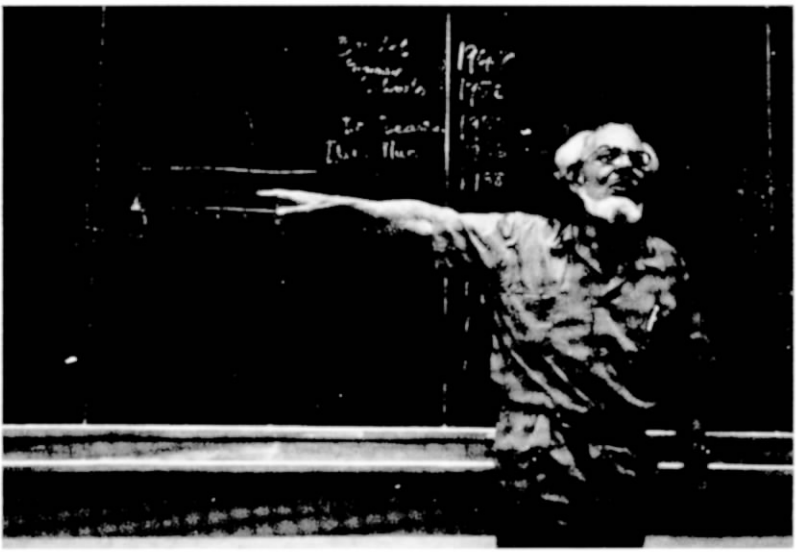
COURSE 7

SINGULARITIES IN WAVES AND RAYS

Michael BERRY

H. H. Wills Physics Laboratory,
Tyndall Avenue, Bristol, BS8 1TL, U.K.



Contents

R. Balian et al., eds.
Les Houches, Session XXXV, 1980 – Physique des Défauts / Physics of Defects

1. Introduction

I plan to discuss singularities in scalar linear waves. It might seem that this is too restricted a context to generate a substantial body of worthwhile theory, in view of the fact that most of the singularities talked about nowadays occur in much more complicated systems such as solids, liquid crystals or non-linear multicomponent fields. In fact, however, scalar linear wave equations conceal a rich variety of singular structures, worth studying for two reasons. Firstly, they describe phenomena in optics, acoustics, hydrodynamics and quantum mechanics that are intrinsically interesting and sometimes dramatic. Secondly, they provide an excellent illustration of the fact, not widely appreciated, that a given mathematical theory can possess singularities of very different sorts and which may manifest themselves on different scales.

Singularities, when considered in the modern way as geometric rather than algebraic structures, are *morphologies*, that is form rather than matter; and waves are morphologies too (it is not matter but form that moves with a wave). Therefore singularities of waves represent a double abstraction – forms of forms, as it were – and so it comes as something of a surprise to learn that they represent observable phenomena in a very direct way, as I shall demonstrate with many illustrations.

In the past, an exposition of wave theory might have consisted of a list of the (few) problems for which wave functions can be expressed exactly in terms of known "special" functions, together with a description of methods (perturbation, variation or asymptotic) for approximating wave functions in those cases (most) where the exact solution is unknown. Nowadays the objectives are completely different. As the philosophers say, there has been a shift to a new "paradigm", in which it is recognised that waves exist in space–time and so possess geometric aspects, whose dominant features are their singularities. In classifying these, the most important new idea, due to Thom [1], is *stability under perturbation* ("structural stability") which, in this context, means we concentrate on singularities occurring in *typical* ("generic") waves, rather than the often misleading cases for which "exact" solutions can be found. These lectures

are set firmly in the new paradigm. An important underlying theme, which I shall make explicit from time to time, is the contrast and interplay between singularities occurring on different scales.

To begin, in section 2, I shall describe the most delicate singularities of waves, namely dislocations in wavefronts; these are manifestations of a wave's phase, and are naturally (but not completely) classified in terms of index or winding number. Then in section 3 I shall describe coarser singularities, characteristic of the shortwave limit, namely caustics (focal surfaces) of ray families; these are classified by catastrophe theory [1, 2]. On the most "macroscopic" scale, caustics can link up into networks (subsection 3.5) whose catastrophe detail cannot be resolved and which constitute a new morphology. On finer scales, the effect of finite wavelength is to cause caustics to be decorated with "diffraction catastrophes". These striking patterns will be described in section 4; as their most "microscopic" features they contain a skeleton of wave front dislocations.

Some of this material has been reviewed before [3–5, 32], but these lectures will present it in a way that is different and, I hope, more accessible.

2. Wavefront dislocations

2.1. Singularities of phase

Consider a travelling wave that has been diffracted or refracted or reflected or scattered so as to have non-trivial structure in space–time. We wish to describe such a wave by a complex scalar function $\psi(\boldsymbol{r}, t)$ of position $\boldsymbol{r}$ and time t, with amplitude $\rho(\boldsymbol{r}, t)$ and phase $\chi(\boldsymbol{r}, t)$, defined by

$$\psi = \rho e^{i\chi}. \tag{2.1}$$

Wavefronts are defined as the contour surfaces of phase; in particular, wavecrests and wavetroughs are defined by

$$\left.\begin{aligned} \chi &= 0 \bmod 2\pi \,(\text{crests}) \\ \chi &= \pi \bmod 2\pi \,(\text{troughs}) \end{aligned}\right\}. \tag{2.2}$$

This mathematically-convenient description in terms of a complex wave function is not easy to relate directly to experiment where, for all waves except those in quantum mechanics only real functions $\psi_R(\boldsymbol{r}, t)$ can be observed. One method is to first write ψ_R as a Fourier transform

over time, i.e.

$$\psi_R(\boldsymbol{r}, t) = \int_{-\infty}^{\infty} d\omega \bar{\psi}(\boldsymbol{r}, \omega) e^{i\omega t}, \tag{2.3}$$

with the reality condition

$$\bar{\psi}(-\omega) = \bar{\psi}^*(\omega). \tag{2.4}$$

Then a complex function whose real part is ψ_R can be produced by retaining only positive frequencies ω, i.e. by defining

$$\psi(\boldsymbol{r}, t) \equiv 2\int_0^{\infty} d\omega \bar{\psi}(\boldsymbol{r}, \omega) e^{i\omega t}. \tag{2.5}$$

Another method [6] applies to the rather common cases where the source of the wave contains a continuously-running oscillator with frequency ω_0 and phase ϕ, modulated smoothly by an amplitude $a(t)$, so that the emitted wave has time-dependence

$$a(t)\cos\{\omega_0 t + \phi\}. \tag{2.6}$$

The complex wave produced by diffraction (and which no longer has the time-dependence (2.6)) is then defined as

$$\psi = \psi_R + i\psi_I, \tag{2.7}$$

where ψ_R and ψ_I are the real diffracted waves when the phase of the reference oscillator is $\phi = 0$ and $\phi = \pi/2$, respectively. Yet another method [7] defines wave crests directly as the instantaneous surfaces in space on which the value of ψ_R as a function of time has a local maximum, i.e. by

$$\partial\psi_R/\partial t = 0, \qquad \partial^2\psi_R/\partial t^2 < 0. \tag{2.8}$$

For the monochromatic or quasimonochromatic waves in which we are interested, the wavefronts, and wavefront singularities, produced by these different procedures are almost identical. Therefore from now on I will employ the scheme based on (2.1) and (2.2) without further discussion.

The most important features of wavefronts are their singularities, which correspond to singularities of the phase function $\chi(\boldsymbol{r}, t)$. The nature of these singularities is determined by the fact that ψ is a smooth single-valued function of its variables. Single-valuedness implies that during a circuit C in space–time χ may change by $2m\pi$, where m is an integer. Suppose m is not zero, and let C be shrunk to a very small loop in such a way that m does not change. Then C encloses a singularity,

because χ is varying infinitely fast. The smoothness of ψ now implies that this can happen only where $\psi = 0$, i.e. where χ [eq. (2.1)] is indeterminate. Since the vanishing of ψ requires two conditions ($\mathrm{Re}\,\psi = \mathrm{Im}\,\psi = 0$), these phase singularities are *lines in space*, or *points in the plane*. Nye and Berry [6] called them "wavefront dislocations", because of their close morphological analogy with crystal dislocations.

Mathematically, the strength S_c of singularity inside the circuit C is defined by regarding $\psi(\boldsymbol{r}, t)$ as a map from space–time to the plane with polar coordinates ρ, χ; under this map, C has the image C', and S_c is defined as the winding number of C' about the origin. Therefore S_c is simply the net number of wavecrests encountered around C, i.e.

$$S_c = \frac{1}{2\pi}\oint_C \mathrm{d}\chi = \frac{1}{2\pi}\oint_C \nabla\chi\cdot\mathrm{d}\boldsymbol{r}, \tag{2.9}$$

where the last form of writing applies when C is spacelike.

Figure 1 shows a snapshot of the crests of a dislocated plane wave travelling up the page, together with some circuits C and, below, their images C'. It is clear that C_1 and C_2 enclose the dislocation (which has unit strength), while C_3 does not.

It is not hard to accept that dislocations are the generic singularities of complex functions. But our functions ψ describe waves, and so must satisfy wave equations. This might prove to be such a strong restriction as to prevent the formation of dislocations, or to cause them to appear in degenerate forms. In fact, as I shall show with numerous examples in subsection 2.2, the wave equation does not impose any strong constraint

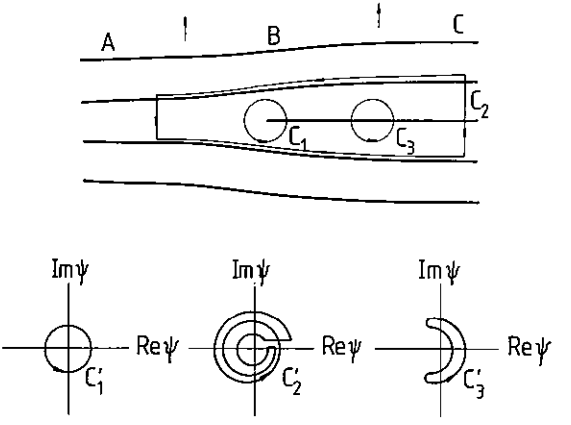


Fig. 1.

on the morphology of phase singularities, so that a typical wave carries a tangle of dislocation lines.

What causes them? The answer is: destructive interference among the different contributions to ψ at a given space–time point P, for example rays crossing at P, or wavelets scattered to P from different parts of a diffracting object. Since dislocations occur on nodal lines of ψ, where the intensity ρ^2 vanishes, they might be confused with the "dark fringes" often discussed in elementary treatments of interference. But these dark fringes are conceived as surfaces rather than lines, and defined as the locus of places where two interfering waves are out of phase. This causes the intensity to be small, but it is only on the dislocation lines, where the two waves have equal amplitude as well as being in antiphase, that the intensity is *exactly* zero. The wavefront dislocation is a more general concept than the dark fringe, because it is applicable in cases where more than two waves interfere, or when two waves interfere with unequal and changing amplitudes.

2.2. *Dislocation morphologies*

I will describe some dislocations satisfied by the non-dispersive linear scalar wave equation in a uniform static isotropic medium; the equation is

$$\nabla^2\psi = \frac{1}{c^2}\frac{\partial^2\psi}{\partial t^2}. \tag{2.10}$$

One of my intentions in restricting myself to such an elementary mathematical framework will be to show how much can come out of so little. With one exception, all the dislocations will occur in a wave which is asymptotically monochromatic with wave number k and travelling entirely in the z-direction. This "undislocated wave" is

$$\psi = \mathrm{e}^{\mathrm{i}k\zeta}, \quad \text{where} \quad \zeta \equiv z - ct. \tag{2.11}$$

Let us first consider dislocations moving rigidly with the "host" wavefronts of the undislocated wave, since these correspond most closely with static dislocations in crystals. The wave function in such cases must take the form

$$\psi = \psi(x, y, \zeta), \quad \text{where} \quad \frac{\partial^2\psi}{\partial x^2} + \frac{\partial^2\psi}{\partial y^2} = 0. \tag{2.12}$$

Our first example is the dislocation of fig. 1, which is situated at $\zeta = x = 0$ and so lies parallel to the y axis. In crystallographic terminology it is a *pure edge dislocation*. Its wave function [6] is

$$\psi = A(kx + \mathrm{i}\beta k\zeta)\mathrm{e}^{\mathrm{i}k\zeta}, \tag{2.13}$$

where A and β are real constants. From (2.1), the phase is

$$\chi = k\zeta + \arctan\{\beta\zeta/x\} + 2n\pi, \tag{2.14}$$

the ambiguity of an odd multiple of π in the arctan function being resolved by requiring the sine of the angle to have the same sign as $\beta k\zeta$. Contours of χ, i.e. wavefronts, are shown in fig. 2. Notice how all wavefronts pour into the singularity at the origin, where a wavecrest ends. Notice also how wavefronts with phase π (troughs) cross at a saddle point

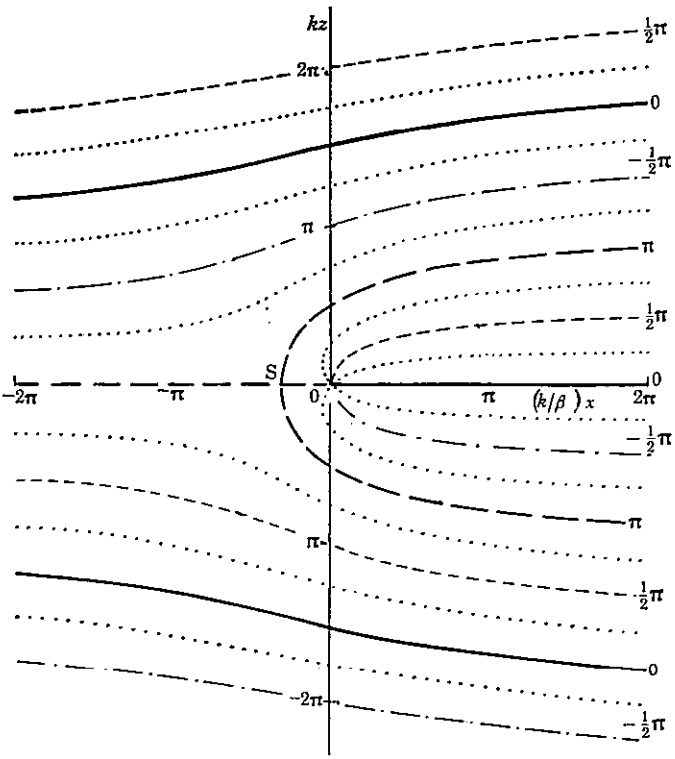


Fig. 2. (From ref. [6].)

of χ situated at $kx = -1$, i.e. $1/2\pi$ wavelengths away from the dislocation. This shows that dislocations are the most delicate features of waves, in the sense that they exhibit interesting phase topology on scales much smaller than a wavelength.

Complementary to the pure edge is the *pure screw dislocation*, which for strength s has wave function

$$\psi = A(kx \pm iky)^s e^{ik\zeta} = A(kR)^s \exp\{i(k\zeta \pm s\theta)\}, \tag{2.15}$$

where R and θ are polar coordinates in the xy plane. The wavecrests consist of s intertwined helicoids whose axis – the dislocation – coincides with the z axis; fig. 3 shows the crests when $s = 2$. Notice that dislocation (2.15) describes a purely monochromatic wave.

Interpolating between these pure cases is the *mixed edge-screw* dislocation. If this is a straight line lying in the yz plane and making an angle δ with the y axis, its equation is

$$\psi = A\{kx + i\beta(k\zeta \cos\delta - ky \sin\delta)\}e^{ik\zeta}. \tag{2.16}$$

$\delta = 0$ and $\delta = \frac{1}{2}\pi$ correspond to pure edge and pure screw, respectively. The wave (2.16) has a dislocation strength of unity. Equation (2.15) can be generalized [7] to produce mixed dislocations with $s > 1$. Such multiple dislocations correspond to multiple zeros of ψ and so would seem to be unstable against perturbation. It is also possible to construct waves containing pure edge dislocations with multiple strength, but only at isolated points or instants at which the dislocation interacts with another.

We conclude that generic wavefront dislocations have strength unity. Now observe that only pure edge dislocations disconnect the wavecrests (figs. 1 and 2); if there is any screw character at all, the wavecrests are all helically connected along the dislocation (cf. fig. 3). These facts suggest

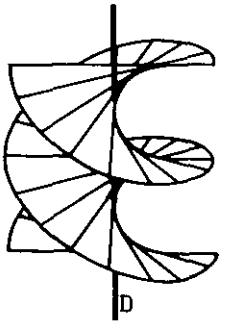


Fig. 3. (From ref. [7].)

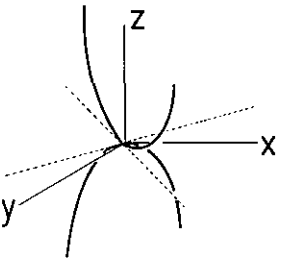


Fig. 4.

the curious conclusion that in three spatial dimensions a typical wave possesses *only one wavecrest*, in the global sense that the points satisfying $\chi = 0 \mod 2\pi$ form a single connected set. A proof of this would be welcomed.

By modulating the undislocated wave (2.11) with terms in x, y that are of higher-than-linear order and which satisfy (2.12), it is possible to generate *curved dislocations* that move rigidly with the host wavefronts. An example is

$$\psi = \{(kR)^2 e^{-2i\theta} + i\beta k\zeta\} e^{ik\zeta}, \tag{2.17}$$

whose dislocation lines have the form of two parabolas lying in orthogonal vertical planes (fig. 4). When $|\zeta| \to \infty$, the dislocations are pure screw, and when $\zeta \to 0$ there are two pure edge dislocations intersecting orthogonally. The limit $\beta \to 0$ is degenerate and corresponds to a double-strength pure screw along the z-axis.

If we now consider solutions of (2.10) which do not have the form (2.12), i.e. which contain z or t separately and not only in the combination ζ, we can produce *dislocations moving relatively to the host wavefronts*. A simple example is

$$\psi = A\{\alpha kx + k^2x^2 + i(\beta k\zeta + kz)\} e^{ik\zeta}, \tag{2.18}$$

where α is a real constant; this has two edge dislocations at $x = 0$ and $x = -\alpha/k$, moving parallel to the z-axis with speed $v = \beta c/(\beta + 1)$ (fig. 5). In crystallographic terminology such motion is *glide*. Note that

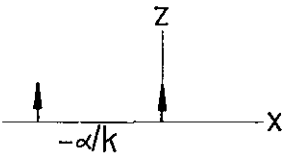


Fig. 5.

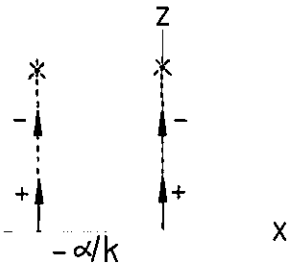


Fig. 6.

$v \to -\infty$ as $\beta \to -1$, indicating that dislocations can flash through the host wavefronts at arbitrarily high speeds; this does not contradict any principle of relativity, because dislocations are forms and not things, and so cannot be used as signals.

A more interesting phenomenon occurs if the coefficient of β is made quadratic, i.e. if

$$\psi = A\{\alpha kx + k^2x^2 + i(\beta k^2\zeta^2 + kz)\}e^{ik\zeta}. \tag{2.19}$$

Now there are pairs of edge dislocations at $x = 0$ and $x = -\alpha/k$. If $\beta > 0$ the members of each pair approach and *annihilate* (fig. 6), while if $\beta \leqslant 0$ they *suddenly appear* and glide apart. These interactions are possible because the two dislocations in each pair are of opposite sign.

A slight modification of form (2.19), namely

$$\psi = A\{k^2x^2 + i(\beta - i\gamma)k^2\zeta^2 + ikz\}e^{ik\zeta}, \tag{2.20}$$

where γ is a real constant, produces a pair of edge dislocations approaching each other along parabolic trajectories and annihilating when they meet (fig. 7). This motion, which in crystallographic terminology is *climb*, corresponds to the disappearance of a strip of wavefront, or to the spontaneous healing of a tear in a wavefront, or to the time-reversed versions of these behaviours, depending on the values of β and γ.

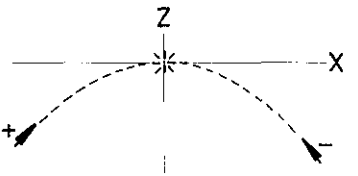


Fig. 7.

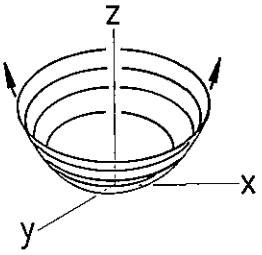


Fig. 8.

The analogue of (2.20) in polar coordinates is

$$\psi = A\{\tfrac{1}{2}k^2R^2 + \mathrm{i}kz + \mathrm{i}(\beta - i\gamma)k^2\zeta^2\}\mathrm{e}^{\mathrm{i}k\zeta}, \tag{2.21}$$

and corresponds to a *circular climbing edge dislocation loop* (fig. 8). Depending on β and γ this could describe the sudden puncture of a wavefront or the spontaneous healing of a puncture or the time-reversed behaviour.

A different sort of interaction is obtained with waves involving cubic terms. For example,

$$\psi = (\tfrac{1}{3}k^3x^3 + \beta\zeta + \mathrm{i}k^2xz)\mathrm{e}^{\mathrm{i}k\zeta}, \tag{2.22}$$

describes two edge dislocations parallel to the y-axis approaching the origin in the x-direction (by climb) and in the z-direction (by glide). At $t = 0$ the dislocations (which have opposite sign) collide and *bounce* off one another (fig. 9) so that each is deflected through a right angle.

To obtain *moving screw dislocations*, the simplest procedure is to seek solutions in which the undislocated wave is modulated by terms involving the combination $x \pm \mathrm{i}y$. For example,

$$\psi = \{kct - \mathrm{i}k^2x^2 - \alpha k(x + \mathrm{i}y)\}\mathrm{e}^{\mathrm{i}k\zeta}, \tag{2.23}$$

describes a screw dislocation gliding along a parabolic path (fig. 10). Replacing α by $-\mathrm{i}\beta$ gives two screw dislocations of opposite sign

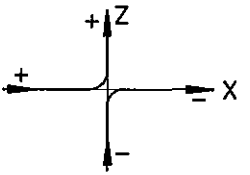


Fig. 9.

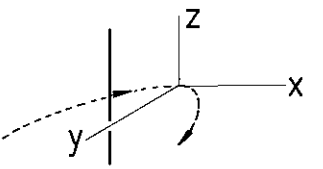


Fig. 10.

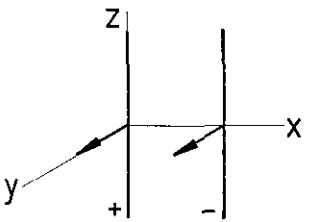


Fig. 11.

(fig. 11) gliding parallel to the y-axis whilst maintaining constant separation. Finally, replacing α in eq. (2.23) by $\alpha - i\beta$ gives two screw dislocations of opposite sign which glide together along a parabolic path (fig. 12) and annihilate.

These solutions show what rich topology is contained in waves that are algebraically quite simple; they also show how changing the sign of a parameter, or changing it from real to imaginary, can dramatically alter the wavefront geometry and its singularity structure.

All our examples so far have been dislocated plane waves. But the undislocated wave need not be plane, and we conclude with an example in which it is cylindrical. The undislocated wave is

$$\psi(R,\theta) = H_0^{(1)}(kr)e^{-i\omega t}, \tag{2.24}$$

where $H_0^{(1)}$ denotes the "outgoing" Bessel function of the third kind [8]. In the xy-plane the wavefronts are expanding circles centred on the origin. A similar wave with an edge dislocation of strength s is

$$\psi(R,\theta) = H_s^{(1)}(kR)\exp\{i(s\theta - \omega t)\}; \tag{2.25}$$

this is a constant angular momentum eigensolution of eq. (2.10), obtained by separation of variables using polar coordinates. Using standard asymptotic forms for the Bessel function [8] it is possible to show that the wavecrests have the form shown in fig. 13 for $s = 10$: s crests emerge

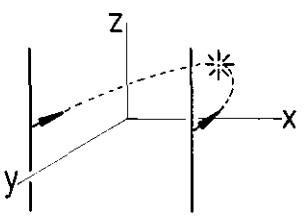


Fig. 12.

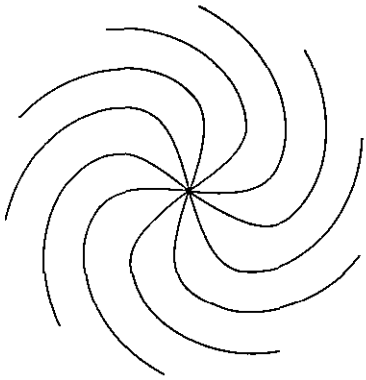


Fig. 13.

radially from the origin and continue almost straight out to the circle $kR = s$, where they curve into equiangular spirals which as $R \to \infty$ approximate circles separated by the wavelength $2\pi/k$; as time proceeds, the whole pattern rotates. I have not seen this picture in any textbook on mathematical physics. My main reason for including this example is the recent upsurge of interest in spiral waves of chemical reaction [9]; of course the waves there are non-linear, but it is surprising to see the same geometry arising in linear waves. The spiral waves (2.25) differ from our other examples in that the dislocation is not merely a phase singularity but a physical singularity of the wave, in this case a line source.

2.3. Dislocations in acoustics, electromagnetism, quantum mechanics and water waves

Wavefront dislocations were discovered [6] in ultrasound reflected from a rough surface. The incident wave was a quasimonochromatic pulse (fig. 14a), and the reflected wave, received at a point and displayed on an oscilloscope as sound pressure vs. time, was an extended train of disorderly oscillations (fig. 14b). On moving the receiver to explore the wave at different places, it was quite common to observe two wavecrests [now defined by eq. (2.8)] move apart and an extra crest appear between them, or the time-reversed sequence of events. Two sequences of this type are shown in figs. 15 and 16. The meaning of such a birth or death of a wavecrest is that a dislocation line has intersected the track of the receiver. This is clear from fig. 1: four crests pass the point A, and five

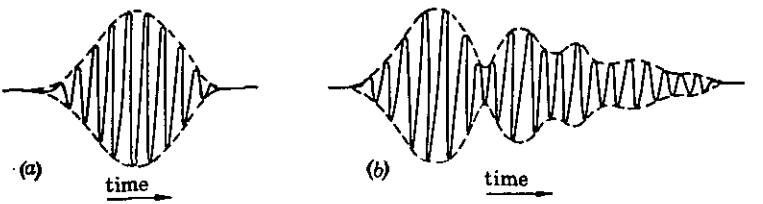


Fig. 14.

pass the point C, so that a receiver moved from A to C must pass a point B at which a crest is born. The time-dependences in figs. 15 and 16 were in fact calculated by taking the real and imaginary parts of ψ for a pure edge dislocation, as given by eq. (2.13).

The ultrasonic experiments were devised as laboratory analogues for the radio echo sounding of polar ice sheets. Echoes (formed in radio waves with length $\lambda \sim 5$ m) come principally from the interface between

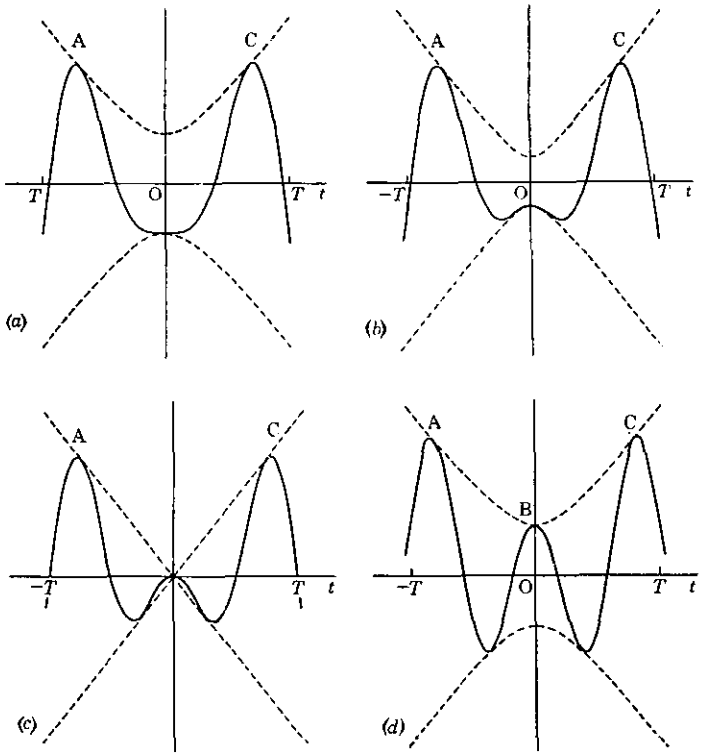


Fig. 15.

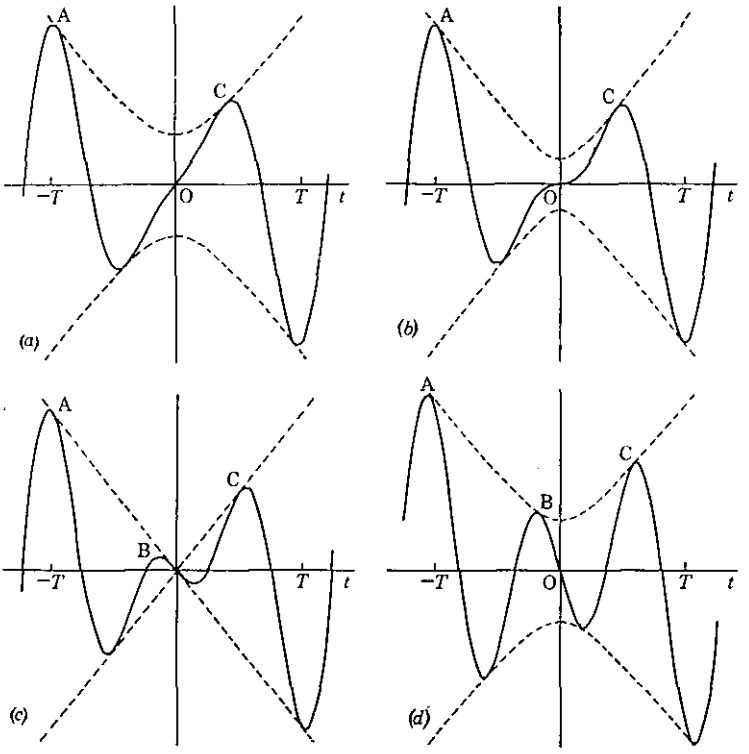


Fig. 16.

ice and bedrock, and were previously used [10] to get information about the geography and nature of the subglacial terrain. Dislocations are probably too sensitively dependent on the details of the topography to be useful as a means of determining that topography. But this very sensitivity makes it easy to detect displacement relative to a fixed bedrock of unknown topography. Such displacement occurs naturally for a radio echo sounder on an ice sheet, because the sheet flows horizontally and also changes its thickness. The echo, and especially its skeleton of dislocation lines, provides a three-dimensional reference frame, fixed relative to the bedrock, with respect to which displacements can be measured. Recent phase-sensitive echo-sounding experiments in the Canadian arctic by Walford et al. [11] have detected wavefront dislocations, and an earlier experiment [12] in Antarctica suggests that displacements as small as a hundredth of a wavelength can be measured in this way.

Of course ultrasound pulses are important in their own right and not just as analogues of radio pulses. They are employed medically, in the study of soft tissue, and industrially, in the non-destructive testing of metal objects (boilers, turbine blades, etc.) for internal cracks. From a theoretical point of view these applications are quite crude; usually, only the times of emission and reception of the pulse as a whole are recorded, and interference effects arising from the presence of several waves within the pulse envelope (quasimonochromaticity) are ignored. In any precise quantitative interpretation of an echo, however, a precondition would be a complete understanding of the wave field in the emitted pulse, dislocations and all. Therefore Wright [13] made a thorough study of the wave emitted by a circular acoustic piston radiator (fig. 17) of radius a, excited by a quasimonochromatic pulse with time-dependence $F(t)$.

His results reveal an astonishing richness of dislocation structure, quite unsuspected in the many earlier studies of this system. The general picture is of circular edge dislocation loops born in the near field and moving out with the pulse into the far field along cones which are close to the asymptotic null surfaces of the pattern formed with purely monochromatic waves. As an example, fig. 18 shows the trajectories of dislocation loops for the case $ka = 10$, i.e. the piston radius is $5/\pi$ wavelengths, and the piston is excited with time-dependence

$$F(t) = \exp\left\{\frac{-t^2}{2}\left(\frac{\omega}{3\pi}\right)^2 + i\omega t\right\}, \tag{2.26}$$

whose intensity $|F|^2$ is a Gaussian which contains three cycles of the carrier wave within its standard deviation, and where $\omega = ck$. Points along the trajectories are labelled in fig. 18 by the times ωt at which the dislocation loop reaches them. Births and deaths of dislocations are labelled B and D, respectively. The dashed lines indicate the asymptotic null cones of the monochromatic wave.

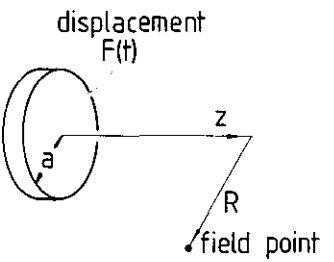


Fig. 17.

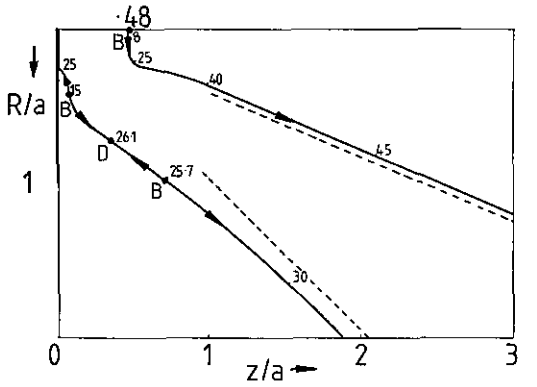


Fig. 18. (After ref. [13].)

There are two dislocation trajectories. One of them starts on the symmetry axis, where a loop is born [cf. fig. 8 and eq. (2.21)] when $\omega t \approx 8$ at an isolated point zero of the monochromatic wave; the loop expands smoothly, although not at constant speed, whilst receding into the far field. The other is more complicated: when $\omega t \approx 14.5$ two loops with opposite signs are born close to the piston at a finite radius; one travels backwards, hits the piston and disappears, while the other travels outwards and when $\omega t \approx 26.4$, annihilates with one loop of a second pair which was born when $\omega t \approx 25.7$; the other loop of the second pair recedes smoothly into the far field.

Figure 18 was constructed by the laborious procedure of computing the wave $\psi(\boldsymbol{r}, t)$ for many times t, by means of an exact diffraction integral, and plotting its wavefronts so as to identify the dislocations "experimentally". Later, Wright [13] devised an ingenious analytical procedure, recently refined by Wright and Nye [14], in which the finite width of the pulse envelope in the quasimonochromatic case is considered as a perturbation of the monochromatic wave. This procedure gives good approximations to dislocation trajectories such as those in fig. 18.

Dislocations are such fundamental features of waves that it is not surprising to learn that their existence has been anticipated several times. The earliest and most unusual reference known to me occurs in two startlingly original papers by Whewell [15] about the ocean tides. I quote from the first paper, written in 1833:

> "Ever since the time of Newton, his explanation of the general phenomena of the tides by means of the action of the moon and the sun has been assented to by all

philosophers who have given their attention to the subject. But even up to the present day this general explanation has not been pursued into its results in detail, so as to show its bearing on the special phenomena of particular places – to connect the actual tides of all the different parts of the world – and to account for their varieties and seeming anomalies... We are, perhaps, not even yet able to answer decisively the enquiry which Bacon suggests to the philosophers of his time, whether the high water extends across the Atlantic so as to affect contemporaneously the shores of America and Africa, or whether it is high on one side of this ocean, when it is low on the other...

It will easily be understood that we may draw a line through all the adjacent parts of the ocean which have high water at the same time; for instance, at 1 o'clock on a given day. We might draw another line through all the places which have high water at 2 o'clock on the same day. Such lines may be called *cotidal* lines; and they will be the principal subject of the present essay."

What Whewell recognizes here is that the tide can be considered as a giant wave, with a period of about twelve hours, moving across the oceans. His cotidal lines are the wavefronts of this wave. He appreciates that a map of the cotidal lines in an ocean would result in an intelligible picture of the pattern of tides around its coasts, but realises that data collected before 1833 were incomplete or misleading. He called for more observations.

By the time he wrote his second paper, in 1836, these observations had been made, in what may have been the first international geophysical collaboration involving several hundred coastguards. He writes

"I have already pointed out the extreme difficulty of forming into a consistent and intelligible scheme the tides of the German Ocean [i.e. the North Sea]. But as we now have a connected series of observations along the whole of its coast, we must make the attempt...

It appears that we may best combine all the facts into a consistent scheme, by dividing this ocean into two *rotatory* systems of tide-waves;...[in each] space the cotidal lines may be supposed to revolve round [a point] where there is no tide, for it is clear that at a point where all the cotidal lines meet, it is high water equally at all hours, that is, the tide vanishes."

In other words, the North Sea contains two wavefront dislocations (stationary and of edge type). Whewell's map, and also modern charts, such as fig. 19, show them very clearly; they play an important part in oceanography [16], in which context they are referred to as "amphidromic points".

Dislocations have also been anticipated in optics. Braunbek and Laukien [17] calculated the wavefronts for Sommerfeld's celebrated exact solution of the problem of a plane wave diffracted by a half-plane. Their results are reproduced in fig. 20; dislocations (in this case also

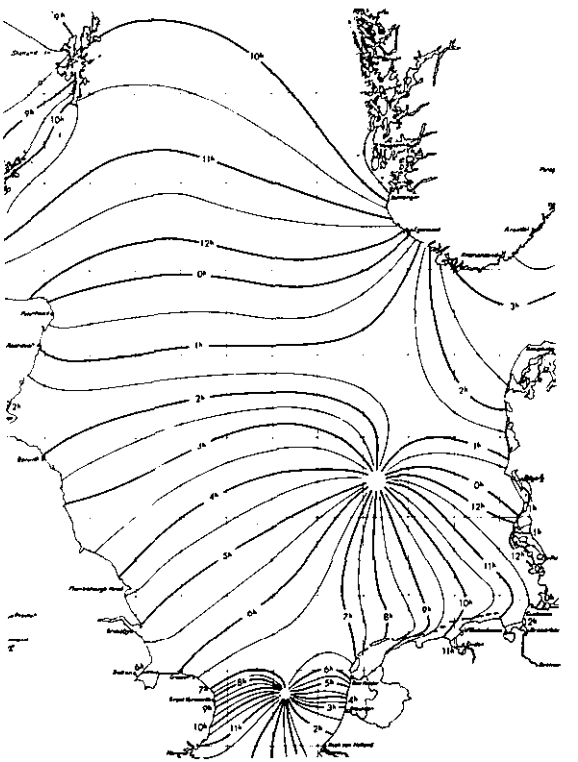


Fig. 19.

stationary and of edge type) are formed by interference between the incident wave (travelling down the page) and the reflected wave. More optical dislocations, with a more elaborate structure, will be described in connection with diffraction catastrophes in subsection 4.3.

Now we come to consider dislocations in the probability waves of quantum mechanics. In a famous paper written in 1931, Dirac [18] studied the phase $\chi(\boldsymbol{r}, t)$ of the wave function for a particle with charge q in an external electromagnetic field with vector potential $\boldsymbol{A}(\boldsymbol{r}, t)$. He showed that a solution $\psi_0(\boldsymbol{r}, t)$ of Schrödinger's equation in the absence of the field could be converted to a solution $\psi(\boldsymbol{r}, t)$ in the presence of the field by multiplication by a phase factor depending on $\boldsymbol{A}$, i.e.

$$\psi(\boldsymbol{r}, t) = \psi_0(\boldsymbol{r}, t)\exp\left\{\mathrm{i}q\int_{\boldsymbol{r}_0}^{\boldsymbol{r}} \boldsymbol{A}(\boldsymbol{r}', t)\cdot \mathrm{d}\boldsymbol{r}'/\hbar\right\}, \tag{2.27}$$

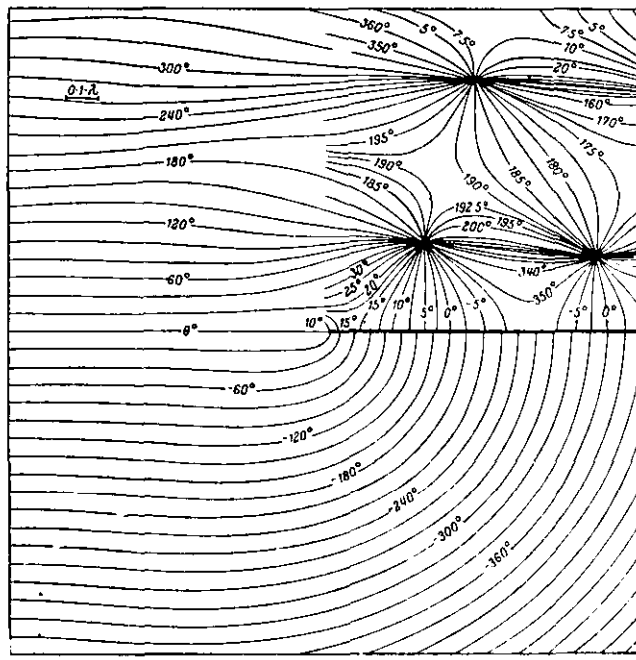


Fig. 20. (From ref. [17].)

where $\hbar$ is Planck's constant and $\boldsymbol{r}_0$ is an arbitrary reference position. In other words, the field acts like an extra contribution χ_1 to the phase, given by

$$\nabla\chi_1 = q\boldsymbol{A}/\hbar. \tag{2.28}$$

If there is a non-vanishing magnetic field $\boldsymbol{B}(= \nabla \wedge \boldsymbol{A})$, this equation will be non-integrable so that χ_1 at a point $\boldsymbol{r}$ will depend on the path from $\boldsymbol{r}_0$ to $\boldsymbol{r}$ and ψ will not be single-valued; Dirac argues that multivaluedness of this sort is not inconsistent with the principles of quantum mechanics. He recognizes that ψ_0 can have nodal lines around which the phase χ_0 in the absence of the field changes by $2s\pi$, i.e. he recognises the existence of wavefront dislocations.

But rather than studying dislocation lines as generic wave morphologies, he asked the deeper question of what would happen if one should come to an end. To answer this, he noted that the total phase change $\Delta\chi$ round a small closed curve C is

$$\Delta\chi = \Delta\chi_0 + \Delta\chi_1 = 2\pi s + \frac{q}{\hbar}\oint \boldsymbol{A}\cdot \mathrm{d}\boldsymbol{r} = 2\pi s + \frac{q}{\hbar}\iint_{\Sigma} \boldsymbol{B}\cdot \mathrm{d}\boldsymbol{S}, \tag{2.29}$$

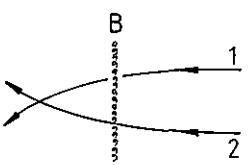
where the last integral is over any surface Σ bounded by C and where s is the total strength of dislocations piercing Σ. If Σ is a *closed* surface, $\Delta\chi = 0$ and s can be non-zero only if a dislocation line ends inside Σ. If this happens, there is a net magnetic flux through Σ, and hence a *magnetic monopole strength* μ inside Σ, given by

$$\mu = \frac{1}{4\pi}\oiint_{\Sigma} \boldsymbol{B}\cdot \mathrm{d}\boldsymbol{S} = \frac{-\hbar s}{2q}. \tag{2.30}$$

As well as predicting the existence of magnetic monopoles, in contradiction to standard formulations of classical electromagnetism, Dirac's argument shows that their strength must be quantized according to eq. (2.30) provided the electric charge q is quantized.

Riess [19–21] suggested an interesting extension of Dirac's interpretation (2.28) of an external magnetic field as an extra phase: the phase gradient $\nabla\chi$ of a wave with no external field is considered to be the vector potential of an *internal* magnetic field produced by the quantal system of charged particles. This interpretation suffers from the difficulty that the $\boldsymbol{B}$-field whose potential is $\nabla\chi$ is non-zero only on isolated (quantized) flux lines (the dislocations) and is different from the $\boldsymbol{B}$-field that Maxwell's equations would generate from the quantum-mechanical current $q\,\mathrm{Im}\,\psi^*\nabla\psi$. Riess also considers dislocations in the wave functions of many-particle systems (he calls them "nodal hypersurfaces"). He proves that a non-zero current density implies the existence of dislocations. He also proves that in the presence of an external field the absence of dislocations implies that the system is diamagnetic, whilst paramagnetism implies that dislocations must exist.

The magnetic phase factor (2.27) can give rise to observable interference effects if waves can travel by different routes to the same point. A striking example is the *Aharonov–Bohm effect* [22], in which a beam of charged particles is diffracted by a thin impenetrable solenoid (fig. 21) containing (ideally) a single flux line of magnetic field $\boldsymbol{B}$. In the most elementary analysis, waves that have travelled round different sides of the solenoid (path 1 and 2 on fig. 21) interfere with a phase difference whose magnetic part is determined by the quantum flux parameter α,



defined by

$$2\pi\alpha \equiv \Delta\chi_1 \frac{q}{\hbar}\left\{\int_{\text{path 1}} \boldsymbol{A}\cdot\mathrm{d}\boldsymbol{r} - \int_{\text{path 2}} \boldsymbol{A}\cdot\mathrm{d}\boldsymbol{r}\right\}$$

$$= \frac{q}{\hbar}\oint \boldsymbol{A}\cdot\mathrm{d}\boldsymbol{r} = \frac{q}{\hbar}\iint \boldsymbol{B}\cdot\mathrm{d}\boldsymbol{S}. \tag{2.31}$$

As α is varied (for example by changing the current in the solenoid), the interference pattern should change, and precisely this predicted behaviour was observed in an experiment by Chambers [23]. The surprising (and controversial) thing about the Aharonov–Bohm effect is its implication that in quantum mechanics the behaviour of charged particles can be affected by vector potentials in a region (the space outside the solenoid) where there is no electromagnetic field.

My purpose in describing the effect is to draw attention to the little-known fact that the Aharonov–Bohm wave function has a wavefront dislocation coinciding with the flux line. The wave function is given not by (2.27) but by the solution of Schrödinger's equation in the presence of the vector potential outside the flux line and vanishing on it. In polar coordinates, for the case where the particles are incident from $x = +\infty$, this solution [22] is given in terms of the flux parameter α by

$$\psi(R,\theta) = \sum_{l=-\infty}^{+\infty} (-\mathrm{i})^{|l-\alpha|}\mathrm{e}^{\mathrm{i}l\theta}J_{|l-\alpha|}(kR). \tag{2.32}$$

Close to the flux line $R = 0$, ψ is dominated by the term involving the Bessel function with smallest order $|l-\alpha|$. This has $l = s$, where s is the closest integer to α. The behaviour of the angular factor in eq. (2.32) during a circuit of the flux line implies that this line is a dislocation with strength s.

As the flux α varies continuously, the strength s is piecewise constant, jumping by unity when α passes through half-integral values. In a study of the Aharonov–Bohm wave function, Berry et al. [24] show that this change of dislocation strength occurs by disconnection and reconnection of the wavecrests along a nodal line stretching from the flux line out into the far field; this behaviour is illustrated in fig. 22.

According to a theorem of Wu and Yang [25], all observable properties of ψ must vary periodically with α. But the dislocation strength increases (discontinuously) with α, and so we come to the disappointing conclusion that the interesting behaviour shown in fig. 22 is unobservable in quantum mechanics. It can, however, be seen in an analogue experiment

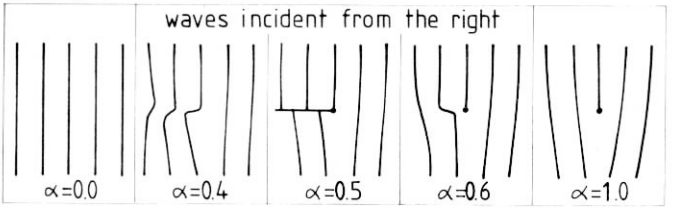


Fig. 22.

employing surface waves on water. If the water has a bulk flow, its velocity can be shown [24] to affect surface waves in the same way as a magnetic vector potential affects charged particles. The analogue of the Aharonov–Bohm effect (potential but no field apart from a single flux line) is therefore a flow which is irrotational everywhere except on a vortex line, and is realizable as ripples encountering a vortex produced by letting water flow out of a tank. We did the experiment [24] and obtained results in good agreement with the predictions of theory; fig. 23 shows an example of a water-wave dislocation with strength $s = 2$.

As a final example of dislocations in quantum mechanics, I draw attention to an important series of papers [26–39] by Hirschfelder and

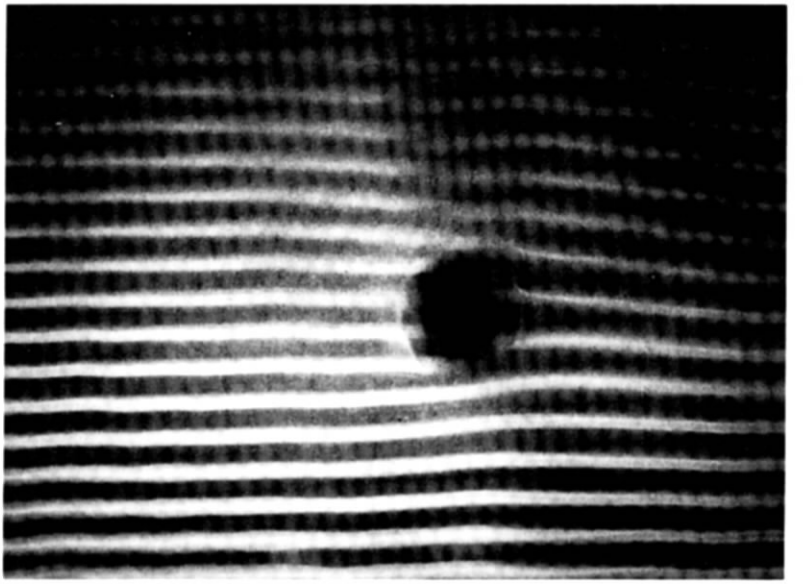


Fig. 23.

his collaborators. They concentrate not on the phase but on its gradient $\nabla\chi$, a vector whose direction defines "quantum-mechanical streamlines". The reason for this terminology is that $\nabla\chi$ is parallel to the local quantum-mechanical current, defined as the expectation value (for a state $|\psi\rangle$) of the velocity density operator, i.e. as

$$\frac{\hbar}{\mu}\rho^2\nabla\chi = \frac{\hbar}{\mu}\mathrm{Im}\,\psi^*\nabla\psi$$

$$= \frac{1}{2\mu}\langle\psi\,|\,\{p_{\mathrm{op}}\delta(r-r_{\mathrm{op}}) + \delta(r-r_{\mathrm{op}})p_{\mathrm{op}}\}\,|\,\psi\rangle. \quad (2.33)$$

Near dislocations, from which wavefronts radiate (fig. 2), $\nabla\chi$ has a *vortex structure*, whose circulation [eq. (2.9)] is quantized. Hirschfelder et al. [26–29] present thorough numerical and analytical studies of the streamlines and quantized vortices for several quantal problems. As one example of their results, fig. 24 shows the streamlines of a particle beam scattered by a spherical attracting object with square-well interaction potential; the particle energy E corresponds to a de Broglie wavelength of 2π times the sphere radius, and the potential strength is $-9.2E$. Three edge dislocation loops are evident; the one inside the interaction sphere has a different sign from the ones outside. This figure shows very clearly yet again how the topological point of view gives new insights into problems which have been studied very often in the past and which we thought we knew all about.

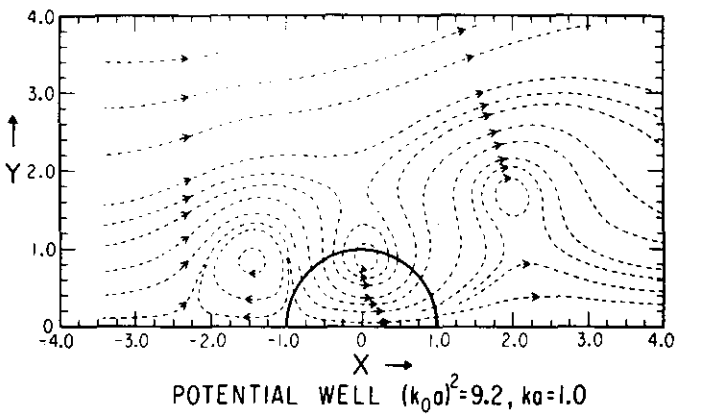


Fig. 24 (From ref. [29].)

3. Caustics as catastrophes

3.1. Singularities of ray families

Let us now consider, not the phase singularities which occur on the sub-wavelength scale, but the singularities arising when a wave is perceived on such a scale that the wavelength is too small to be discerned. In this "shortwave limit" the concept of a wave function is not useful, and is replaced by the notion of *trajectories* along which energy flows: the trajectories corresponding to Maxwell's equations are light rays, and the trajectories corresponding to Schrödinger's equation are the paths of Newtonian particles. Here, I shall employ the language of geometrical optics, and so speak mainly of rays. The subtle connections between wave theories and their ray approximations will be explained in section 4; until then, you should think of rays as entities in their own right.

A natural question to ask is: what are the singularities of ray theory? To answer this, it is crucial to realize that rays occur not in isolation but as *families* (even a laser beam must be many wavelengths wide if it is not to spread by diffraction). It is a family of rays that corresponds, in the shortwave limit, to a wavefield. Now, a family of rays possesses the possibility, not inherent in any individual ray, of *focusing*, that is, of concentrating energy into a region whose dimensionality is smaller than that of the space inhabited by the rays. The general term for such a focal region is a *caustic*, defined as the envelope of the ray family, i.e. the locus of places where neighbouring rays touch. Conservation of energy implies that in geometrical optics the energy density, or intensity, is infinite on caustics, so that the answer to our questions is: *caustics are the singularities of ray theory*. Caustics dominate optical wavefields in the short-wave limit, because it is on caustics that the light is brightest. It would seem obvious that the study of caustics, and in particular the classification of their possible forms, should be an important part of optics, but it is a surprising fact that the systematic investigation of caustics as geometric objects began only recently, and after the appropriate mathematics had been developed. This mathematics is catastrophe theory [1, 2], and I shall describe its main results in subsection 3.2 after a few examples have made us more familiar with caustics.

Before giving these examples I want to note the curious fact that, considered as wave singularities, ray caustics are *complementary* to wavefront dislocations, for the following reasons. On a caustic, the intensity (in the shortwave limit) is infinite, whereas on a dislocation the intensity

is zero (subsection 2.1). To observe a dislocation, waves must be explored on the scale of the wavelength; but on this scale caustic singularities are softened by diffraction (cf. section 4) and so lose their prominence. To observe a caustic, on the other hand, it is natural to explore the short wave limit; but in this limit phase structures such as dislocations are too small to be discerned. This complementarity of singularities must surely embody a deep aspect of the wave–particle duality, but I do not know how to exploit it.

The most familiar conception of a caustic derives from *the focus of a perfect lens* (fig. 25), which is an isolated point in space where all rays cross. Such foci represent an ideal for optical technology, and their existence is assumed in elementary discussions of the formation of images by optical instruments. However the isolated point focus is non-generic: it is *unstable* against perturbations such as moving the source or changing the topography of refracting or reflecting surfaces, in the sense that such perturbations drastically alter the geometry of the caustic. Of course optical scientists are well aware of this, and have developed extensive classification schemes [30] of what they call "lens aberrations". Since these schemes regard every caustic as growing out of the singular limiting case of the isolated point focus, they are not intended to answer, and in fact do not answer, the following questions: Is it possible to classify the forms of those caustics which are stable against perturbation, and which therefore are expected to occur in the absence of any such special circumstance as the cylindrical symmetry and precise figuring of a perfect lens? If the answer is yes (it will be!), what are the stable forms?

I shall illustrate the subtlety of these questions, and also set up a language for talking about them, by working through an instructive example, which will contain a surprise at the end. The example is *looking at an underwater object*. The object acts as a source S (fig. 26) at a depth D below the flat horizontal surface $Z = 0$ of water with refractive index n. Considering the problem as two-dimensional for the present, it is clear that a one-parameter family of rays issues from S; an obvious choice of parameter is the angle i made by a ray with respect to the Z axis (fig. 26). The "image" of S is the caustic of the family of refracted rays. It is a virtual caustic, consisting of the envelope of the backward continuation

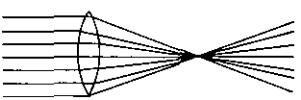


Fig. 25.

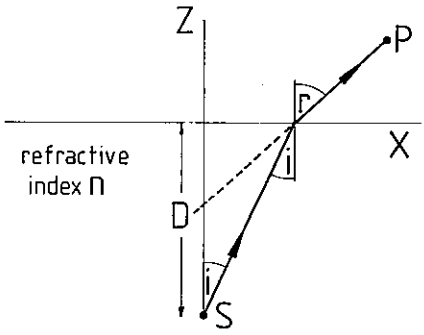


Fig. 26.

for $Z < 0$ of rays refracted through external points P. To an observer at P, the object will appear to be situated on that part of the virtual caustic which is touched by the ray through P.

To find the caustic we must first find the rays. These are determined by Fermat's principle [30]: a trajectory starting from S and ending at P is a ray if and only if its optical distance (i.e. transit time) is stationary under local variations. It is an elementary consequence of this principle that rays in a homogeneous medium are straight, so that only trajectories SP consisting of two straight lines need be considered. The possible trajectories through a point $P = (X, Z)$ are therefore parameterized by the angle i, and from fig. 26 the optical distance $\phi(i; X, Z)$ is simply

$$\phi(i; X, Z) = nD \sec i + \left\{Z^2 + (X - D\tan i)^2\right\}^{1/2}. \tag{3.1}$$

Stationarity then gives the rays by

$$\frac{\partial \phi}{\partial i}(i; X, Z) = \frac{D}{\cos^2 i}\left[n \sin i - \frac{(X - D\tan i)}{\left\{Z^2 + (X - D\tan i)^2\right\}^{1/2}}\right] = 0. \tag{3.2}$$

Recognizing that i is simply the angle of incidence beneath the surface, and introducing the angle of refraction r out of the water (fig. 26), we can rewrite eq. (3.2) as

$$n \sin i = \sin r, \tag{3.3}$$

which of course is Snell's law.

The condition for X, Z to lie on the caustic is that ϕ is stationary with respect to variations in i not just to first order as in eq. (3.2), but at least

to second order; only then will X, Z lie on the envelope of the ray family. To see this, differentiate (3.2) with Z fixed:

$$\frac{\partial^2\phi}{\partial i^2}\mathrm{d}i + \frac{\partial^2\phi}{\partial X\partial i}\mathrm{d}X = 0. \tag{3.4}$$

On a caustic, a bundle of rays $\mathrm{d}i$ is concentrated so that its cross section is of higher than first order in X, so that $\mathrm{d}X = 0$ and (3.4) gives

$$\frac{\partial^2\phi}{\partial i^2}(i; X, Z) = 0 \tag{3.5}$$

as the condition defining a caustic. When applied to (3.2) this gives, on using (3.3)

$$\begin{aligned} n\cos i + \frac{D\sec^2 i}{(Z/\cos r)} - \frac{D(X - D\tan i)^2\sec^2 i}{(Z/\cos r)^3} \\ = \sec^2 i\left\{n\cos^3 i + \frac{D\cos r}{Z} - \frac{D\cos^3 r}{Z^3}\left(\frac{nZ\sin i}{\cos r}\right)^2\right\} \\ = \sec^2 i\left\{n\cos^3 i + \frac{D\cos r}{Z}(1 - n^2\sin^2 i)\right\} = 0. \end{aligned} \tag{3.6}$$

Therefore,

$$Z = -D(1 - n^2\sin^2 i)^{3/2}/n\cos^3 i, \tag{3.7}$$

while (3.2) gives

$$\begin{aligned} X &= D\tan i + Zn\sin i/\cos r \\ &= D(n^2 - 1)\tan^3 i. \end{aligned} \tag{3.8}$$

These are the parametric equations of the caustic, in which X and Z are expressed in terms of i. Elimination of i gives, finally,

$$n^{2/3}Z^{2/3} + (n^2 - 1)^{1/3}X^{2/3} = D^{2/3}, \quad Z < 0. \tag{3.9}$$

This caustic is shown on fig. 27. There are two branches, meeting in a *cusp* at the paraxial image point $(0, -D/n)$. The parametric equation (3.8) shows that, if S emits with equal brightness in all directions i, most of the light is concentrated near this cusp, whose local equation is

$$X = \pm\left\{\frac{2}{3}n\left(Z + \frac{D}{n}\right)\right\}^{3/2} D/\sqrt{(n^2 - 1)}\,. \tag{3.10}$$

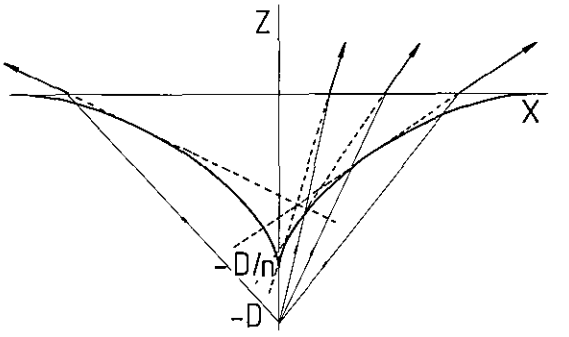


Fig. 27.

In subsection 3.2 we will see that the cusp is a stable form of caustic in two dimensions, so the result embodied by fig. 27 was not unexpected.

Now imagine this as a perturbation of the limit $n \to 1$. In this limit there is no refraction and the caustic is simply the source point S where all rays intersect. This can be seen from (3.7) and (3.8), which show that $Z \to -D$ and $X \to 0$ for all i as $n \to 1$. As soon as n deviates from unity the point focus explodes instantly into the complete cusped caustic curve. It is clear that in this case at least the point focus represents a highly singular limiting case.

Considering now the three-dimensional problem, you might think that because of symmetry the caustic is simply the cusped cone obtained by rotating fig. 27 about the Z-axis. This is false. To see why, let the field point P now depend on a third Cartesian coordinate Y, and let the ray direction be described by an azimuthal angle β in addition to the polar angle i. Instead of (3.1) the optical distance is

$$\phi(i, \beta; X, Y, Z) = nD \sec i + \left\{Z^2 + (X - D \tan i \cos\beta)^2 + (Y - D \tan i \sin\beta)^2\right\}^{1/2} \tag{3.11}$$

This must be stationary under variations of both i and β, i.e.

$$\partial\phi/\partial i = 0; \qquad \partial\phi/\partial\beta = 0, \tag{3.12}$$

leading to ray equations conveniently written

$$\left.\begin{aligned} X &= (D \tan i + Z \tan r)\cos\beta \\ Y &= (D \tan i + Z \tan r)\sin\beta \end{aligned}\right\}, \tag{3.13}$$

where r is given by eq. (3.3).

As before, if X, Y, Z lies on the caustic, ϕ must be stationary to higher order, and now this means that the rank of the matrix of second derivatives of ϕ with respect to i and β must be less than two, i.e.

$$\det\begin{Bmatrix} \dfrac{\partial^2\phi}{\partial i^2} & \dfrac{\partial^2\phi}{\partial i\partial\beta} \\ \dfrac{\partial^2\phi}{\partial i\partial\beta} & \dfrac{\partial^2\phi}{\partial\beta^2} \end{Bmatrix} = 0. \tag{3.14}$$

This follows from an argument analogous to that based on (3.4), and which will later be given in its general form. The calculation is most simply performed by using not eq. (3.14) but the equivalent relation that for fixed Z the Jacobian of $\mathrm{d}X\,\mathrm{d}Y$ with respect to $\sin i\,\mathrm{d}i\,\mathrm{d}\beta$ must vanish (in focusing "a lot goes into a little"). From eq. (3.13)

$$\begin{aligned} \frac{1}{\sin i}\det\begin{Bmatrix} \dfrac{\partial X}{\partial i} & \dfrac{\partial X}{\partial\beta} \\ \dfrac{\partial Y}{\partial i} & \dfrac{\partial Y}{\partial\beta} \end{Bmatrix} &= \frac{1}{\sin i}(D\tan i + Z\tan r)\frac{\partial}{\partial i}(D\tan i + Z\tan r) \\ &= \frac{\sqrt{(X^2+Y^2)}}{\sin i}\frac{\partial}{\partial i}(D\tan i + Z\tan r) = 0. \end{aligned} \tag{3.15}$$

Of the two factors in this caustic condition, the second, involving the derivative with respect to i, gives the expected rotation of eq. (3.9), i.e. a cusped cone. The first factor, $\sqrt{(X^2+Y^2)}$, which was not present in the two-dimensional problem, vanishes when

$$X = Y = 0, \quad Z < 0. \tag{3.16}$$

Therefore the caustic has *another branch*, consisting of a focal line stretching from the cusp to the water surface (fig. 28). On this focal line, infinitely many rays cross, with all azimuths β, whereas off the focal line at most three rays cross, as in the two-dimensional case, with β-values differing by π. Evidently the isolated point focus, when $n \to 1$, is even less stable in three dimensions than in two.

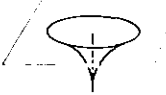


Fig. 28.

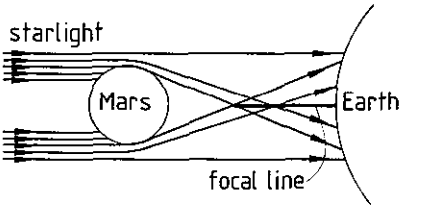


Fig. 29.

The explosion of a focal point into a cusped cone plus focal line is well known in lens theory, where it is called "spherical aberration"; but I shall illustrate this morphology with a different example, giving rise to the biggest caustic I know. This was observed by Elliot et al. [31] during an occultation on 8 April 1976 of the star ϵ Geminorum by the planet Mars, causing a Mars-sized shadow to sweep across the earth's surface. There would be no starlight in this shadow if it were not for Mars' tenuous atmosphere, which acts as a lens with a great deal of spherical aberration, and refracts light (fig. 29) onto the focal line just discussed (the cusped cone is not present because the rays that would give rise to it are obstructed by the solid body of Mars). In the experiment [31], the Kuiper airborne observatory was employed to position a telescope at a point on the track of the predicted centre of the shadow, and the focal line was observed on the photometric trace (fig. 30) as a flash halfway between the instants of disappearance and reappearance of the star's light. I will return to these beautiful observations in subsection 3.3.

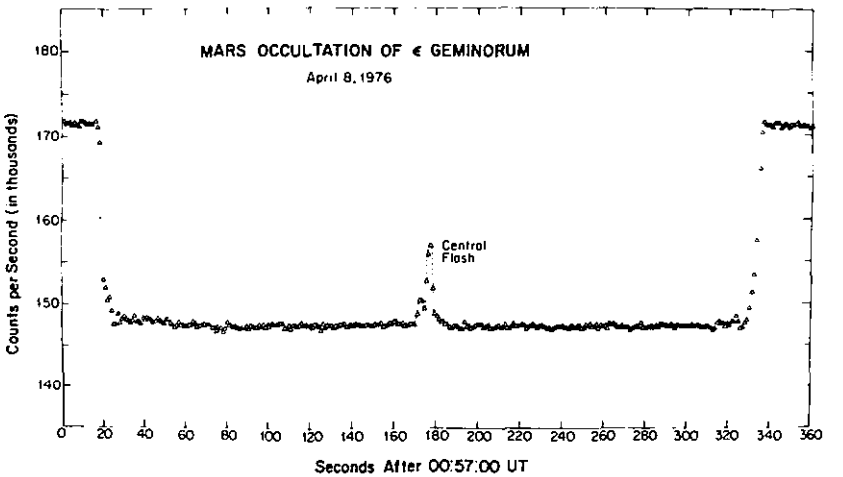


Fig. 30. (From ref. [31].)

Now I return to the underwater object example and deliver the promised surprise: I assert that *the cusped-cone-plus-focal-line of fig. 28 is unstable*. Why is this surprising? Because the caustic of fig. 28 was produced by applying to the three-dimensional problem exactly the same perturbation (n deviating from unity) that in two dimensions caused the focal point to explode into a stable cusped curve. What is different about the three-dimensional case? Answer: the three-dimensional problem has the special (non-generic, atypical, infinitely improbable) feature of *circular symmetry*, and this is what causes the focal line to appear. So what happens to the caustic if the symmetry is broken (e.g. by sending waves across the water surface, or making the refractive index inhomogeneous by local heating of the water)? Answer: this produces a further explosion of the caustic, in which the focal line becomes a complicated surface, which at last is then stable against further perturbation. To understand this we need catastrophe theory.

In order to introduce catastrophe theory, the example given above must be generalized into a mathematical framework applicable to all caustics. The starting point is an optical distance function ϕ: in catastrophe parlance this is called a *generating function*; in mechanics ϕ would be an *action function*. ϕ depends on two quite different types of variable. The first type, called *control parameters*, specify conditions imposed on the rays; in our example, these were the coordinates X, Y, Z of the field point through which the ray must pass, the depth D of the source, and the refractive index n. In our general discussion the set of all such control parameters will be written as $C = (C_1, C_2, \dots)$. The "control space" with coordinates C is the space in which the caustics live. The second type, called *state variables*, label the possible paths compatible with the conditions specified by C, in our example, the state variables were the initial direction coordinates i, β. In our general discussion the set of all such state variables will be written $s = (s_1, s_2, \dots)$. The optical distance function, now written $\phi(s; C)$ is proportional to the transit time along the path s satisfying conditions C.

According to Fermat's principle, s is a ray only if ϕ is stationary with respect to variations of s with C fixed [cf. eqs. (3.2) or (3.12)], i.e.

$$\partial\phi/\partial s_i = 0, \quad \text{for all } i. \tag{3.17}$$

The rays specified by C are the paths s satisfying these equations; they will be denoted by $s^\mu(C)$. The multivaluedness embodied in the index μ is important, and expresses the fact that more than one ray in a family can pass through the same point. It is helpful to think of ϕ for fixed C as the

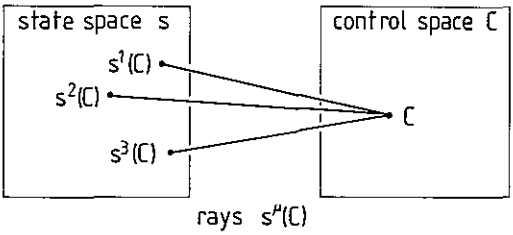


Fig. 31.

height function of a "landscape" with coordinates s; the rays s^μ are the *critical points* (hilltops, saddles, valley bottoms) of this landscape. In catastrophe terminology, the equations (3.17) define a *gradient map* (fig. 31) between control space C and state space s, telling us which ray(s) $s^\mu(C)$ correspond to conditions C.

A typical value of C does not correspond to a caustic. On varying C, however, for example by exploring different field points, it is possible to encounter a caustic, i.e. an envelope of the ray family. To find the condition for this, we differentiate the ray equation (3.17):

$$\sum_i \frac{\partial^2 \phi}{\partial s_i \partial s_j} \, \mathrm{d}s_j + \sum_k \frac{\partial^2 \phi}{\partial s_i \partial C_k} \, \mathrm{d}C_k = 0. \tag{3.18}$$

On a caustic, it is possible to find a set of displacements $\mathrm{d}s_i$ (neighbouring rays) for which the $\mathrm{d}C_k$ vanish (the control point is unchanged to first order). Thus

$$\sum_i \frac{\partial^2 \phi}{\partial s_i \partial s_j} \, \mathrm{d}s_i = 0. \tag{3.19}$$

On a caustic, then, the generating function ϕ is stationary to higher than first order: the gradient map (3.17) is singular. The existence of a set of $\mathrm{d}s_i$ satisfying (3.19) implies the condition

$$\det\left\{\frac{\partial^2 \phi}{\partial s_i \partial s_j}\right\} = 0, \tag{3.20}$$

which together with (3.17) defines the caustics. In terms of the ϕ-"landscape", varying C across a caustic produces a change in the number of critical points s^μ; on the caustic itself, two or more critical points coalesce (rays touch), and s^μ is a *degenerate critical point*. For example, on moving across the caustic in fig. 27 from the "inside" of the cusp, the number of rays through each point changes from three to one.

According to eq. (3.20), the caustic is determined from the rays by one extra condition. Therefore without doing any more mathematics we can expect the caustic to occupy a region in control space whose dimensionality is one less than that of the control space itself: typically, caustics are of "codimension" unity. Therefore we expect generic caustics to be lines in the plane, or surfaces in space, and it is not suprising that we found point foci to be unstable, and that (as we shall learn) the focal line in fig. 28 is unstable too.

3.2. Classification of stable caustics

The upshot of the argument just given is: *caustics are singularities of gradient maps*. Their location in control space C is determined from a generating function $\phi(s, C)$ by the condition (3.20), once the rays $s^{\mu}(C)$ have been determined by eq. (3.17). To classify caustics it is therefore necessary to classify singularities of gradient maps. Precisely this has been achieved by the Thom–Arnol'd theorem [1, 33] of catastrophe theory. I will not prove the theorem, and indeed would not be able to; an outline of the proof is given by Poston and Stewart [2], and the full proof is given by Zeeman [34]; but I shall state the theorem, in a form suitable for application to caustics.

The set of all singularities (caustics) is partitioned into *equivalence classes*. Any two singularities S_1 and S_2 in the same class (and generated by different functions ϕ_1 and ϕ_2) can be locally transformed into one another by diffeomorphism, that is by smooth reversible transformations of the control parameters C. Diffeomorphism can be understood as follows: if S_1 is embedded in a control space made of rubber, then it can be transformed into S_2 by deforming the rubber without tearing or overlap or infinite strain. If S_1 and S_2 are in different equivalence classes, they cannot be transformed into one another by diffeomorphism. It is these equivalence classes that constitute the "catastrophes" with which the theory is concerned, and the classification of caustics can therefore be accomplished by a classification of equivalence classes (catastrophes). In language borrowed from phase transition theory, each catastrophe describes a "universality class" of caustics.

The classification proceeds in terms of *codimension* K, defined as the dimensionality of control space C minus the dimensionality of the singularity S. Alternatively stated, K is the dimensionality of the region in C that must be explored in order to encounter S. The reason for choosing K, rather than the dimensionality of the singularity itself, is that

the dimensionality of the singularity can be trivially altered by augmenting control space with extra parameters C: a point on a line, a curve in the plane, and a surface in space have different dimensionalities but they all have codimension $K = 1$. Loosely stated, K is the minimum number of essential control parameters of a space that contains the singularity.

It is a remarkable result of catastrophe theory that the caustics with which we are concerned are *structurally stable* if $K \leqslant 7$, in the sense that an arbitrary smooth perturbation of ϕ will not change the equivalence class to which the caustic belongs. Moreover, the equivalence classes include all caustics except a set of measure zero; this exceptional set contains isolated point foci, focal lines in space, and other unstable singularities. Therefore the form of almost every caustic is stably described by one of the catastrophes.

What are these forms? Each catastrophe can be represented by any generating function ϕ whose singularity is in its equivalence class. The simplest ϕ of each type is a polynomial whose terms are linear in C but non-linear in s; when written in this way they are called *normal forms* and I will denote them by $\Phi(s; C)$. A list of the normal forms for $K \leqslant 4$ is given in table 1, together with the names given to them by Thom [1] and the symbols by which Arnol'd [33] denotes them. The central theorem of catastrophe theory can now be stated as follows: any stable caustic with $K \leqslant 4$ is locally equivalent to the singularity obtained by eqs. (3.17) and (3.20) from one of the polynomials in table 1.

The forms of these stable singularities are shown on fig. 32 for $K \leqslant 3$. Let us work out the cusp case explicitly, to illustrate the way in which these forms are obtained. From table 1, the ray equation (3.17) gives

$$\partial\Phi/\partial s = s^3 + C_2 s + C_1 = 0, \tag{3.21}$$

Table 1
Standard polynomials Φ for the elementary catastrophes with codimension $K \leqslant 4$

Name	Symbol	K	$\Phi(s; C)$
Fold	A_2	1	$s^3/3 + Cs$
Cusp	A_3	2	$s^4/4 + C_2 s^2/2 + C_1 s$
Swallowtail	A_4	3	$s^5/5 + C_3 s^3/3 + C_2 s^2/2 + C_1 s$
Elliptic umbilic	D_4^-	3	$s_1^3 - 3s_1 s_2^2 - C_3(s_1^2 + s_2^2) - C_2 s_2 - C_1 s_1$
Hyperbolic umbilic	D_4^+	3	$s_1^3 + s_2^3 - C_3 s_1 s_2 - C_2 s_2 - C_1 s_1$
Butterfly	A_5	4	$s^6/6 + C_4 s^4/4 + C_3 s^3/3 + C_2 s^2/2 + C_1 s_1$
Parabolic umbilic	D_5	4	$s_1^4 + s_1 s_2^2 + C_4 s_2^2 + C_3 s_1^2 + C_2 s_2 + C_1 s_1$

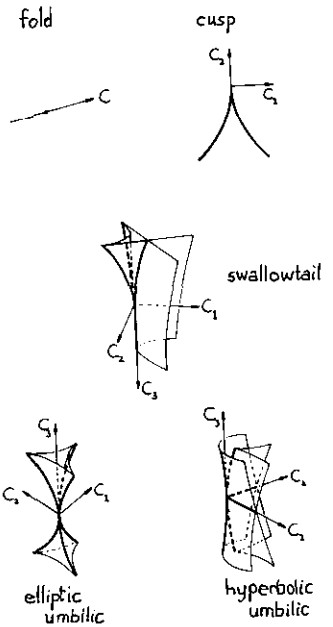


Fig. 32.

and the caustic condition (3.20) gives

$$\partial^2\Phi/\partial s^2 = 3s^2 + C_2 = 0. \tag{3.22}$$

On eliminating s we get the caustic condition

$$C_1 = \pm\left(\frac{-16C_2^3}{27}\right)^{1/2}, \tag{3.23}$$

which is indeed a cusp as shown in fig. 32.

From fig. 32 it is clear that the stable caustics along a line must be *fold* points ($K = 1$). In the plane, stable caustics may be fold curves ($K = 1$), or *cusp* points ($K = 2$). In space, stable caustics may be fold surfaces ($K = 1$), or cusp edges ($K = 2$) – which of course are not isolated focal lines – or singular points ($K = 3$) which may be of swallowtail, elliptic umbilic or hyperbolic umbilic type – of course these are not isolated focal points. As K increases, the singularities get more complicated; this can be understood either in terms of the fact that the higher singularities "contain" the lower ones (e.g., two cusp edges meeting at a swallowtail point) or by noting that at the most singular point in a caustic with

codimension K the number of coalescing rays [extrema $s^{\mu}(C)$] is $K+1$ and so increases with K.

Equivalence under diffeomorphism is a very powerful type of stability with important consequences for caustics. It implies, but is more restrictive than, topological stability, i.e. invariance under homeomorphisms, which are transformations that are merely continuous. To illustrate this idea, consider a curve

$$C_1 = \pm A C_2^{\alpha} \quad (\alpha > 0, C_2 > 0), \tag{3.24}$$

in a control space consisting of a rubber sheet with coordinates C_1, C_2. This has a cusp point at $C_1 = C_2 = 0$. By deforming the rubber it is possible to change A but not the exponent α. Therefore it is impossible to change a cusp catastrophe, where [cf. (3.23)] $\alpha = \frac{3}{2}$, into a finite-angled corner, where $\alpha = 1$, and this implies that no stable caustic curves in the plane have finite-angled corners. If catastrophe theory had operated instead with the weaker concept of topological stability, such corners would have been stable, because any cusp of type (3.24) can be continuously (but not differentiably) transformed into any other, even if the two cusps have different α. The existence of a finite classification scheme for caustics under such a strong stability criterion as diffeomorphism is surprising. Indeed the classification can only be carried out for $K \leqslant 7$. In higher codimensions [33], typical caustics are not stable under diffeomorphism, and a finite classification is only possible in terms of the weaker criterion involving homeomorphism.

The polynomials Φ in table 1 are written in terms of the fewest state variables s necessary to describe the singularity. This number, called the *corank* of the catastrophe, is defined as the reduction in rank of the Hessian matrix in eq. (3.20) as C moves from a typical point (i.e. not on the caustic) to the most singular point of the caustic (i.e. $C = 0$ for the cases in table 1). It is always possible to add extra state variables s to the generating function, in terms which are at most quadratic, without affecting the singularity (because ϕ is non-degenerate in these extra s-directions). The *cuspoid* catastrophes A_2, A_3, A_4 and A_5 have corank 1, and the *umbilics* D_4, D_5 have corank 2.

Considered as singularities in wave theory, it is very helpful, as we shall see, to regard catastrophes as *elemental atomic forms*, playing a role in the physics of waves analogous to atoms in the physics of matter. Like material atoms, these "atoms of form" can link together to form more complicated caustics (analogous to molecules) and, on a larger scale, regular or random "macroscopic" caustic networks (analogous to solids

or liquids) whose catastrophe structure is on too fine a scale to be clearly resolved. And like material atoms the morphological atoms possess a "microscopic" structure, consisting of characteristic patterns of diffraction fringes decorating the caustics. On this analogy, wavefront dislocations are the "elementary particles" of wave physics.

3.3. Optical examples of the simplest catastrophes

Here is the final paragraph of the novel "G", by Berger [35]:

> "The sun is low in the sky and the sea is calm. Like a mirror as they say. Only it is not like a mirror. The waves which are scarcely waves, for they come and go in many directions and their rising and falling is barely perceptible, are made up of innumerable tiny surfaces at variegating angles to one another – of these surfaces those which reflect the sunlight straight into one's eyes, sparkle with a white light during the instant before their angle, relative to oneself and the sun, shifts and they merge again into the blackish blue of the rest of the sea. Each time the light lasts for no longer than a spark stays bright when shot out from a fire. But as the sea recedes towards the sun, the number of sparkling surfaces multiplies until the sea indeed looks somewhat like a silver mirror. But unlike a mirror it is not still. Its granular surface is in continual agitation. The further away the richochetting grains, of which the mass become silver and the visibly distinct minority a dark leaden colour, the greater is their apparent speed. Uninterruptedly receding towards the sun, the transmission of its reflexions becoming ever faster, the sea neither requires nor recognizes any limit. The horizon is the straight bottom edge of a curtain arbitrarily and suddenly lowered upon a performance."

Obviously Berger is describing a familiar phenomenon: *the sparkling of sunlight on the sea*. This provides a very clear illustration of catastrophe theory in optics. The sun (fig. 33), situated at r_s, is a source of rays

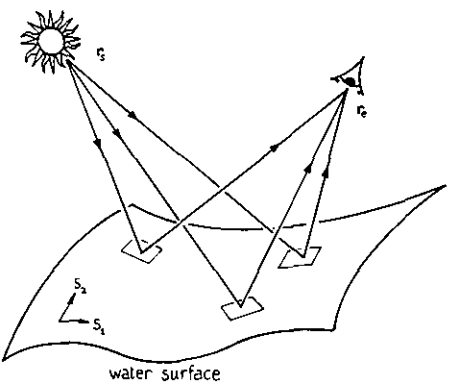


Fig. 33.

reflected by the wavy sea surface into the eye, situated at $\boldsymbol{r}_e$. Two coordinates $s = (s_1, s_2)$ on the water surface may be taken as the state variables, and the control parameters C are the time t of observation together with $\boldsymbol{r}_e$ and $\boldsymbol{r}_s$. The optical distance function $\phi(s_1, s_2; t, \boldsymbol{r}_e, \boldsymbol{r}_s)$ is simply proportional to the distance from the sun to the eye via s on the sea along straight paths. It is obvious by elementary reasoning that the "specular points" $s^u(C)$ which the eye, looking at the sea, sees as reflected images of the sun (fig. 34), are those points for which ϕ is stationary with respect to variations in s_1 and s_2.

Over short times, $\boldsymbol{r}_s$ and $\boldsymbol{r}_e$ may be considered fixed, so that the only effective control is t, which parameterizes the changing form of the water surface. Therefore the only stable catastrophes must be of codimension 1, i.e. *folds*. These correspond to moments when caustics of reflection, which are surfaces in the space above the water, sweep through the eye. At such

Fig. 34.

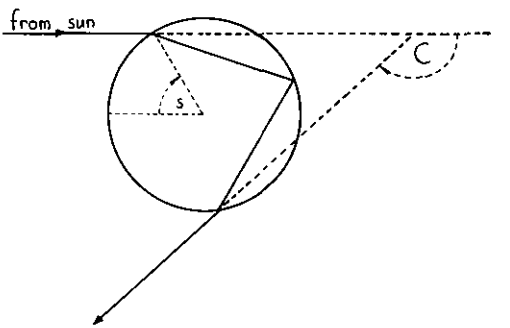


Fig. 35.

moments, which Longuet-Higgins [36] called "twinkles", pairs of specular points annihilate or are born. The vanishing of the determinant (3.20) at a twinkle means that the water surface has not only the right slope to direct sunlight into the eye but also the right curvature to focus it there.

My second example of a fold catastrophe is the *rainbow* [37]. Parallel sun-rays hit a raindrop (fig. 35), suffer two refractions and one internal reflection, and are viewed from afar. Owing to the rotational symmetry, there is only one control parameter C, which can be taken as the angle of deflection of the rays, and there need only be one state variable s, which can be taken as the latitude of the original point of incidence. Elementary optical laws show the graph of $s(C)$ (fig. 36) to be multivalued: there are two rays if $C > C_0 = 138°$ (for orange light) and none if $C < C_0$. The graph of $C(s)$ is, however, single-valued and has a minimum at $s = 59°$. At C_0, where $\partial C(s)/\partial s = 0$, the rays form a directional caustic, and this is a fold catastrophe which is indeed the only stable singularity with $K = 1$. In space the caustic surface is asymptotic to a cone with semiangle $180° - C_0 = 42°$. Each drop radiates such a cone, and what we see brightly lit when looking up into the sky is the locus of all raindrops on whose cones the eye lies, i.e. the rainbow.

Bright light reflected from the inner surface of *an illuminated tea-cup* (fig. 37) produces a caustic surface in space, whose intersection with the

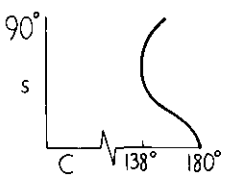


Fig. 36.

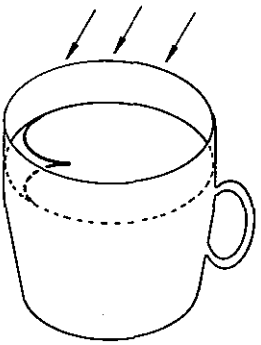


Fig. 37.

tea surface is visible by diffuse reflection as a caustic curve containing a cusp point. This is stable because there are two controls, corresponding to coordinates on the tea, and the stable catastrophes for $K = 2$ are just fold curves and cusp points.

If the cup is made of thin plastic, an extra control parameter becomes available, describing deformations of its circular cross section under slight pressure. Therefore caustics with $K = 3$ become stable, and indeed it is common to observe caustics such as that in fig. 38, which indicate a swallowtail catastrophe (cf. fig. 32).

A rich source of stable caustics is refraction by *irregular water-droplet "lenses"*. These lenses are easily made by allowing water to rain down onto a dusty flat glass surface, and the caustics formed by a laser beam transmitted through the glass (fig. 39) can be projected onto a screen. Since the screen is two-dimensional, the caustics on it will be those with $K = 2$, i.e. fold lines and cusp points. And indeed these are seen – most easily in the far field, as analyzed by Berry [32]. In the three-dimensional space beyond the droplet (and which can be explored section by section

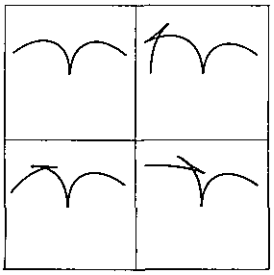


Fig. 38.

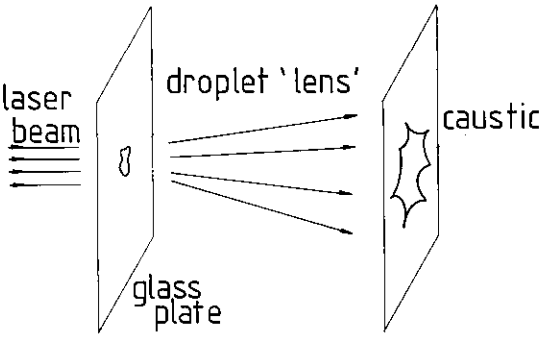


Fig. 39.

by the screen) caustics with $K = 3$ are stable. In addition to swallowtails, elliptic umbilic catastrophes can occur, as will now be shown.

Just beyond the droplet, which lies on the glass plate at $Z = 0$ (fig. 40), the refracted light forms a wavefront W whose height function will be denoted by $f(x, y)$, where x, y, coordinates in the plane of the plate, are the state variables s for this problem. The control parameters are the coordinates X, Y, Z of the field point. Since propagation takes place in free space without further refraction or reflection, the optical distance is

$$\phi(x, y; X, Y, Z) = \left\{(Z - f(x, y))^2 + (X - x)^2 + (Y - y)^2\right\}^{1/2}. \tag{3.25}$$

The rays (normal to W) originate from those positions x, y for which ϕ is stationary [cf. eq. (3.17)]. For droplets with gentle slopes, W is an

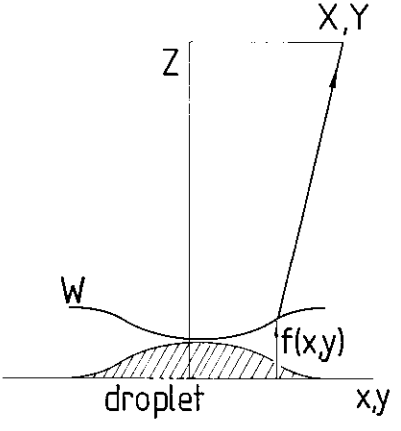


Fig. 40.

inverted scaled image of the droplet surface (fig. 40). The form of this surface is determined by surface tension: the sum of the two principal curvatures is proportional to the excess pressure in the drop; for a thin drop on a horizontal substrate, gravity can be neglected and this pressure is constant. Therefore $f(x, y)$ must satisfy the equation

$$\nabla^2 f(x, y) = 2/R, \tag{3.26}$$

where R is a constant equal to the mean curvature of W.

For our present purposes, it is sufficient to expand f about its minimum (corresponding to the summit of the drop), which will be assumed to lie at $x = y = 0$. Expanding up to quadratic terms is not enough, because that would correspond to a spherical wavefront giving rise to an unstable isolated focal point; it is necessary to include at least third-order terms. The corresponding solution of eq. (3.26) must be (up to rotation of coordinates):

$$f(x, y) = \frac{(x^2 + y^2)}{2R} - B(x^3 - 3xy^2), \tag{3.27}$$

where B is an arbitrary constant. The paraxial focus (i.e. the focal point when $B = 0$) lies at the centre of curvature of W, i.e. at $Z = R$, $X = Y = 0$. Expanding ϕ about this point in control space, and about the origin in state space, gives, up to cubic terms in x, y:

$$\phi(x, y; X, Y, Z) = Z + \frac{X^2 + Y^2}{2Z} + B(x^3 - 3xy^2) + \frac{R - Z}{2RZ}(x^2 + y^2) - \frac{Xx}{Z} - \frac{Yy}{Z}. \tag{3.28}$$

According to table 1, this is exactly the generating function for the elliptic umbilic catastrophe, apart from a trivial additive term, not involving x and y, and a factor B. Catastrophe theory guarantees that this is stable, so it is not necessary to expand to higher order. In this case ϕ is already in normal form – no diffeomorphism is necessary. The control parameters are

$$C_1 = \frac{X}{BZ}; \qquad C_2 = \frac{Y}{BZ}; \qquad C_3 = \frac{R - Z}{2BRZ}. \tag{3.29}$$

The "focus" of the droplet lens therefore has the local form of the elliptic umbilic caustic surface of fig. 32. A screen placed at $Z = R$ would show an isolated focal point; this is neither a fold curve nor a cusp point and so must be unstable, and indeed it is, because the slightest displacement of the screen away from $Z = R$ will cause the point to explode into a three-cusped deltoid curve, which is stable. Nye [38] gives a full analysis

of the caustics for these thin drops without gravity, and shows that there may be several elliptic umbilic foci whose unfoldings interact via swallowtails to give caustics with many-cusped sections close to the droplet and in the far field.

In subsection 3.4 we shall see that without gravity there can be no umbilic foci of hyperbolic type. When gravity is important, however – for example when the droplet lens hangs from a vertical substrate – *hyperbolic umbilic* catastrophes can occur, as will now be shown. With the glass plate vertical, let the x direction in fig. 40 be chosen downwards. Then the pressure in the drop changes linearly with x, and the right-hand side of eq. (3.26) must be augmented by a term Ax, where A is proportional to the acceleration due to gravity. A solution for the wavefront profile is

$$f(x, y) = \frac{x^2 + y^2}{2R} + \frac{Ax}{8}(x^2 + y^2). \tag{3.30}$$

When expanded up to third order in x and y, the optical distance becomes

$$\phi(x, y; X, Y, Z) = Z + \frac{X^2 + Y^2}{2Z} - \frac{Ax}{8}(x^2 + y^2) + \frac{R - Z}{2RZ}(x^2 + y^2) - \frac{Xx}{Z} - \frac{Yy}{Z}. \tag{3.31}$$

This is not identical with any of the normal forms on table 1, but the diffeomorphism

$$x \to x + y + \frac{8(R - Z)}{3ARZ}; \qquad y \to \sqrt{3}(x - y) \tag{3.32}$$

yields, up to trivial constants,

$$\phi(x, y; X, Y, Z) = x^3 + y^3 - \frac{4(Z - R)}{RZA}xy + \frac{2}{AZ}(X + Y\sqrt{3})x + \frac{2}{AZ}(X - Y\sqrt{3})y. \tag{3.33}$$

Reference to table 1 shows that this transformed function does correspond with the normal form for the hyperbolic umbilic catastrophe, with control parameters

$$C_1 = \frac{-2}{AZ}(X + Y\sqrt{3}); \qquad C_2 = \frac{-2}{AZ}(X - Y\sqrt{3}); \qquad C_3 = \frac{4(Z - R)}{RZA}. \tag{3.34}$$

The "focus" of this droplet lens therefore has the local form of the hyperbolic umbilic caustic surface of fig. 32. A screen placed at $Z = R$ would show a V-shaped caustic whose arms lie at 60° to one another; as we have seen, such a finite-angled corner is not a cusp catastrophe and so must be unstable, and indeed it is, because the slightest displacement of the screen away from $Z = R$ causes the V to split into a smooth outer curve and a cusped inner curve, which is stable. Nye [39] gives a full analysis of the caustics of drops under gravity, and shows that higher catastrophes play an important part in elucidating the behaviour of the caustics when a further control is introduced (for example, by rotating the substrate in its own plane and by letting gravity alter the drop shape without changing its perimeter).

For the final example in this selection of optical catastrophes, recall fig. 28, which shows the virtual caustic produced in space when light from a point source is refracted at a plane interface. At the singular point, a focal line meets a cusped cone. This morphology is not one of the three catastrophes with codimension 3 (fig. 32) and so is unstable, as previously asserted. Under perturbation, the caustic will break up into something stable. What will it look like then? The answer is that this cannot be predicted with certainty, because the unstable cone-line combination is the result of circular symmetry, whose breaking requires for its full description infinitely many parameters (e.g., coefficients in a Fourier expansion of the deviation from circularity). Therefore the unstable caustic can be considered as the singular section of a catastrophe with infinite codimension, and different choices of symmetry-breaking parameters will result in different stable "unfoldings" of the caustic in space. In this respect the cone-line combination is analogous to the "glory" effect in optics or potential scattering [32].

The simplest stable unfolding of the cone-line combination occurs in lens theory, when spherical aberration is perturbed with coma or astigmatism. In appendix 2 of ref [4], based on unpublished notes by Hannay, Berry and Upstill show that the caustic unfolds as shown in fig. 41; the structure is more clearly shown in the series of sections in fig. 42. The caustic begins (fig. 42a) with a "lips" event; this is not a catastrophe with $K = 3$ but merely a tangency of the screen (section) with a cusp edge joining two fold surfaces. The lips open, and another is born between them at (b). As these new lips open, their cusps hit the folds of the old lips in two hyperbolic umbilic points (c). The umbilic interchanges cusps and folds on its interpenetrating sheets, so that the new lips' cusps are transferred to the old lips, leaving a smooth curve and a four-cusped curve (d). The smooth curve expands as an almost-conical fold surface

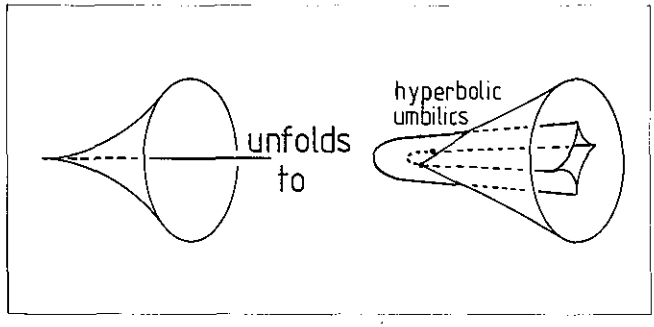


Fig. 41.

surrounding the four-cusped curve which extends into a cusped "needle". This new structure is stable against further perturbations, because it contains only folds, cusps and hyperbolic umbilics, all of which are catastrophes with $K < 3$ and hence stable in space. On the analogy in which catastrophes are regarded as "atoms", the structure in figs. 41 and 42 is a "molecule" consisting of several connected catastrophes.

The four-cusped needle into which the focal line explodes under perturbation was dramatically realized in the caustic produced by the atmosphere of Mars (fig. 29) acting as a lens focusing starlight, because Mars is not a sphere but (to a better approximation) an oblate spheroid, so that the circular symmetry of the focusing system is broken. On the earth's surface the needle appears in section as a four-cusped astroid about 100 km across, shown on fig. 43a together with the path of the telescope. Figure 43b shows the predicted intensity profile, with a small maximum arising from the close approach to the cusp and two sharp maxima produced by crossing the fold caustics. Figure 43c shows the fine structure observed [31] in the central flash of fig. 30. Evidently the agreement is excellent. I do not know a more striking example of the

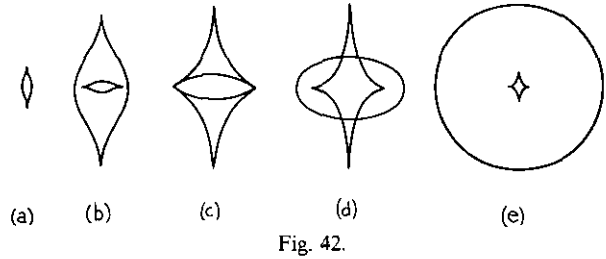


Fig. 42.

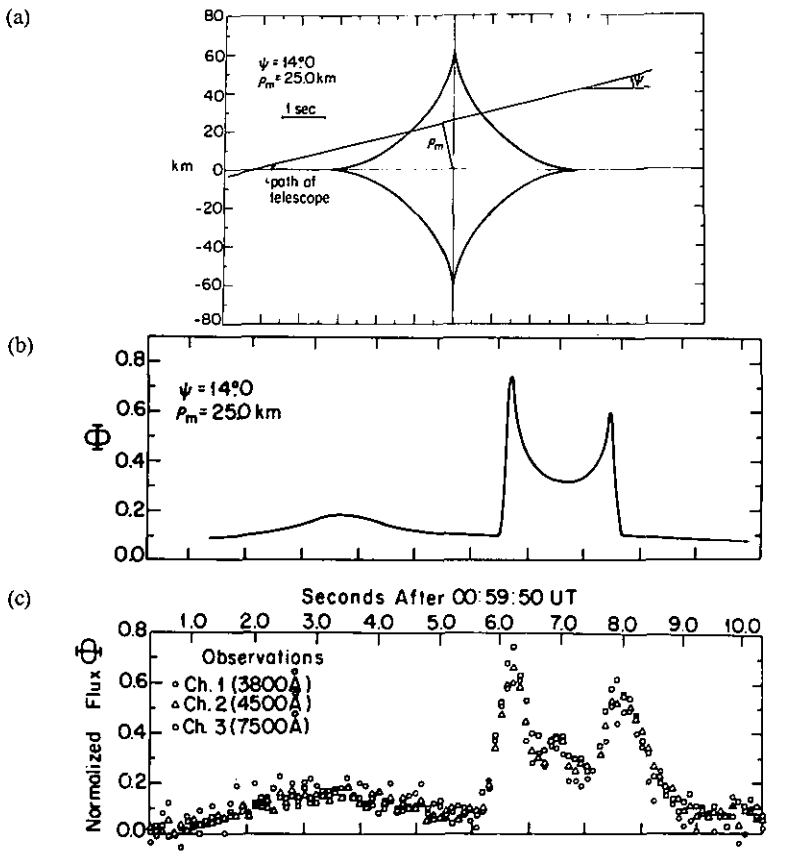


Fig. 43. (From ref. [31].)

emergence of structural stability when a perturbation forces a singularity to adopt a generic morphology.

3.4. Umbilic points: different classifications of the same singularity

It will prove very instructive, not only in terms of understanding catastrophe theory but also more generally, to study the geometry of the elliptic and hyperbolic umbilics. From table 1, we see that these have corank 2 and so occur only as envelopes of ray families containing at least two dimensions worth of rays. A model satisfying this condition is the family of rays consisting of the normals to a wavefront W which is a

gentle deformation of a plane. This has already been introduced (fig. 40) in connection with the optics of droplets. The generating function is given by eq. (3.25), which in view of the assumed gentleness of W can be replaced by the lowest terms of its expansion for $(X-x)/Z$ and $(Y-y)/Z$ small, since the rays will make small angles with the Z axis. Thus

$$\phi(x, y; X, Y, Z) = Z - f(x, y) + \frac{1}{2Z}\left[(x-X)^2 + (y-Y)^2\right]. \tag{3.35}$$

From the point of view of singularity theory this "paraxial" replacement should not be regarded as an approximation as long as it results in stable singularities, because these are locally equivalent to the "exact" singularities generated by eq. (3.25).

When applied to eq. (3.35), the caustic condition (3.20) gives, on denoting derivatives by subscripts,

$$\det\begin{Bmatrix} -f_{xx} + 1/Z & -f_{xy} \\ -f_{xy} & -f_{yy} + 1/Z \end{Bmatrix} = 0. \tag{3.36}$$

Therefore rays from x, y on W touch their caustics at two levels Z which on solving the quadratic equation are found to be

$$1/Z = \tfrac{1}{2}\left\{f_{xx} + f_{yy} \pm \left[\left(f_{xx} - f_{yy}\right)^2 + 4f_{xy}^2\right]^{1/2}\right\}. \tag{3.37}$$

These two values of $1/Z$ are the two *principal curvatures* of W, showing that the bundle of rays (normals) near x, y comes to focus twice, at each of the two centres of curvature of the wavefront.

It is clear that the caustic consists of two sheets. Each sheet is a fold surface which may crease into cusp edges, which in turn may join up at swallowtail points; all these catastrophes have corank 1. In addition, however, the two sheets can be connected; according to eq. (3.37), this event requires two conditions to be satisfied, namely

$$f_{xx} = f_{yy}; \qquad f_{xy} = 0. \tag{3.38}$$

Therefore it occurs with rays emerging from isolated points on W, where the two principal curvatures are equal, so that W is locally spherical. Such points are called *umbilic points* (for an obvious reason), and the corresponding joining of caustic surfaces can occur stably in two ways, which are the elliptic and hyperbolic umbilic catastrophes of fig. 32.

To study an umbilic in more detail, we can assume without loss of generality that it is situated at $x = y = 0$, and that W has zero slope and

two (equal) radii of curvature R. Up to cubic terms, the wavefront is therefore given by

$$f(x, y) = \frac{x^2 + y^2}{2R} + \frac{1}{6}(\alpha x^3 + 3\beta x^2 y + 3\gamma x y^2 + \delta y^3). \qquad (3.39)$$

α, β, γ and δ are the four third derivatives of f at the umbilic, and parameterize the different types of singularity; you should think of $\alpha, \beta, \gamma, \delta$ as the "space of umbilics".

Substitution of eq. (3.39) in eq. (3.37) gives the "curvature landscape" above the umbilic point as

$$1/Z = 1/R + \tfrac{1}{2}\Big\{(\alpha + \gamma)x + (\beta + \delta)y \pm \Big[\{(\alpha - \gamma)x + (\beta - \delta)y\}^2 + 4(\beta x + \gamma y)^2\Big]^{1/2}\Big\}. \qquad (3.40)$$

This shows that in Z, x, y space the curvatures form a double cone connected at the umbilic. For fixed Z the contours of curvature are ellipses or hyperbolae, depending on the values of α, β, γ, δ, which determine the inclination of the cone. In the control space X, Y, Z above W, elliptic contours imply that the caustic is an elliptic umbilic catastrophe (E), and hyperbolic contours imply a hyperbolic umbilic catastrophe (H). In the space of umbilics, these two sorts of singularity – equivalence classes – are separated by the discriminant of (3.40), yielding the following condition:

If

$$C(\alpha, \beta, \gamma, \delta) \equiv 4(\alpha\gamma - \beta^2)(\beta\delta - \gamma^2) - (\alpha\delta - \beta\gamma)^2 \begin{cases} > 0 & \text{then } E, \\ < 0 & \text{then } H. \end{cases} \qquad (3.41)$$

When applied to the elliptic and hyperbolic umbilic normal forms in table 1, this criterion is easily seen to work properly. In the case of the generating function (3.31), which is not in normal form, $\beta = \delta = 0$, $\alpha = -3A/4 = 3\gamma$, eq. (3.41) gives $C < 0$, confirming the result arrived at by the diffeomorphism (3.32), i.e. that the singularity is hyperbolic. For thin water-droplet lenses with gravity neglected, differentiation of the wavefront profile (3.26) gives $\alpha + \gamma = \beta + \delta$ whence eq. (3.41) gives $C = 4(\alpha^2 + \beta^2)^2$, which is always positive, justifying the assertion in subsection 3.3 that the only umbilic caustics produced by such drops are elliptic.

Now we consider umbilic points in a quite different way. At almost every point x, y on W, there exists a pair of orthogonal "principal directions" along which the curvature of $f(x, y)$ is a maximum or a minimum. The exceptional places are the umbilic points, where f is locally spherical so that the principal directions are indeterminate. Therefore umbilic points take on a new significance, as *singularities of the net of lines of curvature*. One way to describe this singularity is by its *index I*, defined as the number of rotations of the curvature cross during a circuit of the umbilic point, i.e. by

$$I \equiv \frac{1}{2\pi}\oint \nabla\theta(\boldsymbol{r})\cdot \mathrm{d}\boldsymbol{r}, \tag{3.42}$$

where $\theta(\boldsymbol{r})$ is the angle made by a line of curvature with the x-axis and $\boldsymbol{r}$ denotes (x, y).

To calculate I we must first find the direction $\theta(\boldsymbol{r})$ at each point. Along an arbitrary direction θ, the curvature of f is

$$\left(\cos\theta\frac{\partial}{\partial x}+\sin\theta\frac{\partial}{\partial y}\right)^2 f = \cos^2\theta f_{xx} + 2\cos\theta\sin\theta f_{xy} + \sin^2\theta f_{yy}. \tag{3.43}$$

Differentiating with respect to θ then gives the two (orthogonal) curvature directions as

$$\tan 2\theta(\boldsymbol{r}) = 2f_{xy}(\boldsymbol{r})/\left[f_{xx}(\boldsymbol{r}) - f_{yy}(\boldsymbol{r})\right]. \tag{3.44}$$

On substituting the form (3.39) for f and performing the (tricky) loop integral (3.42) we get the following criterion:

If

$$J(\alpha, \beta, \gamma, \delta) \equiv \alpha\gamma - \gamma^2 + \beta\delta - \beta^2 \begin{cases} > 0 & \text{then } I = +\frac{1}{2}, \\ < 0 & \text{then } I = -\frac{1}{2}. \end{cases} \tag{3.45}$$

This is a surprising result, because it shows that the index classification of umbilics is different from the catastrophe classification.

Another surprise is in store. The umbilic can also be described in terms of the number of straight lines of curvature passing through it. These are lines for which the curvature direction $\theta(\boldsymbol{r})$ coincides with the direction of the vector $\boldsymbol{r}$ itself. Equations (3.44) and (3.39) then give, for these directions θ,

$$\beta\cos^3\theta + (2\gamma - \alpha)\cos^2\theta\sin\theta - (2\beta - \delta)\cos\theta\sin^2\theta - \gamma\sin^3\theta = 0. \tag{3.46}$$

This cubic equation may have two or six roots with $0 \leqslant \theta < 2\pi$, so that the number of straight curvature lines passing through the umbilic is 1 or 3. This "pattern classification" of the curvature singularities depends on the discriminant of eq. (3.46), and is as follows:

If

$$P(\alpha, \beta, \gamma, \delta) \equiv 4\{3\gamma(\alpha - 2\gamma) - (\delta - 2\beta)^2\} \times \{3\beta(\delta - 2\beta) - (\alpha - 2\gamma)^2\} - \{(\delta - 2\beta)(\alpha - 2\gamma) - 9\beta\gamma\}^2 \begin{cases} > 0 & \text{then 3 lines,} \\ < 0 & \text{then 1 line.} \end{cases} \quad (3.47)$$

This classification differs from both the index classification and the catastrophe classification.

As a result of this analysis, we see that the same singularity – the umbilic point – can be classified in three different ways, i.e. there are three different ways of partitioning the space of umbilics into universality classes. This should serve as a caution to anybody who thinks that a singularity (e.g. in a liquid crystal) can be given a topological assignment in only one way.

The classifications interlock, because the inequalities (3.41), (3.45) and (3.47) imply

$$\left.\begin{array}{r} \text{1 line through umbilic} \rightarrow I = +\frac{1}{2} \\ E \rightarrow I = -\frac{1}{2} \end{array}\right\}. \quad (3.48)$$

The interlocking of the index and pattern classifications leads to three stable arrangements of lines of curvature round an umbilic point, illustrated in fig. 44. The *star* (S) has $I = -\frac{1}{2}$ and 3 lines through the umbilic; the *monstar* (M) has $I = +\frac{1}{2}$ and 3 lines; the *lemon* (L) has $I = +\frac{1}{2}$ and 1 line. In view of (3.48), all elliptic umbilics are stars. The final partitioning of the set of umbilic points is shown in fig. 45.

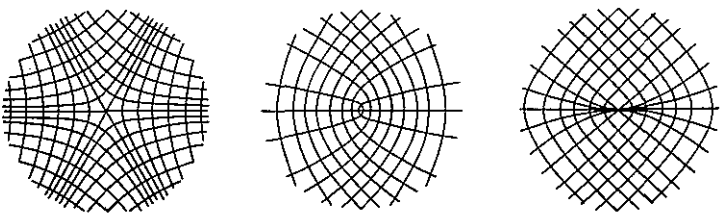


Fig. 44.

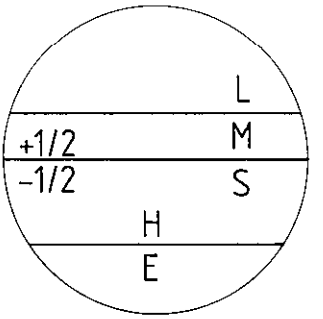


Fig. 45.

The catastrophe classification has an obvious application to caustics: as we have seen, it determines the way in which two caustic sheets can stably join. The index classification has a less obvious application, in restricting the "umbilic reactions" by which these singularities can interact and transform. To see this, consider the (almost obvious) topological theorem that the total index of all the singularities within a closed curve cannot change, provided no singularity crosses the curve. This implies that no isolated singularity may spontaneously change its index, and when singularities interact the total index is conserved. Therefore two elliptic umbilics cannot spontaneously appear or annihilate, and nor can an elliptic and a hyperbolic umbilic do this; according to fig. 45 the only stable reaction with $K < 4$ is

$$H(-\tfrac{1}{2}) + H(+\tfrac{1}{2}) \leftrightarrow \text{nothing}. \tag{3.48a}$$

A transformation is possible across the E–H boundary, since this is all contained in the region $I = -\frac{1}{2}$, and takes place via the ($K = 4$) parabolic umbilic singularity:

$$E(-\tfrac{1}{2}) \leftrightarrow \text{parabolic umbilic} \leftrightarrow H(-\tfrac{1}{2}). \tag{3.49}$$

I do not know whether the pattern classification (3.47) has any application to caustics.

Although the triple classification of umbilics was implicit in classical works on differential geometry, it was first described in 1971 by Porteous [40] (albeit in a form very different from the way I have presented it, and without explicit discriminant formulae). The "monstar" morphology does not appear to be widely known, either as a possible liquid crystal singularity or in fingerprints (where lemons and stars appear). Berry and Hannay [41] studied the statistics of the different types of umbilics on a

Gaussian random surface: there are on average equal numbers with $I = \pm\frac{1}{2}$, and for isotropic randomness 26.8% are elliptic and 5.3% are monstars. Nye and his collaborators [42, 43, 65, 66] have extended the umbilic idea to general tensor fields (which cannot be defined in terms of the second derivatives of a height function).

3.5. Caustic networks

Now I am going to consider the caustics enveloped by the family of rays normal to an extended undulating wavefront W. Such wavefronts generate very complicated caustic surfaces in space, whose intersection with planes ("screens") consist of elaborate networks of caustic lines. An extended undulating wavefront is produced in sunlight refracted by the randomly-rippling surface of water in a swimming pool, and the caustic networks are visible in diffuse reflection as dancing patterns of bright lines on the bottom of the pool. In view of the fact that almost every person who has ever lived must have seen these patterns (on the bottoms of rivers or ponds, or on the sea bed), it is astonishing to discover that they have, so it seems, never been studied before. I suppose that before the "singularity" approach to optics made it possible to concentrate on the morphological aspects of the caustics which are most directly perceived, such a study would have appeared either too trivial ("just Snell's law") or tediously technical ("only with the aid of a computer could the details of the patterns be understood").

Starlight also produces an extended undulating wavefront, after passing through a turbulent atmosphere. For strong turbulence and viewing near the horizon the undulations may be strong enough to focus the light onto caustics that can sweep through the eye. The deflection angles are too small to enable the eye to resolve multiple images of the star as was the case in the sparkling of sunlight on the sea, discussed at the beginning of subsection 3.3, and starlight is too faint for caustic networks to be seen moving across the ground. However, the caustics can be detected by means of the violent polychromatic intensity fluctuations of the star as seen by eye and which are, of course, familiar as "twinkling".

The undulations, which may be regular or random, will as before be described by the deviation $f(x, y)$ of W from a reference plane, and the rays and caustics in the control space X, Y, Z are determined by the generating function (3.35).

In the simplest case, the water waves are all travelling in one direction, and so W can be described by a function $f(x)$ of one variable, and only

corank 1 catastrophes can occur. Because the control space is effectively two-dimensional (X, Z), these are folds and cusps. Rays from x touch the caustic (once) at the centre of curvature

$$Z = 1/f_{xx}. \tag{3.50}$$

The far field caustics ($Z = \infty$) therefore originate from inflections of f ($f_{xx} = 0$). It is not hard to show that the caustic has cusps originating from extrema of the curvature of f ($f_{xxx} = 0$). Therefore the caustic consists of branches (fig. 46) beginning and ending in the far field and containing an odd number of cusps. The branches cross repeatedly en route to the far field.

In three dimensions, such a long-crested train of water waves produces caustics consisting of fold surfaces connected by straight horizontal cusp edges, as shown in fig. 47 for a sinusoidal wave train, where each pair of fold surfaces possesses only one cusp edge. If the pool is too shallow, no caustics will be seen, because the cusps would lie below the level of the bottom. If the depth of the pool is slightly below the level of the cusps, each wave crest will image onto a bright pair of closely-spaced lines. Because the interior of each cusp is lit by three rays, whilst the exterior is lit by only one ray, the region between the lines in each pair is brighter than the regions between adjacent pairs. Figure 48 shows this very clearly.

You might think that by making the pool much deeper, or the waves steeper, it would be possible to see the line pairs widen and cross as a result of the intersection of caustics as shown in fig. 46. But because of a

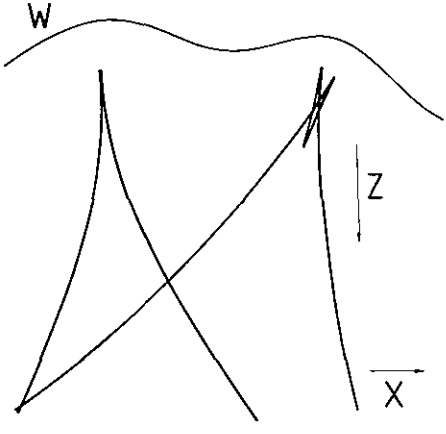


Fig. 46.

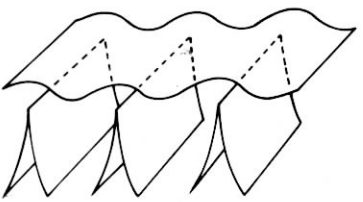


Fig. 47.

curious "uncertainty principle" this is rarely seen in practice. The principle is a consequence of the finite angular width $\delta(\approx \frac{1}{2}°)$ of the sun's disc, which means that caustics at depth Z are blurred over horizontal distances $Z\delta$. If the water waves have length $\sim L$, the caustic pattern at any depth repeats over horizontal distances L, so that the pattern is totally obscured if $Z > L/\delta$, and much of the detail is obscured even if $Z > L/10\delta$. On the other hand, if the waves have amplitude $\sim A$, the cusps occur at a level $Z \sim L^2/A$ and so no caustics are seen if Z is less than this. Therefore caustics can be observed only if

$$L^2/A < Z < L/10\delta. \tag{3.51}$$

Intersections of caustics, and especially the far field, are thus unlikely to be observed, and where individual wave crests can be discerned they will probably appear in the form of line pairs as on fig. 48.

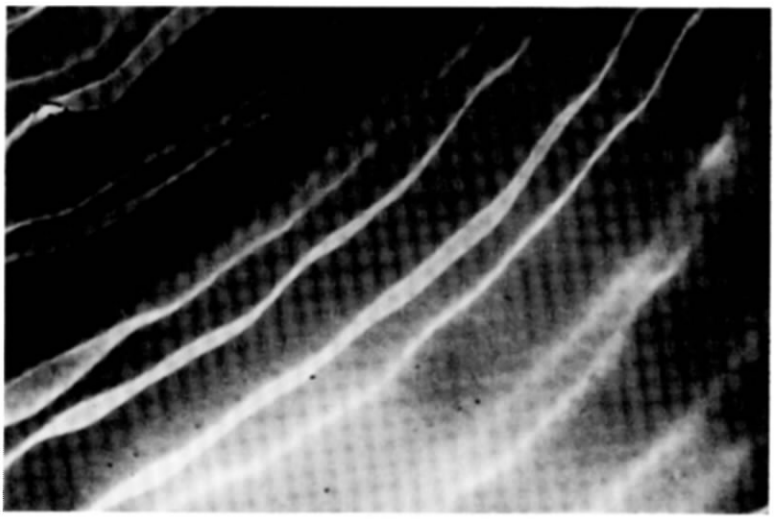


Fig. 48.

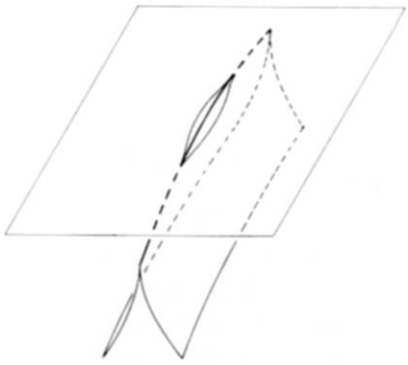


Fig. 49.

Now let the one-dimensional waves be perturbed by a weak modulation in the y-direction. Since the caustic in fig. 47 is stable, the perturbed caustic will have the same form, but the cusp edges will now be curves in space rather than horizontal straight lines. If a cusp edge rises briefly above the bottom of the pool (fig. 49), the caustic will appear as a "lips" singularity (cf. fig. 42). Some fine examples of lips are shown on fig. 50.

If the modulation of the long-crested waves is stronger, then the cusped caustic sheets can interact (instead of merely intersecting), and the problem becomes essentially of corank 2 and codimension 3. Upstill

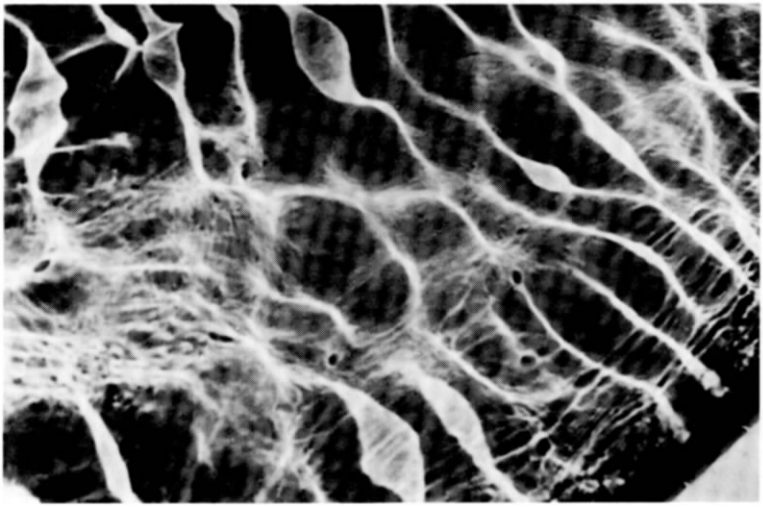


Fig. 50.

[44, 45, 4] analyzed the model where W consists of just two superimposed sinusoidal wave trains, namely

$$f(x, y) = \cos x + A\cos(qx + py). \tag{3.52}$$

Although apparently simple, this model gives rich caustic patterns and has surprising explanatory power.

The first step in understanding the caustic is to find the umbilic points on W. f is a periodic function, so that only one unit cell need be examined. From the conditions (3.38), (3.41) and (3.45) it follows easily that each unit cell has four umbilic points, all hyperbolic, two with index $+\frac{1}{2}$ and two with index $-\frac{1}{2}$. The umbilic foci all lie in the far field ($Z = \infty$) and so they cannot be seen in swimming-pool caustic networks, although they do play a part in organizing the patterns.

I shall discuss just one of Upstill's patterns, corresponding to the (representative) case where $A = 0.95$, $\arctan p/q = 64°$, $p^2 + q^2 = 1$. This is shown on fig. 51. Although at first giving the illusion of being made up of sinuous fold lines, the pattern in fact consists of giant overlapping lip shapes. The illusion introduces an important new principle, which I now discuss more carefully.

Suppose fig. 51 is observed under conditions of poor resolution, as will often be the case in practice because of the blurring due to the sun's disc, embodied in (3.51). Then the cusps and the closely-spaced line pairs will not be discerned, and the pattern will appear as in fig. 52. As fig. 53 indicates, this type of imperfectly resolved caustic network is often observed in practice. A naïve attempt to apply catastrophe theory to it by considering the sinuous curves of fig. 52 to be fold caustics would lead to contradiction, because the principle that the number of contributing rays changes by two across a fold could not be satisfied. Therefore under poor resolution caustic patterns can appear – stably – which are not catastrophes. What we are seeing here is the "macroscopic" level of the catastrophe theory of waves, where many linked catastrophes generate new morphologies.

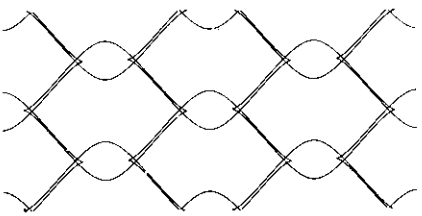


Fig. 51. (From ref. [4].)

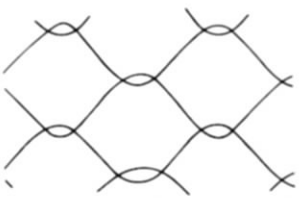


Fig. 52.

A natural next step is to study the caustics produced by three water waves. Unless the three wave vectors generate a lattice, W will not be periodic and so the caustic pattern will never repeat. This is of course the situation in nature, but (although some non-periodic caustics have been studied) it proved easier to understand the morphological elements of the pattern by considering periodic cases. Upstill [45] calculated the caustics of the wave front

$$f(x, y) = \cos x + \cos(\tfrac{1}{2}x + qy) + \cos(\tfrac{1}{2}x - qy), \tag{3.53}$$

whose three waves are of equal amplitude, one travelling in the x direction and the other two travelling in directions making equal angles

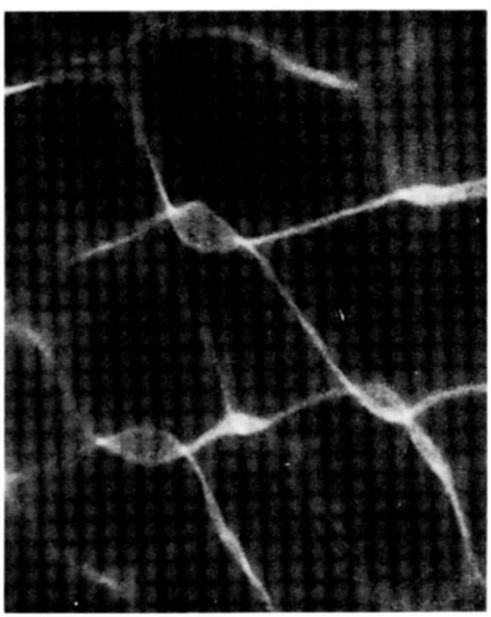


Fig. 53.

arctan $2q$ with it. W is periodic with a hexagonal unit cell (which is regular if $q = \sqrt{3}/2$).

For this wavefront there are umbilic points whose curvature is finite, and which therefore focus in the near field and so dominate the caustics. Each unit cell contains an elliptic umbilic point near each of the six vertices and two hyperbolic umbilic points with index $+\frac{1}{2}$ near the centre. Because each elliptic umbilic can be regarded as "shared" by the three unit cells connected at its vertex, it contributes an index $\frac{1}{3} \times (-\frac{1}{2}) = -\frac{1}{6}$ to the total index of a cell, which is therefore $6 \times (-\frac{1}{6}) + 2 \times (+\frac{1}{2}) = 0$. Therefore, just as in the two-wave case (3.52), the net index of W in the large is zero. This must be the case for topological reasons: the total index of all singularities on any unbounded surface equals $1/2\pi$ times the integrated Gaussian curvature (Gauss–Bonnet theorem) and so must vanish for a wavefront without overall curvature.

An important feature of W is the fact that the sign of the curvature at elliptic umbilic points is opposite to that at the hyperbolic points, so that the caustic patterns in the real and virtual ray families are different. This is a consequence of the fact that the three-wave W [eq. (3.53)], as opposed to the two-wave W [eq. (3.52)], is non-invertible in the sense that $-f$ is not merely a displaced version of $+f$. In water, the real caustics of the refracted wavefront are those on the bottom of the pool, and the virtual caustics of this wavefront are similar to the real caustics of the reflected wavefront, and can often be seen on the sides of boats or the undersides of bridges.

Consider first the real caustics, which are dominated by elliptic umbilics. A typical section of the pattern for the regular-hexagon unit cell is shown in fig. 54. The caustics consist of linked "Olympic rings" with tiny unfolded elliptic umbilic cusped triangles in each junction. These *elliptic*

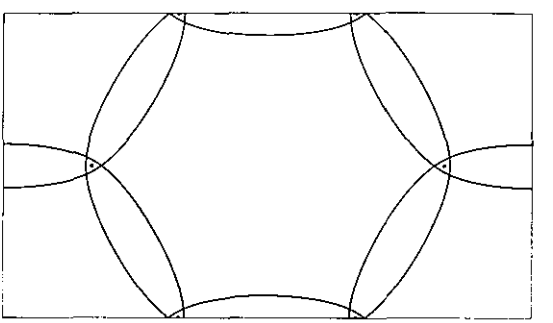


Fig. 54. (From ref. [45].)

umbilic centred triple junctions were studied in detail by Berry and Nye [46] by refracting laser light through a glass "lollipop" on which three smooth grooves (representing three waves on W) had been impressed; fig. 55 shows a junction obtained in this way. For a typical orientation of the lollipop, the focal sequence of patterns as Z varies can contain a complicated sequence of stable events, as fig. 56 illustrates: there is an elliptic umbilic catastrophe, three swallowtail catastrophes, and three "beak-to-beak" events where fold lines touch and separate into two cusps (these are not $K = 3$ catastrophes but simply tangencies of the observation plane with a curved cusp edge). Sequences such as this occur in the junctions of the caustics generated by the model wavefront (3.53).

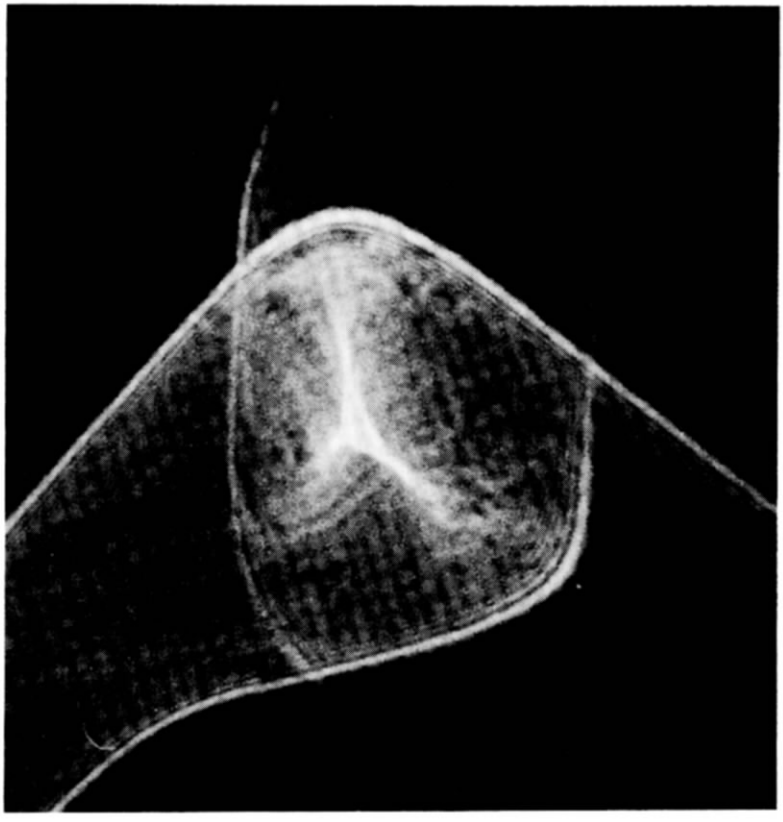


Fig. 55.

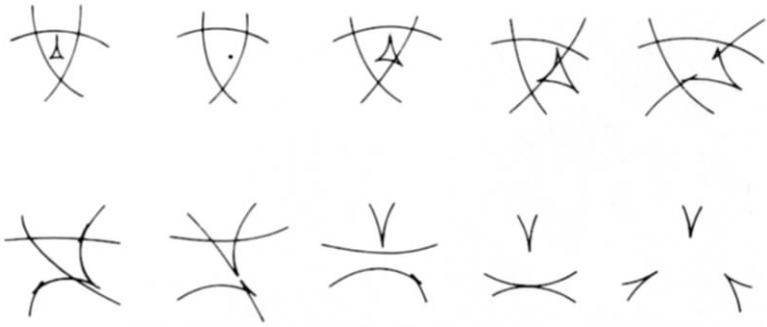


Fig. 56.

Under poor resolution, the detail in the centre of the junction, where the three line pairs meet, cannot be resolved; this is illustrated in fig. 57. Under worse resolution, even the line pairs cannot be resolved, and the caustics appear as a hexagonal network whose lines meet in threes, as in fig. 58, which was produced in laser light focussed by irregular "bathroom window" glass. Such triple junctions are not catastrophes (because cusps are the only stable point foci in the plane), and provide another example of a non-catastrophic morphology emerging stably at the "macroscopic" level.

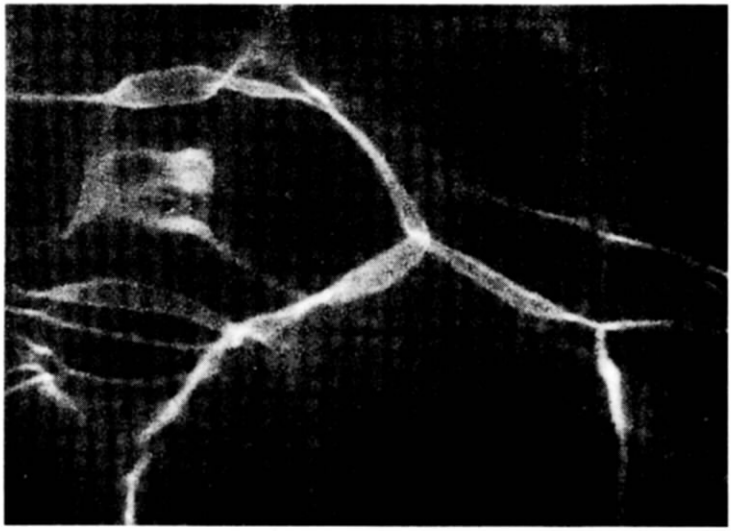


Fig. 57.

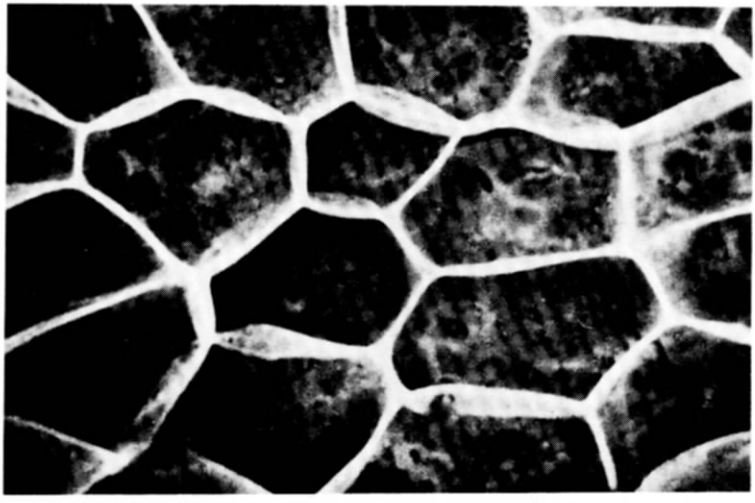


Fig. 58.

Consider now, briefly, the virtual caustics, which are dominated by hyperbolic umbilics. Here a typical morphology (fig. 59) consists of complicated "spangles" linked by closely-spaced line pairs, which can be seen in "bathroom window" caustics (fig. 60).

The caustic networks generated by these two- and three-wave models contain a great complexity which I have not had time to do more than hint at. Caustic networks in nature are more complicated still, and would require for their complete description an n-wave model, where n is large. In the limit $n \to \infty$, f becomes a Gaussian random function of x and y whose caustics have been understood to a certain extent by statistical studies of the distribution of umbilics (cf. subsection 3.4) [41, 4], folds and cusps [4]. Another approach is to concentrate on the junctions, whose complexity (as in fig. 56, for example) is organized to a large extent by the catastrophe X_9 [45, 33] which has codimension 8.

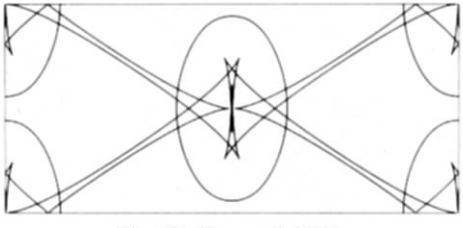


Fig. 59. (From ref. [45].)

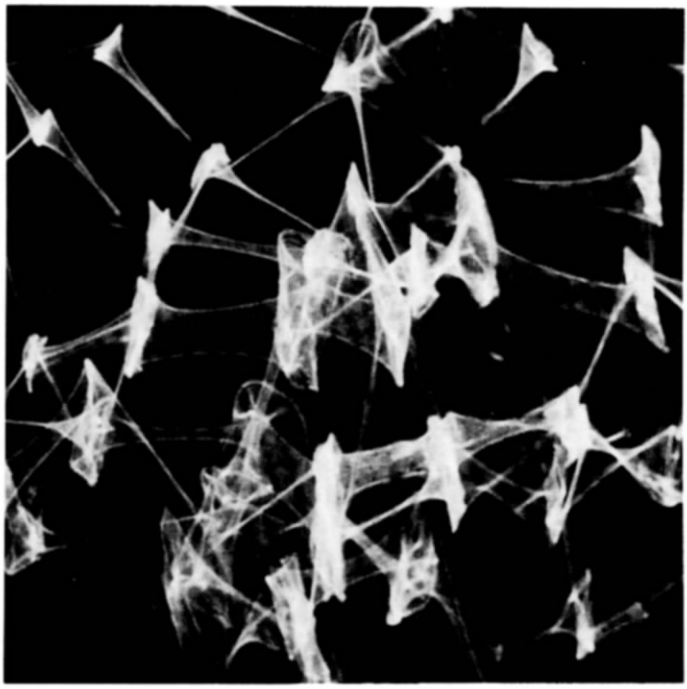


Fig. 60.

4. Diffraction catastrophes

4.1. Integral representations in shortwave asymptotics

In discussing caustics, I have used the language of trajectory theory throughout. However this is an appropriate thing to do only in the shortwave limit, i.e. as $k \to \infty$, where k is 2π/wavelength (or, in quantum mechanics, as $\hbar \to 0$, where $\hbar$ is Planck's constant). What becomes of caustics when k is not infinite but merely large? Then it is necessary to speak about waves, described by wave functions ψ, rather than families of trajectories, described by generating functions ϕ. The answer will be that caustics are not singularities of the wave but places at which the wave intensity $|\psi|^2$ is large and near which the wave pattern (as embodied for example by the contours of $|\psi|^2$) is organised in a manner dependent on which catastrophe describes the caustic. In a very real sense caustics are *emergent* morphologies: as k increases, they become

ever more prominent, and when $k = \infty$ they become genuine singularities of the intensity. In terms of our analogy in which catastrophes are regarded as atoms of form, we are now going to step down to the microscale, to peer into the "subatomic" details of the diffraction patterns that decorate caustics on wavelength scales.

From a mathematical viewpoint, what I am talking about are asymptotic solutions of the wave equation as $k \to \infty$, but I intend to discuss the problem without writing any wave equation. This might seem like a conjuring trick, but I find the physical arguments more instructive than formal asymptotics, which in any case can be found published elsewhere [4, 47, 48, 50].

Wave functions ψ exist in the same space as caustics. Therefore their variables are the control parameters C introduced in subsection 3.1 (time, coordinates in space, parameters describing the shape and refractive index of scatterers, etc.). It is natural to expect that when k is large a first approximation to ψ can be constructed as a superposition of contributions from rays passing through C. These rays are defined in terms of the optical distance function $\phi(s, C)$ by the gradient condition (3.17), whose solutions are $s^{\mu}(C)$, where μ is a label denoting the different rays.

The phase of the contribution of the μth ray must be

$$k\phi(s^{\mu}(C); C) \equiv k\phi_{\mu}(C). \tag{4.1}$$

This is simply 2π times the optical distance, measured in wavelength units along the μth ray, to the point C from some initial wavefront W on which $\phi = 0$. (In mechanics, $k\phi$ would be (action)$/\hbar$.) The initial wavefront depends on what boundary conditions ψ is required to satisfy: for example, W might be an undulating wavefront (as in figs. 40, 46, 47) or it might represent a source point (as in figs. 26 and 27). To find the amplitude of the contribution of the μth ray, realise that the energy density is proportional to $|\psi|^2$, i.e. (amplitude)^2, and that as $k \to \infty$ energy travels along rays and is conserved. Therefore the amplitude is inversely proportional to the square root of the cross-sectional area of a tube of rays.

The wave constructed in this way will not be written in its general form; to do that would complicate the formalism. Instead I shall proceed from now on by using the example, introduced in subsection 3.3, where waves propagate from an initial wavefront W consisting of a gentle deviation from the plane $Z = 0$. It is convenient to introduce the two-dimensional vectors

$$\boldsymbol{r} \equiv (x, y); \qquad \boldsymbol{R} \equiv (X, Y) \tag{4.2}$$

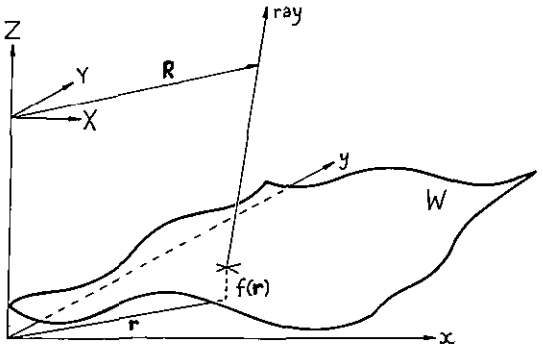


Fig. 61.

(fig. 61), so that the state variables are now $s = r$ and the control parameters are $C = (R, Z)$. If the height function of W is $f(r)$, the optical distance function (3.25) in its paraxial approximation (3.35) can now be written as

$$\phi(r; R, Z) = Z - f(r) + \frac{|R - r|^2}{2Z}. \tag{4.3}$$

The rays through (R, Z) are given [eq. (3.17)] by the solutions $r^\mu(R, Z)$ of

$$\nabla f(r) = \frac{r - R}{Z}. \tag{4.4}$$

For construction the approximate wave function $\psi(R, Z)$, we require the cross section of a ray tube. This is simply the Jacobian determinant describing the mapping (4.4) from $\mathrm{d}r^\mu$ to $\mathrm{d}R$. Therefore, by the arguments of the preceding paragraph:

$$\psi(R, Z) \approx \sum_\mu \left[\det\left\{\frac{\partial r^\mu}{\partial R}(R, Z)\right\}\right]^{1/2} \exp\{\mathrm{i}k\phi_\mu(R, Z)\}. \tag{4.5}$$

As a shortwave (or, in quantum mechanics, semiclassical) approximation, this has considerable merit. Firstly, it describes the interference between the contributions of the rays through R, Z. Secondly, it shows that wave functions ψ are non-analytic in k as $k \to \infty$, so that shortwave approximations cannot be expressed as a series in powers of $1/k$, i.e. deviations from the shortwave limit cannot be obtained by perturbation theory. Thirdly, in the shortwave limit itself, ψ oscillates infinitely fast as R varies and so can be said to average to zero if there is the least

imprecision in the measurement of $\boldsymbol{R}$; in the intensity $|\psi|^2$, the terms from eq. (4.5) with different μ average to zero in this way, leaving the sum of squares of individual ray amplitudes, i.e.

$$|\psi(\boldsymbol{R}, Z)|^2 = \sum_{\mu} \left| \det\left\{ \frac{\partial \boldsymbol{r}^{\mu}}{\partial \boldsymbol{R}}(\boldsymbol{R}, Z) \right\} \right|, \quad \text{when } k = \infty, \tag{4.6}$$

and this of course does not involve k.

But in spite of these merits the interfering-ray-sum (4.5) suffers from a fatal defect. To discover what this is, let us examine the inverse of the amplitude determinant in eq. (4.5). From (4.4), we find that this inverse is proportional to the determinant of second derivatives of the generating function (4.3) with respect to the state variables, i.e. (omitting ray labels μ)

$$\begin{aligned} \det\left\{ \frac{\partial \boldsymbol{R}}{\partial \boldsymbol{r}} \right\} &= 1 - Z(f_{xx} + f_{yy}) + Z^2(f_{xx} f_{yy} - f_{xy}^2) \\ &= Z^2 \det\left\{ \frac{\partial^2 \phi}{\partial s_i \partial s_j} \right\}. \end{aligned} \tag{4.7}$$

Now we recall the fundamental result (3.20), that on a caustic the determinant of second derivatives vanishes. Therefore $\det\{\partial \boldsymbol{R}/\partial \boldsymbol{r}\}$ vanishes too, and the amplitude in eq. (4.5) is infinite. This cannot be correct, because, in the monochromatic and quasimonochromatic waves with which we are dealing, ψ is a smooth function of its arguments C as long as k is not infinite. We conclude that *eq. (4.5) fails precisely where we would need it most*, i.e. on caustics, where the waves are most intense: it cannot tell us how $|\psi|^2$ increases as k increases, or what the wave pattern close to the caustic looks like.

Before I take up these problems, which will be our main concern, let me squeeze one more useful piece of information from eq. (4.5). The determinant was deliberately written without a modulus sign, thereby leaving its phase implicit. From eq. (4.7) it follows that the inverse determinant has a simple zero on a typical (fold) caustic, at which its phase changes by π. If this phase is an increase as the ray touches the caustic, then eq. (4.5) can be written explicitly as

$$\begin{aligned} \psi(\boldsymbol{R}, Z) \simeq \frac{1}{Z} \sum_{\mu} &\left| \det\left\{ \frac{\partial^2 \phi}{\partial s_i \partial s_j} \right\}_{\mu} \right|^{-1/2} \\ &\times \exp\{ik\phi_{\mu}(\boldsymbol{R}, Z) - i\pi m_{\mu}(\boldsymbol{R}, Z)/2\}, \end{aligned} \tag{4.8}$$

where m_μ is the number of caustics touched by the μth ray en route from W to $(\boldsymbol{R}, Z)$. This exploitation of analyticity is, of course, far from a rigorous argument, but in fact it does give the correct assignment of phase to the ray contributions, even in the general case where m_μ can be any positive integer, in contrast to our particular example in which (cf. subsection 3.4) m_μ cannot exceed two.

Now we must return to the main problem of approximating ψ near caustics. The way to do this is once again to use the principle of superposition, but now, instead of adding a finite number of ray contributions we employ a continuous superposition, i.e. an *integral representation*, for ψ. Reverting for a moment to general notation, we shall write ψ in the form

$$\psi(C) = \left(\frac{k}{2\pi \mathrm{i}}\right)^{n/2} \int \ldots \int \mathrm{d}^n s a(s; C) \mathrm{e}^{\mathrm{i}k\phi(s;C)}, \tag{4.9}$$

where ϕ is the generating function for the ray family, n is the number of state variables and a is a non-singular amplitude factor to be made explicit in our example. When k is large, the integrand oscillates rapidly as a function of the integration variables s, so that contributions from neighbouring oscillations will almost always cancel. The exceptional s-values are those for which the phase is stationary, i.e. for which the derivatives $\partial\phi/\partial s_i$ vanish; but this is precisely the "gradient map" condition (3.17) defining the rays $s^\mu(C)$.

It is natural in these circumstances to seek to approximate (4.9) by the *n-dimensional method of stationary phase* [30]. This is valid whenever the different stationary points (rays) $s^\mu(C)$ are not too close together, i.e. precisely when C does not lie close to a caustic, and proceeds by expanding ϕ to second order in $s - s^\mu(C)$. The result is

$$\psi(C) \approx \sum_\mu \frac{a(s^\mu(C); C)}{\left|\det\left\{\partial^2\phi/\partial s_i \partial s_j\right\}_\mu\right|^{1/2}} \exp\left\{\mathrm{i}k\phi_\mu(C) - \mathrm{i}\pi m_\mu(C)/2\right\}, \tag{4.10}$$

where $m_\mu(C)$ is the number of negative eigenvalues of the matrix of second derivatives of ϕ, evaluated at $s^\mu(C)$. In view of the fact that initially, i.e. near W, all eignevalues are positive (because ϕ is a minimum for the unique ray from W to a nearby point), and that at each caustic one eigenvalue changes sign, it is clear that result (4.10) is exactly the same as (4.8) for the particular case to which that equation refers.

This approach via an integral representation reveals the reason for the failure of eqs. (4.8) and (4.10): as C moves onto a caustic, two or more stationary points coalesce, so that the quadratic expansion of ϕ about each of them cannot be carried out, and the method of stationary phase cannot be applied; but the integral (4.9) remains a valid approximation, giving a non-singular description of ψ close to caustics, as will be explained in subsequent sections.

The whole continuous-superposition procedure will, however, remain a fantasy unless it can be shown that ψ really can be written in the form (4.9). Fortunately this is not only possible, but possible in many ways. I shall illustrate this by our now-standard example (fig. 61), and obtain two different integral representations for $\psi(\boldsymbol{R}, Z)$.

Firstly, we write ψ as a diffraction integral of *Kirchhoff type* [30], by superposing spherical waves from virtual sources at each point $\boldsymbol{r}$ on W. The contribution at $\boldsymbol{R}, Z$ from the source at $\boldsymbol{r}$ is given by the free-space Green function, i.e. in the paraxial approximation by

$$\frac{\mathrm{e}^{\mathrm{i}k\phi(\boldsymbol{r};\boldsymbol{R},Z)}}{\phi(\boldsymbol{r};\boldsymbol{R},Z)} \approx \frac{1}{Z}\exp\left\{\mathrm{i}k\left(Z - f(\boldsymbol{r}) + (\boldsymbol{R}-\boldsymbol{r})^2/2Z\right)\right\}. \qquad (4.11)$$

For large k, where the contributions come from the rays, i.e. from the normals to W, it is not necessary to consider the refinement of obliquity factors [30] and the correct superposition is

$$\psi(\boldsymbol{R}, Z) = \frac{k}{2\pi\mathrm{i}Z}\int\int \mathrm{d}\boldsymbol{r}\,\mathrm{e}^{\mathrm{i}k\phi(\boldsymbol{r};\boldsymbol{R},Z)}, \qquad (4.12)$$

where ϕ is given by (4.3). This representation is precisely of the form (4.9), and has the following desirable properties: (i) When W is flat, i.e. $f(\boldsymbol{r}) = 0$, it follows by elementary complex Gaussian integrations that

$$\psi(\boldsymbol{R}, Z) = \mathrm{e}^{\mathrm{i}kZ} \quad (f = 0). \qquad (4.13)$$

(ii) When $Z = 0$, it is easy to show (again by Gaussian integration) that

$$\psi(\boldsymbol{R}, 0) = \mathrm{e}^{-\mathrm{i}kf(\boldsymbol{R})}, \qquad (4.13a)$$

whose phase correctly corresponds to the presence of a gently-varying wavefront with height function f. (iii) When k is large and $\boldsymbol{R}, Z$ does not lie near a caustic, the stationary phase approximation to (4.12) gives precisely the ray sum (4.8).

Secondly, we follow Maslov [48, 49] and obtain a representation of *Fourier type*, which is both more subtle and more instructive than the Kirchhoff integral (4.12). Maslov's method is motivated by the difficulty

of finding globally valid representations such as (4.12) in problems less simple than our illustrative example. It gives a representation whose validity, whilst only local, is guaranteed near caustics. The method employs concepts from Hamiltonian mechanics. Trajectories are defined in terms of the two-dimensional *momentum* $\boldsymbol{P}$ conjugate to $\boldsymbol{R}$, describing the directions of rays through $(\boldsymbol{R}, Z)$. For the μth ray, the momentum is defined by

$$\boldsymbol{P} = \boldsymbol{P}^{\mu}(\boldsymbol{R}, Z) \equiv \nabla_{\boldsymbol{R}}\phi_{\mu}(\boldsymbol{R}, Z), \tag{4.14}$$

and the explicit formulae (4.3) and (4.4) give

$$\boldsymbol{P}^{\mu} = (\boldsymbol{R} - \boldsymbol{r}^{\mu}(\boldsymbol{R}, Z))/Z = -\nabla_{\boldsymbol{r}} f(\boldsymbol{r}^{\mu}(\boldsymbol{R}, Z)). \tag{4.15}$$

The momentum function encodes the caustic structure of the ray family in a very beautiful way. Consider the four-dimensional *phase space* $(\boldsymbol{R}, \boldsymbol{P})$. The equations (4.14), i.e. $\boldsymbol{P} = \boldsymbol{P}^{\mu}(\boldsymbol{R}, Z)$, define for fixed Z a two-dimensional phase space surface ("Lagrangian manifold" [48]). This is a smooth surface, given parametrically in terms of $\boldsymbol{r}$ by

$$\boldsymbol{R} = \boldsymbol{r} + Z\nabla f(\boldsymbol{r}); \qquad \boldsymbol{P} = -\nabla f(\boldsymbol{r}). \tag{4.16}$$

It can fold so that "over" each point $\boldsymbol{R}$ there may be several sheets, corresponding to the rays and labelled by μ. As $\boldsymbol{R}$ moves onto a caustic, two rays become parallel and two sheets of the surface join smoothly. When $Z = 0$, the surface is not folded and indeed is simply the graph of the gradient of f, i.e. from (4.15)

$$\boldsymbol{P} = \nabla f(\boldsymbol{R}) \quad \text{when} \quad Z = 0. \tag{4.17}$$

As Z increases, the surface evolves by shearing in the $\boldsymbol{R}$ plane, as is shown by writing (4.15), together with (4.4), in the form

$$\boldsymbol{P} = -\nabla f(\boldsymbol{R} - Z\boldsymbol{P}). \tag{4.18}$$

This shear inevitably causes the surface to fold, as illustrated in the one-dimensional case by fig. 62. For each Z the caustic in $\boldsymbol{R}$ is the locus of singularities of the projection of the phase space "down" onto $\boldsymbol{R}$, and is given from (4.16) by the locus of divergence of

$$\det\{\partial \boldsymbol{P}/\partial \boldsymbol{R}\} = \det\{\partial \boldsymbol{P}/\partial \boldsymbol{r}\}/\det\{\partial \boldsymbol{R}/\partial \boldsymbol{r}\}. \tag{4.19}$$

This can happen only when the denominator vanishes, which indeed gives the familiar caustic equation [see eq. (4.7) ff.].

Maslov's method is based on two observations. The first is that, just as the trajectory theory is symmetrical in $\boldsymbol{R}$ and $\boldsymbol{P}$, so is the wave theory, so that $\psi(\boldsymbol{R}, Z)$ can be written in terms of a two-dimensional *momentum*

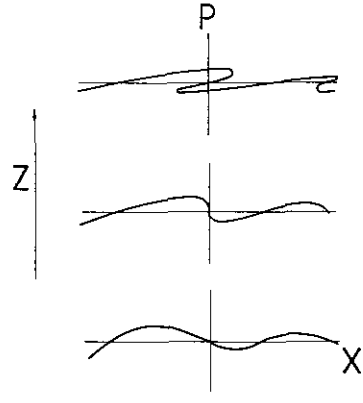


Fig. 62.

representation $\bar{\psi}(\boldsymbol{P}, Z)$, i.e.

$$\psi(\boldsymbol{R}, Z) = \frac{k}{2\pi}\int\int \mathrm{d}\boldsymbol{P}\bar{\psi}(\boldsymbol{P}, Z)\mathrm{e}^{\mathrm{i}k\boldsymbol{P}\cdot\boldsymbol{R}}, \tag{4.20}$$

where

$$\bar{\psi}(\boldsymbol{P}, Z) = \frac{k}{2\pi}\int\int \mathrm{d}\boldsymbol{R}\psi(\boldsymbol{R}, Z)\mathrm{e}^{-\mathrm{i}k\boldsymbol{P}\cdot\boldsymbol{R}}. \tag{4.21}$$

Because of this symmetry, there is a shortwave approximation for $\bar{\psi}$, analogous to eq. (4.5), in the form of a sum over trajectories ν in momentum space, namely

$$\bar{\psi}(\boldsymbol{P}, Z) \approx \sum_{\nu}\left[\det\left\{\frac{\partial \boldsymbol{r}^{\nu}}{\partial \boldsymbol{P}}(\boldsymbol{P})\right\}\right]^{1/2}\exp\left\{\mathrm{i}k\bar{\phi}_{\nu}(\boldsymbol{P}, Z)\right\}, \tag{4.22}$$

where $\boldsymbol{r}^{\nu}(\boldsymbol{P})$ is a solution of the second of equations (4.16) (and is independent of Z in this case because the rays are straight). To find the phase $\bar{\phi}$ on a given branch ν, we proceed by analogy with eq. (4.14), according to which ϕ is given by the area under the phase space surface (fig. 63); in an abbreviated notation,

$$\phi(\boldsymbol{R}) = \int_{\boldsymbol{R}^*}^{\boldsymbol{R}}\boldsymbol{P}(\boldsymbol{R}')\cdot\mathrm{d}\boldsymbol{R}', \tag{4.23}$$

where R^* is arbitrary and leaves ϕ undetermined up to a constant. Thus

$$\bar{\phi}(\boldsymbol{P}) = -\int_{\boldsymbol{P}^*}^{\boldsymbol{P}}\boldsymbol{R}(\boldsymbol{P}')\cdot\mathrm{d}\boldsymbol{P}'. \tag{4.24}$$

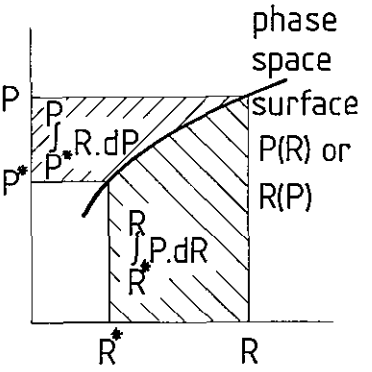


Fig. 63.

Now comes Maslov's second observation, on which the method is crucially dependent. The trajectory sum (4.22) for $\bar{\psi}$ suffers from the same defect as (4.5), namely it diverges on caustics. But now the offending caustics are those in momentum space, where the phase space surface is singular when projected "across" onto $\boldsymbol{P}$ space. *These "**P**-caustics" must be distinct from the "**R**-caustics"*, because a smooth phase–space surface cannot simultaneously be parallel to the $\boldsymbol{R}$-plane and the $\boldsymbol{P}$-plane (fig. 64). When $\boldsymbol{P}(\boldsymbol{R})$ is multiple-valued because of the

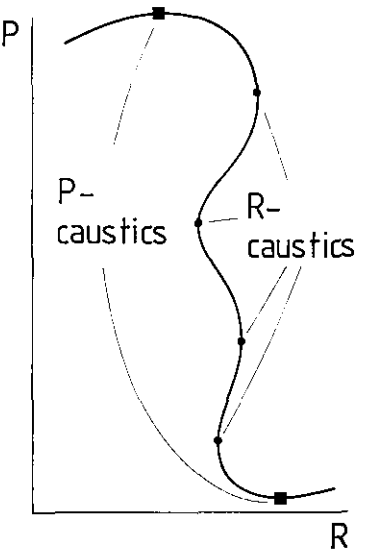


Fig. 64.

presence of $\boldsymbol{R}$-caustics, $\boldsymbol{R}(\boldsymbol{P})$ will be single-valued, and vice versa. There may be several $\boldsymbol{R}$-caustics within a branch ν of the phase–space surface, bounded by two $\boldsymbol{P}$-caustics. We conclude that if the approximation (4.22) is substituted into (4.20), it is possible to get a representation of ψ, namely

$$\psi(\boldsymbol{R}, Z) \approx \frac{k}{2\pi} \iint \mathrm{d}\boldsymbol{P} \sum_{\nu} \left[\det \left\{ \frac{\partial \boldsymbol{r}^{\nu}}{\partial \boldsymbol{P}}(\boldsymbol{P}) \right\} \right]^{1/2} \times \exp\left\{ \mathrm{i}k \left[\bar{\phi}_{\nu}(\boldsymbol{P}, Z) + \boldsymbol{P}\cdot\boldsymbol{R} \right] \right\}, \tag{4.25}$$

which does not diverge on caustics.

We must check that the Maslov representation indeed has the form (4.9), i.e. that

$$\phi_{\nu}(\boldsymbol{P}; \boldsymbol{R}, Z) \equiv -\int_{\boldsymbol{P}^*}^{\boldsymbol{P}} \boldsymbol{R}^{\nu}(\boldsymbol{P}', Z)\cdot \mathrm{d}\boldsymbol{P}' + \boldsymbol{P}\cdot\boldsymbol{R} \tag{4.26}$$

acts like a local generating function with state variable $\boldsymbol{P}$ and control parameters $\boldsymbol{R}$, Z. To see that it does, we first note that ϕ_{ν} is a smooth function, because $\boldsymbol{R}^{\nu}$ is a smooth function of $\boldsymbol{P}$ near $\boldsymbol{R}$-caustics. Next, we form the gradient map

$$\nabla_{\boldsymbol{P}} \phi_{\nu} = -\boldsymbol{R}^{\nu}(\boldsymbol{P}, Z) + \boldsymbol{R} = 0 \tag{4.27}$$

whose solutions are the momenta $\boldsymbol{P}^{\mu}(\boldsymbol{R}, Z)$ [eq. (4.14)] defining the rays. Therefore ϕ_{ν} is a generating function, and (4.25) does have the form (4.9).

Maslov's representation (4.25) is quite different from the "exact" Kirchhoff integral (4.12). To see this explicity, let us use (4.12) to evaluate the momentum wave function (4.21). With the aid of (4.3), $\bar{\psi}$ becomes, exactly

$$\bar{\psi}(\boldsymbol{P}, Z) = \frac{k}{2\pi} \exp\left\{ \mathrm{i}kZ(1 - P^2/2) \right\} \times \iint \mathrm{d}\boldsymbol{r} \exp\left\{ -\mathrm{i}k\left(f(\boldsymbol{r}) + \boldsymbol{P}\cdot\boldsymbol{r} \right) \right\}. \tag{4.28}$$

The stationary phase condition gives precisely (4.16), i.e. $\boldsymbol{r} = \boldsymbol{r}^{\nu}(\boldsymbol{P})$ and the stationary phase method gives, for the final result (4.25),

$$\psi(\boldsymbol{R}, Z) \approx \frac{k}{2\pi} \mathrm{e}^{\mathrm{i}kZ} \iint \mathrm{d}\boldsymbol{P} \mathrm{e}^{-\mathrm{i}kZP^2/2} \times \sum_{\nu} \frac{\exp\left\{ \mathrm{i}k\left[(\boldsymbol{R} - \boldsymbol{r}^{\nu}(\boldsymbol{P}))\cdot\boldsymbol{P} - f(\boldsymbol{r}^{\nu}(\boldsymbol{P})) \right] - \mathrm{i}\alpha_{\nu} \right\}}{\left| \left(f_{xx} f_{yy} - f_{xy}^2 \right)_{\boldsymbol{r}=\boldsymbol{r}^{\nu}(\boldsymbol{P})} \right|^{1/2}}, \tag{4.29}$$

where α_{ν} are phases depending on the curvatures of W at $\boldsymbol{r}^{\nu}(\boldsymbol{P})$.

This is a strange formula. Because it fails on ***P***-caustics it cannot even give ψ correctly at $Z = 0$, whereas the Kirchhoff result (4.12) and even the trajectory sum (4.8) gives the exact wave (4.14). And it fails when $Z \to \infty$ because then (and only then) the denominator $|f_{xx}f_{yy} - f_{xy}^2|$ vanishes at the ***R***-caustics [cf. (4.7)] (I let the reader work out why). Although Maslov's formula correctly describes ψ near all ***R***-caustics for $0 < Z < \infty$, it seems a poor competitor to the Kirchhoff integral. But if the space $Z > 0$ contained a medium with continuously-varying refractive index, the Kirchhoff method could not be applied at large distances, and then Maslov's technique might be the most convenient way of representing ψ in the form (4.9).

4.2. Classification of diffraction patterns near stable caustics

We have found that it is always possible to obtain non-singular asymptotic integral representations (4.9) for waves $\psi(C)$ near caustics, in the form of superpositions of contributions labelled by state variables s. There still remains the nontrivial problem of evaluating these integrals when k is large. If C is not near a caustic, the stationary points $s^{\alpha}(C)$ (rays) are well separated and the integral can be evaluated by the method of stationary phase to give the interfering sum (4.10). As C moves onto a caustic, however, stationary points coalesce and the method becomes inapplicable.

In these circumstances it might appear that there is no alternative to evaluating the integral (4.9) afresh for every case, which in view of the rapid oscillations of the integrand as $k \to \infty$ would be an onerous task. But catastrophe theory comes to the rescue with the observation that, provided only stable caustics are considered, every generating function $\phi(s; C)$ belongs to one of the equivalence classes described in subsection 3.2. All ϕ within a class share the same caustic geometry, and hence the same topology of coalescence of stationary points as C varies. As explained in subsection 3.2, any ϕ in an equivalence class can be transformed into the corresponding normal form $\Phi(s; C)$ in table 1 by a diffeomorphism of the s and C spaces, as long as the codimension $K \leqslant 4$.

This suggests changing the integration variable s in (4.9), and transforming the controls C parameterising the space of observation, so as to deform ϕ into Φ. The very fact that this transformation is a diffeomorphism guarantees that the Jacobian relating the two sets of state variables is non-singular; therefore this Jacobian has the effect of smoothly modifying the amplitude factor $a(s; C)$. Near a caustic, however, all the

contributing points $s^n(C)$ lie close together, so that a may be approximated by a constant and taken outside the integral. As a result of this procedure, the infinity of possible ψ is deformed into *a finite set of standard "diffraction catastrophes"* which I will denote $\Psi(C; k)$. If the standard polynomial is $\Phi(s; C)$, and the corank is n, the corresponding diffraction catastrophe is

$$\Psi(C;k) = \left(\frac{k}{2\pi}\right)^{n/2} \int_{-\infty}^{\infty} \cdots \int \mathrm{d}^n s \mathrm{e}^{\mathrm{i}k\Phi(s;C)}. \tag{4.30}$$

Diffraction near caustics in any practical case will be described by a wave ψ which is a diffeomorphism of one of the diffraction catastrophes Ψ. I do not have space to describe exactly how the mapping of ψ onto Ψ is carried out, but I have done so elsewhere [32], and so, with more rigour, have Duistermaat [47] and Guillemin and Sternberg [50].

In order to understand the catastrophe theory of waves on the microscale, then, it is necessary to understand the finite set of integrals (4.30). With the exception of the simplest integral, corresponding to the fold catastrophe, the $\Psi(C; k)$ are not familiar as "special functions" in applied mathematics, and they have not been tabulated. Their understanding is therefore a challenging problem for the near future. At present our information about the diffraction catastrophes is of three sorts. Firstly, the k-dependence can be removed from (4.30) by rescaling the wave amplitude and the control space, to reveal a fascinating set of scaling exponents that will be discussed in subsection 4.5. Until then, I shall replace k by unity and write $\Psi(C; 1)$ simply as $\Psi(C)$. Secondly, the geometric structure of the intensity $|\Psi(C)|^2$ in the control space can be understood by a combination of asymptotic analysis and computation, described in subsection 4.3 for three catastrophes. This will take us down to the "ultramiscroscopic" scale where we shall renew our acquaintance with the wavefront dislocations of section 2. And thirdly, a surprising series of non-linear functional identities exists, relating different diffraction catastrophes as well as different aspects of the same diffraction catastrophe; these will be described in subsection 4.4.

The differences and similarities between trajectories and waves are brought out very clearly by comparing (4.30) with the gradient condition (3.17). Both begin with a generating function $\Phi(s; C)$ and then eliminate s, but in the wave theory the elimination is by integration over s, i.e., superposition of contributions from all s, whereas the trajectory theory selects those s for which Φ is stationary for a given C.

4.3. Architecture of the simplest diffraction catastrophes

The *fold diffraction catastrophe* is given by (4.30) and table 1 as

$$\psi_{\text{fold}}(C) = \frac{1}{\sqrt{(2\pi)}} \int_{-\infty}^{\infty} \mathrm{d}s \exp\{\mathrm{i}(s^3/3 + Cs)\}. \tag{4.31}$$

This is $\sqrt{(2\pi)}$ times the Airy function of analysis [8] – a real function of one variable – and indeed was invented by Airy [51] in 1838 precisely to describe diffraction near a caustic. Of course Airy could not know that his function was the first in a hierarchy. The gradient map is

$$\partial\Phi/\partial s = s^2 + C = 0. \tag{4.32}$$

This shows that when $C < 0$ there are two interfering rays $s^{\mu} = \pm\sqrt{(-C)}$ which coalesce on the caustic at $C = 0$. For $C > 0$ there are no real rays, but there are two "complex rays", with $s^{\mu} = \pm\mathrm{i}\sqrt{C}$. Ignoring the contribution of the complex rays, which decays rapidly on the "dark" side $C > 0$ of the caustic, and evaluating the contributions for the two rays with $C < 0$ by the method of stationary phase, we obtain

$$\Psi_{\text{fold}}(C) \approx \frac{\sqrt{2}}{(-C)^{1/4}} \sin\left\{\frac{2}{3}(-C)^{3/2} + \frac{\pi}{4}\right\} \quad (C < 0). \tag{4.33}$$

For the special case with which we are dealing, this is simply the interfering ray sum (4.10). Figure 65 shows a comparison of the intensity $|\Psi_{\text{fold}}(C)|^2$ (full line), obtained by computing (4.31), with the intensity in the approximation (4.33) (dashed line), showing that the ray sum gives an

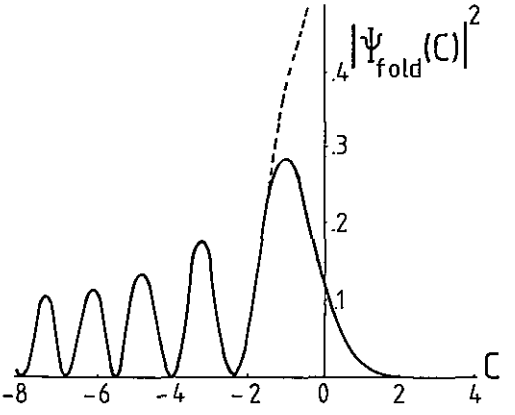


Fig. 65.

accurate description of the interference oscillations but fails very close to the caustic at $C = 0$. Figure 65 also shows a photograph of diffraction fringes near a fold caustic.

The *cusp diffraction catastrophe* is given by (4.30) and table 1 as

$$\Psi_{\text{cusp}}(C_1, C_2) = \frac{1}{\sqrt{(2\pi)}} \int_{-\infty}^{\infty} \mathrm{d}s \exp\left\{\mathrm{i}\left(s^4/4 + C_2 s^2/2 + C_1 s\right)\right\}. \tag{4.34}$$

This is a complex function of two variables, first studied by Pearcey [52] in 1946 in connection with caustics (but again in isolation and not as part of a hierarchy). The cusped caustic has the form shown in fig. 32 and described by equation (3.23). Within the cusp there are three interfering rays; outside and sufficiently close to the caustic [67], there is one real ray and one complex ray. The resulting interference pattern is surprisingly complicated, as is shown by fig. 66. Figure 66a is a photograph of the pattern, obtained with one of the water-droplet "lenses" discussed in subsection 3.3; fig. 66b is a magnification of the region close to the cusp point. Finally, fig. 66c is a computer simulation, obtained by drawing a contour map of $|\psi_{\text{cusp}}(C_1, C_2)|$. It is obvious that the agreement between experiment and theory is astonishingly good, even down to the smallest details.

To see what these "smallest details" are, look closely at fig. 66c, and note the tiny elongated black spots. These are none other than the wavefront dislocations we studied in section 2, reached at last after a journey that has taken us from poorly-resolved "macroscopic" caustic networks, through the "atomic" catastrophes and down through the "microscopic" "subatomic" diffraction catastrophes to the "elementary particles" of waves. Within the cusp, they occur in pairs in a triangular array resulting

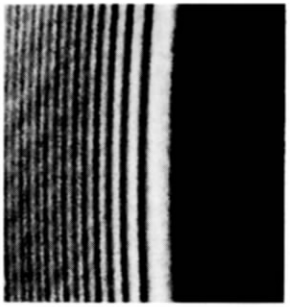
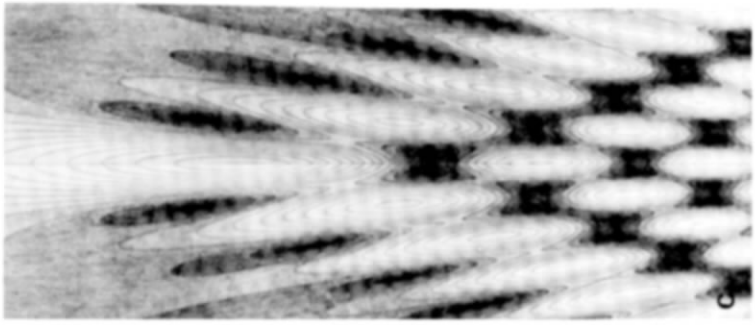
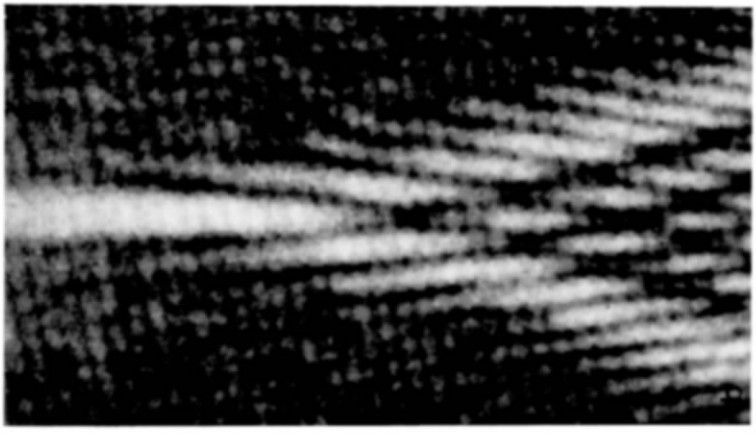
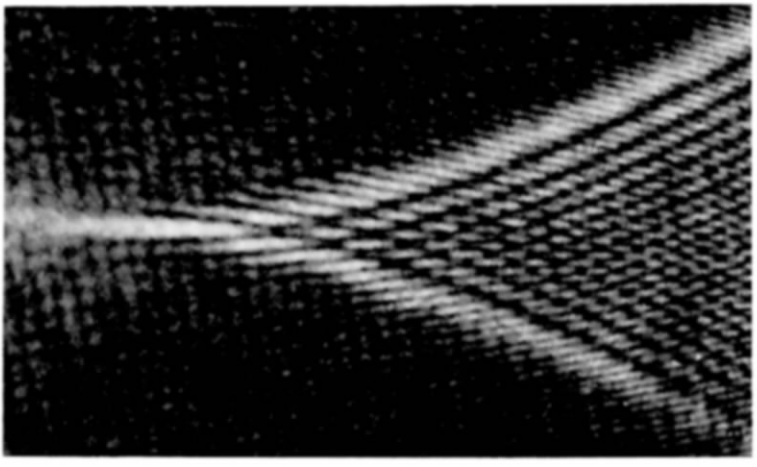


Fig. 66.







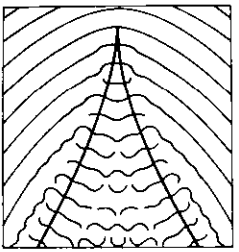


Fig. 67. (From ref. [7].)

from three-wave interference, and indeed their positions are given with high accuracy [53] by the zeroes of the stationary-phase approximation to (4.34). Outside the cusp, they occur as a single row flanking each branch of the caustic, resulting from interference between the single real ray and a complex ray that remains after the coalescence of two of the three inner real rays on each fold line. The dislocation point shows up very clearly as the ends of wavecrests passing through a cusp (fig. 67). It is remarkable that the diffraction catastrophe wave function is able to describe the most delicate wave singularities in spite of being constructed from a generating function whose original purpose was to describe the caustics, i.e. the ray singularities. In this connection it is worth noting that on the experimental photograph (fig. 66b) the distance between neighbouring dislocations within the cusp is only a few wavelengths of light.

The *elliptic umbilic diffraction catastrophe* is given by (4.30) and table 1 as

$$\Psi_{\text{E.U.}}(C_1, C_2, C_3) = \frac{1}{2\pi}\iint_{-\infty}^{\infty} \mathrm{d}s_1\,\mathrm{d}s_2$$

$$\times \exp\left\{\mathrm{i}\left(s_1^3 - 3s_1s_2^2 - C_3\left(s_1^2 + s_2^2\right) - C_2s_2 - C_1s_1\right)\right\}. \quad (4.35)$$

This is a complex function of three variables, realizable as a diffraction pattern in space. In view of the much greater richness of diffraction near a cusp than diffraction near a fold, we can expect the architecture of the elliptic umbilic diffraction catastrophe to be extremely complicated, and so it turns out to be. $\Psi_{\text{E.U.}}$ has been studied in detail by Berry et al. [53], both experimentally (by focusing a microscope on various planes in the space above a water droplet "lens" through which a parallel beam of light has been refracted) and theoretically [by computation and stationary phase analysis of (4.35)].

The caustic is shown in fig. 32; it consists of three sheets joined along three cusped edges which touch at the elliptic umbilic point $C_1 = C_2 = C_3 = 0$. Within the space bounded by the caustic surface there are four rays; outside the caustic there are two. Some idea of the complexity of the resulting interference pattern can be obtained from fig. 68 which shows experiment and computer simulation of an "unfolded" section with coordinates C_1, C_2 with C_3 held fixed; once again there is excellent agreement between the two pictures. The hexagons within the cusped triangle are the result of slicing through a distorted lattice of intensity maxima resembling hexagonal prisms, formed by four-wave interference. The lines outside the triangle are two-wave interference fringes.

As expected from section 2, the wavefront dislocations are lines in C_1, C_2, C_3 space. Within the caustic surface and not too near the cusped edges these take the form of puckered rings arrayed in space; they can be seen as dark rings within the hexagons in fig. 68. Two rows of these are shown in projection in fig. 69a, which is a "vertical" section in the C_1, C_3 plane with $C_2 = 0$, and these rows are shown in plan in fig. 69b. Moving along the row labelled $M = 1$, the rings tilt together and eventually merge into a "hairpin" structure clearly seen in fig. 69b. How many rings are there in the Mth such row before it becomes a hairpin? By an elaborate stationary-phase analysis of (4.35) [53], this number, $N_{\max}(M)$, is found to be

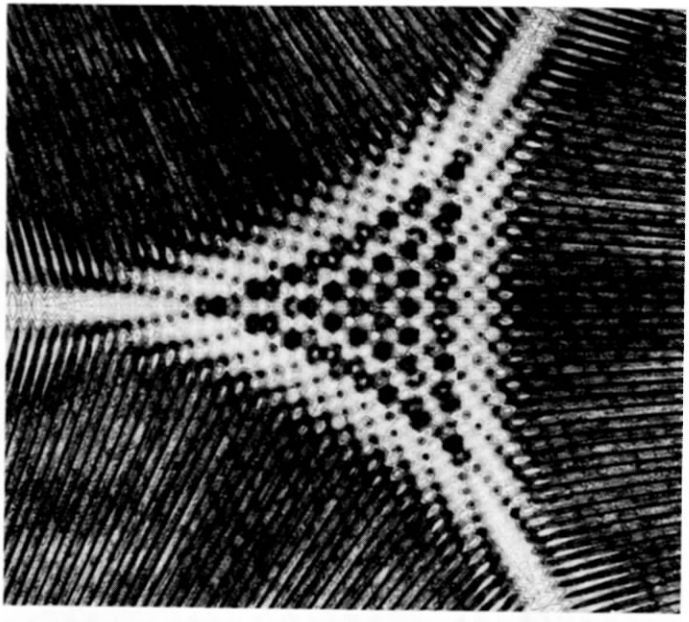
$$N_{\max}(M) = \text{integer part of } \left(\frac{512M}{27} - \frac{539}{104} \right). \tag{4.36}$$

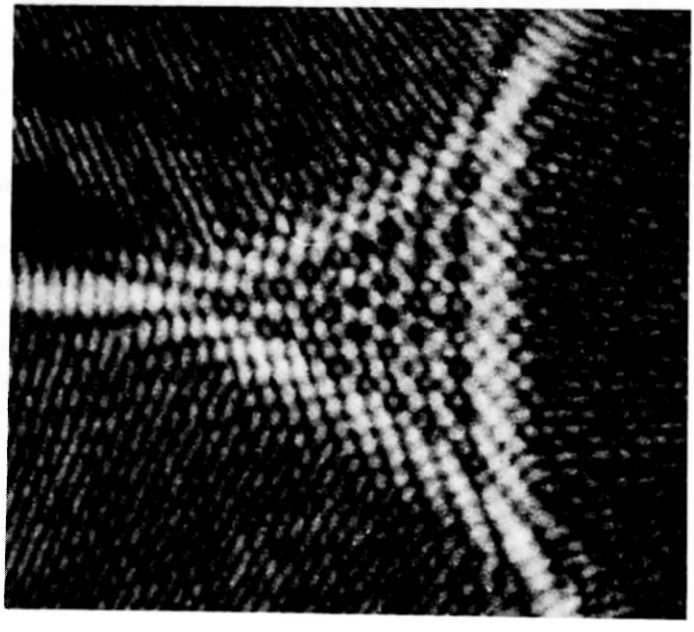
I draw attention to this astonishing result in order to show that numerical as well as geometric richness is encoded by the diffraction catastrophe integrals. Outside the caustic, the dislocations (fig. 70) consist of a set of curves in the plane $C_3 = 0$, and, above, a series of curly "antelope horns" close to the caustic.

Similar studies of the hyperbolic umbilic and swallowtail diffraction catastrophes are in preparation. The architecture of the higher diffraction catastrophes is almost certainly so complicated that it would be of dubious value to explore these in detail, but it is worthwhile to study certain important sections obtained by setting all but two of the control parameters equal to zero.

4.4. Projection identities

In this section I shall give a brief account of a series of strange non-linear identities, derived in collaboration with Wright [54], in which the intensi-





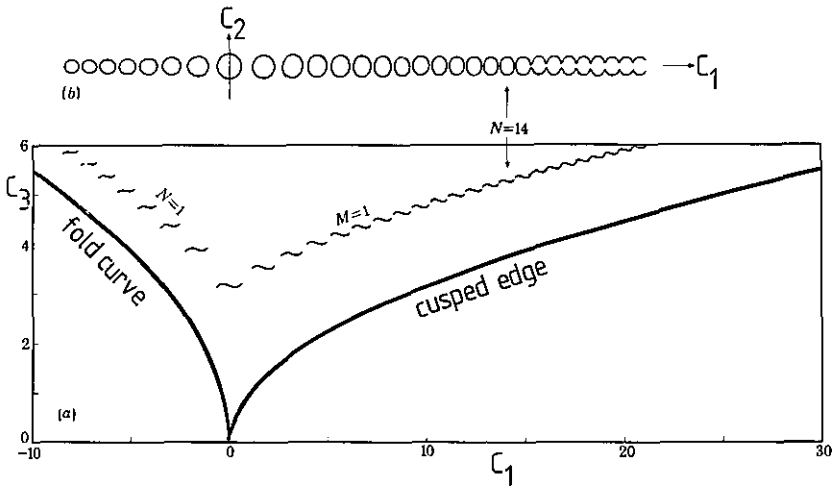


Fig. 69. (From ref. [53].)

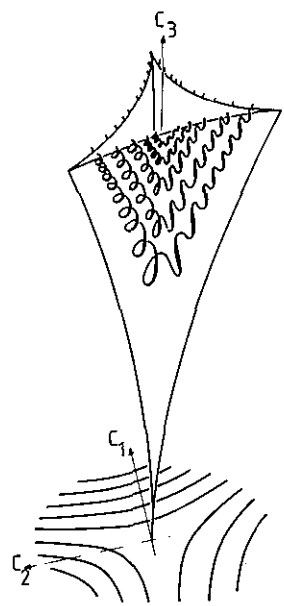


Fig. 70. (From ref. [53].)

ties $|\Phi(C)|^2$ of the diffraction catastrophes (4.30) are expressed as integrals over the wave functions Ψ (not the intensities) for the same catastrophe or one of lower codimension.

From (4.30), the intensity is

$$|\Psi(C)|^2 = \frac{1}{(2\pi)^n}\int_{-\infty}^{\infty}\cdots\int \mathrm{d}^n s \int_{-\infty}^{\infty}\cdots\int \mathrm{d}^n s' \times \exp\{\mathrm{i}(\Phi(s;C) - \Phi(s';C)\}. \tag{4.37}$$

Defining new n-dimensional variables u and v by

$$s \equiv u + v; \qquad s' \equiv u - v, \tag{4.38}$$

and introducing the notation

$$\Theta(u, v; C) \equiv \Phi(u + v; C) - \Phi(u - v; C), \tag{4.39}$$

the intensity becomes

$$|\Psi(C)|^2 = \left(\frac{2}{\pi}\right)^{n/2}\int_{-\infty}^{\infty}\cdots\int \mathrm{d}^n u \left[\frac{1}{(2\pi)^{n/2}}\int_{-\infty}^{\infty}\cdots\int \mathrm{d}^n v \times \exp\{\mathrm{i}\Theta(u, v; C)\}\right]. \tag{4.40}$$

To derive the identities, it is necessary to evaluate the functions [ ], using the explicit forms in table 1 for Φ. This is straightforward, although tedious, but has the surprising result that the functions [ ] can themselves be expressed as diffraction catastrophes. A different series of identities can be obtained by reversing the order of u and v integrations in (4.40).

The simplest example is the fold, for which table 1 and (4.39) lead to

$$\Theta(u, v; C) = 2v^3/3 + 2v(C + 2u^2). \tag{4.41}$$

In (4.40) the function [ ] is

$$[\,] = \frac{1}{\sqrt{(2\pi)}}\int_{-\infty}^{\infty}\mathrm{d}v \exp\{\mathrm{i}(2v^3/3 + 2v(C + 2u^2)\} = 2^{-1/3}\Psi_{\mathrm{fold}}\{2^{2/3}(C + 2u^2)\}. \tag{4.42}$$

Therefore the identity (4.40) is

$$|\Psi_{\mathrm{fold}}(C)|^2 = 2^{-1/3}\sqrt{\frac{2}{\pi}}\int_{-\infty}^{\infty}\mathrm{d}u\Psi_{\mathrm{fold}}\{2^{2/3}(C + 2u^2)\}. \tag{4.43}$$

In the case of the cusp, the same argument gives

$$|\Psi_{\text{cusp}}(C_1, C_2)|^2 = \sqrt{\frac{2}{\pi}} \int_{-\infty}^{\infty} du |6u|^{-1/3} \Psi_{\text{fold}} \left\{ 2 \frac{u}{|u|} |6u|^{-1/3} \left(u^3 + C_2 u + C_1 \right) \right\}, \tag{4.44}$$

showing that the diffraction intensity all over the control plane of the cusp can be derived from a lower diffraction catastrophe, namely the fold.

For the swallowtail, a different result is found, namely

$$|\Psi_{\text{sw}}(C_1, C_2, C_3)|^2 = 2^{-1/5} \sqrt{\frac{2}{\pi}} \int_{-\infty}^{\infty} du \cdot \times \Psi_{\text{sw}} \left\{ 2^{4/5} \left(u^4 + C_3 u^2 + C_2 u + C_1 \right), 0, 2^{2/5} \left(6u^2 + C_3 \right) \right\}, \tag{4.45}$$

showing that the diffraction intensity throughout the control space of the swallowtail can be derived from the swallowtail wave function in its particular plane section with $C_2 = 0$. A similar result holds for the elliptic umbilic, namely

$$|\Psi_{\text{E.U.}}(C_1, C_2, C_3)|^2 = \frac{2^{1/3}}{\pi} \int_{-\infty}^{\infty} du_1 \int_{-\infty}^{\infty} du_2 \times \Psi_{\text{E.U.}} \left\{ 2^{2/3} \left(C_1 + 2C_3 u_1 + 3 \left(u_2^2 - u_1^2 \right) \right), \times 2^{2/3} \left(C_2 + 2C_3 u_2 + 6 u_1 u_2 \right), 0 \right\}, \tag{4.46}$$

so that here the intensity can be derived from the wave function in its plane section $C_3 = 0$.

These relations, and other similar ones forming two infinite hierarchies, have a physical meaning [54] in terms of the "Maslov" interpretation of the diffraction integrals in subsection 4.1, where the control parameters include the components of the position vector $\boldsymbol{R}$ and the state variables are the components of the momentum vector $\boldsymbol{P}$. The identities then relate $|\Psi|^2$ expressed as the square of a Fourier integral to $|\Psi|^2$ expressed as the projection of a quantum-mechanical Wigner function [54] from phase space "down" $\boldsymbol{P}$ onto $\boldsymbol{R}$. This is the reason for the term "projection identities" that heads this subsection.

4.5. Scaling laws for diffraction catastrophes

What happens to the diffraction patterns as $k \to \infty$? Obviously the intensity on the caustic must rise to infinity, and the fringe spacings must shrink to zero. To get a quantitative description of these effects, it is necessary to analyze the k-dependence of the integrals (4.30). By rescaling in s and C spaces, it turns out to be possible to express Ψ at any value of k in terms of Ψ for any other k, say k_0. Written explicitly in terms of the separate control parameters C_j ($1 \leqslant j \leqslant K$), the scaling equation is

$$\Psi\{C_j;k\} = (k/k_0)^{\beta}\Psi\{(k/k_0)^{\sigma_j}C_j;k_0\}. \tag{4.47}$$

All scaling information about the diffraction catastrophes is therefore embodied in $K+1$ positive numbers: $\beta, \sigma_1, \ldots, \sigma_K$. In this way the singularities, which we have been considering as primarily geometrical, acquire an algebraic character.

The divergence of $|\Psi|$ at the most singular part of the caustic ($C = 0$) is governed by the exponent β; this is the "singularity index" introduced by Arnol'd [33] and calculated for a large number of higher catastrophes by Varchenko [55]. In terms of β, $|\Psi|$ scales as k^{β}. The shrinking fringe spacings in the different control directions C_j are governed by the exponents σ_j; these indices were introduced by Berry [56]. In terms of σ_j, the spacings scale as $k^{-\sigma_j}$. The sum

$$\gamma \equiv \sum_{j=1}^{K} \sigma_j, \tag{4.48}$$

governs the shrinking in control space of the K-dimensional hypervolume of the main diffraction maximum; this scales as $k^{-\gamma}$. A list of the exponents β, σ_j and γ for the diffraction catastrophes with $K \leqslant 4$ is given in table 2. Their derivation, and an account of some difficulties encountered when K is large, is given by Berry [56]. Here I give a brief description of the derivation of (4.47) and work through the simplest case.

In carrying out the scaling, a crucial observation is that the generating functions Φ in table 1 consist of two parts: a "germ" consisting of a term or terms involving s but not C, and "unfolding terms" linear in C. The first stage in the scaling is to remove the k-dependence from the germ in (4.30) by a k-dependent scaling of s to a new variable s', giving a factor k^{β} outside the integral. The next stage is to remove the k-dependence from the unfolding terms, and this leads to (4.47).

Table 2
Exponents governing the scaling of wave amplitude and fringe spacings as $k \to \infty$

Catastrophe	β	σ_j	γ
Fold	$\frac{1}{6}$	$\sigma_1 = \frac{2}{3}$	$\frac{2}{3}$
Cusp	$\frac{1}{4}$	$\sigma_1 = \frac{3}{4}, \sigma_2 = \frac{1}{2}$	$\frac{5}{4}$
Swallowtail	$\frac{3}{10}$	$\sigma_1 = \frac{4}{5}, \sigma_2 = \frac{3}{5}, \sigma_3 = \frac{2}{5}$	$\frac{9}{5}$
Elliptic umbilic	$\frac{1}{3}$	$\sigma_1 = \frac{2}{3}, \sigma_2 = \frac{2}{3}, \sigma_3 = \frac{1}{3}$	$\frac{5}{3}$
Hyperbolic umbilic	$\frac{1}{3}$	$\sigma_1 = \frac{2}{3}, \sigma_2 = \frac{2}{3}, \sigma_3 = \frac{1}{3}$	$\frac{5}{3}$
Butterfly	$\frac{1}{3}$	$\sigma_1 = \frac{5}{6}, \sigma_2 = \frac{2}{3}, \sigma_3 = \frac{1}{2}, \sigma_4 = \frac{1}{3}$	$\frac{7}{3}$
Parabolic umbilic	$\frac{3}{8}$	$\sigma_1 = \frac{5}{8}, \sigma_2 = \frac{3}{4}, \sigma_3 = \frac{1}{2}, \sigma_4 = \frac{1}{4}$	$\frac{17}{8}$

In the simplest example of the fold diffraction catastrophe, we have

$$\Psi_{\text{fold}}(C; k) = \sqrt{\frac{k}{2\pi}} \int_{-\infty}^{\infty} \mathrm{d}s \exp\left\{ik(s^3/3 + Cs)\right\}. \tag{4.49}$$

Writing

$$ks^3 \equiv k_0 s'^3, \tag{4.50}$$

this becomes

$$\Psi_{\text{fold}}(C; k) = \left(\frac{k}{k_0}\right)^{1/6} \sqrt{\frac{k_0}{2\pi}} \times \int_{-\infty}^{\infty} ds' \exp\left\{ik_0\left(s'^3/3 + \left(\frac{k}{k_0}\right)^{2/3} Cs'\right)\right\}, \tag{4.51}$$

which is precisely of the form (4.47) with the exponents as in table 2. On a fold caustic, therefore, the intensity $|\Psi|^2$ diverges as $k^{1/3}$ and the fringe spacings on the bright side of the caustic shrink as $k^{-2/3}$.

The singularity index β increases with codimension K (and faster for umbilics than cuspoids). This is reasonable, since the larger K is, the greater is the number of rays coalescing at the most singular point of the caustic (actually, this number is $K + 1$). For waves in three-dimensional space, where the diffraction integrals (4.30) involve at most two variables, s, the greatest possible value of β is unity, and this is attained at the (structurally unstable) perfect point focus of a patch of spherical wave front (when the "patch" is a complete sphere, the focusing wave is simply $\sin kr/r$).

The index γ also increases with K, showing that the intense region near the singularity, where all trajectories interfere destructively, shrinks very rapidly as $k \to \infty$ for the higher catastrophes. This shrinking is anisotropic in control space, because for each catastrophe the exponents σ_j are not all the same. Thus for the cusp the spacing along the symmetry axis shrinks as $k^{-1/2}$ while that across the cusp shrinks as $k^{-3/4}$, explaining why the diffraction patterns near cusps formed by objects large in comparison with the wavelength of light are greatly elongated in the cusp direction.

Because of the structural stability of the diffraction catastrophes, each set of exponents is "universal" in the sense that it describes the scaling of intensity and fringe spacings for any caustic in its equivalence class. Further, according to (4.47), the exponents capture the non-analyticity of waves near caustics. In these respects the exponents are analogous to the "critical exponents" of phase transition theory, which describe the universal non-analytic behaviour of thermodynamic functions at critical points.

5. Conclusion

In these lectures I have ranged widely over several physically different types of wave and studied them on several very different scales. The point of view I have adopted, in which waves are to be understood in terms of the nature and structure of their singularities, often gives an intuitive grasp of observed phenomena in a way that might seem surprising in view of the fact that it is mathematically based. It is worth emphasising that the "singularity" point of view represents a departure from the reductionist philosophy often assumed to underly physics. According to that, we should begin at the deepest level for which we know the laws – Maxwell's equations, say, or even quantum electrodynamics – and derive all higher-level phenomena. But that philosophy, despite a century of intensive development, did not lead to the catastrophes and to an understanding of the way they organise stable diffraction patterns: you will not (yet) find them discussed in any textbook on optics or waves. Notwithstanding this, the "singularity" approach, however successful, must not be allowed to set into a dogma: it is complementary to, rather than incompatible with, more conventional methods.

How restricted are the concepts of wavefront dislocations, caustics and diffraction catastrophes? Within the class of scalar linear waves, they

survive even if the waves propagate in a medium that is inhomogeneous, anisotropic, changing and dispersive. Outside this class, I am not so sure. In vector and tensor waves, trajectories and hence caustics will still exist, and these must surely be decorated with diffraction catastrophes; but in addition to dislocations (which can be observed in the case of light, cf. figs. 66 and 68) there may be new types of phase singularity, related to the degree of mappings of space–time onto multicomponent complex planes. In non-linear waves, dislocations should still exist, because close to zeroes of amplitude the waves may be approximated as linear; but although caustics may still exist (e.g. as singularities in shock fronts [2]) the breakdown of superposition for such intense regions of non-linear wave fields means that the diffraction catastrophes must surely be profoundly modified. There is much scope for research here.

Finally, I point out that there are some geometrical aspects of wave motion which I have ignored completely. For example, I tacitly assumed that as the wavelength vanishes ($k \to \infty$) all diffracting objects and refracting media vary smoothly on wavelength scales, so that a trajectory picture becomes valid. But suppose waves encounter a "fractal" object, i.e. something with structure on all scales [57], such as a landscape, or a fluid at its critical point, or a turbulent atmosphere. Then as $k \to \infty$ the waves discriminate ever finer details in the object, and a trajectory picture is never valid. In this "diffractal" regime there are no rays, no caustics, and no diffraction catastrophes, and the limit $k \to \infty$ is characterized by a series of scaling laws [58, 3] involving k and the "fractal dimension" [57, 59] of the object.

I have also tacitly restricted myself to "scattering" problems where waves propagate in unbounded domains. But for waves in bounded domains, such as modes of vibration in auditoriums or microwave resonators, or bound states in quantum mechanics, new concepts arise. Trajectories, which now repeatedly explore the same territory, may be predictable or chaotic (i.e. integrable or non-integrable [60, 61]), and the morphology of wave functions and the distribution of eigenfrequencies or energy levels in the shortwave limit depends crucially on this distinction [62–64]. These are fascinating and deep questions, which I propose to explore in my lectures at next year's Les Houches Summer School.

References

[1] R. Thom, Structural Stability and Morphogenesis (Benjamin, Reading, MA, 1975).
[2] T. Poston and I. N. Stewart, Catastrophe Theory and its Applications (Pitman, London, 1978).

[3] M. V. Berry, Proc. Symp. App. Maths. 36 (1980) 13.
[4] M. V. Berry and C. Upstill, Catastrophe Optics: Morphologies of Caustics and Their Diffraction Patterns, in: Progress in Optics, ed. E. Wolf, vol. 18 (North-Holland, Amsterdam, 1980) p. 257.
[5] T. Poston, Catastrophe Theory in Physics, to be published in Physics Reports.
[6] J. F. Nye and M. V. Berry, Proc. R. Soc. Lond. A336 (1974) 165.
[7] F. J. Wright, Wavefront dislocations and their analysis using catastrophe theory, in: Structural stability in physics, eds. W. Güttinger and H. Eikemeier (Springer, Berlin, 1979) 141.
[8] M. Abramowitz and I. A. Stegun, eds., Handbook of mathematical functions (Dover, New York, 1964).
[9] A. T. Winfree, The geometry of biological time (Springer, New York, 1980).
[10] G. de Q. Robin, S. Evans and J. T. Bailey, Phil. Trans. R. Soc. Lond. A265 (1969) 437.
[11] M. E. R. Walford, P. C. Holdorf and R. G. Oakberg, J. Glaciol. 18 (1977) 217.
[12] M. E. R. Walford, Nature 239 (1972) 95.
[13] F. J. Wright, Wavefield Singularities, Ph.D. Thesis (H. H. Wills Physics Laboratory, Bristol Univ., U.K., 1977).
[14] F. J. Wright and J. F. Nye, submitted to Phil. Trans. R. Soc. Lond. (1981).
[15] W. Whewell, Phil. Trans. R. Soc. Lond. 123 (1833) 147; (1836) 289.
[16] A. Defant, Physical Oceanography (Pergamon, London, 1961) vol. 2.
[17] W. Braunbek and G. Laukien, Optik 9 (1952) 174.
[18] P. A. M. Dirac, Proc. R. Soc. Lond. A133 (1931) 60.
[19] J. Riess, Ann. Phys. N. Y. 57 (1970) 301.
[20] J. Riess, Phys. Rev. D2 (1970) 647.
[21] J. Riess, Phys. Rev. B13 (1976) 3862.
[22] Y. Aharonov and D. Bohm, Phys. Rev. 115 (1959) 485.
[23] R. G. Chambers, Phys. Rev. Lett. 5 (1960) 3.
[24] M. V. Berry, R. G. Chambers, M. D. Large, C. Upstill and J. C. Walmsley, Europ. J. Phys. 1 (1980) 154; see also M. V. Berry, Europ. J. Phys. 1 (1980) 240.
[25] T. T. Wu and C. N. Yang, Phys. Rev. D12 (1975) 3845.
[26] J. O. Hirschfelder, A. C. Christoph and W. E. Palke, J. Chem. Phys. 61 (1974) 5435.
[27] J. O. Hirschfelder, C. J. Goebel and L. W. Bruch, J. Chem. Phys. 61 (1974) 5456.
[28] J. O. Hirschfelder and K. T. Tang, J. Chem. Phys. 64 (1976) 760.
[29] J. O. Hirschfelder and K. J. Tang, J. Chem. Phys. 65 (1976) 470.
[30] M. Born and E. Wolf, Principles of Optics, 5th ed. (Pergamon, Oxford, 1975).
[31] J. L. Elliot, R. G. French, E. Dunham, P. J. Gierasch, J. Veverka, C. Church and Carl Sagan, Astrophys. J. 217 (1977) 661.
[32] M. V. Berry, Adv. Phys. 25 (1976) 1.
[33] V. I. Arnol'd, Russ. Math. Surys. 30 (5) (1975) p. 1.
[34] E. C. Zeeman, Catastrophe Theory: Selected Papers 1972–1977. (Addison-Wesley, Reading, MA, 1977).
[35] J. Berger, G. (Weidenfeld and Nicholson, London, 1972).
[36] M. S. Longuet-Higgins, J. Opt. Soc. Am. 50 (1960) 838.
[37] H. M. Nussenzveig, Sci. Am. 236 (1977) 116.
[38] J. F. Nye, Proc. R. Soc. Lond. A361 (1978) 21.
[39] J. F. Nye, Phil. Trans. R. Soc. Lond. A292 (1979) 25.
[40] I. R. Porteous, J. Diff. Geom. 5 (1971) 543.

[41] M. V. Berry and J. H. Hannay, J. Phys. A. 10 (1977) 1809.
[42] A. S. Thorndike, C. R. Cooley and J. F. Nye, J. Phys. A, 11 (1978) 1455.
[43] J. F. Nye and A. S. Thorndike, J. Phys. A. 13 (1980) 1.
[44] C. Upstill, Proc. R. Soc. Lond. A365 (1979) 95.
[45] C. Upstill, Catastrophe optics and caustic networks, Ph.D. Thesis (H. H. Wills Physics Laboratory, Bristol University, U.K., 1979).
[46] M. V. Berry and J. F. Nye, Nature 267 (1977) 34.
[47] J. J. Duistermaat, Commun. Pure App. Math. 27 (1974) 207.
[48] V. P. Maslov, Théorie des perturbations et méthodes asymptotiques (Dunod, Paris, 1972).
[49] Yu A. Kravtsov, Sov. Phys. Acoust. 14 (1968) 1.
[50] V. Guillemin and S. Sternberg, Geometric Asymptotics, Am. Math. Soc. Surv. No. 14 (Providence, U.S.A., 1977).
[51] G. B. Airy, Trans. Camb. Phil. Soc. 6 (1838) 379.
[52] T. Pearcey, Phil. Mag. 37 (1946) 311.
[53] M. V. Berry, J. F. Nye and F. J. Wright, Phil. Trans, R. Soc. Lond. A291 (1979) 453.
[54] M. V. Berry and F. J. Wright, J. Phys. A 13 (1980) 149.
[55] A. N. Varchenko, Funct. Anal. Appl. 10 (1976) 13.
[56] M. V. Berry, J. Phys. A. 10 (1977) 2061.
[57] B. B Mandelbrot, Fractals (Freeman, San Francisco, 1977).
[58] M. V. Berry, J. Phys. A 12 (1979) 781.
[59] M. V. Berry and Z. V. Lewis, Proc. R. Soc. A370 (1980) 459.
[60] M. V. Berry, Regular and Irregular Motion, in: Topics in Nonlinear Dynamics, ed. S. Jorna, Am. Inst. Phys. Conf. Proc. 46 (1978) 16.
[61] V. I. Arnol'd, Mathematical methods of classical dynamics (Springer, New York, 1978).
[62] M. V. Berry, J. Phys. A 10 (1977) 2083.
[63] S. W. McDonald and A. N. Kaufman, Phys. Rev. Lett. 42 (1979) 1181.
[64] M. V. Berry, Ann. Phys. N.Y. 131 (1981) 163.
[65] J. F. Nye, this volume.
[66] A. S. Thorndike, Special Features of Two-dimensional Flow Fields and their Evolution in Time, Ph.D Thesis (Polar Science Centre, Univ. of Washington, Seattle, U.S.A.).
[67] F. J. Wright, J. Phys. A13 (1980) 2913.