Huygens' Principle 1690–1990: Theory and Applications
H. Blok, H.A. Ferwerda, H.K. Kuiken (editors)

Rays, wavefronts and phase: a picture book of cusps

Michael Berry
H.H. Wills Physics Laboratory
Tyndall Avenue
Bristol BS8 1TL
UK

Abstract

Pictures of the cusp illustrate that in optics there are two sorts of wavefronts: contour surfaces of the distance function (eikonal) of geometrical optics, and contour surfaces of the phase of the optical wavefunction. In regions traversed by only one family of rays, the geometrical and phase wavefronts resemble each other. However, where rays cross (i.e. where there is interference) they are very different in form. In particular they have different singularities. Geometrical wavefronts are singular at caustics (focal surfaces), that is on the envelopes of the underlying family of rays. Phase wavefronts are singular at zeros (phase dislocations) of the wavefunction. Caustics and phase dislocations are complementary to each other.

1 Introduction

My purpose here is pedagogical. I shall give a brief description, with the aid of pictures (relegating equations to the appendix), of several contrasting pictures of light propagation, associated with the names of Huygens, Newton, Young, Airy and Pearcey. The only modern feature of the approach is the emphasis on the different singularities that appear in these approaches, and the choice of an illustrative example that is generic, i.e. stable under perturbation.

I do not claim to present any new ideas. Everything here is well known to those who know well. Nevertheless the collection of concepts as a whole is not widely known, and I have not seen it illustrated by a set of pictures like the ones which follow.

The example is a beam of monochromatic light, made convergent by refraction or reflection, which emerges into a uniform isotropic medium ('vacuum') and propagates (in the plane) through an imperfect focus in the form of a cusp.

2 Newton's rays

Here the beam is regarded as a collection of rays along which the light energy travels. The rays are straight lines whose directions are determined at the moment of refraction or reflection into the vacuum.

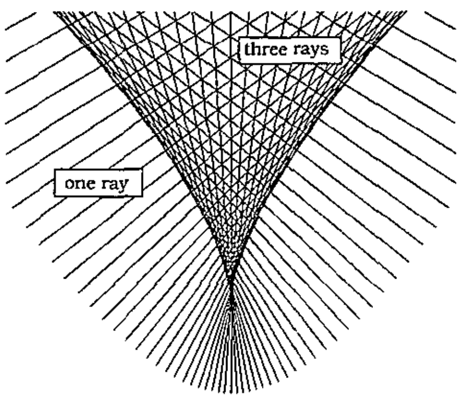

Figure 1. Converging beam of rays propagating upwards through a cusped caustic.

Figure 1 is drawn for ray deflections that vary linearly transverse to the beam. The cusped caustic, formed by the envelope of the rays, is obvious. It is the singularity of the family of rays. Note that no single ray possesses a caustic: the singularity is a holistic property of the beam. The cusp is one of the stable forms that caustics can take. 'Stable' means that the new caustic produced by perturbation of the ray family can be smoothly deformed into the old one. The general classification of these forms is provided by the catastrophe theory of Thom and Arnold [1–3]. The classification is in terms of codimension, which is the number of parameters that must be varied in order to find the singularity. The cusp in figure 1 has codimension two; it is a point where two 'fold catastrophe' curves (codimension one) meet. (In space these would be fold surfaces, and the cusp would be a line.)

The caustic is singular in two ways. First, as a locus of high light intensity – indeed the term 'caustic' comes from the Greek word for burning. In geometrical optics, and for a perfectly monochromatic and collimated initial beam, the intensity (energy density) would be infinite on the caustic. And second, as the curve across which the number of rays changes discontinuously (from one to three in figure 1).

3 Huygens' geometrical wavefronts

On this view, the effect of refraction or reflection is to produce a wavefront that is concave forwards. Each point on the wavefront is a source of secondary wavelets, whose envelope at a subsequent instant gives the new position of the front. In our example, the law of geometrical wavefront propagation is that the initial front moves perpendicular to itself, indeed in an isotropic medium the wavefronts are curves (surfaces in space) perpendicular to the rays.

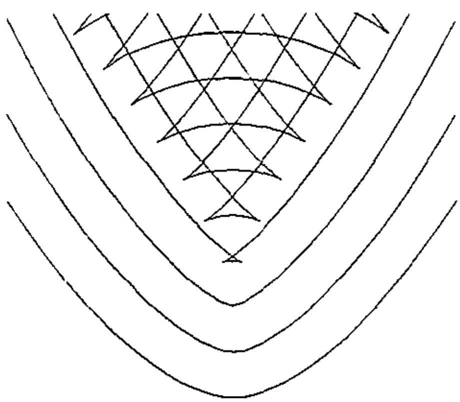

Figure 2. Geometrical wavefronts, propagating upwards.

This is shown in figure 2; the initial wavefront is a parabola. After encountering the cusp point the wavefronts crease and become three-valued inside the caustic, as they must in order to be perpendicular to the three rays there. For later reference, figure 3 shows a magnification of the central portion of figure 2, with a smaller spacing between the wavefronts. The singularities of the geometrical wavefronts are cusps whose locus is the caustic of the rays. This is clearer in figure 4, where the rays and wavefronts are superimposed.

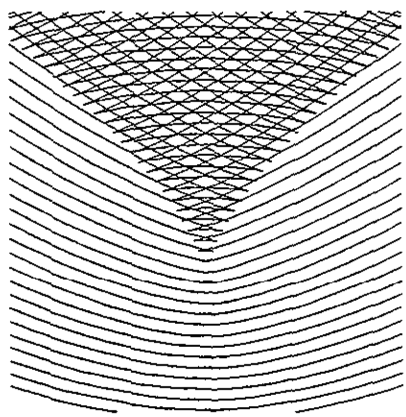

Figure 3. Magnification of central region of figure 2.

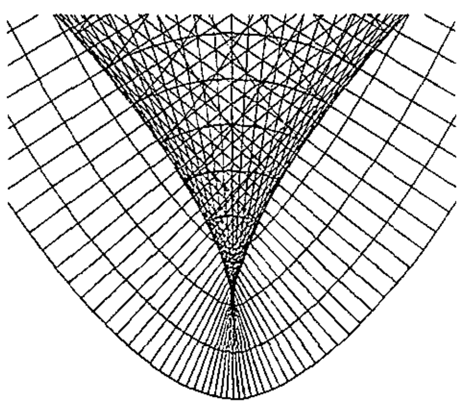

Figure 4. Rays and geometrical wavefronts superimposed.

Note that individual wavefronts, unlike individual rays, do possess singularities. However, the singularities in the wavefronts and the ray families, although both cusped, are only superficially similar: the wavefronts' cusps lie on the smooth (fold) parts of the ray caustic, not on its cusp. In catastrophe theory [2,4], the caustics are Lagrangean singularities (projections, from phase space to configuration space, of smooth manifolds representing the ray family), and wavefront cusps are Legendre singularities (contours of equal optical distance in space).

Huygens was well aware of the relation between the ray and wavefront singularities [5], namely that the caustic is the evolute (envelope of normals, i.e. locus of centres of curvature) of the initial wavefront.

4 Young's interfering waves

Neither Newton's nor Huygens' picture describes correctly what happens inside the caustic, where ray cross and wavefronts are multivalued. Young made an important step with the principle of superposition, which we can state in modern terms as follows: waves are represented by wavefunctions consisting of contributions associated with the different rays, which add where rays cross. (I am not sure whether Young had a clear notion of a wavefunction. Probably that concept emerged only gradually during the early decades of the nineteenth century. At least we can say that it was implicit in his work.) Although Young apparently rejected Huygens' wavefronts, his principle is in a sense a 'superposition' of both Newton's and Huygens' ideas. The individual disturbances travel along rays, and are periodic functions whose phase is constant across wavefronts.

Quantitative expression of this idea requires specification of the contribution associated with a ray. Nowadays it is convenient to regard this as a complex function of position, from which the physical wave is obtained (after multiplication by $\exp(-i\omega t)$ where ω is the frequency) by taking the real part. This wavefunction has an amplitude and a phase. In our example the amplitude is the square root of the wavefront curvature, which allows energy to be conserved along the ray if the energy density is regarded as proportional to the square of the amplitude. The amplitude diverges on the caustic, because the wavefront curvature is infinite there. The phase is proportional to the distance along the ray, the proportionality constant being the wavenumber $k = 2\pi/\lambda$, where λ is the wavelength. There is an additional contribution $-\pi/2$ if the ray has touched the caustic. Young was not aware of this contribution. Its existence was known in particular cases by Airy and Stokes but its generality in optics was appreciated only at the end of the nineteenth century, by Gouy [11]. (More generally still, it is a special case of the Maslov [20,21] and geometric [22] phases associated with dynamical tori, closed orbits, and parameter cycles.)

With superposition comes wave interference, leading to the prediction that inside the cusp are complicated fringes produced by the addition of the complex contributions from the three rays. These are clearly visible as arrays of maxima and minima in figure 5, which is a contour map of the amplitude near the cusp point. Also clear is the caustic singularity where two of the three component amplitudes diverge.

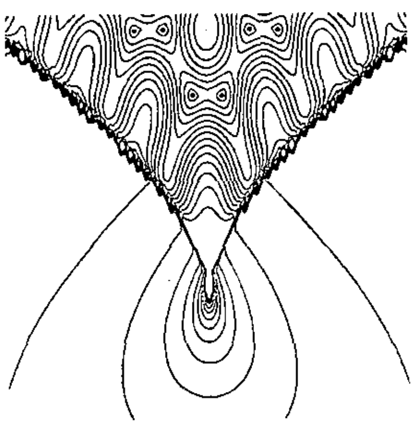

Figure 5. Amplitude contours of superposition of waves associated with the geometrical rays. There are six zeros, marked with small black circles.

What is not obvious from figure 5 is that the minima (six small black circles) are in fact zeros — points of total destructive interference. Actually their existence is not surprising, because the vanishing of a complex function typically requires two conditions and so should occur at points in the plane.

Moreover, these amplitude zeros are another type of singularity. This is clearer in figure 6, which is a contour map of the phase of the wave. This quantity is singular not only on the caustic (because of the discontinuity in the number of rays) but also at amplitude zeros, because these are points of indefinite phase, where contours of all phases meet. The contour lines are the phase wavefronts. Figure 6 shows them for phases spaced by $\pi/2$, corresponding to the crests, troughs and zeros of the real wave. Outside the caustic, where there is only one ray contribution, the phase wavefronts coincide with the geometrical wavefronts (figures 2 and 3). Inside the caustic,

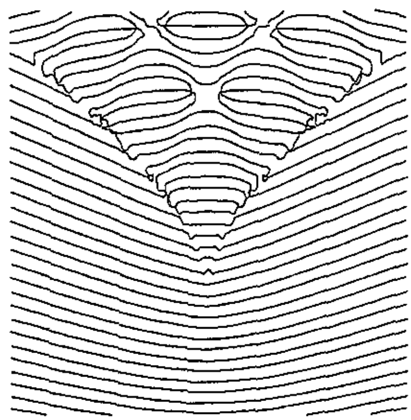

Figure 6. Phase wavefronts (at intervals of $\pi/2$) corresponding to figure 5, showing six phase dislocations where the wavefronts cross.

however, they differ fundamentally. Instead of three superposed sets of lines, there is only one set: phase is a nonlinear property of a wavefunction, so the wavefronts of a superposition are not the superposed wavefronts of the separate contributions.

The 'hedgehog' structure of these phase singularities is evident in figure 7, which shows a magnification of two of them. The singularities are called phase dislocations [6–8], because their generalisations to lines in space, and the way these lines can move, mimic crystal dislocations surprisingly closely (the analogy is between wavefronts and sheets of atoms).

5 Airy's and Pearcey's diffraction patterns

The coexistence in figures 5 and 6 of the caustic singularities of ray optics and the dislocation singularities of wave optics is illusory, because the physics underlying these pictures is wrong. Only after Young did it become clear that functions representing monochromatic waves depend smoothly on position because they have to satisfy wave equations. Therefore the caustic singularities must be artefacts of the geometrical optics limit, and must disappear in a thoroughgoing wave theory. The first explicit representation of a wave whose geometrical rays have a caustic was constructed by Airy [9] in 1838, in the form of an integral (of the sort later generalised by Kirchhoff) summing infinitesimal contributions from all points on the ini-

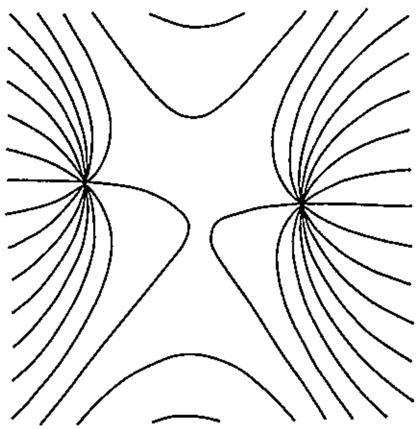

Figure 7. Magnification of figure 6 (with phase intervals $\pi/8$), showing one of the off-axis pairs of phase dislocations.

tial wavefront. This was for the wave decorating a smooth caustic, that is for the fold diffraction catastrophe. However, it was only in 1946 that the corresponding construction was carried out, by Pearcey [10], for the cusp diffraction catastrophe. Of course there had been many studies of waves near the focus of an imperfect lens (see e.g. ref. [11]) but these concerned either finite-aperture (edge diffraction) effects or the (unstable) 'spun cusp' in three dimensions, representing a lens with spherical aberration. Pearcey's function can be regarded as the exact solution for waves near a cusp, because it has been shown [12,13] that it can be deformed to fit waves even near a (smoothly) distorted cusp, in the short-wave limit.

Figure 8 shows the amplitude of this wavefunction close to the cusp. Comparison with figure 5 shows that Young's interfering waves give a good description everywhere except on the caustic and in its immediate neighbourhood. One important qualitative difference is that a new row of phase dislocations has appeared just outside the caustic on each side. At first these seem to be a consequence of the exact diffraction physics, with no counterpart in Young's superposition. However, the dislocations can be described accurately [14] as the result of destructive interference between the wave associated with the single ray outside the caustic and the complex ray persisting (as the result of analytic continuation) outside the caustic after the coalescence of two real rays on it. This unusual example shows that complex rays can affect the topology of a wavefield.

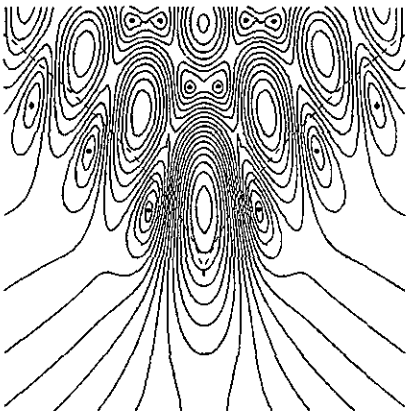

Figure 8. Amplitude contours of the cusp diffraction catastrophe. The dashed line shows the caustic. There are twelve zeros, marked with small black circles.

The corresponding phase map is shown in figure 9. Comparison with figure 6 shows that again the agreement is close except near the caustic where the glitches have disappeared. and just outside the caustic where there are the two new rows of phase dislocations. The phase wavefronts are smooth except at the phase dislocations where they cross.

6 Concluding remarks

In the geometrical optics pictures (figures 1–4), the caustic was the only singularity. In the intermediate model, of interfering contributions associated with individual rays (figures 5 and 6), the caustic is still present but in addition phase dislocations have appeared. In the final pictures, of the exact Pearcey cusp diffraction catastrophe, the caustic no longer features, but the dislocations persist as the true singularities of wave theory. As has been explained elsewhere [3], the relationship of caustics to dislocations is one of complementarity: ray caustics have infinite intensity, while at phase dislocations the wave intensity is zero; and when a caustic can be seen clearly – i.e. in the short-wave limit – the dislocations represent fine structure, too small to see, while at high magnification, when the dislocations can be resolved, the caustic is broadened by diffraction and no longer stands out clearly.

The Pearcey (cusp) and Airy (fold) diffraction catastrophes are just the first members of a hierarchy of wave patterns decorating the caustic sin-

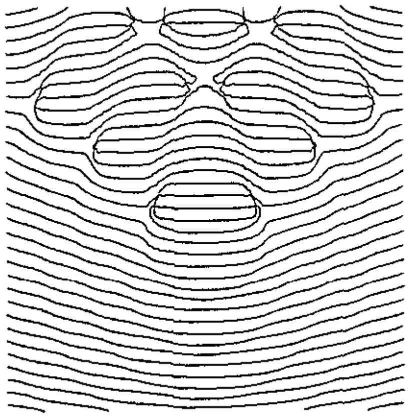

Figure 9. Phase wavefronts of the cusp diffraction catastrophe, at intervals of $\pi/2$.

gularities classified by catastrophe theory. The diffraction catastrophes are very complicated [3,14], but useful information is contained in scaling laws describing how the intensity increases and the fringe spacings decrease as the wavelength λ gets smaller. For the cusp, the intensity increases as $\lambda^{-1/2}$ and the fringes shrink as $\lambda^{1/2}$ along the cusp and as $\lambda^{3/4}$ across it. This anisotropy explains why the cusps seen at night through raindrop 'lenses' on spectacles, when looking at distant lights, are greatly elongated [3,13,15]. (Because of the wavelength dependence, the fringes should look coloured when formed with white light, but they do not. Why?)

Although Pearcey's diffraction pattern in a certain sense completes the story of waves near a cusp, it is of course far from the latest theory of light. For a start, it is based on scalar waves and so cannot describe the polarisation effects which exist in reality and in Maxwell's more accurate electromagnetic theory. In vector waves, phase dislocations persist, but are now accompanied by disclination singularities in the electric and magnetic field directions [16–18]. Beyond Maxwell's theory lies quantum optics, where the wavefunction serves to generate the probability of detecting photons. When passing to this deeper and more accurate level of description, will the phase dislocations dissolve and get replaced by new singularities, as the dislocations replaced the caustic singularities in the passage from ray to wave theory?

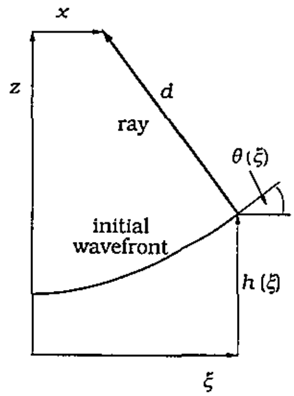

Figure 10. Notation for ray and wavefront points.

Acknowledgement

I am very happy to thank the Dutch Association of Mathematical Physics for their hospitality at the meeting in Scheveningen celebrating the 300th birthday of Huygens' principle.

Appendix

How the pictures were produced

Let the initial wavefront be described by its deviation $z = h(\xi)$ from a straight line (figure 10). For the pictures, the initial front was the parabola

$$h(\xi) = \xi^2/2. \tag{1}$$

The rays are the normals to this curve, and are obtained by drawing lines from the point $(\xi, h(\xi))$ to the point (x, z), where (denoting by θ the slope of the initial front, and indicating derivatives by primes)

$$\begin{aligned} x &= \xi - (z-h)\tan\theta \\ &= \xi - (z-h)h' \\ &= (1-z)\xi + \xi^3/2. \end{aligned} \tag{2}$$

The caustic is the envelope of this family of rays, which depends parametrically on ξ, and is obtained by setting $x'(\xi) = 0$ and then eliminating ξ using

(2). For the parabola (1) this yields the caustic

$$c(x,z) = x^2 - \tfrac{8}{27}(z-1)^3 = 0. \tag{3}$$

The discriminant $c(x,z)$ is positive outside the cusp, and negative inside.

The wavefront labelled by its distance d along the ray from the initial wavefront is given parametrically in terms of ξ by

$$\begin{aligned} x &= \xi - d\sin\theta(\xi), \\ z &= h(\xi) + d\cos\theta(\xi). \end{aligned} \tag{4}$$

In figures 1, 2 and 4, the z range is (0.6, 2.28) and the x range is (−1, 1); the wavefront spacing in figures 2 and 4 is 0.2. In figure 3 the z range is (0.625, 1.75) and the x range is (−0.3, 0.3); the wavefront spacing is 0.03.

Young's superposition gives the wave for wavelength

$$\lambda = 2\pi/k \tag{5}$$

as the sum of contributions from the geometrical rays passing through a point. Thus

$$\psi_g(x,z) = \sum_j A_j(x,z)\exp[ikd_j(x\cdot z)]. \tag{6}$$

Here j labels the real solutions $\xi_j(x,z)$ of the ray equation (2).

In the phase, d_j denotes the distance along the jth ray from the initial wavefront to (x,z), namely

$$d(\xi;x,z) \equiv [(z-h(\xi))^2 + (x-\xi)^2]^{1/2}, \tag{7}$$

evaluated at $\xi_j(x,z)$.

The amplitude A_j is the square root of the curvature of the jth wavefront at (x,z), divided by the curvature K_j of the initial wavefront on the same ray (this ensures correct normalisation for a plane wave). Curvature is the reciprocal of radius of curvature, and the radius of curvature at (x,z) is $K_j^{-1} - d_j$, so that

$$\begin{aligned} A_j &= [(K_j^{-1} - d_j)^{-1}K_j^{-1}]^{1/2} \\ &= [1 - K_j d_j]^{-1/2}. \end{aligned} \tag{8}$$

Now the curvature of the initial wavefront is

$$K_j = \frac{h''}{(1+h'^2)^{3/2}} = h''\cos^3\theta, \tag{9}$$

and this, together with (4), leads to

$$\begin{aligned} A_j &= \left(\frac{1+h'^2(\xi_j)}{1+h'^2(\xi_j)-[z-h(\xi_j)]\,h''(\xi_j)}\right)^{1/2} \\ &= \left(\frac{1+\xi_j^2}{1-z+3\xi_j^2/2}\right)^{1/2}. \end{aligned} \tag{10}$$

This expression is imaginary when the denominator inside the square root is negative, that is when the ray has touched the caustic. For reasons soon to be apparent, we should then choose the phase of A_j to be $-\pi/2$; this is the Gouy phase shift mentioned previously.

After calculating the superposition (6) with the substitutions (7–10) we obtain the complex number

$$\psi_g(x,z) = \rho(x,z)\exp[\mathrm{i}\chi(x,z)]. \tag{11}$$

Figure 5 shows the amplitude contours $\rho =$ constant, and figure 6 the phase wavefronts, that is the contours $\chi =$ constant (mod 2π) (at intervals of $\pi/2$). The ranges of x and z are the same as in figure 3, and the wavelength is $\lambda = 0.12$. Figure 7 is a magnification of figure 6 to show the structure near two of the phase dislocations; the x range is (0.05, 0.13) and the z range is (1.16, 1.74).

Figures 8 and 9, representing Pearcey's function, were calculated (for the same k, and x and z ranges, as figures 5 and 6) via the the two-dimensional diffraction integral over sources on the initial wavefront (with length element $\mathrm{d}\xi/\cos\theta$), namely

$$\psi(x,z) = \left(\frac{k}{2\pi\mathrm{i}}\right)^{1/2} \int_{-\infty}^{\infty} \frac{\mathrm{d}\xi}{\cos\theta(\xi)} \frac{\exp[\mathrm{i}kd(\xi;x,z)]}{[d(\xi;x,z)]^{1/2}}, \tag{12}$$

where $d(\xi;x,z)$ is given by (7). No obliquity factor is necessary because we are interested in the short-wave behaviour, when contributions are associated with the rays and there is no significant oblique propagation. Application of the method of stationary phase [11] yields precisely Young's superposition (6), with the rays (2) emerging as the real stationary points of the distance function $d(\xi;x,z)$. The sign of the second derivative d'' determines the phase of the contributions, and thereby disambiguates the sign of the square root in (10), leading to the Gouy phase inside the caustic. Including the complex stationary point outside the caustic, as well as the single real one, leads via interference to the single rows of dislocations on either side, as previously stated.

Near the cusp, however, all three rays originate from neighbouring points ξ on the initial wavefront and the method of stationary phase is not valid. We can obtain a description appropriate to this region by expanding the exponent d in (12) for small ξ, x, and $z-1$. From (7) and (1) we obtain

$$\begin{aligned} d(\xi,x,z) &= [R^2+\xi^4/4+(1-z)\xi^2-2x\xi]^{1/2} \\ &\approx R+\frac{1}{2R}[\xi^4/4+(1-z)\xi^2-2x\xi], \end{aligned} \tag{13}$$

where

$$R \equiv \sqrt{x^2+z^2}. \tag{14}$$

Thus

$$\psi(x,z) \approx \frac{1}{(\pi \mathrm{i})^{1/2}}\left(\frac{k}{2R}\right)^{1/4} \exp(\mathrm{i}kR)\, P(u,v), \tag{15}$$

where $P(u,v)$ denotes Pearcey's function

$$P(u,v) \equiv \int_{-\infty}^{\infty} \mathrm{d}t\, \exp[\mathrm{i}(\tfrac{1}{4}t^4+\tfrac{1}{2}ut^2+vt)], \tag{16}$$

and the variables u and v are given by

$$u=(1-z)\left(\frac{2k}{R}\right)^{1/2}, \quad v=-2x\left(\frac{k}{2R}\right)^{3/4}. \tag{17}$$

As it stands, the oscillatory integral (16) is difficult to compute, but it can be made honestly convergent by contour deformation [19], and can then be computed fairly easily, at least over the ranges and to the accuracy necessary to produce the amplitude and phase wavefront plots of figures 8 and 9.

All computations and plotting were carried out with the aid of the package Mathematica.

References

[1] T. Poston and I.N. Stewart, Catastrophe theory and its applications (Pitman, London, 1978).

[2] V.I. Arnold, Catastrophe theory (2nd Edition, Springer, Berlin, 1986).

[3] M.V. Berry and C. Upstill, Progress in Optics XVIII (1980) 257–346.

[4] V.I. Arnold, Huygens and Barrow, Newton and Hooke (Birkhauser, Basel, 1990).

[5] Y.G. Yoder, Unrolling time; Christiaan Huygens and the mathematisation of nature (Cambridge University Press, Cambridge, 1988).

[6] J.F. Nye and M.V. Berry Proc. Roy. Soc. Lond. A336 (1974) 165–90.

[7] M.V. Berry, in Les Houches Lecture Series Session XXXV, eds. R. Balian, M. Kléman and J.-P. Poirier (North-Holland, Amsterdam, 1981) pp. 453–543.

[8] F.J. Wright and J.F. Nye, Phil. Trans. Roy. Soc. Lond. A305 (1982) 339–382.

[9] G.B.Airy, Trans. Camb. Phil. Soc. 6 (1838) 379–403.

[10] T. Pearcey, Phil. Mag. 37 (1946) 311–317.

[11] Mi Born and E. Wolf, Principles of Optics (Pergamon, London, 1959).

[12] J.J. Duistermaat, Commns. Pure App. Math. 27 (1974) 207–281.

[13] M.V. Berry, Advances in Physics 25 (1976) 1–26.

[14] M.V. Berry, J.F. Nye and F.J. Wright, Phil. Trans. Roy. Soc. A291 (1979) 453–84.

[15] J.F. Nye, Proc. Roy. Soc. Lond. A361 (1978) 24–41.

[16] J.F. Nye, Proc. Roy. Soc. Lond. A387 (1983) 105–132.

[17] J.F. Nye, Proc. Roy. Soc. Lond. A389 (1983) 279–290.

[18] J.F. Nye and J.V. Hajnal, Proc. Roy. Soc. Lond. A409 (1987) 21–36.

[19] J.N.L. Connor and P.R. Curtis, J. Phys. A 15 (1982) 1179–1190.

[20] V.P. Maslov and M.V. Fedoriuk, Semiclassical approximation in quantum mechanics (North-Holland, Amsterdam, 1981).

[21] J.R. Robbins, Nonlinearity, in press, 1991

[22] A. Shapere and F. Wilczek, Geometric phases in physics (World Scientific, Singapore, 1989).