Van Kampen, N. G.
1958

Physica XXIV
Zernike issue
437–444

THE METHOD OF STATIONARY PHASE AND THE METHOD OF FRESNEL ZONES

N. G. VAN KAMPEN

Rijksuniversiteit te Utrecht, Nederland *)

Synopsis

The method of stationary phase is discussed and reasons are given why it is often preferable to the method of steepest descent, in spite of the mathematicians' partiality to the latter. The method of Fresnel zones is compared with the method of stationary phase and is shown to be a primitive version of it. The purpose of the zones is to reduce the double integral to a single integral; this same artifice underlies the various attempts to prove the method of stationary phase for double integrals.

1. *Introduction.* The method of Zernike and Nijboer [1]) for calculating the diffraction pattern in the presence of optical aberrations, is based on an expansion in powers of the aberration. Hence it is suitable when the distance of the actual wave front from the ideal spherical form is not more than a few wave lengths. This is, of course, the case in which the influence of diffraction on the image formation is of most interest from a practical point of view. In the opposite case of large aberration, the image of a point source can be found to first approximation by means of geometrical optics. However, Nienhuis' experiments with large aberrations [2]) exhibit a variety of detailed structures in the image, which cannot be fully explained in this way. It is therefore not surprising that in 1947 the question arose of how to calculate the influence of diffraction when the aberration is large compared to the wave length.

Mathematically, the problem is the following. Huygens' principle states that the light amplitude u in a point Q of the image space is a sum of contributions from all points in the exit pupil of the optical system. According to Kirchhoff, this can be formulated as an integral, [3])

$$u(Q) = k \iint g(x, y) e^{-ikf(x,y)} \, dxdy. \tag{1}$$

Here x and y are coordinates in the exit pupil, and $g(x, y)$ measures the amplitude in each point (x, y) of the exit pupil.**) $kf(x, y)$ is the phase with which

*) Temporarily at Columbia University, New York, United States.

**) Actually $g(x, y)$ contains two more factors, but they are practically constant in many cases, and at any rate irrelevant for the the present discussion.

a light ray from (x, y) arrives in Q. Alternatively one may say that k is the wave number, and $f(x, y)$ the distance between Q and the projection of (x, y) on some fixed wave front, for instance the one through the centre of the exit pupil. Our problem is to evaluate this integral for large aberrations, or what amounts to the same, for large k.

2. *The method of stationary phase (MSP).* The appropriate mathematical method for dealing with integrals involving rapidly varying phase factors is the method of stationary phase [4] [5]. This method dates back to Stokes and has been used by Kelvin to derive the concept of group velocity [4]. It is a valuable tool in many different fields of physics, wherever one has to do with a superposition of a continuous distribution of coherent waves [6].

The basic idea is that for large k the exponential varies so rapidly that the contributions from the various elements $dxdy$ to the integral almost completely cancel each other, so that the net result is extremely small. This cancellation, however, will not be effective if f does not vary with x and y; therefore the contribution from the vicinity of any point where f is stationary i.e.,

$$\partial f/\partial x = \partial f/\partial y = 0, \tag{2}$$

must be calculated separately. It turns out that these are the main contributions to the integral. Such points are called "critical points of the first kind". In addition, the cancellation may also become ineffective in the neighbourhood of the limits of integration, which gives rise to critical points of the second and third kinds. Their contributions are of a higher order in $1/k$ and represent the diffraction at the edge of the aperture [7].

Until recently, mathematicians were little interested in the method of stationary phase [8]. The reason is that by shifting the integration path into the complex plane, one can (at least for integrals with one variable) make the exponent real, so that the method of steepest descent ("Sattelpunktmethode", reference (4), p. 235) applies. It is then much easier to prove that the result is actually an asymptotic series for large k (in the sense of Poincaré [9]). Yet I feel that it is incorrect to regard the method of stationary phase as an illegitimate child of the steepest descent, for the following reasons.

(i) Even in simple examples the construction of the correct path of steepest descent often requires a cumbersome investigation of the analytic behavior of the exponent f, whereas the MSP only takes a few lines. It is true that the latter gives less general results, but usually it gives just those results in which the physicist is interested.

(ii) The method of steepest descent cannot readily be applied to integrals between finite limits, unless these limits happen to be essential singularities of the integrand. In mathematics this is usually the case, for instance, when the integral arises from the solution of a differential equation. But in inte-

grals arising from physical problems the limits are often arbitrary numbers.

(iii) The generalization of the method of steepest descent to multiple integrals seems hardly practicable.

(iv) There is in general no justification to assume of a function arising in a physical problem (such as in our case the aberration function) that it has an analytic continuation. It is true that one always approximates such functions by analytical expressions, but the behaviour of these approximating functions in the complex plane depends strongly on the special form of the approximation one chooses. Hence no direct physical meaning can be attributed to the contributions of the various saddle points in the complex plane. On the other hand, when using the MSP the contributions from the different critical points have an immediate physical interpretation *). This will be made clear in the following section by showing that the MSP, when applied to the Kirchhoff integral, can be regarded as an improved version of the well known "method of Fresnel zones".

3. *Comparison with the method of Fresnel zones (MFZ).* Since $f(x, y)$ is the distance between the fixed point Q and the point (x, y), the condition (2) tells that each critical point of the first kind is the foot of a normal through Q on the wave front. These are just the points from which a geometrical ray passes through Q. In the neighbourhood of such a critical point, (x_0, y_0) say, f can be expanded

$$\Delta f \equiv f(x_0 + \xi, y_0 + \eta) - f(x_0, y_0) = a\xi^2 + 2b\xi\eta + c\eta^2 + \ldots .$$

The following possibilities must be distinguished.

(*a*) f has a maximum in (x_0, y_0), Δf is negative definite;

(*b*) has a minimum in (x_0, y_0), Δf is positive definite;

(*c*) Δf takes both positive and negative values, f is only stationary;

(*d*) Δf is semi-definite, there is one direction ξ/η along which f does not vary.

For each of these cases we shall compare the MFZ with the MSP.

(*a*) The method of stationary phase yields for the contribution of a point (x_0, y_0) of this type, in lowest approximation [7]),

$$kg_0\, e^{-ikf_0} \int\int_{-\infty}^{+\infty} e^{-ik(a\xi^2+2b\xi\eta+c\eta^2)}\, \mathrm{d}\xi\, \mathrm{d}\eta$$
$$= g_0\, e^{-ikf_0}\, e^{\frac{1}{2}\pi i}\, \pi/\sqrt{ab - c^2} \qquad (3)$$

The first factor is the amplitude in the point x_0, y_0, the second factor is the phase lag due to the distance from (x_0, y_0) to Q. The phase factor $e^{\frac{1}{2}\pi i}$ is connected with the fact that the rays coming from the neighbourhood of (x_0, y_0) have a focus between Q and the wave front. The remaining amplitude

*) It should be emphasized, however, that the MSP only applies when f is real. This remark removes the arbitrariness in the result which Eckart [6]) found. In dispersion theory complex exponents do occur, so that one is forced to shift the integration path. Fortunately in this case the causality principle guarantees that an analytic continuation exists!

factor measures the area on the wave front that is effective in contributing to the amplitude in Q.

According to the method of Fresnel zones [10]) one would of course write the first and second factor, and also the third one, although this "anomalous" phase has been the subject of some discussion [11]). Instead of the last factor one would calculate the area A of the first Fresnel zone, that is, the area of the ellipse (fig. 1)

$$\Delta f \equiv a\xi^2 + 2b\xi\eta + c\eta^2 = -\tfrac{1}{2}\lambda;$$

which is

$$A = \tfrac{1}{2}\pi\lambda(ab - c^2)^{-\frac{1}{2}}. \qquad (4)$$

One then has to take half the contribution of this first zone, which leads to the result (3) with a spurious factor $\frac{1}{4}\lambda$. Thus the correct amplitude is

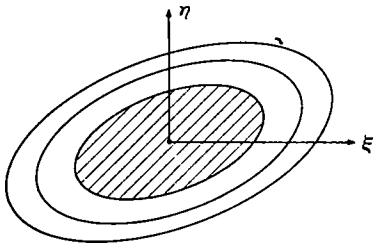

Fig. 1

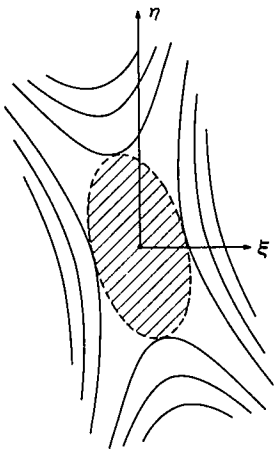

Fig. 2

obtained by taking the result of the MFZ and adding the factor $4/\lambda = 2k/\pi$.

(b) This case is identical with (a), but for the anomalous phase factor, which is now $e^{-\frac{1}{2}\pi i}$. To this category belongs the propagation of light in free space; one then has to take for the wave front a sphere around the light source [12]).

(c) The MSP gives almost the same expression (3); one only has to write $|ab - c^2|$ under the square root, because the determinant is now negative,

and moreover the phase factor $e^{\frac{1}{4}\pi i}$ has to be omitted. Fokker [13]) suggested the application of the method of Fresnel zones to this case, but as the Fresnel zones are hyperbolic and extend to infinity, it is not possible to guess the factor that measures the effective area on the wave front. The MSP, however, gives the following recipe. Take the hyperbolas given by

$$\Delta f = \pm \tfrac{1}{2}\lambda;$$

then (4) is the area of the inscribed ellipse (fig. 2), supposing that the absolute value of the square root is taken. Thus the light amplitude arising from a critical point of the first kind with hyperbolic Fresnel zones may be written

$$(2k/\pi)\, g_0\, e^{-ikf_0} \,.\, \tfrac{1}{2}A.$$

(*d*) There is a "ridge" on the wave front, i.e., a curve on which $f(x, y)$ is constant and greater (say) than in the neighbourhood of the curve (fig. 3).

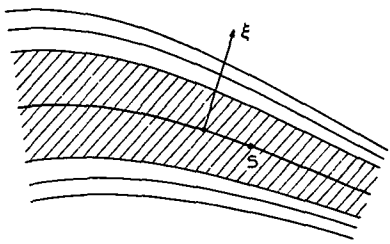

Fig. 3

All points on the curve have stationary phase. Let s measure the length along the curve and ξ the distance from it, then

$$f(x, y) = f_1(s) - \xi^2 f_2(s) + \ldots.$$

The contribution of this stationary curve to (1) is

$$k \int g_1(s)\, e^{-ikf_1(s)}\, \mathrm{d}s \int_{-\infty}^{+\infty} e^{ikf_2(s)\xi^2}\, \mathrm{d}\xi =$$
$$= \tfrac{1}{2}\, k \int g_1(s)\, e^{-ikf_1(s)}\, B(s)\, \mathrm{d}s \,.\, e^{\frac{1}{4}\pi i}.$$

Here $B(s) = \sqrt{2\lambda/f_2(s)}$ is the breadth of the first Fresnel zone, i.e., the strip along the curve in which the variation of f is less than $\frac{1}{2}\lambda$. From the MFZ one would find half the integral, so that the correct result differs from it by an additional factor k.

A critical point of the second kind is a point on the edge of the exit pupil for which f is stationary along the edge: $\partial f/\partial s = 0$. These points give rise to "diffraction rays", which were first derived from the Kirchhoff integral by Rubinowicz [14]). In the neighbourhood of such a point (x_0, y_0) one has

$$\Delta f = a\xi^2 + b\eta.$$

There are two possibilities to be distinguished, according as the parabolas of constant Δf look like in fig. 4 or like in fig. 5 (a and b of the same or the opposite sign respectively).

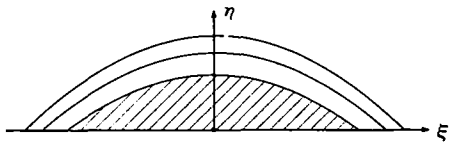

Fig. 4

(*e*) In the first case (fig. 4) the MSP gives for the contribution of (x_0, y_0)

$$kg_0\, e^{-ikf_0} \int_{-\infty}^{+\infty} e^{-ika\xi^2}\, \mathrm{d}\xi \int_0^{\infty} e^{-ikb\eta}\, \mathrm{d}\eta =$$

$$= kg_0\, e^{-ikf_0} \sqrt{\frac{\pi}{k\,|a|}}\, e^{\mp \frac{1}{4}\pi i}\, \frac{1}{ikb}. \tag{5}$$

The area of the first Fresnel zone is $A = \frac{1}{3}\sqrt{2}\, \lambda^{3/2}\, |a|^{-\frac{1}{2}}\, |b|^{-1}$. Thus (5) may be written

$$[3k/2\pi i]\, g_0\, e^{-ikf_0} . \tfrac{1}{2} A\, . e^{\mp \frac{1}{4}\pi i}, \tag{6}$$

exhibiting the factor that has to be added to the MFZ to obtain the exact result.

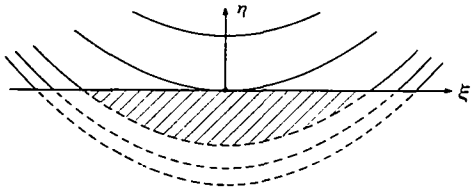

Fig. 5

(*f*) The case of fig. 5 need not be computed separately, because the contribution of this critical point must be just the opposite of the wave that is suppressed by the diaphragm. Hence the contribution of such a point is (6) with a minus sign, where A is now the area of the first "hidden Fresnel zone".

4. *Discussion.* In the preceding section it was explicitly shown for a number of examples that the MFZ gives the same result as the first order MSP. The only factor which the MFZ is unable to determine depends on the wave length and the nature of the critical point, but not on the values of the coefficients in f; these are taken care of in the factor A. What is the general reason why these coefficients always occur in a combination that can be interpreted as the area of a Fresnel zone?

To answer this question, we consider the contribution from the neighbourhood of one particular critical point (x_0, y_0). Let $G(v)\, \mathrm{d}v$ be the area between the curves $\Delta f(x, y) = v$ and $\Delta f(x, y) = v + \mathrm{d}v$. Suppose that for $v \to 0$ the function $G(v)$ behaves like Cv^p, where p is a number depending on the nature

of the singularity. For example, $p = 0$ in case (a) and $p = \frac{1}{2}$ in case (e). The area of the first Fresnel zone is

$$A = \int_0^{\frac{1}{2}\lambda} G(v)\,\mathrm{d}v = C(p+1)^{-1}\,(\tfrac{1}{2}\lambda)^{p+1}. \tag{7}$$

On the other hand, using $G(v)$ one can write the contribution to (1) from the neighbourhood of (x_0, y_0) as a single integral:

$$u = kg_0\, e^{-ikf_0} \int_0^{\varepsilon} G(v)\, e^{-ikv}\,\mathrm{d}v.$$

The MSP for single integrals can now be applied, giving for the contribution at $v = 0$ to first approximation

$$kg_0\, e^{-ikf_0}\, C \int_0^{\varepsilon} v^p\, e^{-ikv}\,\mathrm{d}v \sim kg_0\, e^{-ikf_0}\, C\Gamma(p+1)/(ik)^{p+1}$$

On account of (7) this can be written in the form

$$[2k(\pi i)^{-p-1}\,\Gamma(p+2)]\; g_0\, e^{-ikf_0}\;.\;\tfrac{1}{2}A.$$

The last two factors constitute the expression one would find from the MFZ. The factor i^{-p-1} in the brackets is the anomalous phase, and the other quantities in [ ] are the factor that cannot be found from the MFZ.

In the above argument the investigation of a critical point of a double integral has been reduced to applying the MSP to a single integral. This artifice has been used already by Kirchhoff [12]) in the case of a spherical wave front (in which case there is just one critical point of the first kind). More generally, it provides a means for proving the correctness of the MSP for double integrals *), because the MSP for single integrals is much easier to handle and has been thoroughly studied by Van der Corput [15]). Several proofs along this line have been published [16]) **) However, there is one difficulty.

The artifice does not apply to critical points of "hyperbolic type", such as case (c), fig. 2. Here the Fresnel zones are not confined to the vicinity of the critical point, but extend far into the exit pupil. Hence one must first isolate the critical point (by means of a neutralisator [15])), which makes it possible to define $G(v)$; it then turns out that $G(v)$ has a logarithmic singularity at $v = 0$. This can again be handled by applying the MSP for single integrals, if one deals with the positive and negative values of v separately. Recently a proof along these lines was obtained by Berghuis [17]). However by separating the positive and negative values of v one disregards the fact that they largely compensate each other, so that the results are not as simple as

*) It is easy to show that the MSP gives rise to a power series in $1/k$, but what has to be proved is that this series is an asymptotic expansion in the sence of Poincaré. Of course, one should not confuse a heuristic derivation with a proof.[19]).

**) The obvious approach would be to treat (1) as a repeated integral and to apply the MSP to the x and y integrations separately. This, however, leads to difficulties, because the first step gives rise to zeros in the denominator. These difficulties can be partly overcome in the lowest order [20]), but probably not in the higher terms of the asymptotic expansion.

might be desired. A straightforward proof for critical points of this type does not yet exist to my knowledge.

In addition to the simple examples of critical points considered here, there are many more complicated types, which may be regarded as arising from the coincidence of two or more simple critical points. Such a coincidence gives rise to a discontinuity in the asymptotic expansion ("Stokes' phenomenon"), which shows up physically as a focus or a caustic, or as the boundary of the geometrical shadow. The problem of finding asymptotic expansions that are valid in the vicinity of such singularities has been studied by several authors [14] [7] [18], but will not be discussed here.

Note added in proof. Recently another treatment of the method of stationary phase was given by D. S. Jones and M. Kline [Journ. math. Phys. **38** (1958) 1].

Received 7-3-58

REFERENCES

1) Zernike, F., Physica **1** (1934) 689; Nijboer, B. R. A., Physica **10** (1943) 679; **13** (1947) 605; Nienhuis, K., and Nijboer, B.R.A., Physica **14** (1948) 590.
2) Nienhuis, K., Thesis (Groningen 1948).
3) Born, M., *Optik* (Berlin 1933) p. 147.
4) Watson, G. N., *Theory of Bessel Functions* (Cambridge 1922) p. 229.
5) Van der Corput, J. G., Proc. kon. Ned. Akad. Amsterdam **42**[2] (1939) 468; Bijl, J., Nieuw Arch. Wisk. (2) **19** (1936) 63; Erdélyi, A., *Asymptotic Expansions* (New York 1956) p. 51.
6) Eckart, C., Revs mod. Phys. **20** (1948) 399.
7) Van Kampen, N. G., Physica **14** (1948) 575; **16** (1950) 817.
8) See for instance Jeffreys, H. and Jeffreys, B. S., *Methods of mathematical Physics* (Cambridge 1946) p. 474.
9) Whittaker, E. T., and Watson, G. N., *A Course of modern Analysis* (Cambridge 1927) p. 150.
10) Fresnel, A., *Oeuvres Complètes* **1** (Paris 1866) p. 295; Schuster, A., *An introduction to the theory of optics* (London 1904) Ch. V.; Mach, E., *Principien der physikalischen Optik* (Leipzig 1921) p. 385.
11) Reiche, F., Ann. der Phys. (4) **29** (1909) 65 and 401; **30** (1909) 182; Debye, P., Ann. der Phys. (4) **30** (1909) 755; Fokker, A. D., Physica, Ned. T.Ntk. **3** (1923) 334; Rubinowicz, A., Phys. Rev. **54** (1938) 931.
12) Kirchhoff, G., *Vorlesungen über mathematische Optik* (Leipzig 1891) p. 33.
13) Fokker, A. D., Physica, Ned. T. Ntk., **4** (1924) 116.
14) Rubinowicz, A., Ann. der Phys. (4) **73** (1924) 339.
15) Van der Corput, J. G., Compos. Math. **1** (1934) 15; **3** (1936) 328; Indagationes Math. **17** (1955) 139.
16) Hlawka, E., Monatshefte für Mathem. **54** (1950) 1; Focke, J., Berichte über die Verhandl. der sächs. Akad. Wissens., Mathem.-naturw. Klasse **101**, Heft 3 (1954); G. Braun, Acta Physica Austriaca **10** (1957) 8.
17) Berghuis, J., Thesis Delft ('s-Gravenhage 1955).
18) Van der Waerden, B. L., Appl. sci. Research (B) **2** (1951) 33.
19) Wolf, E., Reports on Progress in Physics **14** (1951) 95.
20) Jones, D. S., Proc. Camb. phil. Soc. **48** (1952) 733.