Journal of Computational and Applied Mathematics 30 (1990) 213–225
North-Holland

Multidimensional stationary phase approximation: boundary stationary point *

J.P. McCLURE
Department of Mathematics and Astronomy, University of Manitoba, Winnipeg, Canada R3T 2N2

R. WONG
Department of Applied Mathematics, University of Manitoba, Winnipeg, Canada R3T 2N2

Received 1 June 1989
Revised 26 September 1989

Abstract: A necessary and sufficient condition is established for the existence of a nonsingular matrix which simplifies a linear form to a single coordinate and at the same time retains a quadratic form. A version of Morse's lemma is also derived. These results are then used in a rigorous derivation of the asymptotic expansion of the oscillatory integral $I(\lambda) = \int_D g(x)\, e^{i\lambda f(x)}\, dx$ $(x \in \mathbb{R}^n)$, where a stationary point of $f(x)$ lies on the boundary of D.

Keywords: Oscillatory integral, asymptotic expansion, stationary phase approximation.

1. Introduction

Oscillatory integrals of the form

$$I(\lambda) = \int_D g(x)\, e^{i\lambda f(x)}\, dx, \quad x \in \mathbb{R}^n, \tag{1.1}$$

occur frequently in diffraction theory [8,1] and partial differential equations [3, §7.7], [7, p.428]. In (1.1), D is a bounded domain, f and g are C^∞-functions in D, and λ is a large positive parameter. Such integrals are usually evaluated asymptotically by the method of stationary phase (see, e.g., [7, p.431] and [2, p.75]), which gives

$$I(\lambda) \sim g(x_0)\, |\det A|^{-1/2} \exp\{i\lambda f(x_0) + i\tfrac{1}{4}\pi\sigma\} \left(\frac{2\pi}{\lambda}\right)^{n/2}, \tag{1.2}$$

where x_0 is a stationary point of f (i.e., $\nabla f(x_0) = 0$), A is the Hessian matrix of f defined by

$$A = \left(\frac{\partial^2 f}{\partial x_i\, \partial x_j} \right)\Bigg|_{x = x_0} \tag{1.3}$$

* This research was partially supported by the Natural Sciences and Engineering Research Council of Canada under grants A-8069 and A-7359.

and σ is the signature of the matrix A. (Recall that the *signature* of a matrix is the number of positive eigenvalues minus the number of negative eigenvalues.) The validity of this approximation requires that the stationary point x_0 is nondegenerate (i.e., $\det A \neq 0$) and lies in the interior of D.

If x_0 is on the boundary of D, then formula (1.2) no longer holds. In [4, p.386], Jones has shown that in this case the asymptotic approximation of $I(\lambda)$ is exactly half of that given in (1.2). However, Jones' argument is not quite rigorous and requires further justification. We are particularly concerned with (i) the transformation which he has used to simplify a linear form to a single coordinate and at the same time retain a diagonalized quadratic form, and (ii) the approximations of $g(x)$ and $f(x)$ by $g(x_0)$ and $f(x_0) + \frac{1}{2}(x - x_0)^{\mathrm{T}} A(x - x_0)$, respectively. For convenience, we shall present a brief outline of Jones' analysis in the following section. Throughout, the superscript "T" on a vector or matrix indicates the operation of transposition.

2. Jones' argument

Let the boundary of D be given by $h(x) = 0$ with $\nabla h(x) \neq 0$, and expand f and h at x_0 to obtain

$$f(x) = f(x_0) + \tfrac{1}{2}(x - x_0)^{\mathrm{T}} A(x - x_0) + \cdots,$$

$$h(x) = b^{\mathrm{T}}(x - x_0) + \tfrac{1}{2}(x - x_0)^{\mathrm{T}} B(x - x_0) + \cdots,$$

where b denotes the gradient of h evaluated at x_0, and A and B are respectively the Hessian matrices of f and h at x_0. Let L be the orthogonal (i.e., real and unitary) matrix which diagonalizes the quadratic form $(x - x_0)^{\mathrm{T}} A(x - x_0)$, i.e., if $y = L(x - x_0)$, then

$$(x - x_0)^{\mathrm{T}} A(x - x_0) = \sum_{j=1}^{n} \mu_j y_j^2,$$

where the μ_j's are the eigenvalues of A. In terms of y, we have

$$f(x) = f(x_0) + \tfrac{1}{2} \sum_{j=1}^{n} \mu_j y_j^2 + \cdots, \tag{2.1}$$

$$h(x) = b^{\mathrm{T}} L^{\mathrm{T}} y + \tfrac{1}{2} y^{\mathrm{T}} L B L^{\mathrm{T}} y + \cdots. \tag{2.2}$$

Now Jones claims that there exists a matrix M such that $y = Mu$ gives

$$h(x) = k_0 u_1 + \tfrac{1}{2} u^{\mathrm{T}} M^{\mathrm{T}} L B L^{\mathrm{T}} M u + \cdots, \tag{2.3}$$

$$f(x) = f(x_0) + \tfrac{1}{2} \sum_{j=1}^{n} \mu_j u_j^2 + \cdots, \tag{2.4}$$

where k_0 is a constant. We assume that the interior of D corresponds to $u_1 > 0$. Observe that this transformation eliminates all linear terms in (2.2) except for the first and at the same time retains the quadratic form in (2.1). The existence of the transformation is, however, not clear. In the next section, we shall show that in fact the matrix M exists if and only if

$$\sum_{j=1}^{n} \frac{d_j^2}{\mu_j} \neq 0, \quad d \equiv Lb \equiv (d_1, \ldots, d_n). \tag{2.5}$$

A geometrical interpretation of (2.5) in the case $n = 2$ is that the boundary curve $h(x) = 0$ meets the level set $f(x) = f(x_0)$ *non*tangentially. To see this, we suppose $x_0 = 0$. If μ_1 and μ_2 have the same sign (i.e., f has a local extremum at 0), then (2.5) holds for all $d \neq 0$ and hence is not a restriction. However, if μ_1 and μ_2 have opposite signs, then $f(x) = f(0)$ can be shown to consist of two smooth curves intersecting at 0, and the pair of straight lines determined by $\mu_1 x_1^2 + \mu_2 x_2^2 = 0$ can be shown to be tangent to these curves. Condition (2.5) says that the boundary curve $h(x) = 0$ is tangent to *neither* of the straight lines of $\mu_1 x_1^2 + \mu_2 x_2^2 = 0$.

Jones' next step is to introduce a new coordinate system $X = (X_1, \ldots, X_n)$ in which $h = 0$ becomes $X_1 = 0$ and the positive X_1-axis is inside D. Thus he puts $k_0 X_1 = h(x)$ and $X_j = u_j$ $(j \neq 1)$. Solving for u_j's in terms of X_j's, one obtains

$$u_1 = X_1 + \text{quadratic terms} + \cdots, \tag{2.6}$$

$$f(x) = f(x_0) + \tfrac{1}{2} \sum_{j=1}^{n} \mu_j X_j^2 + \cdots. \tag{2.7}$$

Let $F(X)$ denote the transform of $f(x)$ and define

$$G(X) = g(x) \left| \frac{\partial(x_1, \ldots, x_n)}{\partial(X_1, \ldots, X_n)} \right|. \tag{2.8}$$

The integral (1.1) becomes

$$I(\lambda) = \int G(X)\, \mathrm{e}^{\mathrm{i}\lambda F(X)}\, \mathrm{d}X, \tag{2.9}$$

where the integration is over the half-space $X_1 > 0$. From (2.7), we have

$$F(X) = f(x_0) + \tfrac{1}{2} \sum_{j=1}^{n} \mu_j X_j^2 + \cdots. \tag{2.10}$$

Jones now replaces $G(X)$ by $G(0)$, and $F(X)$ by its second-degree Taylor polynomial. Formally this gives

$$I(\lambda) \sim G(0)\, \mathrm{e}^{\mathrm{i}\lambda f(x_0)} \int \exp\left\{ \mathrm{i}\tfrac{1}{2}\lambda \sum_{j=1}^{n} \mu_j X_j^2 \right\} \mathrm{d}X. \tag{2.11}$$

The last integral can be written as a product of n one-dimensional integrals, each of which can be evaluated in closed form. Since the interval of integration in X_1 is only from 0 to ∞, the result is expected to be

$$I(\lambda) \sim \tfrac{1}{2} \left(\frac{2\pi}{\lambda} \right)^{n/2} G(0)\, |\mu_1 \cdots \mu_n|^{-1/2} \exp\left\{ \mathrm{i}\lambda f(x_0) + \mathrm{i}\tfrac{1}{4}\pi \sum_{j=1}^{n} \operatorname{sgn} \mu_j \right\}. \tag{2.12}$$

Note that the matrix L in the change of variable from x to y is unitary, hence the Jacobian of the transformation $y = L(x - x_0)$ is equal to 1. Also, from (2.6), it is easily seen that the value of the Jacobian of the transformation $u \to X$ is one. Thus, if one can choose the matrix M in (2.3) to satisfy $\det M = 1$, then

$$\left. \frac{\partial(x_1, \ldots, x_n)}{\partial(X_1, \ldots, X_n)} \right|_{X=0} = 1, \tag{2.13}$$

which in turn gives $G(0) = g(x_0)$. This together with the facts $\det A = \mu_1 \cdots \mu_n$ and $\sigma = \sum_{j=1}^{n} \operatorname{sgn} \mu_j$ yields

$$I(\lambda) \sim \tfrac{1}{2}\left(\frac{2\pi}{\lambda}\right)^{n/2} g(x_0)\,|\det A|^{-1/2} \exp\{i\lambda f(x_0) + i\tfrac{1}{4}\pi\sigma\}, \tag{2.14}$$

which is indeed half of the approximation given in (1.2).

Relationship (2.11) must be justified, before (2.14) can be claimed to be an asymptotic approximation. At the moment it is not clear why the error in this approximation is $o(\lambda^{-n/2})$ as $\lambda \to \infty$. In Section 5, we shall show that the error is in fact $O(\lambda^{-n/2-1})$.

One of the difficulties in justifying (2.11) is the approximation of $F(X)$ by its second-degree Taylor polynomial. This difficulty could be avoided by mapping the phase function $f(x)$ in (1.1) to a sum of squares right at the beginning by using Morse's lemma [6, p.54]. But the complication will resurface once we introduce the new coordinate X in which the boundary of the domain becomes $X_1 = 0$, since the new phase function $F(X)$ will no longer be a sum of squares. On the other hand, if we apply Morse's lemma to the phase function $F(X)$ in (2.9), instead of to $f(x)$ in (1.1), then the domain of integration in (2.9) may no longer be the half-space $X_1 > 0$. Thus neither of these approaches can be used to easily justify the step in (2.11). In Section 4, we derive a version of Morse's lemma, which shows that near the critical point, there is a change of variable $X \to t$ which transforms $F(X)$ to a sum of squares, *and* maps the half-space $X_1 > 0$ to the half-space $t_1 > 0$.

3. Existence of the matrix M

In this section we shall prove the following general result, from which our claim in (2.5) will follow.

Theorem 1. *Let d be any nonzero vector in $\mathbb{R}^n$, and let C be a nonsingular and symmetric $n \times n$ matrix. A necessary and sufficient condition for the existence of a nonsingular $n \times n$ matrix N such that the change of variable $u = Ny$ gives*

$$d^{\mathrm{T}}y = k_0 u_1, \quad k_0 \text{ being a constant}, \tag{3.1}$$

and

$$y^{\mathrm{T}}Cy = \beta_1 u_1^2 + \cdots + \beta_n u_n^2, \quad \beta_i \text{ also being constants}, \tag{3.2}$$

is that

$$d^{\mathrm{T}}C^{-1}d \neq 0. \tag{3.3}$$

If C is the diagonal matrix $\operatorname{diag}(\mu_1, \ldots, \mu_n)$, then condition (3.3) reduces to (2.5). (Recall that C is nonsingular, so $\mu_i \neq 0$ for all i.)

Proof. We first prove the *necessity*. Suppose that N exists. Then (3.1) implies that the first row of N is proportional to d, and (3.2) implies that the matrix $(N^{-1})^{\mathrm{T}}CN^{-1}$ is diagonal. Thus, if $v_1, \ldots, v_n$ are the column vectors of N^{-1}, then we must have

$$v_i^{\mathrm{T}}Cv_j = 0 \quad \text{if } i \neq j. \tag{3.4}$$

Since the first row of N is proportional to d, in view of the relationship between N and N^{-1}, we must also have

$$d^{\mathrm{T}}v_j = 0 \quad \text{if } j \neq 1, \tag{3.5}$$

and since the vectors v_j are linearly independent (columns of an invertible matrix), it follows that

$$d^{\perp} = \operatorname{span}\{v_2, \ldots, v_n\}. \tag{3.6}$$

Now, C being symmetric, $v_1^{\mathrm{T}}C = v_1^{\mathrm{T}}C^{\mathrm{T}} = (Cv_1)^{\mathrm{T}}$. Thus (3.4) with $i = 1$ implies that

$$(Cv_1)^{\mathrm{T}}v_j = 0 \quad \text{for } j = 2, \ldots, n. \tag{3.7}$$

From (3.6) and (3.7), we see that both d and Cv_1 are orthogonal to $\operatorname{span}\{v_2, \ldots, v_n\}$. Hence, Cv_1 is a constant multiple of d, or equivalently v_1 is a constant multiple of $C^{-1}d$. But the linear independence of $\{v_1, \ldots, v_n\}$ implies $v_1 \notin \operatorname{span}\{v_2, \ldots, v_n\} = d^{\perp}$. Therefore, $C^{-1}d$ cannot be orthogonal to d, i.e., $d^{\mathrm{T}}C^{-1}d \neq 0$, as required.

We next prove the *sufficiency*. First we observe that it suffices to obtain linearly independent vectors $v_1, \ldots, v_n$, such that conditions (3.4) and (3.6) hold, since the matrix M whose columns are the vectors $v_1, \ldots, v_n$ is then invertible and its inverse $N \equiv M^{-1}$ meets all the requirements. To see this, we observe that (3.4) guarantees the diagonalization of the quadratic form. Furthermore, since NM is the identity matrix, the first row of N is orthogonal to $v_2, \ldots, v_n$. This together with (3.6) implies that the first row of N is proportional to d, and hence (3.1) is satisfied.

Take $v_1 = C^{-1}d$. Then $Cv_1 = d$, so if $\{v_2, \ldots, v_n\}$ is any basis for $d^{\perp}$, then we immediately have (3.4) for all cases where $i = 1$ or $j = 1$. Recall that we are assuming $d^{\mathrm{T}}C^{-1}d \neq 0$. This means $v_1 \notin d^{\perp}$, so $\{v_1, \ldots, v_n\}$ is a linearly independent set. Thus, it only remains to show that a basis $\{v_2, \ldots, v_n\}$ may be chosen for $d^{\perp}$ so that (3.4) holds.

We shall proceed by induction. Let $1 \leqslant p < n$, and suppose that a linearly independent set $\{v_1, \ldots, v_p\}$ has been chosen such that $v_1 = C^{-1}d$ as before, (3.4) holds for $1 \leqslant i, j \leqslant p$ (vacuous if $p = 1$), and

$$v_j^{\mathrm{T}}Cv_j \neq 0 \quad \text{for } j = 1, \ldots, p, \tag{3.8}$$

(this is equivalent to our assumption $d^{\mathrm{T}}C^{-1}d \neq 0$ if $p = 1$). Write O_p for the orthogonal complement $\{Cv_1, \ldots, Cv_p\}^{\perp}$, and observe that O_p has dimension $n - p$ and is a subspace of $d^{\perp}$ (since $Cv_1 = d$).

We claim that $\operatorname{span}\{v_1, \ldots, v_p, O_p\} = \mathbb{R}^n$. If not, then there exists a nonzero vector v in O_p such that

$$v = \sum_{i=1}^{p} c_i v_i$$

for some constants c_i. Since $v \in O_p$, we have $v^{\mathrm{T}}Cv_j = 0$ for $j = 1, \ldots, p$, which together with (3.4) implies

$$0 = \left(\sum_{i=1}^{p} c_i v_i\right)^{\mathrm{T}} Cv_j = c_j v_j^{\mathrm{T}}Cv_j, \quad j = 1, \ldots, p.$$

From (3.8) it follows that $c_j = 0$ for each j. This contradicts the assumption that v is a nonzero vector, and establishes our claim.

We next claim that we can choose v_{p+1} in O_p such that $v_{p+1}^T C v_{p+1} \neq 0$. Observe that, once this is proved, our main result will have been established by induction; for, we will then have obtained (by the previous claim) linearly independent vectors $v_1, \ldots, v_{p+1}$ such that (3.4) holds for $i, j \leqslant p+1$ (since $v_{p+1} \in O_p$) and $v_2, \ldots, v_{p+1}$ all lie in $d^\perp$ (since $d = Cv_1$). If the last claim is false, then $v^T C v = 0$ for all $v \in O_p$. Take v and w in O_p. Then, $v^T C v = 0$, $w^T C w = 0$, and $(v+w)^T C (v+w) = 0$. Consequently, $w^T C v = 0$, using the symmetry of C. Let x be any vector in $\mathbb{R}^n$. By the previous claim, we can write $x = u + v$, with $u \in \text{span}\{v_1, \ldots, v_p\}$ and $v \in O_p$. For any $w \in O_p$, we then have $w^T C u = 0$, by the definition of O_p, and $w^T C v = 0$, from above. Hence, $w^T C x = 0$. Since x is arbitrary and C is invertible, it follows that $w = 0$. But, $p < n$ and therefore O_p contains a nonzero vector. This again gives a contradiction. Thus, v_{p+1} may be chosen as indicated. The induction continues, and the proof is complete. □

Remark. In our problem discussed in Section 2, the coefficients β_i in (3.2) are the eigenvalues μ_i of the Hessian matrix in (1.3), and C in (3.2) is the diagonal matrix $\text{diag}(\mu_1, \ldots, \mu_n)$; cf. (2.1) and (2.4). Since the choice of the vectors $v_1, \ldots, v_n$ in the above proof was made on the basis of their orientations and not on their magnitudes, these vectors may be rescaled, if necessary, so that $v_i^T C v_i = \mu_i$ for $i = 1, \ldots, n$; cf. (3.8). If this is done, then the matrix N satisfies

$$(N^{-1})^T C N^{-1} = \text{diag}(\mu_1, \ldots, \mu_n)$$

and hence

$$(\det N^{-1})^2 \det C = \mu_1 \cdots \mu_n = \det C.$$

By interchanging two columns in N^{-1} if necessary, we can also have

$$\det M = \det N^{-1} = 1. \tag{3.9}$$

This result has already been used in (2.13).

4. A version of Morse's lemma

Morse's lemma [6, p.54] states that in a neighborhood of a nondegenerate critical point of a C^∞-function $F(X_1, \ldots, X_n)$, there is a C^∞-change of variable $(X_1, \ldots, X_n) \to (Y_1, \ldots, Y_n)$ which transforms F into a (constant plus a) "sum of squares" in terms of the new variables. Without loss of generality, we may take the critical point to be the origin. For our purpose, we need slightly more; we need the transformation to preserve the half-space $X_1 > 0$, at least near the critical point 0. This could be achieved if we could set X_1 equal to Y_1 multiplied by a positive function. In this section, we prove two results which, when coupled together, show that such a requirement on X_1 can be included, provided that the quadratic terms of $F(X)$ already form a sum of squares.

We first introduce some notations. Recall that we are writing X, Y, etc., for points $(X_1, \ldots, X_n)$, $(Y_1, \ldots, Y_n)$, etc.. We shall write $D_X(i_1, \ldots, i_m)$ for the operation of mth-order partial differentiation with respect to variables $X_{i_1}, \ldots, X_{i_m}$; and a similar convention applies to the symbol $D_Y(p_1, \ldots, p_r)$. Since all functions considered in this section will be infinitely differentiable, these operations are independent of the order of the indices $i_1, \ldots, i_m$ (or $p_1, \ldots, p_r$).

Theorem 2. *Assume* (i) $F(X)$ *is a real-valued* C^∞*-function in a neighborhood of the origin, and* $F(X)$ *has a critical point at* 0; (ii) $(D_X(i, j)F)(0)$ *is zero if* $i \neq j$, *and nonzero if* $i = j$. *Then there exists a transformation of the form*

$$\begin{aligned}
X_1 &= Y_1,\\
X_2 &= Y_2,\\
X_3 &= Y_3 + c^{(3)}(1,2)Y_1Y_2,\\
&\vdots\\
X_k &= Y_k + \sum\left\{c^{(k)}(l_1, l_2)Y_{l_1}Y_{l_2}\colon l_1 < l_2 < k\right\}\\
&\quad + \sum\left\{c^{(k)}(l_1, l_2, l_3)Y_{l_1}Y_{l_2}Y_{l_3}\colon l_1 < l_2 < l_3 < k\right\}\\
&\quad + \cdots + c^{(k)}(1, 2, \ldots, k-1)Y_1 \cdots Y_{k-1}, \qquad (4.1)\\
&\vdots\\
X_n &= Y_n + \sum\left\{c^{(n)}(l_1, l_2)Y_{l_1}Y_{l_2}\colon l_1 < l_2 < n\right\}\\
&\quad + \cdots + c^{(n)}(1, 2, \ldots, n-1)Y_1 \cdots Y_{n-1},
\end{aligned}$$

such that $\Phi(Y) = F(X)$ *has a critical point at* $Y = 0$, *and also satisfies*

$$(D_Y(p_1, \ldots, p_r)\Phi)(0) = 0, \quad 2 \leqslant r \leqslant n, \qquad (4.2)$$

whenever $p_i \neq p_j$ *if* $i \neq j$. *The coefficients* $c^{(k)}(l_1, \ldots, l_m)$ $(2 \leqslant m \leqslant k-1,\ 3 \leqslant k \leqslant n)$ *are constants.*

Proof. Since the functions are infinitely differentiable, we can suppose that $p_1 < p_2 < \cdots < p_r$. We also set the following values from the beginning:

$$c^{(k)}(l) = (D_Y(l)X_k)(0) = \delta_{kl} \quad \text{(Kronecker delta)} \qquad (4.3)$$

and

$$\begin{aligned}
&c^{(k)}(l_1, \ldots, l_m) = (D_Y(l_1, \ldots, l_m)X_k)(0) = 0\\
&\quad \text{if } l_m \geqslant k \quad \text{or} \quad \text{if } l_i = l_j \text{ for any } i \neq j.
\end{aligned} \qquad (4.4)$$

Note that, if $l_i \neq l_j$ for $i \neq j$, then $c^{(k)}(l_1, \ldots, l_m) = (D_Y(l_1, \ldots, l_m)X_k)(0)$. Also note that by (4.3),

$$\left.\frac{\partial(X_1, \ldots, X_n)}{\partial(Y_1, \ldots, Y_n)}\right|_{Y=0} = 1 \qquad (4.5)$$

and hence the transformation is invertible near 0.

By the chain rule,

$$D_Y(p)\Phi = \sum_i (D_X(i)F)(D_Y(p)X_i). \qquad (4.6)$$

Since 0 is a critical point of F, (4.6) shows that 0 is also a critical point of Φ. Differentiation of (4.6) gives

$$D_Y(p, q)\Phi = \sum_{i,j} (D_X(i, j)F)(D_Y(p)X_i)(D_Y(q)X_j) + \sum_i (D_X(i)F)(D_Y(p, q)X_i). \qquad (4.7)$$

Evaluating at 0, and using (i) and (4.3), we see that $(D_Y(p, q)\Phi)(0) = (D_X(p, q)F)(0)$. By (ii), the case $r = 2$ in (4.2) is proved.

The coefficients $c^{(k)}(l_1, \ldots, l_m)$, for $3 \leqslant k \leqslant n$ and $l_1 < \cdots < l_m < k$, will be determined by induction. To see the idea of the following argument, the reader is urged to consider separately the cases $n = 3$ and $n = 4$. (There is nothing to prove for $n = 2$.) Also, it may be helpful if we first determine the coefficient $c^{(3)}(1, 2)$ by using the requirement $(D_Y(1, 2, 3)\Phi)(0) = 0$. In (4.7), we set $p = 1$ and $q = 2$, and differentiate with respect to Y_3. This yields

$$\begin{aligned} D_Y(1, 2, 3)\Phi &= \sum_{i,j,k} (D_X(i, j, k)F)(D_Y(1)X_i)(D_Y(2)X_j)(D_Y(3)X_k) \\ &\quad + \sum_{i,j} (D_X(i, j)F)\big[(D_Y(1, 3)X_i)(D_Y(2)X_j) + (D_Y(2, 3)X_j)(D_Y(1)X_i) \\ &\qquad + (D_Y(1, 2)X_i)(D_Y(3)X_j)\big] \\ &\quad + \sum_i (D_X(i)F)(D_Y(1, 2, 3)X_i). \end{aligned} \tag{4.8}$$

Evaluating at 0, and using (i), (ii), (4.3) and (4.4), we get

$$(D_Y(1, 2, 3)\Phi)(0) = (D_X(1, 2, 3)F)(0) + (D_Y(3, 3)F)(0)(D_Y(1, 2)X_3)(0). \tag{4.9}$$

Since $(D_X(3, 3)F)(0) \neq 0$ by (ii), (4.9) determines $c^{(3)}(1, 2) = (D_Y(1, 2)X_3)(0)$.

We now make our inductive hypotheses.

(H_1) *Suppose* $2 \leqslant r < n$, *and for any* $s < r$, *coefficients* $c^{(i)}(l_1, \ldots, l_s)$ *have been determined for any* $i = 1, \ldots, n$ *and* $l_1 < \cdots < l_s < i$, *so that* $(D_Y(l_1, \ldots, l_s, i)\Phi)(0) = 0$.

(Note that, with $r = 2$, (H_1) is satisfied by (4.3); see the remark following (4.7).) With r such that (H_1) holds, we make a further assumption.

(H_2) *Suppose* $2 \leqslant r < k \leqslant n$, *and for any* $i < k$, *the coefficients* $c^{(i)}(l_1, \ldots, l_r)$ *have been determined for any choice of* $l_1 < \cdots < l_r < i$, *so that* $(D_Y(l_1, \ldots, l_r, i)\Phi)(0) = 0$.

(Note that if $i < k \leqslant r$, then $c^{(i)}(l_1, \ldots, l_r) = 0$ by (4.4).) In what follows, it will be shown that for any choice $l_1 < \cdots < l_r < k$, the condition $(D_Y(l_1, \ldots, l_r, k)\Phi)(0) = 0$, together with the values of the coefficients already found, uniquely determines $c^{(k)}(l_1, \ldots, l_r)$. Then, by induction on k based on (H_2), the coefficients $c^{(i)}(l_1, \ldots, l_r)$ can be determined for all $i = 1, \ldots, n$ and $l_1 < \cdots < l_r < i$. This allows induction on r, based on (H_1), to continue, and proves the result.

The expression for $D_Y(l_1, \ldots, l_r, k)\Phi$ can be arranged as a sum of $r + 1$ sums $\Sigma_1, \Sigma_2, \ldots, \Sigma_{r+1}$; cf. (4.7) and (4.8). Σ_1 is a sum of terms of the form $(D_X(i)F)(D_Y(l_1, \ldots, l_r, k)X_i)$, and vanishes at 0 by (i). Σ_2 is more complicated. It is a sum of terms of the form $(D_X(i, j)F)(D_Y(S)X_i)(D_Y(T)X_j)$, where S and T are disjoint, nonempty sets such that $S \cup T = \{l_1, \ldots, l_r, k\}$. If both S and T have fewer than r members, then the value of this term is already determined by (H_1) and (4.4). Thus, we can suppose that S has r members and T has just one, say $T = \{l\}$, where $l \in \{l_1, \ldots, l_r, k\}$. Now we evaluate the term $(D_X(i, j)F)(D_Y(S)X_i)(D_Y(l)X_j)$ at 0. By (ii), $(D_X(i, j)F)(0) = 0$ if $i \neq j$, and by (4.3), $(D_Y(l)X_j)(0) = \delta_{j,l}$; hence the term is zero unless $i = j = l$. Also, by (4.4), $(D_Y(S)X_i)(0) = 0$ if S contains a member greater than i. Putting all these together with $l_1 < \cdots < l_r < k$, we see that the only term in Σ_2 which is not already determined is $(D_X(k, k)F)(0)(D_Y(l_1, \ldots, l_r)X_k)(0)$. Since $(D_X(k, k)F)(0) \neq 0$ by (ii), it follows that $(D_Y(l_1, \ldots, l_r, k)\Phi)(0) = 0$ will determine

$c^{(k)}(l_1,\dots,l_r) = (D_Y(l_1,\dots,l_r)X_k)(0)$, as long as the values $\Sigma_3,\dots,\Sigma_{r+1}$ are known. But Σ_m is a sum of terms of the form $(D_X(i_1,\dots,i_m)F)(D_Y(S_1)X_{i_1})\cdots(D_Y(S_m)X_{i_m})$, where $S_1,\dots,S_m$ are pairwise disjoint, nonempty sets whose union is $\{l_1,\dots,l_r,\ k\}$. If $m \geqslant 3$, then each S_j must have fewer than r members; thus, by (H_1), the value of such a term at $Y=0$ is already determined. This completes the proof of the theorem. □

Theorem 3. *Suppose that $\Phi(Y)$ $(Y=(Y_1,\dots,Y_n))$ is a C^∞-function in a neighborhood of the origin, and that $\Phi(0) = (D_Y(l)\Phi)(0) = (D_Y(l_1,\dots,l_r)\Phi)(0) = 0$ whenever $l_i \neq l_j$ for $i \neq j$. Then there are functions $\Phi_1,\dots,\Phi_n$, each C^∞ near 0, such that*

$$\Phi(Y) = \sum_l Y_l^2 \Phi_l(Y).$$

Proof. We first consider the case $n=1$. Here the assumption is just $\Phi(0) = \Phi'(0) = 0$. By l'Hôpital's rule, $\Phi_1(u) \equiv \Phi(u)/u^2$ extends to a C^∞-function near 0.

Now let $N > 1$, and suppose that the result holds for $k < N$. Assume that $\Phi(Y_1,\dots,Y_N)$ satisfies the hypotheses of the theorem. By induction hypothesis, there are C^∞-functions $G_l(Y_1,\dots,Y_{N-1})$ and $H_l(Y_1,\dots,Y_{N-1})$ in a neighborhood of the origin $(l=1,\dots,N-1)$ such that

$$\Phi(Y_1,\dots,Y_{n-1},0) = \sum_{l=1}^{N-1} Y_l^2 G_l(Y_1,\dots,Y_{N-1}),$$

$$(D_Y(N)\Phi)(Y_1,\dots,Y_{N-1},0) = \sum_{l=1}^{N-1} Y_l^2 H_l(Y_1,\dots,Y_{N-1}).$$

Therefore,

$$\Phi(Y_1,\dots,Y_{N-1},0) + Y_N(D_Y(N)\Phi)(Y_1,\dots,Y_{N-1},0) = \sum_{l=1}^{N-1} Y_l^2 \Phi_l(Y),$$

where $\Phi_l(Y) = G_l(Y_1,\dots,Y_{N-1}) + Y_N H_l(Y_1,\dots,Y_{N-1})$ for $l=1,\dots,N-1$. Since it is easily verified that the function $\Psi(Y) \equiv \Phi(Y) - \Phi(Y_1,\dots,Y_{N-1},0) - Y_N(D_Y(N)\Phi)(Y_1,\dots,Y_{N-1},0)$ satisfies $\Psi(0) = (D_Y(N)\Psi)(0) = 0$, we have, as in the one-dimensional case, $\Psi(Y) = Y_N^2\Phi_N(Y)$, with $\Phi_N(Y)$ being C^∞ near 0. The theorem now follows. □

Remark. If $\Phi(Y) = \Sigma Y_l^2 \Phi_l(Y)$, then $(D_Y(p,q)\Phi)(0) = 0$ if $p \neq q$, and $(D_Y(p,p)\Phi)(0) = 2\Phi_p(0)$. So, if 0 is a nondegenerate critical point of Φ, then we may assume, by restricting to a sufficiently small neighborhood of 0, that each $\Phi_l(y)$ is nonvanishing. Consequently, we can write $\Phi_l(Y) = \Phi_l(0)[1+\Psi_l(Y)]$, with $\Psi_l(0) = 0$ and $1+\Psi_l(Y)$ positive near 0. The equations

$$t_l = Y_l\big[1+\Psi_l(Y)\big]^{1/2}, \quad l=1,\dots,n,$$

then define a transformation which is C^∞ near 0, and represents Φ as a "sum of squares": $\Phi(Y) = \Sigma\Phi_l(0)t_l^2$. Thus, Theorems 2 and 3 together imply the existence of a C^∞-transformation which represents the original function $F(X)$ as (a constant plus) a sum of squares, as does Morse's lemma. Since $X_1 = Y_1$ and $Y_1 > 0$ if and only if $t_1 > 0$, this transformation also maps the half-space $X_1 > 0$ onto $t_1 > 0$, at least near 0. However, note that we do *not* claim that *any* $F(X)$

with a nondegenerate critical point can be transformed to a sum of squares while preserving a specified half-space. Theorem 2 has the extra assumption that the quadratic terms in the Taylor expansion of $F(X)$ were already in the form of a sum of squares. In general, the linear transformation needed to achieve this might not preserve any coordinate half-space in terms of the original variables.

5. Derivation of the asymptotic expansion

We begin with equation (2.9):

$$I(\lambda) = \int G(X)\, e^{i\lambda F(X)}\, dX, \tag{5.1}$$

where the domain of integration is the half-space $X_1 > 0$ and

$$F(X) = f(x_0) + \tfrac{1}{2}\sum_{j=1}^{n} \mu_j X_j^2 + \cdots. \tag{5.2}$$

We know that by Theorem 1, (5.1) and (5.2) can be achieved if (2.5) holds. By using a partition of unity, the amplitude function $g(x)$ in (1.1) is often assumed to vanish off an arbitrarily small neighborhood of the stationary point x_0; cf. [7, p.429]. Thus we may suppose that the amplitude function $G(X)$ also vanishes off an arbitrarily small neighborhood of the origin. Next we observe that by Section 4, there is a C^∞-transformation $X \to Y$ such that the expansion (5.2) can be written in the form

$$F(X) = f(x_0) + \tfrac{1}{2}\mu_1 Y_1^2\left[1 + \Psi_1(Y)\right] + \cdots + \tfrac{1}{2}\mu_n Y_n^2\left[1 + \Psi_n(Y)\right],$$

where each $\Psi_i(Y)$ is infinitely differentiable in Y with $\Psi_i(0) = 0$. Thus, making the change of variables

$$t_i = Y_i\left[1 + \Psi_i(Y)\right]^{1/2}, \quad i = 1, \ldots, n, \tag{5.3}$$

the integral in (5.1) becomes

$$I(\lambda) = e^{i\lambda f(x_0)} \int \phi(t) \exp\left\{ i\tfrac{1}{2}\lambda \sum_{j=1}^{n} \mu_j t_j^2 \right\} dt, \tag{5.4}$$

where $\phi(t)$ is the product of G and the absolute value of the Jacobian of the transformation involved. Since the domain of integration in (5.1) is only a small neighborhood of the origin in the half-space $X_1 > 0$, the integration in (5.4) can be assumed to be over a similar neighborhood in the half-space $t_1 > 0$. From (5.3), it is easily seen that

$$\left.\frac{\partial(Y_1, \ldots, Y_n)}{\partial(t_1, \ldots, t_n)}\right|_{t=0} = 1. \tag{5.5}$$

Hence we may also assume that the Jacobian $\partial(Y_1, \ldots, Y_n)/\partial(t_1, \ldots, t_n)$ is positive in the domain of integration. Consequently, $\phi(t)$ is a C^∞-function. Combining (5.5) with (2.8), (2.13) and (4.5), we also have

$$\phi(0) = g(x_0). \tag{5.6}$$

Put $\psi_\lambda(t) = \exp\{i\frac{1}{2}\lambda(\Sigma\mu_j t_j^2)\}$. The integral in (5.4) becomes

$$\int_{t_1>0} \phi(t) \exp\left\{i\tfrac{1}{2}\lambda \sum_{j=1}^{n} \mu_j t_j^2\right\} dt = \int_{t_1>0} \phi(t)\psi_\lambda(t)\, dt. \tag{5.7}$$

If $\hat{\phi}$ denotes the Fourier transform of $\phi(t)$ then by inversion

$$\phi(t) = \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} \hat{\phi}(\eta)\, e^{-it\cdot\eta}\, d\eta. \tag{5.8}$$

Since ϕ is a C^∞-function with compact support, $\hat{\phi}$ is a rapidly decreasing function, i.e., it decreases faster than any polynomial in η. Inserting (5.8) in (5.7) and interchanging the order of integration, we obtain

$$\int_{t_1>0} \phi(t) \exp\left\{i\tfrac{1}{2}\lambda \sum_{j=1}^{n} \mu_j t_j^2\right\} dt = \int_{\mathbb{R}^n} \hat{\phi}(\eta)\Psi_\lambda(\eta)\, d\eta, \tag{5.9}$$

where

$$\Psi_\lambda(\eta) = \frac{1}{(2\pi)^{n/2}} \int_{t_1>0} \psi_\lambda(t)\, e^{-i\eta\cdot t}\, dt. \tag{5.10}$$

To evaluate $\Psi_\lambda(\eta)$ asymptotically, we first observe that this function can be written as a product of n one-dimensional Fourier integrals. For $j = 2, \ldots, n$, the integral with respect to dt_j can be evaluated in a closed form, and the result gives

$$\begin{aligned} &\frac{1}{(2\pi)^{(n-1)/2}} \int_{\mathbb{R}^{n-1}} e^{-i\eta\cdot t} \exp\left\{i\tfrac{1}{2}\lambda \sum_{j=2}^{n} \mu_j t_j^2\right\} dt_2 \cdots dt_n \\ &= \frac{e^{-i\eta_1 t_1}}{\lambda^{(n-1)/2}} \left|\prod_{j=2}^{n} \mu_j\right|^{-1/2} \exp\left\{i\tfrac{1}{4}\pi \sum_{j=2}^{n} \operatorname{sgn} \mu_j - \frac{i}{2\lambda} \sum_{j=2}^{n} \frac{1}{\mu_j}\eta_j^2\right\}. \end{aligned} \tag{5.11}$$

We are thus left with the integral

$$\frac{1}{(2\pi)^{1/2}} \int_0^\infty \exp\{-i\eta_1 t_1 + \tfrac{1}{2}i\lambda\mu_1 t_1^2\}\, dt_1, \tag{5.12}$$

which, upon making the substitution $\tau = t_1^2$, can be viewed as the Fourier transform of

$$\tfrac{1}{2}\tau^{-1/2} \exp\{-i\eta_1\tau^{1/2}\}$$

on the interval $(0, \infty)$ with the variable $\frac{1}{2}\lambda\mu_1$. By applying the asymptotic theory given in [5], complete with error analysis, to this transform, it can be shown that the integral in (5.12) has the asymptotic expansion

$$\frac{e^{i(\pi/4)\operatorname{sgn}\mu_1}}{(\lambda|\mu_1|)^{1/2}} \left[\frac{1}{2} - \frac{i\, e^{i(\pi/4)\operatorname{sgn}\mu_1}\eta_1}{\sqrt{2\pi|\mu_1|}} \frac{1}{\lambda^{1/2}} - \frac{e^{i(\pi/2)\operatorname{sgn}\mu_1}\eta_1^2}{2^2|\mu_1|} \frac{1}{\lambda} + \frac{i\, e^{i(3\pi/4)\operatorname{sgn}\mu_1}\eta_1^3}{3\sqrt{2\pi}\,|\mu_1|^{3/2}} \frac{1}{\lambda^{3/2}} + \cdots\right] \tag{5.13}$$

as $\lambda \to \infty$. The pth remainder associated with this expansion is bounded by $M_p(\eta_1)\lambda^{-p/2}$, and $M_p(\eta_1)$ is a polynomial in η_1 of degree p. Combining this result with (5.11) gives

$$\Psi_\lambda(\eta) \sim \frac{1}{\lambda^{n/2}} \left| \prod_{j=1}^{n} \mu_j \right|^{-1/2} \exp\left\{ i\tfrac{1}{4}\pi \sum_{j=1}^{n} \operatorname{sgn} \mu_j \right\} \sum_{s=0}^{\infty} p_s(\eta)\lambda^{-s/2}, \tag{5.14}$$

where $p_s(\eta)$ is a homogeneous polynomial in η of degree s and the remainder is bounded by $\lambda^{-p/2}$ multiplied by a polynomial in η. The first few coefficients are given by

$$\begin{aligned} p_0(\eta) &= \tfrac{1}{2}, \qquad p_1(\eta) = -\frac{i\eta_1}{\sqrt{2\pi\,|\mu_1|}} e^{i(\pi/4)\operatorname{sgn}\mu_1}, \\ p_2(\eta) &= -\tfrac{1}{2}i\left(\sum_{j=2}^{n} \mu_j^{-1}\eta_j^2 \right) - \frac{e^{i(\pi/2)\operatorname{sgn}\mu_1}\eta_1^2}{2^2\,|\mu_1|}, \\ p_3(\eta) &= i\frac{e^{i(3\pi/4)\operatorname{sgn}\mu_1}\eta_1^3}{3\sqrt{2\pi}\,|\mu_1|^{3/2}} - \frac{e^{i(\pi/4)\operatorname{sgn}\mu_1}\eta_1}{2\sqrt{2\pi\,|\mu_1|}} \left(\sum_{j=2}^{n} \mu_j^{-1}\eta_j^2 \right). \end{aligned} \tag{5.15}$$

Inserting (5.14) in (5.9) and integrating term by term, we obtain

$$\begin{aligned} &\int_{t_1>0} \phi(t) \exp\left\{ i\tfrac{1}{2}\lambda \sum_{j=1}^{n} \mu_j t_j^2 \right\} dt \\ &\quad \sim \frac{1}{\lambda^{n/2}} \left| \prod_{j=1}^{n} \mu_j \right|^{-1/2} \exp\left\{ i\tfrac{1}{4}\pi \sum_{j=1}^{n} \operatorname{sgn} \mu_j \right\} \sum_{s=0}^{\infty} d_s \lambda^{-s/2}, \end{aligned} \tag{5.16}$$

where

$$d_s = \int_{\mathbb{R}_n} \hat{\phi}(\eta) p_s(\eta)\, d\eta.$$

Recall that $\det A = \mu_1 \cdots \mu_n$ and $\sigma = \operatorname{sgn} \mu_1 + \cdots + \operatorname{sgn} \mu_n$, and observe that d_s can be viewed as the inverse Fourier transform of $(2\pi)^{n/2}[p_s(-D)\phi]\hat{}$ evaluated at $t = 0$, where $p_s(D)$ is the differential operator obtained from the polynomial $p_s(\eta)$ by replacing $\eta = (\eta_1, \ldots, \eta_n)$ by

$$D = \left(i\frac{\partial}{\partial x_1}, \ldots, i\frac{\partial}{\partial x_n} \right).$$

Coupling (5.4) and (5.16) yields

$$I(\lambda) \sim \left(\frac{2\pi}{\lambda} \right)^{n/2} |\det A|^{-1/2} \exp\{ i\lambda f(x_0) + i\sigma\tfrac{1}{4}\pi \} \sum_{s=0}^{\infty} p_s(-D)\phi(0)\lambda^{-s/2}. \tag{5.17}$$

In view of (5.6) and the first equation in (5.15), the desired approximation (2.14) now follows by taking the leading term in (5.17).

Acknowledgements

We are grateful to the referees for a careful reading of the earlier version of this paper, and for suggestions which led to the improvement of the presentation.

References

[1] N. Chako, Asymptotic expansions of double and multiple integrals arising in diffraction theory, *J. Inst. Math. Appl.* **1** (1965) 372–422.
[2] M.V. Fedoryuk, The stationary phase method and pseudo-differential operators, *Russian Math. Surveys* **26** (1971) 65–115.
[3] L. Hörmander, *The Analysis of Linear Partial Differential Operators I* (Springer, Berlin, 1983).
[4] D.S. Jones, *The Theory of Generalized Functions* (Cambridge Univ. Press, Cambridge, 1982).
[5] F.W.J. Olver, Error bounds for stationary phase approximations, *SIAM J. Math. Anal.* **5** (1974) 19–29.
[6] T. Poston and I. Stewart, *Catastrophe Theory and Its Applications* (Pitman, Boston, MA, 1978).
[7] F. Treves, *Introduction to Pseudodifferential and Fourier Integral Operators, Vol. 2* (Plenum, New York, 1980).
[8] E. Wolf, The diffraction theory of aberrations, *Rep. Progr. Phys.* **XIV** (1951) 95–120.