1.9 Simple theory of electromagnetic waves and biaxial birefringence: From Lorentz force to conical refraction in 37 pages

Gavin R. Putland
Melbourne, Australia
Version 2017-05-19.

Abstract

A time-variation in magnetic flux density **B** may occur because the field *changes* and/or because the field *moves* relative to the observation point. Faraday's law for a fixed circuit makes no distinction between these causes. But the latter cause is isolated by the magnetic term in the Lorentz force law, which, in a reference frame fixed with respect to the particle, implies that a field **B** moving at velocity **r** induces an electric field $\mathbf{E} = -\mathbf{r} \times \mathbf{B}$. In the case of a traveling electromagnetic wave, **r** is the *ray* velocity (hence the symbol).

Similarly, a time-variation in the electric displacement field **D** may occur because the field changes and/or because the field moves. The Maxwell-Ampère law makes no distinction between these causes. But, by analogy with the Lorentz force law, the latter cause can be isolated by saying that a **D** field moving at velocity **r** induces a magnetizing field $\mathbf{H} = \mathbf{r} \times \mathbf{D}$.

The two "moving field" laws, combined with the relations between **D** and **E** and between **B** and **H**, yield an unusually simple theory of electromagnetic waves, including a derivation of Fresnel's equation for the ray-velocity surface of a non-chiral birefringent crystal. Taking cross-products of the "moving field" laws with the wave-slowness vector, we obtain two more "moving field" equations in terms of wave slowness (generalizing the conventional formulation in terms of the wave *vector*). The last two equations, by analogy with the first two, yield Hamilton's wave-slowness surface. Comparing the results, we can conclude that the ray-velocity and wave-slowness surfaces of a biaxial crystal have curves of contact with tangent planes, and deduce the associated polarizations. Eigenvectors are introduced to show that, in general, the permitted polarizations for a given propagation direction are orthogonal. A coordinate transformation (simpler than Hamilton's) shows that the curves of contact are circles and yields their linear and angular diameters.

Among the footnotes are interpretations of the Poynting vector and the Minkowski momentum density. The text includes introductory material intended to make it comprehensible to high-school graduates.

Let us investigate the conditions under which an electromagnetic (EM) field can be a **simple wavelike field**, which we shall define as a spatial variation moving through the medium with a constant translational velocity—that is, a function of position but not time, as defined in a reference frame moving at a constant translational velocity w.r.t. the medium, causing a time-variation to appear in a frame fixed w.r.t. the medium. We shall define a general *wavelike field* as a field that is at least *approximately* a simple wavelike field. Similarly, we shall define a (simple or general) *electromagnetic wave* as a (simple or general) wavelike electromagnetic field.

If a wavelike field forms a narrow beam, we think of it as approximating a **ray**, regardless of whether the "ray" direction is normal (perpendicular) to any "wavefront" that we might claim to discern within the beam. So, to allow for any difference between the ray direction and the direction normal to the wavefront, the aforesaid "constant translational velocity" must mean the **ray velocity**, which we shall denote by $\mathbf{r}$. The **wave-normal velocity**, which we shall denote by $\mathbf{v}_n$, is the component of $\mathbf{r}$ in the direction normal to the "wavefront".

But what is a "wavefront"? Intuitively and loosely, we may think of a wavefront as a surface on which the wavelike function reaches a peak (a wave "crest") or a minimum (a wave "trough"), or has its maximum derivative w.r.t. time (a "rising edge"), or changes sign. What these notions have in common is that they refer to a variation of the wave function *across* the wavefront, while tacitly assuming that there is little or no such variation *within* the wavefront. So a **wavefront**—if we can identify such a thing—is a *surface within which the spatial variation of the wavelike function is minimized*.

Rectilinear propagation is implicit in the constancy of $\mathbf{r}$: if a wavelike field is refracted, it is not simple, but may be *piecewise* simple (for abrupt refraction at a surface) or *approximately* simple (for continuous refraction). While $\mathbf{r}$ is initially assumed to be **uniform** (independent of location), it is *not* initially assumed to be **isotropic** (independent of direction).

The locus of $\mathbf{r}$ is the **ray-velocity surface**; it is the surface whose "distance" from the origin in any direction is the ray velocity in that direction (where the word "distance" is in quotes because the "surface" is in velocity space, not position space). Similarly, the locus of $\mathbf{v}_n$ is the **wave-velocity surface** (also called the *normal-velocity surface*); it is the surface whose "distance" from the origin in any direction is the wave-normal velocity in that direction.

A strictly *simple* wavelike field cannot accommodate any convergence or divergence of rays, because that would contradict uniformity of the ray velocity, and because the associated convergence or divergence of energy would violate the time-invariance of the field in the reference frame moving with the wave. Nor can it accommodate any convergence or divergence of a wavefront due to

curvature, because that too would contradict the time-invariance of the field in the moving reference frame. But if the convergence or divergence is sufficiently *gradual* compared with the variation of the field across the wavefront, it is compatible with *approximately* simple wavelike behavior. Hence the concept of a "general" wavelike field.

Consider, for example, a wavefront diverging from a point-source. At sufficiently large distances from the source, the field is approximately simple and wavelike. Close to the source, the field may depart significantly from simple wavelike behavior; but at sufficiently large distances, the cumulative effect of that departure (compared with subsequent behavior) becomes negligible. As the expansion of the wavefront represents energy transport, the wavefront recedes from the source at the *ray* velocity—which, at sufficiently large distances, becomes the ray velocity of a simple wavelike field. Hence, if the wavefront expands for a sufficiently long time in a uniform medium, the distance of the wavefront from the source in a given direction becomes near enough to the expansion time multiplied by the ray velocity in that direction. In the case of *unit* expansion time, we have the following result:

1.9.1 Theorem: *In a uniform medium, the ray-velocity surface is the surface reached by a wavefront expanding from the origin in unit time.*

For this reason, William Rowan Hamilton conveniently and evocatively referred to the ray-velocity surface as the **unit-wave** [12, p.142].

As the ray-velocity surface is a wavefront, we can use it to establish the wave-normal direction:

1.9.2 Theorem: *The normal to the ray-velocity surface at point* $\mathbf{r}$ *is the wave-normal direction for a ray in the direction of* $\mathbf{r}$.

Hence the wave-normal velocity $\mathbf{v}_n$, being the component of $\mathbf{r}$ in the wave-normal direction, is given (in magnitude and direction) by the perpendicular from the origin to the plane tangent to the ray-velocity surface at $\mathbf{r}$. Hence the wave-velocity surface, being the locus of $\mathbf{v}_n$, is the **pedal** of the ray-velocity surface—where *pedal* comes from the Latin word for *foot*, because the pedal is the locus of the *foot* of the perpendicular to the tangent plane.

This relation matters because, when a wavefront is refracted or reflected at a surface separating two media, the incident and refracted/reflected portions of the wavefront must meet at a common curve on the surface; as the wavefront propagates, the common curve moves across the surface. This requirement of **wavefront continuity** governs the refraction and reflection of *waves* and, when combined with the relation between the wave-normal direction and the ray direction, determines the laws of refraction and reflection of *rays*.

If $\mathbf{r}$ is not isotropic, the normal to the unit-wave does not generally pass through the origin. Hence, if we could trace an **orthogonal trajectory** (common normal) through all the previous positions of the wavefront expanding from the origin, that trajectory would generally need to be curved. It follows that in a non-isotropic medium, rectilinearity of the rays does *not* generally imply rectilinearity of the wave-normals: when we speak of "rectilinear propagation", we mean rectilinearity of the *rays*. However, at large distances from the origin, the curvature of the wave-normals is slight. And in the extreme case in which the wavefronts are perfectly *planar*, simple translational motion of the wavefronts allows no change in direction, hence no curvature, of the wave-normals.

From the foregoing discussion of waves and rays in general, we can now proceed to the particulars of *electromagnetic* waves.

If there is no applied electric field, the Lorentz force $\mathbf{F}$ on a particle of charge q moving at velocity $\mathbf{v}$ relative to a magnetic field with flux density $\mathbf{B}$ is

$$\mathbf{F} = q\mathbf{v} \times \mathbf{B} \,. \tag{1.9.3}$$

If, instead, the particle is stationary while the field moves with ray velocity $\mathbf{r}$, the velocity of the particle relative to the field becomes $-\mathbf{r}$, so the force becomes

$$\mathbf{F} = -q\mathbf{r} \times \mathbf{B} \,. \tag{1.9.4}$$

Dividing by q, we find that the moving magnetic field induces the *electric* field

$$\mathbf{E} = -\mathbf{r} \times \mathbf{B} \,. \tag{1.9.5}$$

This relation, applied to a fixed circuit (closed curve), yields a familiar law:

1.9.6 Faraday's law for a fixed circuit: *The integral of* $\mathbf{E}$ *around a fixed circuit is minus the integral of* $\dot{\mathbf{B}}$ *through (a surface enclosed by) the circuit.*

The dot over $\mathbf{B}$ indicates differentiation w.r.t. time, and it is understood that the direction *around* the loop is clockwise about the direction *through* the loop.

Of course, a time-variation in $\mathbf{B}$ may be caused not only by the field *moving* relative to the observation point, by also by the field *changing* (in-situ or as it moves). For a fixed circuit, statement 1.9.6 does not distinguish between these causes, and neither does nature! But for a simple wavelike field, we have only the first cause ("field *moving*"), for which the induced $\mathbf{E}$ is given by (1.9.5). Thus equation (1.9.5) is the simple-wavelike-field form of statement 1.9.6.

If there is no conduction current, the electric *displacement* field $\mathbf{D}$ and the magnetizing field $\mathbf{H}$ are related by another familiar law:

1.9.7 Maxwell-Ampère law (with no conduction): *The integral of* $\mathbf{H}$ *around a fixed circuit is the integral of* $\dot{\mathbf{D}}$ *through (a surface enclosed by) the circuit.*

This statement is exactly analogous to 1.9.6 without the minus sign. Like 1.9.6, it does not care whether the time-variation occurs because the field moves or changes or both (although, strangely, this is more often said of Faraday's law than of the Maxwell-Ampère law). So, by the corresponding analogy with equation (1.9.5), the simple-wavelike-field form of 1.9.7 is

$$\mathbf{H} = \mathbf{r} \times \mathbf{D} \,. \tag{1.9.8}$$

Simple electromagnetic (EM) waves, if they exist, must be simultaneous solutions of (1.9.5) and (1.9.8); these equations and all their implications are necessary conditions for the existence of such waves.

As a cross product is normal to both of its factors, equations (1.9.5) and (1.9.8) immediately yield

$$\mathbf{E}, \mathbf{H} \perp \mathbf{r} \tag{1.9.9}$$

$$\mathbf{E} \perp \mathbf{B} \tag{1.9.10}$$

$$\mathbf{H} \perp \mathbf{D} \,. \tag{1.9.11}$$

From (1.9.9) we see that *simple electromagnetic waves are* ***transverse*** *in the sense that* $\mathbf{E}$ *and* $\mathbf{H}$ *are normal to the ray direction*. Two opposite directions are normal to both $\mathbf{E}$ and $\mathbf{H}$. To see which is the ray direction, we can cross-multiply both sides of (1.9.8) on the left by $\mathbf{E}$, obtaining

$$\begin{aligned} \mathbf{E} \times \mathbf{H} &= \mathbf{E} \times (\mathbf{r} \times \mathbf{D}) \\ &= \mathbf{E}.\mathbf{D}\,\mathbf{r} - \mathbf{E}.\mathbf{r}\,\mathbf{D} && \text{(by a standard identity)} \\ &= \mathbf{E}.\mathbf{D}\,\mathbf{r} && \text{(since } \mathbf{E} \perp \mathbf{r}\text{);} \end{aligned} \tag{1.9.12}$$

and we expect $\mathbf{E}.\mathbf{D}$ to be positive, in which case $\mathbf{E} \times \mathbf{H}$ is in the ray direction. Similarly, cross-multiplying (1.9.5) on the right by $\mathbf{H}$ yields

$$\mathbf{E} \times \mathbf{H} = \mathbf{H}.\mathbf{B}\,\mathbf{r}, \tag{1.9.13}$$

and we expect $\mathbf{H}.\mathbf{B}$ to be positive, so that $\mathbf{E} \times \mathbf{H}$ is again in the ray direction.[21]

[21] The vector $\mathbf{E} \times \mathbf{H}$, usually denoted by $\mathbf{S}$, is called the **Poynting vector**. Adding equations (1.9.12) and (1.9.13), we find that $\mathbf{S} = (\frac{1}{2}\mathbf{E}.\mathbf{D} + \frac{1}{2}\mathbf{H}.\mathbf{B})\,\mathbf{r}$, where the first term in parentheses is the electric energy density and the second is the magnetic energy density. So $\mathbf{S}$, being the product of the total energy density (energy per unit volume) and the ray velocity, is the **intensity** (power per unit area).

That raises an obvious suspicion: could the "other" cross-product, $\mathbf{D}\times\mathbf{B}$, give the *wave-normal* direction, so that $\mathbf{D}$ and $\mathbf{B}$ are tangential to the wavefront?

Consider the contrary propositions. If $\mathbf{D}$ has a component normal to the wavefront, this component will vary in time as the wave passes; and this time-variation, according to the Maxwell-Ampère law, will be proportional to the circulation (spatial rotation) of $\mathbf{H}$ within the wavefront. So, to minimize the spatial variation of $\mathbf{H}$ within the wavefront, there must be no component of $\mathbf{D}$ normal to the wavefront; that is, $\mathbf{D}$ must be *tangential* to the wavefront. Similarly, if $\mathbf{B}$ has a component normal to the wavefront, this component will have a time-variation which, according to Faraday's law, will be proportional to the circulation of $\mathbf{E}$ within the wavefront. So, to minimize the spatial variation of $\mathbf{E}$ within the wavefront, there must be no component of $\mathbf{B}$ normal to the wavefront; that is, $\mathbf{B}$ must be *tangential* to the wavefront. So $\mathbf{D}$ and $\mathbf{B}$, being tangential to the wavefront, are perpendicular to the wave-normal velocity $\mathbf{v}_\mathrm{n}$:

$$\mathbf{D},\mathbf{B}\perp\mathbf{v}_\mathrm{n}\,. \tag{1.9.14}$$

Hence $\mathbf{v}_\mathrm{n}$ has the direction of $\pm\,\mathbf{D}\times\mathbf{B}$. To see which sign is applicable, we can dot-multiply (1.9.5) by $\mathbf{D}$, or (1.9.8) by $\mathbf{B}$, and use the properties of the scalar triple product; either way, it becomes apparent that $\mathbf{r}.\mathbf{D}\times\mathbf{B}$ is positive, so that $\mathbf{D}\times\mathbf{B}$, like $\mathbf{v}_\mathrm{n}$, points to the same side of the wavefront as $\mathbf{r}$. So $(+)\,\mathbf{D}\times\mathbf{B}$ is the direction of the wave-normal velocity.[22]

In summary, simple EM waves are transverse waves for which the electric field $\mathbf{E}$ and the magnetizing field $\mathbf{H}$ are transverse w.r.t. the rays, while the electric displacement field $\mathbf{D}$ and the magnetic flux density $\mathbf{B}$ are transverse w.r.t. the wave-normals—where "transverse" means precisely perpendicular.

Part of our problem, then, is to work out how the magnitudes of $\mathbf{r}$ and $\mathbf{v}_\mathrm{n}$ might depend on the directions (**polarizations**) of $\mathbf{E}$ and $\mathbf{H}$ within the plane normal to $\mathbf{r}$, or the directions of $\mathbf{D}$ and $\mathbf{B}$ within the plane normal to $\mathbf{v}_\mathrm{n}$.

Ray reversibility requires one further (weak) assumption. *If* the properties of the medium are such that a sign-change in $\mathbf{E}$ (as a function of time) causes a sign-change in $\mathbf{D}$, we can change the signs of $\mathbf{E}$, $\mathbf{D}$, and $\mathbf{r}$ in (1.9.5) and (1.9.8). Alternatively, *if* a sign-change in $\mathbf{H}$ causes a sign-change in $\mathbf{B}$, we can change the signs of $\mathbf{B}$, $\mathbf{H}$, and $\mathbf{r}$ in the same equations. Either way, we reverse $\mathbf{r}$. We also reverse $\mathbf{D}\times\mathbf{B}$ and hence the direction of $\mathbf{v}_\mathrm{n}$; and because $\mathbf{v}_\mathrm{n}$ is the component of $\mathbf{r}$ in the direction of $\mathbf{D}\times\mathbf{B}$, reversing both $\mathbf{r}$ and that direction means reversing $\mathbf{v}_\mathrm{n}$. And reversing $\mathbf{v}_\mathrm{n}$ preserves the synchronization of

[22] The vector $\mathbf{D}\times\mathbf{B}$, denoted by $\mathbf{g}_\mathrm{M}$, is called the **Minkowski momentum density**. It has the dimensions of momentum per unit volume, and is related to the "force exerted by light on an object within a medium" [1].

wavefronts at boundaries, so that the reversed rays and waves remain consistent with the laws of refraction and reflection.[23]

The "alternative" condition for ray reversibility is implicit in the following assumption, which we retain from now on.

1.9.15 Assumption: $\mathbf{B} = \mu\mathbf{H}$, *where* μ *is a positive constant of the medium.*

The constant μ is called the **permeability**. Indeed, any reasonably transparent medium probably needs to be non-ferromagnetic in order to avoid significant hysteresis losses; and in a non-ferromagnetic medium, $\mathbf{B}$ is very close to $\mu_0\mathbf{H}$, where μ_0 is the **magnetic constant** (the physical constant formerly known as the "magnetic permeability of a vacuum"). So assumption 1.9.15 is almost implicit in the assumption of simple EM waves (which, by definition, propagate without loss), and it is not greatly generalized by writing μ instead of μ_0.

Assumption 1.9.15 also implies that the proportionality between $\mathbf{H}$ and $\mathbf{B}$ is *instantaneous* (that is, $\mathbf{H}$ and $\mathbf{B}$ are *in phase*). Again, this is very nearly the case for a non-ferromagnetic medium.

Combining assumption 1.9.15 with relations (1.9.10) and (1.9.9), and then with relations (1.9.11) and (1.9.14), we have

$$\begin{array}{c} \mathbf{E} \perp \mathbf{H} \perp \mathbf{r} \perp \mathbf{E} \\ \qquad \| \\ \mathbf{D} \perp \mathbf{B} \perp \mathbf{v}_\mathrm{n} \perp \mathbf{D}, \end{array} \tag{1.9.16}$$

where the symbol $\|$ indicates parallelism (*not* antiparallelism in this context). Under the same assumption 1.9.15, the directions of $\mathbf{r}$ and $\mathbf{v}_\mathrm{n}$ become

$$\mathbf{r} \parallel \mathbf{E}\times\mathbf{H} \parallel \mathbf{E}\times\mathbf{B} \tag{1.9.17}$$

$$\mathbf{v}_\mathrm{n} \parallel \mathbf{D}\times\mathbf{H} \parallel \mathbf{D}\times\mathbf{B}\,. \tag{1.9.18}$$

Relations (1.9.16) to (1.9.18) may be summarized as follows.[24]

1.9.19 Theorem: *For simple EM waves in a medium in which* $\mathbf{B} \parallel \mathbf{H}$, *the vectors* $\mathbf{E}$, $\mathbf{H}$, *and* $\mathbf{r}$ *(the ray velocity) form a right-hand orthogonal triad, as do the vectors* $\mathbf{D}$, $\mathbf{B}$, *and* $\mathbf{v}_\mathrm{n}$ *(the wave-normal velocity), so that* $\mathbf{D}$, $\mathbf{E}$, $\mathbf{v}_\mathrm{n}$, *and* $\mathbf{r}$ *are all in the same plane normal to* $\mathbf{B}$ *and* $\mathbf{H}$.

---

[23] Of course, a sufficient condition for ray & wave reversibility is that the wavelike function be governed by a differential equation in which all derivatives w.r.t. time are of even order, so that one solution may be time-reversed to obtain another solution. But we cannot use that argument here, because we have avoided the use of differential equations!

[24] Cf. Fig. 3 in Lunney & Weaire [19], where $\mathbf{k}$ and $\mathbf{S}$ have the directions of our $\mathbf{v}_\mathrm{n}$ and $\mathbf{r}$.

We can now begin the simultaneous solution of equations (1.9.5) and (1.9.8). Substituting assumption 1.9.15 into (1.9.5) gives

$$\begin{aligned}\mathbf{E} &= -\mathbf{r}\times\mu\mathbf{H}\\ &= -\mathbf{r}\times\mu(\mathbf{r}\times\mathbf{D}) && \text{by (1.9.8)}\\ &= \mu\,\mathbf{r}\times(\mathbf{D}\times\mathbf{r})\\ &= \mu(\mathbf{r}.\mathbf{r}\,\mathbf{D}-\mathbf{r}.\mathbf{D}\,\mathbf{r})\,;\end{aligned}$$

that is,

$$\mathbf{E} = \mu r^2\mathbf{D} - \mu\,\mathbf{r}.\mathbf{D}\,\mathbf{r}, \tag{1.9.20}$$

where r is the magnitude of $\mathbf{r}$. This confirms that $\mathbf{E}$, $\mathbf{D}$, and $\mathbf{r}$ are coplanar, as stated in theorem 1.9.19. Moreover, we shall see that equation (1.9.20), when combined with the relation between $\mathbf{D}$ and $\mathbf{E}$ for the medium, completely specifies the permitted ray speed(s) and polarization(s) as functions of direction.

First note that equation (1.9.20) can be arranged as $\mu r^2\mathbf{D} = \mathbf{E} + \mu\,\mathbf{r}.\mathbf{D}\,\mathbf{r}$, showing that $\mathbf{D}$ consists of a component parallel to $\mathbf{E}$ and a component normal to $\mathbf{E}$. But *in any direction in which the medium requires that* $\mathbf{D}\parallel\mathbf{E}$, the normal component must be zero, so that the equation reduces to

$$\mathbf{D} = \epsilon\mathbf{E}, \tag{1.9.21}$$

where

$$\epsilon = \frac{1}{\mu r^2}\,. \tag{1.9.22}$$

We shall call ϵ the **directional permittivity**. It is obviously real and positive, and has a value for each direction in which $\mathbf{D}\parallel\mathbf{E}$. So *if* $\mathbf{D}$ is parallel to $\mathbf{E}$, there is an instantaneous proportionality between them (under the assumption of simple EM waves and assumption 1.9.15). Solving (1.9.22) for r gives the classical formula for the speed of light:

$$r = 1\Big/\sqrt{\mu\epsilon}\,. \tag{1.9.23}$$

This formula depends only on equations (1.9.5) and (1.9.8), assumption 1.9.15, and the condition that $\mathbf{D}\parallel\mathbf{E}$, and applies to any propagation direction in which that condition holds. Notice that the speed depends on the coefficients of proportionality between the fields.

A "propagation direction" for which $\mathbf{D}\parallel\mathbf{E}$ is not only a ray direction but also a wave-normal direction, and the resulting r is not only the ray speed but

also the wave-normal speed. This can be confirmed by setting $\mathbf{E} \parallel \mathbf{D}$ in relation (1.9.16), which then implies that $\mathbf{v}_n \parallel \mathbf{r}$. As $\mathbf{v}_n$ is the component of $\mathbf{r}$ in the direction of $\mathbf{v}_n$, this means $\mathbf{v}_n = \mathbf{r}$.

If $\mathbf{E}$ happens to be a *sinusoidal* function of time, the instantaneous proportionality between $\mathbf{D}$ and $\mathbf{E}$ does not exclude the possibility that the coefficient ϵ depends on frequency—as is generally the case, in any medium except a vacuum. The resulting frequency-dependence of r is called **dispersion**.

For each direction in which $\mathbf{D} \parallel \mathbf{E}$, we can substitute (1.9.21) into (1.9.8), obtaining

$$\mathbf{H} = \mathbf{r} \times \epsilon \mathbf{E} \,. \tag{1.9.24}$$

As $\mathbf{E}, \mathbf{H}, \mathbf{r}$ form a right-hand orthogonal triad, their magnitudes are related by

$$H = r \epsilon E \,, \tag{1.9.25}$$

where $H = |\mathbf{H}|$; $E = |\mathbf{E}|$. Substituting from (1.9.23) and rearranging, we find

$$E/H = \sqrt{\mu/\epsilon} \,. \tag{1.9.26}$$

This ratio is an *impedance*: in SI units, E is in volts per metre and H in amps per metre, so that E/H is in ohms. As ϵ can vary with frequency, so can E/H.

In an *isotropic* medium, $\mathbf{D} \parallel \mathbf{E}$ for any direction of $\mathbf{E}$, and r is independent of direction, so that equations (1.9.21) to (1.9.26) hold in all directions, and r is both the ray speed and the wave-normal speed in all directions. The isotropic ϵ is called simply the *permittivity* of the medium, and the isotropic E/H is called the *intrinsic impedance* (or *characteristic impedance*) of the medium.

For a *non-isotropic* medium we make the following assumption, which will be retained from now on.

1.9.27 Assumption: *The dependence of* $\mathbf{D}$ *on* $\mathbf{E}$ *is* ***linear****. In other words:* **(a)** *multiplying* $\mathbf{E}$ *by a scalar causes* $\mathbf{D}$ *to be multiplied by the same scalar;* **(b)** *the* $\mathbf{D}$ *field due to a (vector) sum of* $\mathbf{E}$ *fields is the (vector) sum of the* $\mathbf{D}$ *fields due to the separate terms in the sum of* $\mathbf{E}$ *fields.*

In particular, part (a) implies that a sign-change in $\mathbf{E}$ is associated with a sign-change in $\mathbf{D}$.

To make $\mathbf{D}$ *parallel* to $\mathbf{E}$ requires two degrees of freedom, and two are available in choosing the direction of $\mathbf{E}$. So, of all the directions in which we could apply an $\mathbf{E}$ field, we would expect (at least) one in which the resulting $\mathbf{D}$ is parallel to $\mathbf{E}$. Let that direction be the z direction (choosing the coordinate

axes to fit the problem). The parallelism, together with 1.9.27(a), indicates a symmetry about planes normal to the z axis, i.e. parallel to the (x, y) plane, such that if $\mathbf{E}$ is in that plane, so is $\mathbf{D}$. Now repeat the argument in two dimensions. Of all the directions in the (x, y) plane in which we could apply an $\mathbf{E}$ field, obtaining a $\mathbf{D}$ field in the same plane, we would expect (at least) one direction in which $\mathbf{D} \parallel \mathbf{E}$. Let it be the y direction. This parallelism, together with 1.9.27(a), indicates a symmetry about planes normal to the y axis, i.e. parallel to the (z, x) plane. Due to the symmetries about the (x, y) and (z, x) planes, both of which contain the x axis, an $\mathbf{E}$ field in the x direction produces a parallel $\mathbf{D}$ field. Thus there are three mutually perpendicular directions in which $\mathbf{D} \parallel \mathbf{E}$, hence three planes of symmetry normal to those directions.[25]

Let $\mathbf{i}, \mathbf{j}, \mathbf{k}$ be the unit vectors in the x, y, z directions, respectively. These are the directions in which $\mathbf{D} \parallel \mathbf{E}$, so that equations (1.9.21) and (1.9.22) apply. Let the respective values of ϵ be $\epsilon_x, \epsilon_y, \epsilon_z$, and let the corresponding values of r be a, b, c. Then, by (1.9.21), the displacement fields $D_x\mathbf{i}$, $D_y\mathbf{j}$, and $D_z\mathbf{k}$ are produced respectively by the electric fields $D_x\mathbf{i}/\epsilon_x$, $D_y\mathbf{j}/\epsilon_y$, and $D_z\mathbf{k}/\epsilon_z$, where, by (1.9.22),

$$\epsilon_x = \frac{1}{\mu a^2} \; ; \quad \epsilon_y = \frac{1}{\mu b^2} \; ; \quad \epsilon_z = \frac{1}{\mu c^2} \; ; \tag{1.9.28}$$

that is, the displacement fields $D_x\mathbf{i}$, $D_y\mathbf{j}$, and $D_z\mathbf{k}$ are given respectively by the electric fields $\mu a^2 D_x\mathbf{i}$, $\mu b^2 D_y\mathbf{j}$, and $\mu c^2 D_z\mathbf{k}$. Hence, by assumption 1.9.27(b), the electric field

$$\mathbf{E} = \mu a^2 D_x\,\mathbf{i} + \mu b^2 D_y\,\mathbf{j} + \mu c^2 D_z\,\mathbf{k} \tag{1.9.29}$$

gives the displacement field

$$\mathbf{D} = D_x\,\mathbf{i} + D_y\,\mathbf{j} + D_z\,\mathbf{k}\,. \tag{1.9.30}$$

Now, if we let the ray velocity be

$$\mathbf{r} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}, \tag{1.9.31}$$

so that

$$r^2 = x^2 + y^2 + z^2, \tag{1.9.32}$$

[25] This explanation neglects the possibility that a purely *alternating* $\mathbf{E}$ field produces a $\mathbf{D}$ field with a *rotating* component. Media with this property are described as **chiral**, and can cause a gradual rotation of the polarization of a wave as it propagates (known as **optical activity**). These phenomena are not covered by the present "simple" theory.

we can find the locus of $\mathbf{r}$. Substituting (1.9.29), (1.9.30), and (1.9.31) into equation (1.9.20) and canceling μ, we obtain

$$a^2D_x\mathbf{i}+b^2D_y\mathbf{j}+c^2D_z\mathbf{k} = r^2(D_x\mathbf{i}+D_y\mathbf{j}+D_z\mathbf{k}) - (xD_x+yD_y+zD_z)(x\mathbf{i}+y\mathbf{j}+z\mathbf{k}) \tag{1.9.33}$$

or, equating components in the $\mathbf{i}$, $\mathbf{j}$, and $\mathbf{k}$ directions,

$$\begin{bmatrix} r^2-x^2-a^2 & -xy & -xz \\ -xy & r^2-y^2-b^2 & -yz \\ -xz & -yz & r^2-z^2-c^2 \end{bmatrix} \begin{bmatrix} D_x \\ D_y \\ D_z \end{bmatrix} = \mathbf{0}\,. \tag{1.9.34}$$

The existence of a non-trivial solution $\mathbf{D}$, whose components D_x, D_y, D_z are *not all zero*, requires the determinant of the 3×3 coefficient matrix to be zero. If we temporarily name the subexpressions

$$K = r^2-a^2\,,\quad L = r^2-b^2\,,\quad M = r^2-c^2\,, \tag{1.9.35}$$

we find that the determinant expands and simplifies to

$$KLM - LMx^2 - MKy^2 - KLz^2. \tag{1.9.36}$$

Setting this to zero, dividing through by KLM, and reinstating subexpressions, we obtain the short form of the equation for the ray-velocity surface:

$$\frac{x^2}{r^2-a^2}+\frac{y^2}{r^2-b^2}+\frac{z^2}{r^2-c^2} = 1\,. \tag{1.9.37}$$

Multiplying this by (1.9.32) and collecting terms, we get the alternative form

$$\frac{a^2x^2}{r^2-a^2}+\frac{b^2y^2}{r^2-b^2}+\frac{c^2z^2}{r^2-c^2} = 0\,, \tag{1.9.38}$$

which is useful because it can be "multiplied through" without complicating the right-hand side. Multiplying it through by the common denominator, we might expect to obtain an equation of the 6^{th} degree. But in fact, if we expand the products of two denominators, we can collect the term $a^2b^2c^2(x^2+y^2+z^2)$, in which the parenthesized factor is r^2. We can then cancel this factor in all terms, reducing the degree by two and obtaining

$$\begin{aligned} r^2\left(a^2x^2+b^2y^2+c^2z^2\right) &- a^2(b^2+c^2)x^2 \\ &- b^2(c^2+a^2)y^2 \\ &- c^2(a^2+b^2)z^2 + a^2b^2c^2 = 0\,. \end{aligned} \tag{1.9.39}$$

This form—which is the same as Fresnel's [21, p.233], except that we retain one instance of r^2 for abridgment—is more verbose than the preceding forms, but shows at a glance that the surface is only of the 4^{th} degree.

In any of the forms (1.9.37) to (1.9.39), the even powers of the coordinates indicate mirror-symmetry in the coordinate planes.

For some purposes, however, it is most convenient to work directly from (1.9.34), which includes the components of $\mathbf{D}$, hence the polarization. For example, if the ray velocity is in the (x, y) plane, the symmetry about that plane requires $\mathbf{D}$ to be either normal to that plane or *in* it. In the former case, we put $z = 0$ (hence $r^2{=}x^2{+}y^2$) and $D_x = D_y = 0$ in system (1.9.34), reducing it to

$$(x^2{+}y^2{-}c^2)\,D_z = 0\,, \tag{1.9.40}$$

in which a non-trivial solution ($D_z \neq 0$) requires that $x^2{+}y^2 = c^2$. In the latter case, we put $z = 0$ (hence $r^2{=}x^2{+}y^2$) and $D_z = 0$ in (1.9.34), reducing it to

$$\begin{bmatrix} y^2{-}a^2 & -xy \\ -xy & x^2{-}b^2 \end{bmatrix} \begin{bmatrix} D_x \\ D_y \end{bmatrix} = \mathbf{0}\,. \tag{1.9.41}$$

For a non-trivial solution (in which D_x and D_y are not both zero), the $2{\times}2$ determinant must be zero. When that determinant is expanded, the terms in x^2y^2 cancel and we can divide through by a^2b^2, obtaining $x^2/b^2 + y^2/a^2 = 1$. In summary, the ray-velocity surface intersects the (x, y) plane in the curves

$$x^2 + y^2 = c^2 \qquad \text{for } \mathbf{D} \perp \text{the plane} \tag{1.9.42}$$

$$x^2/b^2 + y^2/a^2 = 1 \qquad \text{for } \mathbf{D} \text{ in the plane}\,. \tag{1.9.43}$$

These describe a circle with radius c, and an ellipse with semi-principal axes b in the x direction and a in the y direction, both centered on the origin. Similarly, the ray-velocity surface intersects the (y, z) plane in the curves

$$y^2 + z^2 = a^2 \qquad \text{for } \mathbf{D} \perp \text{the plane} \tag{1.9.44}$$

$$y^2/c^2 + z^2/b^2 = 1 \qquad \text{for } \mathbf{D} \text{ in the plane}\,, \tag{1.9.45}$$

and intersects the (z, x) plane in the curves

$$z^2 + x^2 = b^2 \qquad \text{for } \mathbf{D} \perp \text{the plane} \tag{1.9.46}$$

$$z^2/a^2 + x^2/c^2 = 1 \qquad \text{for } \mathbf{D} \text{ in the plane}\,. \tag{1.9.47}$$

Equations (1.9.42), (1.9.44), and (1.9.46) indicate that the speed of light is isotropic within the three planes normal to the axes on which $\mathbf{D} \parallel \mathbf{E}$.

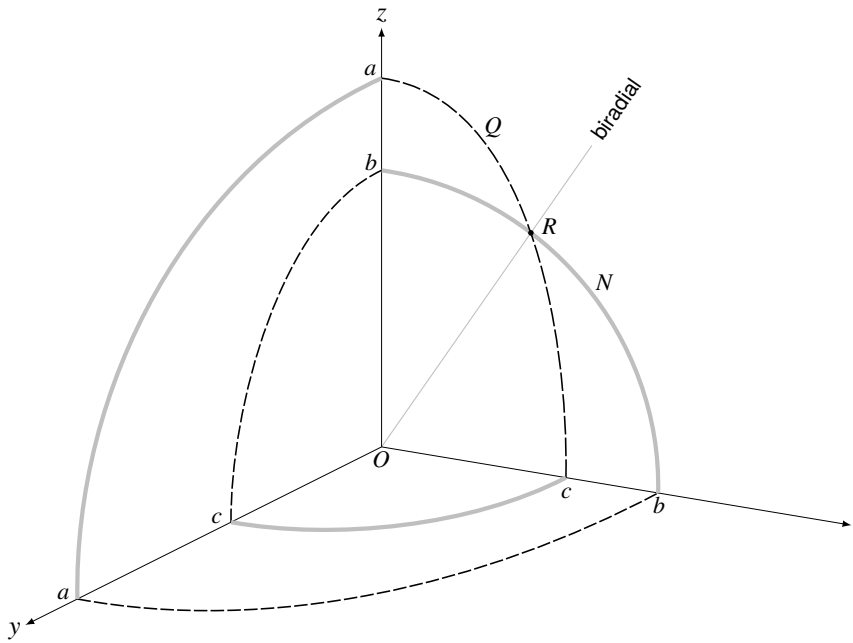

Fig. 1.15: One octant of the two-sheeted ray-velocity surface for a non-isotropic medium.

Concerning polarization for the elliptical intersections, the conditions on equations (1.9.43), (1.9.45), and (1.9.47) tell us only that **D** is in the plane of the intersection. But because the ray-velocity surface is also a wavefront, to which **D** must be tangential, confining **D** to the plane of the intersection means making it *tangential* to the curve of intersection.

Now let the three radii a, b, c be distinct, and assume, without further loss of generality (renaming axes if necessary), that $a > b > c$. In the *first octant* (for which $x, y, z \geq 0$), the circles of intersection become quarter-circles, shown as the gray arcs aa, bb, and cc in Fig. 1.15, where the coordinates marked on the axes also serve as names of points. (The axes are left-handed, as seems to be traditional in this context, but could be made right-handed by changing the y label to $-y$, in which case the octant shown becomes the fourth octant.) The quarter-circles do not meet on the axes—and need not, because their polarizations differ. Hence, even if we did not have equations (1.9.43), (1.9.45), and (1.9.47), we would know that we need three other arcs to complete the intersections of the unit-wave with the quarter-planes. We would also know that on each arc of intersection, **D** is either normal to the plane or tangential to the arc; and as the quarter-circles account for the former case, the other arcs must be the latter. Hence, to match the polarizations where the other arcs meet

the quarter-circles, the other arcs must meet the axes at right angles. On that information, we could make a respectable sketch of the other arcs even if we did not know their exact shapes. But we do: by (1.9.43), (1.9.45), and (1.9.47), they are quarter-ellipses aligned with the axes. In Fig. 1.15, they are drawn as black dashed curves, in which the *black dashes show the direction of* **D**.

By interpolation between the curves of intersection, we infer that the ray-velocity surface is **two-sheeted**, with the inner sheet *Rccb* and the outer sheet *RNbaaQ* meeting at point *R*, like two spinnakers pinned together. Then, if we extend the surface into the other seven octants by successive reflection in the coordinate planes, we infer that the surface encloses the origin *twice*. Double enclosure is consistent with the degree of equation (1.9.39): if we choose two coordinates, there are at most four solutions for the third coordinate; and the even powers of the coordinates imply that these solutions are in equal and opposite pairs, giving 0, 2, or 4 solutions. In other words, a line normal to one of the coordinate planes will cut the surface at 0, 2, or 4 points, arranged symmetrically about that coordinate plane.

The axis *OR*, on which both sheets have the same radius (ray velocity), is called the **biradial** axis (or the *ray axis*). It has the *direction in which the ray velocity is independent of polarization*. For that ray direction, the permitted directions of **D** are in the plane tangential to the ellipse *aQRc* and normal to the (x, z) plane, because **D** can have a component normal to the circle by the condition on (1.9.46), *and* a component tangential to the ellipse by the condition on (1.9.47). For each direction of **D**, the corresponding direction of **E** is normal to the ray and in the plane of **D** and the ray (by theorem 1.9.19).

By the symmetry about the coordinate planes, there are *two biradial axes* passing through O in the (x, z) plane, containing a total of four points like R, equidistant from O. The inner and outer sheets of the ray-velocity surface both enclose the origin O and touch each other at those four points. But we need to know precisely *how* they touch each other, because that determines the relations between the radius vectors and normal vectors around the contact points, hence the relations between the ray and wave-normal directions around those points, hence the refractive and reflective behavior of the medium for ray and wave-normal directions around the axes through those points. To investigate the nature of the contact points, let us first consider qualitative arguments based on symmetry, continuity, and interpolation, and then check our inferences against quantitative arguments based on analytic geometry.

According to the symmetry, the intersections of the ray-velocity surface with the (x, y) and (y, z) planes can be extended into the back octant of Fig. 1.15 by reflecting them in the (x, z) plane. As this plane contains the contact point R, we will then have enough of the surface to surround that point—as shown

in Fig. 1.16, where the inner sheet has become $-c\,ccb$, and the outer sheet has become $-a\,baa$, and the two sheets touch at R.

Now consider the continuity requirements. As the ray-velocity surface encloses the origin twice, each direction (from the origin) generally corresponds to two velocities, represented by one point on the inner sheet and one on the outer. At each point, the polarization (of **D**) must be *tangential to the sheet and in the plane of the radius and the normal* (that is, tangential to the wavefront and in the plane of the ray and the wave-normal). Thus, for each direction, there are generally two permitted polarizations. For each polarization, the continuity of the medium suggests that *a smooth variation in ray direction corresponds to a smooth variation in ray speed and polarization*.

This is true for the intersections of the unit-wave with the coordinate planes, provided that, as we pass through R in the (x, z) plane, we stick to the circle or the ellipse and do not switch from one to the other—in other words, provided that we stick to the corresponding polarization rather than the corresponding sheet; indeed, in the (x, z) plane, the circle and the ellipse *pass smoothly from one sheet to the other* as they pass through R. Hence, appealing again to continuity, we should expect that the unit-wave, adjacent to the circle *bNRb* and the ellipse *aQRc*, contains other smooth curves along which one can pass through R from the outer sheet to the inner, with a smooth variation in polarization.

We have just seen that in the (x, z) plane, we can pass from the outer sheet along a smooth curve through R to the inner sheet, whether we approach R from the N direction or the Q direction. Interpolating between those directions, we might infer that *in the plane of R and the y axis*, we can likewise pass from the outer sheet along a smooth curve through R to the inner sheet. (But because this inference comes from interpolation rather than continuity, we shall confirm it analytically in due course.) As such a curve passes from the outer sheet on one side of the (x, z) plane to the inner sheet on the other side, it is not symmetrical about that plane, and would therefore be expected *not* to meet that plane at right angles, but rather to be inclined towards the origin. (This too will be confirmed in due course.) Hence, in Fig. 1.16, the intersection of the unit-wave with the plane of R and the y axis would be expected to include a curve like $-a\,SRc$, and its mirror-image in the (x, z) plane, the curve $aPR,-c$.

On the outer sheet, point R is surrounded by points N, P, Q, S, from each of which a smooth curve passes along the outer sheet through R onto the inner sheet. Interpolating between the curves, we would expect that the quarter of the unit-wave depicted in Fig. 1.16 consists of an infinite number of smooth curves extending from the outer semicircle $-a\,aa$ through R to the inner semicircle $cc,-c$ (with the points listed in corresponding directions), and from the outer semi-ellipse $-a\,ba$ through R to the inner semi-ellipse $cb,-c$, and that near

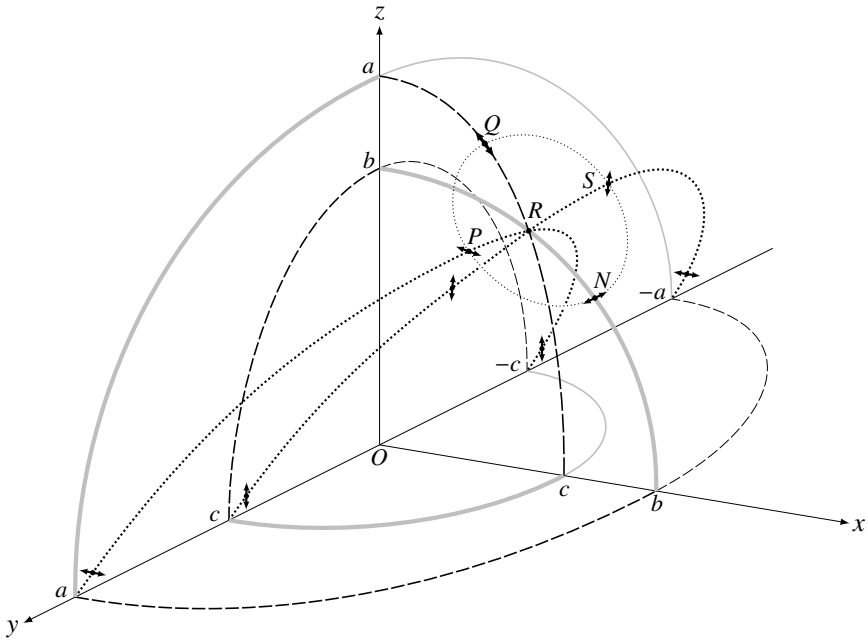

Fig. 1.16: Minimal wireframe drawing of the same two-sheeted ray-velocity surface as in Fig. 1.15.

point R, the two-sheeted surface approximates a *double cone*—not necessarily a right circular cone, but a cone with, at worst, a simple oval cross-section. In that case, R is a roughly conical peak on the inner sheet and a roughly conical dimple in the outer sheet; and as the outer sheet is generally convex, the dimple must be surrounded by a ridge, whose crest is represented by the curve $NPQS$.

At N and Q, and at the ends of arcs $aPR{-}c$ and $-a\,SRc$, the polarizations (already known) are shown by double-ended arrows. On the former arc, as we move from a to $-c$, the polarization points at an acute angle to the *right* of the direction of movement (looking towards O). So, invoking continuity at R and interpolating, we can sketch the polarization at P. Applying the same argument to the other arc, or reflecting in the (x, z) plane, we can sketch the polarization at S. The last two results yield another interpolation: as we make a full turn about the loop $NPQS$, the polarization makes a *half*-turn in the same direction. This conclusion does not depend on the assumption that the loop is the "crest" of the "ridge"; the reasoning still holds if the loop is tightened or loosened by moving points N, P, Q, S towards or away from R. On the arc $-a\,SRc$, we can interpolate one more polarization on the line from O to P, showing how the direction OP is associated with two rays speeds with crossed polarizations.

If the ray-velocity surface (unit-wave) is double-conical at infinitesimal distances from R, it has a cone of normals at that point, so that a ray in the

direction OR is associated with a cone of wave-normals. As the wavefronts pass out of this medium into another, the wave-normals will still form a cone due to wavefront continuity at the boundary. But if the second medium is isotropic, the wave-normal directions will coincide with ray directions. Thus, if a ray passing along the biradial axis OR, containing all permitted polarizations, exits the medium into (e.g.) air or water, it will break into a hollow cone of rays; this is **external conical refraction**. And a half-turn in the polarization of the ray will cause a full turn of the ray about the cone; this is **conical polarization**.

External conical refraction was predicted by Hamilton in October 1832, and confirmed experimentally by his colleague, the Rev. Humphrey Lloyd, in December of that year [21]. But it was Lloyd who first noticed the polarization pattern and explained it in terms of previous theory [16, pp. 149–50]. Only after the event did Hamilton give his own explanation. This was included in the printed version of his paper [12, s.30], which appeared together with Lloyd's report [16] in the summer of 1833 [25, p.157]. At this time, of course, light waves were not yet known to be EM waves, but were assumed to be mechanical.

Let us complete our own explanation of conical refraction by adding some analytical evidence to the preceding arguments that were based on symmetry, continuity, and the dubious art of interpolation.

First we confirm that the intersection of the unit-wave with the plane of R and the y axis consists of smooth curves passing obliquely through the (x, z) plane, from one sheet to the other. By simultaneous solution of (1.9.46) and (1.9.47), the coordinates of R in the (x, z) plane are found to be

$$x_R = c\left(\frac{a^2-b^2}{a^2-c^2}\right)^{1/2} \quad ; \quad z_R = a\left(\frac{b^2-c^2}{a^2-c^2}\right)^{1/2}. \tag{1.9.48}$$

For finding directions of tangents of curves through point R, we can equate the radial coordinate r with the normal distance from the y axis (the error being proportional to y^2 for small y). In the plane of R and the y axis, interpreting r as the distance from the y axis, and noting that b is the distance of R from that axis, we can write

$$x = \frac{x_R}{b}\, r \quad ; \quad z = \frac{z_R}{b}\, r\,. \tag{1.9.49}$$

The remaining steps are trivial but tedious, and therefore best accomplished with the aid of a computer algebra package. Substituting (1.9.49), then (1.9.48), into equation (1.9.38), multiplying through by the common denominator, writing

$$r = b + \rho\,, \tag{1.9.50}$$

expanding completely, and retaining terms of up to 2nd degree in y and ρ, we find that the only such terms are in y^2 and ρ^2. Hence we can solve for ρ. Then,

back-substituting into (1.9.50), we obtain

$$r \approx b \pm \frac{\sqrt{(a^2-b^2)(b^2-c^2)}}{2ac}\, y\,. \tag{1.9.51}$$

And this is a first-order approximation (for small y) to the intersection of the unit-wave with the plane of R and the y axis. Due to the non-zero coefficient of y (for $a > b > c$), it describes an X-shaped pattern in which two smooth curves pass at opposite oblique angles through the (x, z) plane. So the smooth curves pass from one sheet to the other—because staying on one sheet would mean taking the top branches of the 'X' (for the outer sheet) or the bottom branches (for the inner sheet), causing an abrupt change in direction at the junction.

Now we adduce some evidence, other than interpolation, to show that the unit-wave approximates a double cone near R. At the intersections of the unit-wave with a sphere centered on the origin, we treat r as constant in, e.g., equation (1.9.38). To project those intersections onto the (x, y) plane, we put $z^2 = r^2 - x^2 - y^2$ in (1.9.38) and collect terms, obtaining

$$\frac{x^2}{c^2(a^2-r^2)/(a^2-c^2)} - \frac{y^2}{c^2(r^2-b^2)/(b^2-c^2)} = 1\,. \tag{1.9.52}$$

For constant r, this clearly describes a *conic section* symmetrical about the x and y axes; in particular, since $a > b > c$, it describes a hyperbola for $a > r > b$ and an ellipse for $b > r > c$ (because the denominator of x^2 remains positive, whereas the denominator of y^2 changes from positive to negative as r falls through b). Similarly, to project the intersections onto the (y, z) plane, we put $x^2 = r^2 - y^2 - z^2$ in (1.9.38), obtaining

$$\frac{z^2}{a^2(r^2-c^2)/(a^2-c^2)} + \frac{y^2}{a^2(r^2-b^2)/(a^2-b^2)} = 1\,. \tag{1.9.53}$$

This describes an ellipse for $a > r > b$ and a hyperbola for $b > r > c$ (the reverse of the previous case). The conic sections described by the last two equations are not literally intersections between a plane and a cone, but rather projections of the intersections of a sphere with the unit-wave, for which the departure of the unit-wave from a cone compensates for the departure of the sphere from a plane, so that the projections of the intersections on two different planes are exact conic sections. But, as we approach point R and the three other similar points, the departure of the sphere from a plane becomes less and less, and requires the unit-wave to depart less and less from a cone in order to compensate. Seeing that the denominators in the above two equations are smooth functions, and having previously found four smooth curve along which we can pass from the outer sheet to the inner sheet through R, we know that the limiting case is a

double cone and not merely two different single cones with a common apex. As a double cone is somewhat reminiscent of a juggler's diabolo, the point of contact is called (I jest not...) a **diabolical point**![26]

It does *not* follow that the limiting cone is right-circular, because a conic section remains a conic section if it is stretched along one axis—that is, if the generating cone and plane are stretched along the same axis, with the result that the cone is no longer right-circular. But clearly the cross-section is at worst a simple oval, as initially claimed. When the unit-wave is extended into the other octants, the two sheets resemble two inflated balloons, one inside the other, spot-welded together at four points (cf. Fig. 2 in [4]), with the further property that the surface approximates a *double* cone around each point of contact.

External conical refraction occurs when an internal ray with all possible polarizations travels in a biradial direction. But usually a single internal ray direction corresponds to two different ray speeds (one per sheet), each with its own polarization and wave-normal direction. If such a "ray" exits the medium into an isotropic medium, the corresponding external wave-normal directions become ray directions, giving external *double* refraction. Similarly, if a ray of mixed polarization comes *from* an isotropic medium into a non-isotropic one, the incident wavefront usually gives two refracted wave speeds (one per sheet), hence two wave-normal directions (due to wavefront continuity), each with its own polarization and ray direction, causing double refraction of the internal kind. Double refraction is also called **birefringence**.

A medium in which the principal speeds a, b, c are all different, giving two biradial axes, is called a **biaxial** birefringent medium. But now let us consider the special cases in which the principal speeds a, b, c are *not* all different.

First let b increase until $b = a$. Then, in Fig. 1.15 (p. 60), point b on the z axis moves up to a, and the quarter-circle bb expands proportionally, and the quarter-ellipse ab becomes a quarter-circle of radius a. Hence, by interpolation, we might guess that the outer sheet of the unit-wave becomes a *sphere* of radius a, and that the inner sheet becomes a *prolate spheroid* with semi-major axis a and semi-minor axis c. Whatever the exact shapes may be, it is clear that the biradial axes converge on the z axis, and that the four diabolical points merge into two simple points of contact on the z axis.

Alternatively, let b decrease until $b = c$. Then point b on the x axis moves in to c, and the quarter-circle bb shrinks proportionally, and the quarter-ellipse bc becomes a quarter-circle of radius c. Hence, by interpolation, we might guess that the *inner* sheet of the unit-wave becomes a *sphere* of radius c, and

[26] Hamilton [12] called it a *conoidal cusp*. This term departs from the standard mathematical meanings of both *conoid* and *cusp*, and seems to have fallen out of use.

that the outer sheet becomes an *oblate spheroid* with semi-major axis a and semi-minor axis c. The biradial axes converge on the x axis, and the four diabolical points merge into two simple points of contact on the x axis.

A birefringent medium in which the biradial axes merge is described as **uniaxial**. If the outer sheet of the unit-wave is spherical and the inner sheet prolate $(b = a)$, the medium is described as **positive**; an example is *rutile* (a form of TiO_2). If the inner sheet of the unit-wave is spherical and the outer sheet oblate $(b = c)$, the medium is described as **negative**; an example is *calcite* (a form of $CaCO_3$). In either case, the two sheets meet on their common axis—at the poles, as it were, and *not* at the equators. For the spherical sheet, on which the ray is normal to the wavefront, the ray and the wave are described as **ordinary**. For the spheroidal sheet, on which the ray departs from the wave-normal (except at the equator and the poles), the ray and the wave are described as **extraordinary**. Interpolating the polarizations (and continuing the geographic analogy), we would expect the direction of **D** to be along the lines of latitude for the spherical (ordinary) sheet, and along the lines of longitude for the spheroidal (extraordinary) sheet.

Our expectations can be confirmed by algebra. For the case in which $b = a$, equation (1.9.39) can be put in the form

$$\begin{aligned} r^2(a^2x^2 + a^2y^2 + c^2z^2) - a^2c^2(x^2+y^2+z^2) \\ - a^2(a^2x^2 + a^2y^2 + c^2z^2 - a^2c^2) = 0\,. \end{aligned} \tag{1.9.54}$$

Recognizing $(x^2+y^2+z^2)$ as r^2, we can factor the left-hand side, obtaining

$$(r^2 - a^2)(a^2x^2 + a^2y^2 + c^2z^2 - a^2c^2) = 0\,; \tag{1.9.55}$$

that is,

$$r = a \qquad \text{or} \qquad \frac{x^2+y^2}{c^2} + \frac{z^2}{a^2} = 1\,. \tag{1.9.56}$$

The first option is the expected outer sphere, and the second is the expected inner prolate spheroid. The axial symmetry about the z axis confirms the expected polarization pattern, because it implies that the patterns in the (x, z) and (y, z) planes are replicated in every plane through the z axis. Alternatively, for the case in which $b = c$, a similar procedure yields

$$r = c \qquad \text{or} \qquad \frac{x^2}{c^2} + \frac{y^2+z^2}{a^2} = 1\,, \tag{1.9.57}$$

where the two options are the inner sphere and the outer oblate spheroid. In this case the axial symmetry is about the x axis, so that the polarization patterns in the (x, z) and (x, y) planes are replicated in every plane through the x axis.

In either case, if $a = c$, the spheroid merges with the sphere, so that the unit-wave becomes a single spherical sheet on which both polarizations (and all combinations thereof) are permitted. This case has already been treated in the commentary following equations (1.9.21) to (1.9.26).

At this stage, for the more complicated biaxial case, our explanation of the polarization at P and S (in Fig. 1.16) still relies on interpolation. We could try to rectify that situation by further studying the analytic geometry of the "ridge" $NPQS$. But we can establish the essential fact about the geometry and learn far more about the physics, with less effort, by starting a new line of inquiry.

Recall that the vector $\mathbf{r}$ (with magnitude r) is the ray velocity, while the vector $\mathbf{v}_\mathrm{n}$ (with magnitude v_n) is the wave-normal velocity. Now let us define the **ray slowness** vector $\mathbf{s}_\mathrm{r}$ as the vector in the direction of $\mathbf{r}$ with magnitude

$$s_\mathrm{r} = 1/r \,. \tag{1.9.58}$$

Similarly, let us define the **wave slowness** vector $\mathbf{s}$ (also called the *normal slowness*) as the vector in the direction of $\mathbf{v}_\mathrm{n}$ with magnitude

$$s = 1/v_\mathrm{n} \,. \tag{1.9.59}$$

As the ray-velocity surface (unit-wave) is the locus of $\mathbf{r}$, so the **ray-slowness surface** is the locus of $\mathbf{s}_\mathrm{r}$. And as the wave-velocity surface is the locus of $\mathbf{v}_\mathrm{n}$, so the **wave-slowness surface** is the locus of $\mathbf{s}$. To find $\mathbf{s}_\mathrm{r}$ from $\mathbf{r}$ or vice versa, we keep the direction but take the reciprocal of the magnitude. This transformation is called *inversion in the unit sphere*. Hence we say that *the ray-slowness surface is the **inverse** of the ray-velocity surface*, and vice versa. Similarly, *the wave-slowness surface is the inverse of the wave-velocity surface*, and vice versa. As the wave-velocity surface is the pedal of the ray-velocity surface, it follows that *the wave-slowness surface is the **inverse of the pedal** of the ray-velocity surface*.[27]

To see how we would find that "inverse of the pedal", let the ray velocity $\mathbf{r}$ be $\mathbf{r}(u, v)$, with magnitude $r(u, v)$, so that u and v are parameters of the ray-velocity surface, and let a subscript u or v indicate partial differentiation w.r.t. that parameter. (A **partial derivative** of a function of several variables is a derivative w.r.t. *one* of them while the others are held constant.) Similarly, let

[27] I avoid the term *wave surface* because some authors (e.g., Berry & Jeffrey [4]) use that term for the wave-slowness surface or a scaled version thereof, while others (e.g., Buchwald [7] and de Witte [8]) use it for the ray-velocity surface. Lunney and Weaire [19] call the wave-slowness surface the "wave-normal surface" in four places and the "wave surface" in one place, where they belatedly note that its relation to the wave velocity involves a reciprocal.

$\mathbf{v}_\mathrm{n}$, $\mathbf{s}$, and $\mathbf{s}_\mathrm{r}$ (and their magnitudes v_n, s, and s_r) be corresponding functions of u and v. Then, as $\mathbf{r}_u$ and $\mathbf{r}_v$ are tangential to the ray-velocity surface, the cross product $\mathbf{r}_u \times \mathbf{r}_v$ is normal to that surface and is therefore in the direction of $\mathbf{v}_\mathrm{n}$ and $\mathbf{s}$ (if we suppose, without loss of generality, that we have chosen u and v so that the cross product is in the outward sense). Dividing that cross product by its magnitude yields the unit vector in the direction of $\mathbf{v}_\mathrm{n}$ and $\mathbf{s}$, denoted by $\hat{\mathbf{s}}$:

$$\hat{\mathbf{s}} = \frac{\mathbf{r}_u \times \mathbf{r}_v}{|\mathbf{r}_u \times \mathbf{r}_v|} \,. \tag{1.9.60}$$

Now $\mathbf{v}_\mathrm{n}$ is the component of $\mathbf{r}$ in that direction, so its magnitude is

$$v_\mathrm{n} = \mathbf{r}.\hat{\mathbf{s}} = \frac{\mathbf{r}.\mathbf{r}_u \times \mathbf{r}_v}{|\mathbf{r}_u \times \mathbf{r}_v|} \,. \tag{1.9.61}$$

As the magnitude of the wave slowness is $1/v_\mathrm{n}$, the wave slowness *vector* is $\mathbf{s} = \hat{\mathbf{s}}/v_\mathrm{n}$. Substituting from the above yields

$$\mathbf{s}(u, v) = \frac{\mathbf{r}_u \times \mathbf{r}_v}{\mathbf{r}.\mathbf{r}_u \times \mathbf{r}_v} \,, \tag{1.9.62}$$

which is the wave-slowness vector corresponding to $\mathbf{r}(u, v)$.

Now, considering that (i) the radii of the ray-velocity surface give the ray velocities and directions, and (ii) the corresponding normals of the ray-velocity surface give the corresponding wave-normal directions, and (iii) the radii of the wave-slowness surface give the wave slownesses and wave-normal directions, we might well ask: (iv) *what, if anything, is the significance of the corresponding normals of the wave-slowness surface?*

To answer this, we could try to find a vector in the direction of interest, namely $\mathbf{s}_u \times \mathbf{s}_v$. But before we get that far, we shall discover a shortcut. First we differentiate (1.9.62) w.r.t. u by the quotient rule, obtaining

$$\mathbf{s}_u = \frac{(\mathbf{r}.\mathbf{r}_u \times \mathbf{r}_v)(\mathbf{r}_{uu} \times \mathbf{r}_v + \mathbf{r}_u \times \mathbf{r}_{vu}) - (\mathbf{r}_u \times \mathbf{r}_v)(\mathbf{r}.\mathbf{r}_{uu} \times \mathbf{r}_v + \mathbf{r}.\mathbf{r}_u \times \mathbf{r}_{vu})}{(\mathbf{r}.\mathbf{r}_u \times \mathbf{r}_v)^2} \,, \tag{1.9.63}$$

where the last parenthesized factor in the numerator comes from the identity $(\mathbf{a}.\mathbf{b} \times \mathbf{c})_u = \mathbf{a}_u.\mathbf{b} \times \mathbf{c} + \mathbf{a}.\mathbf{b}_u \times \mathbf{c} + \mathbf{a}.\mathbf{b} \times \mathbf{c}_u$ (from which, in this case, one term vanishes due to a repeated factor). Expanding the numerator and applying the identity $(\mathbf{a}.\mathbf{c})\mathbf{b} - (\mathbf{a}.\mathbf{b})\mathbf{c} = \mathbf{a} \times (\mathbf{b} \times \mathbf{c})$, in two places, we get

$$\mathbf{s}_u = \frac{\mathbf{r} \times \big[(\mathbf{r}_{uu} \times \mathbf{r}_v) \times (\mathbf{r}_u \times \mathbf{r}_v)\big] + \mathbf{r} \times \big[(\mathbf{r}_u \times \mathbf{r}_{vu}) \times (\mathbf{r}_u \times \mathbf{r}_v)\big]}{(\mathbf{r}.\mathbf{r}_u \times \mathbf{r}_v)^2} \,, \tag{1.9.64}$$

in which we can factor the numerator, obtaining

$$\mathbf{s}_u = \frac{\mathbf{r} \times \big[(\mathbf{r}_{uu} \times \mathbf{r}_v + \mathbf{r}_u \times \mathbf{r}_{vu}) \times (\mathbf{r}_u \times \mathbf{r}_v)\big]}{(\mathbf{r}.\mathbf{r}_u \times \mathbf{r}_v)^2} \,. \tag{1.9.65}$$

Now for the "shortcut": It immediately follows from (1.9.62) that $\mathbf{r}.\mathbf{s} = 1$ and $\mathbf{r}_u.\mathbf{s} = 0$, and from (1.9.65) that $\mathbf{r}.\mathbf{s}_u = 0$. Because u is general, these results can be written

$$\mathbf{r}.\mathbf{s} = 1 \; ; \quad \mathbf{r}'.\mathbf{s} = 0 \; ; \quad \mathbf{r}.\mathbf{s}' = 0 \, , \tag{1.9.66}$$

where the prime ($'$) denotes differentiation w.r.t. *any* corresponding parameter of the surfaces. The third equation in this set tells us that every tangent to the wave-slowness surface at s is normal to r. In other words:

1.9.67 Theorem: *The normal to the wave-slowness surface at point* $\mathbf{s}$ *is the ray direction for a wave-normal in the direction of* $\mathbf{s}$.

The theorem relates the *directions* of $\mathbf{s}$ and $\mathbf{s}_\mathrm{r}$. For their *magnitudes*, the first equation of (1.9.66) can be written $1 = \mathbf{s}.\mathbf{r}$ and multiplied by (1.9.58) to obtain

$$s_\mathrm{r} = \mathbf{s}.\hat{\mathbf{r}} \, , \tag{1.9.68}$$

where $\hat{\mathbf{r}}$ is the unit vector in the $\mathbf{r}$ direction—which, by 1.9.67, is normal to the wave-slowness surface at $\mathbf{s}$. So (1.9.68) means that the ray slowness is the component of the wave slowness normal to the wave-slowness surface; that is, *the ray-slowness surface is the pedal of the wave-slowness surface*. It follows (by taking inverses) that *the ray-velocity surface is the inverse of the pedal of the wave-slowness surface*. But we already knew that the wave-slowness surface is the inverse of the pedal of the ray-velocity surface. So, *to repeat the inverse-of-the-pedal transformation is to undo it.*

The last three results (in *italics*) were discovered by Hamilton and reported in the paper already cited, although he stated them differently (see [12, p.143], lines 4–7, 32–36, 12–16, respectively). They are various ways of saying that the wave-slowness surface bears the same relation to the ray-velocity surface as the latter to the former—or, in terms of radius vectors, that $\mathbf{s}$ bears the same relation to $\mathbf{r}$ as $\mathbf{r}$ to $\mathbf{s}$. The last statement is obvious from the symmetry of equations (1.9.66), provided of course that these equations are sufficient to define each vector in terms of the other. In principle, if we need to find $\mathbf{s}(u, v)$ given $\mathbf{r}(u, v)$, we know the three components of $\mathbf{r}$ as functions of u and v, and need to find the three components of $\mathbf{s}$ as functions of u and v, for which purpose we need three equations. So it seems *prima facie* that equations (1.9.66), which correspond to equations (P^{20}) of Hamilton [12, p.143], are indeed sufficient.

Because two iterations of the inverse-of-the-pedal transformation take us back where we started from, the inverse of the pedal surface is appropriately called the **reciprocal** surface. So *the wave-slowness surface is the reciprocal of the ray-velocity surface*, and vice versa.

Recall that *wavefront continuity* governs the refraction and reflection of waves, and hence the refraction and reflection of *rays* via the relation between the wave-normal and ray directions. Theorem 1.9.67 expresses that relation in terms of the wave-slowness surface, no less conveniently than theorem 1.9.2 expresses it in terms of the ray-velocity surface. Let us now see how wavefront continuity is most conveniently expressed in terms of wave slowness.

At a refractive/reflective surface, consider a wavefront (incident, refracted, or reflected) propagating with wave-normal velocity v_{n}. Where the wavefront intersects the surface, let the angle between the two (and hence between their normals) be θ, and let the curve of intersection move across the surface with normal velocity v_{t}, where the subscript 't' means *tangential to the surface* (but normal to the curve). Then v_{n} is the component of v_{t} in the wave-normal direction; that is, $v_{\mathrm{n}} = v_{\mathrm{t}} \sin\theta$. Taking reciprocals and rearranging, we find

$$s \sin\theta = 1/v_{\mathrm{t}}, \tag{1.9.69}$$

where the right-hand side is the *slowness* with which the curve of intersection moves across the surface, in the direction tangential to the surface and normal to the curve, while the left-hand side is the *component of the wave-slowness* in that direction.

The condition for wavefront continuity is that v_{t} is the same for the incident, refracted, and reflected portions of the wavefront;[28] that is, the right-hand side of (1.9.69) is the same for all three. In particular, if we use primed symbols for the refracted portion and unprimed for the incident portion, we have

$$s \sin\theta = s' \sin\theta', \tag{1.9.70}$$

which has the form of "**Snell's law**", except that the angles of incidence and refraction are defined in terms of the *wave-normal* directions. For the special case of **ordinary** refraction, in which the wave-normal directions are the ray directions, this law becomes applicable to the rays. Similarly, if we use double-primed symbols for the *reflected* portion of the wavefront, we have

$$s \sin\theta = s'' \sin\theta''. \tag{1.9.71}$$

For the special case of an isotropic medium (*ordinary* reflection), we have $s'' = s$, so that the law reduces to $\theta = \theta''$ ("the angle of incidence equals the angle of reflection"), and becomes applicable to the rays.[29]

[28] Thomas Young stated and exploited this condition as early as 1814, at least for the incident and refracted waves [29, p.263], but did not express it in terms of wave slowness.

[29] For some purposes it is useful to adopt a sign convention that introduces a minus sign into equation (1.9.71) and its corollaries; but I do not pursue that issue here.

Equation (1.9.69) with wavefront continuity (common v_t) implies that the wave slownesses have the same *tangential component*—i.e., the same *projection* on the plane tangential to the surface at the point of incidence. Hence, *if the incident, refracted, and reflected wave slownesses are represented by radii of the respective wave-slowness surfaces centered on the point of incidence, the endpoints of the radii lie on a common perpendicular to the tangent plane*. Moreover, *the radii give the wave-normal directions, while the normals to the wave-slowness surfaces at the endpoints give the respective ray directions*.

The "common perpendicular" rule was noted in the same paper by Hamilton [12, p. 144], in which he called the wave-slowness surface the "surface of components of normal slowness" [p. 142] or simply the *surface of components*.[30]

Having seen the usefulness of the wave-slowness surface, let us recast our basic equations in terms of wave slowness. Cross-multiplying both sides of (1.9.8) on the left by $\mathbf{s}$, we obtain

$$\mathbf{s} \times \mathbf{H} = \mathbf{s} \times (\mathbf{r} \times \mathbf{D}) = \mathbf{s}.\mathbf{D}\,\mathbf{r} - \mathbf{s}.\mathbf{r}\,\mathbf{D}\,. \tag{1.9.72}$$

But $\mathbf{s}.\mathbf{D} = 0$ (because $\mathbf{D}$ is tangential to the wavefront, therefore normal to $\mathbf{s}$), and $\mathbf{s}.\mathbf{r} = 1$ by (1.9.66), so we have

$$\mathbf{D} = -\mathbf{s} \times \mathbf{H}\,. \tag{1.9.73}$$

Similarly, cross-multiplying both sides of (1.9.5) on the left by $\mathbf{s}$, we obtain

$$\mathbf{s} \times \mathbf{E} = -\mathbf{s} \times (\mathbf{r} \times \mathbf{B}) = \mathbf{s}.\mathbf{r}\,\mathbf{B} - \mathbf{s}.\mathbf{B}\,\mathbf{r}\,. \tag{1.9.74}$$

But $\mathbf{s}.\mathbf{B} = 0$ (as $\mathbf{B}$ is tangential to the wavefront), and again $\mathbf{s}.\mathbf{r} = 1$, so we have

$$\mathbf{B} = \mathbf{s} \times \mathbf{E}\,. \tag{1.9.75}$$

Equations (1.9.73) and (1.9.75) have the same form as (1.9.5) and (1.9.8).[31]

Hence we can investigate $\mathbf{D} \times \mathbf{B}$ as we previously investigated $\mathbf{E} \times \mathbf{H}$: cross-multiplying (1.9.75) on the left by $\mathbf{D}$, or (1.9.73) on the right by $\mathbf{B}$, we confirm that $\mathbf{D} \times \mathbf{B}$ is in the direction of $\mathbf{s}$.[32]

---

[30] The summary of Hamilton's research in the *Report of the Third Meeting of the British Association for the Advancement of Science* [5, pp. 366–9] makes an obvious error, which I have corrected in the bibliographic entry [5].

[31] Note for skeptical specialists: In the case of sinusoidal oscillations with wave vector $\mathbf{k}$, we have $\mathbf{s} = \mathbf{k}/\omega$ in (1.9.73) and (1.9.75), which then become $\omega\mathbf{D} = -\mathbf{k} \times \mathbf{H}$ and $\omega\mathbf{B} = \mathbf{k} \times \mathbf{E}$, in agreement with equations (2.2) of Berry & Jeffrey [4].

[32] And adding the results gives $\mathbf{D} \times \mathbf{B} = (\frac{1}{2}\mathbf{E}.\mathbf{D} + \frac{1}{2}\mathbf{H}.\mathbf{B})\,\mathbf{s}$; that is, the Minkowski momentum density is the product of the total energy density and the wave slowness (cf. footnotes 21, 22).

Of course we still have relations (1.9.9), (1.9.10), (1.9.11), and (1.9.14). And on assumption 1.9.15 ($\mathbf{B} = \mu\mathbf{H}$), we still have theorem 1.9.19, in which we can now treat $\mathbf{s}_\mathrm{r}$ as interchangeable with $\mathbf{r}$, and $\mathbf{s}$ with $\mathbf{v}_\mathrm{n}$ (because the theorem concerns only directions, not magnitudes).

Making the same assumptions as for the ray-velocity surface, we would expect the wave-slowness surface derived from equations (1.9.73) and (1.9.75) to have the same form as the ray-velocity surface derived from the analogous equations (1.9.5) and (1.9.8), except that the principal dimensions a, b, c are replaced by their reciprocals. Indeed, multiplying (1.9.73) by μ gives

$$\begin{aligned}
\mu\mathbf{D} &= -\mathbf{s}\times\mu\mathbf{H} \\
&= -\mathbf{s}\times\mathbf{B} && \text{by 1.9.15} \\
&= -\mathbf{s}\times(\mathbf{s}\times\mathbf{E}) && \text{by (1.9.75)} \\
&= \mathbf{s}.\mathbf{s}\,\mathbf{E} - \mathbf{s}.\mathbf{E}\,\mathbf{s}\,;
\end{aligned}$$

that is,

$$\mu\mathbf{D} = s^2\mathbf{E} - \mathbf{s}.\mathbf{E}\,\mathbf{s}\,, \tag{1.9.76}$$

confirming that $\mathbf{D}$, $\mathbf{E}$, and $\mathbf{s}$ are coplanar (cf. theorem 1.9.19). Then, rewriting equations (1.9.29) and (1.9.30) in terms of the components of $\mathbf{E}$, we find that the electric field

$$\mathbf{E} = E_x\mathbf{i} + E_y\mathbf{j} + E_z\mathbf{k} \tag{1.9.77}$$

is associated with the displacement field

$$\mathbf{D} = \tfrac{1}{\mu a^2}E_x\mathbf{i} + \tfrac{1}{\mu b^2}E_y\mathbf{j} + \tfrac{1}{\mu c^2}E_z\mathbf{k}\,. \tag{1.9.78}$$

If we now let the wave slowness be

$$\mathbf{s} = X\mathbf{i} + Y\mathbf{j} + Z\mathbf{k}\,, \tag{1.9.79}$$

we can substitute (1.9.77), (1.9.78), and (1.9.79) into (1.9.76), obtaining

$$\begin{aligned}
\tfrac{1}{a^2}E_x\mathbf{i} + \tfrac{1}{b^2}E_y\mathbf{j} + \tfrac{1}{c^2}E_z\mathbf{k} = s^2(E_x\mathbf{i} + E_y\mathbf{j} + E_z\mathbf{k}) \\
-(XE_x + YE_y + ZE_z)(X\mathbf{i} + Y\mathbf{j} + Z\mathbf{k})
\end{aligned} \tag{1.9.80}$$

—which has the form of (1.9.33), except that a, b, and c are replaced by their reciprocals, and the components of $\mathbf{D}$ are replaced by the components of $\mathbf{E}$, and the magnitude and components of $\mathbf{r}$ are replaced by the magnitude and components of $\mathbf{s}$. The derivation of the wave-slowness surface then proceeds

as for the ray-velocity surface, leading to a surface of the same form with a, b, and c replaced by $1/a$, $1/b$, and $1/c$.

Hence, if the wave-slowness surface is scaled up by a factor b^2, that surface and the ray-velocity surface have the same circle of intersection with the (x, z) plane, namely *bRNb* in Fig. 1.17 (which shows one octant of the ray-velocity surface). Let the line *QNL* be tangential to *aQRc* at Q and to *bRNb* at N (this is how Q and N were chosen in Figs. 1.15 and 1.16). Then, by symmetry, the plane through that line and parallel to the y axis is tangential to the ray-velocity surface at Q and N, and normal to ON, so that *the vector ON is the wave-normal velocity corresponding to the vector ray velocities ON and OQ*. The axis ON is called the **binormal** axis (or the **optic axis**). In the direction of that axis, both sheets of the ray-velocity surface give the same wave-normal velocity, hence the same wave slowness, hence the same radius for the wave-slowness surface, so that N is a *diabolical point of the scaled wave-slowness surface*. There are four such points equidistant from O, on two binormal axes passing through O in the (x, z) plane; indeed, the word *biaxial* (or, in old literature, *biaxal*) originally referred to these axes rather than the biradial axes.

As ON is the wave-normal velocity for *one* ray velocity on the arc *aQRc* (namely OQ), the curve on the wave-velocity surface corresponding to that arc passes through N. Of course it also passes through the endpoints of that arc (because the ray-velocity surface is normal to its radii at those points) and does so at right angles to the axes (for symmetry and smoothness). On that information we can roughly sketch the wave-velocity curve *aNc* in the (x, z) plane. But we can sketch it more accurately with reference to the corresponding wave-slowness curve. Let the scaled wave-slowness magnitude be $S = b^2 s = b^2/v_\mathrm{n}$. Then $S/b = b/v_\mathrm{n}$; that is, in any direction, we have a geometric progression from the wave-normal velocity to the radius of the circle to the scaled wave slowness. Hence we can mark the intersections of the scaled wave-slowness surface with the x and z axes (points C and A), and complete the quarter-ellipse ANC, which is the *other* curve of intersection of the scaled wave-slowness surface with the (x, z) plane. Then we can exploit the same geometric progression in order to locate intermediate points on the wave-velocity curve *aNc*.

At a diabolical point, the wave-slowness surface has a cone of normals, which are ray directions; thus a wavefront normal to ON is associated with a cone of permitted ray directions. So if a ray, containing all polarizations, passes from an isotropic medium into the current (non-isotropic) medium, at such an angle that the refracted wave-normal direction is the binormal, then it will break into a hollow cone of rays; this is **internal conical refraction**.

The corresponding ray-velocity vectors (including ON and OQ) form a closed curve on the ray-velocity surface. These ray velocities have a common

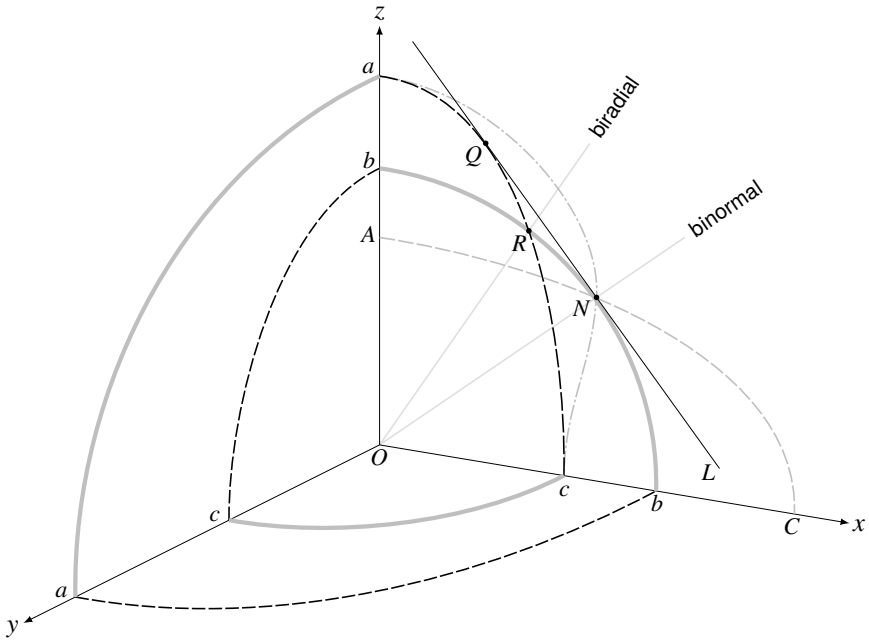

Fig. 1.17: Ray-velocity surface *RNbaaQRccb* and scaled wave-slowness curves *bRNb* & *ANC*.

wave-normal velocity, in magnitude and direction. The common magnitude means that the tangent planes to the ray-velocity surface at all points on the curve have the same normal distance from the origin, while the common direction—the binormal—means that the tangent planes are parallel. So the tangent planes are one: *the ray-velocity surface has a curve of contact with a tangent plane normal to the binormal axis*. That plane is parallel to the y axis and cuts the (x, z) plane at the line QNL.

Similarly, for the internal ray that suffers external conical refraction, the wave-slowness vectors form a closed curve on the wave-slowness surface. As these wave slownesses have a common ray slowness in the biradial direction, *the wave-slowness surface has a curve of contact with a tangent plane normal to the biradial axis*. The intersection of the tangent plane with the (x, z) plane (*not* shown in Fig. 1.17) is a line tangential to the circle *bRNb* at R, and to the ellipse *ANC* at an unmarked point.

Thus the range of internal ray directions for internal conical refraction includes the common internal wave-normal direction (ON) and revolves around the internal ray direction for *external* conical refraction (OR), while the range of internal wave-normal directions for *external* conical refraction includes the common internal ray direction (OR) and revolves around the internal wave-normal direction for internal conical refraction (ON).

The entire wave-slowness surface, having the same form as the ray-velocity surface, can be sketched in the same manner, except that the dashes in the ellipses (like the gray dashes on *ANC*) show the direction of **E**, not **D**. The binormal axis *ON* has the *direction in which the wave-normal velocity is independent of polarization*. For the wave-normal direction *ON*, the permitted directions of **E** are in the plane tangential to the ellipse *ANC* and normal to the (x, z) plane, because **E** can have a component normal to the circle *bRNb and* a component tangential to the ellipse *ANC*.

After internal conical refraction, the rays have a common wave-normal direction, which they retain as they propagate (in a uniform medium). If they exit the medium via a flat surface into another uniform medium, they still have a common wave-normal direction (because the wavefront remains planar). If the new medium is also isotropic (e.g., air or water), the parallel wave-normals correspond to parallel rays, so that the emergent beam is a hollow cylinder (not necessarily right-circular), from which the internal cone can be inferred.

Internal conical refraction was predicted by Hamilton in October 1832, in the same paper as the external form [12, s.29]. Lloyd, having confirmed the external form in December, observed the internal form and its polarization pattern early in the new year [21, p.239], by which time Hamilton had explained the external polarization and predicted its internal counterpart [16, p.156].

Concerning the polarizations, theorem 1.9.19 has the following corollaries, the first of which we have already seen:

- For the *ray-velocity* surface, the direction of **D** is tangential to the surface and in the plane of the radius and the normal—that is, normal to the wave-normal and in the plane of the ray and the wave-normal, respectively.
- For the *wave-slowness* surface, the direction of **E** is tangential to the surface and in the plane of the radius and the normal—that is, normal to the ray and in the plane of the wave-normal and the ray, respectively.
- In each case, the direction of the *other* electric vector (**E** or **D**) is normal to the radius and in the plane of the radius and the normal to the surface—that is, normal to the *other* direction of propagation (ray or wave-normal) and in the plane of the ray and the wave-normal.
- In each case, if the surface is two-sheeted, one propagation direction (ray or wave-normal) will generally give two polarizations (one per sheet).

In Fig. 1.16 (p.63), the curve of contact between the ray-velocity surface and the tangent plane is *NPQS*. (That is how points *P* and *S* were chosen, although the curve of contact was then described merely as the crest of a ridge.) At each point on that curve, the normal to the surface is parallel to *ON*. Hence, by the

first of the above corollaries, the direction of **D** at each point on the curve *NPQS* is toward *N*, so that that *a full turn about the closed curve corresponds to a half-turn in the polarization*—the law of internal conical polarization. This confirms the interpolation argument by which we initially sketched the polarizations at *P* and *S*. If we abandon tangency and allow the closed curve to converge on *R*, the ray directions converge on that of the internal ray for external conical refraction, which therefore shows the same polarization pattern.

Indeed, in Fig. 1.17 (p.75), at each point on the loop of contact between the wave-slowness surface and the tangent plane, the normal to the surface is parallel to *OR*. Hence, by the second of the above corollaries, the direction of **E** at each point on the loop is toward *R*, so that (again) a full turn about the loop corresponds to a half-turn in the polarization—the law of external conical polarization. If we abandon tangency and allow the loop to converge on *N*, the wave-normal directions converge on that of the internal wave-normal for internal conical refraction, which therefore shows the same polarization pattern.

Now reconsider the *uniaxial* case, in which the ray-velocity or wave-slowness surface consists of a spherical sheet and a spheroidal sheet, centered on the origin and making contact at their "poles". By the first or second of the above corollaries, the polarization (of **D** or **E**) on the spheroidal sheet is along the lines of "longitude". To account for the different velocity or slowness, the polarization on the spherical sheet must be different; hence, by the axial symmetry, it must be along the lines of "latitude". Thus, for a given direction from the origin, the polarizations of the two sheets are *orthogonal*.

The proof of this result for the *biaxial* case is slightly more sophisticated. Let **d** be the projection of **D** on the plane normal to the ray velocity **r**. Then **d** is given by **D** minus the component of **D** in the direction of **r**; that is,

$$\mathbf{d} = \mathbf{D} - \mathbf{D}.\hat{\mathbf{r}}\,\hat{\mathbf{r}}, \tag{1.9.81}$$

where, as usual, $\hat{\mathbf{r}}$ is the unit vector in the direction of **r**. Now let

$$\hat{\mathbf{r}} = \alpha\mathbf{i} + \beta\mathbf{j} + \gamma\mathbf{k} = \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix}, \tag{1.9.82}$$

so that

$$\alpha^2 + \beta^2 + \gamma^2 = 1 \tag{1.9.83}$$

since $|\hat{\mathbf{r}}| = 1$; in other words, let α, β, γ be the *direction cosines* of **r**. Then if we put (say) $\mathbf{D} = D_x\mathbf{i} + D_y\mathbf{j} + D_z\mathbf{k}$, we can write (1.9.81) in the matrix form

$$\mathbf{d} = \mathbf{P}\mathbf{D} \tag{1.9.84}$$

where

$$\mathbf{P} = \begin{bmatrix} 1-\alpha^2 & -\alpha\beta & -\alpha\gamma \\ -\alpha\beta & 1-\beta^2 & -\beta\gamma \\ -\alpha\gamma & -\beta\gamma & 1-\gamma^2 \end{bmatrix} . \tag{1.9.85}$$

So **P**, the matrix of the projection, is **real**, meaning that its elements are real; and it is **symmetric**, meaning that it is unchanged when **transposed** (flipped about the diagonal, so that the rows become columns and vice versa).

The derivation of (1.9.84) and (1.9.85) is quite general. So the projection of **E** on the plane normal to **r** is **PE**. In this case, however, **E** is already in the plane normal to **r**, so that its projection is equal to itself:

$$\mathbf{PE} = \mathbf{E} . \tag{1.9.86}$$

And the projection of $\hat{\mathbf{r}}$ on the plane normal to **r** is the zero vector:

$$\mathbf{P}\hat{\mathbf{r}} = \mathbf{0} . \tag{1.9.87}$$

The last result can be confirmed algebraically using (1.9.83), but the geometric argument is simpler. In general, if

$$\mathbf{Ax} = \lambda\mathbf{x} , \tag{1.9.88}$$

where **A** is a matrix and **x** is a vector and λ is a scalar, we say that **x** is an **eigenvector** of **A** and that λ is the associated **eigenvalue** of **A**. It is easily shown that any scalar multiple of **x** is also an eigenvector, with the same eigenvalue. In this terminology, equations (1.9.86) and (1.9.87) tell us that the matrix **P** has an eigenvector **E** with eigenvalue 1, and an eigenvector $\hat{\mathbf{r}}$ with eigenvalue 0. Notice that the two eigenvalues are real and distinct and that the associated eigenvectors are orthogonal. Indeed, a well-known theorem of linear algebra (the generalized study of linear transformations of vectors) says that *if a matrix is real and symmetric, its eigenvalues are real and the eigenvectors belonging to distinct eigenvalues are orthogonal*. (*N.B.*: The theorem does *not* guarantee that the eigenvalues are distinct, but tells us something *if* they are.)

So, to show that the permitted polarizations for a particular ray or wave-normal direction are orthogonal, we might look for a real, symmetric matrix of which **E** or **D** is an eigenvector. Matrix **P**, although real and symmetric, is evidently unsuitable, because *any* **E** in the plane normal to **r** satisfies (1.9.86), so that **P** has a whole plane of eigenvectors with the same eigenvalue.[33] We

[33] This situation is caused by a *repeated eigenvalue*; from the definition (1.9.88), it is easily shown that if two or more eigenvectors have the same eigenvalue, any *linear combination* (sum of scalar multiples) of those eigenvectors is also an eigenvector with the same eigenvalue.

should expect **P** to be too permissive because it fails to account for the relation between **E** and **D**, which was combined with (1.9.20) to find the ray-velocity surface. That relation, namely that the electric field $\mathbf{E} = E_x\mathbf{i} + E_y\mathbf{j} + E_z\mathbf{k}$ gives the displacement field $\mathbf{D} = \epsilon_x E_x\mathbf{i} + \epsilon_y E_y\mathbf{j} + \epsilon_z E_z\mathbf{k}$, can be written

$$\mathbf{D} = \boldsymbol{\epsilon}\mathbf{E}, \tag{1.9.89}$$

where

$$\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_x & 0 & 0 \\ 0 & \epsilon_y & 0 \\ 0 & 0 & \epsilon_z \end{bmatrix}. \tag{1.9.90}$$

If we now project both sides of (1.9.20) onto the plane normal to **r**, we obtain

$$\begin{aligned} \mathbf{E} &= \mu r^2 \mathbf{d} \\ &= \mu r^2 \mathbf{P}\mathbf{D} && \text{by (1.9.84)} \\ &= \mu r^2 \mathbf{P}\boldsymbol{\epsilon}\mathbf{E} && \text{by (1.9.89)} \\ &= \mu r^2 \mathbf{P}\boldsymbol{\epsilon}\mathbf{P}\mathbf{E} && \text{by (1.9.86)}; \end{aligned}$$

that is,

$$(\mathbf{P}\boldsymbol{\epsilon}\mathbf{P})\,\mathbf{E} = \frac{1}{\mu r^2}\,\mathbf{E}. \tag{1.9.91}$$

So **E** is an eigenvector of $\mathbf{P}\boldsymbol{\epsilon}\mathbf{P}$ with eigenvalue $1/(\mu r^2)$. We already know that this eigenvalue is real, and that in any direction except the biradials, there are two values of r, hence two *distinct* eigenvalues. And clearly the matrices are real. So, to show that the eigenvectors **E** for the different eigenvalues are orthogonal according to the theorem, it suffices to show that $\mathbf{P}\boldsymbol{\epsilon}\mathbf{P}$ is symmetric. We could multiply it out and see, but there is an easier way. In general, the transpose of a product of matrices is the product of their transposes in reverse order. In this case, transposing the individual matrices does not change them, because **P** and $\boldsymbol{\epsilon}$ are symmetric; and reversing the order also has no effect. So the product $\mathbf{P}\boldsymbol{\epsilon}\mathbf{P}$ is equal to its transpose; that is, $\mathbf{P}\boldsymbol{\epsilon}\mathbf{P}$ is symmetric—as required.

[From (1.9.87) it is easily verified that $\mathbf{P}\boldsymbol{\epsilon}\mathbf{P}$ has a third eigenvector $\hat{\mathbf{r}}$, with eigenvalue 0. This eigenvector is orthogonal to the other two, in accordance with the theorem and with the condition that $\mathbf{E} \perp \mathbf{r}$ (1.9.9).]

So, from (1.9.20), we have shown that *the permitted directions of* **E** *for a given ray direction are orthogonal* (for different ray speeds). Similarly, from (1.9.76), or because the wave-slowness surface is related to **D** as the ray-velocity surface to **E**, *the permitted directions of* **D** *for a given wave-normal direction are orthogonal* (for different wave slownesses).

From the orthogonality of the two directions of **E** for the same ray direction, together with the requirement that **D** is tangential to the ray-velocity surface and in the plane of **E** and the ray, we can deduce the polarization pattern on the inner sheet of Fig. 1.16 (p. 63) from the pattern on the outer sheet, confirming what we would expect from interpolation.

Finally, for comprehensiveness, let us extract a few results on the shapes and sizes of the curves of contact between the unit-wave and its tangent planes.

By the symmetry about the (x, z) plane, the equation of each plane with a curve of contact is of the form

$$z = h - mx \,. \tag{1.9.92}$$

For the quadrant shown in Fig. 1.16, the constants h and m are positive. In Fresnel's equation for the unit-wave (1.9.39), let us consider z^2 as a function of x^2 and y^2. Usually, if we wanted to differentiate that equation (partially) w.r.t. y^2, we would treat x^2 as a constant. But at points of tangency with the plane (1.9.92), we can also treat z^2 as a constant, because z is independent of y^2 in (1.9.92). And if both x^2 and z^2 are treated as constants, the derivative of r^2 is simply the derivative of y^2, namely 1 (w.r.t. y^2, not y). With all these simplifications, differentiating (1.9.39) w.r.t. y^2 gives

$$r^2b^2 + a^2x^2 + b^2y^2 + c^2z^2 - b^2(a^2{+}c^2) = 0 \,. \tag{1.9.93}$$

Putting $r^2 = x^2 + y^2 + z^2$ and rearranging, we obtain

$$(a^2{+}b^2)x^2 + 2b^2y^2 + (b^2{+}c^2)z^2 = b^2(a^2{+}c^2) \,. \tag{1.9.94}$$

The surface described by this is obviously an ellipsoid. Its intersection with the plane (1.9.92), with the appropriate values of h and m, is a simple closed curve, which must be the curve of tangency between the unit-wave and the plane.[34]

Let us therefore find h and m. We know that the tangent plane is normal to ON in Fig. 1.17 (p. 75) at point N, which is the diabolical point on the scaled wave-slowness surface. The corresponding point on the ray-velocity surface is R, whose coordinates we already know from equations (1.9.48). So, to find the coordinates of N from the coordinates of R, we replace a, b, and c by their reciprocals, multiply by the scale factor b^2, and simplify, obtaining

$$x_N = b\Big(\frac{a^2-b^2}{a^2-c^2}\Big)^{1/2} \;\; ; \;\; z_N = b\Big(\frac{b^2-c^2}{a^2-c^2}\Big)^{1/2} . \tag{1.9.95}$$

[34] Equations (1.9.92) and (1.9.94) by themselves do not prove the existence of a curve of tangency, because (1.9.94) is merely a *necessary* condition of tangency. But, having *previously* established that such a curve exists, we now know that (1.9.94) imposes constraints on it.

Then $m = x_N/z_N$; that is,

$$m = \left(\frac{a^2-b^2}{b^2-c^2}\right)^{1/2}. \tag{1.9.96}$$

As the plane (1.9.92) passes through N, we can substitute (1.9.95) and (1.9.96) into (1.9.92) and solve for h, obtaining

$$h = b\left(\frac{a^2-c^2}{b^2-c^2}\right)^{1/2}. \tag{1.9.97}$$

(We get the same result with less algebra if we mark h and z_N on the z axis in Fig. 1.17, obtaining two similar triangles in which $h/b = b/z_N$.)

Now we can find the intersection of (1.9.92) and (1.9.94). Squaring and expanding (1.9.92) and substituting from (1.9.96) and (1.9.97), we get

$$z^2 = \frac{(a^2-b^2)x^2 - 2b\sqrt{(a^2-b^2)(a^2-c^2)}\,x + b^2(a^2-c^2)}{b^2-c^2}\,. \tag{1.9.98}$$

Substituting this into (1.9.94), multiplying through by (b^2-c^2), regrouping terms, and canceling the common factor $2b$, we obtain,

$$b(a^2-c^2)x^2 - (b^2+c^2)\sqrt{(a^2-b^2)(a^2-c^2)}\,x + b(b^2-c^2)y^2 = bc^2(b^2-a^2), \tag{1.9.99}$$

which is the projection of the curve of tangency (*NPQS* in Fig. 1.16) on the (x, y) plane. To describe the curve itself, let the coordinates within the tangent plane be (ξ, y), where ξ is measured from the (y, z) plane in the direction parallel to the line *QNL* (Fig. 1.17), so that

$$\xi = x\,(1+m^2)^{1/2}. \tag{1.9.100}$$

Substituting from (1.9.96) and solving for x, we find

$$x = \left(\frac{b^2-c^2}{a^2-c^2}\right)^{1/2}\xi\,. \tag{1.9.101}$$

If we substitute this into (1.9.99), divide through by the coefficient of ξ^2, and "complete the square" on the terms in ξ^2 and ξ, we can put the remaining constant terms over a common denominator, obtaining

$$\left[\xi - \frac{b^2+c^2}{2b}\left(\frac{a^2-b^2}{b^2-c^2}\right)^{1/2}\right]^2 + y^2 = \frac{a^2b^4 - 2a^2b^2c^2 + a^2c^4 - b^6 + 2b^4c^2 - b^2c^4}{4b^2(b^2-c^2)}\,, \tag{1.9.102}$$

which shows that the "closed curve" of tangency is a *circle*! In the numerator on the right-hand side, the sum of the first three terms has a factor a^2, and

the sum of the remaining terms has a factor $-b^2$, and in each case the other factor is $(b^2-c^2)^2$. Hence we can factor the numerator and cancel one factor, simplifying the equation of the circle to

$$\left[\xi - \frac{b^2+c^2}{2b}\left(\frac{a^2-b^2}{b^2-c^2}\right)^{1/2}\right]^2 + y^2 = \frac{(a^2-b^2)(b^2-c^2)}{4b^2} \,. \tag{1.9.103}$$

The diameter is twice the square root of the right-hand side, i.e.

$$QN = \frac{\sqrt{(a^2-b^2)(b^2-c^2)}}{b} \,, \tag{1.9.104}$$

where QN refers to the dimension in Fig. 1.16.

Note that the perpendicular dimension PS (in Fig. 1.16) is *not* generally an exact diameter. Points P and S have been defined as being in the plane of R and the y axis, so that PS is a chord in that plane. If this chord were a diameter, it would pass through the midpoint of the other diameter QN, and that midpoint would therefore be on the line OR, which is *not* always true—as may be verified by sketching the circle $bRNb$ and the ellipse $aQRc$ for extreme cases in which a is very large or c very small compared with b.

Hamilton uses a more sophisticated coordinate transformation to find the circle of tangency [12, p.134], which leads to the prediction of internal conical refraction [p.136]. But, as he acknowledges on the last page of his paper [12], the order of presentation of his results is not the order of discovery. While investigating a different problem, he initially came to the diabolical points of the wave-slowness surface, which implied curves of contact on the ray-velocity surface. From the relation between the surfaces, he concluded that the latter also had diabolical points, which implied curves of contact on the former. His discovery that the "curves" are circles came later [5, pp. 368–9].

For the ray-velocity surface, the set of lines through the origin and the circle of tangency defines a cone, and one of those lines is the binormal axis, which is normal to the tangent plane. Similarly, for the wave-slowness surface, one of the lines comprising the cone is the biradial axis, which is normal to the plane of the circle of tangency. In each case, the axis of the cone is therefore *not* normal to the circular base. Hence the cone of rays for internal conical refraction, and the cone of internal wave-normals for external conical refraction, are *circular, but not right-circular*.

However, for each cone, the right angle between the circular base and *one* of the generating lines yields a formula for the opening angle of the cone in the (x, z) plane. In Fig. 1.17, QNO is the right angle and $ON = b$, and we know QN from (1.9.104); so the tangent of the opening angle is

$$\tan \angle NOQ = \frac{\sqrt{(a^2-b^2)(b^2-c^2)}}{b^2} \,. \tag{1.9.105}$$

At point R in Fig. 1.17, the acute angle between the circle and the ellipse, loosely called $\angle aRb$, is also the angle between their normals in the (x, z) plane, which is the opening angle of the corresponding cone of the wave-slowness surface. So, to find its tangent, we replace a, b, and c in (1.9.105) by their reciprocals and simplify, obtaining

$$\tan\angle aRb = \frac{\sqrt{(a^2-b^2)(b^2-c^2)}}{ac} \,. \tag{1.9.106}$$

This is also twice the tangent of half the angle between the arcs $aPR,{-}c$ and $cRS,{-}a$ in Fig. 1.16; compare equation (1.9.51).

The tangent of the angle between the z axis and the biradial axis is x_R/z_R, which may be found from equations (1.9.48):

$$\tan\angle zOR = \frac{c}{a}\left(\frac{a^2-b^2}{b^2-c^2}\right)^{1/2}. \tag{1.9.107}$$

We have already done the same for the binormal axis in (1.9.96):

$$\tan\angle zON = \left(\frac{a^2-b^2}{b^2-c^2}\right)^{1/2}. \tag{1.9.108}$$

The *angle between the biradial and binormal axes* is $\angle RON = \angle zON - \angle zOR$. Hence, using the identity $\tan(u-v) = (\tan u - \tan v)/(1+\tan u \tan v)$, we find

$$\tan\angle RON = \frac{\sqrt{(a^2-b^2)(b^2-c^2)}}{b^2+ac} \,. \tag{1.9.109}$$

Comparing this with (1.9.105) and (1.9.106), we have

$$\cot\angle RON = \cot\angle NOQ + \cot\angle aRb \,. \tag{1.9.110}$$

Taking reciprocals, we see that *the tangent of the angle between the biradial and binormal axes is the reciprocal of the sum of the reciprocals of the tangents of the opening angles of the ray cone and the wave-normal cone*. More concisely, the tangent of the angle between the axes is *half the harmonic mean* of the tangents of the opening angles. Hence it is close to half the *arithmetic* mean if the tangents are not too different (that is, if b^2 is comparable to ac); this approximation applies to the *tangents* of the angles, but extends to the angles themselves if they are small.

The subtractions in (1.9.105) to (1.9.109) magnify the percent uncertainties in the principal speeds a, b, and c. Notice, however, that the angle between the biradial and binormal axes, although obtainable by subtracting two of the other angles, is no more sensitive to the differences between principal speeds.

The **refractive index** is defined as $n = s/s_0$, where s_0 is the wave slowness in a uniform isotropic *reference medium* (usually air or a vacuum). If v_0 is the isotropic speed of light in the reference medium, we can successively put $s_0 = 1/v_0$ and $s = 1/v_n$, obtaining three equivalent definitions:

$$n = s/s_0 = v_0 s = v_0/v_n \,. \tag{1.9.111}$$

While the third definition is the most widely known, the first and second are more instructive: the refractive index is the *normalized wave slowness*, or the scaled wave slowness whose scale factor is the speed of light in the reference medium. If we scale the wave-slowness *surface* by the same factor, we get the surface whose "distance" from the origin in any direction is the refractive index in that direction; this is naturally called the **index surface**.

The refractive indices in the x, y, and z directions (in which $\mathbf{D} \parallel \mathbf{E}$) are called the **principal refractive indices** (or simply **principal indices**) and are respectively given by $n_a = v_0/a$, $n_b = v_0/b$, and $n_c = v_0/c$, so that $n_a < n_b < n_c$ in the biaxial case.[35] Hence

$$a = v_0/n_a \;\;;\;\; b = v_0/n_b \;\;;\;\; c = v_0/n_c \,. \tag{1.9.112}$$

Making these substitution in equations (1.9.105) to (1.9.109), we can express the tangents in terms of the principal indices (and because the angles are obviously insensitive to scale, we can take $v_0 = 1$ for convenience). The results are

$$\tan \angle NOQ = \frac{\sqrt{(n_c{}^2 - n_b{}^2)(n_b{}^2 - n_a{}^2)}}{n_a n_c} \tag{1.9.113}$$

$$\tan \angle aRb = \frac{\sqrt{(n_c{}^2 - n_b{}^2)(n_b{}^2 - n_a{}^2)}}{n_b{}^2} \tag{1.9.114}$$

$$\tan \angle zOR = \left(\frac{n_b{}^2 - n_a{}^2}{n_c{}^2 - n_b{}^2}\right)^{1/2} \tag{1.9.115}$$

$$\tan \angle zON = \frac{n_c}{n_a}\left(\frac{n_b{}^2 - n_a{}^2}{n_c{}^2 - n_b{}^2}\right)^{1/2} \tag{1.9.116}$$

$$\tan \angle RON = \frac{\sqrt{(n_c{}^2 - n_b{}^2)(n_b{}^2 - n_a{}^2)}}{n_b{}^2 + n_a n_c} \,. \tag{1.9.117}$$

To express the general law of refraction in terms of refractive indices, we divide (1.9.70) through by v_0, obtaining

$$n \sin\theta = n' \sin\theta' \,, \tag{1.9.118}$$

[35] Other widely-used symbols for the principal indices include n_α, n_β, n_γ, and n_1, n_2, n_3.

where θ and θ' are the angles of incidence and refraction of the *wave-normals*, and n and n' are the refractive indices of the two media in the wave-normal directions. Hence the law is applicable to the rays on the condition that the rays are normal to the wavefronts. We have seen that in a biaxial birefringent medium, this condition holds in each coordinate plane provided that the electric polarization is normal to the plane. So, in that plane, with that polarization, ray refraction between that medium and an isotropic medium satisfies (1.9.118) if the refractive surface is normal to that plane. Using this fact, one can measure the principal refractive indices.

Lloyd, in his experimental search for conical refraction, used a crystal of *aragonite* (another form of $CaCO_3$), whose principal indices had recently been found by Rudberg [21, pp. 241, 255, 257] to be

$$n_a = 1.5326 \;\; ; \;\; n_b = 1.6863 \;\; ; \;\; n_c = 1.6908 \; .$$

These values yield $NOQ = 1.916^\circ$ (for the internal ray cone), $aRb = 1.747^\circ$ (for the internal wave-normal cone), $zOR = 80.059^\circ$ (leaving 9.941° between the biradial and x axes), $zON = 80.973^\circ$ (leaving 9.027° between the binormal and x axes), and $RON = 0.914^\circ$ (whereas half the mean of the two cone angles is 0.916°). So the opening angles of the internal cones are less than 2° (though the internal wave-normal cone gives a somewhat wider external ray cone in air).

As of 2007, according to Berry & Jeffrey [4, pp. 24, 46], the medium with the largest known angle of conical refraction is *naphthalene*, whose principal indices are $n_a = 1.525$, $n_b = 1.722$, and $n_c = 1.945$. For these values, the opening angles NOQ and aRb are both close to 13.7°. So it is finally revealed that Figs. 1.16 and 1.17, for clarity, grossly exaggerate the dissimilarities between the principal speeds a, b, and c, causing the illustrated angles NOQ and aRb to be much larger than in any real material.

It is said that the discovery of conical refraction was the first occasion in the history of science on which a new phenomenon, *qualitatively* different from anything previously observed or suspected, was predicted by mathematics and confirmed by experiment. The significance of it was not lost on Lloyd; concerning the predicted external and internal cones, he remarked:

> Here, then, are two singular and unexpected consequences of the undulatory theory, not only unsupported by any facts hitherto observed, but even opposed to all the analogies derived from experience. If confirmed by experiment, they would furnish new and almost convincing proofs of the truth of that theory... [16, p.147].

In short, if the cones duly appeared, the proponents of the wave theory of light would get their smoking gun. And they got it.

Bibliography

[1] S. M. Barnett and R. Loudon, "The enigma of optical momentum in a medium", *Philosophical Transactions of the Royal Society A*, vol. 368, no. 1914 (Mar. 13, 2010), pp. 927–939; rsta.royalsocietypublishing.org/content/368/1914/927.

[2] E. Bartholin, *Experimenta Crystalli Islandici Disdiaclastici: Quibus mira et insolita Refractio detegitur*, Copenhagen ("Hafniæ"): D. Paulli, 1669, books.google.com/books?id=F7RAAAAAcAAJ; translated by W. Brandt as *Experiments with the Double Refracting Iceland Crystal: Which led to the discovery of a Marvelous and Strange Refraction*, Westtown, Pa., 1959; translated by T. Archibald as *Experiments on Birefringent Icelandic Crystal*, Copenhagen: Danish National Library of Science and Medicine, 1991.

[3] E. Bartholin (ed. H. Oldenburg?), "An Accompt of sundry Experiments made and communicated by that Learn'd Mathematician, Dr. Erasmus Bartholin, upon a Chrystal-like Body, sent to him out of Island", *Philosophical Transactions of the Royal Society*, vol. 5, no. 67 (1670/71), pp. 2039–2048; rstl.royalsocietypublishing.org/content/5/57-68/2039.

[4] M.V. Berry and M.R. Jeffrey, "Conical diffraction: Hamilton's diabolical point at the heart of crystal optics", in E. Wolf (ed.), *Progress in Optics*, vol. 50, Amsterdam: Elsevier, 2007, pp. 13–50; michaelberryphysics.files.wordpress.com/2013/07/berry400.pdf.

[5] British Assoc. for the Advancement of Science, *Report of the Third Meeting of the British Association for the Advancement of Science* (held at Cambridge in 1833), London: J. Murray, 1834; archive.org/details/reportofthirdmee34lond. *Correction:* On p. 367, "the radius vector of the Huygenian spheroid, and... the undulatory velocity" should be (e.g.) "the inverse of the radius vector of the Huygenian spheroid, and... the reciprocal of the undulatory velocity".

[6] J. Z. Buchwald, "Experimental investigations of double refraction from Huygens to Malus", *Archive for History of Exact Sciences*, vol. 21, no. 4 (Dec. 1980), pp. 311–373.

[7] J. Z. Buchwald, *The Rise of the Wave Theory of Light: Optical Theory and Experiment in the Early Nineteenth Century*, University of Chicago Press, 1989.

[8] A. J. de Witte, "Equivalence of Huygens' principle and Fermat's principle in ray geometry", *American Journal of Physics*, vol. 27, no. 5 (May 1959), pp. 293–301. *Erratum*: In Fig. 7(b), each instance of "ray" should be "normal" (noted in vol. 27, no. 6, p. 387).

[9] F. J. Dijksterhuis, *Lenses and Waves: Christiaan Huygens and the Mathematical Science of Optics in the Seventeenth Century* (doctoral thesis), University of Twente, 1999; doc.utwente.nl/33764. (See also the book with the same author and title, Dordrecht: Kluwer Academic Publishers, 2004.)

[10] E. Frankel, "The search for a corpuscular theory of double refraction: Malus, Laplace and the price [sic] competition of 1808", *Centaurus*, vol. 18, no. 3 (Sep. 1974), pp. 223–245.

[11] E. Frankel, "Corpuscular optics and the wave theory of light: The science and politics of a revolution in physics", *Social Studies of Science*, vol. 6, no. 2 (May 1976), pp. 141–184.

[12] W. R. Hamilton, "Third supplement to an essay on the theory of systems of rays", *Transactions of the Royal Irish Academy*, vol. 17, pp. v–x, 1–144, read Jan. 23 and Oct. 22, 1832; jstor.org/stable/30078785 (author's introduction dated June 1833; volume started 1831(?), completed 1837).

[13] T. Hobbes, *A Minute or first Draught of the Optiques*, Paris, 1646; British Library: Harleian Manuscript No. 3360, www.bl.uk/manuscripts/FullDisplay.aspx?ref=Harley_MS_3360.

[14] C. Huygens, *Treatise on Light* (Leiden: Van der Aa, 1690), tr. S.P. Thompson, London: Macmillan, 1912; archive.org/details/treatiseonlight031310mbp. See also "Errata in various editions of Huygens' *Treatise on Light*" at www.grputland.com or grputland.blogspot.com, June 2016.

[15] C. Huygens, *Oeuvres Complètes* (22 vols.), The Hague: Dutch Society of Sciences / M. Nijhoff, 1888–1950.

[16] H. Lloyd, "On the phenomena presented by light in its passage along the axes of biaxal crystals", *Transactions of the Royal Irish Academy*, vol. 17, pp. 145–157, read Jan. 28, 1833; jstor.org/stable/30078786 (volume started 1831(?), completed 1837).

[17] J. Lohne, "Thomas Harriott (1560–1621): The Tycho Brahe of optics", *Centaurus*, vol. 6, no. 2 (June 1959), pp. 113–121.

[18] J. Lohne, "The fair fame of Thomas Harriott: Rigaud versus Baron von Zach", *Centaurus*, vol. 8, no. 1 (Mar. 1963), pp. 69–84.

[19] J. G. Lunney and D. Weaire, "The ins and outs of conical refraction", *Europhysics News*, vol. 37, no. 3 (May–June 2006), pp. 26–29; doi.org/10.1051/epn:2006305.

[20] I. Newton, *Opticks* (4th Ed., London, 1730), with Foreword by A. Einstein and Introduction by E.T. Whittaker (London: George Bell & Sons, 1931), with Preface by I. B. Cohen and Analytical Table of Contents by D.H.D. Roller, Mineola, NY: Dover, 1952.

[21] J. G. O'Hara, "The prediction and discovery of conical refraction by William Rowan Hamilton and Humphrey Lloyd (1832–1833)", *Proceedings of the Royal Irish Academy, Section A: Mathematical and Physical Sciences*, vol. 82A, no. 2 (1982), pp. 231–257.

[22] G. R. Putland, "The observation by Huygens that should have discredited Newton's 'rule' for the extraordinary refraction of calcite", www.grputland.com or grputland.blogspot.com, Oct. 2016.

[23] R. Rashed, "A pioneer in anaclastics: Ibn Sahl on burning mirrors and lenses", *Isis*, vol. 81, no. 3 (Sep. 1990), pp. 464–491.

[24] H. G. J. Rutten and M.A.M. van Venrooij, *Telescope Optics: A Comprehensive Manual for Amateur Astronomers*, Richmond, VA: Willmann–Bell, 1988 (fifth printing, 2002).

[25] G. Sarton, "Discovery of conical refraction by William Rowan Hamilton and Humphrey Lloyd (1833)", *Isis*, vol. 17, no. 1 (1932), pp. 154–170.

[26] J.A. Schuster, "Descartes *opticien*: The construction of the law of refraction and the manufacture of its physical rationales, 1618–29", in S. Gaukroger, J.A. Schuster, and J. Sutton (eds.), *Descartes' Natural Philosophy*, London: Routledge, 2000, pp. 258–312.

[27] A.E. Shapiro, "Kinematic optics: A study of the wave theory of light in the seventeenth century", *Archive for History of Exact Sciences*, vol. 11, no. 2/3 (June 1973), pp. 134–266.

[28] R. H. Silliman, "Fresnel, Augustin Jean", *Complete Dictionary of Scientific Biography*, vol. 5, Detroit: Charles Scribner's Sons, 2008, pp. 165–171.

[29] T. Young (ed. G. Peacock), *Miscellaneous Works of the Late Thomas Young*, vol. I, London: J. Murray, 1855; archive.org/details/miscellaneouswo01youngoog.

[30] A. Ziggelaar, "The sine law of refraction derived from the principle of Fermat—prior to Fermat? The theses of Wilhelm Boelmans S.J. in 1634", *Centaurus*, vol. 24, no. 1 (Sep. 1980), pp. 246–262.