane is superimposed on the x, y-plane.

The convolution of these two factors, $\phi_1(p,q) * *\phi_2(p,q) = w(p,q)$, is an infinite series of parallel walls, spaced $1/a$ apart, lying on lines parallel to the line $lp = -mq$. On superposition of the two spaces, these walls are perpendicular to the original fence line. Diagrammatically:

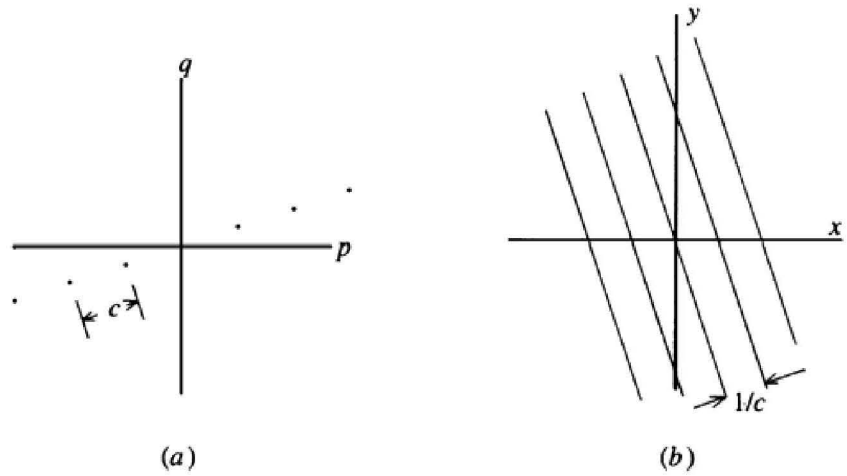


Fig. 7.5. A line of fence-posts of spacing c and its Fourier transform, a series of parallel walls a distance $1/c$ apart.

7.5 The 'bed of nails'

Now consider the convolution of two fences, f_1 and f_2. Let each lie on a line through the origin, at angles θ_1 and θ_2 and with spacings a_1 and a_2. The convolution, $f_1 * *f_2$, will be a two-dimensional array of δ-functions – a 'bed of nails'.

The Fourier transform of this convolution is the product w_1w_2 of the two transforms, each one a series of parallel walls, and differs from zero only when both factors are different from zero. This gives another 'bed of nails'.

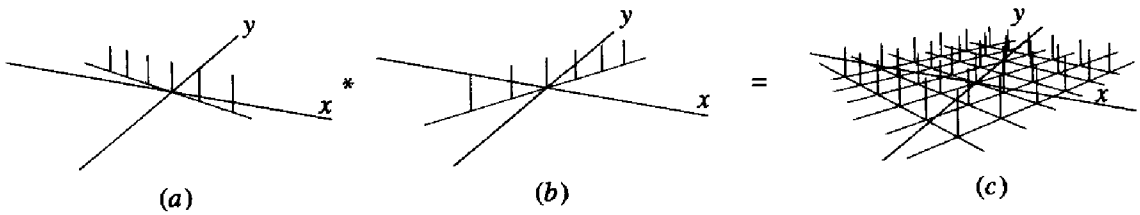


Fig. 7.6. The convolution of two lines of fence-posts to give a 'bed of nails'.

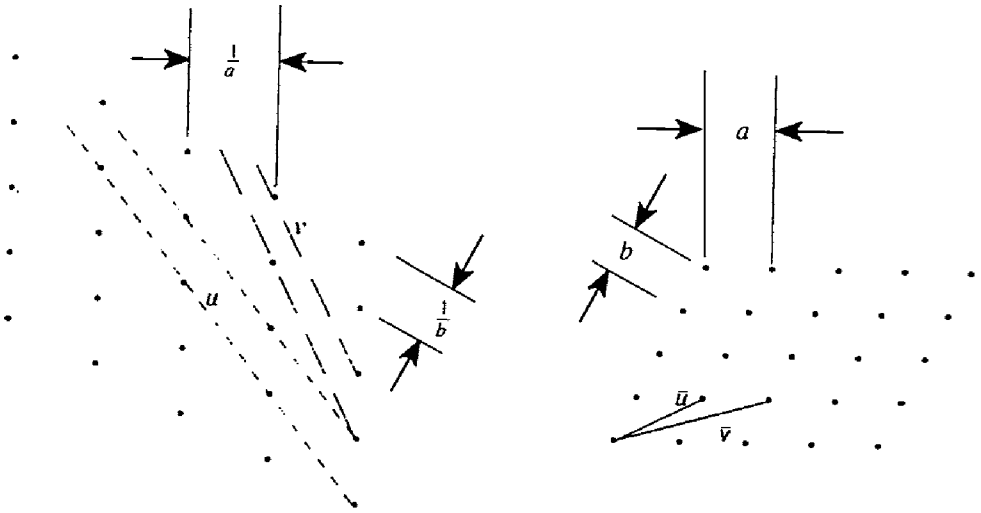


Fig. 7.7. Reciprocal lattices: the correspondence between a bed of nails and its Fourier pair. The pair are not unique: dashed lines show other possible Dirac walls, with different spacings, and the letters u and v show the corresponding directions of the Dirac fences which are their Fourier transforms. $\bar{u}$ and $\bar{v}$ are reciprocals of u and v: the narrow spacing of the walls implies a greater spacing between the fence-posts.

The interesting thing is that the route to w_1w_2 from $f_1 * f_2$ is not unique. The two-dimensional array w_1w_2 could have been composed from two different factors, both again parallel sets of walls, but transformed from different fences f_1' and f_2' with different spacings $a_1'$and a_1' and different angles θ_1' and θ_2'. But the convolution of this new pair will necessarily yield the same function $f_1 * f_2$ as before.

The correspondence between the two beds of nails is this: corresponding to *any* set of parallel lines that can be drawn through points in one plane there is a point[9] in the other. In Fig. 7.6, parallel lines separated by $1/a$ in one plane are matched by a point distance a in the other: another set separated by $1/c$ correspond to the point c, and so on. The whole thing is the two-dimensional analogue of the 'reciprocal lattice' idea in crystallography.

There is a familiar illustration: seats in a theatre or cinema are arranged regularly, often staggered so that people do not sit directly behind someone.

[9] Actually a pair of points – one either side of the origin.

Alignments of seat-backs can be seen in different directions, and these correspond to the lines that can be drawn through beds of nails.

7.6 Parallel plane delta-functions

In three dimensions the function $\delta(lx + my + nz)$ describes a function of unit amplitude which is zero except on the plane $lx + my + nz = 0$.

Its three-dimensional Fourier transform is

$$\phi(p, q, r) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \delta(lx + my + nz)e^{2\pi ipx}e^{2\pi iqy}e^{2\pi irz}dxdydz$$

and after the x-integration:

$$\begin{aligned}\phi(p, q, r) &= \frac{1}{l}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{2\pi i(p/l)(-my-nz)}e^{2\pi iqy}e^{2\pi irz}dydz \\ &= \frac{1}{l}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{2\pi iy(q-mp/l)}e^{2\pi iz(r-np/l)}dydz\end{aligned}$$

which is separable into

$$\frac{1}{l}\int_{-\infty}^{\infty} e^{2\pi iy(q-mp/l)}dy \int_{-\infty}^{\infty} e^{2\pi iz(r-np/l)}dz$$

so that

$$\phi(p, q, r) = l\delta(lq - mp)\delta(lr - np)$$

A δ-function which, when the coordinate systems are superimposed, is zero except on the line $p/l = q/m = r/n$, a line through the origin, perpendicular to the original plane in the x, y, z frame.

The extension is intuitive: a pair of parallel planes equally disposed about the origin and each at a distance a from the origin, will have as a Fourier transform a line along which the amplitude varies sinusoidally with period $1/a$. An infinite sequence of equally separated parallel planes will transform to a row of equally-spaced points along a line passing through the origin and perpendicular to the planes. It is the three-dimensional version of a Dirac comb but the function differs from zero at isolated *points*.

7.7 Point arrays

The ideas are even more apparent when transforms are done in three dimensions, when point-arrays are defined by products of three three-dimensional *III*-functions. For example, $III_a(l_1x + m_1y + n_1z)$ defines a set of parallel planes

on which the function is not zero. The planes have equations $l_1x + m_1y + n_1z - \lambda a = 0$ where l, m, n are direction cosines, λ is any integer and a is the perpendicular distance between two adjacent planes.

Two other sets of parallel planes can be defined similarly by $Ш_b(l_2x + m_2y + n_2z)$ and $Ш_c(l_3x + m_3y + n_3z)$ and the point array or lattice is defined by the product of these three functions.

The Fourier transform of one of these functions is simple:

$$\phi(p,q,r) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\sum_{\lambda=-\infty}^{\infty}\delta(lx + my + nz - \lambda a)e^{2\pi i(px+qy+rz)}dxdydz$$

do the x-integral first:

$$\phi(p,q,r) = \frac{1}{l}\sum_{\lambda=-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}e^{2\pi ip(\lambda a - nz - my)/l}e^{2\pi i(qy+rz)}dydz$$

where the λ-sum provides the $Ш$ function and the integral as before is merely the substitution of the value in the δ-function argument which makes it non-zero.

The integral is now separable:

$$\phi(p,q,r) = \frac{1}{l}\sum_{\lambda=-\infty}^{\infty}e^{2\pi ipa\lambda/l}\cdot\int_{-\infty}^{\infty}e^{-2\pi i(\frac{pn}{l}-r)z}dz\int_{-\infty}^{\infty}e^{-2\pi i(\frac{pm}{l}-q)y}dy$$

$$= \frac{1}{l}\sum_{\lambda=-\infty}^{\infty}e^{2\pi ipa\lambda/l}\cdot\frac{1}{n}\delta\left(\frac{p}{l}-\frac{r}{n}\right)\cdot\frac{1}{m}\delta\left(\frac{p}{l}-\frac{q}{m}\right)$$

The last two factors, the δ-functions, define two planes. The intersection of the planes defines a line. The sum over λ defines those points on the line where the lattice points exist in p, q, r-space[10].

Again if the (p, q, r)-space is superimposed on the (x, y, z)-space, we find that $\phi(p, q, r)$ is a set of equispaced points along a line perpendicular to the set of planes defined by $\delta(lx + my + nz - \lambda a)$ and that the spacing between the points is $1/a$.

7.8 Lattices

A complete three-dimensional lattice, described by the product of three planar $Ш$-functions of the type $Ш_a(lx + my + nz)$ has as its Fourier transform the

[10] By analogy with all the other entities to which the prefix 'Dirac' has been attached, the idea of a 'Dirac string' might be advanced to describe a spatial curve on which a three-dimensional function $f(x, y, z)$ is defined, on the understanding that it is zero everywhere except on that curve. For example $f(x, y, z)\delta(l_1x + m_1y + n_1z)\delta(l_2x + m_2y + n_2z)$ describes a function which is zero everywhere except on the line $x/(n_1m_2 - n_2m_1) = y/(l_1n_2 - l_2n_1) = z/(l_1n_2 - l_2n_1)$.

triple convolution of three lines of equispaced points. This gives a new lattice – the *reciprocal lattice*[11] in p, q, r-space, used in crystallography. Points on this reciprocal lattice define various planes in x, y, z-space, which contain two-dimensional arrays of lattice points. Lines from the origin to points on the reciprocal lattice define both the orientation and separation of the corresponding planes in x, y, z space.

This now clears up a fundamental problem in describing crystals. The three *III*-functions used to define the crystal lattice in x, y, z-space are not the only possible ones. Other sets of planes can be used – an infinite number of possibilities exists. The points in the reciprocal lattice define uniquely such sets of parallel planes. The lines ('vectors') from the origin to these points in p, q, r-space are normal to the lattice-planes in x, y, z-space and the length of each vector is inversely proportional to the separation of the planes in x, y, z-space. The coordinates of the lattice points in p, q, r-space, when multiplied by a factor to make them integer, are the *Miller Indices*, beloved of crystallographers, of the x, y, z-planes.

[11] Vide, e.g. H. M. Rosenberg *The Solid State* 3rd edn. Oxford University Press, 1988.

Chapter 8

The formal complex Fourier transform

In physics we are usually concerned with functions of real variables, often experimental curves, data strings, or shapes and patterns. Generally the function is asymmetric about the y-axis and so its Fourier transform is a complex function of a real variable; that is, for any value of p, a complex number is defined.

Any function obeying the Dirichlet conditions can be divided into a symmetric and an antisymmetric part. In Fig. 8.1, for example, and generally, $f_s(x) = \frac{1}{2}[f(x) + f(-x)]$ and $f_a(x) = \frac{1}{2}[f(x) - f(-x)]$. The symmetric part is synthesized only from cosines and the antisymmetric part only from sines. We write:

$$f(x) = f_s(x) + f_a(x); \qquad f_s(x) \rightleftharpoons \phi_s(p); \qquad f_a(x) \rightleftharpoons \phi_a(p)$$

where $\phi_s(p)$, being made of cosines, is real and symmetric and $\phi_a(p)$ is imaginary and, being made of sines, is antisymmetric.

We can also define:

(a) the phase-transform of $f(x)$ which is the function $\theta(p)$ where

$$\tan\theta(p) = \phi_a(p)/\phi_s(p)$$

(b) The power transform

$$P(p) = \mid \phi(p) \mid^2 = \phi_a(p)^2 + \phi_s(p)^2$$

(c) The modular transform:

$$M(p) = \mid \phi(p) \mid = \sqrt{\phi_a(p)^2 + \phi_s(p)^2}$$

All these have their practical uses although none of them has an unique inverse.

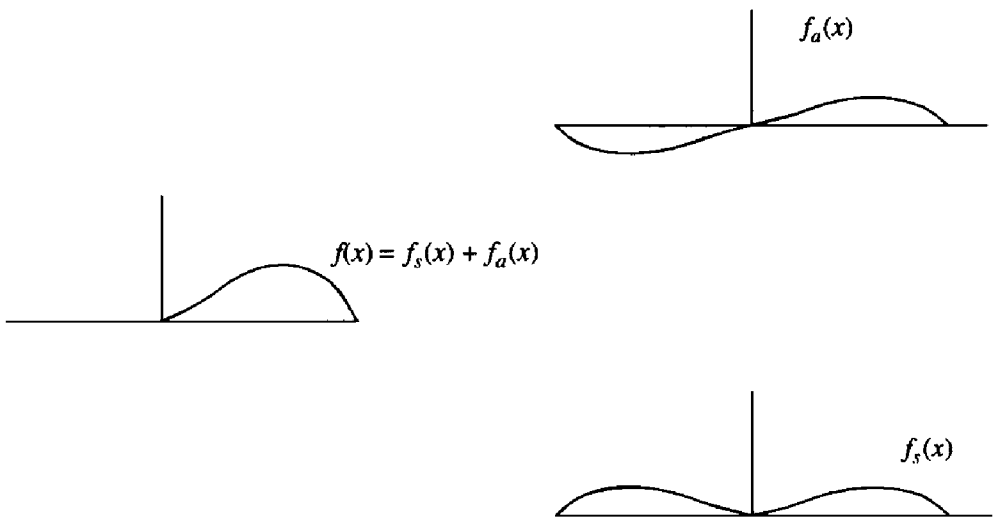


Fig. 8.1. Dividing a function into symmetric and antisymmetric parts.

A useful corollary of the convolution theorem is that if $C(x) = f_1(x) * f_2(x)$ and if $C(x) \rightleftharpoons \Gamma(p)$ then the power transforms of C, f_1 and f_2, given by $|\ \Gamma\ |^2$, $|\ \phi_1\ |^2$ and $|\ \phi_2\ |^2$ are related by:

$$|\ \Gamma\ |^2 = |\ \phi_1\ |^2 \cdot |\ \phi_2\ |^2$$

A simple example shows the use of phase transforms. Consider for instance, a displaced top-hat function (any function would do, in fact), of width a and displaced sideways a distance b.

The function is:

$$f(x) = \Pi_a(x) * \delta(x - b)$$

Its Fourier transform is

$$\begin{aligned}\phi(p) &= a\ \mathrm{sinc}(\pi ap).e^{2\pi ibp} \\ &= a\ \mathrm{sinc}(\pi ap)\,[\cos(2\pi bp) + i\ \sin(2\pi bp)]\end{aligned}$$

and its phase transform is

$$\theta(p) = \tan^{-1}(\sin 2\pi pb / \cos 2\pi pb)$$

so that $\theta(p) = 0$ when $p = 0$, $\theta(p) = 2\pi$ when $p = 1/b \ldots$

Phase transforms are useful when an experimentally measured function, which should have been symmetrical, has been displaced by an unknown

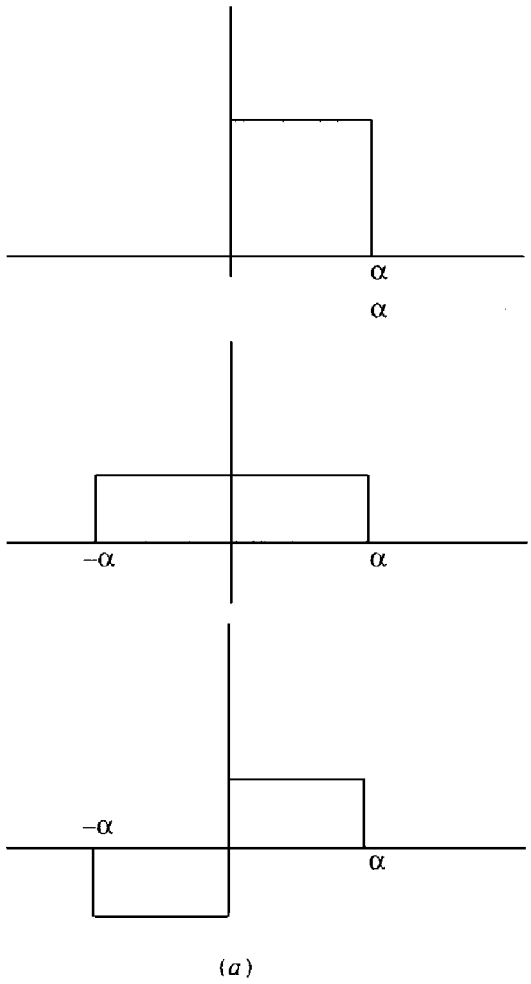
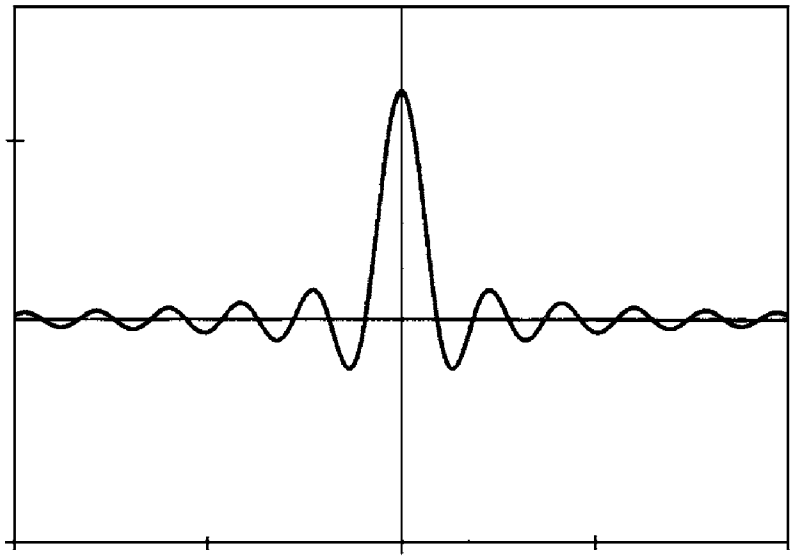


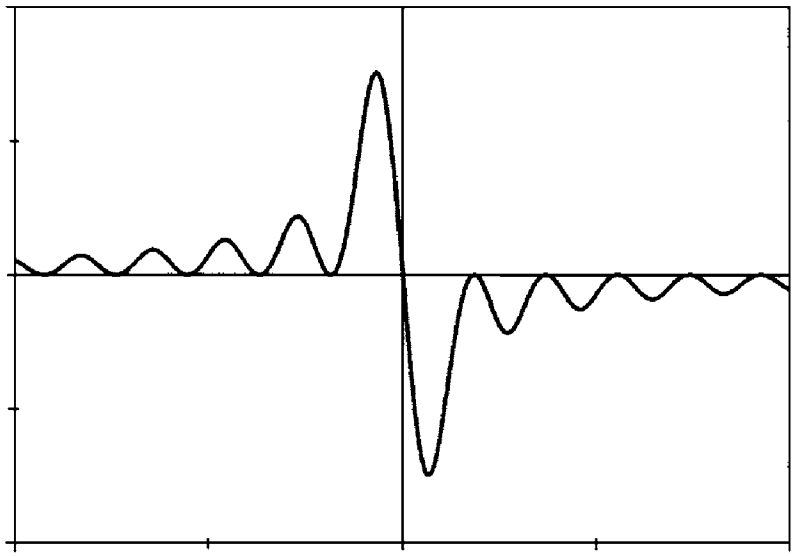
Fig. 8.2. A top-hat function displaced by its own width. (a) The dissection of the top-hat into symmetric and antisymmetric parts. (*continued overleaf*)

amount from its axis of symmetry – for example by sampling it in the wrong places. A quick calculation of a few points on the phase transform will find the displacement and allow any adjustments to be made or the true, symmetrical samples to be computed by interpolation. It also confirms (or not!) that the function really is symmetric, since only then is its phase transform a straight line.

It is worth including here something which will be useful later when considering computing Fourier transforms. Since it is easy to separate the real and imaginary parts of a complex function of x or p and then to divide these into their symmetric and antisymmetric parts, it is possible to combine two real functions of x into a complex function and then separate the combined complex Fourier transform into its constituent parts. This is a useful technique when computing digital Fourier transforms: one can do two transforms for the



(*b*)



(*c*)

Fig. 8.2 (*cont.*). (b) the cosine transform; (c) the sine transform.

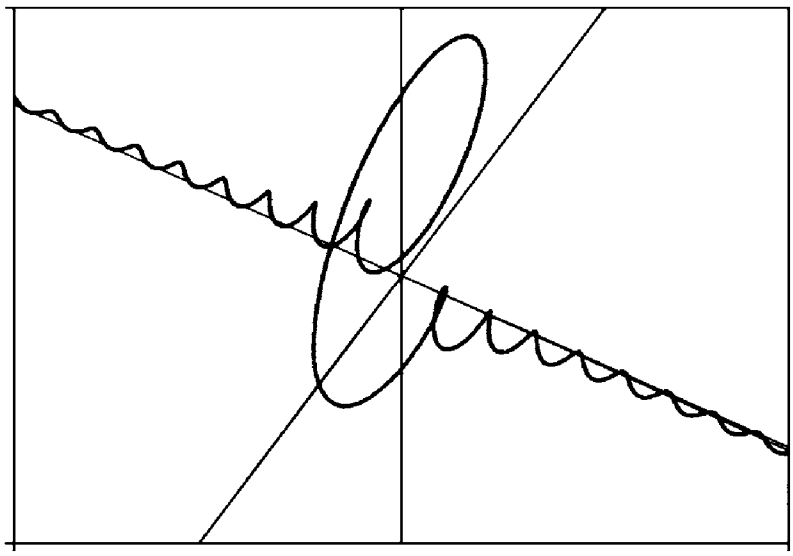
price of one. Written analytically, let the two functions be $f_1(x)$ and $f_2(x)$ and separate each into its symmetric and antisymmetric parts:

$$f_1(x) = f_{1s}(x) + f_{1a}(x); \quad f_2(x) = f_{2s}(x) + f_{2a}(x)$$

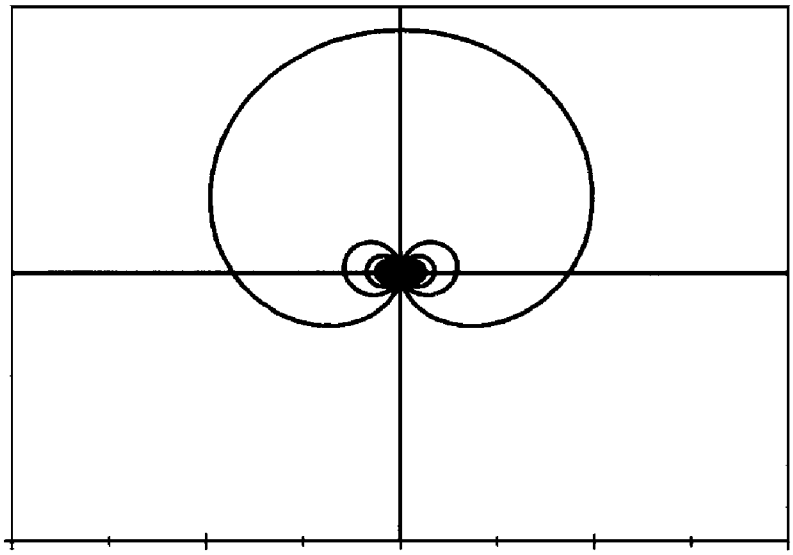
$$\text{let } F(x) = f_1(x) + if_2(x)$$

$$\text{and let } F(x) \rightleftharpoons \Phi(p). \text{ Then}$$

$$\Phi(p) = \int_{-\infty}^{\infty} [f(_1(x) + if_2(x)]e^{2\pi ipx} dx$$



(*d*)



(*e*)

Fig. 8.2 (*cont.*). (d) the transform in perspective; (e) the Nyquist diagram – the view looking along the ν-axis.

and remember that a symmetric function has only a cosine transform, etc.

$$\begin{aligned}\Phi(p) &= \int f_{1s}(x)\cos 2\pi px\,dx + i\int f_{1a}(x)\sin 2\pi px\,dx \\ &\quad + i\int f_{2s}(x)\cos 2\pi px\,dx - \int f_{2a}(x)\sin 2\pi px\,dx \\ &= \phi_{1s}(p) + i\phi_{1a}(p) + i\phi_{2s}(p) - \phi_{2a}(p)\end{aligned}$$

where the meaning of the suffixes is the same as in the f-functions.

Then:

$$\Phi(p) = [\phi_{1s}(p) - \phi_{2a}(p)] + i[\phi_{1a}(p) + \phi_{2s}(p)]$$

Both the real and imaginary parts of $\Phi(p)$ now have symmetrical and antisymmetrical components. When $\Phi(p)$ has been computed, it has a real part, $\Phi_r(p)$ and an imaginary part $\Phi_i(p)$.

The symmetric real part is $\frac{1}{2}[\Phi_r(p) + \Phi_r(-p)] = \phi_{1s}(p)$

and the antisymmetric part is $\frac{1}{2}[\Phi_r(p) - \Phi_r(-p)] = -\phi_{2a}(p)$.

Similarly, $\frac{1}{2}[\Phi_i(p) - \Phi_i(-p)] = \phi_{1a}(p)$

and $\frac{1}{2}[\Phi_i(p) + \Phi_i(-p)] = \phi_{2s}(p)$.

so that, finally:

$$f_1(x) \rightleftharpoons \frac{1}{2}[\Phi_r(p) + \Phi_r(-p)] + \left(\frac{i}{2}\right)[\Phi_i(p) - \Phi_i(-p)]$$

$$\rightleftharpoons \frac{1}{2}\phi_{1s}(p) + \left(\frac{i}{2}\right)\phi_{1a}(p)$$

and similarly,

$$f_2(x) \rightleftharpoons \frac{1}{2}\phi_{2s}(p) + \left(\frac{i}{2}\right)\phi_{2a}(p)$$

In other words, the Fourier transform of $f_1(x)$ is $\frac{1}{2} \times$ (the symmetrical part of the real component of $\Phi(p)$ plus $i \times$ the antisymmetrical part of the imaginary component of $\Phi(p)$), and the Fourier transform of $f_2(x)$ is $\frac{1}{2} \times$ (the symmetrical part of the imaginary component $+ i \times$ the antisymmetric part of the real component). The computer sorts these out without difficulty!

Notice that all the F's, Φ's, f's and ϕ's with suffixes are *real* quantities. This is because a computer deals ultimately in real numbers, although its program may include complex arithmetic. This level of complication is not commonly met when discussing analytic Fourier transforms. However computing algorithms compute the complex transform whether you like it or not, and the relations above can be used to do tricks in shortening computing time when you know that the data represent only real functions.

Diagrammatically, the process can be represented by:

$$f_{1s}(x) \leftarrow \cos \rightarrow \phi_{1s}(p)$$

$$f_{1a} \leftarrow i \sin \rightarrow i\phi_{1a}$$

$$if_{2s} \leftarrow \cos \rightarrow i\phi_{2s}$$

$$if_{2a} \leftarrow i \sin \rightarrow -\phi_{2a}$$

A function is said to be Hermitian if its real part is symmetric and its imaginary part is antisymmetric. So if $f_1(x)$ is symmetric and $f_2(x)$ is antisymmetric, then $\phi_{1a} \equiv 0$ and $\phi_{2s} \equiv 0$. Then

$$\mathbf{\Phi}(p) = \phi_{1s}(p) + \phi_{2a}(p)$$

and is real. Alternatively, the Fourier transform of a real but asymmetric function is Hermitian:

$$f_1(x) \rightleftharpoons \phi_{1s}(p) + i\phi_{1a}(p)$$

Chapter 9

Discrete and digital Fourier transforms

9.1 History

Fourier transformation is formally an analytical process which uses integral calculus. In experimental physics and engineering, however, the integrand may be a set of experimental data, and the integration is necessarily done artificially. Since a separate integration is needed to give each point of the transformed function, the process would become exceedingly tedious if it were to be attempted manually and many ingenious devices have been invented for performing Fourier transforms mechanically, electrically, acoustically and optically. These are all now part of history since the arrival of the digital computer and more particularly since the discovery – or invention – of the 'Fast Fourier Transform' algorithm or FFT as it is generally called. Using this algorithm, the data are put ('read') into a file (or 'array', depending on the computer jargon in use); the transform is carried out, and the array then contains the points of the transformed function. It can be achieved by a software program, or by a purpose-built integrated circuit. It can be done very quickly so that vibration-sensitive instruments with Fourier transformers attached can be used for tuning pianos and motor engines, for aircraft and submarine detection and so on. It must not be forgotten that the ear is Nature's own Fourier transformer[1], and, as used by an expert piano-tuner for example, is probably the equal of any electronic simulator in the 20–20 000 Hz range. The diffraction grating too, is a passive Fourier transformer device provided that it is used as a spectrograph taking full advantage of the simultaneity of outputs.

The history of the FFT is complicated and has been researched by Brigham[2] and, as with many discoveries and inventions, it arrived before the (computer) world was ready for it. Its digital apotheosis came with the publication of

[1] It detects the *power* transform, and is not sensitive to phase.
[2] E. Oran Brigham *The Fast Fourier Transform*. Prentice-Hall, 1974.

the 'Cooley–Tukey' algorithm[3] in 1965. Since then other methods have been virtually abandoned except for certain specialized cases and this chapter is a description of the principles underlying the FFT and how to use it in practice.

9.2 The discrete Fourier transform

There is a pair of formulae by which sets of numbers $[a_n]$ and $[A_m]$, each set having N elements, can be mutually transformed:

$$A(m) = \frac{1}{N}\sum_{0}^{N-1} a(n)e^{2\pi inm/N}; \qquad a(n) = \sum_{0}^{N-1} A(m)e^{-2\pi inm/N} \tag{9.1}$$

In appearance and indeed in function, these are very similar to the formulae of the analytic Fourier transform and are generally known as a 'discrete Fourier transform' (DFT). They can be associated with the true Fourier transform by the following argument.

Suppose, as usual, that $f(x)$ and $\phi(p)$ are a Fourier pair. If $f(x)$ is multiplied by a *Ш*-function of period a then the Fourier transform becomes:

$$\Phi(p) = \int_{-\infty}^{\infty} f(x)Ш_a(x)e^{2\pi ipx}dx = 1/a[\phi(p) * Ш_{1/a}(p)]$$

Now suppose that $f(x)$ is negligibly small for all x outside the limits $-a/2 \rightarrow (N-1/2)a$, so that there are N teeth in the Dirac comb, and $f(x)$ extends over a range $\leq Na$. We rewrite the integral and use the properties of δ-functions so that

$$\Phi(p) = \int_{-\infty}^{\infty}\sum_{n=-\infty}^{\infty} f(x)e^{2\pi ipx}\delta(x-na)\,dx$$

$$= \sum_{n=-\infty}^{\infty}\int_{-\infty}^{\infty} f(x)e^{2\pi ipx}\delta(x-na)\,dx$$

Because there are only N teeth in the comb, the sum is finite and the integral means substituting the argument of the δ-function as usual.

$$\Phi(p) = \sum_{n=0}^{N-1} f(na)e^{2\pi ipna}$$

$$= 1/a[\phi(p) * Ш_{\frac{1}{a}}(p)]$$

[3] J. W. Cooley and J. W. Tukey An algorithm for the machine calculation of complex Fourier series. *Math. Computation.* **19** 297–301, April 1965.

This in turn, is periodic in p with period $1/a$, and can be written:

$$\begin{aligned}\Phi(p) &= (1/a)\phi(p) * \mathit{III}_{\frac{1}{a}}(p) \\ &= (1/a)[\phi(p) + \phi(p + 1/a) + \phi(p - 1/a) \\ &\quad + \phi(p + 2/a) + \phi(p - 2/a) + \cdots)\end{aligned}$$

And in its first period $\Phi(p)$ is the same as the analytic function $(1/a)\phi(p)$.

Now consider n small intervals of p each of width $1/Na$. At the mth such interval the equation becomes:

$$\Phi(m/Na) = \sum_{n=0}^{N-1} f(na)e^{2\pi ina(m/Na)} = (1/a)\phi(m/Na)$$

or, more succinctly:

$$\sum_{n=0}^{N-1} f(n)e^{2\pi inm/N} = (1/a)\phi(m)$$

and this approximates to the analytic Fourier transform. The approximation is that in its first period the periodic $\Phi(p) = \phi(p)$. Theoretically it is not – there is bound to be some overlap since $\phi(p)$ is not zero – but practically it can be ignored[4].

The choice of the interval $-a/2 \rightarrow (N - 1/2)a$ for $f(x)$ is so as to have exactly N teeth in the Dirac comb without the embarrassment of having teeth at the very edge – where a top-hat function changes from 1 to 0, for example. in theory *any* interval of the same length would do.

9.3 The matrix form of the DFT

One way of looking at the formula for the DFT is to set it out as a matrix operation. The data set $[a(n)]$ can be written as a column matrix or 'vector' (in an N-dimensional space), to be multiplied by a square matrix containing all the exponentials and giving another column matrix with N components, $[A(m)]$ as its result.

$$\begin{bmatrix} A(0) \\ A(1) \\ A(2) \\ A(3) \\ \vdots \\ A(N-1) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & e^{2\pi i/N} & e^{4\pi i/N} & \cdots & e^{2(N-1)\pi i/N} \\ 1 & e^{4\pi i/N} & e^{8\pi i/N} & \cdots & e^{4(N-1)\pi i/N} \\ 1 & e^{6\pi i/N} & e^{12\pi i/N} & \cdots & e^{6(N-1)\pi i/N} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \cdots & \cdots & \cdots & e^{(N-1)^2 2\pi i/N} \end{bmatrix} \begin{bmatrix} a(0) \\ a(1) \\ a(2) \\ a(3) \\ \vdots \\ a(N-1) \end{bmatrix}$$

[4] It is not possible for a function and its Fourier pair both to be finite in extent – one at least must extend to $\pm\infty$ but the condition that both be small compared with the values in the region of interest is allowable.

The process of matrix multiplicaton requires n^2 multiplications for its completion. If large amounts of data are to be processed, this can become inordinate, even for a computer. Some people like to process columns of data with 10^6 numbers occasionally, but normally experimenters make do with 1024, although they often require the transform in a few microseconds.

The secret of the FFT is that it reduces the number of multiplications to be done from N^2 to about $2N\log_2(N)$. A data 'vector' 10^6 numbers long then requires 4.2×10^7 multiplications instead of 10^{12}, a gain in speed of approximately $\times 26\,200$. In this year of grace 2002, the computation time on a desktop computer is reduced from about 20 min to a few milliseconds.

The way it does this is, in essence, to factorize the matrix of exponentials, but there are easier ways of looking at the process. For example: suppose that the number N of components in the vector is the product of two numbers k and l. Instead of writing the subscript of each number in the vector to denote its position $(0 \ldots N-1)$ it can be given two subscripts s and t, and written $a(s,t)$, $= a(sk+t)$, where s takes values from 0 to $(l-1)$ and t runs from 0 to $(k-1)$. In this way all the numbers in the vector are labelled, but now with two suffixes instead of one. There is absolutely no point in doing this except for computational purposes: it is purely a piece of computer-mathematical manipulation, and would have struck mathematicians of pre-computer days as ludicrous. However we now write the digital transform as:

$$A(u,v) = \sum_{s=0}^{l-1} \sum_{t=0}^{k-1} a(s,t) e^{2\pi i(sk+t)(ul+v)/kl}$$

where the suffix m in the transformed vector has similarly been dissected into u and v, with $m = ul + v$. u runs from 0 to $(k-1)$ and v from 0 to $(l-1)$.

The exponent is now multiplied out and gives

$$A(u,v) = \sum_{s=0}^{l-1} \sum_{t=0}^{k-1} a(s,t) e^{2\pi isu} e^{2\pi isv/l} e^{2\pi itu/k} e^{2\pi ivt/kl}$$

The first exponential factor is unity and is discarded. The double sum can be rewritten now as:

$$A(u,v) = \sum_{t=0}^{l-1} e^{2\pi itu/k} e^{2\pi ivt/kl} \sum_{s=0}^{k-1} a(s,t) e^{2\pi isv/l}$$

which is legitimate since only the last exponent contains a factor s.

This sum over k terms gives a new set of numbers $[g(v,t)]$ and we write:

$$A(u,v) = \sum_{t=0}^{l-1} [g(v,t) e^{2\pi ivt/kl}] e^{2\pi itu/k}$$

The array $[g(v, t)]$ is multiplied by $[e^{2\pi ivt/kl}]$ to give an array $[g'(v, t)]$ and finally the sum:

$$g''(v, u) = \sum_{t=0}^{l-1} g'(v, t)e^{2\pi itu/k}$$

and $g''(v, u) = A(u, v)$. (The reversing of the order of v and u is important.)

The transform has been split into two stages: there are k transforms, each of length l, followed by N multiplications by the exponential factors $e^{2\pi ivt/kl}$ (the 'twiddle-factors'), followed by l transforms, each of length k: a total of $kl^2 + lk^2 = N(k + l)$ multiplications, apart from the relatively small number, N of multiplications [by $e^{2\pi ivt/kl}$] in the middle.

The lesson is that, provided N can be factorized, the vector $[a(n)]$ can be turned into a rectangular $k \times l$ matrix and treated column by column as a set of shorter transforms. For example, if there were a factor 2, the *even* numbered a's could be put into one vector of length $N/2$ and the odd-numbered a's into another. Then each is subjected to a Fourier transform of half the length to give two more vectors, and these, after multiplying by the 'twiddle-factors' as above, can be recombined into a vector of length N.

The same process can be repeated provided that $N/2$ can be factored; and if the factors are always 2, it continues until only 2×2 matrices are left, with trivially easy Fourier transforms (and a multiplicity of 'twiddle-factors'!) The interesting thing is that each number in the transformed vector has its address in bit-reversed order. In the example given earlier the final outcome was $g''(v, u)$, so that the two indices have to be reversed – the number $g''(v, u)$ is in the wrong place in the array. This effect is multiplied until, in the 2^N transform, the transformed data appear in the wrong addresses, the true address being the bit-reversed order of the apparent address.

The FFT is thus usually done with N a power of two. Not only is it very efficient in terms of computing time, but is ideally suited to the binary arithmetic of digital computers. The details of the way programs are written are given by Brigham[5] and a BASIC listing of an FFT routine is given at the end of this chapter. There are many such routines, the results of many hours of research, and sometimes very efficient. This one is not particularly fast but will suffice for practice and is certainly suitable for student laboratory work.

The data file for this program must be 2048 words long (1024 complex numbers, alternately real and imaginary parts), and if only real data are to be transformed, they should go in the even-numbered elements of the array, from 0 to 2046. Some caution is needed: zero frequency is at array element 0. If

[5] E. Oran Brigham, *The Fast Fourier Transform.* Prentice Hall, 1974.

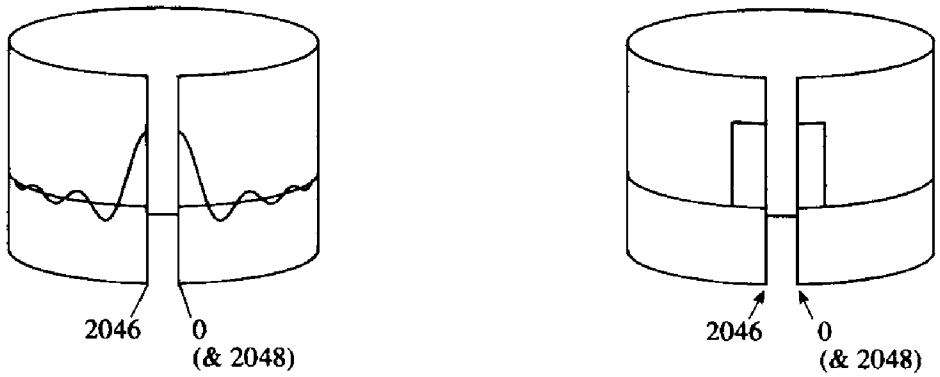


Fig. 9.1. The implementation of the FFT using a sinc-function as an example. The two cylinders unwrapped, represent the input and output data arrays. Do not expect zero to be in the middle as in the analytic case of a Fourier transform. If the input data are symmetrical about the centre, these two halves must be exchanged (en-bloc, not mirror-imaged) before and after doing the FFT.

you want to Fourier transform a sinc-function for example, the positive part of the function should go at the beginning of the array and the negative part at the end. The diagram illustrates the point: the output will similarly contain the zero-frequency value in element 0, so that the top-hat appears to be split between the beginning and the end.

Alternatively, you can arrange to have zero-frequency at point 1024 in the array, in which case the input and output arrays must both be transposed, by having the first and second halves interchanged (but not flipped over) before and after the FFT is done.

Attention to these details saves a lot of confusion! It helps to think of the array as wrapped around a cylinder, with the beginning of the array at zero-frequency and the end at point (-1) instead of $(+1023)$.

9.3.1 Two-dimensional FFT's

Two-dimensional transforms can be done using the same routines. The data are in a rectangular array of 'pixels' which form the picture which is to be transformed. Each row should first have its right and left halves transposed. Then each column must have the top and bottom halves transposed, so that what was perhaps a circle in the middle of the picture becomes four quadrants, one in each corner. Then each row is given the FFT treatment. Then each column in the resulting array gets the same. Then the rows and finally the columns are transposed again to give the complete FFT. At this stage periodic features such as a TV raster for example, will appear as Dirac nails (provided that the original picture has been sampled often enough) and can be suppressed by altering the

contents of the pixels where they appear. Then the whole procedure is reversed to give the whole 'clean' picture.

Apodising functions can similarly be applied to remove false information, to smooth edges and to improve the picture cosmetically.

Obviously far more elaborate techniques than this have been developed, but this is the basis of the whole process.

The output can be used in a straightforward way to give the power, phase or modular transforms, and the data can be presented graphically with simple routines which need no description here.

9.4 The BASIC FFT routine

The listing below is of a simple BASIC Routine for the FFT of 1024 complex numbers[6].

This is a routine which can be incorporated into a program which you can write for yourself.

The data to be transformed are put in an array D(I) declared at the beginning of the programme as 'DIM D(2047)', and the reals go in the even-numbered places, beginning at 0, and the imaginaries in the odd-numbered places. The transformed data are found similarly in the same array. The variable G on line 4 should be set to 1 for a direct transform and to -1 for an inverse transform. Numbers to be entered into the D(I) array should be in ASCII format. The programme should fill the D(I) array with data; call the FFT as a routine with a 'GOSUB 100' statement (The 'RETURN' is the last statement, on line 10) and this can be followed by instructions for displaying the data.

It is well worth while incorporating a routine for transposing the two halves of the D(I) array before and after doing the transform, as an aid to understanding what is happening.

```
100   N = 2048                    REM for 1024 complex points
      PRINT"BEGIN FFT"              transform.
      J=1
      G=1                         REM For direct transform. G = -1
      FOR I = 1 TO N STEP 2         for inverse
      IF (I-J) < 0 GOTO 1
      IF I=J GOTO 2
      IF (I-J)>0 GOTO 2
```

[6] But N can be changed by changing the first line of the program.

```
1   T=D(J-1)
    S= D(J)
    D(J-1)=D(I-1)
    D(J)=D(I)
    D(I-1)=T
    D(I)=S
2   M=N/2
3   IF (J-M)<0 GOTO 5
    IF J=M GOTO 5
    IF (J-M)>0 GOTO 4
4   J=J-M
    M=M/2
    IF (M-2)<0 GOTO 5
    IF M=2 GOTO 3
    IF(M-2)>0 GOTO 3
5   J=J+M
    NEXT I
    X=2
    IF (X-N)<0 GOTO 7
6   IF X=N GOTO 8
    IF (X-N)>0 GOTO 8
7   F=2*X
    H = - 6.28319/(G*X)
    R = SIN(H/2)
    W= -2*R*R
    V = SIN(H)
    P = 1
    Q = 0
    FOR M = 1 TO X STEP 2
    FOR I = M TO N STEP F
    J=I+X
    T=P*D(J-1)-Q*D(J)
    S=P*D(J)+Q*D(J-1)
    D(J-1)=D(I-1)-T
    D(J)=D(I)-S
    D(I-1)=D(I-1)+T
    D(I)=D(I)+S
    NEXT I
    T=P
    P=P*W-Q*V+P
```

```
      Q=Q*W+T*V+Q
      NEXT M
      X=F
      GOTO 6
 8    CLS
      FOR I = 0 TO N-1
      D(I)=D(I)/(SQR(N/2))
      NEXT I
      PRINT "FFT DONE"
10    RETURN
```

And here is a short program to generate a file with .DAT extension which will contain a top-hat function of any width you choose. The data are generated in ASCII and can be used directly with the FFT program above.

```
REM Programme to generate a 'Top-hat' function.
INPUT 'input desired file name', A$
INPUT 'Top-hat Half-width ?', N
PI=3.141 592 654
DIM B(2047)
FOR I = 1024-N TO 1024+N STEP 2
B(I) = 1/(2 * N)
NEXT I
C$=".DAT"
C$=A$+C$
PRINT
OPEN C$ FOR OUTPUT AS #1
FOR I=0 TO 2047
PRINT #1,B(I)
NEXT I
CLOSE #1
```

The simple file-generating arithmetic in lines 6–8 can obviously be replaced by something else, and this sort of 'experiment' is of great help in understanding the FFT process.

The file thus generated can be read into the FFT program with:

```
REM Subroutine FILELOAD
REM To open a file and load contents into D(I)
GOSUB 24
```

```
    (insert the next stage of the program, e.g."gosub 100",here)
24  CLS:LOCATE 10,26,0
    PRINT"NAME OF DATA FILE ?"
    LOCATE 14,26,0
    INPUT A$
    ON ERROR GOTO 35
    OPEN "I",#1,A$
    FOR I = 0 TO 2047
    ON ERROR GOTO 35
    INPUT#1,D(I)
    NEXT I
    CLOSE
35  RETURN
```

Appendix

A.1.1 The Heaviside step-function

This has the properties that:

$$H(x) = 0, x < 0 \text{ and } H(x) = 1, x > 0$$

and it is convenient to assume that $H(0) = \frac{1}{2}$.

Its Fourier transform is obtained easily. It can be regarded as the integral of the δ-function and the integral theorem (q.v.) can be used to derive it.

$$\partial H(x)/\partial x = \delta(x)$$

$$\text{Therefore } \partial H(x)/\partial x \rightleftharpoons 1$$

$$\text{hence} \qquad H(x) \rightleftharpoons \frac{1}{2\pi i p}$$

It can be manipulated in the usual way:

$$H(x - a/2) = H(x) * (\delta(x - a/2)$$

and

$$H(x - a/2) - H(x + a/2) = H(x) * [\delta(x + a/2) - \delta(x - a/2)]$$

Fourier transforming the right-hand side gives

$$\begin{aligned} H(x - a/2) - H(x + a/2) &\rightleftharpoons \frac{-1}{2\pi i p}[e^{-i\pi pa} - e^{i\pi pa}] \\ &= \frac{1}{\pi p} \sin \pi pa \\ &= a\text{sinc}\pi pa \end{aligned}$$

and the left-hand side is clearly a top-hat function of unit height and width a.

The step-function is chiefly used, much as a top-hat function is used, to isolate parts of another function. For example a sinusoidal wave switched on at time $t = 0$ can be written as $f(t) = \cos 2\pi\nu t.H(t)$.

A.1.2 Parseval's theorem and Rayleigh's theorem

Parseval's theorem states that:

$$\int_{-\infty}^{\infty} f(x)g^*(x)\,dx = \int_{-\infty}^{\infty} F(p)G^*(p)\,dp$$

This proof relies on the fact that if

$$g(x) = \int_{-\infty}^{\infty} G(p)e^{2\pi ipx}dp$$

then

$$g^*(x) = \int_{-\infty}^{\infty} G^*(p)e^{-2\pi ipx}dp$$

(simply by taking complex conjugates of everything).

Then it follows that:

$$G^*(p) = \int_{-\infty}^{\infty} g^*(x)e^{2\pi ipx}dx$$

The argument of the integral on the left-hand side of the theorem can now be written as:

$$f(x)g^*(x) = \int_{-\infty}^{\infty} F(q)e^{2\pi iqx}dq \int_{-\infty}^{\infty} G^*(p)e^{-2\pi ipx}dp$$

We integrate both sides with respect to x. If we choose the order of integration carefully, we find:

$$\int_{-\infty}^{\infty} f(x)g^*(x)\,dx = \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{\infty} F(q) \left[\int_{-\infty}^{\infty} G^*(p)e^{-2\pi ipx}dp \right] e^{2\pi iqx}dq \right\} dx$$

and changing the order of integration:

$$= \int_{-\infty}^{\infty} \left\{ F(q) \int_{-\infty}^{\infty} g^*(x)e^{2\pi iqx}dx \right\} dq$$

$$= \int_{-\infty}^{\infty} F(q)G^*(q)dq$$

The theorem is often seen in a simplified form, with $g(x) = f(x)$ and $G(p) = F(p)$. Then it is written:

$$\int_{-\infty}^{\infty} |f(x)|^2 dx = \int_{-\infty}^{\infty} |F(p)|^2 dp$$

This is **Rayleigh's theorem**.

Another version of Parseval's theorem involves the coefficients of a Fourier *series*. In words, it states that the average value of the square of $F(t)$ over one period is the sum of the squares of all the coefficients of the series.

The proof, using the half-range series, is simple:

$$F(t) = \frac{A_0}{2} + \sum_0^\infty A_n \cos\frac{2\pi nt}{T} + B_n \sin\frac{2\pi nt}{T}$$

and since all cross-products vanish on integration and

$$\int_0^T \cos^2 2\pi ntdt = \int_0^T \sin^2 2\pi ntdt = \frac{1}{2}$$

$$\int_0^T [F(t)]^2 dt = T\left[\frac{A_0^2}{4} + \sum_1^\infty \frac{\left(A_n^2 + B_n^2\right)}{2}\right]$$

A.1.3 Useful formulae from Bessel function theory

A.1.3.1 The Jacobi expansion

$$e^{ix\cos y} = J_0(x) + 2\sum_{n=1}^\infty i^n J_n(x)\cos ny$$

$$e^{ix\sin y} = \sum_{z=-\infty}^\infty J_z(x)e^{izy}$$

A.1.3.2 The integral expansion

$$J_0(2\pi\rho r) = \frac{1}{2\pi}\int_0^{2\pi} e^{2\pi i\rho r\cos\theta}d\theta$$

which is a particular case of the general formula:

$$J_n(x) = \frac{i^{-n}}{2\pi}\int_0^{2\pi} e^{in\theta}e^{ix\cos\theta}d\theta$$

$$\frac{d}{dx}\left(x^{n+1}J_{n+1}(x)\right) = x^{n+1}J_n(x)$$

A.1.3.3 The Hankel Transform

This is similar to a Fourier transform, but with polar coordinates, r, θ. The Bessel functions form a set with orthogonality properties similar to those of the trigonometrical functions and there are similar inversion formulae.

These are:

$$F(x) = \int_0^\infty pf(p)J_n(px)\,dp$$

$$f(p) = \int_0^\infty xF(x)J_n(px)\,dx$$

where J_n is a Bessel function of any order.

Bessel functions are analogous in many ways to the trigonometric functions sin and cos. In the same way as sin and cos are the solutions of the SHM equation $\frac{d^2y}{dx^2} + k^2y = 0$, they are the solutions of *Bessel's equation*, which is:

$$x^2\frac{d^2y}{dx^2} + x\frac{dy}{dx} + (x^2 - n^2)y = 0$$

In its full glory, n need not be integer and neither x nor n need be real. The functions are tabulated in various books[1] for real x and for integer and half-integer n, and can be calculated numerically, as are sines and cosines, by computer.

In its simpler form, as shown, it occurs with θ as variable when Laplace's equation is solved in cylindrical polar coordinates and variables are separated to give functions $R(r)\Theta(\theta)\Phi(\phi)$, and this is why it proves useful in Fourier transforms with circular symmetry.

A.1.4 Conversion of Fourier series coefficients to complex exponential form

We use De Moivre's theorem to do the conversion. Write $2\pi\nu_0 t$ as θ. Then, expressed as a half-range series, $F(t)$ becomes:

$$F(t) = A_0/2 + \sum_{m=1}^{\infty} A_m \cos m\theta + B_m \sin m\theta$$

This can also be written as a full-range series:

$$F(t) = \sum_{m=-\infty}^{\infty} a_m \cos m\theta + b_m \sin m\theta$$

where $A_m = a_m + a_{-m}$ and $B_m = b_m - b_{-m}$

Then by De Moivre's theorem the full-range series becomes:

$$F(t) = \sum_{m=-\infty}^{\infty} \frac{a_m}{2}(e^{im\theta} + e^{-im\theta}) + \frac{b_m}{2i}(e^{im\theta} - e^{-im\theta})$$

$$= \sum_{m=-\infty}^{\infty} \frac{a_m - ib_m}{2}e^{im\theta} + \sum_{m=-\infty}^{\infty} \frac{a_m + ib_m}{2}e^{-im\theta}$$

[1] For example, in Jahnke & Emde (see bibliography).

The two sums are independent and m is a dummy suffix, which means that it can be replaced by any other suffix not already in use. Here, we replace $m = -m$ in the second sum. Then:

$$\begin{aligned} F(t) &= \sum_{m=-\infty}^{\infty} \frac{a_m - i b_m}{2} e^{im\theta} + \sum_{m=-\infty}^{\infty} \frac{a_{-m} + i b_{-m}}{2} e^{im\theta} \\ &= \sum_{m=-\infty}^{\infty} e^{im\theta} \left\{ \frac{A_m - i B_m}{2} \right\} \\ &= \sum_{m=-\infty}^{\infty} e^{im\theta} C_m \end{aligned}$$

and $C_{-m} = C_m^*$.

Bibliography

The most popular books on the practical applications of Fourier theory are undoubtedly those of Champeney and Bracewell and they cover the present ground more thoroughly and in much more detail than here. E. Oran Brighan, on the Fast Fourier Transform, is the classic work on the subjects dealt with in Chapter 9.

Of the more theoretical works, the 'bible' is Titchmarsh, but a more readable (and entertaining) work is Körner's. Whittaker's (not to be confused with the more prolific E. T. Whittaker) book is a specialized work on interpolation, but that is a subject which is getting more and more important, especially in computer graphics.

Many writers on Quantum Mechanics, Atomic Physics and Electronic Engineering like to include an early chapter on Fourier theory. One or two (who shall be nameless) get it wrong! They confuse ω with ν or leave out a 2π when there should be one, or something like that. The specialist books, such as those below, are much to be preferred.

Abramowitz, M. and Stegun, I. A. *Handbook of Mathematical Functions*. Dover, New York. 1965
A more up-to-date version of Jahnke & Emde, below.

Bracewell, R. N. *The Fourier Transform and its Applications*. McGraw-Hill, New York. 1965
This is one of the two most popular books on the subject. Similar in scope to this book, but more thorough and comprehensive.

Brigham, E. O. *The Fast Fourier Transform*. Prentice Hall, New York. 1974
The standard work on digital Fourier transforms and their implementation by various kinds of FFT programs

Champeney, D. C. *Fourier Transforms and Their Physical Applications*. Academic Press, London and New York. 1973
Like Bracewell, one of the two most popular books on practical Fourier tranforming. Covers similar ground, but with some differences.

Champeney, D. C. *A Handbook of Fourier Theorems*. Cambridge University Press. 1987

Herman, Gabor T. *Image Reconstruction From Projections*. Academic Press, London and New York. 1980
Includes details of Fourier methods (among others) for computerized tomography, including theory and applications.

Jahnke, E and Emde, F. *Tables of Functions with Formulae and Curves*. Dover, New York. 1943
The classic work on the functions of mathematical physics, with diagrams, charts and tables, of Bessel functions, Legendre polynomials, spherical harmonics etc.

Körner, T. W. *Fourier Analysis*. Cambridge University Press. 1988
One of the more thorough and entertaining works on analytic Fourier theory, but plenty of physical applications: expensive, but firmly recommended for serious students.

Titchmarsh, E. C. *An Introduction to the theory of Fourier Integrals*. Clarendon Press, Oxford. 1962
The theorists' standard work on Fourier theory. Unnecessarily difficult for ordinary mortals, but needs consulting occasionally.

Watson, G. N. *A Treatise of the Theory of Bessel Functions*, Cambridge University Press. 1962
Another great theoretical classic: chiefly for consultation by people who have equations they can't solve, and which seem likely to involve Bessel functions.

Whittaker, J. M. *Interpolary Function Theory*, Cambridge University Press. 1935
A slim volume dealing with (among other things) the sampling theorem and problems of interpolating points between samples of band-limited curves.

Index