A Hybrid Numerical Asymptotic Method for Scattering Problems

Eldar Giladi* and Joseph B. Keller†

*2124 Rock Street, #3, Mountain View, California 94043; †Departments of Mathematics and Mechanical Engineering, Stanford University, Stanford, California 94305 E-mail: egiladi@hotmail.com

Received November 13, 2000; revised August 6, 2001

We develop a hybrid numerical asymptotic method for the Helmholtz equation. The method is a Galerkin finite element method in which the space of trial solutions is spanned by asymptotically derived basis functions. The basis functions are very "efficient" in representing the solution because each is the product of a smooth amplitude and an oscillatory phase factor, like the asymptotic solution. The phase is determined a priori by solving the eiconal equation using the ray method, while the smooth amplitude is represented by piecewise polynomials. The number of unknowns necessary to achieve a given accuracy with this new basis is dramatically smaller than the number necessary with a standard method, and it is virtually independent of the wavenumber k. We apply the method to the problems of scattering from a parabola and from a circle and compare the results with those of a standard finite element method. (© 2001 Elsevier Science)

Key Words: Helmholtz equation; wave propagation; WKB method; geometrical optics; geometrical theory of diffraction; asymptotic methods; Galerkin method; finite element method.

1. INTRODUCTION

Scattering problems for the Helmholtz equation involving large values of the wavenumber k are notoriously difficult to solve numerically by standard finite element methods. See for example [1–3]. Accurate solutions require very refined meshes due to the highly oscillatory nature of the solutions. If k is large enough, the number of unknowns associated with the mesh is so large that the problem is computationally intractable. But for such k, asymptotic solutions of scattering problems, valid for k large, are provided by the geometrical theory of diffraction [4–6]. These asymptotic solutions are sums of products of rapidly oscillating phase factors multiplied by slowly varying amplitude factors. The phase factors can be determined easily by ray tracing. However, the determination of the amplitude factors is



0021-9991/01 \$35.00 © 2001 Elsevier Science All rights reserved often difficult because it can involve the construction and matching of several different expansions. To overcome the difficulty of finding the amplitude factors in the asymptotic method, and to avoid the necessity for a fine mesh in the finite element method, we combine the best features of both methods.

To do so we develop a hybrid numerical asymptotic method which yields accurate solutions with a number of unknowns that is virtually independent of the wavenumber k. Our method is based on the introduction of asymptotically derived basis functions, which we use to define the space of trial solutions in a Galerkin finite element method. The basis functions are very "efficient" in representing the solution because each is the product of a smooth amplitude function and an oscillatory phase factor, like the terms in the asymptotic solutions. The phase factor is determined a priori by solving the eiconal equation of geometrical optics for the phase, using ray tracing. The smooth amplitude, which is represented by a standard finite element basis, is determined by the Galerkin method.

The resolution of the computational grid is commensurate with the variation of the smooth amplitude, and therefore it is relatively coarse, and in general virtually independent of the wavenumber k. In some scattering problems, the amplitude develops a sharp transition in a small portion of the domain, such as a shadow boundary or caustic. In these cases, only local refinement is necessary, and the problem remains computationally tractable.

This method extends our results in [7] and [8]. In one dimension it is similar in spirit to the method of Melenk and Babuška [9, 10], and Babuška and Sauter [1], which employs basis functions involving trigonometric functions. Laghrouche and Bettess [11], and Laghrouche *et al.* [12] extended that method to two dimensions using basis functions involving plane waves. A different hybrid of asymptotic and finite elements, closer to ours, was introduced by Barbone *et al.* [13]. It employs a field constructed by using the geometrical theory of diffraction [4] in most of the domain, together with fields represented by finite element basis functions in neighborhoods of points of diffraction, such as corners on the scattering object.

Although our method can be used to treat a wide variety of wave propagation problems, here we just introduce it and apply it to a few examples of scattering involving reflection and diffraction from smooth surfaces. In Section 2 we introduce the method in the context of a one-dimensional problem. We extend it to two space dimensions in Section 3. In Section 3.2.1 we review the elements of the asymptotic theory needed to develop the asymptotically derived basis functions, and in Section 3.2.2 we introduce these new basis functions. Then in Section 3.3, we apply the method to scattering from a parabola, and in Section 3.4 we apply it to scattering from a circle. In section 3.5 we summarize the method and describe how to use it in three dimensions. In the Appendix we show that the new finite elements are asymptotically orthogonal for k large.

2. A MODEL PROBLEM

2.1. Variational Formulation and Galerkin Approximation

We consider the propagation of a time harmonic wave traveling to the right along the positive x axis. Its motion is described by the one-dimensional Helmholtz equation

$$u'' + k^2 n^2(x)u = 0, (1)$$

subject to the conditions that *u* is outgoing at $x = +\infty$, and that

$$u(0) = 1.$$
 (2)

Here k is the wave number and n(x) is the index of refraction, assumed to be positive. We assume that n(x) - 1 is negligible for $x \ge \pi$ and replace the outgoing condition by

$$u'(\pi) - iku(\pi) = 0.$$
 (3)

In order to state the variational formulation of problem (1)–(3) for $0 \le x \le \pi$, we introduce the space S of trial solutions and the space V of test functions:

$$S = \{ u \mid u \in H^1([0, \pi]), u(0) = 1 \},$$
(4)

$$\mathcal{V} = \{ w \mid w \in H^1([0,\pi]), w(0) = 0 \}.$$
(5)

The Sobolev space $H^1([0, \pi])$ consists of all complex-valued functions on [0, 1] whose first derivative is square integrable. The variational formulation of problem (1)–(3) is:

find
$$u \in S$$
 such that for all $w \in V$, $a(u, w) = 0$. (6)

The hermitian bilinear form a in (6) is defined by

$$a(u,w) = -\int_0^{\pi} u'(t)\bar{w}'(t)\,dt + k^2 \int_0^{\pi} n^2(t)u(t)\bar{w}(t)\,dt + iku(\pi)\bar{w}(\pi). \tag{7}$$

The formulation (6) is derived in standard fashion (see [14, 15]) by multiplying (1) by the complex conjugate \bar{w} of a test function $w \in \mathcal{V}$, integrating by parts, and using (3).

We shall approximate the solution to this variational problem by the Galerkin method. Hence we introduce finite dimensional spaces $S_h \subset S$ and $V_h \subset V$. The subscript *h* in S_h and V_h refers to the relation of these spaces to a discretization of the domain $[0, \pi]$. Indeed, in the finite element method, we consider the mesh $0 = x_0 < x_1 < \cdots < x_m = \pi$ with mesh parameter $h = \max_{j=1,\dots,m} (x_j - x_{j-1})$. The functions in V_h consist of piecewise polynomials. Specifically, in the linear finite element method there is one basis function N_j associated with each node *j*. The function N_j is 1 at node *j*, 0 at all other nodes, and linear on every interval $[x_j, x_{j+1}], j = 1, \dots, m-1$. The spaces V_h and S_h are defined in terms of these basis functions as

$$\mathcal{V}_h = \left\{ w_h \, \middle| \, w_h = \sum_{j=1}^m a_j N_j \right\},\tag{8}$$

$$S_h = \{ u_h = v_h + g_h \mid v_h \in \mathcal{V}_h, g_h = N_0(x) \}.$$
(9)

The original variational problem is then replaced by the Galerkin problem:

find
$$u_h = v_h + g_h \in S_h$$
 such that $\forall w_h \in V_h$, $a(v_h, w_h) = -a(g_h, w_h)$. (10)

The Galerkin finite element theory [14–16] ensures that the solution u_h of the Galerkin problem (10) converges in the Sobolev norm to the solution u of the variational problem (6) as $h \rightarrow 0$. However, u is highly oscillatory for large values of k, and an accurate approximation

 u_h is obtained only for very small values of h. Hence, the number of unknowns in the Galerkin method grows rapidly as $k \to \infty$, and the problem is eventually computationally intractable. Our goal is to introduce a more efficient solution space S_h which gives accurate approximations with relatively large values of h, even for very large values of k. Towards this goal, we first study the asymptotic solution to the original problem (1)–(3).

2.2. Asymptotic Finite Elements

2.2.1. Elements of the Asymptotic Theory

We shall now determine the asymptotic form of u(x) in (1) by using the W.K.B. method [17]. Hence we seek a solution of the form

$$u(x) = K(x) \exp(ikS(x)), \tag{11}$$

where K(x) is the amplitude, S(x) is the phase, and $\exp(ikS(x))$ is the phase factor. Both K and S are slowly varying, so they can be resolved numerically on a grid which is much coarser than that required for u.

To get an equation for S we introduce (11) into (1) and divide the resulting expression by $k^2 e^{ikS(x)}$ to obtain

$$K(n^2 - S_x^2) + i\frac{1}{k}(2K_xS_x + KS_{xx}) + \frac{1}{k^2}K_{xx} = 0.$$
 (12)

Upon equating to zero the coefficient of k^0 , the leading order term in (12), we obtain the eiconal equation of geometrical optics

$$S_x^2(x) = n^2(x).$$
 (13)

The solution of (13) for which (11) satisfies (3) to leading order in k is $S(x) = \int_0^x n(s) ds$. We substitute this expression for S(x) into (11) and (12). From (12) we get

$$\frac{1}{k}K_{xx} + i(2nK_x + n_xK) = 0.$$
(14)

Then we use (11) in (2) and (3) to obtain

$$K(0) = 1, \quad K_x(\pi) = 0.$$
 (15)

Now, we seek an expansion for K of the form

$$K \sim \sum_{j=0}^{\infty} K_j (ik)^{-j}.$$
 (16)

To determine the coefficients K_j , we substitute (16) for K into (14), gather coefficients of equal powers of k, and then equate the coefficient of each power of k to zero. This yields a system of equations for K_j , j = 0, 1, 2... The equation for K_0 is

$$K_{0,x} + \frac{n_x}{2n} K_0 = 0. (17)$$

The solution of (17) which satisfies (15) is $\sqrt{n(0)/n(x)}$ provided that $n_x(\pi) = 0$, and then (16) yields

$$K = \sqrt{\frac{n(0)}{n(x)}} + O(1/k).$$
 (18)

2.2.2. Asymptotically Derived Basis Functions

From (18) we see that for k large, K is a smooth function of x. Hence K can be represented accurately by a piecewise linear finite element basis on a coarse mesh. Moreover, the degree of accuracy is retained as $k \to \infty$, and no mesh refinement is necessary. This leads us to introduce the asymptotic finite element basis functions

$$M_j(x,k) = N_j(x) \exp\left(ik \int_0^x n(s) \, ds\right),\tag{19}$$

where $N_j(x)$ is the piecewise linear basis function associated with node *j*. We also define new spaces of test functions and trial solutions for the Galerkin problem (10):

$$\mathcal{V}_h^k = \left\{ w_h^k \, \middle| \, w_h^k = \sum_{j=1}^m a_j M_j \right\},\tag{20}$$

$$S_{h}^{k} = \left\{ u_{h}^{k} = v_{h}^{k} + g_{h}^{k} \, \middle| \, v_{h}^{k} \in \mathcal{V}_{h}^{k}, \, g_{h}^{k} = M_{0}(x, k) \right\}.$$
(21)

The new Galerkin problem is

find
$$u_h^k = v_h^k + g_h^k \in \mathcal{S}_h^k$$
 such that $\forall w_h^k \in \mathcal{V}_h^k$, $a(v_h^k, w_h^k) = -a(g_h^k, w_h^k)$, (22)

with a defined in (7). The solution,

$$u_h^k = \sum_{j=1}^m a_j M_j + M_0, \tag{23}$$

to problem (22) is obtained in a standard fashion (see [14]) by solving the system

$$\begin{pmatrix} a(M_1, M_1) & a(M_2, M_1) & \dots & a(M_m, M_1) \\ a(M_1, M_2) & a(M_2, M_2) & \dots & a(M_m, M_2) \\ \dots & \dots & \dots & \dots \\ a(M_1, M_m) & a(M_2, M_m) & \dots & a(M_m, M_m) \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \dots \\ a_m \end{pmatrix} = \begin{pmatrix} -a(M_0, M_1) \\ 0 \\ \dots & \dots \\ 0 \end{pmatrix}. \quad (24)$$

The functions $u \in S_h^k$ and $w \in \mathcal{V}_h^k$ have the forms

$$u(x) = p(x) \exp\left(ik \int_0^{\pi} n(t) dt\right), \quad w = q(x) \exp\left(ik \int_0^{\pi} n(t) dt\right), \quad (25)$$

with $p(x) \in S$ and $q(x) \in V$. We substitute (25) for u and w into (7) and use the identity

$$-p(x)n(x)\bar{q}'(x) = -(p(x)\bar{q}(x)n(x))' + p(x)'\bar{q}(x)n(x) + p(x)\bar{q}(x)n'(x)$$
(26)

in the resulting expression to obtain the bilinear form \tilde{a} :

$$\tilde{a}(p,q) = \int_0^{\pi} \left[-p'\bar{q}' + ik(2p'\bar{q}n + p\bar{q}n') \right] dx.$$
(27)

Then we note that the variational problem,

find
$$p \in S$$
 such that for all $q \in \mathcal{V}$, $\tilde{a}(p,q) = 0$, (28)

is the weak formulation of the amplitude equation (14) for K, subject to (15). It follows that the system of equations (24) can also be obtained by applying the Galerkin method, with piecewise linear finite elements, to problem (14)–(15).

2.3. Numerical Calculations

We shall now solve the Galerkin problem using the new asymptotically derived basis functions, and again using the standard piecewise linear basis functions. Then we shall compare both results with the exact solution to problem (1)–(3). We shall use the wavenumber k = 25 and the index of refraction

$$n(x) = \exp[-10(x-1)^2)] + 1.$$
(29)

For $x \ge \pi$, $n(x) - 1 \le 1.2 \times 10^{-20}$.



FIG. 1. The real part of *u* as a function of *x*. The Galerkin approximation for k = 25 with asymptotic basis functions (solid lines) is indistinguishable from the exact solution (\diamond). The number of grid points *N* is shown above each graph. The approximation with the standard linear basis (dashed lines) agrees with the exact solution only when N = 1000.

Three sets of computations were performed on uniform grids with N = 50, N = 250, and N = 1000 points. On each grid the solution was computed by a standard linear finite element method and by a Galerkin method with the asymptotically derived basis functions. The results of both computations are compared graphically with the exact solution in Fig. 1. The "exact" solution is in fact an accurate approximation obtained with the linear basis on a uniform grid with 2000 points. The maximum error in this accurate approximation was estimated to be less than 0.008 by comparing it to the solution obtained with 4000 points. The solution based on the asymptotic elements is indistinguishable from the exact solution even for the smallest value N = 50. The solution based on the linear elements achieves the same "graphical accuracy" as the solution based on the asymptotic elements only when the number of grid points N = 1000. Thus with k = 25, the use of linear elements requires 20 times as many points as using asymptotic elements for the same accuracy.

3. PROBLEMS IN TWO DIMENSIONS

3.1. Variational Formulation and Galerkin Approximation

We now extend the method of Section 2 to two dimensions. We consider scattering by a simple curve Γ , of a time harmonic plane wave of unit amplitude, traveling to the right along the *x* axis. We set $n(x, y) \equiv 1$ for simplicity, although $n(x, y) \neq 1$ can be treated in a similar way. Hence we seek a solution of

$$\Delta u + k^2 u = 0 \tag{30}$$

in the unbounded domain exterior to the simple curve Γ . We seek u of the form

$$u = e^{ikx} + u^s, (31)$$

where e^{ikx} and u^s are the incident and scattered fields, respectively. We impose on u the boundary condition

$$u(x, y) = 0, \quad (x, y) \in \Gamma, \tag{32}$$

and on u^s the Sommerfeld radiation condition

0

$$\lim_{r \to \infty} r^{1/2} \left(\frac{\partial u^s}{\partial r} - iku^s \right) = 0.$$
(33)

In practice, u^s is computed in a bounded computational domain Ω , bounded internally by the scatterer Γ and externally by an artificial external boundary which we refer to as $\partial \Omega$. Usually a nonreflecting boundary condition is used on $\partial \Omega$ to replace (33) (Givoli [18]). However, in the examples we consider the exact solution u_{ex}^s is known, so on $\partial \Omega$ we impose the value of

$$f(x, y) = \frac{\partial u_{ex}^s}{\partial n} - iku_{ex}^s.$$
(34)

Hence the computational problem which we solve is (30) for u^s , subject to the conditions

$$u^{s} = -\exp(ikx), \quad (x, y) \in \Gamma, \tag{35}$$

$$\frac{\partial u^s}{\partial n} - iku^s = f(x, y), \quad (x, y) \in \partial\Omega.$$
(36)

We now introduce the variational formulation of the problem for u^s . First we introduce the spaces

$$\mathcal{S} = \{ u \mid u \in H^1(\Omega); u(x, y) = -\exp(ikx), \quad (x, y) \in \Gamma \}$$
(37)

and

$$\mathcal{V} = \{ w \mid w \in H^1(\Omega); \, w(x, y) = 0, \ (x, y) \in \Gamma \}.$$
(38)

The Sobolev space $H^1(\Omega)$ consists of all complex valued functions on Ω whose first order partial derivatives are square integrable. The variational formulation of (30) for u^s , subject to (35) and (36), is

find
$$u \in S$$
 such that for all $w \in \mathcal{V}$, $a(u, w) = -(f, w)_{\partial}$, (39)

$$a(u, w) = -\iint_{\Omega} \nabla u(x, y) \cdot \nabla \bar{w}(x, y) \, dx \, dy$$

$$+ k^2 \iint_{\Omega} u(x, y) \bar{w}(x, y) \, dx \, dy + \int_{\partial \Omega} iku \bar{w} \, ds.$$
(40)

$$(f,w)_{\partial} = \int_{\partial\Omega} f \bar{w} \, ds. \tag{41}$$

This formulation is derived in standard fashion by multiplying both sides of (30) by \bar{w} and integrating by parts [14].

In order to define the Galerkin approximation to problem (39), we introduce the finite dimensional approximations S_h and V_h of S and V, respectively. All members of $u_h \in V_h$ vanish or vanish approximately on Γ , and each member of S_h admits the representation

$$u_h = v_h + g_h, \tag{42}$$

where $v_h \in V_h$ and g_h satisfies exactly or approximately the Dirichlet condition (35) on Γ . The Galerkin formulation is:

find
$$u_h = v_h + g_h \in S_h$$
 such that for $\forall w_h \in V_h$,

$$a(v_h, w_h) = -a(g_h, w_h) - (f_h, w_h)_{\partial}.$$
(43)

The subscript *h* in \mathcal{V}_h and \mathcal{S}_h refers to a discretization of Ω . Here, we consider a triangulation of Ω , as depicted for example in Fig. 2. The mesh parameter *h* is the maximum edge length over all triangles *T* in the mesh. The number of nodes in this triangulation is \mathcal{N} , and the set of node indices η is $\eta = \{1, \ldots, \mathcal{N}\}$. The set η_d contains indices of nodes on Γ . With node *j*, we associate the basis function N_j , which is 1 at node *j*, 0 at all other nodes, and linear on every triangle [14]. Higher order basis functions can also be considered. With this notation, $w_h \in \mathcal{V}_h$ and g_h in (42) have the forms

$$w_h = \sum_{j \in \eta - \eta_d} a_j N_j, \quad g_h = \sum_{j \in \eta_d} g_j N_j, \tag{44}$$



FIG. 2. Left: the rays reflected from the surface of the parabola. Right: the triangulated computational domain.

where g_j is the value of g at node j. The solution v_h in (42),

$$v_h = \sum_{j \in \eta - \eta_d} X_j N_j, \tag{45}$$

is determined by solving the linear system

$$\sum_{j \in \eta - \eta_d} a(N_j, N_i) X_j = -\sum_{j \in \eta_d} a(N_j, N_i) g_j - (f, N_i)_{\partial}, \quad i \in \eta - \eta_d.$$
(46)

3.2. Asymptotically Derived Finite Elements

3.2.1. Elements of the Asymptotic Theory

The asymptotic theory for problem (30)–(33), based on geometrical optics and the geometrical theory of diffraction, is described in detail in [4–6]. Here, we review elements of that theory necessary for the development of asymptotic basis functions. In the asymptotic theory, one seeks the solution u^s in (31) as a superposition of the form

$$u^{s}(x, y) = \sum_{l=1}^{L} A_{l}(x, y) \exp(ikS_{l}(x, y)).$$
(47)

Here L is the number of fields. In the examples we consider here, the sum (47) consists of a reflected field and possibly one or two diffracted fields, and a shadow forming field, as we shall see. Each field

$$A(x, y) \exp(ikS(x, y)) \tag{48}$$

is a solution to Eq. (30). In (48), A is called the amplitude, k the wavenumber, $e^{ikS(x,y)}$ the phase factor, and S the phase. We shall use the methods of the asymptotic theory to determine the phase of each field.

First we substitute (48) into (30) and cancel the phase factor to obtain

$$-k^{2}[(\nabla S)^{2} - 1]A + 2ik\nabla S \cdot \nabla A + ikA\Delta S + \Delta A = 0.$$
⁽⁴⁹⁾

Equating the leading order term of (49) to zero yields the eiconal equation of geometrical optics:

$$(\nabla S)^2 - 1 = 0. \tag{50}$$

This is a first order nonlinear partial differential equation for the phase *S*, which we solve by the method of characteristics [6]. To do so we introduce the two-parameter family of characteristic curves or rays $X(\sigma, \tau)$, $Y(\sigma, \tau)$, and the equations

$$\frac{dX(\sigma,\tau)}{d\sigma} = S_x, \quad \frac{dY(\sigma,\tau)}{d\sigma} = S_y.$$
(51)

In this case, $S_x \equiv S_x(\tau)$ and $S_y \equiv S_y(\tau)$ are independent of the arclength σ because the index of refraction is a constant. Hence the rays propagate along straight lines in the direction ∇S :

$$X(\sigma,\tau) = X(0,\tau) + \sigma S_x(\tau), \quad Y(\sigma,\tau) = Y(0,\tau) + \sigma S_y(\tau).$$
(52)

On each ray, the phase is determined from the equation

$$\frac{dS(\sigma,\tau)}{d\sigma} = 1,$$
(53)

so

$$S(\sigma, \tau) = S_0(\tau) + \sigma.$$
(54)

The initial condition $S_0(\tau) = S|_{\Gamma}$ is determined on the surface of the scatterer $\Gamma(\tau)$,

$$\Gamma(\tau) = (X(0,\tau), Y(0,\tau)).$$
(55)

From the boundary condition (35) for the reflected field, $S_0(\tau) = X(0, \tau)$. Arguments pertaining to the geometrical theory of diffraction yield $S_0(\tau)$ for the diffracted fields. To determine the ray direction $\nabla S_0(\tau)$ we use (50) and the strip condition on Γ ,

$$\frac{dS_0(\tau)}{d\tau} = S_x(\tau)\frac{\partial X}{\partial \tau} + S_y(\tau)\frac{\partial Y}{\partial \tau}.$$
(56)

3.2.2. Asymptotically Derived Basis Functions

We define the asymptotic finite element basis functions associated with field l as the collection

$$\left\{M_{j}^{l}(x, y, k)\right\}_{j=1}^{N}$$
 (57)

Here \mathcal{N} is the number of nodes in the grid and

$$M_{i}^{l}(x, y, k) = N_{j}(x, y) \exp(ikS_{l}(x, y)).$$
(58)

Here N_j is the linear nodal finite element function for node j, and the phase factor $\exp(ikS_l)$ is determined by the method of the previous section. To represent u^s in (47) efficiently, the space of trial solutions must represent each field of u^s efficiently. Hence we define

$$\mathcal{S}_h^k = \left\{ u_h^k = v_h^k + g_h^k \, \middle| \, v_h^k \in \mathcal{V}_h^k, \, g_h^k = \sum_{j \in \eta_d} g_j M_j^r \right\},\tag{59}$$

$$\mathcal{V}_h^k = \left\{ w_h^k \, \middle| \, w_h^k = \sum_{l=1}^L \sum_{j \in \eta - \eta_d} a_j^l M_j^l \right\}.$$
(60)

The function g_h^k in (59) is represented in terms of basis functions associated with the reflected field M_j^r because only the reflected field is nonzero on Γ . Hence to satisfy (35), g_j must be given by

$$g_j = -\exp(ikx_j) / M_j^r(x_j, y_j, k),$$

with (x_j, y_j) the coordinates of node j on Γ . The Galerkin problem is

find
$$u_h^k = v_h^k + g_h^k \in \mathcal{S}_h$$
 such that for $\forall w_h^k \in \mathcal{V}_h^k$,
 $a(v_h^k, w_h^k) = -a(g_h^k, w_h^k) - (f_h^k, w_h^k).$
(61)

Its solution satisfies the linear system

$$\sum_{l=1}^{L} \sum_{j \in \eta - \eta_d} a(M_j^l, M_i^m) X_j^l = -\sum_{j \in \eta_d} a(M_j^r, M_i^m) g_j - (f_h^k, M_i^m)_{\vartheta},$$

$$i \in \eta - \eta_d, \quad m = 1, \dots, L.$$
(62)

3.3. Scattering from a Parabolic Cylinder

3.3.1. Rays and Phase of Scattered Field

We begin by studying the exterior problem of scattering of the plane wave $\exp(ikx)$ from the surface of the parabola

$$\Gamma(y) = \left(\frac{y^2}{2} - \frac{1}{2}, y\right).$$
 (63)

In this case, the scattered field u^s given by (47) consists only of a reflected field, so we write

$$u^{s} = u^{r} = A_{r}(x, y) \exp(ikS_{r}(x, y)).$$
 (64)

We shall determine the rays and phase S_r of this field following the method of Section 3.2.1. The boundary condition (35) implies that $S_r(x, y) = x$ on Γ . Thus

$$S_r\left(\frac{y^2}{2} - \frac{1}{2}, y\right) \equiv S_0(y) = \frac{y^2}{2} - \frac{1}{2},$$
 (65)

in view of the parameterization (63). Now we use (50), (56), and (65) to determine the ray direction $\nabla S(y)$ on Γ :

$$S_x(y) = \frac{y^2 - 1}{y^2 + 1}, \quad S_y(y) = \frac{2y}{y^2 + 1}.$$
 (66)

By using (63) and (66), we find that on Γ

$$\frac{Y(0, y)}{X(0, y)} = \frac{S_y(y)}{S_x(y)} = \frac{2y}{y^2 - 1}.$$
(67)

This implies that the rays propagate in the radial direction, as depicted in Fig. 2. We also note that

$$S_0(y) = r_0(y) - 1,$$
(68)

where $r_0(y)$ is the distance from the origin to the point $\Gamma(y)$. By using (68) in (54) with $\sigma = r - r_0(y)$, we find that $S(\sigma, \tau) = r - 1$. Thus

$$\exp(ikS_r(x, y)) = e^{-ik}\exp(ikr), \quad r = \sqrt{x^2 + y^2}.$$
 (69)

3.3.2. Numerical Calculations

We choose the computational domain Ω to be exterior to the parabola Γ and interior to the circle

$$y^{2} + (x + 3/8)^{2} = .5^{2},$$
 (70)

as is shown in Fig. 2. The mesh parameter h, which is the maximum over all triangles T of the largest edge length of T, is h = 0.0937.

In (34) we use the exact expression for the scattered field, which is (see [19])

$$u_{ex}^{s}(x, y) = -\exp(ikx)\frac{F(\eta\sqrt{k})}{F(\sqrt{k})}.$$
(71)

Here F is the Fresnel integral, which we compute numerically using the formula

$$F(w) = \frac{1}{2}\sqrt{\pi}e^{i\pi/4} - \left[\int_0^w \cos t^2 dt + i\int_0^w \sin t^2 dt\right].$$
 (72)

The parabolic cylinder coordinates ξ and η are defined by

$$x = \frac{1}{2}(\xi^2 - \eta^2), \quad y = \eta\xi.$$
 (73)

We compute the Galerkin approximation to the problem (30), (35), (36) for u^s in Ω with Γ defined by (63), using both the asymptotic basis and the standard linear basis, for k = 1, 10, 20, 30. The initial mesh with 67 unknowns is depicted in Fig. 2. In each computation, the mesh is uniformly refined until the maximum relative error at the node

k	Linear basis		Asymptoic basis	
	N	$\ $ Relative error $\ _{\infty}$	N	$\ $ Relative error $\ _{\infty}$
1	67	0.00066	67	0.00021
10	1309	0.039	67	0.0024
20	5401	0.020	67	0.00081
30	21937	0.015	67	0.00084

 TABLE I

 The Number of Unknowns N and the Maximum Relative Error at the Node Points, in the Galerkin Approximation^a

^{*a*} The approximation with piecewise linear basis functions is in the left column, and that with the asymptotic basis functions is in the right column.

points is less than 0.04. The numbers of unknowns required by the two methods are compared in Table I. This computation was done with the exact Neumann boundary condition on $\partial \Omega$. We see that the number of unknowns necessary to achieve the required accuracy with the piecewise linear basis grows rapidly with *k*, while the computation with the asymptotic basis requires a fixed number of unknowns. The maximum errors at the node points are also compared in Table I. We see that in all cases the solution with the asymptotic basis yields a higher accuracy for a substantially lower cost.

The top panel of Fig. 3 shows \log_{10} of the maximum absolute value of the relative error as a function of *k*, using the asymptotic basis. In the bottom panel, \log_{10} of the condition



FIG. 3. The top panel shows the \log_{10} of the maximum absolute value of the relative error as a function of *k*, using the asymptotic basis. In the bottom panel, \log_{10} of the condition number $\kappa(A)$ of the matrix *A*, which arises in the Galerkin method using the asymptotic basis, is shown as a function of *k*.

number $\kappa(A)$ of the matrix A, which arises in the Galerkin method using the asymptotic basis, is shown as a function of k. The relative error is computed by subtracting the computed solution from u_{ex}^s given by (71) and dividing by u_{ex}^s . The top graph shows that it does not increase with k, and the bottom graph shows that the condition number does not increase with k.

3.4. Scattering from a Circular Cylinder

3.4.1. Rays and Phase of Scattered Field

We now consider the exterior problem for scattering of an incident plane wave by a circle of radius *a*. With *a* as the unit of length, and with k' denoting the dimensionless product ka, we write the plane wave as $u^i = \exp(ikx)$, omitting the prime. Then Γ is the unit circle

$$\Gamma(\tau) = (\cos \tau, \sin \tau). \tag{74}$$

The asymptotic theory for this problem is described in detail in [5, 6]. The scattered field consists of four terms

$$u^{s} = u^{r} + u^{sf} + u^{d+} + u^{d-}, (75)$$

which are the reflected field u^r , the shadow forming field u^{sf} , the diffracted field u^{d+} originating at (0, 1), and the diffracted field u^{d-} originating at (0, -1). The fields u^r and u^{sf} satisfy the boundary conditions

$$u^{r}(x, y) = -\exp(ikx), \quad (x, y) \in \Gamma, \quad x < 0,$$
 (76)

$$u^{sf}(x, y) = -\exp(ikx), \quad (x, y) \in \Gamma, \quad x \ge 0.$$
 (77)

Both diffracted fields vanish on the surface of the cylinder. The field u^{sf} in the region $x \ge 0$, $|y| \le 1$ is called the shadow forming field because $u^{sf} + u^i = 0$ in that region, as we shall see.

We determine the reflected rays and phase by the method of Section 3.2.1. The rays

$$X(\sigma,\tau) = \cos\tau - \sigma\cos 2\tau, \quad \pi/2 \le \tau \le 3\pi/2, \tag{78}$$

$$Y(\sigma, \tau) = \sin \tau - \sigma \sin 2\tau, \quad \pi/2 \le \tau \le 3\pi/2 \tag{79}$$

are depicted in Fig. 4. The phase factor is

$$\exp[ikS_r(x, y)] = \exp[ik(\cos\tau + \sigma)], \quad \pi/2 \le \tau \le 3\pi/2.$$
(80)

Similarly for u^{sf} the rays and phase factor are

$$X(\sigma,\tau) = \cos\tau + \sigma, \quad -\pi/2 \le \tau \le \pi/2, \tag{81}$$

$$Y(\sigma, \tau) = \sin \tau, \qquad -\pi/2 \le \tau \le \pi/2, \tag{82}$$

$$\exp(ikS_r(x, y)) = \exp(ikx), \quad -\pi/2 \le \tau \le \pi/2.$$
(83)

In the shadow region, $|y| \le 1$, x > 0, $u^{sf} + u^i = 0$ [6]. However, the exact solution is nonzero there, in view of diffraction effects. The additional terms u^{d+} and u^{d-} in (47)



FIG. 4. The left panel shows the rays reflected by the circular cylinder, and the rays of the shadow forming beam. The right panel shows the diffracted rays emanating from (0, 1).

account for the diffracted field. The rays associated with u^{d+} , u^{d-} are called diffracted rays [4]. Each incident ray which is tangent to the cylinder gives rise to a surface diffracted ray. Here the incident rays are tangent at (0, 1) and (0, -1) and bound the shadow region. Each surface diffracted ray travels along the surface of the cylinder starting at the point of diffraction. As it travels along the surface, it sheds additional diffracted rays into the domain. These new rays leave the surface of the cylinder tangentially.

The surface diffracted ray emanating from (0, 1) travels in the clockwise direction along the surface of Γ and sheds the family of rays

$$X(\sigma, \tau) = \sin \tau + \sigma \cos \tau, \quad \tau, \sigma \ge 0, \tag{84}$$

$$Y(\sigma,\tau) = \cos\tau - \sigma\sin\tau, \tag{85}$$

depicted in Fig. 4. The associated phase is

$$S_{d_+} = \tau + \sigma. \tag{86}$$

The surface ray emanating from (0, -1) travels in the anticlockwise direction and sheds the family of diffracted rays

$$X(\sigma,\tau) = \sigma \cos \tau + \sin \tau, \quad \tau,\sigma \ge 0, \tag{87}$$

$$Y(\sigma,\tau) = \sigma \sin \tau - \cos \tau. \tag{88}$$

The phase is

$$S_{d_{-}} = \tau + \sigma. \tag{89}$$

The fact that τ is unbounded in (84)–(89) implies that a surface diffracted ray travels an infinite number of times around the cylinder and sheds infinitely many diffracted rays in the same direction. Since their phases differ by integer multiples of $2\pi ka$, we choose k' = ka to be an integer. Then their phase factors are all equal:

$$\exp(ikS_{d_+}(x, y)) = \exp(ik(\tau + \sigma)), \quad 0 \le \tau < 2\pi$$
(90)

$$\exp(ikS_{d_{-}}(x, y)) = \exp(ik(\tau + \sigma)), \quad 0 \le \tau < 2\pi.$$
(91)



FIG. 5. The illuminated computational subdomain $\pi \le \theta \le 3\pi/4$, $1 \le r \le 1.5$ (left panel) and the shaded subdomain $0 \le \theta \le \pi/4$, $1 \le r \le 1.5$ (right panel).

3.4.2. Numerical Calculations

We begin by presenting computational results for u^s on two subdomains of the exterior of the unit disk. They are the sector $0 \le \theta \le \pi/4$, $1 \le r \le 1.5$ and the sector $\pi \le \theta \le 3\pi/4$, $1 \le r \le 1.5$, depicted in Fig. 5. We refer to these subdomains as the shaded and the illuminated subdomains, respectively, because the first is in the shadow and the second is in the illuminated region. Their mesh parameters are h = 0.1106 and h = 0.1043, respectively. The purpose of these computations is to demonstrate that our basis functions can provide an accurate representation of the solution. Therefore in each computational domain, we impose the Dirichlet condition $u^s = -e^{ikx}$ on the unit circle. On the rest of the boundary of each computational domain, we impose the Robin condition (34). The exact solution is

$$u_{ex}^{s}(x, y) = -\sum_{n=0}^{\infty} \epsilon_{n} i^{n} \frac{J_{n}(k)}{H_{n}^{1}(k)} H_{n}^{1}(k\rho) \cos n\phi, \qquad (92)$$

where ρ and ϕ are the polar coordinates of the point (x, y), and J_n and H_n^1 are the Bessel function and the Hankel function of the first kind, respectively [19].

Table II presents the maximum relative error at the node points in each subdomain for several integer values of the wavenumber k. In each case the computation was performed twice, once with basis functions associated with the reflected or shadow forming field only,

TABLE II

	Illuminated region		Shaded region	
k	Reflected field	All fields	Shadow forming field	All fields
20	0.0015	0.0017	0.0091	0.0025
40	0.0026	0.0018	0.0189	0.0059
60	0.0035	0.0044	0.0252	0.0052
80	0.0031	0.0055	0.0326	0.0055

The Maximum Absolute Value of the Error in the Illuminated and Shaded Subdomains, with One Field (Reflected or Shadow Forming) and All Fields

and once with all fields. The \log_{10} of these data is presented in Fig. 6. The relative error is nearly independent of k.

We now present computational results for u^s in the annulus $1 \le r \le 1.5$. The solution is symmetric about the axis y = 0, so it suffices to compute the solution in the half annulus $y \ge 0$ shown in Fig. 7.

In this domain the amplitude varies rapidly in thin regions or layers, in contrast to the previous cases. The region of rapid variation occurs near the surface of the circle, and in particular, near the point of diffraction (0, 1). As $k \to \infty$, the neighborhood of the shadow



FIG. 6. The maximum relative error at the node points in each subdomain for several integer values of the wavenumber k. Computations were performed with one and three fields in both the shaded and the illuminated region.



FIG. 7. The computational domain for scattering from a circle.

boundary, x > 0, y = 1 also requires local refinement. We use the grid depicted in Fig. 7 for all the computations. The mesh parameter of the outer part of the grid is h = 0.1018, but near the surface of the circle it is h = 1/50.

Table III presents the maximum relative error at the nodes of the grid for different values of *k*. Each computation was performed once with the standard linear basis and once with the asymptotic basis. Here, the computation was done with the exact Neumann boundary condition on $\partial \Omega$. We see that the asymptotic basis yields substantially better results.

Figure 8 shows \log_{10} of the maximum absolute value of the relative error using the asymptotic basis. The error is computed by using u_{ex}^s given by (92). The graph shows that it does not increase much with k. The bottom panel indicates the condition number of the asymptotic finite element matrix.

TABLE III
The Maximum Relative Error at the Node
Points in the Full Annulus

k	Linear basis $\ relative error\ _{\infty}$	Asymptotic basis ∥relative error∥∞
1	0.0013	0.0005
10	5.9493	0.0091
20	3.5240	0.0061
40	63.1651	0.0080



FIG. 8. The \log_{10} of the maximum absolute value of the relative error using the asymptotic basis (top). The condition number of the asymptotic finite element basis (bottom).

3.5. Summary of the Method and Extension to Three Dimensions

We now summarize the asymptotic finite element method for scattering problems:

1. Determine the components of the scattered field by using the ideas of geometrical optics and the geometrical theory of diffraction [4]. In our examples the parabola scattered a reflected wave, while the circle scattered a reflected wave, a shadow forming wave, and two diffracted waves.

2. Determine the rays and the phase of each wave by the method of characteristics. In a medium with a variable index of refraction, this may have to be done numerically.

3. Assemble the system of equations for the Galerkin method using the asymptotic basis functions, replacing (34) by a nonreflecting boundary condition [18].

- 4. Solve the system of equations.
- 5. Determine the solution by adding together all the waves.

Steps 1–5 all apply in any number of dimensions and they can be automated. For simplicity we have illustrated them in one and two dimensions. We have shown that the condition number of the resulting equations, although large, does not increase with the wavenumber k. The error does not increase much with k either.

In more complicated problems more waves will have to be included in the solution. Some will result from multiple reflection, if the scatterer is not convex. Some will arise because of refraction, if the medium is not uniform. Others will result from edge or surface diffraction, if the scatterer is not smooth. They can all be multiply reflected, refracted, and diffracted. In

addition, caustics will occur, necessitating the use of complex phase functions on the dark side of the caustic. The complex phase can be found by using complex rays or by solving the eiconal equation in fractional powers of distance from the caustic.

All of these problems have been encountered in many problems of wave propagation. Consequently computer programs have been written to calculate rays which are reflected, refracted, diffracted, and multiply reflected, refracted, or diffracted. The programs also calculate the phase S(x) at points on these rays. The main complications arise in calculating the amplitudes of these waves, which is done numerically by our hybrid method. Many examples of such calculations, in two and three dimensions, can be found in the book *Geometric Theory of Diffraction*, edited by Hansen [20].

In addition, methods have been devised to calculate the phase function S(x) directly by solving the eiconal equation without using rays. Of course multiple valued solutions must be found, corresponding to the various waves.

In the case of an opaque convex scattering object in three dimensions, the only waves which occur in u^s are the reflected, shadow forming, and surface diffracted waves. Caustics will occur in the surface diffracted waves. This will require mesh refinement in the neighborhood of the caustic, and the use of a complex phase on its dark side. Aside from this difference, the other steps in the method are the same as those in the example of scattering by a circular cylinder.

We plan to present the method and results for some three-dimensional problems in a subsequent paper.

APPENDIX: ASYMPTOTIC ORTHOGONALITY OF BASIS FUNCTIONS

In the examples we consider, basis functions associated with different components of the scattered field are asymptotically orthogonal, as $k \to \infty$. Hence, they are linearly independent for large k. Indeed, the directions of the rays associated with fields l and m are different except perhaps on a set of measure 0. Hence, for any $\epsilon > 0$, there is a subset $V_{\epsilon} \subset \Omega$ such that

$$\int_{V_{\epsilon}} \left| M_{j}^{l} \bar{M}_{i}^{m} \right| \leq \epsilon/2, \quad l \neq m,$$
(A.1)

and $\nabla S_l \neq \nabla S_m$ in $\Omega - V_{\epsilon}$. Then

$$\begin{pmatrix} M_j^l, M_i^m \end{pmatrix} = \int_{V_{\epsilon}} M_j^l \bar{M}_i^m \, dx \, dy$$

$$+ \frac{1}{ik} \int_{\partial(\Omega - V_{\epsilon})} N_j N_i \exp(ik(S_l - S_m)) \frac{\nabla(S_l - S_m)}{|\nabla(S_l - S_m)|^2} \cdot \vec{n} \, ds$$

$$- \frac{1}{ik} \int_{\Omega - V_{\epsilon}} \exp(ik(S_l - S_m)) \nabla \cdot \left(N_j N_i \frac{\nabla(S_l - S_m)}{|\nabla(S_l - S_m)|^2} \right) dx \, dy, \quad (A.2)$$

by the divergence theorem. Hence, for any $\epsilon > 0$ and k sufficiently large,

$$\left| \left(M_{j}^{l}, M_{i}^{m} \right) \right| \leq \epsilon, \quad l \neq m.$$
(A.3)

This proves the asymptotic orthogonality of M_i^l and M_i^m .

For small values of the wavenumber k, basis functions could be linearly dependent and as a result the condition number of the finite element matrix could be very large. If necessary, the condition number could be improved by removing redundant basis functions as we describe now.

Consider basis function M_j^l . Let $\{\Phi_i\}_{i=1}^m$, $\Phi_i = M_{j_i}^{m_i}$, $m_i \neq l$ be basis functions associated with node j or one of its neighbors. The fact that $m_i \neq l$ indicates that $M_{j_i}^{m_i}$ represents a field different from M_j^l . Then, the projection of M_j^l on span $\{\Phi_i\}$,

$$P\left[M_{j}^{l}\right] = \sum_{i=1}^{m} y_{i} \Phi_{i}, \qquad (A.4)$$

is determined by solving the system (see [21])

$$\begin{pmatrix} (\Phi_1, \Phi_1) & (\Phi_2, \Phi_1) & \dots & (\Phi_m, \Phi_1) \\ (\Phi_1, \Phi_2) & (\Phi_2, \Phi_2) & \dots & (\Phi_m, \Phi_2) \\ \dots & \dots & \dots & \dots \\ (\Phi_1, \Phi_m) & (\Phi_2, \Phi_m) & \dots & (\Phi_m, \Phi_m) \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_m \end{pmatrix} = \begin{pmatrix} (M_j^l, \Phi_1) \\ (M_j^l, \Phi_2) \\ \dots & \dots \\ (M_j^l, \Phi_m) \end{pmatrix}.$$
(A.5)

The basis function M_i^l is considered redundant and is removed if

$$\frac{\left\|P\left[M_{j}^{l}\right]-M_{j}^{l}\right\|_{2}}{\left\|M_{j}^{l}\right\|_{2}} \leq \delta, \tag{A.6}$$

where δ is a small parameter. We choose $\delta = Ch^2$ to be commensurate with the truncation error of linear interpolation associated with the grid.

The removal of redundant basis functions can be done simultaneously with the assembly of the system for low computational cost. Indeed, all inner products (Φ_j, Φ_i) , (Φ_i, M_j^l) in (A.5) are computed during the assembly, in view of the second term in the bilinear form (40). Hence, no additional computations are necessary to create the system (A.5). In a regular mesh, the dimension of this system is very small and is bounded, since the number of neighbors for each node is small and is bounded. Hence the determination of $\{y_i\}$ in (A.5), and the subsequent test (A.6), requires a very small number of additional computations. It is also possible to explicitly orthogonalize basis functions belonging to different fields by the Gram–Schmidt process.

REFERENCES

- I. M. Babuška and S. A. Sauter, Is the pollution effect of the FEM avoidable for the Helmholtz equation considering high wave numbers?, *SIAM J. Numer. Anal.* 34, 2392 (1997). Reprinted in *SIAM Rev.* 42, 451 2000.
- J. R. Stewart and T. J. R. Hughes, A tutorial in elementary finite element error analysis: A systematic presentation of a priori and a posteriori error estimates, *Comput. Meth. Appl. Mech. Eng.* 158, 1 (1998).
- F. Ihlenburg and I. Babuška, Dispersion analysis and error estimation of Galerkin finite element methods for the Helmholtz equation, *Int. J. Numer. Meth. Eng.* 38, 3745 (1995).
- 4. J. B. Keller, Geometrical theory of diffraction, J. Opt. Soc. Am. 52, 116 (1962).
- 5. J. B. Keller, Diffraction by a convex cylinder, Trans. IRE AP-4, 312 (1956).
- 6. E. Zauderer, Partial Differential Equations of Applied Mathematics (Wiley-Interscience, New York, 1989).

- 7. E. Giladi, Hybrid Numerical Asymptotic Methods, Ph.D. thesis (Stanford University, 1995).
- E. Giladi and J. B. Keller, Iterative solution of elliptic problems by approximate factorization, *J. Comp. Appl. Math.* 85, 287 (1997).
- J. M. Melenk and I. M. Babuška, The partition of unity finite element method: Basic theory and application, Comput. Meth. Appl. Mech. Eng. 139, 289 (1996).
- 10. I. M. Babuška and J. M. Melenk, The partition of unity method, Int. J. Numer. Meth. Eng. 40, 727 (1997).
- O. Laghrouche and P. Bettess, Short wave modelling using special finite elements, J. Comput. Acoust. 8, 189 (2000).
- O. Laghrouche, P. Bettess, and R. Sugimoto, Analysis of the conditioning number of the plane wave approximation for the Helmholtz equation, in *Proceedings of the European Congress on Computational Methods in Applied Science and Engineering, ECCOMAS 2000*, 2000, PP. 1–14.
- P. E. Barbone, J. M. Montgomery, O. Michael, and I. Harari, Scattering by a hybrid asymptotic/finite element method, *Comput. Meth. Appl. Mech. Eng.* 164, 141 (1998).
- 14. T. J. R. Hughes, The Finite Element Method (Prentice Hall, New York, 1987).
- C. Johnson, Numerical Solution of Partial Differential Equations by the Finite Element Method (Cambridge University Press, Cambridge, UK, 1987).
- 16. G. Strang and G. J. Fix, An Analysis of the Finite Element Method (Prentice Hall, New York, 1973).
- C. M. Bender and S. Orszag, Advanced Mathematical Methods for Scientists and Engineers (McGraw–Hill, New York, 1978).
- 18. D. Givoli, Numerical Methods for Problems in Infinite Domains (Elsevier, Amsterdam/New York, 1992).
- P. L. E. Uslenghi, J. H. Bowman, and T. B. A. Senior, *Electromagnetic and Acoustic Scattering by Simple Shapes* (North-Holland, Amsterdam, 1969).
- 20. R. C. Hansen, Ed., Geometric Theory of Diffraction (IEEE Press, New York, 1981).
- 21. N. I. Akhiezer and I. M. Glazman, Theory of Linear Operators in Hilbert Space (Dover, New York, 1993).