4 Electromagnetic wave propagation in dielectrics

4.1 Introduction

It is easily demonstrated that the fields associated with an electromagnetic wave propagating through a uniform dielectric medium of dielectric constant ϵ satisfy

$$\left(\frac{\epsilon}{c^2}\frac{\partial^2}{\partial t^2} - \nabla^2\right)\boldsymbol{E} = 0, \tag{4.1}$$

and

$$\nabla \wedge \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t}. \tag{4.2}$$

The plane wave solutions to these equations are well known:

$$\boldsymbol{E} = \boldsymbol{E}_0\, \mathrm{e}^{\mathrm{i}\,(\boldsymbol{k}\cdot\boldsymbol{r}-\omega t)}, \tag{4.3a}$$

$$\boldsymbol{B} = \boldsymbol{B}_0\, \mathrm{e}^{\mathrm{i}\,(\boldsymbol{k}\cdot\boldsymbol{r}-\omega t)}, \tag{4.3b}$$

where $\boldsymbol{E}_0$ and $\boldsymbol{B}_0$ are constant vectors, with

$$\frac{\omega^2}{k^2} = \frac{c^2}{\epsilon}, \tag{4.4}$$

and

$$\boldsymbol{B}_0 = \frac{\boldsymbol{k} \wedge \boldsymbol{E}_0}{\omega}. \tag{4.5}$$

The phase velocity of the wave is given by

$$v = \frac{\omega}{k} = \frac{c}{n}, \tag{4.6}$$

where

$$n = \sqrt{\epsilon} \tag{4.7}$$

is called the *refractive index* of the medium. It is clear that an electromagnetic wave propagates with a phase velocity which is slower than the velocity of light in a conventional (*i.e.*, ϵ real and greater than unity) dielectric medium.

In some dielectric media ϵ is complex. This leads, from Eq. (4.4), to a complex wave vector $\boldsymbol{k}$. For a wave propagating in the x-direction we obtain

$$\boldsymbol{E} = \boldsymbol{E}_0 \exp[\,\mathrm{i}\,(\mathrm{Re}(k)\,x - \omega t)]\exp[-\mathrm{Im}(k)\,x]. \tag{4.8}$$

Thus, a complex dielectric constant leads to the attenuation (or amplification) of the wave as it propagates through the medium in question.

Up to now, we have tacitly assumed that ϵ is the same for waves of all frequencies. In practice, this is not the case. In dielectric media ϵ is, in general, complex, and varies (in some cases, strongly) with the wave frequency, ω. Thus, waves of different frequencies propagate through a dielectric medium with different phase velocities. This phenomenon is known as *dispersion*. Moreover, there may exist frequency bands in which the waves are attenuated (*i.e.*, absorbed). All of this makes the problem of determining the behaviour of a wave packet as it propagates through a dielectric medium far from straightforward. Recall, that the solution to this problem for a wave packet traveling through a vacuum is fairly trivial. The packet propagates at the velocity c without changing its shape. What is the equivalent result for the case of a dielectric medium? This is an important question, since nearly all of our information regarding the universe is obtained from the study of electromagnetic waves emitted by distant objects. All of these waves have to propagate through dispersive media (*e.g.*, the interstellar medium, the ionosphere, the atmosphere) before reaching us. It is, therefore, vitally important that we understand which aspects of these wave signals are predominantly determined by the wave sources, and which are strongly modified by the dispersive media through which they have propagated in order to reach us.

The study of wave propagation through dispersive media was pioneered by two scientists, Arnold Sommerfeld and Léon Brillouin, during the first half of this century. In the following discussion, we shall stick as close as possible to Sommerfeld and Brillouin's original analysis.

4.2 The form of the dielectric constant

Let us investigate an electromagnetic wave propagating through a transparent, isotropic, non-conducting, medium. The electric displacement inside the medium

is given by

$$\boldsymbol{D} = \epsilon_0 \boldsymbol{E} + \boldsymbol{P}, \tag{4.9}$$

where $\boldsymbol{P}$ is the electric polarization. Since electrons are much lighter than ions (or atomic nuclei), we would expect the former to displace further than the latter under the influence of an electric field. Thus, to a first approximation the polarization $\boldsymbol{P}$ is determined by the electron response to the wave. Suppose that the electrons displace a distance $\boldsymbol{s}$ from their rest positions in the presence of the wave. It follows that

$$\boldsymbol{P} = -N e\, \boldsymbol{s}, \tag{4.10}$$

where N is the number density of electrons.

Let us assume that the electrons are bound "quasi-elastically" to their rest positions, so that they seek to return to these positions when displaced from them by a field $\boldsymbol{E}$. It follows that $\boldsymbol{s}$ satisfies the differential equation of the form

$$m\, \ddot{\boldsymbol{s}} + f\, \boldsymbol{s} = -e \boldsymbol{E}, \tag{4.11}$$

where m is the electron mass, $-f\,\boldsymbol{s}$ is the restoring force, and $\dot{}$ denotes a partial derivative with respect to time. The above equation can also be written

$$\ddot{\boldsymbol{s}} + g\, \omega_0\, \dot{\boldsymbol{s}} + {\omega_0}^2 \boldsymbol{s} = -\frac{e}{m}\, \boldsymbol{E}, \tag{4.12}$$

where

$${\omega_0}^2 = \frac{f}{m} \tag{4.13}$$

is the characteristic oscillation frequency of the electrons. In almost all dielectric media this frequency lies in the far *ultraviolet* region of the electromagnetic spectrum. In Eq. (4.12) we have added a phenomenological damping term $g\,\omega_0\,\dot{\boldsymbol{s}}$, in order to take into account the fact that an electron excited by an impulsive electric field does not oscillate for ever. In general, however, electrons in dielectric media can be regarded as high-Q oscillators, which is another way of saying that the dimensionless damping constant g is typically much less than unity. Thus, an electron "rings" for a long time after being excited by an impulse.

Let us assume that the electrons oscillate in sympathy with the wave at the wave frequency ω. It follows from Eq. (4.12) that

$$\boldsymbol{s} = -\frac{(e/m)\,\boldsymbol{E}}{{\omega_0}^2 - \omega^2 - \mathrm{i}\,g\,\omega\,\omega_0}. \tag{4.14}$$

Note that we have neglected the response of the electrons to the magnetic component of the the wave. It is easily demonstrated that this is a good approximation provided that the electrons do not oscillate with relativistic velocities (*i.e.*, provided that the amplitude of the wave is sufficiently small). Thus, Eq. (4.10) yields

$$\boldsymbol{P} = \frac{(Ne^2/m)\,\boldsymbol{E}}{{\omega_0}^2 - \omega^2 - \mathrm{i}\,g\,\omega\,\omega_0}. \tag{4.15}$$

Since, by definition,

$$\boldsymbol{D} = \epsilon_0 \epsilon\,\boldsymbol{E} = \epsilon_0 \boldsymbol{E} + \boldsymbol{P}, \tag{4.16}$$

it follows that

$$\epsilon(\omega) \equiv n^2(\omega) = 1 + \frac{(Ne^2/\epsilon_0 m)}{{\omega_0}^2 - \omega^2 - \mathrm{i}\,g\,\omega\,\omega_0}. \tag{4.17}$$

Thus, the index of refraction is frequency dependent. Since ω_0 typically lies in the ultraviolet region of the spectrum (and since $g \ll 1$), it is clear that the denominator ${\omega_0}^2 - \omega^2 - \mathrm{i}\,g\,\omega\,\omega_0 \simeq {\omega_0}^2 - \omega^2$ is positive in the entire visible spectrum, and is larger at the red end than at the blue end. This implies that *blue light is refracted more than red light.* This is normal dispersion. Incidentally, an expression, like the above, which specifies the dispersion of waves propagating through some dielectric medium is usually called a *dispersion relation.*

Let us now suppose that there are N molecules per unit volume with Z electrons per molecule, and that instead of a single oscillation frequency for all electrons, there are f_i electrons per molecule with oscillation frequency ω_i and damping constant g_i. It is easily demonstrated that

$$n^2(\omega) = 1 + \frac{Ne^2}{\epsilon_0 m} \sum_i \frac{f_i}{{\omega_i}^2 - \omega^2 - \mathrm{i}\,g_i\,\omega\,\omega_i}, \tag{4.18}$$

where the *oscillator strengths* f_i satisfy the sum rule,

$$\sum_i f_i = Z. \tag{4.19}$$

A more exact quantum mechanical treatment of the response of an atom, or molecule, to a low amplitude electromagnetic wave also leads to a dispersion relation of the form (4.18), except that the quantities f_i, ω_i, and g_i can, in principle, be calculated from first principles. In practice, this is too difficult except for the very simplest cases.

Since the damping constants g_i are generally small compared to unity, it follows from Eq. (4.18) that $n(\omega)$ is a predominately real quantity at most wave frequencies. The factor $({\omega_i}^2 - \omega^2)^{-1}$ is positive for $\omega < \omega_i$ and negative for $\omega > \omega_i$. Thus, at low frequencies, below the smallest ω_i, all of the terms in the sum in (4.18) are positive, and $n(\omega)$ is consequently greater than unity. As ω is raised so that it passes successive ω_i values, more and more negative terms occur in the sum, until eventually the whole sum is negative and $n(\omega)$ is less than unity. Thus, at very high frequencies electromagnetic waves propagate through dielectric media with phase velocities which exceed the velocity of light in a vacuum. For $\omega \simeq \omega_i$, Eq. (4.18) predicts a rather violent variation of the refractive index with frequency. Let us examine this phenomenon more closely.

4.3 Anomalous dispersion and resonant absorption

When ω is approximately equal to ω_i the dispersion relation (4.18) reduces to

$$n^2 = {n_i}^2 + \frac{N e^2 f_i/\epsilon_0 m}{{\omega_i}^2 - \omega^2 - \mathrm{i}\, g_i\, \omega\, \omega_i}, \tag{4.20}$$

where n_i is the average contribution in the vicinity of $\omega = \omega_i$ of all other resonances (also included in n_i is the contribution 1 of the vacuum displacement current, which was previously written down separately). The refractive index is clearly complex. For a wave propagating in the x-direction

$$\boldsymbol{E} = \boldsymbol{E}_0\, \exp[\mathrm{i}\,(\omega/c)(\mathrm{Re}(n)\, x - c\, t)]\, \exp[-(\omega/c)\, \mathrm{Im}(n)\, x]. \tag{4.21}$$

Thus, the phase velocity of the wave is determined by the real part of the refractive index via

$$v = \frac{c}{\mathrm{Re}(n)}. \tag{4.22}$$

Note that a positive imaginary component of the refractive index leads to the attenuation of the wave as it propagates.

Let

$$a^2 = \frac{N e^2 f_i}{\epsilon_0 m {\omega_i}^2}, \tag{4.23a}$$

$$x = \frac{\omega^2 - {\omega_i}^2}{{\omega_i}^2}, \tag{4.23b}$$

$$y = \frac{\mathrm{Re}(n)^2 - \mathrm{Im}(n)^2}{a^2}, \tag{4.23c}$$

$$z = \frac{2\,\mathrm{Re}(n)\,\mathrm{Im}(n)}{a^2}, \tag{4.23d}$$

where a, x, y, z are all dimensionless quantities. It follows from Eq. (4.20) that

$$y = \frac{{n_i}^2}{a^2} - \frac{x}{x^2 + {g_i}^2(1+x)}, \tag{4.24a}$$

$$z = \frac{g_i\sqrt{1+x}}{x^2 + {g_i}^2(1+x)}. \tag{4.24b}$$

Let us adopt the physical ordering $g_i \ll 1$. The extrema of the function y occur at $x = \pm g_i$. It is easily demonstrated that

$$y_{\min} = y(x = g_i) = \frac{{n_i}^2}{a^2} - \frac{1}{2g_i}, \tag{4.24c}$$

$$y_{\max} = y(x = -g_i) = \frac{{n_i}^2}{a^2} + \frac{1}{2g_i}. \tag{4.24d}$$

The maximum value of the function z occurs at $x = 0$. In fact,

$$z_{\max} = \frac{1}{g_i}. \tag{4.25}$$

Note that

$$z(x = \pm g_i) = \frac{1}{2g_i}. \tag{4.26}$$

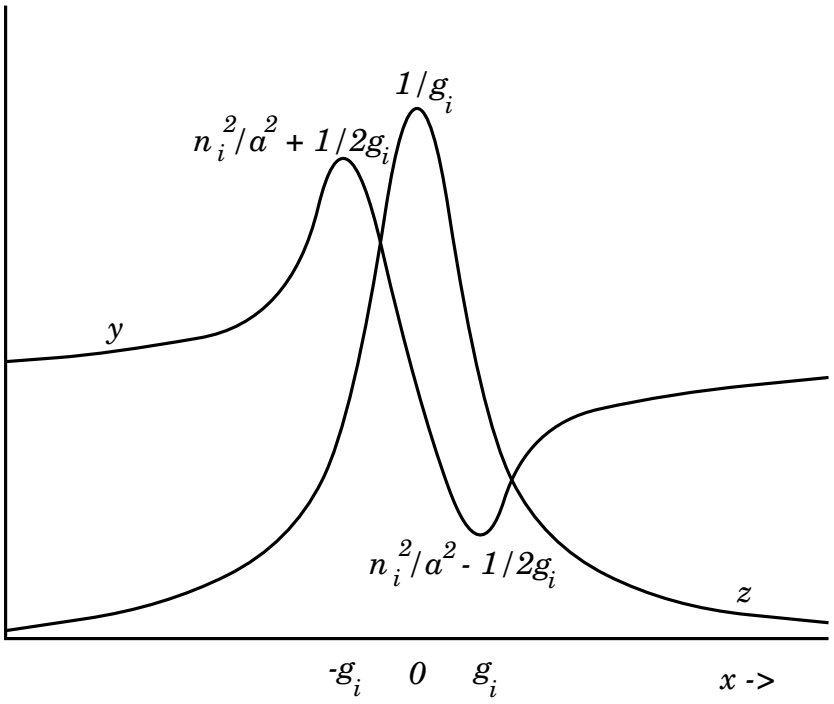

Figure 5: Sketch of the variation of the functions y and z with x

Figure 5 shows a sketch of the variation of the functions y and z with x. These curves are also indicative of the variation of $\mathrm{Re}(n)$ and $\mathrm{Im}(n)$, respectively, with frequency ω in the vicinity of the resonant frequency ω_i. Recall that normal dispersion is associated with an increase in $\mathrm{Re}(n)$ with increasing ω. The reverse situation is termed *anomalous dispersion*. It is clear from the figure that normal dispersion occurs everywhere except in the immediate neighbourhood of the resonant frequency ω_i. It is also clear that the imaginary part of the refractive index is only appreciable in those regions of the electromagnetic spectrum where anomalous dispersion takes place. A positive imaginary component of the refractive index implies that the wave is absorbed as it propagates through the medium, so the regions of the spectrum where $\mathrm{Im}(n)$ is appreciable are called regions of *resonant absorption*. Anomalous dispersion and resonant absorption take place in the vicinity of the ith resonance when $|\omega - \omega_i| \lesssim O(g_i)$. Since the damping constants g_i are, in practice, very small compared to unity, the regions of the spectrum in which resonant absorption takes place are strongly localized in the vicinity of the various resonant frequencies.

The dispersion relation (4.18) only takes electron resonances into account. Of course, there are also resonances associated with displacements of the ions (or atomic nuclei). The off-resonance contributions to the right-hand side of Eq. (4.18) from the ions are smaller than those from the electrons by a factor of order m/M (where M is a typical ion mass). Nevertheless, the ion contributions are important because they give rise to anomalous dispersion and resonant absorption close to the ion resonant frequencies. The ion resonances associated with the stretching and bending of molecular bonds typically lie in the infrared region of the electromagnetic spectrum. Those associated with molecular rotation (these resonances only affect the dispersion relation if the molecule is polar) occur in the microwave region of the spectrum. Thus, both air and water exhibit strong resonant absorption of electromagnetic waves in both the ultraviolet and infrared regions of the spectrum. In the first case this is due to electron resonances, and in the second to ion resonances. The visible region of the spectrum exists as a narrow window lying between these two regions in which there is comparatively little attenuation of electromagnetic waves.

4.4 Wave propagation through a conducting medium

In the limit $\omega \rightarrow 0$, there is a significant difference in the response of a dielectric medium, depending on whether the lowest resonant frequency is zero or non-zero. For insulators the lowest resonant frequency is different from zero. In this case, the low frequency refractive index is predominately real, and is also greater than unity. Suppose, however, that some fraction f_0 of the electrons are "free," in the sense of having $\omega_0 = 0$. In this situation, the low frequency dielectric constant takes the form

$$\epsilon(\omega) \equiv n^2(\omega) = {n_0}^2 + \mathrm{i}\,\frac{N e^2}{\epsilon_0 m}\frac{f_0}{\omega\,(\gamma_0 - \mathrm{i}\,\omega)}, \tag{4.27}$$

where n_0 is the contribution to the refractive index from all of the other resonances, and $\gamma_0 = \lim_{\omega_0 \rightarrow 0} g_0\,\omega_0$. Note that for a conducting medium the contribution to the refractive index from the free electrons is singular at $\omega = 0$. This singular behaviour can be explained as follows. Consider the Ampère-Maxwell

equation

$$\nabla \wedge \boldsymbol{B} = \mu_0 \left(\boldsymbol{j}_t + \frac{\partial \boldsymbol{D}}{\partial t} \right). \tag{4.28}$$

Let us assume that the medium in question obeys Ohm's law, $\boldsymbol{j}_t = \sigma \boldsymbol{E}$, and has a "normal" dielectric constant ${n_0}^2$. Here, σ is the conductivity. Assuming an $\exp(-\mathrm{i}\,\omega t)$ time dependence of all field quantities the above equation yields

$$\frac{\nabla \wedge \boldsymbol{B}}{\mu_0} = -\mathrm{i}\,\epsilon_0\,\omega \left({n_0}^2 + \mathrm{i}\,\frac{\sigma}{\epsilon_0\,\omega} \right) \boldsymbol{E}. \tag{4.29}$$

Suppose, however, that we do not explicitly use Ohm's law but, instead, attribute all of the properties of the medium to the dielectric constant. In this case, the effective dielectric constant of the medium is equivalent to the term in round brackets on the right-hand side of the above equation. Thus,

$$\epsilon(\omega) \equiv n^2(\omega) = {n_0}^2 + \mathrm{i}\,\frac{\sigma}{\epsilon_0\,\omega}. \tag{4.30}$$

A comparison of this term with Eq. (4.27) yields the following expression for the conductivity

$$\sigma = \frac{f_0 N e^2}{m(\gamma_0 - \mathrm{i}\,\omega)}. \tag{4.31}$$

Thus, at low frequencies conductors possess predominately real conductivities (*i.e.*, the current remains in phase with the electric field). However, at higher frequencies the conductivity becomes complex. At these sorts of frequencies there is little meaningful distinction between a conductor and an insulator, since the "conductivity" contribution to $\epsilon(\omega)$ appears as a resonant amplitude just like the other contributions. For a good conductor, such as Copper, the conductivity remains predominately real for all frequencies up to and including those in the microwave region of the electromagnetic spectrum.

The conventional way in which to represent the complex refractive index of a conducting medium (in the low frequency limit) is to write it in terms of a real "normal" dielectric constant, $\epsilon = {n_0}^2$, and a real conductivity, σ. Thus, from Eq. (4.30)

$$n^2(\omega) = \epsilon + \mathrm{i}\,\frac{\sigma}{\epsilon_0\,\omega}. \tag{4.32}$$

For a poor conductor ($\sigma/\epsilon\,\epsilon_0\,\omega \ll 1$) we find

$$k = n\,\frac{\omega}{c} \simeq \sqrt{\epsilon}\,\frac{\omega}{c} + \mathrm{i}\,\frac{\sigma}{2\sqrt{\epsilon}\,\epsilon_0\,c}. \tag{4.33}$$

In this limit $\mathrm{Re}(k) \gg \mathrm{Im}(k)$, and the attenuation of the wave, which is governed by $\mathrm{Im}(k)$ [see Eq. (4.8)], is independent of the frequency. Thus, for a poor conductor the wave is basically the same as a wave propagating through a conventional dielectric with dielectric constant ϵ, except that the wave attenuates gradually over a distance of very many wavelengths. For a good conductor ($\sigma/\epsilon\,\epsilon_0\,\omega \gg 1$)

$$k \simeq e^{\mathrm{i}\,\pi/4}\sqrt{\mu_0\,\sigma\,\omega}. \tag{4.34}$$

It follows from Eq. (4.5) that

$$\frac{cB_0}{E_0} = \frac{kc}{\omega} = e^{\mathrm{i}\,\pi/4}\sqrt{\frac{\sigma}{\epsilon_0\,\omega}}. \tag{4.35}$$

Thus, the phase of the magnetic field *lags* that of the electric field by 45°. Moreover, the magnitude of cB_0 is much larger than that of E_0 (since $\sigma/\epsilon_0\,\omega \gg \epsilon \gtrsim 1$). It follows that the field energy is almost entirely magnetic in nature. It is clear that an electromagnetic wave propagating through a good conductor has markedly different properties to a wave propagating through a conventional dielectric. For a wave propagating in the x-direction, the amplitudes of the electric and magnetic fields attenuate like $\exp(-x/d)$, where

$$d = \sqrt{\frac{2}{\mu_0\,\sigma\,\omega}}. \tag{4.36}$$

This quantity is known as the *skin depth*. It is clear that an electromagnetic wave incident on a conducting medium will not penetrate more than a few skin depths into that medium.

4.5 The high frequency limit

Consider the behaviour of the dispersion relation (4.18) in the high frequency limit $\omega \gg \omega_i$ (for all i). In this limit, the relation simplifies considerably to give

$$n^2(\omega) = 1 - \frac{{\omega_p}^2}{\omega^2}, \tag{4.37}$$

where the quantity

$$\omega_p = \sqrt{\frac{NZe^2}{\epsilon_0 m}} \tag{4.38}$$

is called the *plasma frequency*. The wave-number in the high frequency limit is given by

$$k = n\,\frac{\omega}{c} = \frac{\sqrt{\omega^2 - {\omega_p}^2}}{c}. \tag{4.39}$$

This expression is only valid in dielectrics when $\omega \gg \omega_p$. Thus, the refractive index is real and slightly less than unity, giving waves which propagate without attenuation with a phase velocity slightly larger than the velocity of light in vacuum. However, in certain ionized media (in particular, in tenuous plasmas such as occur in the ionosphere) the electrons are free and the damping is negligible. In this case, Eqs. (4.37) and (4.39) are valid even when $\omega < \omega_p$. It is clear that a wave can only propagate through a tenuous plasma if its frequency exceeds the plasma frequency (in which case it has a real wave-number). If wave frequency is less than the plasma frequency then the wave-number is purely imaginary, according to Eq. (4.39), and the wave is therefore attenuated. This accounts for the fact that long-wave and medium-wave radio signals can be received even when the transmitter lies over the horizon. The frequency of these waves is less than the plasma frequency of the ionosphere, which reflects them, so they are trapped between the ionosphere and the surface of the Earth (which is also a good reflector of radio waves), and can, in certain cases, travel many times around the Earth before being attenuated. Unfortunately, this scheme does not work very well for medium-wave signals at night. The problem is that the plasma frequency of the ionosphere is proportional to the square root of the number density of free ionospheric electrons. These free electrons are generated through the ionization of neutral molecules by ultraviolet radiation from the Sun. Of course, there is no radiation from the Sun at night so the density of free electrons starts to drop as the electrons gradually recombine with ions in the ionosphere. Eventually, the plasma frequency of the ionosphere falls below the frequency of medium-wave radio signals allowing them to be transmitted through the ionosphere into outer space. The ionosphere appears almost completely transparent to high frequency signals such as TV and FM radio signals. Thus, this type of signal is not reflected

by the ionosphere. Consequently, to receive such signals it is necessary to be in the line of sight of the relevant transmitter.

4.6 Faraday rotation

The electromagnetic force acting on an electron is given by

$$\boldsymbol{f} = -e\,(\boldsymbol{E} + \boldsymbol{v} \wedge \boldsymbol{B}). \tag{4.40}$$

If the $\boldsymbol{E}$ and $\boldsymbol{B}$ fields in question are due to an electromagnetic wave propagating through a dielectric medium then

$$|B| = \frac{n}{c}\,|E|. \tag{4.41}$$

It follows that the ratio of the magnetic to the electric forces acting on the electron is $n\,v/c$. In other words, the magnetic force is completely negligible unless the wave amplitude is sufficiently high that the electron moves relativistically in response to the wave. This state of affairs is rare, but can occur when intense laser beams are made to propagate through plasmas.

Suppose, however, that the dielectric medium contains an externally generated magnetic field $\boldsymbol{B}$. This can easily be made much stronger than the optical magnetic field. In this case, it is possible for a magnetic field to affect the propagation of low amplitude electromagnetic waves. The electron equation of motion (4.11) generalizes to

$$m\,\ddot{\boldsymbol{s}} + f\,\boldsymbol{s} = -e(\boldsymbol{E} + \dot{\boldsymbol{s}} \wedge \boldsymbol{B}), \tag{4.42}$$

where any damping of the motion has been neglected. Suppose that the direction of $\boldsymbol{B}$ is in the positive z-direction, and that the wave propagates in the same direction. With these assumptions the $\boldsymbol{E}$ and $\boldsymbol{s}$ vectors lie in the x-y plane. The above equation reduces to

$$(\omega_0{}^2 - \omega^2)\,s_x - \mathrm{i}\,\omega\Omega\,s_y = -\frac{e}{m}\,E_x, \tag{4.43a}$$

$$(\omega_0{}^2 - \omega^2)\,s_y + \mathrm{i}\,\omega\Omega\,s_x = -\frac{e}{m}\,E_y, \tag{4.43b}$$

provided that all perturbed quantities have an $\exp(-\mathrm{i}\,\omega t)$ time dependence. Here,

$$\Omega = \frac{eB}{m} \tag{4.44}$$

is the electron cyclotron frequency. Let

$$E_\pm = E_x \pm \mathrm{i}\, E_y, \tag{4.45}$$

and

$$s_\pm = s_x \pm \mathrm{i}\, s_y. \tag{4.46}$$

Note that

$$E_x = \frac{1}{2}\,(E_+ + E_-), \tag{4.47a}$$

$$E_y = \frac{1}{2\,\mathrm{i}}\,(E_+ - E_-). \tag{4.47b}$$

Equations (4.43) reduce to

$$({\omega_0}^2 - \omega^2 - \omega\,\Omega)\, s_+ = -\frac{e}{m}\, E_+, \tag{4.48a}$$

$$({\omega_0}^2 - \omega^2 + \omega\,\Omega)\, s_- = -\frac{e}{m}\, E_-. \tag{4.48b}$$

Defining $P_\pm = P_x \pm \mathrm{i}\, P_y$, it follows from Eq. (4.10) that

$$P_\pm = \frac{(N e^2/m)\, E_\pm}{{\omega_0}^2 - \omega^2 \mp \omega\,\Omega}. \tag{4.49}$$

Finally, from Eq. (4.15), we can write

$$\epsilon_\pm \equiv {n_\pm}^2 = 1 + \frac{P_\pm}{\epsilon_0\, E_\pm}, \tag{4.50}$$

giving

$${n_\pm}^2(\omega) = 1 + \frac{(N e^2/\epsilon_0 m)}{{\omega_0}^2 - \omega^2 \mp \omega\,\Omega}. \tag{4.51}$$

According to the dispersion relation (4.51), the refractive index of a magnetized dielectric medium can take one of two possible values, which presumably correspond to two different types of wave propagating along the z-axis. The first wave has the refractive index n_+ and an associated electric field [see Eqs. (4.45)]

$$E_x = E_0 \cos[(\omega/c)(n_+ z - ct)], \tag{4.52a}$$

$$E_y = E_0 \sin[(\omega/c)(n_+ z - ct)]. \tag{4.52b}$$

This corresponds to a *left-handed circularly polarized wave* propagating in the z-direction with the phase velocity c/n_+. The second wave has the refractive index n_- and an associated electric field

$$E_x = E_0 \cos[(\omega/c)(n_- z - ct)], \tag{4.53a}$$

$$E_y = -E_0 \sin[(\omega/c)(n_- z - ct)]. \tag{4.53b}$$

This corresponds to a *right-handed circularly polarized wave* propagating in the z-direction with the phase velocity c/n_-. It is clear from Eq. (4.51) that $n_+ > n_-$. Thus, we conclude that in the presence of a z-directed magnetic field, a z-directed left-handed circularly polarized wave propagates with a phase velocity which is slightly *less* than that of the corresponding right-handed wave. It should be remarked that the refractive index is always real (in the absence of damping), so the magnetic field gives rise to no net absorption of electromagnetic radiation. This is not surprising since the magnetic field *does no work* on charged particles, and can therefore transfer no energy to the particles from any waves propagating through the medium.

We have seen that right-handed and left-handed circularly polarized waves propagate with different phase velocities through a magnetized dielectric medium. But, what does this imply for the propagation of a plane polarized wave? Let us superimpose the left-handed wave whose electric field is given by Eqs. (4.52) on the right-handed wave whose electric field is given by Eqs. (4.53). In the absence of a magnetic field $n_+ = n_- = n$, and we obtain

$$E_x = 2E_0 \cos[(\omega/c)(nz - ct)], \tag{4.54a}$$

$$E_y = 0. \tag{4.54b}$$

This, of course, is the field of a plane polarized wave (polarized along the x-direction) propagating along the z-axis with the phase velocity c/n. In the presence of a magnetic field we obtain

$$E_x = 2E_0 \cos[(\omega/c)(nz - ct)] \cos[(\omega/2c)(n_+ - n_-)z], \tag{4.55a}$$

$$E_y = 2E_0 \cos[(\omega/c)(nz - ct)] \sin[(\omega/2c)(n_+ - n_-)z], \tag{4.55b}$$

where

$$n = \frac{1}{2}(n_+ + n_-) \tag{4.56}$$

is the mean index of refraction. Equations (4.55) can be recognized as the field of a plane polarized wave whose angle of polarization with respect to the x-axis,

$$\chi = \tan^{-1}(E_y/E_x), \tag{4.57}$$

rotates as the wave propagates along the z-axis with the phase velocity c/n. In fact, the angle of polarization is given by

$$\chi = \frac{\omega}{2c}(n_+ - n_-)\, z, \tag{4.58}$$

which clearly increases linearly with the distance traveled by the wave along the direction of the magnetic field. This rotation of the plane of polarization of a linearly polarized wave propagating through a magnetized dielectric medium is known as *Faraday rotation* (since it was discovered by Michael Faraday in 1845).

Assuming that the cyclotron frequency Ω is relatively small compared to the wave frequency ω, and also that ω does not lie close to the resonant frequency ω_0, it is easily demonstrated that

$$n \simeq 1 + \frac{(Ne^2/\epsilon_0 m)}{{\omega_0}^2 - \omega^2}, \tag{4.59}$$

and

$$n_+ - n_- \simeq \frac{Ne^2}{\epsilon_0 m\, n} \frac{\omega\Omega}{({\omega_0}^2 - \omega^2)^2}. \tag{4.60}$$

It follows that the rate at which the plane of polarization of an electromagnetic wave rotates with the distance traveled by the wave is given by

$$\frac{d\chi}{dl} = \frac{\kappa(\omega)\, N B_\parallel}{n(\omega)}, \tag{4.61}$$

where $B_\parallel$ is the component of the magnetic field along the direction of propagation of the wave, and

$$\kappa(\omega) = \frac{e^3}{2\epsilon_0 m^2 c}\frac{\omega^2}{(\omega_0^2-\omega^2)^2}. \tag{4.62}$$

If the medium in question is a tenuous plasma then $n \simeq 1$ and $\omega_0 = 0$. Thus,

$$\frac{d\chi}{dl} \simeq \frac{e^3}{2\epsilon_0 m^2 c}\frac{N B_\parallel}{\omega^2} \tag{4.63}$$

Clearly, the rate at which the plane of polarization rotates is proportional to the product of the electron number density and the parallel magnetic field strength. Moreover, the plane of rotation rotates faster for low frequency waves than for high frequency waves. The total angle by which the plane of polarization is twisted after passing through a magnetized plasma is given by

$$\Delta\chi \simeq \frac{e^3}{2\epsilon_0 m^2 c\,\omega^2}\int N(l) B_\parallel(l)\, dl, \tag{4.64}$$

provided that N and $B_\parallel$ vary on length-scales which are large compared to the wavelength of the radiation. This formula is regularly employed in radio astronomy to infer the magnetic field-strength in interstellar space.

4.7 Wave propagation through a magnetized plasma

For a plasma ($\omega_0 = 0$) the dispersion relation (4.51) reduces to

$$n_\pm^2(\omega) = 1 - \frac{{\omega_p}^2}{\omega(\omega \mp \Omega)}. \tag{4.65}$$

The upper sign corresponds to a left-handed circularly polarized wave and the lower sign to a right-handed polarized wave. Of course, Eq. (4.65) is only valid for wave propagation along the direction of the magnetic field. Wave propagation through the Earth's ionosphere is well described by the above dispersion relation. There are wide frequency intervals where one of n_+^2 or n_-^2 is positive and the other negative. At such frequencies one state of circular polarization cannot propagate

through the plasma. Consequently, a wave of that polarization incident on the plasma is totally reflected. The other state of polarization is partially transmitted.

The behaviour of $n_-^2(\omega)$ at low frequencies is responsible for a strange phenomenon known to radio hams as "whistlers." As the frequency tends to zero, Eq. (4.65) yields

$$n_-^2 \simeq \frac{\omega_p^2}{\omega\,\Omega}. \tag{4.66}$$

At this sort of frequency n_+^2 is negative, so only right-hand polarized waves can propagate. The wave-number of such waves is given by

$$k_- = n_-\,\frac{\omega}{c} \simeq \frac{\omega_p}{c}\sqrt{\frac{\omega}{\Omega}}. \tag{4.67}$$

Energy transport is governed by the *group velocity* (see later)

$$v_g(\omega) = \frac{d\omega}{dk_-} \simeq 2c\,\frac{\sqrt{\omega\,\Omega}}{\omega_p}. \tag{4.68}$$

Thus, low frequency waves transmit energy *slower* than high frequency waves. A lightning strike in one hemisphere of the Earth generates a wide spectrum of radiation, some of which propagates along the dipolar field lines of the Earth's magnetic field in a manner described approximately by the dispersion relation (4.68). The high frequency components of the signal return to the surface of the Earth before the low frequency components (since they travel faster along the magnetic field). This gives rise to a radio signal which begins at a high frequency and then "whistles" down to lower frequencies.

4.8 The propagation of electromagnetic radiation through a dispersive medium

Let us now investigate the propagation of electromagnetic radiation through a dispersive medium by studying a simple one-dimensional problem. Suppose that our dispersive medium extends from $x = 0$, where it interfaces with a vacuum, to $x = \infty$. Suppose further that a wave is incident *normally* on the medium,

so that the field quantities only depend on x and t. The wave is specified as a given function of t at $x = 0$. Since we are not interested in the reflected wave, let this function, $f(t)$, say, give the wave amplitude *just inside* the surface of the dispersive medium. Suppose that the wave arrives at this surface at $t = 0$, and that

$$f(t) = \begin{cases} 0 & \text{for } t < 0, \\ \sin\left(\frac{2\pi t}{\tau}\right) & \text{for } t \geq 0. \end{cases} \tag{4.69}$$

How does the wave subsequently develop in the region $x > 0$? In order to answer this question we must first of all decompose $f(t)$ into harmonic components of the form $\exp(-\mathrm{i}\,\omega t)$ (*i.e.*, Fourier harmonics). Unfortunately, if we attempt this using only real frequencies, ω, we encounter convergence difficulties, since $f(t)$ does not vanish at $t = \infty$. For the moment, we can circumvent these difficulties by only considering *finite* (in time) wave forms. In other words, we now imagine that $f(t) = 0$ for $t < 0$ *and* $t > T$. Such a wave form can be thought of as the superposition of two infinite (in time) wave forms, the first beginning at $t = 0$ and the second at $t = T$ with the opposite phase, so that the two cancel for all time $t > T$.

According to standard Fourier transform theory

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} d\omega \int_{-\infty}^{\infty} f(t')\, \mathrm{e}^{-\mathrm{i}\,\omega(t-t')}\, dt'. \tag{4.70}$$

Since $f(t)$ is a real function of t which is zero for $t < 0$ and $t > T$, we can write

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} d\omega \int_{0}^{T} f(t')\, \cos[\omega(t-t')\,]\, dt'. \tag{4.71}$$

Finally, it follows from symmetry (in ω) that

$$f(t) = \frac{1}{\pi} \int_{0}^{\infty} d\omega \int_{0}^{T} f(t')\, \cos[\omega(t-t')\,]\, dt'. \tag{4.72}$$

Equation (4.69) yields

$$f(t) = \frac{1}{\pi} \int_{0}^{\infty} d\omega \int_{0}^{T} \sin\left(\frac{2\pi t'}{\tau}\right) \cos[\omega(t-t')\,]\, dt', \tag{4.73}$$

or

$$f(t) = \frac{1}{2\pi}\int_0^\infty d\omega \left\{ \frac{\cos[2\pi t'/\tau + \omega(t-t')\,]}{\omega - 2\pi/\tau} - \frac{\cos[2\pi t'/\tau - \omega(t-t')\,]}{\omega + 2\pi/\tau} \right\}_{t'=0}^{t'=T}. \tag{4.74}$$

Let us assume, for the sake of simplicity, that

$$T = N\tau, \tag{4.75}$$

where N is a positive integer. This ensures that $f(t)$ is continuous at $t = T$. Equation (4.74) reduces to

$$f(t) = \frac{2}{\tau}\int_0^\infty \frac{d\omega}{\omega^2 - (2\pi/\tau)^2}\,(\,\cos[\omega(t-T)\,] - \cos\omega t\,)\,. \tag{4.76}$$

This expression can be written

$$f(t) = \frac{1}{\tau}\int_{-\infty}^\infty \frac{d\omega}{\omega^2 - (2\pi/\tau)^2}\,(\,\cos[\omega(t-T)\,] - \cos\omega t\,)\,, \tag{4.77}$$

or

$$f(t) = \frac{1}{2\pi}\,\mathrm{Re}\int_{-\infty}^\infty \frac{d\omega}{\omega - 2\pi/\tau}\left(\mathrm{e}^{-\mathrm{i}\,\omega(t-T)} - \mathrm{e}^{-\mathrm{i}\,\omega t}\right). \tag{4.78}$$

It is not entirely obvious that Eq. (4.78) is equivalent to Eq. (4.77). However, we can easily prove that this is the case by taking Eq. (4.78) and using the standard definition of a real part (*i.e.*, half the sum of the quantity in question and its complex conjugate) to give

$$\begin{aligned} f(t) &= \frac{1}{4\pi}\int_{-\infty}^\infty \frac{d\omega}{\omega - 2\pi/\tau}\left(\mathrm{e}^{-\mathrm{i}\,\omega(t-T)} - \mathrm{e}^{-\mathrm{i}\,\omega t}\right) \\ &\quad + \frac{1}{4\pi}\int_{-\infty}^\infty \frac{d\omega}{\omega - 2\pi/\tau}\left(\mathrm{e}^{+\mathrm{i}\,\omega(t-T)} - \mathrm{e}^{+\mathrm{i}\,\omega t}\right). \end{aligned} \tag{4.79}$$

Replacing the dummy integration variable ω by $-\omega$ in the second integral and then making use of symmetry, it is easily seen that the above expression reduces to Eq. (4.77).

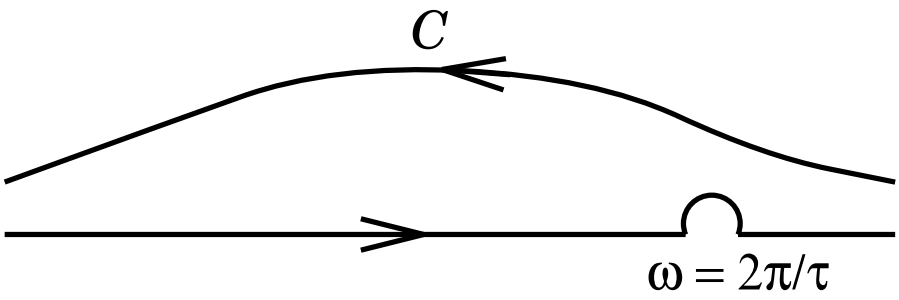

Figure 6: Sketch of the integration contours used to evaluate Eqs. (4.78) and (4.81)

Equation (4.77) can be written

$$f(t) = \frac{2}{\tau} \int_{-\infty}^{\infty} d\omega \, \sin[\omega(t - T/2)] \, \frac{\sin(\omega T/2)}{\omega^2 - (2\pi/\tau)^2}. \tag{4.80}$$

Note that the integrand is finite at $\omega = 2\pi/\tau$, since at this point the vanishing of the denominator is compensated for by the simultaneous vanishing of the numerator. It follows that the integrand in Eq. (4.78) is also not infinite at $\omega = 2\pi/\tau$, as long as we do not separate the two exponentials. Thus, we can replace the integration along the real axis through this point by a small semi-circle in the upper half of the complex plane. Once this has been done, we can deform the path still further and can integrate the two exponentials in Eq. (4.78) separately:

$$f(t) = \frac{1}{2\pi} \, \mathrm{Re} \int_C \mathrm{e}^{-\mathrm{i}\,\omega t} \frac{d\omega}{\omega - 2\pi/\tau} - \frac{1}{2\pi} \, \mathrm{Re} \int_C \mathrm{e}^{-\mathrm{i}\,\omega(t-T)} \frac{d\omega}{\omega - 2\pi/\tau} \tag{4.81}$$

The contour C is sketched in Fig. 6. Note that it runs from $+\infty$ to $-\infty$, which accounts for the change of sign between Eqs. (4.78) and (4.81).

We have already noted that a finite wave form which is zero for $t < 0$ and $t > T$ can be through of as the superposition of two out of phase infinite wave forms, one starting at $t = 0$ and the other at $t = T$. It is plausible, therefore, that the first term in the above expression corresponds to the infinite wave form starting at $t = 0$, and the second to the infinite wave form starting at $t = T$. If this is the case then the signal (4.69), which starts at $t = 0$ and ends at $t = \infty$,

can be written in the form

$$f(t) = \frac{1}{2\pi}\,\mathrm{Re}\int_C \mathrm{e}^{-\mathrm{i}\,\omega t}\frac{d\omega}{\omega - 2\pi/\tau}. \tag{4.82}$$

Let us test this proposition. In order to do this we must replace the original path of integration C by two equivalent paths.

First, consider $t < 0$. In this case, $-\mathrm{i}\,\omega t$ has a negative real part in the upper half plane which increases indefinitely with increasing distance from the axis. Thus, we can replace the original path of integration by the path A (see Fig. 7). The integral clearly vanishes along this path if we let A approach infinity in the upper half plane. Consequently,

$$f(t) = 0 \tag{4.83}$$

for $t < 0$.

Next, consider $t > 0$. Now, $-\mathrm{i}\,\omega t$ has a negative real part in the lower half plane, so that the exponential vanishes at infinity in this half plane. If we attempt to deform C to infinity in the lower half plane, the path of integration "catches" on the singularity of the integrand at $\omega = 2\pi/\tau$ (see Fig. 7). The path of integration B therefore consists of three parts: the part at infinity, B_1, where the integral vanishes due to the exponential factor $\mathrm{e}^{-\mathrm{i}\,\omega t}$; B_2, the two parts leading to infinity which cancel each other and thus contribute nothing to the integral; the path B_3 around the singularity. This latter contribution can easily be evaluated using the Cauchy residue theorem:

$$B_3 = \frac{1}{2\pi}\,\mathrm{Re}\,(2\pi\mathrm{i}\ \mathrm{e}^{-2\pi\mathrm{i}\,t/\tau}) = \sin\left(\frac{2\pi t}{\tau}\right). \tag{4.84}$$

Thus, it is proven that the expression (4.82) actually describes a wave form beginning at $t = 0$ whose subsequent motion is specified by Eq. (4.69).

Equation (4.82) can immediately be generalized to give the wave motion in the region $x > 0$:

$$f(x,t) = \frac{1}{2\pi}\,\mathrm{Re}\int_C \mathrm{e}^{\mathrm{i}\,(kx-\omega t)}\frac{d\omega}{\omega - 2\pi/\tau}. \tag{4.85}$$

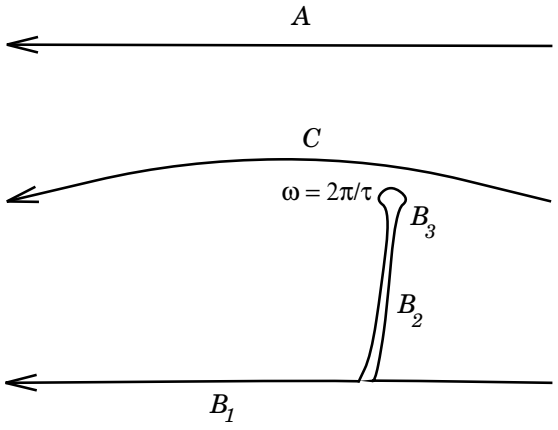

Figure 7: Sketch of the integration contours used to evaluate Eq. (4.82)

This follows from standard wave theory, because we know that an unterminated wave motion at $x = 0$ of the form $\mathrm{e}^{-\mathrm{i}\,\omega t}$ takes the form $\mathrm{e}^{\,\mathrm{i}\,(kx-\omega t)}$ after moving a distance x in the dispersive medium, provided that k and ω are related by the appropriate dispersion relation. For a medium consisting of a single resonant species this dispersion relation is written (see Eq. (4.17))

$$k^2 = \frac{\omega^2}{c^2}\left(1 + \frac{(Ne^2/\epsilon_0 m)}{{\omega_0}^2 - \omega^2 - \mathrm{i}\,g\,\omega\,\omega_0}\right). \tag{4.86}$$

4.9 Propagation of the wave front in a dispersive medium

It is helpful to define

$$s = t - \frac{x}{c}. \tag{4.87}$$

Let us consider the two cases $s < 0$ and $s > 0$ separately.

Suppose that $s < 0$. In this case we distort the path C, used to evaluate the integral (4.85), into the path A shown in Fig. 8. This is only a sensible thing to do if the real part of $\mathrm{i}\,(kx - \omega t)$ is negative at infinity in the upper half plane. It is clear from the dispersion relation (4.86) that $k = \omega/c$ in the limit $|\omega| \to \infty$. Thus,

$$\mathrm{i}\,(kx - \omega t) = -\mathrm{i}\,\omega(t - x/c) = -\mathrm{i}\,\omega s. \tag{4.88}$$

It follows that $\mathrm{i}\,(kx-\omega t)$ possesses a large negative real part along path A provided that $s<0$. Thus, Eq. (4.85) yields

$$f(x,t)=0 \tag{4.89}$$

for $s<0$. In other words, *it is impossible for the wave front to propagate through the dispersive medium with a velocity greater than the velocity of light in a vacuum.*

Suppose that $s>0$. In this case we distort the path C into the *lower* half plane, since $\mathrm{i}\,(kx-\omega t)=-\mathrm{i}\,\omega s$ has a negative real part at infinity in this region. In doing this, the path becomes stuck not only at the singularity of the denominator when $\omega=2\pi/\tau$, but also at the branch points of the expression for k. After a little algebra, the dispersion relation (4.86) yields

$$k=\frac{\omega}{c}\sqrt{\frac{\omega_{1+}-\omega}{\omega_{0+}-\omega}}\sqrt{\frac{\omega_{1-}-\omega}{\omega_{0-}-\omega}}, \tag{4.90}$$

where

$$\omega_{0\pm}=-\mathrm{i}\,\rho\pm\sqrt{{\omega_0}^2-\rho^2}, \tag{4.91}$$

and

$$\omega_{1\pm}=-\mathrm{i}\,\rho\pm\sqrt{{\omega_0}^2+{\omega_p}^2-\rho^2}. \tag{4.92}$$

Here,

$$\omega_p=\sqrt{Ne^2/\epsilon_0 m} \tag{4.93}$$

is the plasma frequency, and

$$\rho=\frac{g\,\omega_0}{2}\ll\omega_0 \tag{4.94}$$

parameterizes the damping. In order to prevent multiple roots of Eq. (4.90) it is necessary to place branch cuts between ω_{0+} and ω_{1+} and also between ω_{0-} and ω_{1-} (see Fig. 8).

The path of integration B is conveniently split into the parts B_1 through B_5. The contribution from B_1 is negligible since the exponential in Eq. (4.85) is vanishingly small on this part of the integration path. Likewise, the contribution

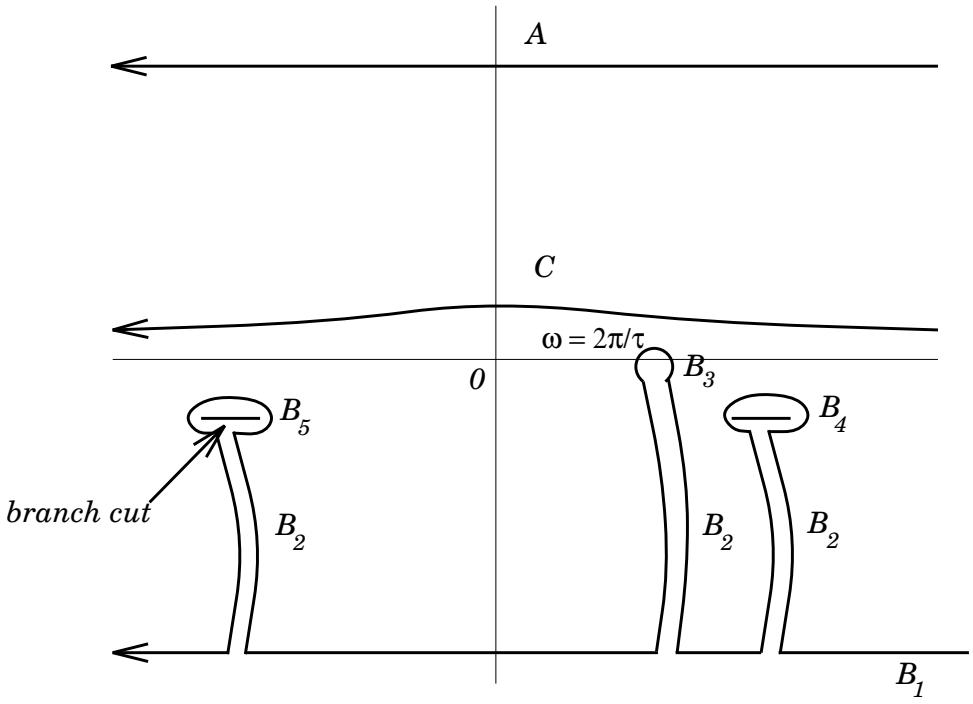

Figure 8: Sketch of the integration contours used to evaluate Eq. (4.85)

from B_2 is zero since its two sections always cancel. The contribution from B_3 follows from the residue theorem:

$$B_3 = \frac{1}{2\pi}\,\mathrm{Re}\,(2\pi\mathrm{i}\,\mathrm{e}^{\mathrm{i}\,[k_\tau x - 2\pi t/\tau]}). \tag{4.95}$$

Here, k_τ denotes the value of k obtained from the dispersion relation (4.86) in the limit $\omega \to 2\pi/\tau$. Thus,

$$B_3 = \mathrm{e}^{-\mathrm{Im}(k_\tau)\,x}\,\sin\left(2\pi\,\frac{t}{\tau} - \mathrm{Re}(k_\tau)\,x\right). \tag{4.96}$$

In general, the contributions from B_4 and B_5 cannot be simplified further. For the moment we denote them as

$$B_4 = \frac{1}{2\pi}\,\mathrm{Re}\oint_{B_4} \mathrm{e}^{\mathrm{i}\,(kx-\omega t)}\,\frac{d\omega}{\omega - 2\pi/\tau}, \tag{4.97}$$

and

$$B_5 = \frac{1}{2\pi}\,\mathrm{Re}\oint_{B_5} \mathrm{e}^{\mathrm{i}\,(kx-\omega t)}\,\frac{d\omega}{\omega - 2\pi/\tau}, \tag{4.98}$$

where the paths of integration circle the appropriate branch cuts. In all, we have

$$f(x,t) = \mathrm{e}^{-\mathrm{Im}(k_\tau)\,x} \sin\left(2\pi\,\frac{t}{\tau} - \mathrm{Re}(k_\tau)\,x\right) + B_4 + B_5 \tag{4.99}$$

for $s > 0$.

Let us now look at the special case $s = 0$. For this value of s we can change the original path of integration to one at infinity in either the upper or the lower half plane, since the integrand vanishes in each case, through no longer exponentially, but rather as $1/\omega^2$. We can see this from Eq. (4.82), which can be written in the form

$$f(t) = \frac{1}{4\pi}\left(\int_C \mathrm{e}^{-\mathrm{i}\,\omega t}\frac{d\omega}{\omega - 2\pi/\tau} + \int_C \mathrm{e}^{+\mathrm{i}\,\omega t}\frac{d\omega}{\omega - 2\pi/\tau}\right). \tag{4.100}$$

Substitution of ω for $-\omega$ in the second integral yields

$$f(t) = \frac{1}{\tau}\int \mathrm{e}^{-\mathrm{i}\,\omega t}\,\frac{d\omega}{\omega^2 - (2\pi/\tau)^2}. \tag{4.101}$$

Now, applying dispersion theory, we get from the above equation, just as we got Eq. (4.85) from Eq. (4.82),

$$f(x,t) = \frac{1}{\tau}\int \mathrm{e}^{\mathrm{i}\,(kx-\omega t)}\,\frac{d\omega}{\omega^2 - (2\pi/\tau)^2}. \tag{4.102}$$

Clearly, the integrand vanishes as $\mathrm{e}^{-\mathrm{i}\,\omega s}/\omega^2$ as ω becomes very large. Thus, it vanishes as $1/\omega^2$ for $s = 0$. Since we can calculate $f(x,t)$ by using either path A or path B, we can see that

$$f(x,t) = \mathrm{e}^{-\mathrm{Im}(k_\tau)\,x} \sin\left(2\pi\,\frac{t}{\tau} - \mathrm{Re}(k_\tau)\,x\right) + B_4 + B_5 = 0 \tag{4.103}$$

for $s = 0$. Thus, there is continuity in the transition from the region $s < 0$ to the region $s > 0$.

We are now in a position to make some meaningful statements about the behaviour of the signal at depth x inside the dispersive medium. Prior to the time $t = x/c$ there is no motion. Even if the phase velocity is superluminal, no

electromagnetic signal can arrive earlier than one propagating with the velocity of light in vacuum c. The wave motion for $t > x/c$ is conveniently divided into two parts: *free oscillations* and *forced oscillations*. The former are given by $B_4 + B_5$, and the latter by

$$\mathrm{e}^{-\mathrm{Im}(k_\tau)\,x}\,\sin\left(2\pi\,\frac{t}{\tau} - \mathrm{Re}(k_\tau)\,x\right) = \mathrm{e}^{-\mathrm{Im}(k_\tau)\,x}\,\sin\left(\frac{2\pi}{\tau}\left[t - \frac{x}{v_p}\right]\right), \qquad (4.104)$$

where

$$v_p = \frac{2\pi}{\tau\,\mathrm{Re}(k_\tau)} \qquad (4.105)$$

is termed the *phase velocity*. The forced oscillations have the same sine wave characteristics and oscillation frequency as the incident wave. However, the wave amplitude is diminished by the damping coefficient, although, as we have seen, this is generally a negligible effect unless the frequency of the incident wave closely matches one of the resonant frequencies of the dispersive medium. The phase velocity v_p determines the velocity with which a point of constant phase (*e.g.*, a peak or trough) of the forced oscillation signal propagates into the medium. However, *the phase velocity has no effect on the velocity with which the forced oscillation wave front propagates into the medium*. This latter velocity is equivalent to the velocity of light in vacuum c . The phase velocity v_p can be either greater or less than c, in which case peaks and troughs either catch up with or fall further behind the wave front. Of course, peaks can never overtake the wave front.

It is clear from Eqs. (4.91), (4.92), (4.97), and (4.98) that the free oscillations oscillate with real frequencies which are somewhere between the resonant frequency ω_0 and the plasma frequency ω_p. Furthermore, the free oscillations are *damped* in time like $\exp(-\rho\,t)$. The free oscillations, like the forced oscillations, begin at time $t = x/c$. At $t = x/c$ the free and forced oscillations just cancel (see Eq. (4.103)). As t increases both the free and forced oscillations set in, but the former rapidly damp away, leaving only the forced oscillations. Thus, the free oscillations can be regarded as some sort of *transient* response of the medium to the incident wave, whereas the forced oscillations determine the time asymptotic response. The real frequency of the forced oscillations is that imposed externally by the incident wave, whereas the real frequency of the free oscillations is determined by the nature of the dispersive medium, quite independently of the frequency of the incident wave.

One slightly surprising result of the above analysis is the prediction that the wave front of the signal propagates into the dispersive medium with the velocity of light in vacuum, irrespective of the dispersive properties of the medium. Actually, this is a fairly obvious result. As is well described by Feynman in his famous *Lectures on Physics*, when an electromagnetic wave propagates through a dispersive medium, the electrons and ions which make up that medium oscillate in sympathy with the incident wave and in doing so emit radiation. Both the radiation from the electrons and ions and the incident radiation travel at the velocity c. However, when these two radiation signals are superposed the net effect is as if the incident signal propagates through the dispersive medium with a phase velocity which is different from c. Consider the wave front of the incident signal, which clearly propagates into the medium with the velocity c. Prior to the arrival of this wave front the electrons and ions are at rest, since no information regarding the arrival of the incident wave at the surface of medium can propagate faster than c. After the arrival of the wave front the electrons and ions are set into motion and emit radiation which can affect the apparent phase velocity of radiation which arrives somewhat later. But this radiation certainly cannot affect the propagation velocity of the wave front itself, which has already passed by the time the electrons and ions are set into motion (because of the finite inertia of the electrons and ions).

4.10 The Sommerfeld precursor

Let us consider the situation immediately after the arrival of the signal; *i.e.*, when s is small and positive. Let us start from Eq. (4.102), which can be written in the form

$$f(x,t) = \frac{1}{\tau}\int_C \mathrm{e}^{\,\mathrm{i}\,([k-\omega/c]x-\omega s)}\,\frac{d\omega}{\omega^2-(2\pi/\tau)^2}. \tag{4.106}$$

We can deform the original path of integration C into a large semi-circle of radius R in the upper half-plane, plus the segments of the real axis, as shown in Fig. 9. Because of the denominator $\omega^2-(2\pi/\tau)^2$, the integrand tends to zero as $1/\omega^2$ on the real axis. We may add the path in the lower half plane which is shown as a dotted line in the figure, for if the radius of the semi-circular portion of this lower path is increased to infinity, the integrand vanishes exponentially because

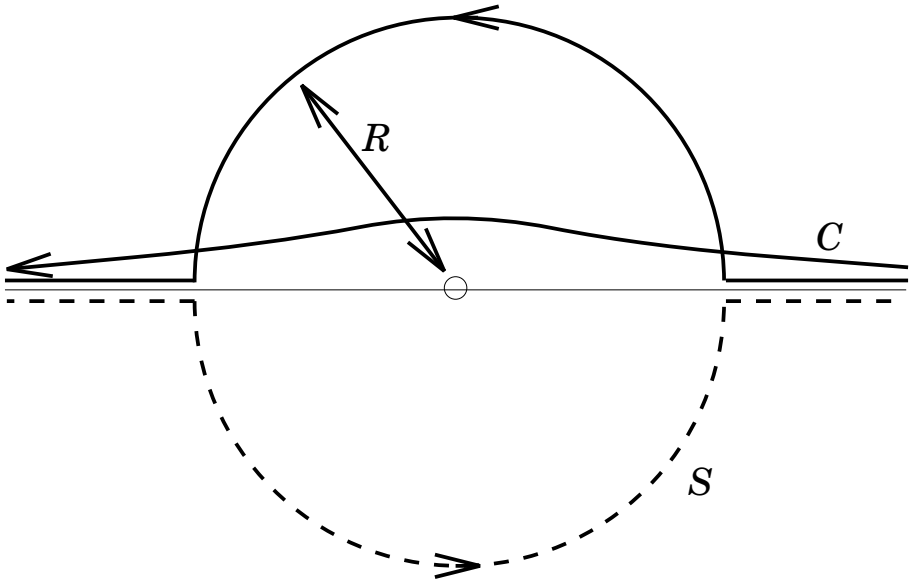

Figure 9: Sketch of the integration contour used to evaluate Eq. (4.107)

$s > 0$. Therefore, we may replace our original path of integration by the entire circle S. Thus,

$$f(x,t) = \frac{1}{\tau} \oint_S \mathrm{e}^{\mathrm{i}\,([k-\omega/c]x-\omega s)} \frac{d\omega}{\omega^2 - (2\pi/\tau)^2} \tag{4.107}$$

in the limit that the radius of the circle R tends to infinity.

The dispersion relation (4.86) yields

$$k - \frac{\omega}{c} \simeq \frac{\omega}{c}\left(\sqrt{1 - \frac{{\omega_p}^2}{\omega^2}} - 1\right) \simeq -\frac{{\omega_p}^2}{2\,c\,\omega} \tag{4.108}$$

in the limit $|\omega| \to \infty$. Using the abbreviation

$$\xi = \frac{{\omega_p}^2}{2\,c}\,x, \tag{4.109}$$

and henceforth neglecting $2\pi/\tau$ with respect to ω, we obtain from Eq. (4.107)

$$f(x,t) = f_1(\xi,t) \simeq \frac{1}{\tau} \oint_S \exp\left[\mathrm{i}\left(-\frac{\xi}{\omega} - \omega\, s\right)\right] \frac{d\omega}{\omega^2}. \tag{4.110}$$

This expression can also be written

$$f_1(\xi,t) = \frac{1}{\tau}\oint_S \exp\left[-\mathrm{i}\sqrt{\xi\,s}\left(\frac{1}{\omega}\sqrt{\frac{\xi}{s}}+\omega\sqrt{\frac{s}{\xi}}\right)\right]\frac{d\omega}{\omega^2}. \tag{4.111}$$

Let

$$\omega\sqrt{\frac{s}{\xi}} = \mathrm{e}^{\mathrm{i}\,u}. \tag{4.112}$$

It follows that

$$\frac{d\omega}{\omega} = \mathrm{i}\,du, \tag{4.113}$$

giving

$$\frac{d\omega}{\omega^2} = \mathrm{i}\sqrt{\frac{s}{\xi}}\,\mathrm{e}^{-\mathrm{i}\,u}\,du. \tag{4.114}$$

Substituting the angular variable u for ω as the integration variable in Eq. (4.111) yields

$$f_1(\xi,t) = \frac{\mathrm{i}}{\tau}\sqrt{\frac{s}{\xi}}\int_0^{2\pi}\exp(-2\mathrm{i}\sqrt{\xi s}\,\cos u)\,\mathrm{e}^{-\mathrm{i}\,u}\,du. \tag{4.115}$$

Here, we have taken $\sqrt{\xi/s}$ as the radius of the circular integration path in the ω-plane. This is indeed a large radius, since $s \ll 1$. From symmetry, Eq. (4.115) simplifies to

$$f_1(\xi,t) = \frac{\mathrm{i}}{\tau}\sqrt{\frac{s}{\xi}}\int_0^{2\pi}\exp(-2\mathrm{i}\sqrt{\xi s}\,\cos u)\,\cos u\,du. \tag{4.116}$$

The following mathematical identity is very well-known[11]

$$J_n(z) = \frac{\mathrm{i}^{-n}}{2\pi}\int_0^{2\pi}\mathrm{e}^{\mathrm{i}\,z\,\cos\theta}\cos(n\theta)\,d\theta, \tag{4.117}$$

where $J_n(z)$ is Bessel function of order n. It follows from Eq. (4.115) that

$$f_1(\xi,t) = \frac{2\pi}{\tau}\sqrt{\frac{s}{\xi}}\,J_1(2\sqrt{\xi s}). \tag{4.118}$$

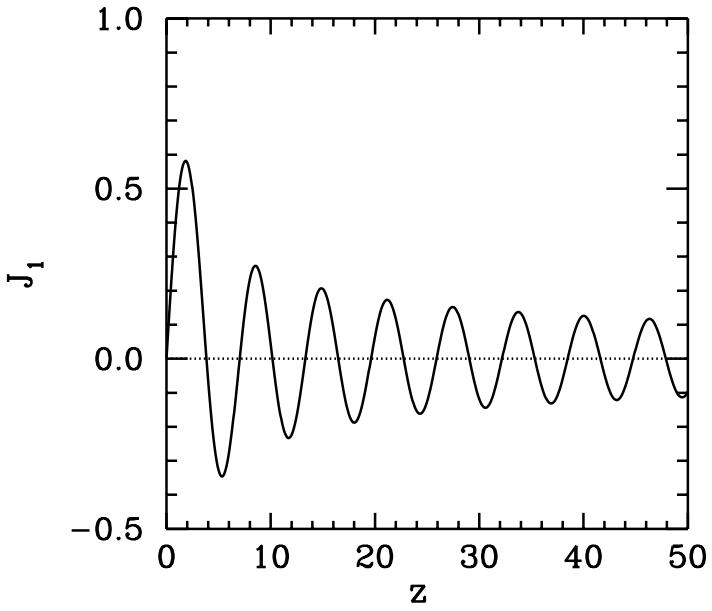

Figure 10: The Bessel function $J_1(z)$

Here, we have made use of the fact that $J_1(-z) = -J_1(z)$.

The properties of Bessel functions are well-known and are listed in many standard references on mathematical functions (see, for instance, Abramowitz and Stegun). In the small argument limit $z \ll 1$ we find that

$$J_1(z) = \frac{z}{2} + O(z^3). \tag{4.119}$$

On the other hand, in the large argument limit $z \gg 1$ we obtain

$$J_1(z) = \sqrt{\frac{2}{\pi z}}\cos(z - 3\pi/4) + O(z^{-3/2}). \tag{4.120}$$

The behaviour of $J_1(z)$ is further illustrated in Fig. 10.

We are now in a position to make some quantitative statements regarding the signal which first arrives at depth x in the dispersive medium. This signal propagates at the velocity of light in vacuum and is called the *Sommerfeld*

[11] M. Abramowitz, and I.A. Stegun, *Handbook of mathematical functions*, (Dover, New York, 1965), Eq. 9.1.21.

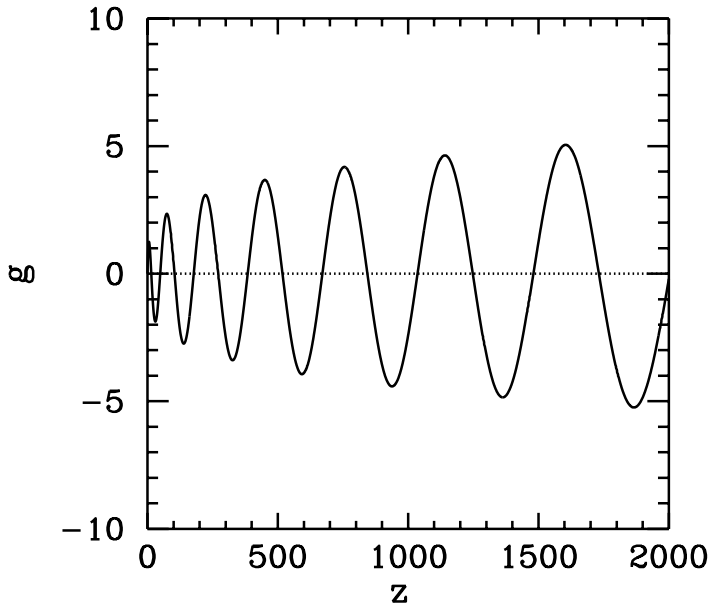

Figure 11: The Sommerfeld precursor

precursor. The first important point to note is that the amplitude of the Sommerfeld precursor is very small compared to that of the incident wave (whose amplitude is normalized to unity). We can easily see this because in deriving Eq. (4.118) we assumed that $|\omega| = \sqrt{\xi/s} \gg 2\pi/\tau$ on the circular integration path S. Since the magnitude of J_1 is always less than, or of order, unity, it is clear that $|f_1| \ll 1$. This is a comforting result, since in a naive treatment of wave propagation through a dielectric medium the wave front propagates at the group velocity v_g (which is usually less than c) and, therefore, no signal should reach depth x in the medium before time x/v_g. We are finding that there is, in fact, a precursor which arrives at $t = x/c$, but that this signal is fairly small. Note from Eq. (4.109) that ξ is proportional to x. Clearly, the amplitude of the Sommerfeld precursor decreases like one over the distance traveled by the wave front through the dispersive medium (since J_1 attains its maximum value when $s \sim 1/\xi$). Thus, the Sommerfeld precursor is likely to become undetectable after the wave has traveled a long distance through the medium.

Equation (4.118) can be written

$$f_1(\xi, t) = \frac{\pi}{\xi\,\tau}\, g(s/s_0), \tag{4.121}$$

where $s_0 = 1/4\,\xi$, and

$$g(z) = \sqrt{z}\, J_1(\sqrt{z}). \tag{4.122}$$

The normalized Sommerfeld precursor $g(z)$ is shown in Fig. 11. It can be seen that both the amplitude and the oscillation period of the precursor gradually increase. The roots of $J_1(z)$ [*i.e.*, the solutions of $J_1(z) = 0$] are spaced at distances of approximately π apart. Thus, the time interval for the mth half period of the precursor is approximately given by

$$\Delta t_m \sim \frac{m\pi^2}{2\xi}. \tag{4.123}$$

Note that the initial period of oscillation,

$$\Delta t_0 \sim \frac{\pi^2}{2\xi}, \tag{4.124}$$

is extremely small compared to the incident period τ. Moreover, the initial period of oscillation is *completely independent* of the frequency of the incident wave. In fact, Δt_0 depends only on the depth x and on the dispersive power of the medium. The period decreases with increasing distance x traveled by the wave front though the medium. So, when visible radiation is incident on some dispersive medium it is quite possible for the first signal detected well inside the medium to lie in the X-ray region of the electromagnetic spectrum.

4.11 The method of stationary phase

Equation (4.102) can be written in the form

$$f(x, t) = \int_C \mathrm{e}^{\mathrm{i}\,\phi(\omega)} F(\omega)\, d\omega \tag{4.125}$$

where

$$F(\omega) = \frac{1}{\tau}\frac{1}{\omega^2 - (2\pi/\tau)^2}, \tag{4.126}$$

and

$$\phi(\omega) = k(\omega)\, x - \omega t. \tag{4.127}$$

It is clear that $F(\omega)$ is a relatively slowly varying function of ω (except in the immediate vicinity of the singular points $\omega = \pm 2\pi/\tau$), whereas the phase $\phi(\omega)$ is generally large and rapidly varying. The rapid oscillations of $\exp(\,\mathrm{i}\,\phi)$ over most of the range of integration means that the integrand averages to almost zero. Exceptions to this cancellation rule occur only when $\phi(\omega)$ is *stationary*; *i.e.*, when $\phi(\omega)$ has an extremum. The integral can therefore be estimated by finding places where $\phi(\omega)$ has a vanishing derivative, evaluating (approximately) the integral in the neighbourhood of each of these points, and summing the contributions. This procedure is called the *method of stationary phase.*

Suppose that $\phi(\omega)$ has a vanishing first derivative at $\omega = \omega_s$. In the neighbourhood of this point, $\phi(\omega)$ can be expanded as a Taylor series,

$$\phi(\omega) = \phi_s + \frac{1}{2}\phi_s''(\omega - \omega_s)^2 + \cdots. \tag{4.128}$$

Here, the subscript s is used to indicate ϕ or its second derivative evaluated at $\omega = \omega_s$. Since $F(\omega)$ is slowly varying, the contribution to the integral from this stationary phase point is approximately

$$f_s \simeq F(\omega_s)\, e^{\mathrm{i}\,\phi_s} \int_{\infty}^{-\infty} \mathrm{e}^{(\mathrm{i}/2)\phi_s''(\omega-\omega_s)^2}\, d\omega. \tag{4.129}$$

It is tacitly assumed that the stationary point lies on the real axis in ω-space, so that locally the integral along the contour C is an integral along the real axis in the direction of decreasing ω. The above expression can be written in the form

$$f_s \simeq -F(\omega_s)\, \mathrm{e}^{\mathrm{i}\,\phi_s} \sqrt{\frac{4\pi}{\phi_s''}} \int_0^{\infty} \left[\cos(\pi t^2/2) + \mathrm{i}\, \sin(\pi t^2/2)\right]\, dt, \tag{4.130}$$

where

$$\frac{\pi}{2}\, t^2 = \frac{1}{2}\, \phi_s''\, (\omega - \omega_s)^2. \tag{4.131}$$

The integrals in the above expression are, *Fresnel integrals*[12] and can be shown to take the values

$$\int_0^\infty \cos(\pi t^2/2)\,dt = \int_0^\infty \sin(\pi t^2/2)\,dt = \frac{1}{2}. \tag{4.132}$$

It follows that

$$f_s \simeq -\sqrt{\frac{2\pi\,\mathrm{i}}{\phi_s''}}\,F(\omega_s)\,\mathrm{e}^{\,\mathrm{i}\,\phi_s}. \tag{4.133}$$

It is easily seen that the arc length (in ω-space) of the integration contour which makes a significant contribution to f_s is of order $\Delta\omega/\omega_s \sim 1/\sqrt{k(\omega_s)\,x}$. Thus, the arc length is relatively short provided that the wavelength of the signal is much less than the distance propagated through the dispersive medium. If there is more than one point of stationary phase in the range of integration then the integral is approximated as a sum of terms like the above.

Integrals of the form (4.125) can be calculated exactly using the *method of steepest decent.*[13] The stationary phase approximation (4.133) agrees with the leading term of the method of steepest decent (which is far more difficult to implement than the method of stationary phase) provided that $\phi(\omega)$ is real (*i.e.*, provided that the stationary point lies on the real axis). If ϕ is complex, however, the stationary phase method can yield erroneous results. This suggests that the stationary phase method is likely to break down when the extremum point $\omega = \omega_s$ approaches any poles or branch cuts in the ω-plane (see Fig. 8).

4.12 The group velocity

The point of stationary phase, defined by $\partial\phi/\partial\omega = 0$, satisfies the condition

$$\frac{c}{v_g} = \frac{ct}{x}, \tag{4.134}$$

[12] M. Abramowitz, and I.A. Stegun, *Handbook of mathematical functions*, (Dover, New York, 1965), Sec. 7.3.

[13] Léon Brillouin, *Wave propagation and group velocity*, (Academic press, New York, 1960).

where

$$v_g = \frac{d\omega}{dk} \tag{4.135}$$

is conventionally termed the *group velocity*. Thus, the signal seen at position x and time t is dominated by the frequency range whose group velocity v_g is equal to x/t. In this respect, the signal incident at the surface of the medium ($x = 0$) at time $t = 0$ can be said to propagate through the medium at the group velocity $v_g(\omega)$.

The simple one-resonance dispersion relation (4.86) yields

$$\frac{c}{v_g} \simeq n(\omega)\left[1 + \frac{\omega^2}{{\omega_0}^2 - \omega^2} + \frac{\omega^2}{\omega^2 - {\omega_0}^2 - {\omega_p}^2}\right] \tag{4.136}$$

in the limit $g \to 0$, where

$$n(\omega) = \frac{ck}{\omega} = \sqrt{\frac{{\omega_0}^2 + {\omega_p}^2 - \omega^2}{{\omega_0}^2 - \omega^2}}. \tag{4.137}$$

The variation of c/v_g and the refractive index n with frequency is sketched in Fig. 12. With $g = 0$ the group velocity is less than c for all ω, except for $\omega_0 < \omega < \omega_1 \equiv \sqrt{{\omega_0}^2 + {\omega_p}^2}$, where it is purely imaginary. Note that the refractive index is also complex in this frequency range. The phase velocity $v_p = c/n$ is subluminal for $\omega < \omega_0$, imaginary for $\omega_0 \leq \omega \leq \omega_1$, and superluminal for $\omega > \omega_1$.

The frequency range which contributes to the amplitude at time t is determined graphically by finding the intersection of a horizontal line with ordinate ct/x with the solid curve in Fig. 12. There is no crossing of the two curves for $t < t_0 \equiv x/c$, thus no signal arrives before this time. For times immediately following t_0 the point of stationary phase is seen to be at $\omega \to \infty$. In this large ω limit the point of stationary phase is given by

$$\omega_s \simeq \omega_p \sqrt{\frac{t_0}{2(t - t_0)}}. \tag{4.138}$$

Note that $\omega = -\omega_s$ is also a point of stationary phase. It is easily demonstrated that

$$\phi_s \simeq -2\sqrt{\xi(t - t_0)}, \tag{4.139}$$

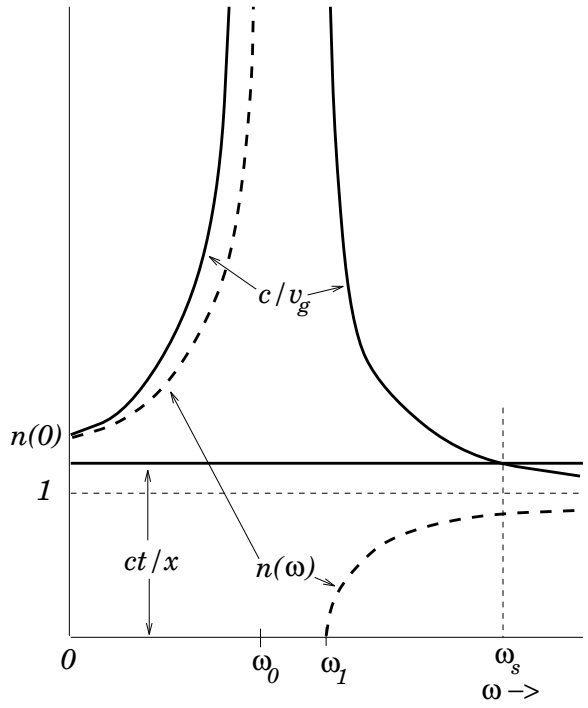

Figure 12: The typical variation of the functions $c/v_g(\omega)$ and $n(\omega)$. Here, $\omega_1 = ({\omega_0}^2 + {\omega_p}^2)^{1/2}$.

and

$$\phi_s'' \simeq -2\,\frac{(t-t_0)^{3/2}}{\xi^{1/2}}, \tag{4.140}$$

with

$$F(\omega_s) \simeq \frac{t-t_0}{\tau\,\xi}. \tag{4.141}$$

Here, ξ is given by Eq. (4.109). The stationary phase approximation (4.133) gives

$$f_s \simeq \sqrt{\frac{\pi\,\xi^{1/2}}{(t-t_0)^{3/2}}}\,\frac{t-t_0}{\tau\,\xi}\,\mathrm{e}^{-2\mathrm{i}\sqrt{\xi(t-t_0)}+3\pi\mathrm{i}/4} + \text{c.c.}, \tag{4.142}$$

where c.c. denotes the complex conjugate of the preceding term (this contribution comes from the second point of stationary phase located at $\omega = -\omega_s$). The above expression reduces to

$$f_s \simeq \frac{2\sqrt{\pi}}{\tau}\,\frac{(t-t_0)^{1/4}}{\xi^{3/4}}\cos\left[2\sqrt{\xi(t-t_0)}-3\pi/4\right]. \tag{4.143}$$

It is easily demonstrated that the above formula is the same as the expression (4.118) for the Sommerfeld precursor in the large argument limit $t - t_0 \gg 1/\xi$. Thus, the method of stationary phase yields an expression for the Sommerfeld precursor which is accurate at all times except those immediately following the first arrival of the signal.

4.13 The Brillouin precursor

As time progresses the horizontal line ct/x in Fig. 12 gradually rises and the point of stationary phase moves to ever lower frequencies. In general, however, the amplitude remains relatively small. Only when the elapsed time reaches

$$t_1 = \frac{n(0)\,x}{c} > t_0 \tag{4.144}$$

is there a qualitative change. This time marks the arrival of a second precursor known as the *Brillouin precursor*. The reason for the qualitative change is evident from Fig. 12. At $t = t_1$ the lower region of the c/v_g curve is intersected

for the first time, and $\omega = 0$ becomes a point of stationary phase. It is clear that the oscillation frequency of the Brillouin precursor is far less than that of the Sommerfeld precursor. Moreover, it is easily demonstrated that the second derivative of $k(\omega)$ vanishes at $\omega = 0$. This means that $\phi_s'' = 0$. The stationary phase result (4.133) gives an infinite answer in such circumstances. Of course, the amplitude of the Brillouin precursor is not infinite, but it is significantly larger than that of the Sommerfeld precursor.

In order to generalize the result (4.133) to deal with a stationary phase point at $\omega = 0$ it is necessary to expand $\phi(\omega)$ about this point, keeping terms up to ω^3. Thus,

$$\phi(\omega) \simeq \omega(t_1 - t) + \frac{x}{6}\, k_0'''\, \omega^3, \tag{4.145}$$

where

$$k_0''' \equiv \left(\frac{d^3k}{d\omega^3}\right)_{\omega=0} = \frac{3\,{\omega_p}^2}{c\,n(0)\,{\omega_0}^4} \tag{4.146}$$

for the simple dispersion relation (4.86). The amplitude (4.125) is therefore given approximately by

$$f(x,t) \simeq F(0) \int_{\infty}^{-\infty} \mathrm{e}^{\mathrm{i}\,\omega(t_1-t)+\mathrm{i}\,(x/6)k_0'''\omega^3}\, d\omega. \tag{4.147}$$

This expression reduces to

$$f(x,t) = \frac{\tau}{\sqrt{2}\,\pi^2}\sqrt{\frac{|t-t_1|}{x\,k_0'''}} \int_0^{\infty} \cos\left[\frac{3}{2} z \left(\frac{v^3}{3} \pm v\right)\right]\, dv, \tag{4.148}$$

where

$$v = \sqrt{\frac{x\,k_0'''}{2\,|t-t_1|}}\;\omega, \tag{4.149}$$

and

$$z = \frac{2\sqrt{2}\,|t-t_1|^{3/2}}{3\sqrt{x\,k_0'''}}. \tag{4.150}$$

The positive (negative) sign in the integrand is taken for $t < t_1$ ($t > t_1$).

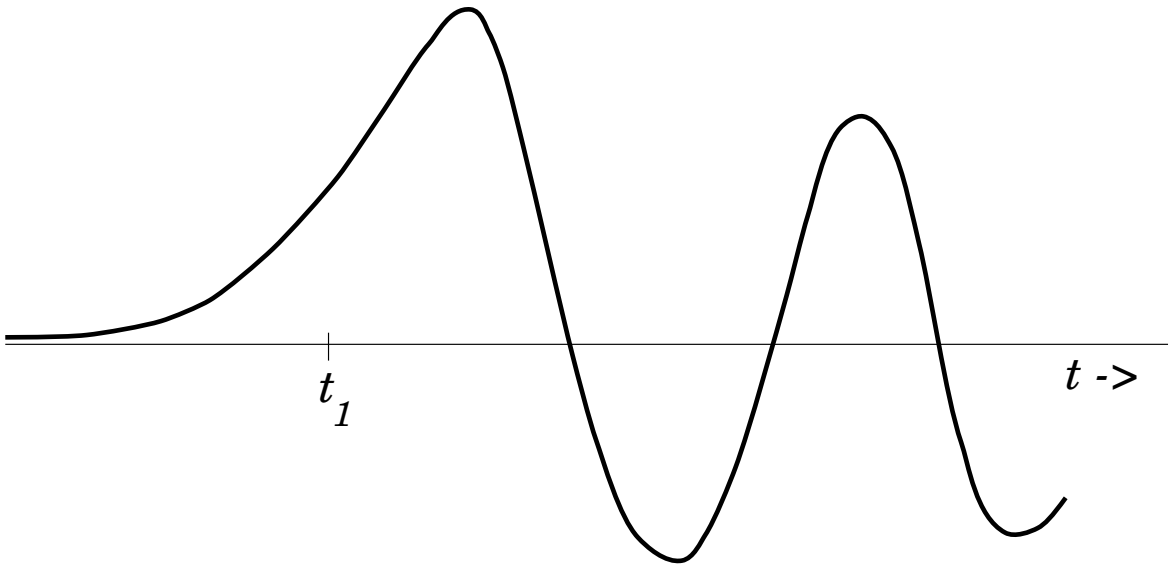

Figure 13: A sketch of the behaviour of the Brillouin precursor as a function of time

The integral in Eq. (4.150) is known as an *Airy integral*. It can be expressed in terms of Bessel functions of order 1/3, as follows:

$$\int_0^\infty \cos\left[\frac{3}{2}z\left(\frac{v^3}{3}+v\right)\right]\,dv = \frac{1}{\sqrt{3}}\,K_{1/3}(z), \tag{4.151}$$

and

$$\int_0^\infty \cos\left[\frac{3}{2}z\left(\frac{v^3}{3}-v\right)\right]\,dv = \frac{\pi}{3}\left[J_{1/3}(z)+J_{-1/3}(z)\right]. \tag{4.152}$$

From the well-known properties of Bessel functions the precursor can be seen to have a growing exponential character for times earlier than $t = t_1$, and an oscillating character for $t > t_1$. The amplitude in the neighbourhood of $t = t_1$ is plotted in Fig. 13.

The initial oscillation period of the Brillouin precursor is crudely estimated (from $z \sim 1$) as

$$\Delta t_0 \sim (x\,k_0''')^{1/3}. \tag{4.153}$$

The amplitude of the Brillouin precursor is approximately

$$|f| \sim \frac{\tau}{(x\,k_0''')^{1/3}}. \tag{4.154}$$

Let us adopt the ordering

$$1/\tau \sim \omega_0 \sim \omega_p \ll \xi, \tag{4.155}$$

which corresponds to most physical situations involving the propagation of electromagnetic radiation through dielectric media. It follows from the above results, plus the results of Section 4.10, that

$$(\Delta t_0\, \omega_p)_{\rm brillouin} \sim \left(\frac{\xi}{\omega_p}\right)^{1/3} \gg 1, \tag{4.156}$$

and

$$(\Delta t_0\, \omega_p)_{\rm sommerfeld} \sim \left(\frac{\omega_p}{\xi}\right) \ll 1. \tag{4.157}$$

Furthermore,

$$|f|_{\rm brillouin} \sim \left(\frac{\omega_p}{\xi}\right)^{1/3} \ll 1, \tag{4.158}$$

and

$$|f|_{\rm sommerfeld} \sim \left(\frac{\omega_p}{\xi}\right) \ll |f|_{\rm brillouin}. \tag{4.159}$$

It is clear that the Sommerfeld precursor is a low amplitude, high frequency signal, whereas the Brillouin precursor is a higher amplitude, low frequency signal. Note that the amplitude of the Brillouin precursor, whilst it is significantly higher than that of the Sommerfeld precursor, is still much less than that of the incident wave.

4.14 Signal arrival

Let us try to establish at what time t_2 a signal first arrives at position x inside the dielectric medium whose amplitude is comparable with that of the wave incident at time $t = 0$ on the surface of the medium ($x = 0$). Let us term this event the "arrival" of the signal. It is plausible from the discussion in Section 4.11 regarding the stationary phase approximation that signal arrival corresponds to the situation where the point of stationary phase in ω-space corresponds to a pole of the function $F(\omega)$. In other words, when ω_s approaches the frequency

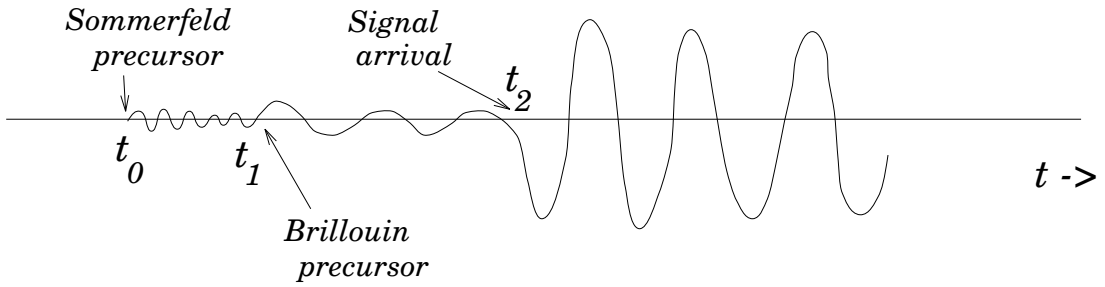

Figure 14: A sketch of the signal amplitude as a function of time as seen inside some dielectric medium subject to an incident wave which starts at some specific time

$2\pi/\tau$ of the incident signal. It is certainly the case that the stationary phase approximation yields a particularly large amplitude signal when $\omega_s \to 2\pi/\tau$. Unfortunately, as has already been discussed, the method of stationary phase becomes inaccurate under these circumstances. However, calculations involving the more robust method of steepest decent[14] confirm that in most cases the signal amplitude first becomes significant when $\omega_s = 2\pi/\tau$. Thus, the signal arrival time is

$$t_2 = \frac{x}{v_g(2\pi/\tau)}, \tag{4.160}$$

where $v_g(2\pi/\tau)$ is the group velocity calculated using the frequency of the incident signal. It is clear from Fig. 12 that

$$t_0 < t_1 < t_2. \tag{4.161}$$

Thus, the main signal arrives later than the Sommerfeld and Brillouin precursors.

The final picture which emerges from our investigations is summarized in Fig. 14. The main signal arrives at the group velocity corresponding to the frequency of the incident wave. However, it is possible to detect the arrival of the signal before this, given sufficiently accurate detection equipment. In fact, the first information regarding the arrival of the incident wave at the vacuum/dielectric interface propagates at the velocity of light in a vacuum.

[14]Léon Brillouin, *Wave propagation and group velocity*, (Academic press, New York, 1960).

4.15 The propagation of radio waves through the ionosphere

We have studied the *transient* behaviour of an electromagnetic wave incident on a spatially *uniform* dielectric medium in great detail. Let us now consider a quite different, but equally important, problem. What is the time asymptotic *steady-state* behaviour of an electromagnetic wave propagating though a spatially *non-uniform* dielectric medium?

As a specific example, let us consider the propagation of radio waves through the Earth's ionosphere. The refractive index of the ionosphere can be written [see Eq. (4.27)]

$$n^2 = 1 - \frac{{\omega_p}^2}{\omega(\omega + \mathrm{i}\,\nu)}, \tag{4.162}$$

where ν is a real positive constant which parameterizes the damping of electron motion (in fact, ν is the collision frequency of free electrons with other particles in the ionosphere), and

$$\omega_p = \sqrt{\frac{N e^2}{\epsilon_0 m}} \tag{4.163}$$

is the plasma frequency. In the above formula, N is the density of free electrons in the ionosphere and m is the electron mass. We shall assume that the ionosphere is horizontally stratified, so that $N = N(z)$, where the coordinate z measures height above the Earth's surface (*n.b.*, the curvature of the Earth is neglected in the following analysis). The ionosphere actually consists of two main layers; the E-layer, and the F-layer. We shall concentrate on the lower E-layer, which lies about 100 km above the surface of the Earth, and is about 50 km thick. The typical day-time number density of free electrons in the E-layer is $N \sim 3 \times 10^{11}\,\mathrm{m}^{-3}$. At night-time, the density of free electrons falls to about half this number. The typical day-time plasma frequency of the E-layer is, therefore, about 5 MHz. The typical collision frequency of free electrons in the E-layer is about 0.05 MHz. According to simplistic theory, any radio wave whose frequency lies below the day-time plasma frequency, 5 MHz, (*i.e.*, any wave whose wavelength exceeds about 60 m) is reflected by the ionosphere during the day. Let us investigate in more detail exactly how this process takes place. Note, incidentally, that for

mega-Hertz frequency radio waves $\nu \ll \omega$, so it follows from Eq. (4.162) that n^2 is predominately real (*i.e.*, under most circumstances, the electron collisions can be neglected).

The problem of radio wave propagation through the ionosphere was of great practical importance during the first half of the 20th Century, since at that time long-wave radio waves were the principle means of military communication. Nowadays, the military have far more reliable ways of communicating. Nevertheless, this subject area is still worth studying because the principle tool used to deal with the problem of wave propagation through a non-uniform medium, the so-called W.K.B. approximation, is of great theoretical importance. In particular, the W.K.B. approximation is very widely used in quantum mechanics (in fact, there is a great similarity between the problem of wave propagation through a non-uniform medium and the problem of solving Schrödinger's equation in the presence of a non-uniform potential).

Maxwell's equations for a wave propagating through a non-uniform, unmagnetized, dielectric medium are:

$$\nabla\cdot\boldsymbol{E} = 0, \tag{4.164a}$$

$$\nabla\cdot c\boldsymbol{B} = 0, \tag{4.164b}$$

$$\nabla\wedge\boldsymbol{E} = \mathrm{i}\,k\,c\boldsymbol{B}, \tag{4.164c}$$

$$\nabla\wedge c\boldsymbol{B} = -\mathrm{i}\,k\,n^2\,\boldsymbol{E}, \tag{4.164d}$$

where n is the non-uniform refractive index of the medium. It is assumed that all field quantities vary in time like $\mathrm{e}^{-\mathrm{i}\,\omega t}$, where $\omega = kc$. Note that, in the following, k is the wavenumber in free space, rather than the wavenumber in the dielectric medium.

4.16 The W.K.B. approximation

Consider a radio wave which is vertically incident, from below, on the horizontally stratified ionosphere. Since the wave normal is initially aligned along the z-axis, and since $n = n(z)$, we expect all field components to be functions of z only, so

that

$$\frac{\partial}{\partial x} \equiv \frac{\partial}{\partial y} \equiv 0. \tag{4.165}$$

In this situation, Eqs. (4.164) reduce to $E_z = cB_z = 0$, with

$$-\frac{\partial E_y}{\partial z} = \mathrm{i}\, k\, cB_x, \tag{4.166a}$$

$$\frac{\partial\, cB_x}{\partial z} = -\mathrm{i}\, k\, n^2\, E_y, \tag{4.166b}$$

and

$$\frac{\partial E_x}{\partial z} = \mathrm{i}\, k\, cB_y, \tag{4.167a}$$

$$-\frac{\partial\, cB_y}{\partial z} = -\mathrm{i}\, k\, n^2\, E_x. \tag{4.167b}$$

Note that Eqs. (4.166) and (4.167) are isomorphic and completely independent of one another. It follows that, without loss of generality, we can assume that the wave is linearly polarized with its electric vector parallel to the y-axis. This means that we are only going to consider the solution of Eqs. (4.166). The solution of Eqs. (4.167) is of exactly the same form, except that it describes a linear polarized wave with its electric vector parallel to the x-axis.

Equations (4.166) can be combined to give

$$\frac{d^2 E_y}{dz^2} + k^2 n^2\, E_y = 0. \tag{4.168}$$

Since E_y is a function of z only, we now use the total derivative sign d/dz instead of the partial derivative sign $\partial/\partial z$. The solution of the above equation for the case of a uniform medium, where n is constant, is straightforward:

$$E_y = A\, \mathrm{e}^{\,\mathrm{i}\,\phi(z)}, \tag{4.169}$$

where A is a constant, and

$$\phi = \pm k\, n\, z. \tag{4.170}$$

Note that the $\mathrm{e}^{-\mathrm{i}\omega t}$ time dependence of all wave quantities is taken as read during this investigation. The solution (4.169) represents a *wave* of constant amplitude A and phase $\phi(z)$. According to Eq. (4.170), there are, in fact, two independent waves which can propagate through the medium in question. The upper sign corresponds to a wave which propagates vertically upwards, and the lower sign corresponds to a wave which propagates vertically downwards. Both waves propagate with the constant phase velocity c/n.

In general, if $n = n(z)$ the solution of Eq. (4.168) does not remotely resemble the wave-like solution (4.169). However, in the limit in which $n(z)$ is a "slowly varying" function of z (exactly how slowly varying is something which we shall establish later), we expect to recover wave-like solutions. Let us suppose that $n(z)$ is indeed a "slowly varying" function, and let us try substituting the wave solution (4.169) into Eq. (4.168). We obtain

$$\left(\frac{d\phi}{dz}\right)^2 = k^2 n^2 + \mathrm{i}\,\frac{d^2\phi}{dz^2}. \tag{4.171}$$

This is a non-linear differential equation which, in general, is very difficult to solve. However, we note that if n is a constant then $d^2\phi/dz^2 = 0$. It is, therefore, reasonable to suppose that if $n(z)$ is a "slowly varying" function then the last term on the right-hand side of the above equation can be regarded as being small. Thus, to a first approximation Eq. (4.171) yields

$$\frac{d\phi}{dz} \simeq \pm k\, n, \tag{4.172}$$

and

$$\frac{d^2\phi}{dz^2} \simeq \pm k\,\frac{dn}{dz}. \tag{4.173}$$

It is clear from a comparison of Eqs. (4.171) and (4.173) that $n(z)$ can be regarded as a "slowly varying" function of z as long as its variation length-scale is far longer than the wavelength of the wave. In other words, provided that $(dn/dz)/(k\,n^2) \ll 1$.

The second approximation to the solution is obtained by substituting Eq. (4.173)

into the right-hand side of Eq. (4.171):

$$\frac{d\phi}{dz} \simeq \pm \left(k^2 n^2 \pm \mathrm{i}\, k \frac{dn}{dz} \right)^{1/2}. \tag{4.174}$$

This gives

$$\frac{d\phi}{dz} \simeq \pm k\, n \left(1 \pm \frac{\mathrm{i}}{k\, n^2} \frac{dn}{dz} \right)^{1/2} \simeq \pm k\, n + \frac{\mathrm{i}}{2n} \frac{dn}{dz}, \tag{4.175}$$

where use has been made of the binomial expansion. The above expression can be integrated to give

$$\phi \sim \pm k \int^z n\, dz + \mathrm{i}\, \log(n^{1/2}). \tag{4.176}$$

Substitution of Eq. (4.176) into Eq. (4.169) yields the final result

$$E_y \simeq A\, n^{-1/2} \exp\left(\pm \mathrm{i}\, k \int^z n\, dz \right). \tag{4.177}$$

It follows from Eq. (4.166a) that

$$c B_x \simeq \mp A\, n^{1/2} \exp\left(\pm \mathrm{i}\, k \int^z n\, dz \right) - \frac{\mathrm{i}\, A}{2k\, n^{3/2}} \frac{dn}{dz} \exp\left(\pm \mathrm{i}\, k \int^z n\, dz \right). \tag{4.178}$$

Note that the second term is small compared to the first, and can usually be neglected.

Let us test to what extent the expression (4.177) is a good solution of Eq. (4.168) by substituting this expression into the left-hand side of the equation. The result is

$$\frac{A}{n^{1/2}} \left\{ \frac{3}{4} \left(\frac{1}{n} \frac{dn}{dz} \right)^2 - \frac{1}{2n} \frac{d^2 n}{dz^2} \right\} \exp\left(\pm \mathrm{i}\, k \int^z n\, dz \right). \tag{4.179}$$

This must be small compared with either term on the left-hand side of Eq. (4.168). Hence, the condition for Eq. (4.177) to be a good solution of Eq. (4.168) becomes

$$\frac{1}{k^2} \left| \frac{3}{4} \left(\frac{1}{n^2} \frac{dn}{dz} \right)^2 - \frac{1}{2n^3} \frac{d^2 n}{dz^2} \right| \ll 1. \tag{4.180}$$

The solutions

$$E_y \simeq A\, n^{-1/2} \exp\left(\pm \mathrm{i}\, k \int^z n\, dz\right), \tag{4.181a}$$

$$cB_x \simeq \mp A\, n^{1/2} \exp\left(\pm \mathrm{i}\, k \int^z n\, dz\right), \tag{4.181b}$$

to the non-uniform wave equations (4.166) are most commonly called the *W.K.B. solutions*, in honor of G. Wentzel, H.A. Kramers, and L. Brillouin, who are credited with independently discovering these solutions (in a quantum mechanical context) in 1926. Actually, H. Jeffries wrote a paper on these solutions (in a wave propagation context) in 1923. Hence, some people call these the W.K.B.J. solutions (or even the J.W.K.B. solutions). In fact, these solutions were first discussed by Liouville and Green in 1837, and again by Rayleigh in 1912. We shall refer to Eqs. (4.181) as the W.K.B. solutions, since this is what they are most commonly called. However, it should be understand that, in doing so, we are not making any statement as to the credit due to various scientists in discovering these solutions. After all, this is not a history of science course!

Recall, that when a propagating wave is normally incident on an *interface*, where the refractive index suddenly changes (for instance, when a light wave propagating in the air is normally incident on a glass slab), there is generally significant reflection of the wave. However, according to the W.K.B. solutions (4.181), when a propagating wave is normally incident on a medium in which the refractive index changes *slowly* along the direction of propagation of the wave, then the wave is not reflected at all. This is true even if the refractive index varies very substantially along the path of propagation of the wave, as long as it varies *slowly*. The W.K.B. solutions imply that as the wave propagates through the medium its wavelength gradually changes. In fact, the wavelength at position z is approximately $\lambda(z) = 2\pi/k\, n(z)$. Equations (4.181) also imply that the amplitude of the wave gradually changes as it propagates. In fact, the amplitude of the electric field component is inversely proportional to $n^{1/2}$, whereas the amplitude of the magnetic field component is directly proportional to $n^{1/2}$. Note, however, that the energy flux in the z-direction, given by the the Poynting vector $-(E_y B_x^* + E_y^* B_x)/(4\mu_0)$, remains constant (assuming that n is predominately real).

Of course, the W.K.B. solutions (4.181) are only *approximations*. In reality, a wave propagating into a medium in which the refractive index is a slowly varying function of position is subject to a small amount of reflection. However, it is easily demonstrated that the ratio of the reflected amplitude to the incident amplitude is of order $(dn/dz)/(k\,n^2)$. Thus, as long as the refractive index varies on a much longer length-scale than the wavelength of the radiation, the reflected wave is negligibly small. This conclusion remains valid as long as the inequality (4.180) is satisfied. There are two main reasons why this inequality might fail to be satisfied. First of all, if there is a localized region in the dielectric medium in which the refractive index suddenly changes (*i.e.*, if there is an interface), then (4.180) is likely to break down in this region, allowing strong reflection of the incident wave. Secondly, the inequality obviously breaks down in the vicinity of a point where $n = 0$. We would, therefore, expect strong reflection of the incident wave from such a point.

4.17 The reflection coefficient

Consider an ionosphere in which the refractive index is a slowly varying function of height z above the surface of the Earth. Let n^2 be positive for $z < z_0$, and negative for $z > z_0$. Suppose that an upgoing radio wave of amplitude E_0 is generated at ground level ($z = 0$). The complex amplitude of the wave in the region $0 < z < z_0$ is given by the upgoing W.K.B. solution

$$E_y = E_0\, n^{-1/2} \exp\left(\mathrm{i}\, k \int_0^z n\, dz \right), \tag{4.182a}$$

$$cB_x = -E_0\, n^{1/2} \exp\left(\mathrm{i}\, k \int_0^z n\, dz \right). \tag{4.182b}$$

The upgoing energy flux is given by $-(E_y B_x^* + E_y^* B_x)/(4\mu_0) = (\epsilon_0/\mu_0)^{1/2}\, |E_0|^2/2$. In the region $z > z_0$ the W.K.B. solutions take the form

$$E_y = A\, \mathrm{e}^{\,\mathrm{i}\,\pi/4}\, |n|^{-1/2} \exp\left(\pm k \int^z |n|\, dz \right), \tag{4.183a}$$

$$cB_x = \pm A\, \mathrm{e}^{-\mathrm{i}\pi/4}\, |n|^{1/2} \exp\left(\pm k \int^z |n|\, dz \right), \tag{4.183b}$$

where A is a constant. These solutions correspond to exponentially growing and decaying waves. Note that the magnetic components of the waves are in *phase quadrature* with the electric components. This implies that the Poynting fluxes of the waves are zero; *i.e.*, the waves do not transmit energy. Thus, there is a non-zero incident energy flux in the region $z < z_0$, and zero energy flux in the region $z > z_0$. Clearly, the incident wave is either absorbed or reflected in the vicinity of the plane $z = z_0$ (where $n = 0$). In fact, as we shall prove later on, the wave is *reflected*. The complex amplitude of the reflected wave in the region $0 < z < z_0$ is given by the downgoing W.K.B. solution

$$E_y = E_0\, R\, n^{-1/2} \exp\left(-\mathrm{i}\, k \int_0^z n\, dz\right), \tag{4.184a}$$

$$cB_x = E_0\, R\, n^{1/2} \exp\left(-\mathrm{i}\, k \int_0^z n\, dz\right), \tag{4.184b}$$

where R is the coefficient of reflection. Suppose, for the sake of argument, that the plane $z = z_0$ acts like a perfect conductor, so that $E_y(z_0) = 0$. It follows that

$$R = -\exp\left(2\mathrm{i}\, k \int_0^{z_0} n\, dz\right). \tag{4.185}$$

In fact, as we shall prove later on, the correct answer is

$$R = -\mathrm{i} \exp\left(2\mathrm{i}\, k \int_0^{z_0} n\, dz\right). \tag{4.186}$$

Thus, there is only a $-\pi/2$ phase shift at the reflection point, instead of the $-\pi$ phase shift which would be obtained if the plane $z = z_0$ acted like a perfect conductor.

4.18 Extension to oblique incidence

We have discussed the W.K.B. solutions for radio waves propagating vertically through an ionosphere whose refractive index varies slowly. Let us now generalize these solutions to allow for radio waves which propagate at an angle to the vertical axis.

The refractive index of the ionosphere varies continuously with height z. However, let us, for the sake of clarity, imagine that the ionosphere is replaced by a number of thin discrete strata in which the medium is homogeneous. By making these strata sufficiently thin and numerous we can approximate as closely as is desired to the real ionosphere. Suppose that a plane wave is incident on the ionosphere, from below, and suppose that the wave normal lies in the x-z plane and makes an angle θ_I with the vertical axis. At the lower boundary of the first stratum the wave is partially reflected and partially transmitted. The transmitted wave is partially reflected and partially transmitted at the second boundary between the strata, and so on. However, in the limit of many strata, where the difference in refractive indices between neighbouring strata is very small, the amount of reflection at the boundaries becomes negligible. In the nth stratum, let n_n be the refractive index, and let θ_n be the angle between the wave normal and the vertical axis. According to Snell's law,

$$n_{n-1} \sin\theta_{n-1} = n_n \sin\theta_n. \tag{4.187}$$

Below the ionosphere $n = 1$, and so

$$n_n \sin\theta_n = \sin\theta_I. \tag{4.188}$$

For a wave in the nth stratum, any field quantity depends on z and x through factors

$$A \exp\left[\,\mathrm{i}\, k\, n_n (\pm z \cos\theta_n + x \sin\theta_n)\right], \tag{4.189}$$

where A is a constant. The $\pm$ signs denote upgoing and downgoing waves, respectively. When the operator $\partial/\partial x$ acts on the above expression, it is equivalent to multiplication by $\mathrm{i}\, k\, n_n \sin\theta_n = \mathrm{i}\, k \sin\theta_I$, which is independent of x and z. It is convenient to use the notation $S = \sin\theta_I$. Hence, we may write symbolically

$$\frac{\partial}{\partial x} \equiv \mathrm{i}\, k\, S, \tag{4.190a}$$

$$\frac{\partial}{\partial y} \equiv 0. \tag{4.190b}$$

This result is true no matter how thin the strata are, so it must also hold for the real ionosphere. Note that, according to Snell's law, if the wave normal starts off

in the x-z plane then it will remain in this plane as it propagates through the ionosphere.

Equations (4.164) and (4.190) can be combined to give

$$-\frac{\partial E_y}{\partial z} = \mathrm{i}\, k\, cB_x, \tag{4.191a}$$

$$\mathrm{i}\, kS\, E_y = \mathrm{i}\, k\, cB_z, \tag{4.191b}$$

$$\frac{\partial\, cB_x}{\partial z} - \mathrm{i}\, kS\, cB_z = -\mathrm{i}\, k\, n^2\, E_y, \tag{4.191c}$$

and

$$\frac{\partial E_x}{\partial z} - \mathrm{i}\, kS\, E_z = \mathrm{i}\, k\, cB_y, \tag{4.192a}$$

$$-\frac{\partial\, cB_y}{\partial z} = -\mathrm{i}\, k\, n^2\, E_x, \tag{4.192b}$$

$$\mathrm{i}\, kS\, cB_y = -\mathrm{i}\, k\, n^2\, E_z. \tag{4.192c}$$

As before, Maxwell's equations can be split into two independent groups, corresponding to two independent polarizations of radio waves propagating through the ionosphere. For the first set of equations, the electric field is always parallel to the y-axis. The corresponding waves are, therefore, said to be *horizontally polarized*. For the second set of equations, the electric field always lies in the x-z plane. The corresponding waves are, therefore, said to be *vertically polarized* (*n.b.*, the term "vertically polarized" does not necessarily imply that the electric field is parallel to the vertical axis). Note that the equations governing horizontally polarized waves are *not* isomorphic to those governing vertically polarized waves, so both types of waves must be dealt with separately.

For the case of horizontally polarized waves, Eqs. (4.191b) and (4.191c) yield

$$\frac{\partial\, cB_x}{\partial z} = -\mathrm{i}\, k\, q^2\, E_y, \tag{4.193}$$

where

$$q^2 = n^2 - S^2. \tag{4.194}$$

The above equation can be combined with Eq. (4.191a) to give

$$\frac{\partial^2 E_y}{\partial z^2} + k^2 q^2 \, E_y = 0. \tag{4.195}$$

Equations (4.193) and (4.195) have exactly the same form as Eqs. (4.166b) and (4.168), except that n^2 is replaced by q^2, so the results of Section 4.16 can be immediately employed to find the W.K.B. solutions, which take the form

$$E_y = A \, q^{-1/2} \exp\left(\pm \mathrm{i} \, k \int^z q \, dz\right), \tag{4.196a}$$

$$cB_x = \mp A \, q^{1/2} \exp\left(\pm \mathrm{i} \, k \int^z q \, dz\right), \tag{4.196b}$$

where A is a constant. Of course, both expressions should also contain a multiplicative factor $\mathrm{e}^{\,\mathrm{i}\,(kS\,x-\omega t)}$, but this is usually omitted for the sake of clarity. By analogy with Eq. (4.180), the W.K.B. solutions are valid as long as

$$\frac{1}{k^2}\left|\frac{3}{4}\left(\frac{1}{q^2}\frac{dq}{dz}\right)^2 - \frac{1}{2q^3}\frac{d^2q}{dz^2}\right| \ll 1. \tag{4.197}$$

This inequality clearly fails in the vicinity of $q = 0$, no matter how slowly q varies with z. Hence, $q = 0$, or $n^2 = S^2$, specifies the height at which reflection takes place. By analogy with Eq. (4.186), the reflection coefficient at ground level ($z = 0$) is given by

$$R = -\mathrm{i} \exp\left(2\mathrm{i} \, k \int_0^{z_0} q \, dz\right), \tag{4.198}$$

where z_0 is the height at which $q = 0$.

For the case of vertical polarization, Eqs. (4.192a) and (4.192c) yield

$$\frac{\partial E_x}{\partial z} = \mathrm{i} \, k \, \frac{q^2}{n^2} \, cB_y. \tag{4.199}$$

This equation can be combined with Eq. (4.192b) to give

$$\frac{\partial^2 B_y}{\partial z^2} - \frac{1}{n^2}\frac{d(n^2)}{dz}\frac{\partial B_y}{\partial z} + k^2 q^2 \, B_y = 0. \tag{4.200}$$

Clearly, the differential equation which governs the propagation of vertically polarized waves is considerably more complicated than the corresponding equation for horizontally polarized waves.

The W.K.B. solution for vertically polarized waves is obtained by substituting the wave-like solution

$$c B_y = A \, \mathrm{e}^{\,\mathrm{i}\,\phi(z)}, \tag{4.201}$$

where A is a constant and $\phi(z)$ is the generalized phase, into Eq. (4.200). The differential equation thus obtained for the phase is

$$\mathrm{i}\,\frac{d^2\phi}{dz^2} - \left(\frac{d\phi}{dz}\right)^2 - \frac{\mathrm{i}}{n^2}\frac{d(n^2)}{dz}\frac{d\phi}{dz} + k^2 q^2\,\phi = 0. \tag{4.202}$$

Since the medium is slowly varying, the first and third term in the above equation are small, and so to a first approximation

$$\frac{d\phi}{dz} = \pm k\,q, \tag{4.203a}$$

$$\frac{d^2\phi}{dz^2} = \pm k\,\frac{dq}{dz}. \tag{4.203b}$$

These expressions can be inserted into the first and third terms of Eq. (4.202) to give the second approximation

$$\frac{d\phi}{dz} = \pm\left[k^2 q^2 \pm \mathrm{i}\,k\left(\frac{dq}{dz} - \frac{2q}{n}\frac{dn}{dz}\right)\right]^{1/2}. \tag{4.204}$$

The final two terms on the right-hand side of the above equation are small, so expanding the right-hand side using the binomial theorem yields

$$\frac{d\phi}{dz} = \pm kq + \frac{\mathrm{i}}{2q}\frac{dq}{dz} - \frac{\mathrm{i}}{n}\frac{dn}{dz}. \tag{4.205}$$

This expression can be integrated, and the result inserted into Eq. (4.201), to give the W.K.B. solution

$$c B_y = A\,n\,q^{-1/2}\,\exp\left(\pm\mathrm{i}\,k\int^z q\,dz\right). \tag{4.206}$$

The corresponding W.K.B. solution for E_x is obtained from Eq. (4.199):

$$E_x = \pm A\, n^{-1}\, q^{1/2}\, \exp\left(\pm \mathrm{i}\, k \int^z q\, dz\right). \tag{4.207}$$

Here, any terms involving derivatives of n and q have been neglected.

Substituting Eq. (4.206) into the differential equation (4.200), and demanding that the result be small compared to the original terms in the differential equation, yields the following condition for the validity of the above W.K.B. solutions:

$$\frac{1}{k^2}\left|\frac{3}{4}\left(\frac{1}{q^2}\frac{dq}{dz}\right)^2 - \frac{1}{2q^3}\frac{d^2q}{dz^2} + \frac{1}{q^2}\left[\frac{1}{n}\frac{d^2n}{dz^2} - 2\left(\frac{1}{n}\frac{dn}{dz}\right)^2\right]\right| \ll 1. \tag{4.208}$$

This criterion fails close to $q = 0$, no matter how slowly n and q vary with z. Hence, $q = 0$ gives the height at which reflection takes place. The condition also fails close to $n = 0$, which does not correspond to the reflection level. If, as is usually the case, the electron density in the ionosphere increases monotonically with height, then the level where $n = 0$ lies above the reflection level, where $q = 0$. If the two levels are well separated then the reflection process is unaffected by the failure of the above inequality at the level $n = 0$, and the reflection coefficient is given by Eq. (4.198), just as for the case of horizontal polarization. If, however, the level $n = 0$ lies close to the level $q = 0$ then the reflection coefficient may be affected, and a more accurate treatment of the differential equation (4.200) is required in order to obtain the true value of the reflection coefficient.

4.19 Pulse propagation in the ionosphere

Suppose that we possess a generator of radio waves which sends radio pulses vertically upwards into the ionosphere. For the sake of argument, we shall assume that these pulses are linearly polarized such that the electric field vector lies parallel to the y-axis. The pulse structure can be represented as

$$E_y(t) = \int_{-\infty}^{\infty} F(\omega)\, \mathrm{e}^{-\mathrm{i}\,\omega\, t}\, d\omega, \tag{4.209}$$

where $E_y(t)$ is the electric field produced by the generator (*i.e.*, the field at $z = 0$). Suppose that the pulse is a signal of roughly constant (angular) frequency ω_0, which lasts a time T, where T is long compared to $1/\omega_0$. It follows that $F(\omega)$ possesses narrow maxima around $\omega = \pm\omega_0$. In other words, only those frequencies which lie very close to the central frequency ω_0 play a significant role in the propagation of the pulse.

Each component frequency of the pulse yields a wave which travels independently up into the ionosphere, in a manner specified by the appropriate W.K.B. solution [see Eqs. (4.181)]. Thus, if Eq. (4.209) specifies the signal at ground level ($z = 0$), then the signal at height z is given by

$$E_y(z,t) = \int_{-\infty}^{\infty} \frac{F(\omega)}{n^{1/2}(\omega,z)}\, \mathrm{e}^{\,\mathrm{i}\,\phi(\omega,z,t)}\, d\omega, \tag{4.210}$$

where

$$\phi(\omega,z,t) = \frac{\omega}{c}\int_0^z n(\omega,z)\, dz - \omega\, t. \tag{4.211}$$

Here, we have used $k = \omega/c$.

Equation (4.210) can be regarded as a contour integral in ω-space. The quantity $F/n^{1/2}$ is a relatively slowly varying function of ω, whereas the phase ϕ is a large and rapidly varying function of ω. As described in Section 4.11, the rapid oscillations of $\exp(\,\mathrm{i}\,\phi)$ over most of the path of integration ensure that the integrand averages almost to zero. However, this cancellation argument does not apply to those points on the path of integration where the phase is *stationary*; *i.e.*, those points where $\partial\phi/\partial\omega = 0$. It follows that the left-hand side of Eq. (4.210) averages to a very small value, expect for those special values of z and t at which one of the points of stationary phase in ω-space coincides with one of the peaks of $F(\omega)$. The locus of these special values of z and t can obviously be regarded as the equation of motion of the pulse as it propagates through the ionosphere. Thus, the equation of motion is specified by

$$\left(\frac{\partial\phi}{\partial\omega}\right)_{\omega=\omega_0} = 0, \tag{4.212}$$

which yields

$$t = \frac{1}{c}\int_0^z \left[\frac{\partial(\omega\, n)}{\partial\omega}\right]_{\omega=\omega_0} dz. \tag{4.213}$$

Suppose that the z-velocity of a pulse of central frequency ω_0 at height z is given by $u_z(\omega_0, z)$. The differential equation of motion of the pulse is then $dt = dz/u_z$. This can be integrated, using the boundary condition $z = 0$ at $t = 0$, to give the full equation of motion:

$$t = \int_0^z \frac{dz}{u_z}. \tag{4.214}$$

A comparison of Eqs. (4.213) and (4.214) yields

$$u_z(\omega_0, z) = c \Bigg/ \left\{\frac{\partial[\omega\, n(\omega, z)]}{\partial\omega}\right\}_{\omega=\omega_0}. \tag{4.215}$$

The velocity u_z is usually called the *group velocity*. It is easily demonstrated that the above expression for the group velocity is entirely consistent with that given previously [see Eq. (4.135)].

The dispersion relation (4.164) yields

$$n(\omega, z) = \left(1 - \frac{{\omega_p}^2(z)}{\omega^2}\right)^{1/2}, \tag{4.216}$$

in the limit where electron collisions are negligible. The phase velocity of radio waves of frequency ω propagating vertically through the ionosphere is given by

$$v_z(\omega, z) = \frac{c}{n(\omega, z)} = c\left(1 - \frac{{\omega_p}^2(z)}{\omega^2}\right)^{-1/2}. \tag{4.217}$$

According to Eqs. (4.215) and (4.216), the corresponding group velocity is

$$u_z(\omega, z) = c\left(1 - \frac{{\omega_p}^2(z)}{\omega^2}\right)^{1/2}. \tag{4.218}$$

It follows that

$$v_z\, u_z = c^2. \tag{4.219}$$

Note that as the reflection point $z = z_0$ [defined as the solution of $\omega = \omega_p(z_0)$] is approached from below, the phase velocity tends to infinity, whereas the group velocity tends to zero.

Let τ be the time taken for the pulse to travel from the ground to the reflection level, and back to the ground again. The product $c\,\tau/2$ is termed the *equivalent height of reflection*, and is denoted $h(\omega)$, since it is a function of the pulse frequency, ω. The equivalent height is the height to which the pulse would have to go if it always traveled with the velocity c. Since we know that a pulse of dominant frequency ω propagates at height z with the z-velocity $u_z(\omega, z)$ (this is true for both upgoing and downgoing pulses), and also that the pulse is reflected at the height $z_0(\omega)$, where $\omega = \omega_p(z_0)$, it follows that

$$\tau = 2 \int_0^{z_0(\omega)} \frac{dz}{u_z(\omega, z)}. \tag{4.220}$$

Hence,

$$h(\omega) = \int_0^{z_0(\omega)} \frac{c}{u_z(\omega, z)}\, dz. \tag{4.221}$$

Note that the equivalent height of reflection, $h(\omega)$, is always *greater* than the actual height of reflection, $z_0(\omega)$, since the group velocity u_z is always less than the velocity of light. The above equation can be combined with Eq. (4.218) to give

$$h(\omega) = \int_0^{z_0(\omega)} \left(1 - \frac{{\omega_p}^2(z)}{\omega^2}\right)^{-1/2} dz. \tag{4.222}$$

Note that the integrand diverges as the reflection point is approached, but the integral remains finite.

4.20 Determining the ionospheric electron density profile

We can measure the equivalent height of the ionosphere in a fairly straightforward manner, by timing how long it takes a radio pulse fired vertically upwards

to return to ground level again. We can, therefore, determine the function $h(\omega)$ experimentally by performing this procedure many times over, using pulses of different central frequencies. But, is it possible to use this information to determine the number density of free electrons in the ionosphere as a function of height? In mathematical terms, the problem is as follows. Does a knowledge of the function

$$h(\omega) = \int_0^{z_0(\omega)} \frac{\omega}{[\omega^2 - {\omega_p}^2(z)]^{1/2}}\, dz, \tag{4.223}$$

where ${\omega_p}^2(z_0) = \omega^2$, necessarily imply a knowledge of the function ${\omega_p}^2(z)$? Note, of course, that ${\omega_p}^2(z) \propto N(z)$.

Let $\omega^2 = v$ and ${\omega_p}^2(z) = u(z)$. Equation (4.223) then becomes

$$v^{-1/2}\, h(v^{1/2}) = \int_0^{z_0(v^{1/2})} \frac{dz}{[v - u(z)]^{1/2}}, \tag{4.224}$$

where $u(z_0) = v$, and $u(z) < v$ for $0 < z < z_0$. Let us multiply both sides of the above equation by $(w - v)^{-1/2}/\pi$ and integrate from $v = 0$ to w. We obtain

$$\frac{1}{\pi}\int_0^w v^{-1/2}\,(w - v)^{-1/2}\, h(v^{1/2})\, dv = \frac{1}{\pi}\int_0^w \left[\int_0^{z_0(v^{1/2})} \frac{dz}{(w - v)^{1/2}\,(v - u)^{1/2}}\right] dv. \tag{4.225}$$

Consider the double integral on the right-hand side. The region of v-z space over which this integral is performed is sketched in Fig. 15. It can be seen that, as long as $z_0(v^{1/2})$ is a *monotonically increasing* function of z, we can swap the order of integration to give

$$\frac{1}{\pi}\int_0^{z_0(w^{1/2})} \left[\int_{u(z)}^w \frac{dv}{(w - v)^{1/2}\,(v - u)^{1/2}}\right] dz. \tag{4.226}$$

Here, we have used the fact that the curve $z = z_0(v^{1/2})$ is identical with the curve $v = u(z)$. Note that if $z_0(v^{1/2})$ is *not* a monotonically increasing function of v then we can still swap the order of integration, but the limits of integration are, in general, far more complicated than those indicated above. The integral over v in the above expression can be evaluated using standard methods (by making the

substitution $v = w\,\cos^2\theta + u\,\sin^2\theta$): the result is simply π. Thus, the expression (4.226) reduces to $z_0(w^{1/2})$. It follows from Eq. (4.225) that

$$z_0(w^{1/2}) = \frac{1}{\pi}\int_0^w v^{-1/2}\,(w-v)^{-1/2}\,h(v^{1/2})\,dv. \tag{4.227}$$

Making the substitutions $v = w\,\sin^2\alpha$ and $w^{1/2} = \omega$, we obtain

$$z_0(\omega) = \frac{2}{\pi}\int_0^{\pi/2} h(\omega\,\sin\alpha)\,d\alpha. \tag{4.228}$$

By definition, $\omega = \omega_p$ at the reflection level $z = z_0$. Hence, the above equation reduces to

$$z(\omega_p) = \frac{2}{\pi}\int_0^{\pi/2} h(\omega_p\,\sin\alpha)\,d\alpha. \tag{4.229}$$

Thus, we can obtain z as a function of ω_p (and, hence, ω_p as a function of z) simply by taking the appropriate integral of the experimentally determined function $h(\omega)$. Since $\omega_p(z) \propto \sqrt{N(z)}$, this means that we can determine the electron number density profile in the ionosphere provided we know the variation of the equivalent height of the ionosphere with pulse frequency. The constraint that $z_0(\omega)$ must be a monotonically increasing function of ω translates to the constraint that $N(z)$ must be a monotonically increasing function of z. Note that we can still determine $N(z)$ from $h(\omega)$ for the case where the former function is non-monotonic, it is just a far more complicated procedure than that outlined above. Incidentally, the technique by which we have inverted Eq. (4.222), which specifies $h(\omega)$ as some integral over $\omega_p(z)$, to give $\omega_p(z)$ as some integral over $h(\omega)$ is known as *Abel inversion*.

4.21 Ray tracing in the ionosphere

Suppose that we possess a radio antenna which is capable of launching radio waves of constant frequency ω into the ionosphere at an angle to the vertical. Let us consider the paths traced out by these waves in the x-z plane. For the sake of simplicity, we shall assume that the waves are horizontally polarized, so that the

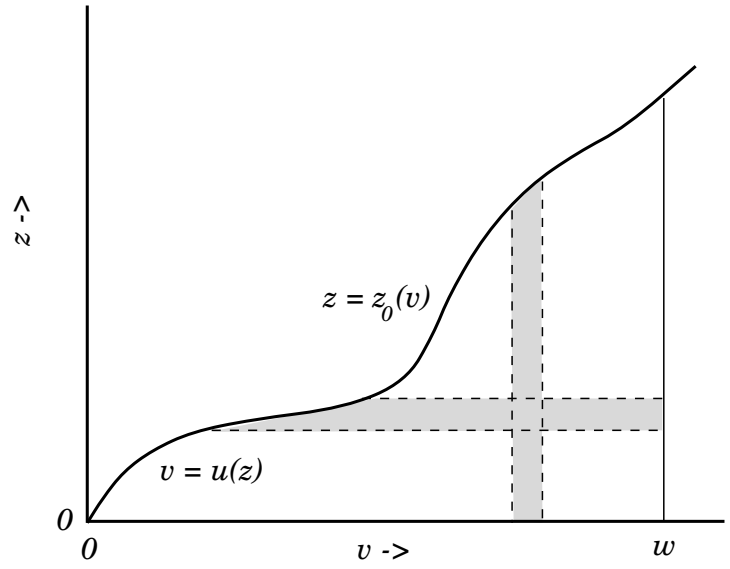

Figure 15: A sketch of the region of v-z space over which the integral on the right-hand side of Eq. (4.223) is evaluated

electric field vector always lies parallel to the y-axis. The signal emitted by the antenna (located at $z = 0$) can be represented as

$$E_y(x) = \int_0^1 F(S)\, \mathrm{e}^{\,\mathrm{i}\, k\, S\, x}\, dS, \tag{4.230}$$

where $k = \omega/c$. Here, the $\mathrm{e}^{-\mathrm{i}\,\omega\, t}$ time dependence of the signal is neglected for the sake of clarity. Suppose that the signal emitted by the antenna is mostly concentrated in a direction making an angle θ_I with the vertical. It follows that $F(S)$ possesses a narrow maximum around $S = S_0$, where $S_0 = \sin\theta_I$.

If Eq. (4.230) represents the signal at ground level, then the signal at height z in the ionosphere is easily obtained by making use of the W.K.B. solutions for horizontally polarized waves described in Section 4.18. We obtain

$$E_y(x, z) = \int_0^1 \frac{F(S)}{q^{1/2}(z, S)}\, \mathrm{e}^{\,\mathrm{i}\,\phi(x, z, S)}\, dS, \tag{4.231}$$

where

$$\phi(x, z, S) = k \int_0^z q(z, S)\, dz + k\, S\, x. \tag{4.232}$$

Equation (4.231) is basically a line integral in S-space. The quantity $F/q^{1/2}$ is a relatively slowly varying function of S, whereas the phase ϕ is a large and rapidly varying function of S. As described in Section 4.11, the rapid oscillations of $\exp(\mathrm{i}\,\phi)$ over most of the path of integration ensure that the integrand averages almost to zero. In fact, only those points on the path of integration where the phase is stationary (*i.e.*, where $\partial\phi/\partial S = 0$) make a significant contribution to the integral. It follows that the left-hand side of Eq. (4.231) averages to a very small value, except for those special values of x and z at which one of the points of stationary phase in S-space coincides with the peak of $F(S)$. The locus of these special values of x and z can clearly be regarded as the trajectory of the radio signal emitted by the antenna as it passes through the ionosphere. Thus, the signal trajectory is specified by

$$\left(\frac{\partial\phi}{\partial S}\right)_{S=S_0} = 0, \tag{4.233}$$

which yields

$$x = -\int_0^z \left(\frac{\partial q}{\partial S}\right)_{S=S_0} dz. \tag{4.234}$$

We can think of this equation as tracing the path of a *ray* of radio frequency radiation, emitted by the antenna at an angle θ_I to the vertical (where $S_0 = \sin\theta_I$), as it propagates through the ionosphere.

Now

$$q^2 = n^2 - S^2, \tag{4.235}$$

so the ray tracing equation becomes

$$x = S\int_0^z \frac{dz}{\sqrt{n^2(z) - S^2}}, \tag{4.236}$$

where S is the sine of the initial (*i.e.*, at the antenna) angle of incidence of the ray with respect to the vertical axis. Of course, Eq. (4.236) only holds for *upgoing* rays. For *downgoing* rays, a simple variant of the previous analysis using the downgoing W.K.B. solutions yields

$$x = S\int_0^{z_0(S)} \frac{dz}{\sqrt{n^2(z) - S^2}} + S\int_z^{z_0(S)} \frac{dz}{\sqrt{n^2(z) - S^2}}, \tag{4.237}$$

where $n(z_0) = S$. Thus, the ray ascends into the ionosphere after being launched from the antenna, reaches a maximum height z_0 above the surface of the Earth, and then starts to descend. The ray eventually intersects the Earth's surface again a horizontal distance

$$x_0 = 2S \int_0^{z_0(S)} \frac{dz}{\sqrt{n^2(z) - S^2}} \tag{4.238}$$

away from the antenna.

The angle ξ which the ray makes with the vertical is given by $\tan\xi = dx/dz$. It follows from Eqs. (4.236) and (4.237) that

$$\tan\xi = \pm \frac{S}{\sqrt{n^2(z) - S^2}} \tag{4.239}$$

where the upper and lower signs correspond to the upgoing and downgoing parts of the ray trajectory, respectively. Note that $\xi = \pi/2$ at the reflection point, where $n = S$. Thus, the ray is horizontal at the reflection point.

Let us investigate the reflection process in more detail. In particular, we wish to prove that radio waves are reflected at the $q = 0$ surface, rather than being absorbed. We would also like to understand the origin of the $-\pi/2$ phase shift of radio waves at reflection which is evident in Eq. (4.198). In order to achieve these goals, we shall need to review the mathematics of asymptotic series.

4.22 Asymptotic series: A mathematical aside

It is often convenient to expand a function of the complex variable $f(z)$ in inverse powers of z:

$$f(z) = \phi(z) \left[A_0 + \frac{A_1}{z} + \frac{A_2}{z^2} + \cdots \right], \tag{4.240}$$

where $\phi(z)$ is a function whose behaviour for large values of z is known. If $f(z)/\phi(z)$ is singular as $|z| \to \infty$ then the above series diverges. Nevertheless, under certain circumstances, the series may still be useful.

The circumstance needed to make this possible is that the difference between $f(z)/\phi(z)$ and the first $n+1$ terms of the series be of order $1/z^{n+1}$, so that for sufficiently large z this difference becomes vanishingly small. More precisely, the series is said to represent $f(z)/\phi(z)$ *asymptotically*, that is

$$f(z) \simeq \phi(z) \sum_{p=0}^{\infty} \frac{A_p}{z^p}, \tag{4.241}$$

provided that

$$\lim_{|z|\to\infty} \left\{ z^n \left[\frac{f(z)}{\phi(z)} - \sum_{p=0}^{n} \frac{A_p}{z^p} \right] \right\} \to 0. \tag{4.242}$$

In other words, for a given value of n, the first $n+1$ terms of the series may be made as close as desired to the ratio $f(z)/\phi(z)$ by making z sufficiently large. For each value of z and n there is an error of order $1/z^{n+1}$. Since the series actually diverges, there is an optimum number of terms in the series used to represent $f(z)/\phi(z)$ for a given value of z. Associated with this is an unavoidable error. As z increases, the optimal number of terms increases and the error decreases.

Consider a simple example. The exponential integral is defined

$$\mathrm{E}_1(x) = \int_x^{\infty} \frac{\mathrm{e}^{-t}}{t}\, dt. \tag{4.243}$$

The asymptotic series for this function can be generated via a series of partial integrations. For example,

$$\mathrm{E}_1(x) = \frac{\mathrm{e}^{-x}}{x} - \int_x^{\infty} \frac{\mathrm{e}^{-t}}{t^2}\, dt. \tag{4.244}$$

Continuing this procedure yields

$$\begin{aligned} \mathrm{E}_1(x) \quad = \quad & \frac{\mathrm{e}^{-x}}{x} \left[1 - \frac{1}{x} + \frac{2!}{x^2} - \frac{3!}{x^3} + \cdots + \frac{(-1)^n n!}{x^n} \right] \\ & + (-1)^{n+1}(n+1)! \int_x^{\infty} \frac{\mathrm{e}^{-t}}{t^{n+2}}\, dt. \end{aligned} \tag{4.245}$$

The infinite series obtained by taking the limit $n \to \infty$ diverges, since the Cauchy convergence test yields

$$\lim_{n\to\infty} \left| \frac{u_{n+1}}{u_n} \right| = \lim_{n\to\infty} \left[\frac{n}{x} \right] \to \infty. \tag{4.246}$$

Note that two successive terms in the series become equal in magnitude for $n = x$, indicating that the optimum number of terms for a given x is roughly the integer nearest x. To prove that the series is asymptotic, we need to show that

$$\lim_{x\to 0} x^{n+1} \mathrm{e}^x (-1)^{n+1} (n+1)! \int_x^\infty \frac{\mathrm{e}^{-t}}{t^{n+2}}\, dt = 0. \tag{4.247}$$

This immediately follows, since

$$\int_x^\infty \frac{\mathrm{e}^{-t}}{t^{n+2}}\, dt < \frac{1}{x^{n+2}} \int_x^\infty \mathrm{e}^{-t}\, dt = \frac{\mathrm{e}^{-x}}{x^{n+2}}. \tag{4.248}$$

Thus, the error involved in using the first n terms is less than $(n+1)!\,\mathrm{e}^{-x}/x^{n+2}$, which is exactly the next term in the series. We can see that as n increases, this estimate of the error first decreases and then increases without limit. In order to visualize this phenomenon more exactly, let $f(x) = x \exp(x)\, \mathrm{E}(x)$, and let

$$f_n(x) = \sum_{p=0}^{n} \frac{(-1)^p p!}{x^p} \tag{4.249}$$

be the asymptotic series representation of this function which contains $n+1$ terms. Figure 16 shows the relative error in the asymptotic series $|f_n(x) - f(x)|/f(x)$ plotted as a function of the approximate number of terms in the series n for $x = 10$. It can be seen that as n increases the error initially falls, reaches a minimum value at about $n = 10$, and then increases rapidly. Clearly, the optimum number of terms in the asymptotic series used to represent $f(10)$ is about 10.

Asymptotic series are fundamentally different to conventional power law expansions, such as

$$\sin z = z - \frac{z^3}{3!} + \frac{z^5}{5!} - \frac{z^7}{7!} + \cdots. \tag{4.250}$$

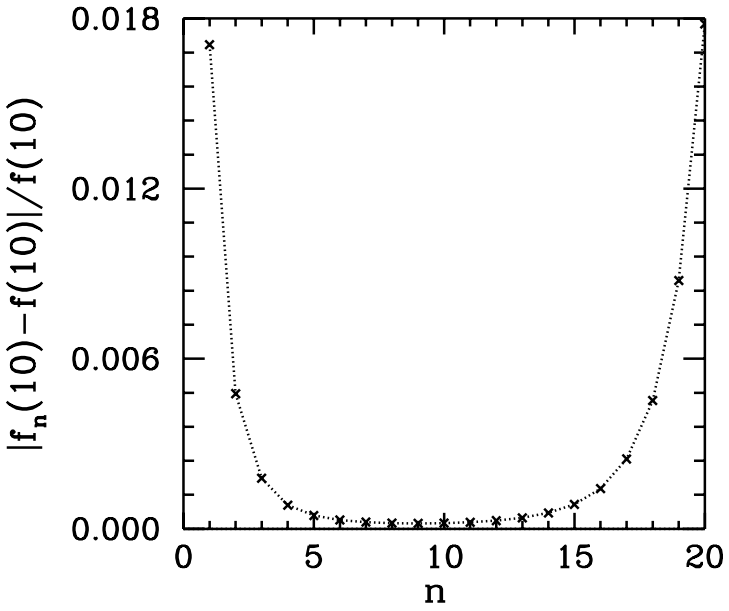

Figure 16: The relative error in a typical asymptotic series plotted as a function of the number of terms in the series

This series representation of $\sin z$ *converges* absolutely for all finite values of z. Thus, at any z the error associated with the series can be made as small as is desired by including a sufficiently large number of terms. In other words, unlike an asymptotic series, there is no intrinsic, or unavoidable, error associated with a convergent series. It follows that a convergent power law series representation of a function is *unique* inside the domain of convergence of the series. On the other hand, an asymptotic series representation of a function is *not unique*. It is perfectly possible to have two different asymptotic series representations of the same function, as long as the difference between the two series is less than the intrinsic error associated with each series. Furthermore, it is often the case that *different* asymptotic series are used to represent the *same* single-valued analytic function in different regions of the complex plane.

For example, consider the asymptotic expansion of the confluent hypergeometric function $F(a, c, z)$. This function is the solution of the differential equation

$$z F'' + (c - z) F' - a F = 0 \tag{4.251}$$

which is analytic at $z = 0$ [in fact, $F(a, c, 0) = 1$]. Here, $'$ denotes d/dz. The

asymptotic expansion of $F(a,c,z)$ takes the form:

$$\frac{\Gamma(a)\Gamma(c-a)}{\Gamma(c)}\,F(a,c,z) \simeq \Gamma(c-a)\,z^{a-c}\,\mathrm{e}^{z}\,[1+O(1/z)] + \Gamma(a)\,z^{-a}\,\mathrm{e}^{-\mathrm{i}\,\pi a}\,[1+O(1/z)] \quad (4.252a)$$

for $-\pi < \arg(z) < 0$, and

$$\frac{\Gamma(a)\Gamma(c-a)}{\Gamma(c)}\,F(a,c,z) \simeq \Gamma(c-a)\,z^{a-c}\,\mathrm{e}^{z}\,[1+O(1/z)] + \Gamma(a)\,z^{-a}\,\mathrm{e}^{\mathrm{i}\,\pi a}\,[1+O(1/z)] \quad (4.252b)$$

for $0 < \arg(z) < \pi$, and

$$\frac{\Gamma(a)\Gamma(c-a)}{\Gamma(c)}\,F(a,c,z) \simeq \Gamma(c-a)\,z^{a-c}\,\mathrm{e}^{-\mathrm{i}\,2\pi(a-c)}\,\mathrm{e}^{z}\,[1+O(1/z)] + \Gamma(a)\,z^{-a}\,\mathrm{e}^{\mathrm{i}\,\pi a}\,[1+O(1/z)] \quad (4.252c)$$

for $\pi < \arg(z) < 2\pi$, *etc.* It can be seen that the expansion consists of a linear combination of two asymptotic series (only the first term in each series is shown). For $|z| \gg 1$, the first series is exponentially larger than the second whenever $\mathrm{Re}(z) > 0$. We say that the first series is *dominant* in this region, whereas the second series is *subdominant*. Likewise, the first series is exponentially smaller than the second whenever $\mathrm{Re}(z) < 0$. We say that the first series is subdominant and the second series is dominant in this region.

Consider a region in which one or other of the series is dominant. Strictly speaking, it is not mathematically consistent to include the subdominant series in the asymptotic expansion because its contribution is actually less than the intrinsic error associated with the dominant series [this error is $O(1/z)$ times the dominant series, since we are only including the first term in this series]. Thus, at a general point in the complex plane the asymptotic expansion simply consists of the dominant series. However, this is not the case in the immediate vicinity of the lines $\mathrm{Re}(z) = 0$: these are called the *anti-Stokes lines*. When an anti-Stokes line is crossed, a dominant series becomes subdominant and *vice versa*. In

the immediate vicinity of an anti-Stokes line neither series is dominant, so it is mathematically consistent to include both series in the asymptotic expansion.

The hypergeometric function $F(a,c,z)$ is a perfectly well behaved, single-valued, analytic function in the complex plane. However, our two asymptotic series are, in general, multi-valued functions in the complex plane [see Eq. (4.252a)]. Can a single-valued function be represented asymptotically by a multi-valued function? The short answer is no. We have to employ different combinations of the two series in different regions of the complex plane in order to ensure that $F(a,c,z)$ remains single-valued. Equations (4.252) show how this is achieved. Basically, the coefficient in front of the subdominant series changes *discontinuously* at certain values of $\arg(z)$. This is perfectly consistent with $F(a,c,z)$ being an analytic function because the subdominant series is "invisible"; *i.e.*, the contribution of the subdominant series to the asymptotic solution falls below the intrinsic error associated with the dominant series, so it does not really matter if the coefficient in front of the former series changes discontinuously. Imagine tracing a large circle, centred on the origin, in the complex plane. Close to an anti-Stokes line, neither series is dominant, so we must include both series in the asymptotic expansion. As we move away from the anti-Stokes line, one series becomes dominant, which means that the other series becomes subdominant and, therefore, drops out of our asymptotic expansion. Eventually, we approach a second anti-Stokes line, and the subdominant series reappears in our asymptotic expansion. However, the coefficient in front of the subdominant series when it reappears is different to that which it had when it disappeared. This new coefficient is carried across the second anti-Stokes line into the region where the subdominant series becomes dominant. In this new region, the dominant series becomes subdominant and disappears from our asymptotic expansion. Eventually, a third anti-Stokes line is approached and the series reappears, but, again, with a different coefficient in front. The jumps in the coefficients of the subdominant series are chosen in such a manner that if we perform a complete circuit in the complex plane then the value of the asymptotic expansion is the same at the beginning and the end points. In other words, the asymptotic expansion is single-valued, despite the fact that it is built up out of two asymptotic series which are not single-valued. The jumps in the coefficient of the subdominant series, which are needed to keep the asymptotic expansion single-valued, are called *Stokes phenomena*, after the celebrated nineteenth century British mathematician Sir George Gabriel Stokes,

who first drew attention to this effect.

Where exactly does the jump in the coefficient of the subdominant series occur? All we can really say is "somewhere in the region between two anti-Stokes lines where the series in question is subdominant." The problem is that we only retain the first term in each asymptotic series. Consequently, the intrinsic error in the dominant series is relatively large and we lose track of the subdominant series very quickly after moving away from an anti-Stokes line. However, we could include more terms in each asymptotic series. This would enable us to reduce the intrinsic error in the dominant series and, thereby, expand the region of the complex plane in the vicinity of the anti-Stokes lines where we can see both the dominant and subdominant series. If we were to keep adding terms to our asymptotic series, so as to minimize the error in the dominant solution, we would eventually be forced to conclude that a jump in the coefficient of the subdominant series can only take place on those lines in the complex plane on which $\mathrm{Im}(z) = 0$: these are called *Stokes lines*. This result was first proved by Stokes in 1857.[15] On a Stokes line the magnitude of the dominant series achieves its *maximum* value with respect to that of the subdominant series. Once we know that a jump in the coefficient of the subdominant series can only take place at a Stokes line, we can retain the subdominant series in our asymptotic expansion in all regions of the complex plane. What we are basically saying is that, although, in practice, we cannot actually see the subdominant series very far away from an anti-Stokes line because we are only retaining the first term in each asymptotic series, we could, in principle, see the subdominant series at all values of $\arg(z)$ provided that we retained a sufficient number of terms in our asymptotic series.

Figure 17 shows the location in the complex plane of the Stokes and anti-Stokes lines for the asymptotic expansion of the hypergeometric function. Also shown is a branch cut, which is needed to make z single-valued. The branch cut is chosen such that $\arg(z) = 0$ on the positive real axis. Every time we cross an anti-Stokes line the dominant series becomes subdominant and *vice versa*. Every time we cross a Stokes line the coefficient in front of the dominant series stays the same, but that in front of the subdominant series jumps discontinuously [see Eqs. (4.252)]. Finally, the jumps in the coefficient of the subdominant series are such as to ensure that the asymptotic expansion is single-valued.

[15] G.G. Stokes, *Trans. Camb. Phil. Soc.* **10**, 106–128 (1857)

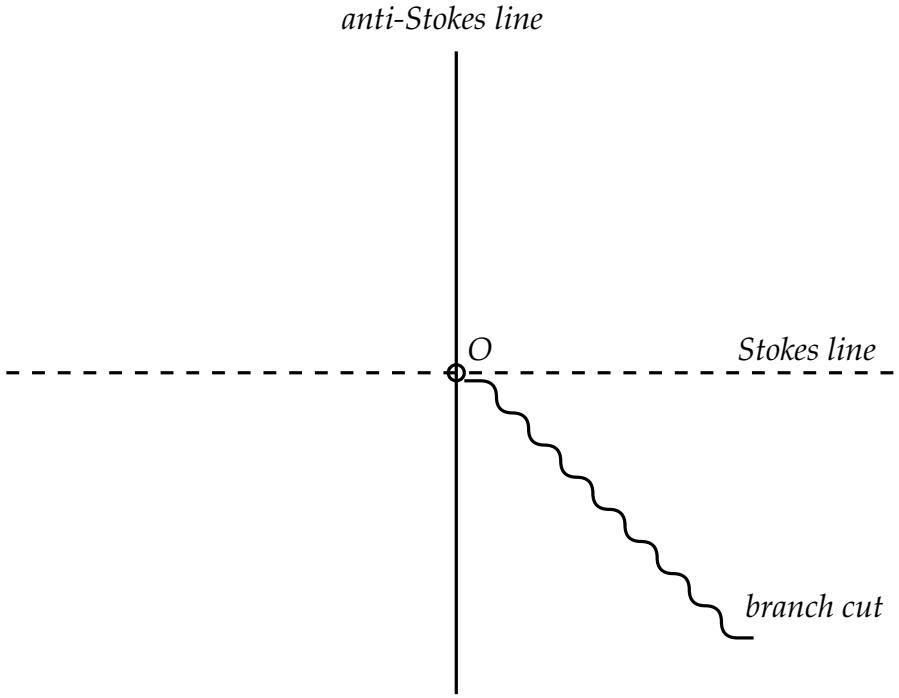

Figure 17: The location of the Stokes lines (dashed), the anti-Stokes lines (solid), and the branch cut (wavy) in the complex plane for the asymptotic expansion of the hypergeometric function

4.23 The W.K.B. solutions as asymptotic series

We have seen that the W.K.B. solution

$$E_y = n^{-1/2} \exp\left(\pm \mathrm{i}\, k \int^z n\, dz\right) \tag{4.253}$$

is an approximate solution of the differential equation

$$\frac{d^2 E_y}{dz^2} + k^2 n^2(z)\, E_y = 0 \tag{4.254}$$

in the limit where the typical wavelength, $2\pi/nk$, is much smaller than the typical variation length-scale of the refractive index. But, what sort of approximation is involved in writing this solution?

It is convenient to define the scaled variable

$$\hat{z} = \frac{z}{L}, \tag{4.255}$$

where L is the typical variation length-scale of the refractive index, $n(z)$. Equation (4.254) can then be written

$$w'' + h^2\, q\, w = 0, \tag{4.256}$$

where $w(\hat{z}, h) \equiv E_y(L\,\hat{z})$, $q(\hat{z}) \equiv n^2(L\,\hat{z})$, $' \equiv d/d\hat{z}$, and $h = kL$. Note that, in general, $q(\hat{z})$, $q'(\hat{z})$, $q''(\hat{z})$, *etc.* are $O(1)$ quantities. The non-dimensional constant h is of order the ratio of the variation length-scale of the refractive index to the wavelength. Let us seek the solutions to Eq. (4.256) in the limit $h \gg 1$.

We can write

$$w(\hat{z}, h) = \exp\left[\,\mathrm{i}\, h\, \phi(\hat{z}, h)\right]. \tag{4.257}$$

Equation (4.256) transforms to

$$\frac{\mathrm{i}}{h}\, \phi'' - (\phi')^2 + q = 0. \tag{4.258}$$

Expanding in powers of $1/h$, we obtain

$$\phi' = \pm q^{1/2} + \frac{\mathrm{i}}{4h} \frac{q'}{q} + O\left(\frac{1}{h^2}\right), \tag{4.259}$$

which yields

$$w(\hat{z},h) = q^{-1/4}\,\exp\left(\pm \mathrm{i}\,h\int^{\hat{z}} q\,d\hat{z}\right)\left[1+O\left(\frac{1}{h}\right)\right]. \tag{4.260}$$

Of course, we immediately recognize this expression as a W.K.B. solution.

Suppose that we keep expanding in powers of $1/h$ in Eq. (4.259). The appropriate generalization of Eq. (4.260) is a series solution of the form

$$w(\hat{z},h) = q^{-1/4}\,\exp\left(\pm \mathrm{i}\,h\int^{\hat{z}} q\,d\hat{z}\right)\left[1+\sum_{p=1}^{\infty}\frac{A_p(\hat{z})}{h^p}\right]. \tag{4.261}$$

This is, in fact, an *asymptotic series* in h. We can now appreciate that a W.K.B. solution is just a highly truncated asymptotic series in h, in which only the first term in the series is retained.

But, why is it so important that we recognize that W.K.B. solutions are highly truncated asymptotic series? The point is that the W.K.B. method was initially rather controversial after it was popularized in the 1920s. A lot of people thought that the method was completely wrong. Let us try to understand what the problem was. Suppose that we have never heard of an asymptotic series. Looking at Eq. (4.261), we would imagine that the expression in square brackets is a power law expansion in $1/h$. The W.K.B. approximation involves neglecting all terms in this expansion except the first. This sounds fine, as long as h is much greater than unity. But, surely, to be mathematically rigorous, we have to check that the sum of all of the terms in the expansion which we are neglecting is *small* compared to the first term? However, if we attempt this we discover, much to our consternation, that the expansion is *divergent*. In other words, the sum of all of the neglected terms is infinite! Thus, if we interpret Eq. (4.261) as a conventional power law expansion in $1/h$, the W.K.B. method is clearly nonsense: the W.K.B. solution is the first approximation to infinity. However, once we appreciate that Eq. (4.261) is actually an asymptotic series in h, the fact that the series diverges becomes irrelevant. If we retain the first n terms in the series, the series approximates the exact solution of Eq. (4.261) with an intrinsic (fractional) error which is of order $1/h^n$ (*i.e.*, the first neglected term in

the series). The error is minimized at a particular value of h. As the number of terms in the series is increased, the intrinsic error decreases, and the value of h at which the error is minimized increases. In particular, we can see that there is an intrinsic error associated with a W.K.B. solution which is of order $1/h$ times the solution.

It is amusing to note that if Eq. (4.261) were not a divergent series then it would be impossible to obtain total reflection of the W.K.B. solutions at the point $q = 0$. As we shall discover, the reflection is directly associated with the fact that the expansion (4.261) exhibits a Stokes phenomenon. It is, of course, impossible for a convergent power series expansion to exhibit a Stokes phenomenon.

4.24 Stokes constants

We have seen that the differential equation

$$w'' + h^2\, q(\hat{z})\, w = 0, \tag{4.262}$$

where $' \equiv d/d\hat{z}$, possesses approximate W.K.B. solutions of the form

$$(a, \hat{z}) = q^{-1/4} \exp\left(\mathrm{i}\, h \int_a^{\hat{z}} q^{1/2}\, d\hat{z} \right) \left[1 + O\left(\frac{1}{h} \right) \right], \tag{4.263a}$$

$$(\hat{z}, a) = q^{-1/4} \exp\left(-\mathrm{i}\, h \int_a^{\hat{z}} q^{1/2}\, d\hat{z} \right) \left[1 + O\left(\frac{1}{h} \right) \right]. \tag{4.263b}$$

Here, we have adopted an arbitrary phase reference level $\hat{z} = a$. The convenient notation $(a, \hat{z})$ is fairly self explanatory: a and $\hat{z}$ refer to the lower and upper bounds of integration, respectively, inside the exponential. It follows that the other W.K.B. solution can be written $(\hat{z}, a)$ (we can reverse the limits of integration inside the exponential to obtain *minus* the integral in $\hat{z}$ from $\hat{z} = a$ to $\hat{z} = \hat{z}$).

Up to now we have thought of $\hat{z}$ as a *real* variable representing scaled height in the ionosphere. Let us now generalize our analysis somewhat and think of $\hat{z}$ as a *complex* variable. There is nothing in our derivation of the W.K.B. solutions

which depends crucially on $\hat{z}$ being a real variable, so we expect these solutions to remain valid when $\hat{z}$ is reinterpreted as a complex variable. Incidentally, we must now interpret $q(\hat{z})$ as some well behaved function of the complex variable. An approximate general solution of the differential equation (4.262) in the complex $\hat{z}$ plane can be written as as a linear superposition of the two W.K.B. solutions (4.263).

The parameter h is assumed to be much larger than unity. It is clear from Eqs. (4.263) that in some regions of the complex plane one of the W.K.B. solutions is going to be exponentially larger than the other. In such regions, it is not mathematically consistent to retain the smaller W.K.B. solution in the expression for the general solution, since the contribution of the smaller W.K.B. solution is less than the intrinsic error associated with the larger solution. Adopting the terminology introduced in Section 4.22, the larger W.K.B. solution is said to be *dominant*, and the smaller solution is said to be *subdominant*. Let us denote the W.K.B. solution (4.263a) as $(a,\hat{z})_d$ in regions of the complex plane where it is dominant, and as $(a,\hat{z})_s$ in regions where it is subdominant. An analogous notation is adopted for the second W.K.B. solution (4.263b).

Suppose that $q(\hat{z})$ possesses a simple zero at the point $\hat{z} = \hat{z}_0$ (chosen to be the origin for the sake of convenience). It follows that in the immediate neighbourhood of the origin we can write

$$q = a_1\,\hat{z} + a_2\,\hat{z}^2 + \cdots, \tag{4.264}$$

where $a_1 \neq 0$. It is convenient to adopt the origin as the phase reference point (*i.e.*, $a = 0$), so the two W.K.B. solutions become $(0,\hat{z})$ and $(\hat{z},0)$. We can define *anti-Stokes lines* in the complex $\hat{z}$ plane (see Section 4.22). These are lines which satisfy

$$\mathrm{Re}\left[\mathrm{i}\int_0^{\hat{z}} q^{1/2}\,d\hat{z}\right] = 0. \tag{4.265}$$

As we cross an anti-Stokes line, a dominant W.K.B. solution becomes subdominant, and *vice versa*. Thus, $(0,\hat{z})_d \leftrightarrow (0,\hat{z})_s$ and $(\hat{z},0)_d \leftrightarrow (\hat{z},0)_s$. In the immediate vicinity of an anti-Stokes line the two W.K.B. solutions have about the same magnitude, so it is mathematically consistent to include the contributions from both solutions in the expression for the general solution. In such a

region, we can drop the subscripts d and s, since the W.K.B. solutions are neither dominant nor subdominant, and write the W.K.B. solutions simply as $(0, \hat{z})$ and $(\hat{z}, 0)$.

It is clear from Eqs. (4.263) that the W.K.B. solutions are not single-valued functions of $\hat{z}$, since they depend on $q^{1/2}(\hat{z})$, which is a double-valued function. Thus, if we wish to write an approximate *analytic* solution to the differential equation (4.262) we cannot express this solution as the *same* linear combination of W.K.B. solutions in all regions of the complex $\hat{z}$-plane. This implies that there must exist certain lines in the complex $\hat{z}$-plane across which the mix of W.K.B. solutions in our expression for the general solution changes discontinuously. These lines are called *Stokes lines* (see Section 4.22), and satisfy

$$\mathrm{Im}\left[\mathrm{i}\int_0^{\hat{z}} q^{1/2}\, d\hat{z}\right] = 0. \tag{4.266}$$

As we cross a Stokes line, the coefficient of the dominant W.K.B. solution in our expression for the general solution must remain unchanged, but the coefficient of the subdominant solution is allowed to change discontinuously. Incidentally, this is perfectly consistent with the fact that the general solution is analytic: the jump in our expression for the general solution due to the jump in the coefficient of the subdominant W.K.B. solution is *less* than the intrinsic error in this expression due to the intrinsic error in the dominant W.K.B. solution. Once we appreciate that the coefficient of the subdominant solution can only change at a Stokes line, we can retain both W.K.B. solutions in our expression for the general solution throughout the complex $\hat{z}$ plane. In practice, we can only see a subdominant solution in the immediate vicinity of an anti-Stokes line, but if we were to evaluate the W.K.B. solutions to higher accuracy [*i.e.* retain more terms in the asymptotic series in Eqs. (4.263)]we could, in principle, follow a subdominant solution all the way to a neighbouring Stokes line.

In the immediate vicinity of the origin

$$\int_0^{\hat{z}} q^{1/2}\, d\hat{z} \simeq \frac{2\sqrt{a_1}}{3}\, \hat{z}^{3/2}. \tag{4.267}$$

It follows from Eqs. (4.265) and (4.266) that *three* Stokes lines and *three* anti-Stokes lines radiate from a zero of $q(\hat{z})$. The general arrangement of Stokes and

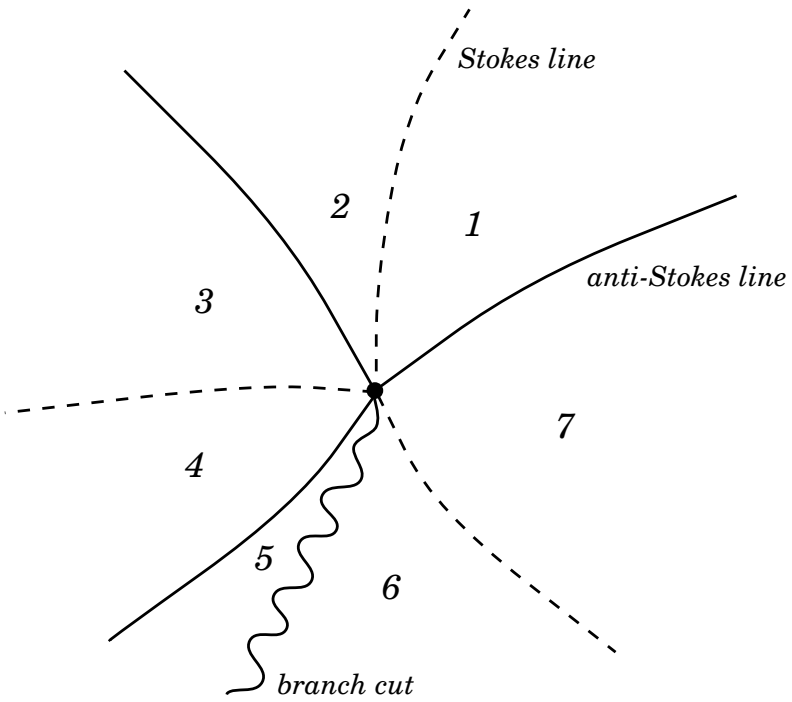

Figure 18: The arrangement of Stokes lines (dashed) and anti-Stokes lines (solid) around a simple zero of $q(\hat{z})$. Also shown is the branch cut (wavy line). All of the lines radiate from the point $q = 0$.

anti-Stokes lines in the vicinity of a $q = 0$ point is sketched in Fig. 18. Note that a branch cut must also radiate from the $q = 0$ point in order to uniquely specify the function $q^{1/2}(\hat{z})$. Thus, in general, *seven* lines radiate from a zero of $q(\hat{z})$, dividing the complex $\hat{z}$ plane into seven domains (numbered 1 through 7).

Let us write our general solution as

$$w(\hat{z}, h) = A\,(0, \hat{z}) + B\,(\hat{z}, 0) \tag{4.268}$$

on the anti-Stokes line between domains 1 and 7, where A and B are arbitrary constants. Suppose that the W.K.B. solution $(0, \hat{z})$ is dominant in domain 7. Thus, in domain 7 the general solution takes the form

$$w(7) = A\,(0, \hat{z})_d + B\,(\hat{z}, 0)_s. \tag{4.269}$$

Let us move into domain 1. In doing so, we cross an anti-Stokes line, so the dominant solution becomes subdominant, and *vice versa*. Thus, in domain 1 the

general solution takes the form

$$w(1) = A\,(0,\hat{z})_s + B\,(\hat{z},0)_d. \tag{4.270}$$

Let us now move into domain 2. In doing so, we cross a Stokes line, so the coefficient of the dominant solution, B, must remain constant, but the coefficient of the subdominant solution, A, is allowed to change. Suppose that the coefficient of the subdominant solution jumps by t times the coefficient of the dominant solution, where t is an undetermined constant. It follows that in domain 2 the general solution takes the form

$$w(2) = (A + t\,B)\,(0,\hat{z})_s + B\,(\hat{z},0)_d. \tag{4.271}$$

Let us now move into domain 3. In doing so, we cross an anti-Stokes line, so the the dominant solution becomes subdominant, and *vice versa*. Thus, in domain 3 the general solution takes the form

$$w(3) = (A + t\,B)\,(0,\hat{z})_d + B\,(\hat{z},0)_s. \tag{4.272}$$

Let us now move into domain 4. In doing so, we cross a Stokes line, so the coefficient of the dominant solution must remain constant, but the coefficient of the subdominant solution is allowed to change. Suppose that the coefficient of the subdominant solution jumps by u times the coefficient of the dominant solution, where u is an undetermined constant. It follows that in domain 4 the general solution takes the form

$$w(4) = (A + t\,B)\,(0,\hat{z})_d + (B + u\,[A + t\,B])\,(\hat{z},0)_s. \tag{4.273}$$

Let us now move into domain 5. In doing so, we cross an anti-Stokes line, so the the dominant solution becomes subdominant, and *vice versa*. Thus, in domain 5 the general solution takes the form

$$w(5) = (A + t\,B)\,(0,\hat{z})_s + (B + u\,[A + t\,B])\,(\hat{z},0)_d. \tag{4.274}$$

Let us now move into domain 6. In doing so, we cross the branch cut in an anti-clockwise direction. Thus, the argument of $\hat{z}$ decreases by 2π. It follows from Eq. (4.264) that $q^{1/2} \rightarrow -q^{1/2}$ and $q^{1/4} \rightarrow -\mathrm{i}\,q^{1/4}$. The following rules for tracing

the W.K.B. solutions across the branch cut (in an anti-clockwise direction) ensure that the general solution is continuous across the cut [see Eqs. (4.261)]:

$$(0,\hat{z}) \quad \rightarrow \quad -\mathrm{i}\,(\hat{z},0), \tag{4.275a}$$

$$(\hat{z},0) \quad \rightarrow \quad -\mathrm{i}\,(0,\hat{z}). \tag{4.275b}$$

Note that the properties of dominancy and subdominancy are preserved when the branch cut is crossed. It follows that in domain 6 the general solution takes the form

$$w(6) = -\mathrm{i}\,(A + t\,B)\,(\hat{z},0)_s - \mathrm{i}\,(B + u\,[A + t\,B])\,(0,\hat{z})_d. \tag{4.276}$$

Let us, finally, move into domain 7. In doing so, we cross a Stokes line, so the coefficient of the dominant solution must remain constant, but the coefficient of the subdominant solution is allowed to change. Suppose that the coefficient of the subdominant solution jumps by v times the coefficient of the dominant solution, where v is an undetermined constant. It follows that in domain 7 the general solution takes the form

$$w(7) = -\mathrm{i}\,(A+t\,B+v\,\{B+u\,[A+t\,B]\})\,(\hat{z},0)_s - \mathrm{i}\,(B+u\,[A+t\,B])\,(0,\hat{z})_d. \tag{4.277}$$

Now, we expect our general solution to be an *analytic* function, so it follows that the solutions (4.269) and (4.277) must be identical. Thus, we can compare the coefficients of the two W.K.B. solutions, $(\hat{z},0)_s$ and $(0,\hat{z})_d$. Since A and B are arbitrary, we can also compare the coefficients of A and B. Thus, comparing the coefficients of $A\,(0,\hat{z})_d$, we find

$$1 = -\mathrm{i}\,u. \tag{4.278}$$

Comparing the coefficients of $B\,(0,\hat{z})_d$ yields

$$0 = 1 + ut. \tag{4.279}$$

Comparing the coefficients of $A\,(\hat{z},0)_s$ gives

$$0 = 1 + vu. \tag{4.280}$$

Finally, comparing the coefficients of $B\,(\hat{z},0)_s$ yields

$$1 = -\mathrm{i}\,(t + v + vut). \tag{4.281}$$

Equations (4.278)–(4.281) imply that

$$t = u = v = \mathrm{i}. \tag{4.282}$$

In other words, if we adopt the simple rule that every time we cross a Stokes line in an anti-clockwise direction the coefficient of the subdominant solution jumps by i times the coefficient of the dominant solution, then this *ensures* that our expression for the general solution (4.268) behaves as an analytic function. Here, the constant i is usually called a *Stokes constant*. Note that if we cross a Stokes line in a clockwise direction then the coefficient of the subdominant solution has to jump by $-\mathrm{i}$ times the coefficient of the dominant solution in order to ensure that our general solution behaves as an analytic function.

4.25 The reflection coefficient

Let us write $\hat{z} = x + \mathrm{i}\,y$, where x and y are real variables. Consider the solution of the differential equation

$$w'' + h^2\,q(x)\,w = 0, \tag{4.283}$$

where $q(x)$ is a real function, h is a large number, $q > 0$ for $x < 0$, and $q < 0$ for $x > 0$. It is clear that $\hat{z} = 0$ represents a simple zero of $q(\hat{z})$. Here, we assume, as seems eminently reasonable, that we can find a well behaved function of the complex variable $q(\hat{z})$ such that $q(\hat{z}) = q(x)$ along the real axis. The arrangement of Stokes and anti-Stokes lines in the immediate vicinity of the point $\hat{z} = 0$ is sketched in Fig. 19. The argument of $q(\hat{z})$ on the positive x-axis is chosen to be $-\pi$. Thus, the argument of $q(\hat{z})$ on the negative x-axis is 0.

On OB, the two W.K.B. solutions (4.261) can be written

$$(0,x) = q^{-1/4}(x)\,\exp\left(\mathrm{i}\,h\int_0^x q^{1/2}(x)\,dx\right), \tag{4.284a}$$

$$(x,0) = q^{-1/4}(x)\,\exp\left(-\mathrm{i}\,h\int_0^x q^{1/2}(x)\,dx\right). \tag{4.284b}$$

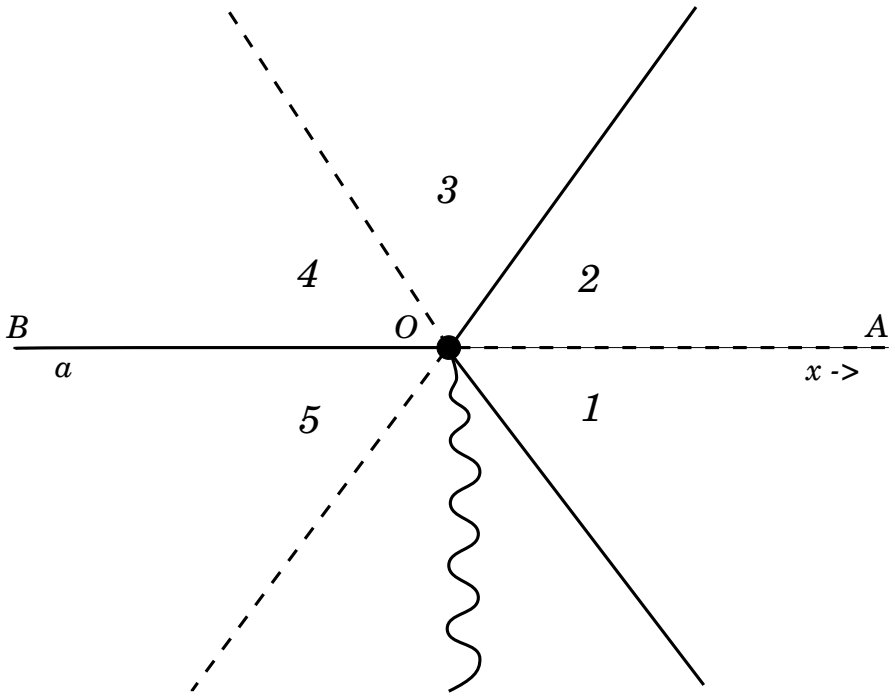

Figure 19: The arrangement of Stokes lines (dashed) and anti-Stokes lines (solid) in the complex $\hat{z}$ plane. Also shown is the branch cut (wavy line).

Here, we can interpret $(0,x)$ as a wave propagating to the right along the x-axis, and $(x,0)$ as a wave propagating to the left. On OA, the W.K.B. solutions take the form

$$(0,x)_d = \mathrm{e}^{\,\mathrm{i}\,\pi/4}\,|q(x)|^{-1/4}\,\exp\left(+h\int_0^x |q(x)|^{1/2}\,dx\right), \tag{4.285a}$$

$$(x,0)_s = \mathrm{e}^{\,\mathrm{i}\,\pi/4}\,|q(x)|^{-1/4}\,\exp\left(-h\int_0^x |q(x)|^{1/2}\,dx\right). \tag{4.285b}$$

Clearly, $(x,0)_s$ represents an evanescent wave which decays to the right along the x-axis, whereas $(0,x)_d$ represents an evanescent wave which decays to the left. If we adopt the boundary condition that there is no incident wave from the region $x\rightarrow+\infty$, the most general asymptotic solution to Eq. (4.283) on OA is written

$$w(x,h) = A\,(x,0)_s, \tag{4.286}$$

where A is an arbitrary constant.

Let us assume that we can find an analytic solution $w(\hat{z},h)$ to the differential equation

$$w'' + h^2\,q(\hat{z})\,w = 0, \tag{4.287}$$

which satisfies $w(\hat{z},h) = w(x,h)$ along the real axis, where $w(x,h)$ is the physical solution. From a mathematical point of view, this seems eminently reasonable. In the domains 1 and 2 the solution (4.286) becomes

$$w(1) = A\,(\hat{z},0)_s, \tag{4.288}$$

and

$$w(2) = A\,(\hat{z},0)_s. \tag{4.289}$$

Note that the solution is continuous across the Stokes line OA, since the coefficient of the dominant solution $(0,\hat{z})$ is zero: thus, the jump in the coefficient of the subdominant solution is zero times the Stokes constant, i; *i.e.*, it is zero. Let us move into domain 3. In doing so, we cross an anti-Stokes line, so the solution becomes

$$w(3) = A\,(\hat{z},0)_d. \tag{4.290}$$

Let us now move into domain 4. In doing so, we cross a Stokes line. Applying the general rule derived in the preceding section, the solution becomes

$$w(4) = A\,(\hat{z}, 0)_d + \mathrm{i}\,A\,(0, \hat{z})_s. \tag{4.291}$$

Finally, on OB the solution becomes

$$w(x, h) = A\,(x, 0) + \mathrm{i}\,A\,(0, x). \tag{4.292}$$

Suppose that there is a point a on the negative x-axis where $q(x) = 1$. It follows from Eqs. (4.286) and (4.292) that we can write the asymptotic solution to Eq. (4.283) as

$$\begin{aligned} w(x,h) &= q^{-1/4}\,\exp\left(\mathrm{i}\,h\int_a^x q^{1/2}\,dx\right) \\ &\quad -\mathrm{i}\,\exp\left(2\,\mathrm{i}\,h\int_a^0 q^{1/2}\,dx\right)\,q^{-1/4}\,\exp\left(-\mathrm{i}\,h\int_a^x q^{1/2}\,dx\right), \end{aligned} \tag{4.293}$$

in the region $x < 0$, and

$$w(x,h) = \exp\left(\mathrm{i}\,h\int_a^0 q^{1/2}\,dx\right)\,\mathrm{e}^{-\mathrm{i}\,\pi/4}\,|q|^{-1/4}\,\exp\left(-h\int_0^x |q|^{1/2}\,dx\right) \tag{4.294}$$

in the region $x > 0$. Here, we have chosen

$$A = -\mathrm{i}\,\exp\left(\mathrm{i}\,h\int_a^0 q^{1/2}\,dx\right). \tag{4.295}$$

If we interpret x as a normalized altitude in the ionosphere, $q(x)$ as the square of the refractive index in the ionosphere, the point a as ground level, and w as the electric field strength of a radio wave propagating vertically upwards into the ionosphere, then Eq. (4.293) tells us that a unit amplitude wave fired vertically upwards from ground level into the ionosphere is *reflected* at the level where the refractive index is zero. The first term in Eq. (4.293) is the incident wave and the second term is the reflected wave. The reflection coefficient (*i.e.*, the ratio of the reflected to the incident wave at ground level) is given by

$$R = -\mathrm{i}\,\exp\left(2\,\mathrm{i}\,h\int_a^0 q^{1/2}\,dx\right). \tag{4.296}$$

Note that $|R| = 1$, so the amplitude of the reflected wave equals that of the incident wave. In other words, there is no absorption of the wave at the level of reflection. The phase shift of the reflected wave at ground level, with respect to that of the incident wave, is that associated with the wave propagating from ground level to the reflection level and back to ground level again, *plus* a $-\pi/2$ phase shift at reflection. According to Eq. (4.294), the wave attenuates fairly rapidly (in the space of a few wavelengths) above the reflection level. Of course, Eq. (4.296) is completely equivalent to Eq. (4.186).

Note that the reflection of the incident wave at the point where the refractive index is zero is directly associated with the Stokes phenomenon. Without the jump in the coefficient of the subdominant solution, as we go from domain 3 to domain 4, there is no reflected wave on the OB axis. Note, also, that the W.K.B. solutions (4.293) and (4.294) break down in the immediate vicinity of $q = 0$ (*i.e.*, the reflection point). Thus, it is possible to demonstrate that the incident wave is totally reflected at the point $q = 0$, with a $-\pi/2$ phase shift, without having to solve for the wave structure in the immediate vicinity of the reflection point. This demonstrates that the reflection of the incident wave at $q = 0$ is an intrinsic property of the W.K.B. solutions, and does not depend on the detailed behaviour of the wave in the region where the W.K.B. solutions break down.

4.26 The Jeffries connection formula

In the preceding section there is a tacit assumption that the square of the refractive index, $q(x) \equiv n^2(x)$, is a *real* function. As is apparent from Eq. (4.162), this is only the case in the ionosphere as long as electron collisions are negligible. Let us generalize our analysis to take electron collisions into account. In fact, the main effect of electron collisions is to move the zero of $q(\hat{z})$ a short distance off the real axis (the distance is relatively short provided that we adopt the physical ordering $\nu \ll \omega$). The arrangement of Stokes and anti-Stokes lines around the new zero point, located at $\hat{z} = \hat{z}_0$, is sketched in Fig. 20. Note that electron collisions only significantly modify the form of $q(\hat{z})$ in the immediate vicinity of the zero point. Thus, sufficiently far away from $\hat{z} = \hat{z}_0$ in the complex $\hat{z}$-plane, the W.K.B. solutions and the locations of the Stokes and anti-Stokes lines are exactly the same as in the preceding section.

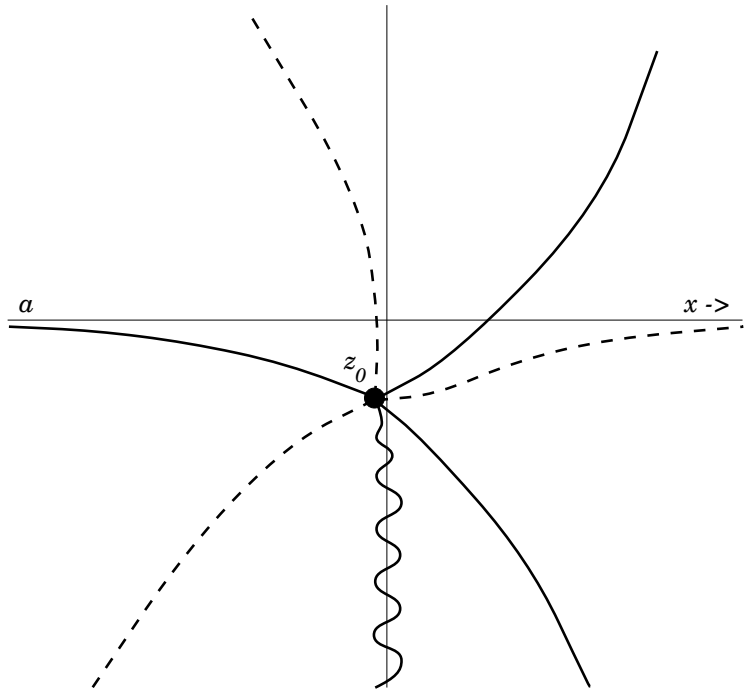

Figure 20: The arrangement of Stokes lines (dashed) and anti-Stokes lines (solid) in the complex $\hat{z}$ plane. Also shown is the branch cut (wavy line).

The W.K.B. solutions (4.284) and (4.285) are valid all the way along the real axis, except for a small region close to the origin where electron collisions significantly modify the form of $q(\hat{z})$. Thus, we can still adopt the physically reasonable decaying solution (4.286) on the positive real axis. Let us trace this solution in the complex $\hat{z}$-plane until we reach the negative real axis. We can achieve this by moving in a semi-circle in the upper half-plane. Since we never move out of the region in which the W.K.B. solutions (4.284) and (4.285) are valid, we conclude, by analogy with the preceding section, that the solution on the negative real axis is given by Eq. (4.292). Of course, in all of the W.K.B. solutions the point $\hat{z} = 0$ must be replaced by the new zero point $\hat{z} = \hat{z}_0$. The new formula for the reflection coefficient, which is just a straightforward generalization of Eq. (4.296), is

$$R = -\mathrm{i}\,\exp\left(2\,\mathrm{i}\,h\int_a^{\hat{z}_0} q^{1/2}\,d\hat{z}\right). \tag{4.297}$$

This is called the *Jeffries connection formula*, after H. Jeffries, who discovered it in 1923. The general expression for the reflection coefficient is incredibly simple. We just integrate the W.K.B. solution in the complex $\hat{z}$-plane from the phase reference level $\hat{z} = a$ to the zero point, square the result, and multiply by $-\mathrm{i}$. Note that the path of integration between $\hat{z} = a$ and $\hat{z} = \hat{z}_0$ does not matter, because of Cauchy's theorem. Note, also, that since $q^{1/2}$ is, in general, *complex* along the path of integration, we no longer have $|R| = 1$. In fact, it is easily demonstrated that $|R| \leq 1$. Thus, when electron collisions are included in the analysis we no longer obtain perfect reflection of radio waves from the ionosphere. Instead, some (small) fraction of the radio energy is *absorbed* at each reflection event. This energy is ultimately transfered to the particles in the ionosphere with which the electrons collide.