PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Phase Integral Methods for Studying the Effect of the Ionosphere on Radio Propagation

K. G. Budden

Phil. Trans. R. Soc. Lond. A 1975 **280**, 111-130
doi: 10.1098/rsta.1975.0095

Phil. Trans. R. Soc. Lond. A. **280**, 111–130 (1975) [111]
Printed in Great Britain

Phase integral methods for studying the effect of the ionosphere on radio propagation

By K. G. Budden, F.R.S.
Cavendish Laboratory, University of Cambridge

The phase integral method is a form of ray theory, extended to use complex values of the space coordinates. Its application to radio propagation studies was pioneered by T. L. Eckersley who showed how to use it for calculating (*a*) the reflexion coefficient of the ionosphere, (*b*) the propagation constant for radio waves guided by the Earth's surface and by the ionosphere or troposphere, and (*c*) the coefficient for coupling of an ordinary and an extraordinary wave in the ionosphere. The method involves the evaluation of integrals along suitably chosen contours in complex space. It is approximate but often capable of high accuracy and often quicker to use than more exact methods. Its justification is based on the physical principles of analytic continuation and of uniform approximation.

For reflexion and coupling problems in a horizontally stratified ionosphere, the contours used for the phase integrals are determined by those real or complex heights called 'reflexion' or 'coupling' points, where two roots of the Booker quartic equation are equal. The study of the behaviour of the governing equations near these points shows when failure of the phase integral method may be expected.

1. Introduction

When studying the theory of the propagation of radio waves in the ionosphere, the methods of geometrical optics or 'ray theory' can often be used if the frequency is great enough or equivalently if the medium varies slowly enough in space. The total electromagnetic field can then be expressed as the sum of a number of contributions called progressive waves or 'W.K.B. solutions' (§ 4), that are propagated independently. But no matter how slowly the medium varies there are always some regions where the independence breaks down, and we say that one (or more) of the progressive waves is there being converted into one or more of the others by a process called 'reflexion' or 'coupling'. In this review the term 'coupling' will be used to refer to both processes. In a very slowly varying medium the coupling regions occupy only a small fraction of space whereas in a rapidly varying medium they may occupy nearly the whole of it. Ray theory is often considered to be inadequate to deal with the coupling process, and the governing differential equations are therefore formulated, and solved with suitable boundary conditions. This technique is widely used for ionospheric radio wave problems and is an example of the method of 'full wave solutions'.

On closer study, however, it is found that ray theory can be used in conditions where, at first, it might be expected to fail. By suitably combining it with the principle of 'analytic continuation' (§ 2 (*b*)) it can often be used to deal with the coupling processes. There are many ionospheric radio propagation problems, sometimes with quite small frequencies in the v.l.f. range, which can be dealt with entirely by ray theory. The phase integral method is the technique of extending ray theory in this way to the limits of its usefulness. Thus it is just ray theory and not, as is sometimes supposed, a method of 'full wave solution'. The phrase 'ray theory' is used here because it

appeared in the title of Eckersley's first paper on the subject, mentioned below. But one ray cannot exist in isolation. It must be one of a pencil of rays that are trajectories of the wave fronts. The essential idea of ray theory is that it studies the properties of progressively travelling wave fronts.

Many authors have extended the phase integral method to use successively higher orders of approximation. Some of these treatments use extremely complicated analysis and it is difficult to link them with any kind of physical interpretation. The subject then becomes merely a mathematical exercise. If solutions are required that are more accurate than are given by the basic form of the phase integral method described here, then it is better to use full wave solutions, involving numerical integration of the governing differential equations. These are widely used, easy to use on modern computers, and known to work well.

The first to use phase integral methods for radio propagation studies was T. L. Eckersley but his papers are often so obscure that many readers have suspected that there may be errors. There are indeed many typographical and minor algebraic errors, but there is no doubt that he correctly understood the underlying ideas. His first main paper on the subject (Eckersley 1931) is entitled 'On the connection between the ray theory of electric waves and dynamics'. In the preceding years the analogy between the motion of a particle in a potential field, and the motion of a wave packet in a medium whose refractive index varies in space, had been used to develop the science of wave mechanics. It was postulated that a particle is represented by a progressive wave whose phase must be a single valued function of the space coordinates. Thus when it goes round a closed orbit, the total change of phase or 'phase integral' must be an integer times 2π. This is the origin of the title 'phase integral method'. Ray theory is closely analogous to classical mechanics with the addition of the concept of phase, leading to some of the quantum conditions. When ray theory fails, 'full wave solutions' of Schrödinger's wave equation are used, and lead to a modification of the quantum conditions. Thus by using results from the theory of wave propagation the science of wave mechanics developed. But the converse is not true. In using the phase integral method for the study of the propagation of radio waves we are not drawing at all on any results from dynamics. We are merely using the results of ray theory, as indicated in the title of Eckersley's paper.

Eckersley's first group of papers (1931, 1932*a*, *b*, *c*) on the subject was concerned with guided waves. The simplest kind of waveguide has two plane reflecting boundaries with free space between them. A single waveguide mode can be considered as two plane progressive waves in this space, with their wave normals making angles θ, $\pi-\theta$ with the normal to the boundary plane. One wave is converted to the other by reflexion at one plane and then converted back to the first wave by a second reflexion at the other plane (Brillouin 1936; Chu & Barrow 1938). For a self-consistent mode, the twice reflected wave must be identical with the original wave. This requires that the total change of phase in the double traverse of the guide width and at the two reflexions must be an integer multiple of 2π. It leads to the well known 'mode condition', and is a very simple example of the use of the phase integral. It is analogous to the phase integral condition for the wave associated with a particle in a closed orbit.

In this example the twice reflected wave must be identical with the original wave in both amplitude and phase. The two boundary planes may have reflexion coefficients $R_1(\theta)$, $R_2(\theta)$ with moduli less than unity for real angles of incidence. The change of amplitude can be interpreted by saying that the changes of phase at the two reflexions have imaginary parts $i \ln |R_1|$, $i \ln |R_2|$ as well as real parts $-\arg R_1$, $-\arg R_2$. In other words the phase of a wave is allowed to take complex

values. A change of its imaginary part then gives a change of the wave's amplitude. The decrease of amplitude at the reflexions must be compensated, which is done by allowing the angles θ, $\pi-\theta$ to take complex values, and the reflexion coefficients must be evaluated for these complex angles. Then each component wave is an inhomogeneous plane wave. The imaginary part of its phase changes as the guide is crossed and this gives the required increase of amplitude. The total change of complex phase in the double traverse and two reflexions must be real and an integer multiple of 2π. This gives the phase integral condition or mode condition

$$\mathrm{i}\ln|R_1|-\arg R_1+\mathrm{i}\ln|R_2|-\arg R_2+2kw\cos\theta=2\pi r, \tag{1}$$

where r is an integer and w is the width of the guide. This is usually written

$$R_1R_2\exp(-2\mathrm{i}kw\cos\theta)=\exp(-2\pi\mathrm{i}r)=1. \tag{2}$$

This simple example is exceptional because the reflexions at the sharp boundary planes cannot be dealt with by ray theory. The reflexion coefficients are derived by imposing boundary conditions on the components of the electromagnetic field, which is a method of 'full wave solution'. But in some other important problems considered by Eckersley, the two reflexions occurred in continuous, slowly varying media, where it was possible to solve the whole problem by ray theory (§ 5). In many guided wave problems for radio propagation one of the reflectors is a sharp boundary – the surface of the Earth, and its reflexion coefficient must be known as a function of angle of incidence. But it often happens that the other reflexion occurs in a slowly varying stratified medium, the ionosphere or the troposphere, and can be dealt with by ray theory. Then it is evaluated by the phase integral method, and the result is combined with the reflexion coefficient of the Earth's surface to give the mode condition.

In this way the phase integral method developed into a method of calculating the reflexion coefficient of a continuously varying medium, for a *given* θ, even when no mode condition or phase integral condition, like (1), is imposed. This idea is implied but not stressed in Eckersley's earlier papers (1931, 1932*a*, *b*, *c*). By the time he wrote his last paper on the subject, however (Eckersley 1950), he had used and extended the idea by showing how it could be used to find the coefficient for coupling of an ordinary to an extraordinary wave in the ionosphere (§ 8).

2. The meaning of 'phase integral methods' in radio propagation

The phase integral method uses two important physical principles namely (i) ray theory, or geometrical optics, and (ii) the principle of analytic continuation. These must now be discussed.

This review considers only media which do not change with time, and with harmonic waves; all components of the electromagnetic field are assumed to vary with time only through a factor $\exp(\mathrm{i}\omega t)$. The standard notation of magnetoionic theory is used including the symbols

X $Ne^2/\epsilon_0 m\omega^2$ where N is the electron concentration.

$\boldsymbol{Y}$ vector $e\boldsymbol{B}/m$ of magnitude Y antiparallel to the Earth's magnetic field $\boldsymbol{B}$, because the electronic charge e is negative.

Z ν/ω where ν is the effective collision frequency for electrons.

(*a*) *Ray theory*

In a medium which varies in space there is no exact solution of the governing equations that represents a progressive or travelling wave (Schelkunoff 1951). But if the spatial variation is slow

enough a progressive wave 'solution' that is a very good approximation can be constructed as follows. The medium may be said to be locally plane stratified. It may be simulated by a number of discrete thin slabs each containing a homogeneous medium with the properties of the actual medium at some point within the slab. By making the slabs thin and numerous enough the actual medium can be simulated as closely as desired. This device was used by Lord Rayleigh(1912) and later by Bremmer (1949).

For the fictitious homogeneous medium of one slab, solutions can be found that are progressive plane waves. One of these is described by a propagation vector $\boldsymbol{k}_1$ of magnitude $k_1 = n_1 k$ where n_1 is the refractive index. The ratios of the components of its electromagnetic field are expressed in terms of n_1 and a polarization ρ_1 given, for the ionosphere, by magnetoionic theory. The direction of $\boldsymbol{k}_1$ is the wave normal. The direction of the ray may be different (§ 8). When a distance $\mathrm{d}\boldsymbol{r}$ is traversed, the change of phase is $\boldsymbol{k}_1 \cdot \mathrm{d}\boldsymbol{r}$. The wave is incident on the boundary of the slab and is transmitted into the medium in the next slab where the propagation vector is $\boldsymbol{k}_2$ with magnitude $k_2 = n_2 k$. Snell's law requires that the projections of $\boldsymbol{k}_1$, $\boldsymbol{k}_2$ on the boundary shall be the same. There may be more than one wave that satisfies this but we choose the one for which $\boldsymbol{k}_2$ is closest to $\boldsymbol{k}_1$, that is the one for which, in the limit of infinitesimally thin slabs, $\boldsymbol{k}$ is a continuous function of the space coordinates. In the second slab the change of phase in distance $\mathrm{d}\boldsymbol{r}$ is $\boldsymbol{k}_2 \cdot \mathrm{d}\boldsymbol{r}$. If many slabs are traversed the phase changes $\boldsymbol{k} \cdot \mathrm{d}\boldsymbol{r}$ have to be added, where $\boldsymbol{k}$ has magnitude nk in the direction of the wave normal. In the limit, for a continuous medium, the total change of phase for any path is

$$\mathscr{E} = \int_{\text{path}} \boldsymbol{k} \cdot \mathrm{d}\boldsymbol{r}. \tag{3}$$

This assumes that the change of phase on crossing any boundary plane between slabs is zero. This is not always true and the effect, though small, must be allowed for (§ 4).

The function $\mathscr{E}$ is the eikonal function and has the property

$$\operatorname{grad} \mathscr{E} = \boldsymbol{k}. \tag{4}$$

It is supposed that the starting value of $\boldsymbol{k}$ is given for some region of space, usually near the source of the waves, or in the free space below the ionosphere. The progressive wave can then be traced through successive regions of the variable medium. Thus any one field component F of the wave is given by

$$F = F_0 \exp\left(-\mathrm{i}\int \boldsymbol{k} \cdot \mathrm{d}\boldsymbol{r}\right). \tag{5}$$

The path of integration in (3) or (5) is not restricted to be in the direction of the wave normal nor of the ray, but can extend to any point in space where the approximations of ray theory are well enough satisfied. For example in the waveguide discussed in § 1 it is simplest to choose a path that is perpendicular to the boundary planes of the guide.

The field component F may change in space not only because of the change of phase in the exponential factor of (5), but also because the other factor, F_0, changes. In the simplest form of ray theory this additional change is ignored completely. Then the only variable in (5) is the phase integral in the exponent, which shows that the change of phase is cumulative as any path is traversed. This has been called the 'phase memory concept'. It is this and Snell's law which are the two most important principles underlying ray theory.

But F_0 cannot stay constant for all field components. The ratios of the various field components F at any point are given by n and ρ, which change with the space coordinates. In some applications of the phase integral method the change of F_0 must be found, though it is a much

smaller change than that of the exponential or phase factor. This can be done by studying the transmission coefficients for the boundaries between the strata. It was done for isotropic media by Rayleigh (1912) and Bremmer (1949). The method could in principle be applied to magnetoionic theory but would be algebraically complicated. An alternative method is discussed in § 4, which deals with W.K.B. solutions.

(*b*) *Analytic continuation*

Many of the variables used in physics can take complex values. For example if the refractive index n, for a progressive wave, is complex, the real part determines the phase velocity and the imaginary part determines the attenuation. An example of waves whose wave normals make complex angles with real coordinate axes was mentioned in § 1 for the waveguide problem. There are many other well known examples. The principle of analytic continuation extends this idea to the study of variable complex functions of one or more complex variables.

In the theory of the differential equations of physics both the dependent variable y and the independent variable z are treated as complex. In physical problems the differential coefficient $\mathrm{d}y/\mathrm{d}z$ exists for nearly all values of z so that $y(z)$ satisfies the Cauchy–Riemann conditions and is then said to be analytic. It is shown in standard text books on the theory of complex variables (see, for example, Whittaker & Watson 1927, ch. 5) that if $y(z)$ is analytic and known for all z within some domain of the z-plane, however small, then all derivatives of y are known, and a Taylor series can be constructed for $y(z)$ which can be used to evaluate y at points outside the domain. By successive use of Taylor series in this way the value or values of $y(z)$ can be found for larger regions of the z-plane and, for the functions encountered in physics, usually for the whole of the z-plane except for the singular points of $y(z)$. This use of the analytic property for some part of the z-plane is implied in most methods for solving ordinary differential equations, for example the method of ascending power series.

Thus the independent variables are allowed to take complex as well as real values. They may be space coordinates, or velocity components or angles or angular momenta, etc. The technique is now so widely used in many different branches of physics, that it has the status of a basic physical principle. The singularities of the function which is continued analytically often occur for complex values of the independent variable z and not for the real values recorded in actual observations. But it is these singularities that play an essential part in the behaviour of the function for all values of z, so that by studying them and their neighbourhoods it is possible to make deductions about the behaviour of the function for the accessible real values of z.

An outstanding example is in the physics of fundamental particles. This includes the study of the complex scattering amplitude $F(\theta, s)$ for a scattering angle θ when one particle collides with another, where s is the energy in the centre of mass frame. The function F can be expressed as a series with coefficients $f(l, s)$, in which each term is associated with an integer l, the total angular momentum quantum number. The series is then converted to a contour integral in the complex l-plane, where l is now a continuous complex variable (see, for example, Eden 1967, ch. 5.) This is done by the Watson transform which was used originally in two famous papers on radio propagation (Watson 1919*a*, *b*), and is here applied to angular momentum treated as a complex variable.

Another example, more familiar to those concerned with the ionospheric plasma, is in the theory of Landau damping, where the velocity distribution function of the electrons is continued analytically to use complex values of the components of the velocities (Clemmow & Dougherty 1969, ch. 8).

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

One of the most powerful features of the phase integral method is the use, in this way, of complex values of the space coordinates. We say that the progressive waves are 'followed' into complex space. Thus we study the components of the propagation vector $\boldsymbol{k}$ and of the phase integral $\int \boldsymbol{k}\cdot \mathrm{d}\boldsymbol{r}$ in (3), (5) as complex analytic functions of the complex space coordinates. The singularities of these functions include the coupling points, which are branch points.

3. Horizontal stratification

In nearly all radio propagation problems that are studied by the phase integral method, it is assumed that the composition of the ionosphere or troposphere is a function only of the height z above the Earth's surface. If the Earth's curvature is neglected the ionosphere is thus a horizontally stratified medium. For a curved Earth it is radially stratified but the curvature can be allowed for by supposing that the Earth is flat and using a modified refractive index (Booker & Walkinshaw 1946; an example is given in § 5 below). Thus the assumption of a horizontally stratified medium will deal with most problems and it is the only case considered here.

Further, the electromagnetic field in the free space above the Earth's surface is assumed to be composed of two progressive plane waves, one travelling obliquely upwards and the other formed from it by one or more reflexions from the ionosphere. The wave normals of these waves define a vertical plane called the 'plane of incidence'. Cartesian coordinates are chosen with the z-axis vertically upwards and the x-axis in the plane of incidence. The wave normals of the two waves in free space make angles θ, $\pi-\theta$ with the z-axis. The x dependence of all field components of both waves in free space is given by a factor $\exp(-ikx\sin\theta)$ and there is no y dependence. Snell's law requires that this applies at all heights z within the ionosphere as well. Hence the propagation vector $\boldsymbol{k}$ has components

$$k(\sin\theta, 0, q), \tag{6}$$

where

$$q^2+\sin^2\theta = n^2. \tag{7}$$

Here q and n are functions of z only, and $\sin\theta$ is a constant, possibly complex. It is only necessary to study the z dependence of the fields at fixed values of x and y. Thus only the coordinate z varies, and it is allowed to take complex values. For fixed x, y, (5) becomes

$$F = F_0 \exp\left(-ik\int^z q\,\mathrm{d}z\right). \tag{8}$$

For an isotropic ionosphere the refractive index n is independent of the direction of $\boldsymbol{k}$, and then q is given by (7) which leads to

$$q^2 = \cos^2\theta - X/(1-iZ). \tag{9}$$

For the anisotropic ionosphere it is necessary to use magnetoionic theory to find n, which now depends on the direction of $\boldsymbol{k}$. Then q is given by the Booker quartic equation (17) below (Booker 1938). The analytic properties of its solutions, in the complex z-plane, are thus of the greatest importance in phase integral methods (Smith 1974).

The assumption, used in (6)–(9), that the incident wave is plane, is adequate for most radio propagation problems, but the correct value of θ must be found. A transmitter of small dimensions radiates a spherical wave which can be expressed as an angular spectrum of plane waves (Booker & Clemmow 1950; Clemmow 1966). It is necessary to pick out from this spectrum the particular plane wave that predominates at a given receiver. This is done by tracing a ray from transmitter

to receiver, usually by trial and error in an iterative process. If the losses caused by electron collisions are allowed for, the ray has complex direction cosines and θ is complex (Budden & Jull 1964). This leads to the technique of complex ray tracing (Budden & Terry 1971) which uses only ray theory. It may therefore be considered to be an application of the phase integral method, but its discussion is beyond the scope of this review. It is here assumed that the direction θ, in general complex, of the wave normal of the incident wave is known.

4. W.K.B. solutions in stratified media

It is now necessary to discuss in more detail the 'slowly varying' factor F_0 in (8). The field component F may be any component of the electromagnetic field of the wave, but it proves useful to consider only the four components

$$E_x, E_y, Z_0 H_x, Z_0 H_y. \tag{10}$$

From these the others are easily found. The factor Z_0 is the characteristic impedance of free space and is inserted to give these four numbers the same physical dimensions. It is convenient to write

$$Z_0 H_x = \mathscr{H}_x, \quad Z_0 H_y = \mathscr{H}_y. \tag{11}$$

Consider first an isotropic medium and waves whose polarization is linear with $\boldsymbol{E}$ horizontal. Then $E_x = \mathscr{H}_y = 0$, and (8) can be written for both E_y and $\mathscr{H}_x$, thus

$$\begin{pmatrix} E_y \\ \mathscr{H}_x \end{pmatrix} = \begin{pmatrix} E_{y0} \\ \mathscr{H}_{x0} \end{pmatrix} \exp\left(-ik \int_0^z q \, dz \right). \tag{12}$$

Now for the conditions of ray theory, (12) must have the properties of a progressive wave in a homogeneous medium whence it is easy to show that

$$\mathscr{H}_x = -qE_y. \tag{13}$$

For a loss free medium no energy can be absorbed. The energy flux must be independent of z. Thus the z-component $-\mathrm{Re}\,(H_x^* E_y)$ of the time averaged Poynting vector is constant. This is satisfied, for real q, if $E_{y0} \propto q^{-\frac{1}{2}}$ and $\mathscr{H}_{x0} \propto q^{\frac{1}{2}}$, so that (11) is

$$\begin{pmatrix} E_y \\ \mathscr{H}_x \end{pmatrix} = A \, e^{i\gamma} \begin{pmatrix} q^{-\frac{1}{2}} \\ q^{\frac{1}{2}} \end{pmatrix} \exp\left(-ik \int_0^z q \, dz \right), \tag{14}$$

where A is a real constant and γ is some unspecified function of z, real when z is real. This method cannot be used in a lossy medium and it provides no justification for the analytic continuation of (13) into the complex z-plane. An alternative derivation, without these limitations, makes use of the differential equation

$$d^2 E_y / dz^2 + k^2 q^2 E_y = 0 \tag{15}$$

satisfied by E_y. It is solved with the approximation that q is a slowly varying function of z. This was done very clearly and simply by Gans (1915), and leads to (14) where γ is now a constant. This solution is called the W.K.B. solution and the method is called the W.K.B. method, where the letters are the initials of authors who derived it much later. There were also some authors who used it earlier. The history is given by Heading (1962). Gans was considering the propagation of light in an inhomogeneous isotropic medium, but his treatment is very close to what is needed for radio waves in the stratified ionosphere, and is very clearly presented. The term W.K.B. will, however, be used here because it is so widely accepted.

For an anisotropic medium the wave (8) must be one of the characteristic waves, ordinary or extraordinary, upgoing or downgoing, with the appropriate solution q of the Booker quartic. The polarization is given by magnetoionic theory, as for a homogeneous medium, so that all four field components (10) must be retained. The analogue of the W.K.B. method for this case was given by Clemmow & Heading (1954). The differential equation satisfied by the components (10) is written

$$\mathrm{d}\boldsymbol{e}/\mathrm{d}z = -\mathrm{i}k\boldsymbol{T}\boldsymbol{e}, \tag{16}$$

where $\boldsymbol{e}$ is the column matrix with elements E_x, $-E_y$, $\mathscr{H}_x$, $\mathscr{H}_y$ and $\boldsymbol{T}$ is a 4×4 matrix function of θ and of the parameters X, Y, Z of the ionosphere. Then it can be shown that the Booker quartic is

$$\mathrm{Det}\,(\boldsymbol{T}-q\boldsymbol{1}) = 0, \tag{17}$$

where $\boldsymbol{1}$ is the unit 4×4 matrix. For any solution q, the W.K.B. solution, analogous to (14), is

$$\boldsymbol{e} = A\,\mathrm{e}^{\mathrm{i}\gamma}\,\boldsymbol{e}_i \exp\left(-\mathrm{i}k\int_0^z q_i\,\mathrm{d}z\right) \quad (i = 1, 2, 3, 4), \tag{18}$$

where $\boldsymbol{e}_i$ is an eigencolumn vector of $\boldsymbol{T}$ belonging to the eigenvalue q_i, and A is a constant. To complete the definition of $\boldsymbol{e}_i$ some further condition must be imposed. A convenient definition, analogous to a normalization, was adopted by Budden & Clemmow (1957). The factor $\mathrm{e}^{\mathrm{i}\gamma}$ cannot always be constant, and is discussed in § 9.

The W.K.B. solutions (18) may be said to be the embodiment of the approximations of ray theory. They each represent a wave that has the correct polarization, wave impedance, phase velocity and attenuation for a progressive wave in a homogeneous medium. They are analytic functions of z and may be continued analytically into the complex z plane. The purpose of the phase integral method is to use these solutions for those problems where they can give useful results.

The accuracy of the W.K.B. solutions (14) or (18) can be tested by substituting them in the governing differential equations (15) or (16) respectively and examining the error. It is found that near points in the z-plane where two roots, say q_1, q_2, of the quartic (17) are equal, the error is large for the two solutions that use q_1, q_2. These points are called 'coupling points'; they include 'reflexion points'. Surrounding each of them is a domain where the error is large. For reasons given in § 6 these domains are called Airy regions. If one of the ray theory solutions (18) is to be continued analytically and remain a good approximation, it is clearly necessary to avoid entering an Airy region associated with it.

If the coupling points are well separated from each other, however, so that their Airy regions do not overlap, it is possible to proceed by analytic continuation from one part to another of the z-plane, using only the approximations of ray theory, that is the W.K.B. solutions, and this is the phase integral method. There is one other very important requirement that sets a limit to this procedure. It is connected with the Stokes phenomenon and is described in § 7.

5. Guided waves: diffraction round the Earth

As an example of a wave guiding system Eckersley (1931) considered a stratified collisionless isotropic ionized medium in which the electron concentration N and the refractive index n are given by

$$N = \alpha z^2, \quad n^2 = 1-\beta z^2, \tag{19}$$

where α and β are related real constants. Suppose that the waves are linearly polarized with $\boldsymbol{E}$ parallel to the y-axis (Eckersley did not specify the polarization). Then its only non-zero component is E_y, which satisfies

$$\mathrm{d}^2E_y/\mathrm{d}z^2 + k^2(\cos^2\theta - \beta z^2)\,E_y = 0. \tag{20}$$

This is exactly analogous to the Schrödinger equation

$$\mathrm{d}^2\psi/\mathrm{d}z^2 + 2m\hbar^{-2}(E - \tfrac{1}{2}m\omega^2 z^2)\,\psi = 0, \tag{21}$$

used in the quantum theory of the harmonic oscillator whose classical angular frequency is ω. The object of both problems is to find the eigenfunctions, and the eigenvalues of $\cos^2\theta$ in (20), and of the energy E in (21).

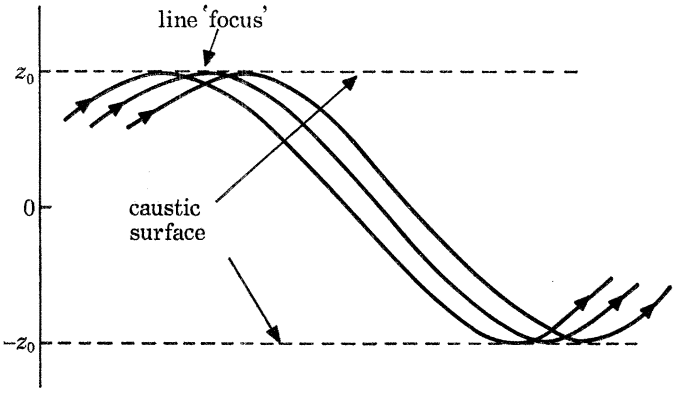

FIGURE 1. The rays for a self-consistent mode in an ionized medium for which the electron concentration is $N = \alpha z^2$. The rays are reflected where $z = \pm z_0$, and these planes are caustic surfaces. There is a phase advance of $\frac{1}{2}\pi$ at each reflexion.

First consider (20). According to ray theory there are, near $z = 0$, two crossing progressive waves whose wave normals make angles θ, $\pi - \theta$ with the z-axis where $z = 0$. The one travelling in the direction of positive z will be called the upgoing wave and it is given by (14), namely

$$E_y = q^{-\frac{1}{2}} \exp\left(-\mathrm{i}k\int_0^z q\,\mathrm{d}z\right), \tag{22}$$

where an unimportant constant factor has been omitted and where, from (7) and (19)

$$q^2 = \cos^2\theta - \beta z^2. \tag{23}$$

The wave normals for this wave make an angle arc sin (q/n) with the x-axis. Where

$$q = 0, \quad z = \beta^{-\frac{1}{2}}\cos\theta = z_0, \tag{24}$$

they are parallel to the x-axis. This z_0 is sometimes called the 'level of reflexion'. If the orthogonal trajectories of the wave fronts, that is the 'rays', are traced, they are as shown in figure. 1. They are all the same shape (pure sine waves in this example) and they touch the plane surface $z = z_0$, which is called a 'caustic surface'. Thereafter they are directed towards negative z. According to ray theory, therefore, the wave (22) is reflected where $z = z_0$ and thus converted into the other progressive wave, the downgoing wave given by

$$E_y = q^{-\frac{1}{2}} \exp\left(-2\mathrm{i}k\int_0^{z_0} q\,\mathrm{d}z\right) \exp\left(\mathrm{i}k\int_0^z q\,\mathrm{d}z\right). \tag{25}$$

This argument ignores the failure of the W.K.B. solutions (22), (25) near $z = z_0$. It is now supposed that a similar reflexion process occurs for the wave (25) near the other reflexion level $z = -z_0$, where it is converted back to the first wave thus

$$E_y = q^{-\frac{1}{2}} \exp\left(-2ik\int_{-z_0}^{z_0} q\,dz\right) \exp\left(-ik\int_0^z q\,dz\right). \tag{26}$$

For a self-consistent mode, (22) and (26) must be identical and this gives the phase integral condition

$$2k\int_{-z_0}^{z_0} q\,dz = 2\pi r, \tag{27}$$

where r is an integer. This leads to

$$k\cos^2\theta = 2r\beta^{\frac{1}{2}}, \tag{28}$$

which is the mode condition determining the eigenvalues of $\cos\theta$. If a similar argument is applied to the Schrödinger equation (21) it gives

$$E = r\hbar\omega. \tag{29}$$

Now Eckersley realized that these results are wrong and should be replaced by

$$k\cos^2\theta = (2r+1)\,\beta^{\frac{1}{2}}, \quad E = (r+\tfrac{1}{2})\,\hbar\omega \tag{30}$$

respectively. He attributed the error to the failure of ray theory near $z = \pm z_0$. There are, however, at least two simpler alternative explanations.

Consider, first, a narrow pencil of the rays in the upgoing wave (22). These all touch the caustic surface $z = z_0$ and thereafter form a pencil of downgoing rays. Near the caustic every ray of the pencil crosses every other ray that is in the same vertical plane. The rays are therefore passing through a line focus near the caustic. In these circumstances it can be shown by a very simple application of Huygens' principle that there must be a phase advance of $\frac{1}{2}\pi$. This occurs for the two reflexions at $z = \pm z_0$. Thus a phase shift π must be added to the right of (27), and a similar change is made in (28). The results give (30) which are the correct eigenvalues obtained from an exact solution of the differential equations (20), (21). This suggests that we can continue to use ray theory for rays that have passed through a focus, or equivalently have been reflected in a slowly varying medium. This is exactly what is done when geometrical optics is used in the study of optical instruments.

The alternative approach, used in the phase integral method, is to avoid the points $z = \pm z_0$ by continuing the expression (22) analytically into the complex z-plane (figure 2). The function q has branch points at $z = \pm z_0$ and surrounding each of them is a region called the Airy region (§ 6), where ray theory fails. But ray theory can be used outside the two Airy regions. We therefore 'follow' the solution (22) from the point P along the path shown with single arrows in figure 2, and back to P. This path encircles the branch point so that on returning to P the function q has its original value but with the opposite sign. To deal with this, a branch cut is inserted as shown in figure 2. When it is crossed clockwise, q changes from its old value q_1 to the new value $q_1 e^{i\pi}$. But the function (22) must be analytic and therefore continuous, so that on crossing the cut, q must be multiplied by $e^{-i\pi}$ wherever it appears. Thus the factor $q^{-\frac{1}{2}}$ becomes $iq^{-\frac{1}{2}}$, and $-q$ in the integrand becomes $+q$. It is now clear that on returning to P, the expression (22) has been converted by analytic continuation into

$$E_y = iq^{-\frac{1}{2}} \exp\left(-ik\int_C q\,dz\right), \tag{31}$$

where the contour C runs from P clockwise round z_0, outside its Airy region, and back to P, and it is implied that q in the integrand is changed to $-q$ after the cut is crossed. This choice of C is to ensure that the approximations of ray theory are good enough at all points on it. Such a contour is called a 'good path'. But q is an analytic function of z except at $z = \pm z_0$, so that the contour may be distorted anywhere provided that the points $\pm z_0$ are not crossed. It can thus be distorted to the real z axis as indicated by double arrows in figure 2. Provided that the good path exists, it is not necessary to use it. Then (31) becomes

$$E_y = \mathrm{i}q^{-\frac{1}{2}} \exp\left(-2\mathrm{i}k\int_0^{z_0} q\,\mathrm{d}z\right) \exp\left(\mathrm{i}k\int_0^{z} q\,\mathrm{d}z\right), \tag{32}$$

which is the downgoing wave and is the same as (25) except that the factor i is now present and correctly gives the $\frac{1}{2}\pi$ phase advance. This has been derived entirely by ray theory and analytic continuation, and has not made any use of full wave solutions. It is tempting to say, in physical language, that the upgoing wave (22) has propagated into complex space, travelled through the region surrounding the reflexion point z_0, but remote from it, and has returned to P as the downgoing wave.

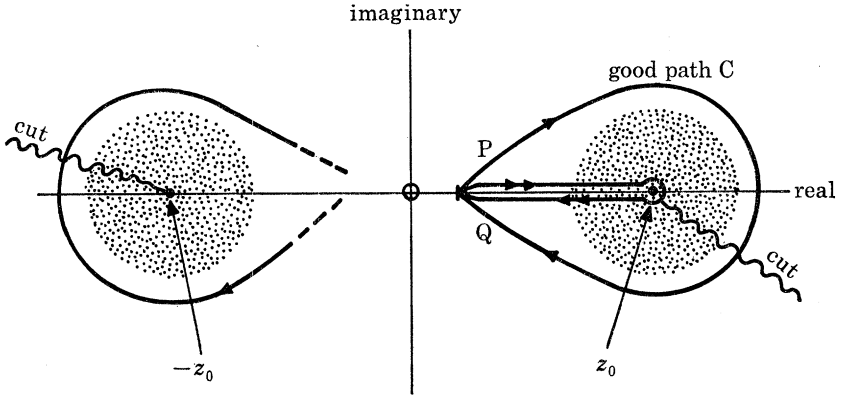

FIGURE 2. The complex z-plane for the guided wave problem of figure 1. The reflexion points $\pm z_0$ are each surrounded by an Airy region shown shaded. The thick lines are contours used in the phase integral method.

It is stressed that the contours must encircle both z_0 and $-z_0$ clockwise. If the wrong direction is used it gives the wrong answer. The reasons are discussed in § 7.

Although the argument has been presented here for the special electron distribution function (19), it would apply equally well to any other slowly varying analytic function with an isolated zero of q^2 at $z = z_0$. The reflexion coefficient is the ratio of the downgoing (reflected) wave (32) to the incident (upgoing) wave (22). If this is measured at the 'reference level' $z = 0$, it gives

$$R(0) = \mathrm{i}\exp\left(-2\mathrm{i}k\int_0^{z_0} q\,\mathrm{d}z\right). \tag{33}$$

This is the phase integral formula for the reflexion coefficient. It can be evaluated when the angle of incidence θ is known, and can be used whether or not the waves form part of a guided wave system.

This technique can now be applied to other guided wave problems. A simpler example occurs if the refractive index n of the air just above the Earth's surface is given by

$$n^2 = 1 - 2\alpha z, \tag{34}$$

which may happen for a height of several hundred metres in the meteorological conditions where a warm air mass comes over a cool sea. This, and other cases where the variation is not exactly linear, were studied by Booker & Walkinshaw (1946) whose techniques included the phase integral method. There is a reflexion level where

$$q = 0, \quad z = \cos^2\theta/2\alpha = z_0. \tag{35}$$

The sea surface can usually be considered to be a perfect reflector. The two reflexions thus form a wave guiding system, and the mode condition can be derived by the phase integral method. Its solutions give real values of $\cos\theta$ and of z_0. This leads to the 'anomalous propagation' of high frequency radio waves (in the range roughly 40–400 MHz) in tropospheric ducts.

If the medium outside the Earth is entirely free space, but if the Earth's curvature is allowed for, the apparent bending of a straight ray away from the Earth's surface can be simulated by supposing that the Earth is flat, and by using a modified form of the refractive index n given by

$$n^2 = 1 + 2z/a, \tag{36}$$

where a is the Earth's radius. The height variation of the refractive index of the air can also be included by using a different value for this a (Booker & Walkinshaw 1946). Thus the propagation of waves round the curved earth can be studied by using the 'Earth flattening approximation' (36).

Eckersley (1932*b*) investigated this problem and it was one of the most remarkable successes of his treatment. He did not use the Earth flattening approximation but his method is, in essentials, equivalent to it. For (36) q is given by

$$q^2 = \cos^2\theta + 2z/a, \tag{37}$$

and the reflexion 'level' z_0 by

$$\cos^2\theta + 2z_0/a = 0. \tag{38}$$

Suppose that the waves are linearly polarized with $\boldsymbol{E}$ horizontal, and that the Earth's surface is a perfect reflector with reflexion coefficient -1 for this polarization. The phase integral condition can be found by starting just above the ground with an upgoing wave given by (22) and (37), continuing it analytically in the complex z plane clockwise round z_0, and returning to the ground at $z = 0$. Then after reflexion at the ground it must give the original wave. This leads to

$$\cos^3\theta = -3\pi(r + \tfrac{3}{4})/ka, \tag{39}$$

where r is an integer, positive or zero. The correct cube root of (39) must be chosen (see Budden 1961*b*, § 12.2), which leads to

$$z_0 = a^{\frac{1}{3}}k^{-\frac{2}{3}}\{3\pi(r + \tfrac{3}{4})\}^{\frac{2}{3}} \exp(-\tfrac{1}{3}\pi\mathrm{i}). \tag{40}$$

If the positions of the transmitter and receiver are given, the received signal can now be expressed as the sum of contribution from the modes $r = 0, 1, 2, \ldots$. Because $\sin\theta$ is complex the modes are attenuated through a factor $\exp(-\mathrm{i}kx\sin\theta)$ as they travel in the x direction. It usually happens that only the least attenuated modes contribute appreciably, and for large x it may be only the one mode $r = 0$, if both transmitter and receiver are near the ground.

The reflexion level (40) is at a complex value of z. There is no easy way in which this reflexion process can be visualized in real space, but the reflexion at the complex z_0 has all the features of the reflexion which occurs in the other examples where z_0 is real. This simple picture is only achieved if the process of continuation into complex space is accepted.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

The $\frac{1}{2}\pi$ phase advance on reflexion is included in deriving (39). Eckersley (1932*b*) included it for reasons based on a comparison of his results with those of Watson (1919*a*) who had given a more exact (full wave) treatment for the spherical Earth.

6. The Airy region and the principle of uniform approximation

Because the methods of ray theory are approximate, it is necessary to make some examination of the errors. This leads to further rules which sometimes restrict the use of the phase integral method, but also help to show where it can safely be used. The behaviour of the solutions near a reflexion point must therefore be examined. The first full treatment was given by Gans (1915). It applies for an isotropic medium and is reviewed here by discussing again the case where the waves are obliquely incident on a stratified ionosphere, and linearly polarized with the electric vector horizontal. Gans discussed also the case where the waves are linearly polarized with the magnetic vector horizontal and showed that usually there are only minor differences. It is now known that in some conditions the difference can be important. A full treatment of the second case requires full wave techniques, beyond the scope of the phase integral method, given by Hirsch & Shmoys (1965).

The electric field component E_y satisfies (15) and there is a reflexion level where $q = 0$, at $z = z_0$ say. Near here the W.K.B. solutions (22), (25) cannot be used, and so the differential equation must be re-examined to find more accurate solutions. If the medium is sufficiently slowly varying the derivative $\mathrm{d}(q^2)/\mathrm{d}z$ must be small so that over a small enough range of z it may be supposed that q^2 varies linearly with z:

$$q^2 = -2\alpha(z - z_0). \tag{41}$$

Gans used only real values of α and z_0 but this restriction is unnecessary. In the ionosphere, with collisions allowed for, they are both complex. Insertion of (41) in (15) gives

$$\mathrm{d}^2E_y/\mathrm{d}z^2 = 2\alpha k^2(z - z_0), \tag{42}$$

which was solved exactly by Gans in terms of Bessel functions of order $\frac{1}{3}$. This method was known much earlier, but is cumbersome. The solutions of (42) used today are the Airy Integral functions (Miller 1946) which are very much simpler. The equation (42) was also studied by Lord Rayleigh (1912) who pointed out that one of its solutions is the integral that had been used by Airy in the study of the rainbow. He also mentioned the solutions $J_{\frac{1}{3}}$ and $J_{-\frac{1}{3}}$ but did not pursue the topic to find the reflexion coefficient.

The required solution of (41) is thus

$$E_y = \mathrm{Ai}\{(2\alpha k^2)^{\frac{1}{3}}(z - z_0)\}, \tag{43}$$

which is selected because it is the only one that satisfies the requirement that E_y decreases with increasing z when z is large, real and positive. Now if $|z - z_0|$ is great enough (43) is given by a linear combination of its two asymptotic forms

$$q^{-\frac{1}{2}} \exp\left(\pm \mathrm{i}k \int_{z_0}^{z} q\,\mathrm{d}z\right), \tag{44}$$

where q is given by (41). But these are proportional to the two W.K.B. solutions (22), (25). Suppose that, in proceeding away from z_0, the function q^2 in (40) remains linear until $|z - z_0|$ is large enough for the asymptotic forms (44) to be used. Then beyond this the W.K.B. solutions

are valid and it does not matter if q^2 departs from linearity. In this way the W.K.B. solutions remote from z_0 are 'fitted on' to the asymptotic forms of the solution (43), and so the ratio of the two W.K.B. solutions can be found. It depends on arg z. If z is real and $\ll |z_0|$, that is near the base of the ionosphere, it can be shown (Budden 1961*a*, § 16.6) that the reflexion coefficient is given by the phase integral formula (33). This includes the factor i which is the phase advance of $\frac{1}{2}\pi$ already mentioned.

How close must $q^2(z)$ be to linearity for this argument to work? This question is best answered by an alternative approach used by Langer (1937). It is supposed that q^2 has a zero at $z = z_0$ but is no longer exactly linear. In the differential equation (15) change the independent variable to

$$\zeta = \left\{\tfrac{3}{2}ik\int_{z_0}^{z} q\,\mathrm{d}z\right\}^{\frac{2}{3}}, \tag{45}$$

and the dependent variable to

$$U = V^{-1}E_y \quad \text{where} \quad V = \zeta^{\frac{1}{4}}q^{-\frac{1}{2}}. \tag{46}$$

Then it becomes

$$\mathrm{d}^2U/\mathrm{d}\zeta^2 - \zeta U = U\{2V^{-2}(\mathrm{d}V/\mathrm{d}\zeta)^2 - V^{-1}(\mathrm{d}^2V/\mathrm{d}\zeta^2)\}. \tag{47}$$

Now if $\{q(z)\}^2$ is exactly linear, V is a constant and the right hand side of (47) is zero. In a sufficiently slowly varying ionosphere it is small enough to be neglected (Budden 1972). Then (47) is the Stokes equation whose solution is the Airy Integral function, so that

$$E_y = \zeta^{\frac{1}{4}}q^{-\frac{1}{2}}\,\mathrm{Ai}\,(\zeta). \tag{48}$$

If now the two asymptotic forms $\zeta^{-\frac{1}{4}}\exp(\mp\frac{2}{3}\zeta^{\frac{3}{2}})$ for $\mathrm{Ai}(\zeta)$ are used in turn in (47), they give just the two W.K.B. solutions (22), (25) even though q^2 is not exactly linear. In this way a solution (48) has been found that is *uniformly* valid, not only in the remote regions where the asymptotic forms give the ray theory or W.K.B. solutions, but also throughout a domain containing z_0, where ray theory fails. This is an example of the 'principle of uniform approximation' which has been used recently in several branches of physics. The accuracy of these solutions is examined by studying the 'error' term on the right hand side of (47). The subject is reviewed by Berry & Mount (1972).

The asymptotic forms can often be used if $|\zeta| \gtrsim 1$. Thus the curve $|\zeta| = 1$ gives approximately the boundary of the Airy region. In the special case of exact linearity (41), this boundary is the circle $|z - z_0| \approx (2\alpha k^2)^{-\frac{1}{3}}$.

The solution (48) may be said to extend the results of the phase integral method into the Airy region near z_0. It is necessary to do this when the components of the electromagnetic field must be known near the level of reflexion, as occurs for example in the study of wave interaction near a reflexion level. The theory has been used in this way recently by Maslin (1975).

7. The Stokes phenomenon

An effect must now be discussed which sets a limit to the use of the phase integral method. Ray theory solutions are approximate. At points of the z-plane that are not in an Airy region they can be identified with the asymptotic forms of a more accurate solution. For a reflexion point which is far enough away from other coupling points, this solution uses the Airy Integral function as in (48). It has two asymptotic forms and is given by a linear combination of them thus

$$\mathrm{Ai}(\zeta) \sim A\zeta^{-\frac{1}{4}}\exp(-\tfrac{2}{3}\zeta^{\frac{3}{2}}) + B\zeta^{-\frac{1}{4}}\exp(\tfrac{2}{3}\zeta^{\frac{3}{2}}), \tag{49}$$

but the 'constants' A, B are not the same for all $\arg\zeta$. For $-\frac{2}{3}\pi \leqslant \arg\zeta \leqslant \frac{2}{3}\pi$ only one of the two terms is present. This property is often displayed in a 'Stokes diagram' (see figure 4*a*).

The term in (49) whose exponential factor is the greater is called the 'dominant' term and the other is the 'subdominant' term. The exponentials have equal moduli where $\arg\zeta = \pm\frac{1}{3}\pi, \pi$ and these are called anti-Stokes lines; on them the dominancies change over. For fixed $|\zeta|$ the ratio of the moduli of the exponential factors is greatest or least where $\arg\zeta = 0, \pm\frac{2}{3}\pi$ and these are called Stokes lines. For the single valued function (49) the Stokes diagram is conveniently drawn as a polar diagram in which the polar angle is $\arg\zeta$ (see figure 4). It shows the Stokes and

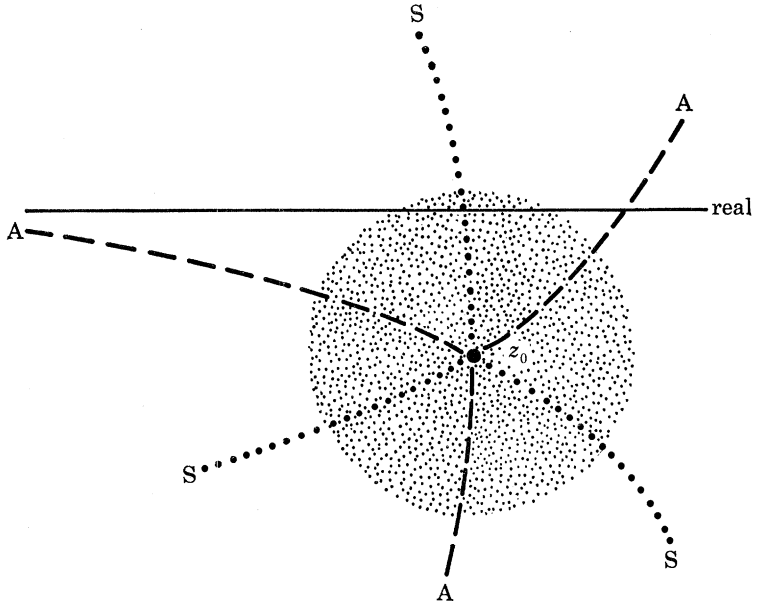

FIGURE 3. The complex z-plane for an isotropic ionosphere with collisions included, where the electron concentration $N(z)$ is monotonically increasing, and real when z is real. For vertically incident waves the refractive index is zero where $z = z_0$, which is the complex reflexion point. The Stokes lines S and anti-Stokes lines A are where $\int_{z_0}^{z} q\,dz$ is purely imaginary and purely real, respectively. The Airy region is shown shaded.

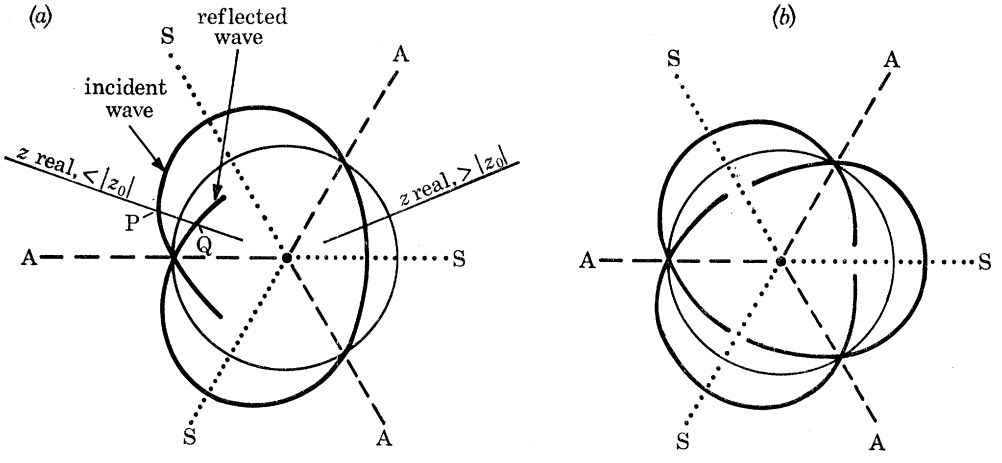

FIGURE 4. Stokes diagrams. (*a*) Applies when the reflexion point z_0 is sufficiently isolated, so that the solution $\mathrm{Ai}(\zeta)$ is good enough. There is no Stokes phenomenon on the Stokes line at $\arg\zeta = 0$. (*b*) Applies when the solution $\mathrm{Ai}(\zeta)$ cannot be used. There is now a dominant term where $\arg\zeta = 0$, and the subdominant term there displays the Stokes phenomenon, as shown by a break in the inner thick line.

anti-Stokes lines and a circle centred at the origin. There is also a thick line which can be cut either once or twice by any radius. For a given $\arg\zeta$, if the radius cuts the thick line inside the circle, the subdominant term is present in (49) and if it cuts the thick line outside the circle, the dominant term is present.

The asymptotic forms in (49) are only approximate. For a fixed $|\zeta|$ the dominant term is greatest on a Stokes line and here its error is greatest. It can be shown that this error is then greater than the modulus of the subdominant term. Thus the multiplier of the subdominant term may change near a Stokes line if the dominant term is present. A full study of the function $\mathrm{Ai}(\zeta)$ shows that this change must occur on the Stokes lines where $\arg\zeta = \pm\frac{2}{3}\pi$, so that the multiplier of the subdominant term drops to zero when $\arg\zeta$ increases through $-\frac{2}{3}\pi$ or decreases through $+\frac{2}{3}\pi$ (figure 4*a*). It cannot occur on the Stokes line $\arg\zeta = 0$ because there is no dominant term there to mask a change in the subdominant term.

This property is an example of the 'Stokes phenomenon of the discontinuity of the arbitrary constants' (Stokes 1857). It operates near Stokes lines, for all values of $|\zeta|$ both inside and outside the Airy region. Thus it applies in the regions of complex space where ray theory has been used. It is not predicted by the analytic continuation of the W.K.B. solutions. It therefore follows that where this analytic continuation is used, the regions where the continued solution undergoes the Stokes phenomenon must be avoided

Figure 3 shows the Stokes and anti-Stokes lines for the complex reflexion point z_0 when the ionosphere has a monotonically increasing electron distribution function $N(z)$ with collisions allowed for. Figure 4*a*. shows the Stokes diagram when there is an upgoing wave vertically incident from below. It can then be shown that on the real z-axis below the level of reflexion, the ray theory solution for the upgoing wave is the dominant term, represented by a point on the outer thick line at P, figure 4*a*. It is now continued analytically clockwise round z_0. Thus the thick line is traversed clockwise, and $\arg\zeta$ decreases through 2π. In this range the Stokes phenomenon does not occur and the thick line is a continuous curve. On completion, the point followed is on the inner thick line at Q and represents the downgoing or reflected wave, which here is subdominant.

If an attempt were made to encircle z_0 anticlockwise, starting at P as before, a difficulty would arise at $\arg\zeta = -\frac{2}{3}\pi$ where the thick line ends. The term has here become subdominant, and because of the Stokes phenomenon it disappears and cannot be followed further when $\arg\zeta$ goes beyond the Stokes line. It is thus important to ensure that when a 'good path' is chosen, it must be one for which the Stokes phenomenon does not occur.

We now have to consider a limitation that further restricts the choice of a good path, and which may underlie some of the criticisms that have been made of the phase integral method. The Stokes diagram figure 4*a* shows that $\mathrm{Ai}(\zeta)$ has only one asymptotic term for

$$-\tfrac{2}{3}\pi < \arg\zeta < \tfrac{2}{3}\pi, \tag{50}$$

so that $B = 0$ in (49) for this range. The requirement for this was imposed because there is no downgoing wave at real heights above $\mathrm{Re}\,(z_0)$. But the asymptotic forms are only approximate. Suppose we examine (49) for the range (50) and for some fixed $|\zeta|$ outside the Airy region. The one term that is present is subject to a small error which can be expressed as a known function of $|\zeta|$. A very small multiple of the other term would be within this error and therefore might be present but undetectable. A value of B satisfying

$$B < 3^{-\frac{1}{2}}|A| \exp\left(-\tfrac{4}{3}|\zeta|^{\frac{3}{2}} - 2|\zeta|\right) \tag{51}$$

could be tolerated throughout the range (50). This is derived by evaluating the smallest term in the asymptotic expansion for $\mathrm{Ai}(\zeta)$ as given, for example, by Miller (1946). If, now, a larger $|\zeta|$ is used, the maximum tolerable $|B|$ gets less according to (51), and tends to zero as $|\zeta| \to \infty$. If q^2 is exactly linear, as in (41), V in (46) is constant and the solution $\mathrm{Ai}(\zeta)$ in (47) is exact, and extends to $|z| \to \infty$, $|\zeta| \to \infty$. Then (51) shows that B *must* be zero for the range (50), so that it can be asserted with certainty that the one asymptotic form is then absent. The Stokes diagram, figure 4*a*, in effect does just this.

But in the actual ionosphere q^2 is never exactly linear, so that the small term on the right of (47) is not exactly zero and the solution (48) is not quite exact. The error may be small if $\mathrm{d}^2(q^2)/\mathrm{d}z^2$ is small at $z = z_0$, and if $|\zeta|$ is small enough. But when q^2 is not linear it must have either another zero as well as z_0, or a singularity, or both. Suppose the singularity or zero nearest to z_0 is at z_1. The error in (48) gets larger as $|z - z_0|$ or equivalently $|\zeta|$ increases, and is very large when z is near to z_1 .Thus (48) and thence (51) cannot be used for indefinitely large values of $|\zeta|$, and so it cannot be asserted that B is exactly zero.

If $|B| \neq 0$ for $\arg \zeta = 0$, the Stokes diagram is as in figure 4*b*. There is now a dominant term present where $\arg \zeta = 0$ and so the subdominant term there displays the Stokes phenomenon. The change of its multiplier depends on B and would modify the phase integral formula (33) for the reflexion coefficient, by introducing an extra factor. To find this factor is very complicated. It can be done in some special cases where the solutions of (15) can be expressed with functions whose asymptotic properties are known, but it has not yet been done for the general fourth order system of equations (16).

But the error in the phase integral formula (33) is very small if the reflexion point z_0 is far enough away from other coupling points and singularities. This has been tested in some special cases by comparing phase integral results with full wave solutions; see, for example, Budden & Cooper (1962), Smith (1973). For these cases it was found that the phase integral formula is still useful provided that the Airy region of the coupling point to which it is applied does not overlap the Airy region of any other coupling point.

No clear *physical* interpretation of the Stokes phenomenon has yet been given. It occurs in regions where the approximations of ray theory might be expected to apply, and it is in a sense a failure of ray theory.

8. Anisotropic medium: coupling

The discussion so far has been mainly for an isotropic ionosphere, for which q is given by (9). Then the reflexion points in the complex z-plane are where the two values of q are equal, and these are also zeros of q. If the Earth's magnetic field is allowed for, the ionosphere is anisotropic and q is a solution of the Booker quartic equation (17). The following two important effects apply in the anisotropic case:

(*a*) (Booker 1938). The ray and the wave normal are not, in general, in the same direction. In particular, for a loss free medium, reflexion occurs where the ray is horizontal, but usually the wave normal is not horizontal there. Thus the two q's, for the incident and reflected waves, are equal at the reflexion level, but not zero. A zero of q is not, in general, a reflexion level. Where $q = 0$ the wave normal is horizontal but this can occur where ray theory applies, at points remote from reflexion or coupling points.

(*b*) (Booker 1936). Near a coupling or reflexion point where two q's are equal, the associated two solutions are strongly interacting because the reflexion or coupling process is going on there.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

But, if the medium is sufficiently slowly varying, the other two solutions are unaffected and are propagated independently. They can be removed from the fourth order system of equations (16) leaving a second order equation which describes the coupling or reflexion process, and is said to be 'embedded' in the more exact system (16).

Eckersley (1950) stressed that the solutions $q(z)$ of the quartic are analytic functions of z nearly everywhere, and can be continued analytically into the complex z-plane. The function $q(z)$ is represented by a Riemann surface of four sheets, two of which touch at each reflexion or coupling point where two q's are equal. These points are branch points of $q(z)$, and are usually at complex values of z. In a loss free ionosphere the reflexion points are on the real z-axis, but there are other coupling points at complex z, where the two equal q's both apply to obliquely upgoing (or both downgoing) waves, one ordinary and the other extraordinary.

Eckersley applied these ideas only to a loss free ionosphere with $N(z)$ increasing slowly and monotonically and for vertical incidence, so that q is the same as the refractive index n. In this case there are two coupling points at values z_p, z_n of z, where $\mathrm{Re}\,(z_p) = \mathrm{Re}\,(z_n)$, and

$$-\mathrm{Im}\,(z_n) = \mathrm{Im}\,(z_p) > 0.$$

The level $\mathrm{Re}\,(z_p)$ is near where $X = 1$. When an ordinary wave travels upwards and passes near $\mathrm{Re}\,(z_p)$ it gives rise, by a coupling process, to some upgoing extraordinary wave in the Z-mode. Eckersley realized that its amplitude could be calculated by the phase integral method. The solution for the upgoing ordinary wave is continued analytically into the complex z-plane on a 'good path' which passes on the side of z_p remote from the real z-axis and returns to the real axis where $z > \mathrm{Re}\,(z_p)$ and where the solution is now the upgoing extraordinary wave or Z-mode. It is reflected at the higher level where $X = 1 + Y$. The resulting downgoing extraordinary wave gives rise by coupling to some downgoing ordinary wave which travels to the ground. Its amplitude is calculated by using the phase integral method on a path which, according to Eckersley, goes on the side of z_n remote from the real axis. In this way Eckersley explained the phenomenon of triple splitting of an ionosonde echo into ordinary, extraordinary and Z-trace, and showed correctly that the Z-trace has the polarization of the ordinary wave. He gave no reason for using z_p on the upward trip and z_n on the downward trip. In fact z_p should have been used for both (Smith 1973). The phase integral method used in this way can be extended to apply to a lossy ionosphere, and gives the correct amplitude for the Z-trace if the ionosphere is really horizontally stratified. But this amplitude is much smaller than is often observed. An alternative explanation given by Ellis (1956) is now believed to be correct. It involves oblique propagation of the waves, and back scattering from irregularities near the reflexion level where $X = 1 + Y$.

To justify the use of the phase integral method near a coupling point, it is necessary to show that the principle of uniform approximation can be applied to the two waves that are coupled, whose fields satisfy an embedded second order equation. It must further be shown that the solutions of this equation can be expressed in terms of an Airy Integral function, whose asymptotic forms, at points outside the Airy region, give the W.K.B. solutions of the two waves. The proof of this was given by Heading (1961) in an extremely important paper. Unfortunately, instead of a quartic equation, with two roots equal at the coupling point, he considered an equation of degree n with r roots equal (n and r are integers). Thus the importance of his work was obscured by the generality. A transcription of the theory for the ionospheric problem $n = 4$, $r = 2$, was given by Budden (1972).

The phase integral method has been successfully used to study vertically incident waves on an

anisotropic ionosphere (Cooper 1961, 1964; Altman 1965; Smith 1973; Maslin 1975). Its use for the general problem of oblique incidence on a lossy ionosphere is laborious because numerous complex solutions q of the Booker quartic must be computed when the phase integrals $k\int q\,\mathrm{d}z$ are evaluated on complex paths. This, however, is becoming easier with modern computers. Some form of the phase integral method will have to be used in studies of wave interaction, when the fields within the ionosphere have to be computed for frequencies in the long wave or medium wave bands.

9. Some recent developments

The simple form of the phase integral method fails if it is applied to a coupling or reflexion point that is too close to another. This raises the question of what happens when two coupling points move close together towards coalescence. The question has been examined by Budden & Smith (1974) who showed that conditions can approach coalescence in some cases that could occur in practice. Any extension of the phase integral method to deal with this would probably be very complicated and impractical. It would be easier to compute the full wave solutions. But the theory of coalescence is useful for providing an interpretation of some unexpected features of these solutions. Budden & Smith studied only two types of coalescence. They drew attention to several others whose further study would be very interesting.

There is a feature of the W.K.B. solutions that does not appear in the theory for isotropic media, but that must occur for the anisotropic case. It has some important influence on the phase integral solutions. It is the factor $\mathrm{e}^{\mathrm{i}\gamma}$ in (18). The other factor $\boldsymbol{e}_i(z)$ is a function of the parameters X, Y, Z of the ionospheric plasma at the point z, and does not depend on $\mathrm{d}X/\mathrm{d}z$, $\mathrm{d}Z/\mathrm{d}z$. But γ is more complicated. It can be written, if Y is assumed to be constant

$$\mathrm{i}\gamma = \int^z \Gamma_{ii}\,\mathrm{d}z = \int \{L(X,Z)\,\mathrm{d}X + M(X,Z)\,\mathrm{d}Z\}, \tag{52}$$

where L and M are known functions. The Γ_{ii} is the same as used by Budden & Clemmow (1957) and is a function of $\mathrm{d}X/\mathrm{d}z$, $\mathrm{d}Z/\mathrm{d}z$. Now the integrand in (52) is not a perfect differential so that γ is not expressible as a function of the local X and Z, but depends on the forms of the functions $X(z)$, $Z(z)$. This has been studied by Bennett (1974). In a loss free medium and at real heights γ is purely real and can affect only the real phase, although there are problems where this phase shift is important. More generally γ affects both the phase and amplitude. It is thus important in complex ray tracing. It influences the application of reciprocity and reversibility to rays in magnetoionic theory, and its presence requires that some of the conclusions of Budden & Jull (1964) must be modified. Its study comes within the province of ray theory and, when properly understood, may reveal a new physical phenomenon.

Another possible use of phase integral methods is in the study of mode conversion in a tapered waveguide (Budden 1975).

The author is particularly grateful to Professor J. Heading who, in correspondence over several years, has helped him to avoid rushing too far from the straight and narrow path into those complex regions where the angels fear to tread.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

References (Budden)

Altman, C. 1965 *J. Res. natn. Bur. Stds* **69**D, 511.
Bennett, J. A. 1974 *Proc. Inst. Elect. Electronic Engng* **62**, 1577.
Berry, M. V. & Mount, K. E. 1972 *Rep. Prog. Phys.* **35**, 313.
Booker, H. G. 1936 *Proc. R. Soc. Lond.* A **155**, 235.
Booker, H. G. 1938 *Phil. Trans. R. Soc. Lond.* A **237**, 411.
Booker, H. G. & Clemmow, P. C. 1950 *Proc. Instn Elect. Engrs* **97** III, 11.
Booker, H. G. & Walkinshaw, W. 1946 *Report on meteorological factors in radio wave propagation*, p. 80. London: Physical Society.
Bremmer, H. 1949 *Physica* **15**, 593.
Brillouin, L. 1936 *Rev. Gen d'Elec.* **40**, 227.
Budden, K. G. 1961*a* *Radio waves in the ionosphere*. Cambridge University Press.
Budden, K. G. 1961*b* *The waveguide mode theory of wave propagation*. London: Logos Press.
Budden, K. G. 1972 *J. atm. terr. Phys.* **34**, 1909.
Budden, K. G. 1975 *Math. Proc. Camb. Phil. Soc.* **77**, 567.
Budden, K. G. & Clemmow, P. C. 1957 *Proc. Camb. Phil. Soc.* **53**, 669.
Budden, K. G. & Cooper, Elisabeth A. 1962 *J. atm. terr. Phys.* **24**, 609.
Budden, K. G. & Jull, G. W. 1964 *Can. J. Phys.* **42**, 113.
Budden, K. G. & Smith, M. S. 1974 *Proc. R. Soc. Lond.* A **341**, 1.
Budden, K. G. & Terry, P. D. 1971 *Proc. R. Soc. Lond* A **321**, 275.
Chu, L. J. & Barrow, W. L. 1938 *Proc. Inst. Radio Engng* **26**, 1520.
Clemmow, P. C. 1966 *The plane wave spectrum representation of electromagnetic fields*. Oxford: Pergamon Press.
Clemmow, P. C. & Dougherty, J. P. 1969 *Electrodynamics of particles and plasmas*. Reading, Mass.: Addison-Wesley.
Clemmow, P. C. & Heading, J. 1954 *Proc. Camb. Phil. Soc.* **50**, 319.
Cooper, Elisabeth A. 1961 *J. atm. terr. Phys.* **22**, 122.
Cooper, Elisabeth A. 1964 *J. atm. terr. Phys.* **26**, 995.
Eckersley, T. L. 1931 *Proc. R. Soc. Lond.* A **132**, 83.
Eckersley, T. L. 1932*a* *J. Instn. Elect. Engrs* **71**, 405.
Eckersley, T. L. 1932*b* *Proc. R. Soc. Lond.* A **136**, 499.
Eckersley, T. L. 1932*c* *Proc. R. Soc. Lond.* A **137**, 158.
Eckersley, T. L. 1950 *Proc. Phys. Soc.* B **63**, 49.
Eden, R. J. 1967 *High energy collisions of elementary particles*. Cambridge University Press.
Ellis, G. R. A. 1956 *J. atm. terr. Phys.* **8**, 43.
Gans, R. 1915 *Ann. Phys. Lpz.* **47**, 709.
Heading, J. 1961 *J. Res. natr. Bur. Stds* **65**D, 595.
Heading, J. 1962 *An introduction to phase-integral methods*. London: Methuen.
Hirsch, P. & Shmoys, J. 1965 *Radio Sci., J. Res. natn. Bur. Stds* **69**D, 521.
Langer, R. E. 1937 *Phys. Rev.* **51**, 669.
Maslin, N. M. 1975 *Proc. R. Soc. Lond.* A **343**, 109.
Miller, J. C. P. 1946 *The Airy Integral* Brit. Ass. Math. tables, part Vol. B. Cambridge University Press.
Rayleigh, Lord 1912 *Proc. R. Soc. Lond.* A **86**, 207.
Schelkunoff, S. A. 1951 *Comm. pure appl. Math.* **4**, 117.
Smith, M. S. 1973 *Proc. R. Soc. Lond.* A **335**, 213.
Smith, M. S. 1974 *Proc. R. Soc. Lond.* A **336**, 229.
Stokes, Sir G. G. 1857 *Trans. Camb. Phil. Soc.* **10**, 106.
Watson, G. N. 1919*a* *Proc. R. Soc. Lond.* A **95**, 83.
Watson, G. N. 1919*b* *Proc. R. Soc. Lond.* A **95**, 546.
Whittaker, E. T. & Watson, G. N. 1927 *Modern analysis*. Cambridge University Press.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES