13

Ray theory results for anisotropic plasmas

13.1 Introduction

We now have to consider how results such as those of ch. 12 for an isotropic plasma must be modified when the earth's magnetic field is allowed for. The dispersion relation (4.47) or (4.51) is now much more complicated so that, even for simple electron height distribution functions $N(z)$, it is not possible to derive algebraic expressions for such quantities as the equivalent height of reflection $h'(f)$ or the horizontal range $D(\theta)$. These and other quantities must now be evaluated numerically. The ray direction is not now in general the same as the wave normal. Breit and Tuve's theorem and Martyn's theorems do not hold. These properties therefore cannot be used in ray tracing. Some general results for ray tracing in a stratified anisotropic plasma have already been given in ch. 10.

The first part of this chapter §§ 13.2–13.6 is concerned with vertically incident pulses of radio waves on a stratified ionosphere, since this is the basis of the ionosonde technique, §1.7, which is widely used for ionospheric sounding. The transmitted pulses have a radio frequency f which is here called the 'probing frequency'. It is also sometimes called the 'carrier frequency'. The study of this subject involves numerical methods that are important because of their use for analysing ionospheric data and for the prediction of maximum usable frequencies. It has already been explained, § 10.2, that the incident pulse splits into two separate pulses, ordinary and extraordinary, that travel independently. This is the phenomenon of magnetoionic splitting. Ionosonde equipments can be carried in satellites above the maximum of $N(z)$ in the F2-region of the ionosphere, so that they send pulses downwards. This is known as topside sounding and has been extensively used since 1962. The main results are briefly described in § 13.5.

Four other topics come naturally into this chapter, namely Faraday rotation, § 13.7, whistlers and ion cyclotron whistlers §§ 13.8, 13.9, absorption § 13.10, and wave interaction §§ 13.11–13.13.

13.2. Reflection levels and penetration frequencies

A radio wave vertically incident on the ionosphere from below travels upwards until it is reflected at a level where the refractive index $n = 0$. For the ordinary wave, when collisions are neglected, this occurs where $X = 1$, that is $f = f_N$, which is the same condition as for an isotropic plasma. For the extraordinary wave the refractive index is zero where $X = 1 - Y$, which requires that

$$f_N^2 = f^2 - f f_H. \tag{13.1}$$

This is a quadratic equation with the solution

$$f = f_E^{(-)} = \tfrac{1}{2}\{(f_H^2 + 4f_N^2)^{\frac{1}{2}} + f_H\}. \tag{13.2}$$

The other solution is negative and therefore of no practical interest. The refractive index is also zero for the extraordinary wave where $X = 1 + Y$ and this similarly gives

$$f = f_E^{(+)} = \tfrac{1}{2}\{(f_H^2 + 4f_N^2)^{\frac{1}{2}} - f_H\}. \tag{13.3}$$

Thus an extraordinary wave travels upwards until it is reflected at the level where either (13.2) or (13.3) is first satisfied. In the free space below the ionosphere (13.3) is zero and (13.2) is f_H. They both increase when the ionosphere is entered and f_N increases, and (13.2) is always the greater.

If $f > f_H$ then as f_N increases it is (13.2) that first reaches the value f of the probing frequency. Thus it might be expected that reflection at a level given by (13.3) can only be observed if $f < f_H$. This is true according to simple ray theory when the transmitter is at the ground or anywhere in free space. But in topside sounding the transmitter is in the ionosphere and may be at a level where X exceeds the value marked RES in fig. 4.3, so that n^2 is positive for the Z-mode. Then reflection at the level given by (13.3) can be observed even when $f > f_H$; see § 13.5.

When observing from the ground, especially at high latitudes, a reflection from the level given by (13.3) is sometimes seen when $f > f_H$, and the resulting branch of the $h'(f)$ curve is known as the 'Z-trace'. The phenomenon is described later, § 16.8.

The penetration frequency of an ionospheric layer is the frequency that makes $n = 0$ at the maximum of the layer. This can only happen if collisions are neglected, but penetration frequencies are customarily defined by neglecting collisions. For the ordinary wave the penetration frequency is the plasma frequency f_N at the maximum of the layer. It is denoted by f_p^O which is the same as the f_p used in ch. 12 for an isotropic ionosphere. When the term 'penetration frequency' is used alone, it normally refers to the ordinary wave.

For the extraordinary wave there are two penetration frequencies. One is where $X = 1 - Y$ at the maximum of the layer. It is given, from (13.2), by

$$f_p^{(X-)} = \tfrac{1}{2}[\{f_H^2 + 4(f_p^{(O)})^2\}^{\frac{1}{2}} + f_H]. \tag{13.4}$$

This is the one that is normally observed when $f > f_H$. If $f_p^{(O)}$ is large compared

with f_H, (13.4) gives

$$f_p^{(X-)} - f_p^{(O)} \approx \tfrac{1}{2} f_H. \tag{13.5}$$

This result was used by Appleton and Builder (1933) to measure f_H in the ionosphere, and to show that it is electrons that are responsible for the reflection of radio waves.

The other extraordinary wave penetration frequency is, from (13.3),

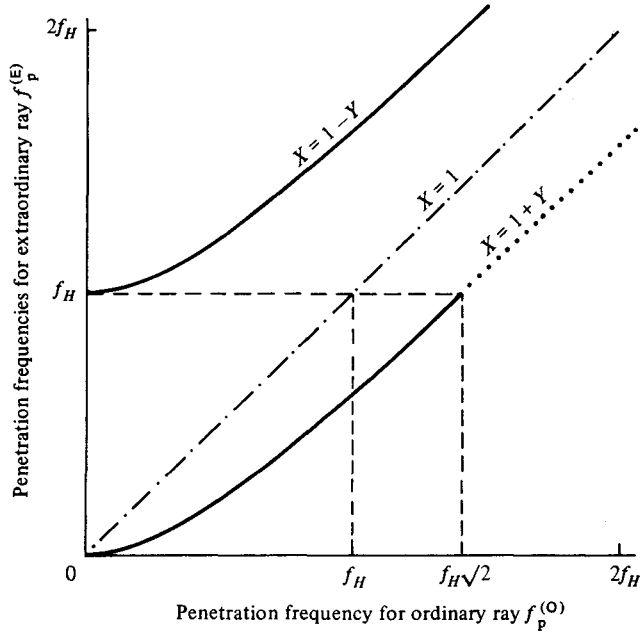
$$f_p^{(X+)} = \tfrac{1}{2}\{(f_H^2 + 4f_p^{(O)2})^{\frac{1}{2}} - f_H\}. \tag{13.6}$$

The relation between the two penetration frequencies (13.4), (13.6) and $f_p^{(O)}$ is shown in fig. 13.1.

In (13.4) and (13.6) the superscript X is used for 'extraordinary' because the label E is needed to designate the E-layer. In the literature that deals with the reduction of ionograms, the penetration frequencies $f_p^{(O)}$, $f_p^{(X-)}$, $f_p^{(X+)}$ for the E-layer are usually written F_0E, F_XE, F_ZE. For the F-layer the E is replaced by F. The most important penetration frequencies are $f_p^{(O)}$, $f_p^{(X-)}$ for the F2-layer and they are commonly written F_0F2, F_XF2.

When the probing frequency f increases and approaches a penetration frequency, the equivalent height $h'(f)$ becomes large and tends to infinity. It is in this way that penetration frequencies are identified on ionogram records. It was pointed out

Fig. 13.1. Shows how the two penetration frequencies for the extraordinary ray, shown as continuous and dotted curves, are related to that for the ordinary ray, shown as a chain line.



above that, except for the Z-trace, the reflection condition (13.3) is only observed when probing from the ground, if $f < f_H$. The same applies to the penetration frequency (13.6). If it is to be observable from the ground it must be less than f_H, and this requires that

$$f_p^{(O)} < f_H\sqrt{2}. \tag{13.7}$$

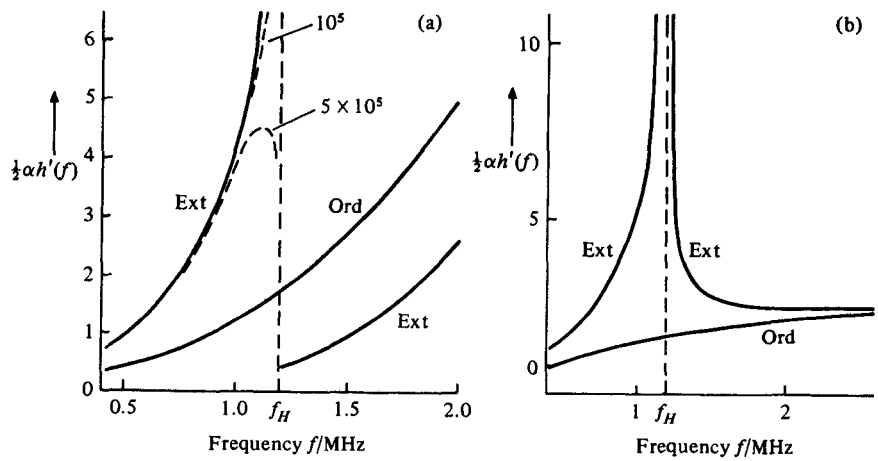
For a layer that satisfies this condition, the $h'(f)$ curve for the extraordinary wave shows two infinities, one given by (13.4) where $f > f_H$, and the other by (13.6) where $f < f_H$. Fig. 13.3 shows an example of this. If (13.7) is not satisfied, the $h'(f)$ curve for the extraordinary wave shows only one infinity that is given by (13.4) where $f > f_H$. In either case $h'(f)$ attains large values when $f \to f_H$ but here it tends to a bounded limit; see fig. 13.2. This is not an infinity and is not associated with penetration; see § 13.4.

13.3. The calculation of equivalent height, $h'(f)$

In the first part of this section it is assumed that electron collisions may be neglected. The effect of collisions is discussed at the end of the section.

The equivalent height of reflection for a vertically incident pulse of radio waves is $h'(f) = \int_0^{z_0} n'\,dz$ from (12.2) and this still holds when the earth's magnetic field is

Fig. 13.2. The $h'(f)$ curves for frequencies near the electron gyro-frequency, 1.2 MHz in these examples. The angle between the earth's magnetic field and the vertical is 23.27°. In (a) $f_N^2 = \alpha(z - h_0)$ for $z > h_0$, and $h'(f)$ is measured from the level $z = h_0$. In (b) f_N^2 is a proportional to $e^{\alpha z}$ and $h'(f)$ is measured from the level where $f_N = 0.5$ MHz. For the continuous curves the collision frequency ν is zero. For the broken curves in (a) the numbers are the values of ν in s^{-1}. These collision frequencies make no appreciable difference except for the extraordinary wave at frequencies just less than the gyro-frequency.



allowed for. It will now be written

$$h'(f) = \int_0^{z_0} n'(f, f_N)\,\mathrm{d}z. \tag{13.8}$$

Here n' is the group refractive index. It is a function of the probing frequency f and the plasma frequency f_N, and it is through f_N that n' depends on z. It also depends on f_H and on the angle Θ between $\boldsymbol{Y}$ and the vertical, but these can be assumed to be constant for any one observing location, and therefore they are not written in (13.8). The reflection level $z = z_0$ is where the refractive index is zero, and here n' tends to infinity; see figs. 5.16, 5.17. For the ordinary wave near the reflection level, n' is given by (5.78) and for the extraordinary wave by (5.79). These can be used to show that the integral (13.8) is bounded even though n' is infinite at the upper limit. The method is the same as in § 12.2.

To evaluate (13.8) it is necessary to use numerical methods. The contribution from the part of the path in free space is simply the length of that part. To find the contribution from the part in the ionosphere it is sometimes convenient to use f_N as the variable of integration, instead of z, thus

$$h'(f) = \int_0^{f_R} n'(f, f_N)\frac{\mathrm{d}z}{\mathrm{d}f_N}\,\mathrm{d}f_N. \tag{13.9}$$

Here f_R is the value of f_N where reflection occurs. For the ordinary wave $f_R = f$, and for the extraordinary wave, the two reflection conditions give

$$\text{for } X = 1 \mp Y: \quad f_R = (f^2 \mp f f_H)^{\frac{1}{2}}. \tag{13.10}$$

The transformation in (13.9) can only be used when $f_N(z)$ is a monotonic function in the range of integration. For numerical integration it is best to use an integrand that does not go to infinity at one limit. This can be achieved by a further transformation. For example, let

$$f_N = f_R \sin\phi: \tag{13.11}$$

Then (13.9) becomes

$$h'(f) = f_R \int_0^{\frac{1}{2}\pi} n'(f, f_R \sin\phi)\frac{\mathrm{d}z}{\mathrm{d}f_N}\cos\phi\,\mathrm{d}\phi. \tag{13.12}$$

It can be shown from (5.78) or (5.79) that $n'\cos\phi$ is now bounded at the upper limit. This can be achieved by other transformations, and the best choice depends on the function $f_N(z)$. Various other forms were used by Millington (1938a) and by Shinn and Whale (1952). An integral such as (13.12) is now suitable for evaluation by one of the standard methods, for example Gaussian quadrature (Abramowitz and Stegun, 1965, table 25.4).

The effect of electron collisions on the equivalent height for an isotropic ionosphere was discussed in § 12.3, and the same general conclusions hold when the earth's magnetic field is allowed for. The process of analytic continuation in the

complex z plane is used. The group refractive index n', the height of reflection z_0 and the equivalent height h' are all complex. It was shown in § 11.15 that it is $\mathrm{Re}\,(h')$ that gives the time of travel of the radio pulse. The complex $h'(f)$ is still given by (13.8) where the integral is now a contour integral in the complex z plane.

The complex reflection height z_0 is where $n = 0$, that is where

$$X(z) = 1 + \eta Y - \mathrm{i}Z(z). \tag{13.13}$$

Here, for the ordinary wave $\eta = 0$, and for the extraordinary wave $\eta = -1$ if $f > f_H$ and $+1$ if $f < f_H$. This equation usually has more than one solution z. The required solution z_0 is nearly always the nearest to the real z_0 that would be used when $Z = 0$. The contribution to the integral (13.8) from within the ionosphere can be evaluated by using a transformation similar to (13.11). Let $Z(z_0) = Z_0$ and let

$$f_N = f(1 + \eta Y - \mathrm{i}Z_0)^{\frac{1}{2}} \sin\phi. \tag{13.14}$$

Then the contribution to (13.8) from within the ionosphere is

$$h'(f) = f(1 + \eta Y - \mathrm{i}Z_0)^{\frac{1}{2}} \int_0^{\frac{1}{2}\pi} n'(f, f_N, \nu) \frac{\mathrm{d}z}{\mathrm{d}f_N} \cos\phi \,\mathrm{d}\phi. \tag{13.15}$$

Here the complex group refractive index depends on the collision frequency ν, which may be a function of z. To use this formula it must be possible to express z and $\nu(z)$ as analytic functions of f_N and thence of ϕ. The variable ϕ may be chosen to be real over the range of integration, so that f_N and $\mathrm{d}z/\mathrm{d}f_N$ and ν are complex. At the top limit, $\phi = \frac{1}{2}\pi$, n' is infinite but $\cos\phi$ is zero. It can again be shown, as for (13.12), that the product $n' \cos\phi$ is bounded. To allow for collisions, the proof needs an extension of the argument of § 5.9 that led to (5.78). It is left as an exercise for the reader. The formula (13.14) and extensions of it have been used by Cooper (1961) and by Piggott and Thrane (1966) to study the effect of collisions on the $h'(f)$ curves for a parabolic model and other models of the ionosphere.

It is sometimes necessary to compute $h'(f)$ when $N(z)$ is given not as an algebraic function but as a table of numerical values, usually at equal intervals of z. Then $N(z)$ for any z is found by interpolation in the table. If collisions are neglected, the path for the integral (13.8) is the real z axis, and $h'(f)$ can be found by numerical integration. It is best not to change to f_N as the variable, as was done in (13.9). To transform the integral so that the integrand is bounded at the upper limit, the change of variable

$$z = z_0 \sin^2\phi \tag{13.16}$$

can be used.

When collisions are allowed for it is first necessary to find the complex z_0 where $n = 0$, and this requires analytic continuation of the tabulated $N(z)$ to complex values of z. To do this we first find the real z where $X = 1$ (ordinary wave) or $X = 1 \mp Y$ (extraordinary wave). Then it is assumed that $N(z)$ and therefore $X(z)$ is a linear function in this neighbourhood, whence the complex z_0 that makes $X = 1 - \mathrm{i}Z$, or

$1 \mp Y - iZ$, respectively, can be found. In a few cases studied by the author it was found better to use a quadratic function instead of a linear function. Then the required value of the equivalent height is given by

$$\mathrm{Re}\{h'(f)\} = \int_0^{\mathrm{Re}(z_0)} \mathrm{Re}(n')\,\mathrm{d}z - \int_0^{\mathrm{Im}(z_0)} \mathrm{Im}(n')\,\mathrm{d}s \tag{13.17}$$

where in the second integral $z = \mathrm{Re}(z_0) + is$, and $N(z)$ is calculated by the same linear or quadratic function that was used to find z_0. In the second integral it is also useful to use

$$s = \mathrm{Im}(z_0)\sin^2\phi, \tag{13.18}$$

so that the integrand is bounded at the upper limit; compare (13.16).

Another way of calculating $h'(f)$ is to compute two values h_1, h_2 of $h(f)$ at neighbouring frequencies f_1, f_2 and then to use $h' = \partial(fh)/\partial f \approx (f_1 h_1 - f_2 h_2)/(f_1 - f_2)$. This method was used by Altman (1965) who included a study of the Z-trace (§16.8).

13.4. Ionograms

'Ionogram' is the name given to the curves of $h'(f)$ obtained with an ionosonde equipment, §1.7. In the following description subscripts will be used to indicate the values $h'_O(f)$, $h'_X(f)$ for the ordinary and extraordinary waves respectively.

Curves of $h'_O(f)$ are of the same general form as those for an isotropic ionosphere. They show similar behaviour near the penetration frequencies of ionospheric layers, and near the plasma frequency in a ledge in the electron height distribution $N(z)$; see §§ 12.2–12.5 and figs. 12.2–12.4. For frequencies f near to the electron gyro-frequency f_H, the refractive index n and group refractive index n' are bounded and continuous functions of f. The $h'_O(f)$ curves for the ordinary wave show no special features when f is near to f_H.

When $f > f_H$ the extraordinary wave is reflected where $X = 1 - Y$, that is at a lower level than the ordinary wave. Thus $h'_X(f)$ is in general smaller than $h'_O(f)$ but there are exceptions. For example, if f very slightly exceeds the extraordinary penetration frequency of the E-layer, the extraordinary wave is delayed near the maximum of the E-layer and has the larger equivalent height. If the extraordinary wave is reflected in a ledge of $N(z)$, the pulse spends a long time near the reflection level; see § 12.5. Then $h'_X(f)$ can be greater than $h'_O(f)$. Examples of this can be seen in the typical $h'(f)$ diagram sketched in fig. 13.4.

Suppose now that the extraordinary wave is reflected from an ionospheric layer with a non-zero value of $G = \mathrm{d}N(z)/\mathrm{d}z$ at its lower edge. Let the frequency f be decreased so that it approaches f_H. The reflection level where $X = 1 - Y$ gets lower as f decreases and in the limit $f \to f_H$ it is where $f_N \to 0$, that is at the base of the layer. But the contribution to $h'_X(f)$ from within the layer does not necessarily decrease. It

tends to a non-zero limit and it can be shown that this is proportional to $1/G$. (See problem 13.1.) For actual ionospheric layers $N(z)$ tails off gradually towards zero as z decreases, so that G is effectively zero, or very small. Then $h'_X(f)$ tends to large values as $f \to f_H(+)$. These effects are illustrated by the two examples in fig. 13.2.

Fig. 13.3. Calculated $h'(f)$ curves for a model ionosphere with two parabolic layers. For the lower layer, $f_p^{(O)} = 0.7$ MHz, half thickness $a = 30$ km, height of maximum = 110 km. For the upper layer $f_p^{(O)} = 2.4$ MHz, half thickness $a = 50$ km, height of maximum = 260 km.

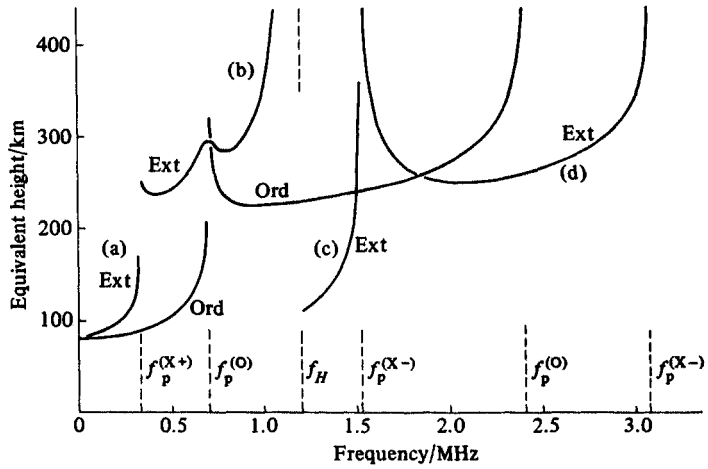
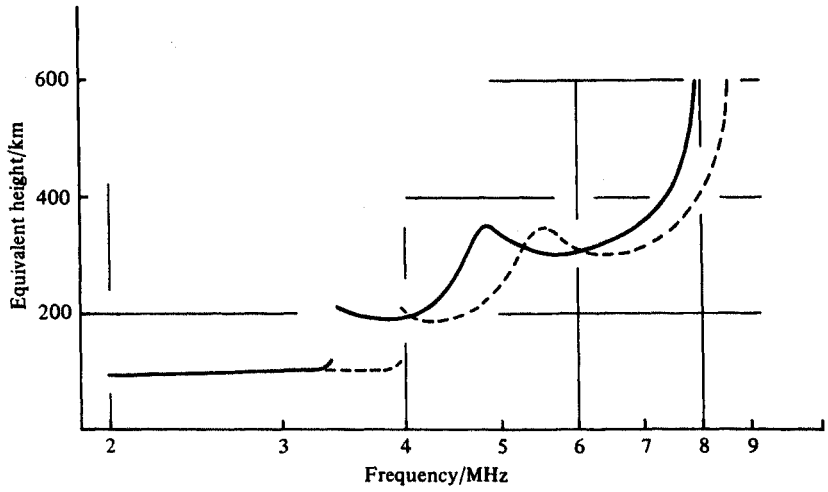


Fig. 13.4. Sketch of a typical ionogram for an undisturbed summer day in temperate latitudes. The trace for the extraordinary wave is shown as a broken line. Note that a logarithmic scale of frequency is used. This is the usual practice in published ionograms.



When $f < f_H$ the extraordinary wave is reflected where $X = 1 + Y$, that is at a higher level than the ordinary wave. Consequently $h'_X(f)$ is in general greater than $h'_O(f)$ but again there can be exceptions. Now let f be increased so that it approaches f_H. Then $h'_X(f)$ gets very large, but this is quite different from the behaviour when f approaches a penetration frequency. It occurs because n' is very large where X and $Y - 1$ are small. There are two examples in fig. 13.2.

Fig. 13.3 shows the calculated ionogram for a model of the ionosphere consisting of two parabolic layers simulating the E- and F-layers. The curve $h'_O(f)$ is similar to that obtained when the earth's magnetic field was neglected, § 12.4 and fig. 12.4. The curve $h'_X(f)$ is in four parts. The left branch (a) is for reflection from the lower layer and goes to infinity at the penetration frequency $f_p^{(X+)}$ (13.6). The next branch (b) is for reflection from the upper layer. It shows an upward turn at the left end because the waves have just penetrated the lower layer but are still retarded as they pass through it. At the penetration frequency of the lower layer for the ordinary wave, the extraordinary branch (b) shows a maximum of $h'_X(f)$. This is because, near the maximum $N(z)$ in the lower layer, X is close to unity. Now fig. 5.17 shows that near $X = 1$, the group refractive index n' for the extraordinary wave has a large value and this occurs over a large range of height z near the maximum of $N(z)$. Hence the pulse is markedly retarded in the lower layer. For other frequencies, the level where $X \approx 1$ is not near the maximum. It extends over a smaller range of height and the retardation is less. For the branch (b), $h'_X(f)$ tends towards infinity at its right-hand end as the frequency f approaches f_H, but this is not associated with penetration of any layer.

The branch (c) is for f just greater than f_H, and is for reflection from the lower layer. It shows penetration at the frequency $f_p^{(X-)}$, (13.4). The remaining branch (d) is for reflection from the upper layer, and shows a large $h'_X(f)$ at its left-hand end because the waves have then just penetrated the lower layer but are still retarded by it. At the right-hand end it goes to infinity as f approaches the penetration frequency $f_p^{(X-)}$ for the upper layer, (13.4).

Although the two-parabola model used for fig. 13.3 is very simple compared with the actual $N(z)$, the calculated ionogram shows many of the features of observed ionograms. In published ionograms the frequency range covered rarely extends down to the electron gyro-frequency or below it. There are several reasons for this. In the day time radio waves with frequencies of order 1 MHz or less are heavily absorbed in the D-region, so that the reflected amplitudes are small. Ionosonde receivers are subject to interference at these frequencies from transmitters in the broadcast frequency bands. As the frequency is decreased it is necessary to use larger and more elaborate aerials to achieve efficient radiation and to avoid interfering with broadcast reception. The absorption is particularly great for the extraordinary wave near the gyro-frequency so that here its behaviour in ionograms is not easy to

observe. At night, however, the absorption is less. Ionograms extending down to 50 kHz have been published (Watts and Brown, 1954; Watts, 1957) that display the maximum in branch (b) of the $h'_X(f)$ curve; fig. 13.3. For frequencies greater than f_H the general form of the ionogram in fig. 13.3 is similar to actual ionograms observed at night. In the day time, however, the penetration frequency of the F-layer is much greater and there is, nearly always, an F1-layer giving a ledge in the function $N(z)$, (see § 12.5 and fig. 12.5) with a resulting maximum in both $h'_O(f)$ and $h'_X(f)$; see fig. 13.4.

The ionosonde technique has been widely used in ionospheric research and the number of published ionograms is very great. For some collections of typical ionograms see, for example, Pickle (1951), Wright and Knecht (1957), Stevens (1961), Rawer and Suchy (1967). For a detailed study of the interpretation of ionograms see Piggott and Rawer (1972, 1978).

The above description applies mainly to the special simple case when the ionosphere is horizontally stratified. In practice ionograms can be very complicated because of disturbing factors. For example when the layers are tilted the transmitted and received rays are oblique. Irregularities in the ionospheric plasma can make the ionogram curves broad and irregular. There is sometimes a thin layer at a height of about 100 km, known as sporadic E, abbreviation E_S, that has a penetration frequency comparable with or greater than that of the F-layer, so that it gives an almost horizontal line at about 100 km on the ionograms, and the reflections from higher levels are masked by it. (See, for example, Smith, E.K. and Matsushita, 1962.) In polar regions the ionosphere is nearly always disturbed and its behaviour differs from that at lower latitudes. These and other disturbances are not considered further in this book. For a general description and references see, for example, Rawer and Suchy (1967), Davies (1969), Ratcliffe (1972).

13.5. Topside sounding

Ionograms are frequently recorded with an ionosonde equipment in a satellite above the maximum of the F2-layer. Then the radio pulses travel downwards and return to the satellite after reflection. The time of travel is $2h'(f)/c$ and $h'(f)$ is often called the 'equivalent depth.' The curves are often presented with the ordinate $h'(f)$ plotted downwards. The following description applies only to the simplest ideal conditions. Fig. 13.6 is a sketch of a typical topside ionogram for this case.

The transmitter and receiver are immersed in the ionospheric plasma. It is here assumed that below them the electron concentration $N(z)$ and thence the plasma frequency $f_N(z)$ increase monotonically as the height z decreases, until the maximum of the F2-layer is reached. There is no intervening maximum and no ledge. It is further assumed that the ionosphere is horizontally stratified so that the wave normals are everywhere vertical. The transmitted pulse splits into ordinary and

extraordinary components, and these suffer lateral deviation, as described in § 10.12, but it is the vertical component $\mathscr{U}_z$ of the group velocity that determines the time of travel.

To describe topside ionograms it is useful to plot a diagram, fig. 13.5, with f as abscissa and f_N as ordinate, increasing downwards. Separate diagrams are shown for the ordinary and extraordinary waves.

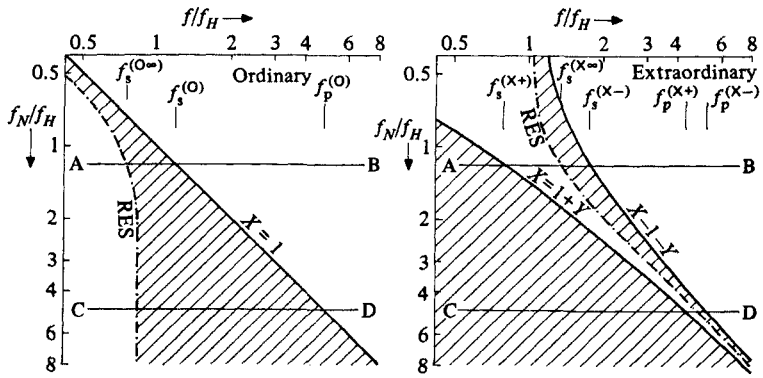
The horizontal lines AB are drawn where f_N is the plasma frequency $f_s^{(O)}$ at the satellite, and the lines CD where f_N is the penetration frequency $f_p^{(O)}$ of the F2-layer. The continuous lines show where the refractive index $n = 0$ and the waves are reflected, namely $X = 1$ for the ordinary wave and $X = 1 \mp Y$ for the extraordinary wave. The chain line shows where n is infinite. It is given by (4.76) with $U = 1$, and depends on the angle Θ between the earth's magnetic field and the vertical. It corresponds to the value of X marked RES in figs. 4.3, 4.6. The shaded regions between these lines show where the waves are evanescent and cannot be propagated. The various transition values of the frequency f at the satellite are indicated by a subscript s. Thus $f_s^{(O)}$ is where $X = 1$, and $f_s^{(X-)}$, $f_s^{(X+)}$ are where $X = 1 \mp Y$ respectively, and $f_s^{(O\infty)}$ $f_s^{(X\infty)}$ are where n and n' are infinite. The equivalent depth of reflection is given by

$$h'(f) = -\int_{f_s^0}^{f_R} n'(f, f_N) \frac{\mathrm{d}z}{\mathrm{d}f_N} \mathrm{d}f_N \tag{13.19}$$

compare (13.9), where f_R is the value of f_N at the reflection level.

Consider now the ordinary wave. If $f > f_p^{(O)}$ the wave penetrates the ionosphere. It is reflected at the ground and returns to the satellite. If f only slightly exceeds $f_p^{(O)}$, the group retardation near the maximum of the F-layer is large. If f is slightly less

Fig. 13.5. The f/f_H vs f_N/f_H diagrams used in the study of topside ionograms. The chain curves marked RES are where one refractive index is infinite. They depend on the angle Θ between the earth's magnetic field and the vertical, and in this example $\Theta = 30°$.

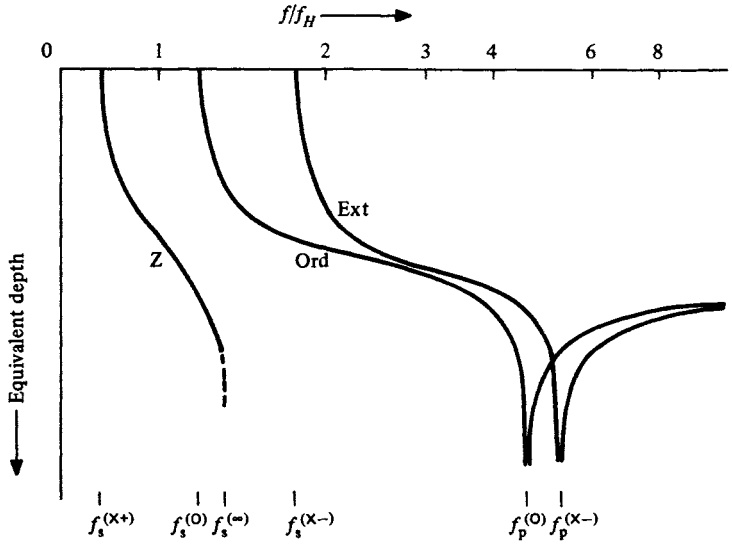


than $f_p^{(O)}$ the wave is reflected just above the maximum of the F-layer and again the group retardation is large, just as in ionograms observed at the ground, fig. 13.3. If f is decreased, the equivalent depth $h'(f)$ gets less. When f slightly exceeds $f_s^{(O)}$ the reflection level is at only a small distance below the satellite and $h'(f)$ is small. It tends to zero when $f \to f_s^{(O)}$. The extraordinary wave is similar. It penetrates the ionosphere and is reflected from the ground when $f > f_p^{(X-)}$. Its $h'(f)$ tends to zero when $f \to f_s^{(X-)}$. These effects are seen in the sketch, fig. 13.6. The ground reflections are at the right-hand end.

An ordinary wave can also be launched from the satellite if $f < f_s^{(O\infty)}$; see fig. 13.5(a). This is where $f < f_H$ and the wave is in the whistler mode; § 13.8. The figure shows that there is no value of f_N where a wave of this frequency can be reflected in the ionosphere. It therefore penetrates through to the lower ionosphere. Here the effect of collisions cannot be ignored for these low frequencies. The wave might be reflected at the ground and return to the satellite but such reflections are rarely seen in topside ionograms.

An extraordinary wave can be launched from the satellite if $f_s^{(X+)} < f < f_s^{(X\infty)}$. It is reflected where $X = 1 + Y$ and its $h'(f)$ curve is called the 'topside Z-trace'. If f slightly exceeds $f_s^{(X+)}$, the value of $h'(f)$ is small and tends to zero when $f \to f_s^{(X+)}$. In the conditions of fig. 13.5(b) this wave could never penetrate the ionosphere. When f increases and approaches the value $f_s^{(X\infty)}$, the value of X near the satellite approaches the value X_∞ that makes the group refractive index n' infinite. It was shown in § 5.9 that n' is then of order $(X - X_\infty)^{-\frac{3}{2}}$; see (5.82). It can thence be shown that the contribution to $h'(f)$, (13.19), from near the lower limit of the integral, tends to

Fig. 13.6. Sketch to show the general form of a topside ionogram.



infinity when $f \to f_s^{(X\infty)}$. This is quite different from the behaviour near a reflection level, where $X = X_R$. Then n' is of order $(X - X_R)^{-\frac{1}{2}}$, see (5.78), (5.79), and $h'(f)$ tends to a bounded limit, as was shown in § 13.3. When f increases and approaches $f_s^{(X\infty)}$, therefore, $h'(f)$ tends to infinity because of the large group retardation near the satellite. This infinity is not connected with penetration. In this respect the topside Z-trace is different from the Z-trace in ionograms recorded at the ground, § 16.8, which show $h'(f) \to \infty$ where penetration of the F-layer is approached.

When f is near to $f_s^{(X\infty)}$, conditions are near resonance and it was shown in § 5.4 item (9) that then the ray and the wave normal are nearly at right angles. Since the wave normal is vertical, the path of the wave packet must be nearly horizontal, and it can be shown that it is towards the equator. It can be seen from fig. 13.5(b) that, for the topside Z-trace, the level where $X = 1$ is between the satellite and the reflection level. Here, from § 5.2 item (8), $n = 1$ whence (5.83) shows that $\alpha = 0$, and the path of the wave packet is vertical. At still lower levels the angle α between the ray path and the wave normal has changed sign, so that the downward path here is directed away from the equator. For some examples of the ray paths for all three topside sounder traces see Davies (1969).

Actual topside ionograms are never as simple as is implied by fig. 13.6. The ionosphere may not be horizontally stratified so that there are oblique reflections. It may be disturbed so that the traces are broadened and irregular. The function $N(z)$ may not be monotonic, or may contain a ledge. The most striking thing of all is that the ionograms show resonances or spikes at certain frequencies. These occur because the transmitted energy excites the plasma surrounding the satellite into some kind of resonance that decays in a time comparable with the interval between transmitted pulses. They show in the ionogram as vertical lines extending down from $h'(f) = 0$. They are sometimes so intense that the actual ionogram is partly masked and difficult to read. They occur at the frequencies $f_s^{(O)}$, $f_s^{(X-)}$, $f_s^{(X+)}$, $f_s^{(O\infty)}$, $f_s^{(X\infty)}$, $2f_H$, $3f_H, \ldots$ and less strongly at f_H, and at the upper hybrid frequency $(f_s^{(O)2} + f_H^2)^{\frac{1}{2}}$. These resonances are not within the realm of radio propagation, but they are extremely important in plasma physics and are studied in books on that subject; see, for example, Lockwood (1963). For the theory see Fejer and Calvert (1964), Dougherty and Monaghan (1965), Sturrock (1965), Nuttall (1965a, b).

The details of all aspects of topside sounding, including the resonances, were discussed in a series of papers in a special issue of *Proc. I.E.E.E* in June 1969. For good general accounts see Rawer and Suchy (1967), Davies (1969), Jackson, Schmerling and Whitteker (1980).

13.6. The calculation of electron concentration $N(z)$ from $h'(f)$

Some examples have been given of $h'(f)$ curves calculated when $N(z)$ is known. Of much greater practical importance is the calculation of $N(z)$ when $h'(f)$ is known.

This problem has already been discussed, in § 12.6, for the case when the earth's magnetic field is neglected, and it was shown that the integral equation (12.29) that relates $h'(f)$ to $N(z)$ can be solved analytically in the range of z where $N(z)$ is a monotonically increasing function. When the earth's magnetic field is allowed for, however, the integral equation is more complicated and can only be solved by numerical methods. The integral (13.8) or (13.9) is expressed as a sum of discrete terms that can be written as a matrix product. The problem is then solved by matrix inversion or some equivalent process.

The first method to be described, or variants of it, formed the basis of calculations that were widely used for many years. It is now superseded but it is useful for illustrating the general principles. It is supposed that the ionosphere is divided into discrete laminae, and this type of method is therefore called a 'lamination' method. It is explained here in terms of the ordinary wave and for ionograms observed at the ground, but it can easily be extended to use the extraordinary wave and for topside ionograms. Electron collisions are neglected.

The integral equation to be studied is (13.9) with $f_R = f$. It is to be solved to give the function $z(f_N)$. Suppose that $h'(f)$ is given as a tabulated function at equal intervals Δf of f, and let

$$h'(r\Delta f) = h'_r \tag{13.20}$$

where r is an integer. The range of integration in (13.9) is divided up into discrete intervals Δf so that for the s^{th} interval

$$(s-1)\Delta f < f_N < s\Delta f \tag{13.21}$$

where $s \leqslant r$. It is assumed that, in each interval, $\mathrm{d}z/\mathrm{d}f_N$ is constant and given by

$$\mathrm{d}z/\mathrm{d}f_N \approx (z_s - z_{s-1})/\Delta f \tag{13.22}$$

where z_s means $z(s\Delta f)$. Now let

$$M_{rs} = \Delta f^{-1} \int_{(s-1)\Delta f_N}^{s\Delta f_N} n'(r\Delta f, f_N)\,\mathrm{d}f_N \quad (s \leqslant r), \tag{13.23}$$

$$M_{rs} = 0 \qquad (s > r).$$

Then equation (13.9) becomes, for all r

$$h'_r = \sum_{s=1}^{r} M_{rs}(z_{s-1}) \tag{13.24}$$

which may be written

$$\begin{bmatrix} h'_1 \\ h'_2 \\ h'_3 \\ \vdots \\ h'_r \end{bmatrix} = \begin{bmatrix} M_{11} & 0 & 0 & \cdots \\ M_{21} - M_{22} & M_{22} & 0 & \cdots \\ M_{31} - M_{32} & M_{32} - M_{33} & M_{33} & \cdots \\ \vdots & \vdots & \vdots & \\ M_{r1} & M_{r2} & M_{r3} & \cdots \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ \vdots \\ z_r \end{bmatrix}. \tag{13.25}$$

Let $\boldsymbol{A}$ denote the square matrix on the right. Then the equation may be written in matrix notation and solved by inversion of $\boldsymbol{A}$, thus

$$\boldsymbol{h}' = \boldsymbol{A}\boldsymbol{z}, \quad \boldsymbol{z} = \boldsymbol{A}^{-1}\boldsymbol{h}'. \tag{13.26}$$

The matrix $\boldsymbol{A}$ is lower triangular, that is, all elements above the leading diagonal are zero. It is therefore possible to find the elements of $\boldsymbol{z}$ by solving the equation (13.25) for $z_1, z_2 \ldots$ in succession. Let

$$B_{rs} = -A_{rs}/A_{rr} \quad \text{for } s \neq r, \quad B_{rr} = 1/A_{rr}. \tag{13.27}$$

Then from (13.25)

$$z_r = \sum_{s=1}^{r-1} B_{rs} z_s + B_{rr} h'_r. \tag{13.28}$$

This process is equivalent to the second equation (13.26) but there is no need to invert $\boldsymbol{A}$.

The numbers B_{rs} depend only on the properties of the plasma and not on $h'(f)$ or $z(f_N)$. To use the method a table of these numbers must first be computed. In practical cases a value $\Delta f = 0.1$ MHz was used. To cover the frequency range 2 to 6 MHz the table contains 861 numbers. For an example see Budden (1961a). A separate table is needed for each ground observing station. Once computed, it can be used with any set $\boldsymbol{h}'$ to find $\boldsymbol{z}$. The $h'(f)$ data from ionograms do not extend down to $f = 0$ and in many practical cases the smallest f used is about 1 to 2 MHz. The method can be extended to deal with this. For these and other details, and for examples of the tables see Thomas, Haselgrove and Robbins (1958), Titheridge (1959), Thomas and Vickers (1959), Budden (1961a), Jackson (1969).

In the above version of the method it was supposed that $h'(f)$ is given at equal intervals Δf of the frequency f. Instead of using f, any monotonic function of it may be used. For example King (1957) has shown that there is some advantage in using $\ln f$. For other ways of increasing the accuracy of the method see, for example, Titheridge (1979).

In more recent methods of calculating $N(z)$ from $h'(f)$ the need for a large table of numbers has been avoided. Instead of using a sequence of values of $z(f_N)$ as in (13.25), this function is expressed in some other way. The idea was suggested first by Titheridge (1961a) who used a polynomial expansion for $z(f_N)$ thus

$$z(f_N) = \sum_{j=1}^{r} \alpha_j f_N^{j+1} \tag{13.29}$$

and his method is known as the 'polynomial method'. Here z is measured from the base of the ionospheric layer, where $f_N = 0$. The term in f_N was not included because it would impose the restriction that $\mathrm{d}z/\mathrm{d}N$ is infinite at the base of the layer where $N = 0$, whereas Titheridge assumed that $\mathrm{d}z/\mathrm{d}N$ is bounded here. Substitution of (13.29) into (13.9) with $f_R = f$ gives

$$h'(f) = \sum_{j=1}^{r} \alpha_j (j+1) \int_0^f n'(f, f_N) f_N^j \, \mathrm{d}f_N. \tag{13.30}$$

Suppose now that observed values of $h'(f)$ are given at r different values $f_1, f_2, \ldots, f_r$ of f. Let $h'(f_i)$ be written h_i and let $\boldsymbol{h}'$ be the column matrix of the r values of h_i. Similarly let $\boldsymbol{\alpha}$ be the column matrix of the r values of α_j in (13.29). Then (13.30) gives

$$\boldsymbol{h}' = \boldsymbol{B}\boldsymbol{\alpha} \tag{13.31}$$

where $\boldsymbol{B}$ is a square matrix with elements

$$B_{ij} = (j+1) \int_0^{f_i} n'(f_i, f_N) f_N^j \, \mathrm{d}f_N. \tag{13.32}$$

Equation (13.29) may be written for r different values $f_{N1}, f_{N2}, \ldots, f_{Nr}$ of f_N. Let z_i denote $z(f_{Ni})$ and let z be the column matrix of the z_i's. Then (13.29) gives

$$z = \boldsymbol{A}\boldsymbol{\alpha} \tag{13.33}$$

where $\boldsymbol{A}$ is the square matrix with elements

$$A_{ij} = f_{Ni}^{j+1} \tag{13.34}$$

and on substitution for $\boldsymbol{\alpha}$ from (13.31)

$$z = \boldsymbol{A}\boldsymbol{B}^{-1}\boldsymbol{h}'. \tag{13.35}$$

This is the required solution. The matrix $\boldsymbol{A}\boldsymbol{B}^{-1}$ must be computed first. Titheridge gave examples where $r = 6$ so that the matrix contains only 36 elements. It can be computed very quickly and there is no need for the large table of numbers used with the older method.

The simplest form of the polynomial method described above has some disadvantages which were discussed by Titheridge (1961a) who showed how to modify and extend the method so as to overcome them. In later papers, Titheridge (1967a, b, 1969, 1975), Titheridge and Lobb (1977), the method has been adapted to deal with the extraordinary wave and with topside ionograms, and to give information about $N(z)$ in the lowest parts of the ionosphere where f_N is less than the smallest f used in the $h'(f)$ data.

The expansion (13.29) for z is a single valued function f_N. For values of z near where $f_N(z)$ has a maximum, $z(f_N)$ is necessarily a two-valued function, so that (13.29) cannot show a maximum of $f_N(z)$. In spite of this the method can still be used for z values that are very close to the maximum of a layer.

No method can give the function $N(z)$ completely in a valley region such as that shown between P and Q in fig. 12.7(b), because no ionogram can contain the information. But some properties of the distribution can be found. For example the minimum value of $N(z)$ can be estimated but its height cannot. For a discussion of this type of problem see Becker (1967), Titheridge (1975), Lobb and Titheridge (1977), and the references given below.

Clearly the polynomial expansion (13.29) can be replaced by other expansions. One recent development should be briefly mentioned. From the ionogram the penetration frequency $f_p^{(O)}$ of the F-layer can first be found. Then an expansion of the form

$$z(f_N) = z_m + \{\ln(f_p^{(O)}/f_N)\}^{\frac{1}{2}} \sum_j A_j \mathscr{P}_j(f_N) \tag{13.36}$$

can be used. Here z_m is the height, at first unknown, of the maximum of the layer. By using the factor $\{\ln(f_p^{(O)}/f_N)\}^{\frac{1}{2}}$, the difficulty of representing the parabolic shape of $N(z)$ near its maximum by a single valued polynomial is avoided. The $\mathscr{P}_j(f_N)$ are polynomials of some function of f_N. The unknown coefficients are the A_js and z_m. These correspond to the α_js in (13.29), and are found by a matrix inversion analogous to $\boldsymbol{B}^{-1}$ in (13.35).

Reinisch and Huang (1983) used an expansion of the type (13.36) for analysing ionograms observed at the ground. For the $\mathscr{P}_j(f_N)$ they used suitably chosen Chebyshev polynomials (Snyder, 1966). Huang and Reinisch (1982) used a similar method for topside ionograms.

13.7. Faraday rotation

When a linearly polarised wave enters an anisotropic plasma it splits into the two characteristic waves, ordinary and extraordinary, and each has its own polarisation ρ_O, ρ_E respectively. The two waves have different phase velocities so that their phase difference changes as they travel. At some point along their path the resultant signal is found by adding the fields of the two waves, but because of the phase difference the resulting polarisation is not the original linear polarisation. For a very simple example of this consider a homogeneous collisionless electron plasma with the superimposed magnetic field parallel to the z axis. Let the wave normal of the incident wave, and therefore also the wave normals of the two characteristic waves, be parallel to the z axis. In these conditions the ray and wave normal directions are the same. Let the incident wave start from $z = 0$ and travel in the direction of positive z and let it be linearly polarised with $\boldsymbol{E}$ parallel to the x axis. Suppose that $\boldsymbol{Y}$ is in the positive z direction, $Y < 1$, and $X < 1 - Y$. Then the ordinary wave is circularly polarised with a right-handed sense; see fig. 4.2. Its electric field has a constant amplitude A_O, and where $z = 0$ the components are

$$E_x^{(O)} = A_O e^{i\omega t}, \quad E_y^{(O)} = -i A_O e^{i\omega t} \tag{13.37}$$

Similarly the extraordinary wave is circularly polarised with a left-handed sense and its electric field, of amplitude A_E, has components

$$E_x^{(E)} = A_E e^{i\omega t}, \quad E_y^{(E)} = i A_E e^{i\omega t}. \tag{13.38}$$

The linearly polarised incident wave is the sum of (13.37), (13.38) so that $A_O = A_E$,

giving

$$E_x = 2A_O e^{i\omega t}, \quad E_y = 0. \tag{13.39}$$

When the wave has travelled a distance z, the fields are, for the ordinary wave

$$E_x^{(O)} = A_O \exp\{i(\omega t - kzn_O)\}, \quad E_y^{(O)} = -iE_x^{(O)}. \tag{13.40}$$

and for the extraordinary wave

$$E_x^{(E)} = A_O \exp\{i(\omega t - kzn_E)\}, \quad E_y^{(E)} = iE_x^{(E)}. \tag{13.41}$$

The total fields are the sum of these, namely

$$\left.\begin{matrix} E_x \\ E_y \end{matrix}\right\} = 2A_O \exp[i\{\omega t - \tfrac{1}{2}kz(n_O + n_E)\}] \left\{\begin{matrix} \cos\{\frac{1}{2}kz(n_O - n_E)\} \\ \sin\{\frac{1}{2}kz(n_O - n_E)\} \end{matrix}\right\}. \tag{13.42}$$

Thus the resultant wave is linearly polarised and its plane of polarisation makes an angle

$$\Psi = \tfrac{1}{2}kz(n_O - n_E) \tag{13.43}$$

with the x axis. This is half the phase difference between the two waves. The plane of polarisation of the resultant wave rotates as z increases. This effect was studied by Faraday (see, for example, Jenkins and White, 1976). It was observed in the laboratory for light waves travelling in a transparent medium with a magnetic field applied in the direction of travel. It occurs not only for the free electrons in a plasma as illustrated here, but for the bound electrons in transparent solids and liquids. It is known as the 'Faraday effect', or 'Faraday rotation'. The spatial rate of rotation is

$$\partial\Psi/\partial z = \tfrac{1}{2}k(n_O - n_E). \tag{13.44}$$

If the waves travel in the opposite direction, the values $\rho_O = -i$, $\rho_E = i$ for the same axes are unchanged; see end of § 4.4. Thus the Faraday rotation is in the same absolute direction. If the wave system is reflected and retraces its path, the plane of polarisation does not go back to its initial plane, but continues rotating in the same sense as before the reflection.

In this example the refractive indices n_O, n_E are given by (4.84). One important application is for high frequencies so that X and Y are small. If X^2, Y^2 and higher powers can be neglected, this gives

$$\tfrac{1}{2}(n_O - n_E) \approx \tfrac{1}{2}XY, \quad \tfrac{1}{2}(n_O + n_E) \approx 1 - \tfrac{1}{2}X. \tag{13.45}$$

The second expression is the average of the two refractive indices and is in the exponential of (13.42). It may be called the 'effective' value of the refractive index, and for this approximation it is the same as for an isotropic plasma. If the first expression is inserted in (13.44) and if (3.5), (3.25) are used for X, Y respectively, then

$$\partial\Psi/\partial z = NBe^3/(8\pi^2\varepsilon_0 cm^2 f^2) \tag{13.46}$$

where B is the strength of the magnetic field. This has been derived for a homogeneous plasma. If N and B are slowly varying functions of z, the equation may

be integrated with respect to z to give the total rotation for a given path; see (13.51) below.

Consider again a homogeneous plasma and suppose, as before, that the source is at the origin and the receiver is on the z axis. Let the vector $\boldsymbol{Y}$ be in the x–z plane at an angle β to the z axis. The emitted linearly polarised signal again splits into the two characteristic component waves. Both ray directions are along the z axis but the two wave normal directions are different. They are in the x–z plane and make angles Θ_O, Θ_E with $\boldsymbol{Y}$, and angles

$$\alpha_O = \Theta_O - \beta, \quad \alpha_E = \Theta_E - \beta \tag{13.47}$$

with the z axis; compare (5.32) and fig. 5.1. The two electric fields now have components E_z, but we here study only the E_x and E_y components. The two polarisations are no longer exactly circular but if β is small enough they are approximately circular and the two waves have approximately equal amplitudes A_O; compare (13.40), (13.41). If they are recombined after a distance z they give a resultant field similar to (13.42) but in general the polarisation is no longer exactly linear. It is an elongated ellipse whose major axis makes an angle Ψ with the x axis. In particular if the phase difference between the waves is $2\pi m$ where m is an integer, their combination has linear polarisation similar to the emitted wave at $z = 0$, but rotated through an angle $\Psi = \pi m$. The rotation angle is half the phase difference, as before.

The two refractive indices n_O, n_E depend on the wave normal angles Θ_O, Θ_E respectively. On the z axis the components E_x for the two electric fields are

$$E_x^{(O)} = A_O \exp\left[\mathrm{i}\{\omega t - kzn_O(\Theta_O)\cos\alpha_O\}\right], \quad E_x^{(E)} = A_O \exp\left[\mathrm{i}\{\omega t - kzn_E(\Theta_E)\cos\alpha_E\}\right]. \tag{13.48}$$

The two products $n\cos\alpha$ are the ray refractive indices, $\mathscr{M}$; see § 5.3. The spatial rate of Faraday rotation is $\partial/\partial z$ of half the phase difference, that is, compare (13.44)

$$\partial\Psi/\partial z = \tfrac{1}{2}k\{n_O(\Theta_O)\cos\alpha_O - n_E(\Theta_E)\cos\alpha_E\}. \tag{13.49}$$

Consider again the application to high frequencies so that X and Y are small. Now (5.83) shows that α is of order XY^2. Note that in (5.83) both $n^2 - 1$ and $An^2 - B$ have a factor X. If X^2, Y^2 and higher powers are negligible, both α_O and α_E are negligibly small so that $\Theta_O = \Theta_E = \beta$. The two refractive indices are now given by (4.113). Hence (13.49) gives

$$\partial\Psi/\partial z \approx \tfrac{1}{2}k\{n_O(\beta) - n_E(\beta)\} \approx \tfrac{1}{2}kXY\cos\beta. \tag{13.50}$$

The only change needed in (13.46) is the insertion of a factor $\cos\beta$. In a slowly varying medium N, B and β may all depend on z. Then integration with respect to z gives for the Faraday rotation

$$\Psi = \frac{e^3}{8\pi^2\varepsilon_0 cm^2 f^2}\int_0^z NB\cos\beta\,\mathrm{d}z. \tag{13.51}$$

This formula is often used in radio astronomy; see, for example Kraus (1966, eqn. 5–76). It is applied to the interstellar and intergalactic plasmas where X and Y are extremely small. The refraction of the rays is then negligible and it can safely be assumed that the two component waves have the same ray path.

Measurements have been made of the Faraday rotation of radio signals received at the ground from a satellite in the upper part of the ionosphere; see for example Titheridge (1966, 1968, 1971a); Garriott, Da Rosa and Ross (1970). The frequencies used were typically 20 MHz to 137 MHz, so that X was small. Hence the refraction of the rays could be neglected, and (13.51) could be used. Now z is used for height and s for distance along the path, so that $\mathrm{d}z$ is replaced by $\mathrm{d}s = \mathrm{d}z \sec\theta$ where θ is the angle between the path and the vertical. Throughout the ionosphere B and β can be taken as constant to a first approximation, so that (13.51) gives:

$$\Psi = \frac{e^3 B \cos\beta \sec\theta}{8\pi^2 \varepsilon_0 c m^2 f^2} \int_0^z N \,\mathrm{d}z. \tag{13.52}$$

The integral in (13.52) is called the 'total electron content' of the ionosphere up to height z, and Faraday rotation is one of the most important ways of measuring it.

When the approximation, that X and Y are very small, cannot be made, the problem is more complicated. When a linearly polarised wave enters the ionosphere, the two component rays undergo lateral deviation in opposite directions; see §§ 10.12–10.14 and figs. 10.11–10.13. The two ray directions at the transmitter are different and must be chosen so that the two rays both reach the receiver. Thus the two ray paths are different. To find the Faraday rotation for this problem it may be necessary to trace the two rays separately and compute their phase paths P_O, P_E. The Faraday rotation angle is then $\frac{1}{2}k(P_O - P_E)$. The only case where the two ray paths are the same is for a homogeneous medium.

Table 13.1

X	Y	β/deg	$n_O(\Theta_O)\cos\alpha_O - n_E(\Theta_E)\cos\alpha_E$	$n_O(\beta) - n_E(\beta)$	$XY\cos\beta$
0.1	0.1	10	0.0105	0.0105	0.0098
0.7	0.1	10	0.1296	0.1298	0.0689
0.1	0.1	50	0.0068	0.0068	0.0064
0.7	0.1	50	0.0863	0.0868	0.0450
0.1	0.5	10	0.0706	0.0706	0.0492
0.4	0.5	10	0.4044	0.4054	0.1970
0.1	0.5	50	0.0479	0.0484	0.0321
0.4	0.5	50	0.3102	0.3193	0.1286
0.1	1.1	10	0.431	0.426	0.108
0.7	1.1	10	1.977	1.600	0.758
0.1	1.1	50	0.286	0.257	0.071
0.7	1.1	50	1.210	0.560	0.495

The use of the approximation $n_O(\beta) - n_E(\beta)$ in (13.50) implies that the two wave normal directions are the same as the ray direction β. It is important to know by how much this approximation differs from the correct expression $n_O(\Theta_O)\cos\alpha_O - n_E(\Theta_E)\cos\alpha_E$ from (13.49). Table 13.1 gives some typical computed values for these two quantities and for the approximation $XY\cos\beta$ from (13.50).

13.8. Whistlers

In §4.12, fig. 4.6 the values of n^2 for a collisionless electron plasma were studied and it was shown that for $f < f_H$ the value of n^2 for the ordinary wave is positive for large X and for sufficiently small values of the angle Θ between the wave normal and the earth's magnetic field. The refractive index and ray surfaces are as shown in figs. 5.5, 5.6. This behaviour of the ordinary wave extends to frequencies down in the audible range. This type of wave is known as the 'whistler mode'. To study it we first neglect the effects of heavy positive ions and of collisions. These two effects are considered in the following section. An effect associated with protons is mentioned briefly at the end of this section.

If an audio-frequency amplifier, with headphones or a loud-speaker, is connected to a suitable aerial, such as a long wire as high as possible, various atmospheric noises can be heard and sometimes these include a whistle of falling pitch lasting for several seconds, known as a 'whistler'. The existence of whistlers has been known for many years (Burton and Boardman, 1933; Eckersley, 1935) but the main stimulus to the study of the subject came from the work of Storey (1953). For a full account of all aspects including the history, see Helliwell (1965), and for reviews see Al'pert (1980a, 1983), Carpenter (1983), Walker (1976).

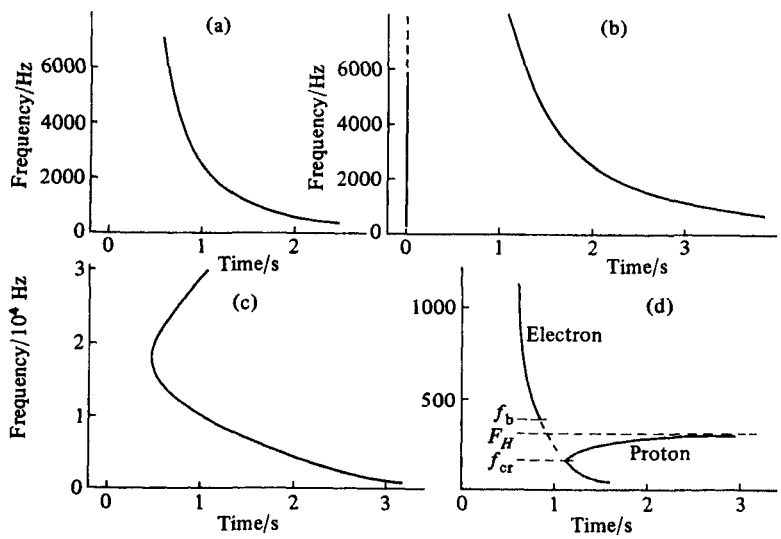
It is now believed that a whistler originates in a lightning flash near the earth's surface. This is an impulse signal that contains a wide range of frequencies including those in the audible range. Some energy in this range can penetrate to the upper ionosphere and magnetosphere. The ray surfaces, fig. 5.6, show that for the low frequencies, that is large X, the ray directions are confined within a cone surrounding the direction of the earth's magnetic field, and Storey (1953) suggested that this causes the rays to be guided roughly in the direction of the magnetic field lines. Thus the signal travels over the equator and returns to earth at a point somewhere near the other end of the magnetic line of force along which it started out. The ray directions in fig. 5.6 may make angles up to 20° or more with the earth's magnetic field, and it is now known that this is not adequate to give the observed guiding. For a study of whistler ray paths see Yabroff (1961), Hashimoto, Kimura and Kumagai (1977). There are ducts in the magnetosphere like tubes containing an enhanced electron concentration, aligned with the magnetic lines of force. These act as wave guides and are now known to cause the guiding.

The two refractive indices are given by (4.46). If collisions are neglected ρ is purely

imaginary and $U \approx 1$. For the extremely small frequencies in whistlers X and Y are large, so that the U and the 1 can be neglected. For upgoing waves in the northern hemisphere $\cos\Theta$ is positive. Then (4.46) shows that if n is positive, ρ must be positive imaginary. Thus upgoing whistler waves are polarised with a left-handed sense of rotation in the northern hemisphere. The field vectors $\boldsymbol{E}$, $\mathscr{H}$ rotate in the same direction as free electrons would rotate about the earth's magnetic field. The convention used here to specify the sense of rotation is as explained in § 4.3.

Because of the dependence of the group velocity $\mathscr{U}$ on frequency, § 5.9, the high frequencies in the audible range travel faster, and the impulse is therefore drawn out into a signal of descending pitch, and is heard as a descending whistle. A whistler that makes one journey over the equator is called a 'short whistler', fig. 13.7(a). When it comes down into the ionosphere it can be reflected and return along roughly the same path to a region near its origin. It is then dispersed twice as much as a short whistler so that its duration is roughly doubled, fig. 13.7(b). It is called a 'long whistler'. The lightning flash that generates the whistler gives an impulsive signal that is often heard as a 'click' preceding a long whistler, and this signal goes directly to the receiver and so is not dispersed. Whistlers can be reflected repeatedly and make many transits over the equator. They are attenuated by collisions, and this would cause the amplitude to fall below the detection level after only a few transits. But

Fig. 13.7. Sketches showing spectrograms of typical whistlers. (a) is a short whistler with $T\sqrt{f} = 50\,\mathrm{s}^{\frac{1}{2}}$. (b) is the same whistler that has made two passages over the equator and returned to near its source. The initiating lightning flash shows as a vertical line at time $t = 0$. $T\sqrt{f} = 100\,\mathrm{s}^{\frac{1}{2}}$. (c) is a nose whistler. (d) is a fractional hop whistler as observed in a satellite, and it has an ion cyclotron (proton whistler) component.



whistler waves can be amplified by interaction with streaming charged particles in the magnetosphere and this compensates for the attenuation. Hence long trains of whistlers are sometimes recorded, all originating from the same lightning flash.

Whistler signals are analysed by feeding them to some form of spectral analyser. This produces a record with frequency as ordinate and time as abscissa and the record is darkened where any frequency is present at a given time. Some typical records are sketched in fig. 13.7. On these records the direct signal from a lightning flash contains all frequencies in the band accepted by the receiver, and appears as a vertical line, fig. 13.7(b).

For the extremely low frequencies of the audible range X may be of order 200 or more, so that the approximation (4.110) can be used. Since Y also is large, the 1 in the denominator can be neglected. Hence the refractive index n for the ordinary wave is given by

$$n \approx \left(\frac{X}{Y}\sec\Theta\right)^{\frac{1}{2}} = f_N \sec^{\frac{1}{2}}\Theta (f f_H)^{-\frac{1}{2}} \tag{13.53}$$

where Θ is the angle between the wave normal and the earth's magnetic field. Thus for these extremely low frequencies one of the two refractive indices is real and positive and the waves are propagated. It can be shown that if the Lorentz term, § 3.3, is included in the dispersion relation, neither refractive index would be real and positive and there would be no whistlers. The existence of whistlers is perhaps the strongest evidence that the Lorentz term must be omitted (Eckersley, 1935). The angle α between the ray and wave normal is given, from (5.33), by

$$\tan\alpha = \tfrac{1}{2}\tan\Theta. \tag{13.54}$$

whence the angle $\beta = \Theta - \alpha$ between the ray and the earth's magnetic field is given by

$$\tan\beta = \tan\Theta/(2 + \tan^2\Theta). \tag{13.55}$$

The maximum value of $\tan\beta = 1/\sqrt{8}$. It occurs where $\Theta = 54°44'$ and gives $\beta(\max) = 19°29'$. This is the semi-vertical angle of the cone within which the ray directions are confined.

The group refractive index n' is given, from (5.65), by

$$n' = \tfrac{1}{2}n \tag{13.56}$$

so that the group velocity is

$$\mathscr{U} = 2c/(n\cos\alpha). \tag{13.57}$$

For rays directed along the earth's magnetic field, $\Theta = \alpha = 0$, this gives

$$\mathscr{U} \approx 2c(f f_H)^{\frac{1}{2}}/f_N. \tag{13.58}$$

For whistlers guided in a duct Θ and α are usually small so that (13.58) is a good approximation. The time of travel T of a whistler from its source to the receiver is the integral of $1/\mathscr{U}$ along the path. Thus for the various frequencies in the signal $T\sqrt{f} = D$ is a constant, called the 'dispersion'. This gives the shape of the spectro-

gram for low enough frequencies. It is known as the Eckersley law (Eckersley, 1935), and is confirmed by measurements. It would be expected to fail for frequencies that are too great to allow the use of the approximation (13.53), or so small that the effect of positive ions cannot be neglected. This behaviour too is confirmed by the measurements.

When the frequency is increased, the group velocity does not continue to increase in accordance with (13.58) but attains a maximum value where f is slightly less than $\frac{1}{4}f_H$ (see problem 5.11). For this frequency the time of travel is a minimum, and for greater frequencies the time of arrival increases with frequency. Thus the spectrogram has the shape of a nose as in fig. 13.7(c), and the whistler is called a 'nose whistler.' In fact all whistlers would be nose whistlers if the frequency were extended to high enough values. The frequency where T is a minimum is called the nose frequency. It is approximately the average of $\frac{1}{4}f_H$ over the path. A long part of the whistler path is near the top of its trajectory, and it is the electron gyro-frequency here that mainly determines the nose frequency. Typical nose frequencies are of order 20 to 30 kHz. The earth's magnetic field, and therefore f_H, is proportional to the inverse cube of the distance from the earth's centre. Whistlers observed at high latitudes can travel to great heights where f_H is small, and then the nose frequency can sometimes be as small as 3 to 4 kHz.

A lightning flash at the ground emits a spherical pulse of waves into free space. Each component frequency in the pulse can be resolved into an angular spectrum of plane waves as in (10.1), and this includes plane waves with real angles of incidence θ from 0° to 90°. In the ionosphere and above, the refractive index for whistler waves is very large, of order 50 or more. All the plane waves with real θ are therefore refracted so that their wave normals are very nearly vertical. Most theories of whistler paths assume that, where a whistler first enters the topside ionosphere from below, the wave normal is exactly vertical. Above this the medium is no longer horizontally stratified. The whistler enters a duct and is guided as already described.

Whistlers of another kind have been observed with receivers in rockets or satellites in the topside ionosphere. They are known as 'subprotonic whistlers' and were described by Carpenter, Dunckel and Walkup (1964). They were first reported by Barrington and Belrose (1963). Their dispersion is less than for a normal whistler and in a spectrogram they appear as several successive traces at equal intervals, suggesting that the signal travels up and down with repeated reflections at levels of about 100 km and 1000 km. They are confined to lower frequencies less than about 4 kHz. They occur mainly at night and at latitudes greater than 45°.

It is believed that in the topside ionosphere up to about 800 km the positive ions in the plasma are mainly atomic oxygen and helium ions. Above this the relative abundance of protons becomes greater and at 1000 km and above it is the protons that predominate. It has been suggested (Smith, R.L., 1964) that the effect of the

protons on the refractive index is great enough to cause the upgoing whistler signals to be refracted into ray paths that become horizontal and are then reflected downwards. The signal thus returns to the ionosphere where the vertical gradients are large and can give an upward reflection. The region above 800 km where the proton concentration becomes appreciable is called the 'protonosphere', and these whistlers are therefore called subprotonic whistlers. Their paths have been studied by Kimura (1966) who used ray tracing methods.

In a horizontally stratified topside ionosphere where the rays become horizontal and reflection is occurring, the wave normals cannot be vertical. These waves cannot, therefore, originate from waves with real θ below the ionosphere if there is horizontal stratification at all levels. The subject has been studied by Walker (1968a, b) who showed that large values of $S = \sin\theta$ of about 10 to 25 are needed to explain the propagation and reflection of subprotonic whistlers. He gave examples of ray paths for various frequencies and values of S, with the wave normal in the magnetic meridian plane, and he was able to explain most of the observed features, assuming that the topside ionosphere is horizontally stratified. But the pulse that initiates these whistlers can only enter this topside ionosphere if there is some region of the lower ionosphere, near the lightning source, where there are strong horizontal variations. One suggestion is that subprotonic whistlers are associated with a strong but localised sporadic E-layer (see §19.8).

13.9. Ion cyclotron whistlers

In the discussion of whistlers in the first part of § 13.8 it was assumed that electrons are the only effective charged particles in the plasma. For frequencies less than about 500 Hz, however, the effect of positive ions cannot be ignored. Hines (1957) was one of the first to consider the effect of heavy positive ions on whistler propagation. This section deals with some effects where the positive ions play an essential part.

When a whistler is received by a satellite above the ionosphere it travels up from the lightning flash at the ground to the receiver. Its dispersion $D = T\sqrt{f}$ is smaller than for a whistler that has travelled over the equator because the distance travelled is smaller. It is called a 'fractional hop' whistler. The spectrogram often shows the effect sketched in fig. 13.7(d). There is a part of the trace for which the frequency increases slowly with time, and approaches a limiting value equal to the proton gyro-frequency F_H at the level of the receiver. Observations have shown that in the northern hemisphere this part of the trace is polarised with a right-handed sense of rotation, and it is called a 'proton whistler'. At lower frequencies still, a similar trace is sometimes observed, associated with helium ions and called a 'helium whistler'. These whistlers are collectively known as ion cyclotron whistlers. Proton whistlers were first reported by Smith, Brice *et al.* (1964), and helium whistlers by Barrington *et al.* (1966).

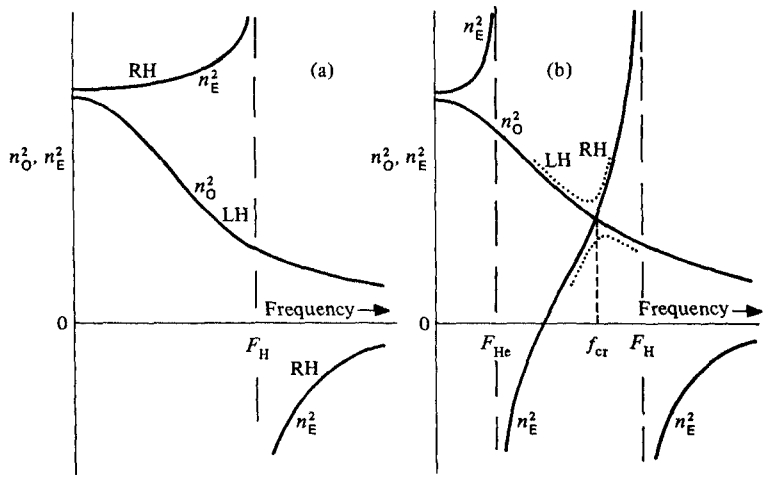
At higher frequencies the dispersion of the trace is caused mainly by electrons. This part of the upgoing signal is polarised with a left-handed sense in the northern hemisphere, and is an ordinary wave of the type discussed in the previous section. It is called an electron whistler. Part of its trace is absent or weak in the frequency range where the proton whistler appears. This suggests that the incident ordinary wave has changed to an extraordinary wave, with a reversal of the sense of the polarisation.

When the ray and wave normal are parallel or antiparallel to the earth's magnetic field, $\Theta = 0$, the two refractive indices are given by

$$n_O^2 = \varepsilon_2, \quad n_E^2 = \varepsilon_1. \tag{13.59}$$

The frequency dependence of these quantities is shown in figs. 3.1, 3.2. If only electrons are effective then at low enough frequencies ε_2 is positive and ε_1 is negative. The extraordinary wave is evanescent. But if ions are allowed for, there is a range of frequencies less than each ion gyro-frequency, where ε_1 is positive; see fig. 3.2. The frequency dependence of n_O^2 and n_E^2 is sketched in fig. 13.8. Collisions are neglected so that both are real. In fig. 13.8(a) there is only one ion species present so that n_O^2 is never equal to n_E^2. In fig. 13.8(b) there are two ion species present and there is now a frequency f_{cr} between the two ion gyro-frequencies where $\varepsilon_1 = \varepsilon_2$, that is $n_O^2 = n_E^2$. This is the 'crossover' frequency discussed in § 3.11 item (15). In both diagrams, for

Fig. 13.8. Sketches, not to scale, to show how n_O^2 and n_E^2 depend on frequency when collisions are neglected and the wave normal is parallel to the earth's magnetic field. In (a) there is only one species of positive ion, protons, with gyro-frequency F_H. In (b) there are two species, protons, and helium ions with gyro-frequency F_{He}, and f_{cr} is the crossover frequency. The dotted curves show how the values change when the wave normal makes a small angle with the earth's magnetic field. RH, right-handed, and LH, left-handed, show the sense of the polarisation for upgoing waves in the northern hemisphere.



the curves marked n_O^2 the polarisation for upgoing waves in the northern hemisphere is circular with a left-handed sense, and for the curves marked n_E^2 it is circular with a right-handed sense.

If Θ is now made slightly different from zero the square root in (4.51) can never be zero when collisions are neglected. Thus n_O^2 and n_E^2 are never equal at any frequency. The curves cannot intersect, but separate as shown by dotted curves in fig. 13.8(b). The polarisations are no longer exactly circular, but the values of ρ are purely imaginary. When the upper dotted curve of fig. 13.8(b) is followed from left to right, ρ is near $-$i for $f < f_{cr}$, giving a left-handed sense as for the electron whistler. When $f = f_{cr}$, ρ has decreased to zero, and when $f > f_{cr}$ it is near to $+$i giving a right-handed sense as for the proton whistler.

The crossover frequency f_{cr} depends on the electron and ion concentrations in the plasma, and therefore in the upper ionosphere f_{cr} depends on the height. For a wave of fixed frequency f travelling vertically upwards there may be some height, called the 'crossover level'. where $f = f_{cr}$. Then, according to ray theory, the polarisation changes continuously from left hand to right hand, or vice versa, as the wave travels upwards. Curves showing how ρ/i depends on height for various values of Θ were given by D. Jones (1970). If collisions are negligible this reversal of the sense of polarisation must happen for all waves that encounter a level where $f = f_{cr}$. This change occurs continuously as the height increases. It does not entail any kind of mode conversion. If the wave starts from the ground as an electron whistler, its name changes to 'proton whistler' after it has passed the crossover level. This change of name is similar to the change from 'ordinary' to 'extraordinary' discussed in §4.16. The change occurs at different levels for the various frequencies in the whistler wave.

Crossover cannot occur unless at least two positive ion species are present. When proton whistlers are observed there must be both protons and some heavier positive ions such as singly charged helium ions in the upper ionosphere. Similarly when helium whistlers are observed there must also be some heavier ions such as singly charged atomic oxygen ions.

The following explanation is given in terms of ray theory alone, and this is adequate for the main features of ion cyclotron whistlers. In practice some mode conversion does occur, and it is mentioned again at the end of this section.

The theory of the formation of ion cyclotron whistlers was given by Gurnett, Shawhan *et al.* (1965). They showed that in the full theory the effect of collision damping plays an important part. The following treatment is based on the work of D. Jones (1970, 1969). The second of his papers, though published earlier, is a sequel to the first.

For the extremely low frequencies in whistlers the refractive indices in the upper ionosphere are large, of order 10 to 100; see fig. 13.10 for some examples. Thus the wave normal of a wave entering from free space is refracted so that it is nearly

vertical even when the incidence is oblique. We may therefore assume that the wave normal is vertical, and for the range of height up to about 3000 km studied here we assume that the ionosphere is horizontally stratified. Now Θ is the angle between the vector $\boldsymbol{Y}$, antiparallel to the earth's magnetic field, and the vertical.

When collisions are allowed for, ε_1 and ε_2 are complex and are never equal for any real frequency. Thus n_O and n_E are never equal when $\Theta = 0$. But they are equal for some real transition value Θ_t of Θ; compare §§ 4.4, 4.16. When this occurs the two wave polarisations ρ_O, ρ_E are also equal. The condition is, from (4.24) with (3.51):

$$\mathfrak{G} = \frac{\varepsilon_3(\varepsilon_1 - \varepsilon_2)}{\varepsilon_1\varepsilon_2 - \frac{1}{2}\varepsilon_3(\varepsilon_1 + \varepsilon_2)} = \pm i\frac{\sin^2\Theta_t}{\cos\Theta_t}. \tag{13.60}$$

It requires that

$$\mathrm{Re}(\mathfrak{G}) = 0, \quad \Theta = \Theta_t \tag{13.61}$$

where

$$\sin^2\Theta_t/\cos\Theta_t = \pm\,\mathrm{Im}(\mathfrak{G}). \tag{13.62}$$

This condition is here called 'exact crossover'. The frequency f_b for which it occurs is here called the 'boundary frequency'. It is a function of the angle Θ. In simple cases there is only one value f_b in the frequency range associated with one ion whistler, but there can be more than one.

In practical cases it is found that, at exact crossover, $\mathrm{Im}(\mathfrak{G})$ is positive. For upgoing waves in the northern hemisphere $\cos\Theta$ is positive, whence (4.24) shows that

$$\rho_O = \rho_E = -1. \tag{13.63}$$

The condition $\mathrm{Re}(\mathfrak{G}) = 0$ alone is called simply 'crossover'. At any given height the frequency for which it occurs is called the 'crossover frequency' for that height, and is denoted by f_{cr}. Then if $\Theta < \Theta_t$, (4.24) shows that the two values of ρ are real, so that the wave polarisations are linear.

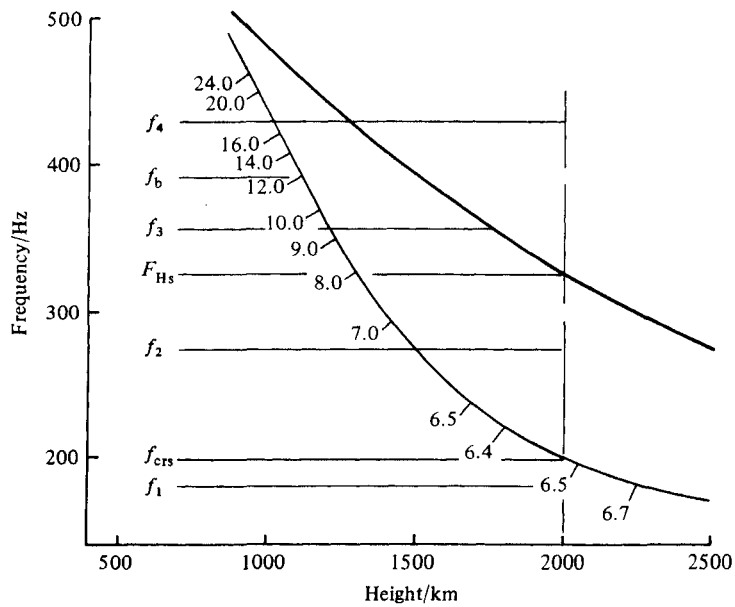
To continue this explanation we now need a model of the upper ionosphere for illustration. The model used here is based on suggestions by Gurnett, Shawhan *et al.* (1965) with some modifications used by D. Jones (1970). It applies for heights above 500 km and here the electron concentration decreases monotonically with increasing height. Three positive ion species are present, namely protons, singly charged helium ions and singly charged atomic oxygen ions, and their concentrations depend on height. The collision frequencies also depend on height. The full details are complicated and not needed here, but are given by D. Jones (1970).

The earth's magnetic field and thence the ion gyro-frequencies are proportional to the inverse cube of the distance from the earth's centre. In fig. 13.9 the upper and lower curves show how the proton gyro-frequency F_H and the crossover frequency f_{cr}, respectively depend on height. The numbers by the lower curve are the values in degrees of Θ_t as found from (13.62).

Suppose, now, that the satellite is in the northern hemisphere at a height of 2000 km as shown by the vertical line, and let the angle Θ between the vector $\boldsymbol{Y}$ and the vertical be 12°. The waves coming up from the earth are electron whistler waves, polarised with a left-handed sense. The other type of wave polarised with a right-handed sense is heavily attenuated at low heights and does not penetrate the ionosphere. The crossover frequency at the satellite is marked f_{crs} in fig. 13.9. Consider a component frequency f_1 less than f_{crs}. In travelling up to the satellite the wave does not encounter a crossover level and so it remains an electron whistler with left-handed polarisation. This accounts for the part of the trace at the lowest frequencies in fig. 13.7(d). Next consider a frequency $f_2 > f_{\text{crs}}$. The wave now encounters a crossover level below the satellite, at about 1500 km in fig. 13.9. Here Θ_t is about 6.9° and Θ is greater than this. Near the crossover level, therefore, the left-handed elliptical polarisation has changed to linear, and at greater heights it becomes elliptical with a right-handed sense; see fig. 13.11 for more details of this. When it arrives at the satellite it is a proton whistler, and appears on the part of the trace marked 'proton' in fig. 13.7(d).

Now consider a frequency f_3 greater than the proton gyro-frequency F_{Hs} at the satellite. This wave encounters a crossover level below the satellite, at about 1200 km in fig. 13.9. Above this it becomes a proton whistler. But before it reaches

Fig. 13.9. Shows how the proton gyro-frequency, upper curve, and the crossover frequency, lower curve depend on height in the upper ionosphere for a typical model. The numbers by the lower curve are the values in degrees of the transition angle Θ_t (13.62).



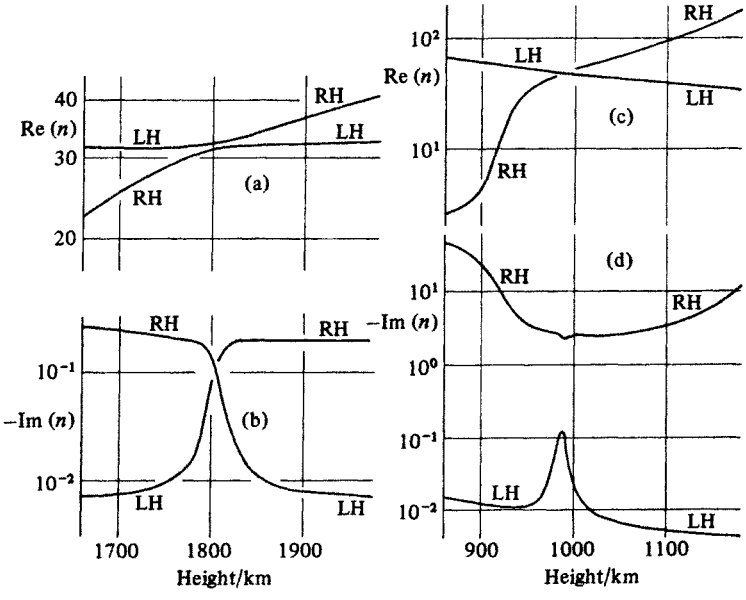
the satellite it attains a level, about 1750 km, where its frequency is equal to the proton gyro-frequency. At greater heights than this $\mathrm{Re}(n^2)$ is negative and the wave is almost evanescent. Hence it is not observed at the satellite. Fig. 13.7(d) shows that, for a frequency range slightly greater than F_{Hs}, the trace is absent.

Finally consider a frequency such as f_4, fig. 13.9. The wave now encounters a crossover level at about 1000 km height but here $\Theta < \Theta_t$. Thus the wave remains an electron whistler and the sense of its polarisation does not reverse. It travels on up and is observed at the satellite. This accounts for the part of the trace marked 'electron' in fig. 13.7(d). There is a 'boundary frequency' marked f_b in figs. 13.7(d), 13.9, that separates the two regions where Θ is greater or less than Θ_t. It is the frequency for which the 'exact crossover' condition (13.49), (13.50) is satisfied.

For the other values of Θ and for other compositions of the topside ionosphere the spectrograms of ion cyclotron whistlers can have different forms from that in fig. 13.7(d). For examples see the papers by D. Jones (1970, 1969), where curves similar to fig. 13.9 for helium whistlers are also given.

The dependence of $\mathrm{Re}(n)$ and $-\mathrm{Im}(n)$ on height near the crossover level for two typical frequencies f is shown in fig. 13.10. For figs. 13.10 (a and b), $f = 220$ Hz and

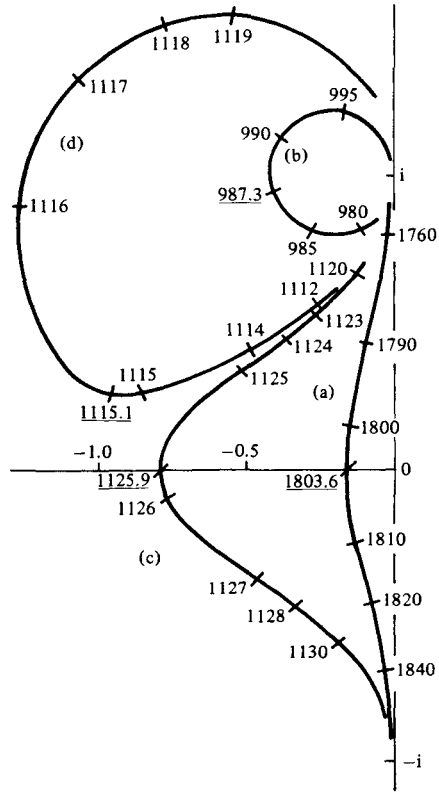
Fig. 13.10. Height dependence of $\mathrm{Re}(n)$ and $-\mathrm{Im}(n)$ near crossover. The labels LH, RH show whether the sense of the wave polarisation is left handed (electron whistler) or right handed (proton whistler) respectively. (a) and (b) are for a frequency $f = 220$ Hz so that $f_{\mathrm{crs}} < f < f_b$. The upgoing electron whistler changes to a proton whistler. (c) and (d) are for $f = 440$ Hz so that $f > f_b$. At the crossover level $\Theta < \Theta_t$ and the upgoing electron whistler remains an electron whistler above the crossover level.



the crossover level is at 1800 km. Here $\Theta_{cr} < \Theta$. The two curves for $-\mathrm{Im}(n)$ cross at this level, and the curves for $\mathrm{Re}(n)$ do not cross. The electron whistler wave, polarised with a left-handed sense, that comes up from lower levels, goes over continuously as the height increases to a proton whistler wave polarised with a right-handed sense. For figs. 13.10(c and d), $f = 440$ Hz and the crossover level is at 987 km. Here $\Theta_{cr} > \Theta$. The two curves for $\mathrm{Re}(n)$ cross, and the curves for $-\mathrm{Im}(n)$ do not cross. The upcoming electron whistler wave remains an electron whistler after it has passed to above the crossover level. It is instructive to compare fig. 13.10 with the sequence figs. 4.7–4.11.

Fig. 13.11 is a diagram of the complex ρ plane and shows how the polarisation ρ of the upcoming electron whistler wave depends on height. For a frequency 440 Hz, curve (b), ρ remains near $+\mathrm{i}$. For a frequency 220 Hz, curve (a), ρ changes

Fig. 13.11. The complex ρ plane. The curves show how the wave polarisation ρ depends on height for an upgoing wave that starts as an electron whistler. The numbers by the curves are the heights in km and the underlined number is the crossover height. Curve (a) is for 220 Hz and curve (b) for 440 Hz; see caption of fig. 13.10. Curve (c) is for frequency 390 Hz and curve (d) for 394 Hz. They are near to, and on opposite sides of the boundary frequency 391.8 Hz.

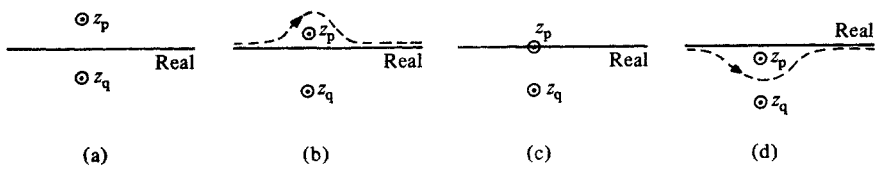


continuously from near $+\mathrm{i}$ through a negative real value at the crossover level to near $-\mathrm{i}$ at greater heights. Curves are also shown for frequencies 390 and 394 Hz, that is just less than and just greater than the boundary frequency $f_b = 391.8$ Hz, for which the exact crossover condition (13.61), (13.62) is satisfied. It is seen that in both cases the value of ρ is near -1 at the crossover level. For 390 Hz $< f_b$, curve (c), it goes over to near $-\mathrm{i}$ and the wave becomes a proton whistler. For 394 Hz $> f_b$, curve (d), it returns to near $+\mathrm{i}$ and the wave remains an electron whistler.

The above description has been given entirely in terms of ray theory, which is the subject of this chapter. Exact crossover occurs for one particular frequency f_b when the crossover conditions (13.61), (13.62) are both true at some point on the real z axis. It has been assumed that the approximations of ray theory hold at all points on the real z axis. This, however, is not strictly true. When $f \neq f_b$ the conditions (13.61) (13.62) hold at points z_p, z_q in the complex z plane, known as 'coupling points'. When f is changed, z_p and z_q move, and when $f = f_b$ one of them is on the real z axis, fig. 13.12(c). The approximations of ray theory break down near a coupling point; this is studied in § 16.3. The two refractive indices are nearly equal there, and we say that the two waves are strongly coupled. One wave as it travels gives rise to some of the other. Thus the transition from an electron whistler to a proton whistler does not occur abruptly for frequencies near to f_b.

Consider again the case where collisions are neglected, as in fig. 13.8. Then when z is real, $\varepsilon_1, \varepsilon_2, \varepsilon_3$ and thence $\mathfrak{G}$ are all real. If $\Theta \neq 0$, (13.60), (13.61) cannot be satisfied for any real z, but they are satisfied for two complex values $z_p = z_r + \mathrm{i}z_i$, and $z_q = z_r - \mathrm{i}z_i$, at equal and opposite distances from the real z axis, fig. 13.12(a). These are the two coupling points. At each of them the two upgoing waves are strongly coupled. The real z axis runs between the two coupling points. On it the two refractive indices are real and continuous. For one of them the wave at low heights is an electron whistler and at great heights it has changed its name to proton whistler. The other refractive index similarly is for a proton whistler at low heights and an electron whistler at great heights.

Fig. 13.12. The complex z plane showing the positions of the coupling points z_p, z_q, for ion cyclotron whistlers (not to scale). Collisions are neglected in (a) but not in (b), (c), (d). (b) is typical of low latitudes where the angle Θ between the earth's magnetic field and the vertical is large. (c) is for exact crossover. (d) is typical of high latitudes where Θ is small and ion whistlers can be produced only by mode conversion. The broken curves in (b) and (d) are possible paths for use with the phase integral method.



If, now, Θ is made smaller, z_i decreases. The coupling points move closer together, and when $\Theta = 0$ they coalesce on the real z axis. There is then a coalescence C2, § 17.4, and this example is used in § 17.5 for illustration.

In practice collisions cannot be neglected so $\mathfrak{G}$ is complex when z is real. Equations (13.60), (13.61) still give the position of the coupling points z_p, z_q but they are no longer at equal and opposite distances from the real z axis. For the model used earlier in this section it can be shown that they are both displaced in the direction of negative imaginary z. If Θ is small enough, however, they are still on opposite sides of the real z axis, fig. 13.12(b). The two refractive indices now have imaginary parts. Their height dependence is as shown in fig. 13.10(a, b).

For a given Θ there can be some frequency $f = f_b$ for which one of the coupling points is on the real z axis. This is the condition of exact crossover, fig. 13.12(c). It is also called 'critical coupling', § 16.6. For larger Θ both coupling points are on the negative imaginary side of the real z axis, fig. 13.12(d). The coupling points are branch points of the two refractive indices $n(z)$. Now the real z axis runs on the positive imaginary side of the branch point z_p so that on it an electron whistler at low heights remains an electron whistler at great heights, and similarly for a proton whistler. The height dependence of the two refractive indices is now as shown in fig. 13.10(c, d).

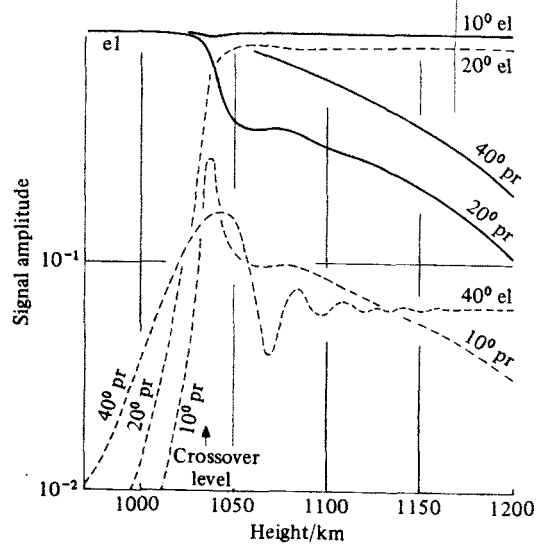
In this case, the mode conversion from an electron whistler to a proton whistler can be studied by using, instead of the real z axis, a path in the z plane that runs on the negative imaginary side of z_p, as sketched in fig. 13.12(d). It may be possible to find such a path that does not go too near to z_p so that the W.K.B. solutions are good approximations at all points on it. Then ray theory can still be used by following the ray to complex values of z. The field amplitude where z is large and real is the amplitude of the wave produced by mode coupling. This is an example of the phase integral method for coupling, § 16.7, and may be regarded as a special case of the use of a complex ray, §§ 14.9–14.12.

Similarly, when z_p is on the positive imaginary side of the real z axis, it may be possible to find a good path in the z plane that runs on the positive imaginary side of z_p, as sketched in fig. 13.12(b). This could then be used to find the amplitude at great heights of the electron whistler wave produced by mode coupling from the incident wave that is an electron whistler at low heights.

To study the dependence of the received whistler signal amplitudes on frequency f and angle Θ it is therefore necessary to extend the use of ray theory by using paths in the complex z plane; see §§ 14.9–14.12, 16.7. An alternative method is to use full wave solutions of the governing differential equations, ch. 18. Figs. 13.13, 13.14 show some results calculated in this way for a model of the upper ionosphere similar to that used in figs. 13.9–13.11. The equations (7.80), with (7.82), were integrated through a downward range of height from 1200 km to 950 km for vertically incident waves and

using the methods of §§ 18.6, 18.7. For each f and Θ two integrations are needed because there are two different possible upgoing waves at the top. The two solutions are combined so that at the bottom there is an upgoing electron whistler wave of unit amplitude. Fig. 13.13 shows how the amplitudes of the two upgoing waves depend on height z for various values of Θ. The signal amplitudes here are $|f_i|$ where the two f_is are the elements of $\mathbf{f}$, (6.53), for the two upgoing waves. It was found that the amplitudes of the downgoing waves were negligible at all heights so there is no appreciable reflection in this part of the ionosphere. The results of figs. 13.13, 13.14 all apply for a frequency of 422.5 Hz, which is less than the proton gyro-frequency F_H for the whole of the height range in the figures. Fig. 13.9 shows that the transition value Θ_t is near 16°, and the crossover height is just above 1000 km. The continuous curves of figs. 13.13, 13.14 apply for the wave that is propagated without mode conversion. The broken curves are the waves produced by mode conversion; these would be absent if the approximations of the simple ray theory of this chapter were used. The label el indicates an electron whistler whose polarisation has a left-handed sense, and pr indicates a proton whistler with a right-handed sense. It can be seen that for $\Theta = 10°$, that is less than Θ_t, the amplitude of the electron whistler is close to unity at all heights, but the growth of the proton whistler, by mode conversion near the coupling level, can be seen. For $\Theta = 20°$, that is greater than Θ_t, the continuous curve represents an electron whistler at low heights and a proton whistler at great heights. Its amplitude decreases with increasing height because of

Fig. 13.13. Dependence of signal amplitude on height for electron and proton whistlers for a frequency 422.5 Hz and various values of Θ as marked by the curves. The model used is slightly different from that of figs. 13.9–13.11. The transition value of Θ is $\Theta_t \approx 15°$ and the crossover height is 1037 km.

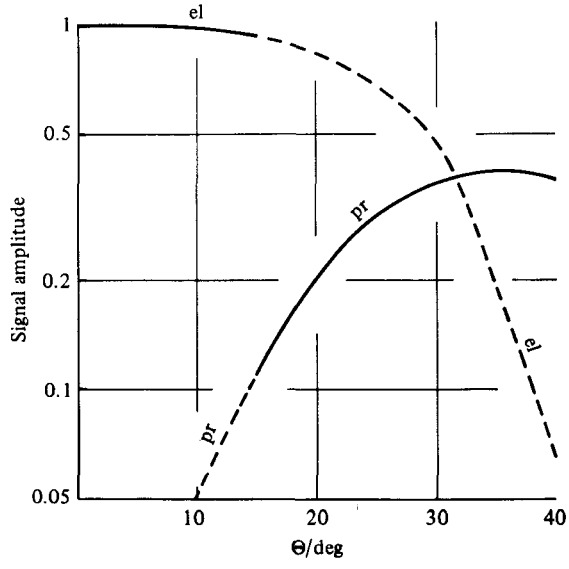


attenuation. The other wave is a proton whistler at low heights and an electron whistler at great heights. The broken curve shows its growth, and it also shows oscillations. These are even more marked for $\Theta = 40°$. They occur because of interference between contributions to the second wave from different levels; their relative phases depend on the difference of the phase velocities of the two upgoing waves.

Fig. 13.14 shows how the amplitudes of the two waves depend on Θ at a height of 1150 km, that is above the crossover level. For small Θ the wave here is mainly an electron whistler and the proton whistler has very small amplitude. This is what is observed at high latitudes. For large Θ, $\approx 40°$, the predominant wave is a proton whistler but the electron whistler is also present. For smaller Θ, although still $> \Theta_t$, the electron whistler formed by mode conversion may have the larger amplitude because the proton whistler is more heavily attenuated. Thus there are ranges of Θ and of frequency f where both electron and proton whistlers can be present with comparable amplitudes. Experimental confirmation of this was given by Rodriguez and Gurnett (1971).

Several workers have used full wave solutions to study ion cyclotron whistlers. Thus Wang (1971) investigated the intermode coupling for vertical incidence when collisions are neglected. This subject is of some theoretical interest; see § 17.5. Fijalkow, Altman and Cory (1973) allowed for collisions and for oblique incidence, and included a study of the propagation of the electron whistler from the ground up

Fig. 13.14. Dependence on Θ of the amplitudes of the two waves of fig. 13.13, for a height of 1150 km. Other details as in fig. 13.13.



into the topside ionosphere. Their conclusions seem to be generally in agreement with those given here. Arantes and Scarabucci (1975) allowed for collisions and studied vertical incidence. They discussed the physical processes associated with changes of wave polarisation in the region where the coupling is strong. Their computing method was very similar to that used here for figs. 13.13, 13.14.

In this section it has been assumed that the wave normals are vertical. The receiver was assumed to be high in the topside ionosphere somewhere above the source but its exact position was not considered. In a fuller treatment allowance must be made for the lateral deviation of the signals, §§ 10.12–10.14. This means that for a ray to reach a given receiver it must leave the source with the wave normal not vertical. Because collisions are important in this theory, the refractive indices are complex and the rays are refracted through complex angles. Thus the wave normals at the source have complex directions that are different for the electron whistler and the ion whistler. A theory that makes allowance for this was given by Terry (1971, 1978).

13.10. Absorption, non-deviative and deviative

The effect of collisions in a plasma is to cause attenuation of radio waves. This is because some of the electromagnetic energy is absorbed by the medium and causes heating, mainly of the electrons. The present section is concerned with attenuation. Some consequences of heating are studied in §§ 13.11–13.13. The term 'absorption' is commonly used to mean both these aspects.

The effect of electron collisions on the phase height $h(f)$ and the equivalent height $h'(f)$ for wave packets vertically incident on an anisotropic ionosphere has been discussed in § 13.3. The present section also is concerned with vertically incident wave packets. The subject is important in the theory of ionospheric probing using the ionosonde technique. Oblique incidence is discussed in §§ 10.15, 10.16, 14.7, and in problem 10.6.

For an isotropic ionosphere, the effect of electron collisions on the attenuation of a wave packet was discussed in § 12.3 where the important formula (12.14) was derived. There is an analogous relation, (13.67) below, for the anisotropic case. For a fuller treatment of this subject see Dyson and Bennett (1979). The formulae (12.3), (12.9) can still be used. It must be remembered that a vertically incident wave packet is deviated laterally, § 10.12, but the group refractive index n' in (12.9) gives the vertical component c/n' of the group velocity. The predominant wave normal is everywhere vertical. The attenuation in nepers is $-\ln|R| = -2k \operatorname{Im}\{h(f)\}$, and is given by (12.10), also for the anisotropic case. It is zero if $Z = 0$. It is now instructive to study the effect of a small value of Z, as was done in § 12.3. Then (12.11) can still be used and $(\partial n/\partial Z)_{Z=0}$ must now be found. For $Z = 0$, the refractive index n is a function of X and Y. In the form (4.65) of the dispersion relation let the coefficients A, B, C be divided by U^3. Then each of them is a function of X/U, Y/U. This shows

that n is a function of these two variables. It then follows that

$$\frac{\partial n}{\partial Z} = \mathrm{i}X\frac{\partial n}{\partial X} + \mathrm{i}Y\frac{\partial n}{\partial Y} \tag{13.64}$$

where all derivatives have their values for $Z = 0$. Then from (12.10), (12.11)

$$\ln|R| = \frac{1}{c}\int_0^{z_0} \nu\left(X\frac{\partial n}{\partial X} + Y\frac{\partial n}{\partial Y}\right)\mathrm{d}z. \tag{13.65}$$

For $Z = 0$, n depends on frequency f only through X and Y. Now (5.65) shows that $n' = \partial(nf)/\partial f$ whence

$$n' - n = f\frac{\partial n}{\partial f} = -2X\frac{\partial n}{\partial X} - Y\frac{\partial n}{\partial Y} \tag{13.66}$$

and if this is put in (13.65) it gives for the attenuation

$$-\ln|R| = \frac{1}{c}\int_0^{z_0} \tfrac{1}{2}\nu\frac{n' - n}{1 + g}\mathrm{d}z \tag{13.67}$$

where

$$g = \frac{-\frac{1}{2}Y\dfrac{\partial n}{\partial Y}}{X\dfrac{\partial n}{\partial X} + Y\dfrac{\partial n}{\partial Y}}. \tag{13.68}$$

Some special cases of this may be noted. For purely longitudinal propagation, $\Theta = 0$, the two values of n^2 are given by (4.84). For the first, (13.68) gives $g = \frac{1}{2}Y$. In the ionosphere this may be assumed to be independent of z. If ν also is independent of z, (13.67) gives

$$-\ln|R| = \frac{1}{2}\frac{\nu}{c}\frac{1}{1 + \frac{1}{2}Y}(h' - h). \tag{13.69}$$

The wave is ordinary where $X < 1$, extraordinary where $X > 1$ (see § 4.16), and is reflected where $X = 1 + Y$. The other value of n^2 in (4.84) gives the same result with the sign of Y reversed. The wave is extraordinary and reflected where $X = 1 - Y$.

For purely transverse propagation, $\Theta = \frac{1}{2}\pi$, the value of n^2 for the ordinary wave is the same as for an isotropic medium and is independent of Y. Then $g = 0$ and if ν is independent of z, (13.67) is the same as (12.14). For the extraordinary wave n^2 is given by (4.81), second formula. Then it can be shown that

$$\frac{1}{1 + g} = \frac{Y^2 + (1 - X)^2}{XY^2 + (1 - X)^2}. \tag{13.70}$$

This depends on height z and therefore cannot be taken outside the integral in (13.67).

For the more general case Bennett and Dyson (1978), and Dyson and Bennett (1979) give a very full study of the function g, including curves showing how it depends on X for various Θ.

The electron collision frequency ν decreases with increasing height; see fig. 1.2. It often happens that a vertically incident radio pulse is reflected at a height where ν is small, but travels up and down through a large thickness of ionosphere lower down where ν is larger, but $N(z)$ and X are small. Then most of the attenuation occurs where X is small. For example the pulse may be reflected in the F-region but most of its attenuation occurs in the D- and E-regions. In this type of situation the attenuation occurs where n is close to unity. Here an oblique ray path would be almost straight. The lateral deviation of a vertical ray here is very small. The absorption is then said to be 'non-deviative'. The topic is of some interest for high frequencies that are not near the penetration frequency of any layer. Then Z is negligible near the reflection level and small in the absorbing region. In these conditions (13.65) can be used where the derivatives have their values for $Z = 0$ and small X. For this approximation (4.113) applies and (13.65) gives

$$-\ln|R| = \frac{1}{c}\int_0^{z_0} \nu X(\tfrac{1}{2} \mp Y\cos\Theta)\,\mathrm{d}z \tag{13.71}$$

where the upper sign is for the ordinary wave. The condition $|\cos\Theta/\sin^2\Theta| \gg \frac{1}{2}Y$ for (4.112) must be imposed here, so that (13.71) cannot be used for near transverse propagation. In (4.113) it is assumed that Z and Y^2 are negligible. More accurate formulae can be derived if this assumption is not made. For examples see Rawer and Suchy (1967, § 9).

The approximation that X is small, used in (13.71), cannot be made for a pulse that is reflected near the maximum of an ionospheric layer. Here $\mathrm{d}N/\mathrm{d}z$ is small near the reflection level. There is a range of z with nearly constant N where the refractive index is near to zero and the group velocity is very small. The wave packet then spends a long time near the reflection level and most of its attenuation occurs there. This is where the ray is being deviated, and the absorption here is called 'deviative absorption'. For this case also there are approximate formulae for $\mathrm{Im}(n)$ near the reflecting level. For details see Rawer and Suchy (1967, § 9) where further references are given.

For discussions of many aspects of absorption, both experimental and theoretical, see *Journal of Atmospheric and Terrestrial Physics* (1962, special volume on absorption).

13.11. Wave interaction 1. General description

The phenomenon of wave interaction, also called cross modulation or the Luxembourg effect, was first observed by Tellegen (1933). He used a receiver in Eindhoven and was receiving a broadcast signal from Beromünster (Switzerland, 652 kHz). It was observed that the radio program from Luxembourg (252 kHz) was heard at the same time. It was established that the modulation of the Luxembourg signal, called the 'disturbing wave', was impressed as amplitude modulation of the

Beromünster signal, called the 'wanted wave', and that the process was occurring within the ionosphere.

This subject has aroused great interest and there are numerous published papers dealing both with experiment and with the theory. It is of practical importance for two main reasons. First, it is a source of interference with broadcast transmissions, which is becoming more serious as more and more transmitters are being built that send disturbing energy into the D- and E-regions. This aspect of the subject has been surveyed by Maslin (1976e). Second, it is a useful means of investigating the D- and E-regions and particularly as a method of measuring the effective collision frequency. It was used in this way most successfully by Fejer (1955) with pulse modulated signals for both the wanted and the disturbing waves. The timing of the pulses could be adjusted so as to choose the height where the interaction occurred. Fejer's method has been used with various elaborations by several groups of later workers (Barrington and Thrane, 1962; Landmark and Lied, 1962; Weisbrod, Ferraro and Lee, 1964).

Most of the earlier theories of wave interaction ignored the earth's magnetic field so that the ionosphere is assumed to be isotropic. An excellent summary, with a detailed discussion of the energy balance of the electrons, is given by Huxley and Ratcliffe (1949). A theory that allows for the earth's magnetic field has been given by Maslin (1974, 1975a, b, 1976c, d, e).

The explanation, in outline, is as follows. The disturbing wave is attenuated by the ionospheric plasma so that some of its energy is converted to heat. It is the electrons that are heated first, so that their average random thermal speed is increased. When the disturbing wave is amplitude modulated the heating rate is greatest when the modulation amplitude is greatest, and varies in step with the modulation envelope of the wave. The heated electrons can lose energy with a time constant of about 1 ms, so that for modulation frequencies less than about 1 kHz the modulation causes the electron temperature to vary. There is a time lag as explained later; see (13.76). Now the collision frequency of the electrons depends on their average speed and is increased when the electron temperature is increased. When a wanted wave goes through the same plasma it is attenuated at a rate that depends on the collision frequency so that its attenuation varies in step with the modulated electron temperature. Thus the wanted wave, through this attenuation, acquires a modulation that originates from the modulation of the disturbing wave.

The changed collision frequency also means that the attenuation of the disturbing wave itself is changed. Usually it increases when the collision frequency increases so that the modulation amplitude is reduced and the disturbing wave is said to be 'self-demodulated'. Observations of this have been made (Aitchison, 1957; Hibberd, 1957; King, 1959). It can happen that if the average value of ν is large enough, an increase of

ν can cause a decrease of attenuation; see, for example, problem 4.6. Then the modulation of the disturbing wave is enhanced.

Wave interaction is, in one sense, a non-linear process, which would place it outside the subject range of this book. But, for radio waves in the ionosphere from communication or broadcasting transmitters, the power is small enough for attenuation to be treated as a linear process, and for the heating to be found by using only the linear form, ch. 3, of the constitutive relations. Attenuation, also called absorption, has been discussed in §§ 10.15, 10.16, 12.3, 12.11, 13.3, 13.10. A brief discussion of the heating process is given in the following section. A full account of wave interaction is included in a text book by Gurevich (1978) on non-linear phenomena.

13.12. Wave interaction 2. Outline of theory

The version of the theory in this section is greatly simplified and gives only the barest essentials. There are many refinements for which the original papers must be consulted.

In any heating process in the ionosphere it is the electrons that acquire energy first, from the heating wave. They already have random thermal speeds of the order of $2 \times 10^5\,\mathrm{m\,s^{-1}}$ (energy of order 0.1 eV). Superimposed on this is an ordered motion caused by the electric field of the heating wave. The ordered speeds are typically about $3 \times 10^3\,\mathrm{m\,s^{-1}}$, that is very much less than the random speed. This figure applies at 100 km from a 1 MW transmitter of frequency 1 MHz. The ordered motion is an elliptical orbit and is periodic with the period of the heating wave. The resultant orbits, comprising both ordered and disordered motion, are complicated because of the earth's magnetic field, and because the wave is in general elliptically polarised. They have been discussed by Ratcliffe (1959, ch. 14).

The electrons make collisions which, in the ionosphere, are predominantly with neutral particles and may be assumed to be instantaneous. Because the particles that they collide with are massive, the change of the electron's speed in one collision is very small. It is the direction of motion that is changed. Immediately after a collision, the electron's new velocity is the resultant of a new thermal velocity and the same ordered velocity as immediately before the collision. It is easy to show that the average resultant speed is greater than the thermal speed alone. Thus, on the average, a collision occurs when the resultant speed is greater than the thermal speed. The result is that, on the average, the thermal energy after a collision is greater than it was before it. Through collisions, the ordered energy imparted by the heating wave is converted to disordered electron energy, that is to heat.

The collision frequency of an electron in the ionosphere is of order $3 \times 10^3\,\mathrm{s^{-1}}$ at 300 km to $10^6\,\mathrm{s^{-1}}$ at 80 km. Thus, at the lower levels, an electron makes an average

of one or several collisions in one cycle of the heating wave. Although the electron's ordered energy is small, it can be continuously converted to heat by collisions. Let H be the rate at which heat is imparted to the electrons in unit volume by the disturbing wave. Methods of calculating it are discussed below.

In the electron heating process just described it was assumed that an electron does not change its energy when it makes a collision. In fact it does transfer a small fraction of its energy to the heavy particle with which it collides. Thus electrons heated by a disturbing wave can cool down again by transfer of their excess energy to other particles of the plasma. For the electrons in a small volume let Q be the average energy per electron at any fixed time, and let Q_0 be the value that Q would have if the electrons were in thermal equilibrium with the other particles. For thermal equilibrium the electrons would have a Maxwellian velocity distribution with temperature $T_0 = \frac{2}{3}Q_0/K$ where K is Boltzmann's constant. For disturbed electrons the distribution is not in general Maxwellian but we can speak of an effective electron temperature $T = \frac{2}{3}Q/K$. The disturbed electrons tend to return to thermal equilibrium so that on average an electron loses energy $G(Q - Q_0)$ in each collision, where G is a constant. The time dependence of Q is then given by

$$\partial Q/\partial t = -Gv_{\mathrm{av}}(Q - Q_0) + H/N \tag{13.72}$$

where v_{av} is the average collision frequency of an electron. The cooling process has a time constant $1/Gv_{\mathrm{av}}$. The constant G can be estimated from a study of the mechanics of a collision; see for example, problem 3.7. It is found, however, that the value from these estimates may be too small. Experimental measurements for air show that G is of order 1.3×10^{-3} (Huxley and Ratcliffe, 1949), so that in the lower ionosphere $1/Gv_{\mathrm{av}}$ is of order 1 ms. If the disturbing wave is amplitude modulated with a frequency less than about 1 kHz, the electron temperature has time to rise and fall as the heating rate H changes. Thus the modulation is impressed on the electron temperature. For greater modulation frequencies, however, the electrons cannot lose much energy within the short modulation cycles. The electron temperature then simply increases until the average energy loss rate $Gv_{\mathrm{av}}(Q - Q_0)$ is equal to the average rate of heating by the disturbing wave.

It was shown in § 2.11 that when a harmonic wave system is present in a plasma with electric intensity $\boldsymbol{E}$ at any point, the heating rate at that point is, from (2.52)

$$H = -\tfrac{1}{2}\omega \operatorname{Im}(\boldsymbol{E}^* \cdot \boldsymbol{D}) = -\tfrac{1}{2}\omega\varepsilon_0 \operatorname{Im}(\boldsymbol{E}^* \cdot \boldsymbol{\varepsilon} \boldsymbol{E}). \tag{13.73}$$

Now let $\boldsymbol{E}$ for the disturbing wave have a sinusoidal modulation so that

$$\boldsymbol{E} = \boldsymbol{E}_0(1 + M \cos \Omega t). \tag{13.74}$$

Here Ω is the angular modulation frequency and is very much less than the angular wave frequency ω. Then

$$H = H_0(1 + \tfrac{1}{2}M^2 + 2M \cos \Omega t + \tfrac{1}{2}M^2 \cos 2\Omega t), \quad H_0 = -\tfrac{1}{2}\omega_0\varepsilon_0 \operatorname{Im}(\boldsymbol{E}_0^* \cdot \boldsymbol{\varepsilon} \boldsymbol{E}_0). \tag{13.75}$$

This is now inserted in (13.72). The term $2M\cos\Omega t$ in H then contributes to the solution a term

$$Q=\frac{2MH_0}{N}\{(Gv_{av})^2+\Omega^2\}^{-\frac{1}{2}}\cos(\Omega t-\Phi),\quad \tan\Phi=\Omega/Gv_{av}. \tag{13.76}$$

This shows that the modulation of Q includes a part with angular frequency Ω having a phase lag Φ. There is also a term with angular frequency 2Ω that is not present in the original modulation (13.74). The phase shift Φ is transferred to the modulation imposed on the wanted wave. By measuring Φ for various modulation frequencies Ω, Ratcliffe and Shaw (1948) could locate the region of the ionosphere where the interaction was occurring.

Consider now the method of Fejer (1955) in which the disturbing and wanted waves are radio pulses. In the original version of the method these pulses travel vertically as in the ionosonde technique, § 1.7. It is arranged that the disturbing pulse is upgoing and the wanted pulse is downgoing after reflection at a higher level. The disturbing pulse is transmitted only for alternate wanted pulses so that the received amplitudes of the wanted pulse with and without disturbance can be compared. When the disturbing pulse is present the two pulses first meet at a level $z=z_0$ chosen by the timing of the pulses. The wanted pulse then travels on down through the lower levels recently heated by the disturbing pulse. Let the duration τ of the disturbing pulse be much less than $1/Gv_{av}$ and for any given height $z_1<z_0$ let it apply heating at the rate $H(z_1)$, assumed constant during the pulse. Then (13.72) shows that, after the pulse has passed this height, Q has acquired an increment $H(z_1)\tau/N$ that subsequently decays thus

$$Q=\frac{H(z_1)\tau}{N}\exp(-Gv_{av}t). \tag{13.77}$$

From this the increase of effective temperature and thence the increment $\Delta\nu_{eff}$ of the effective collision frequency for the electrons are found, at the time t when the wanted pulse passes through the level z_1. This depends on how $\nu(v)$ depends on electron velocity v, and on the velocity distribution function of the disturbed electrons. Then by an integration through a height range below z_0 the extra attenuation of the wanted wave, caused by $\Delta\nu_{eff}$ is found and expressed in terms of G, ν_{av} and N. By measuring this extra attenuation estimates can be made of these three quantities, and their height dependence can be studied.

13.13. Wave interaction 3. Kinetic theory

In much published work on the theory of wave interaction it was assumed that an electron's mean free path λ_e is independent of its speed v so that $\nu(v)\approx v/\lambda_e$ and ν is proportional to v. But laboratory experiments have suggested that for electrons of low energy the assumption $\nu\propto v^2$ is better. This topic has been discussed in § 3.12. For slow electrons the correct form for $\nu(v)$ is not known with certainty. But the

occurrence of wave interaction shows that $\nu(v)$ must be some increasing function of v to give an increase of the average ν when heating occurs.

When the electrons are in thermal equilibrium it is believed that, at least in the lower ionosphere, their velocity distribution function is Maxwellian. The faster electrons are making more collisions so that, when a heating wave is applied, the fast electrons are heated more than the slow electrons. The velocity distribution is therefore disturbed and no longer Maxwellian. It then changes with time and in the cooling process, when the heating is removed, it must relax back to Maxwellian. In the disturbed plasma the extra attenuation of a wanted wave occurs while the distribution function is not Maxwellian, and a full theory must allow for this. To pursue these problems the Boltzmann distribution function for electrons is used. It satisfies the Boltzmann equation. The study of this is in the realm of plasma kinetic theory. It is a large subject with a very extensive literature. A brief summary of the problem for wave interaction was given by Budden (1983b). For recent work in this field and further references see Garrett (1982a, b, 1983, 1985).

PROBLEMS 13

13.1. If the earth's magnetic field is vertical and electron collisions are neglected, then for waves with vertical wave normals in the ionosphere, the two group refractive indices n' are given (see problem 5.10) by $\{1 + XY/2(1-Y)^2\}\{1 - X/(1-Y)\}^{-\frac{1}{2}}$ and the same expression with the sign of Y reversed. Let $h'(f)$ be the contribution to the equivalent height of reflection from the part of the ray path within the ionosphere. Express it as an integral, for a frequency slightly greater than the electron gyrofrequency f_H, and show that

$$\lim_{\varepsilon \to 0} h'\{f_H(1+\varepsilon)\} = \tfrac{2}{3}/(\mathrm{d}X/\mathrm{d}z)_0$$

where $(\mathrm{d}X/\mathrm{d}z)_0$ is the value of $\mathrm{d}X/\mathrm{d}z$ at the base of the ionosphere. (A useful transformation is $X = \varepsilon(1-t^2)$. See Millington (1938a). The author is indebted to Dr. D.H. Shinn of the G.E.C.-Marconi Research Laboratories (formerly Marconi's Wireless Telegraph Co., Ltd.) for bringing this and some similar problems to his attention.)

13.2. Some substances show natural optical activity. A well known example is a solution of an organic compound whose molecule occurs in two enantiomorphic forms that are mirror images of each other and not identical. For an isotropic medium, containing only one of the two types of molecule, the constitutive relation is

$$\boldsymbol{D} = \varepsilon_0\{\varepsilon \boldsymbol{E} + (\eta/k)\,\mathrm{curl}\,\boldsymbol{E}\}.$$

Note that this displays spatial dispersion (§§ 3.1, 5.5) because the operator curl contains spatial derivatives. Show that when a linearly polarised wave travels in a

homogeneous medium of this kind, its plane of polarisation is rotated through an angle $\frac{1}{2}k\eta$ in unit length of its path. Compare and contrast this effect with Faraday rotation in a magnetoplasma. (See Landau and Lifshitz, 1960, p. 337 ff.).

13.3. Let ρ_O and ρ_E be the two values of the wave polarisation ρ for waves of extremely low frequency with vertical wave normals in the ionosphere. The effect of collisions and of heavy positive ions is allowed for. Show that in the complex ρ plane, at the crossover frequency (not exact crossover) the two points that represent ρ_O, ρ_E lie either both on the real axis, or both on a circle of unit radius with centre at the origin.