Iterative Techniques for Rough Surface Scattering Problems

P. Cao and C. Macaskill
School of Mathematics and Statistics
The University of Sydney
NSW, Australia, 2006.

December 1, 1994

Abstract

The problem of wave scattering at a randomly rough surface gives rise to a non-Hermitian linear system, which corresponds to the discretization of a linear *Fredholm integral equation of the first or second kind.* In this paper, we describe the use of various iterative methods for the solution of the real linear non-symmetric system, corresponding to the complex system of the scattering problem, and the regions of convergence of the methods in term of the basic parameters of the problem. It has been found that Generalized Conjugate Gradient method is the only approach valid for very rough surfaces, $h/l > 1.0$ say (here h/l is the ratio of the RMS surface height to surface correlation length). The regions of validity for the iterative methods in terms of convergence and energy conservation are also found and it is shown that this depends not only on h/l and kl (here k is the wavenumber of the incident field) but also on the density of mesh points in the discretization. Some convergence properties are discussed by analyzing the spectral radius of the iterative matrix. It is found for small h/l that the spectral radius is linearly proportional to h/l and approximately proportional to $\sqrt{kl}$ for large kl. It is also found numerically that when $h/l \cdot \sqrt{kl}$ is greater than one, the iterative matrix for the linear integral equation of the second kind treated here has spectral radius greater than one, and therefore, the Generalized Conjugate Gradient method is not guaranteed to converge since the corresponding linear real system of the second kind may not have positive definite symmetric part. Finally, we present a comparison of iterative techniques for the linear integral equations of the first and second kinds, and find that when such iterative methods are to be employed, it is more efficient to use the second kind integral equation formulation than the first kind formulation.

1 Introduction

Numerical simulation techniques are now well-established for wave scattering from randomly rough surfaces (Fung and Chen [1], Kachoyan and Macaskill [2], Thorsos [3], Thorsos and Jackson [4] and Macaskill and Kachoyan [5]). Direct inversion methods are unsuitable for dealing with the very large matrices that arise in large 2-D problems (i.e. surfaces varying in one dimension only) or essentially any 3-D problem, so there is considerable interest in the use of iterative methods, which often require fewer computational operations and may also need less computer memory. On the other hand iterative solvers are not generally as robust as direct solvers, with the possibility of no or very slow convergence, depending on the form of the iterative matrix.

Rough surface scattering problems can be formulated as either first or second kind Fredholm linear integral equations (see e.g. [3]). We refer to the discretization of the integral equations as the *first kind linear system* or *second kind linear system* respectively. The matrix of the first kind linear system is complex symmetric, while the matrix of the second kind linear system is non-Hermitian. By rewriting these complex equations using the real and imaginary parts, we find the equivalent real systems have no special symmetry or structure and so the difficulty is to find a successful iteration scheme; as we shall see later, a variety of schemes will work in practice if the ratio of the rms surface height h to surface correlation length l is small, and if the second kind linear system is employed. As the scaled rms surface height is increased, the problem becomes more taxing numerically.

A Neumann series iteration scheme, corresponding to a basic iteration of the second kind linear system, is a natural choice for an iterative solution to the scattering problem. This method, and some extensions using a minimum residual relaxation method and preconditioning techniques have been explored in Macaskill and Kachoyan [5]. These techniques have the virtue of simplicity of implementation and speed, but as was indicated in [5] the rate of convergence decreases as h/l increases, and for large h/l, say $h/l > 1$, they found that sometimes convergence was not achieved at all.

An alternative approach is the Hestenes-Steifel conjugate gradient method (henceforth CG) described in [6], [7], [8], [9], [10] and [11]. It has been found that the standard CG method and various modifications of it (see [12]) can deal with surfaces of limited roughness, which give rise to matrices that are close to symmetric if the second kind formulation is used. This is to be expected as the CG algorithms were developed for real positive definite symmetric matrices. We find (see later) that the standard CG method breaks down even for quite moderate roughness (say $h/l = 0.1$) because of the lack of symmetry. Moreover, any two term algorithm of the CG type that does not adjust the search direction will terminate prematurely in theory if the coefficient matrix is not positive definite or semi-definite [12].

Thus it is of interest to look for more robust procedures. Since the late 1970's, iterative procedures for non-symmetric linear systems, based on searching using a generalized gradient derived from the symmetric part of the metrix have been developed in [19], [20], [21] and [22]. These methods are known as *Generalized Conjugate Gradient* (GCG) methods.

In this paper, we concentrate on the GCG method for scattering from a Gaussian rough surface with a Gaussian correlation function. We establish the conditions of convergence for this problem and solve the second kind linear system. A comparison of the CG, GCG and simple iterative methods is presented. Significantly improved convergence is obtained using the GCG method in all mildly and highly non-symmetric cases. The region of validity of the iterative approach is given in terms of the basic parameters of the problem (h, l and the incident wavenumber k). The effects of h/l, kl and the resolution of the surface on the energy conservation is also discussed in sections 4.3 and 4.4.

Finally, we give a brief discussion of the application of these techniques to the equivalent first kind linear system. Using the ITPACK package [25] we find that there are considerable difficulties involved, and that only one iterative technique is successful in this case. In addition, the first kind formulation is considerably more sensitive to the method used to truncate the infinite surface.

2 Solution of the second kind linear system

The scattering geometry is shown in Fig 1, where $\boldsymbol{k}_i$ and $\boldsymbol{k}_s$ represent the incident wavevector and the scattered wavevector respectively. The formulation of the scattering problem in the two-dimensional (2-D) case is considered in the present work, the surface varies in one dimension only, with the incident field and the scattered field in the $x - z$ plane.

Fig. 1 near here

We consider a Gaussian surface $z = \zeta(x)$ with the wavefield satisfying a Dirichlet boundary condition. The pressure field away from the surface can be written as

$$p(x, z) = p_{inc}(x, z) + p_s(x, z), \tag{2.1}$$

where the incoming plane wave p_{inc} is given by

$$p_{inc} = e^{ik(\alpha_0 x - \gamma_0 z)}, \tag{2.2}$$

with $\alpha_0 = \cos\theta$, and$\gamma_0 = \sin\theta$, θ being the incident angle measured from the horizontal.

The scattered field p_s is given by the Helmholtz integral formula

$$p_s(x, z) = \frac{k}{4}\int_{-\infty}^{\infty} \psi(x') H_0^{(1)}(k|\boldsymbol{r} - \boldsymbol{r}'|)dx', \tag{2.3}$$

where ψ, the normal-derivative of the wavefield, is given by

$$\psi(x) = \frac{i}{k}\left(-\frac{\partial p}{\partial z} + \zeta'(x)\frac{\partial p}{\partial x}\right)_{z=\zeta(x)}. \tag{2.4}$$

In Eq. (2.3), $\boldsymbol{r}$ is the position vector for a point away from the surface and $\boldsymbol{r'}$ is the position vector for a point on the surface, respectively, i.e., $\boldsymbol{r} = (x, z)$ and $\boldsymbol{r'} = (x', \zeta(x'))$. $H_0^{(1)}$ is the zero-order Hankel function of the first kind. As discussed by Meecham [13], the normal-derivative operator applied to Eq (2.3), with use of the normal-derivative of (2.1) gives

$$\psi(x) + \int_{-\infty}^{\infty} \psi(x')K(x', x)dx' = 2\psi_{inc}(x). \tag{2.5}$$

which is a Fredholm integral equation of the second kind for ψ. In Eq. (2.5),

$$\psi_{inc}(x) = -[\gamma_0 + \alpha_0\zeta'(x)]e^{ik[\alpha_0 x - \gamma_0\zeta(x)]} \tag{2.6}$$

and

$$K(x, x') = \frac{ik}{2}\frac{H_1^{(1)}(k\rho)}{\rho}[\zeta(x') - \zeta(x) - (x' - x)\zeta'(x)], \tag{2.7}$$

where $H_1^{(1)}(y) = -dH_0^{(1)}(y)/dy$, and both x and x' are confined to the surface.

Equation (2.5) has been extensively discussed in previous papers. We can consider (2.5) truncated to a finite domain, either by imposing a finite incident beam width (see e.g. Thorsos [3]) or by simply truncating as in Kachoyan and Macaskill [2]. For the 2-D case, since the kernel $K(x', x)$ is non-singular and it is an $\mathcal{L}_2$-function, the existence of the solution is guaranteed by the classical Fredholm theory and the solution may be derived as a uniformly convergent power series - the *Neumann series* (see [14]).

Furthermore, we may regard the equation (2.5) as the operator equation

$$\psi + A\psi = f, \tag{2.8}$$

where $A\psi = \int K(x', x)\psi(x')dx'$ and $f = 2\psi_{inc}$. Then the existence and uniqueness of a solution can be established by Neumann series

$$\psi = \lim_{n\to\infty} \psi_n = f + \sum_{k=1}^{n}(-1)^k A^k f \tag{2.9}$$

provided $\| A \| < 1$ for some suitable norm. The iterated operator A^n is defined by $A^0 := I$ (identity operator), and $A^n = AA^{n-1}$ for $n = 1, 2, \ldots$. Hence, the Neumann series is related to successive approximations by the result [15] that

$$\psi_{n+1} := f - A\psi_n, \qquad n = 0, 1, 2, ..., \tag{2.10}$$

converges to the unique solution ψ of the operator equation (2.8). We note that the existence of a solution to (2.8) is assured only when

$$\max_{x\in G} \int_G |K(x, x')|dx' < 1$$

i.e., the norm of the kernel is sufficiently small, and that this iterative form cannot be summed in closed form in general. Corresponding to the operator equation (2.8), there is an equivalent theory for the matrix equation, which is a discrete representation of the continuous operator (see [23]).

For the solution of the matrix equation

$$(I + A)\psi = f, \tag{2.11}$$

we have the same conclusion as for the continuous kernel, namely that for the matrix A arising from the discretization of the continuous system, if $\rho(A) < 1$ (the spectral radius of A less than 1), then the series

$$I - A + A^2 - A^3 + \cdots$$

converges to the sum $(I + A)^{-1}$.

Note that $\rho(A) < \| A \|$ for every matrix norm $\| \cdot \|$ so that the unique solution of the equation (2.11) can be obtained exactly using the power series (2.9) or equivalently the recursion formulas (2.10).

The above results suggest a very close correspondence between the integral equation and its discrete representation. Methods of solution based on the Neumann series will be henceforth referred to as the *basic iterative method.*

3 Generalized Conjugate Gradient Method

Before establishing the conditions of convergence we consider the structure of the matrix A in the linear system

$$(I + A)\psi = f \tag{3.12}$$

In rough surface scattering problems the matrix A corresponding to the kernel function $K(x', x)$ is a non-Hermitian complex matrix (see e.g. [2], [3]). By letting $A = A_1 + iA_2$, $\psi = \psi_1 + i\psi_2$ and $f = f_1 + if_2$, the complex case can be reformulated in terms of the equivalent real system of the second kind:

$$\begin{bmatrix} I + A_1 & -A_2 \\ A_2 & I + A_1 \end{bmatrix} \begin{bmatrix} \psi_1 \\ \psi_2 \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} \tag{3.13}$$

or simply

$$(I + B)\, u = f \tag{3.14}$$

where $u = (\psi_1, \psi_2)^\top$ and $f = (f_1, f_2)^\top$, and

$$B = \begin{bmatrix} A_1 & -A_2 \\ A_2 & A_1 \end{bmatrix} \tag{3.15}$$

Thus, $(I + B)$ is a general real matrix with no special structure.

As noted earlier, the conjugate gradient method, as a member of a larger class of methods that are referred to as *conjugate direction* algorithms, was originally developed for solving linear systems by Hestenes and Stiefel. Solving the real second kind linear system (3.14) is equivalent to solving the optimization problem

$$\min_u F(u) = \frac{1}{2} u^\top (I + B) u - f^\top u \tag{3.16}$$

One of the most important requirements for CG-algorithms was that the matrix $(I + B)$ should be positive definite. Gradually, the methods were extended to deal with positive semidefinite matrices, and then non-symmetric matrices with positive definite symmetric part. There are many different modified versions of the CG-algorithm.

In the present paper we suppose that all methods that use the CG directions or "quasi-CG directions" of the quadratic function in (3.16), belong in the class of *Conjugate Gradient Type Methods*, and divide them into two categories:

Category 1: the standard CG algorithm and its modified versions where the search-step along the conjugate gradient is modified without altering its direction at any stage. Typical examples are the scaled CG-algorithm and the augmented CG-algorithm.

Category 2: the GCG type methods which involve altering the conjugate gradient search directions instead of "quasi-CG directions", such as the GCR method (Generalized Conjugate Residual), GMRES method (Generalized Minimal Residual), IGCG (idealized generalized CG) and its alternative forms such as ORTHODIR, ORTHOMIN, and ORTHORES.

The first category is adequate for *symmetric* matrices of the form $(I + B)$ which may be positive definite or nonnegative. More generally, these CG methods are applicable also to the case where $(I + B)$ is "almost" symmetric or $(I + B)$ is mildly non-symmetric. The second category will converge in at most $2n$ iterations for non-symmetric problems so long as $(I + B)$ is PR (*positive real*), i.e. its symmetric part is positive definite (see Young and Jea [22], THEOREM 3.1, pp. 165 - 172, or Eisenstat al [24], COROLLARY 3.2, pp. 350). In practice though it has been found that the GCG methods are often suitable for problems with indefinite symmetric parts and that the number of iterations is far less than $2n$ in most circumstances.

The symmetric part of $I + B$, $H(I + B)$, is positive definite if the spectral radius of $H(B)$, the symmetric part of B, is less than one. This can be easily proved (See Appendix A). Hence to guarantee convergence of the GCG method we require:

$$\rho(H(B)) < 1, \tag{3.17}$$

whereas basic iteration requires

$$\rho(B) < 1. \tag{3.18}$$

In general these two requirements are not simply related for the non-symmetric matrices corresponding to the scattering problem.

Convergence properties of some versions of the CG or GCG methods have been discussed in the literature. See, for example, [8], [10], [16], [18] and [24].

An algorithm should not, in general, be preferred on the basis of convergence rate only but other aspects, such as the amount of computation needed per iteration and the computer storage requirement should be considered. Both *accuracy* and *efficiency* must be considered in choosing an appropriate numerical method for solving a physical problem. The CG type methods belong to a class of methods which perform a small number of "expensive" iterations, their convergence depends on accurate one-dimensional optimizations and each iteration is quite computationally intensive. By contrast, the basic iterative methods perform a large number of "cheap" iterations. Thus the particular choice of method will depend on the precise form of the problem under consideration. Basic iterative methods should be more effective for surfaces with small roughness as they converge quickly and cheaply for h/l small. An alternative for these cases may be to use standard CG methods (category 1 above). Both these techniques work well for symmetric matrices; we expect adequate performance if the non-symmetric part is small in magnitude, corresponding to small h/l in the second kind formulation.

For large h/l the GCG methods (category 2) should be considered, as these have been designed to deal with strongly non-symmetric problems. We can expect convergence over a wider range of parameters, and a moderate number of expensive iterations is acceptable under such conditions.

It is shown in [17] that if $(I + B)$ is a symmetric positive definite matrix and $\text{rank}(B) = r$ then the standard CG method converges in at most $r + 1$ steps. In practice, we found that the rank of B is not small, for example, $\text{rank}(B) > 60$ for a surface with 256 points and with $kl = 25$ and $h/l = 1.0$. It is also found that the rank of B increases significantly as kl increases. However, by using a preconditioning technique to convert the problem $(I + B)u = b$ into a related problem $(I + \tilde{B})\tilde{u} = \tilde{b}$ where $\tilde{B}$ is close to the zero matrix, it is sometimes possible to reduce the rank of $\tilde{B}$. Although this idea was developed for the standard CG method the preconditioner technique is applicable for improving the rate of convergence of GCG methods as well.

4 Analysis of the numerical results

Except where noted, the results in this section used 256-point surfaces obeying Gaussian statistics with a Gaussian correlation function and 32 intervals per correlation length (see Fig. 2). The incident angle of the wave, $\theta_i = 45°$. All the numerical work was done on an Apollo 10000.

Fig. 2 near here

4.1 Iterative methods

A brief description of the algorithms used is as follows:

BASIC (basic iterative method): This algorithm runs the unaccelerated classical iterative method corresponding to the Neumann series.

CG (conjugate gradient acceleration): This is the two-term form of conjugate gradient for symmetric and positive definite (SPD) matrices and mildly non-symmetric matrices.

MRR (minimum residual relaxation method): This is one kind of Neumann series iterative method based on *Richardson iteration* with the relaxation parameter. A left spectral (64-point) preconditioning technique which improves convergence significantly is used, as described in [5].

GCG (generalized conjugate gradient): This is a common truncated/restarted method for non-symmetric systems [22], and also includes the popular GCR method (Generalized Conjugate Residual or restarted ORTHOMIN) [24].

All the iteration algorithms except for the MRR were run using ITPACK [25]. To do this the complex matrix is interpreted as the corresponding real matrix with row and column size twice that of the original complex matrix. As mentioned, this real linear system is non-symmetric and may even be indefinite (i.e. the symmetric part is not positive definite).

Fig. 3 near here

In Figs. 3, we present the behaviour of RMS errors with respect to the number of iterations for CG-, MRR, BASIC and GCG methods respectively. Each figure shows that the RMS height of the surface is clearly a dominant factor in determining the rate of convergence for the method given.

By comparing the results in Fig. 3, it has been found that all the methods work well except for the standard CG method if the roughness of the surfaces is not too large, say $h/l < 0.7$. The standard CG is adequate only for very small roughness, but it has better than linear convergence, so that fast convergence may be obtained by the CG method for very weak roughness cases, say $h/l < 0.1$, i.e., the RMS slope angle ($= \arctan(\sqrt{2}h/l)$) is less than $8°$. By contrast, the speed of convergence of the BASIC algorithm is slower than the standard CG method in the very small roughness case. However, the BASIC method is valid over a greater region of parameter space than the CG algorithm and this algorithm is very stationary if $\rho(B) < 1$. Its weakness is that it is likely to fail in highly rough cases, such as $h/l > 0.7$.

The preconditioned MRR method as used in [5] is valid for a wide range of parameters and converges quickly. In particular the region of convergence is expanded to RMS roughness of order one ($h/l = 1.0$) for which the BASIC- and CG methods are not applicable, and when an optimal relaxational parameter and preconditioner are used the MRR method performs a number of 'cheap' iterations and is economic in its use of computer storage. However, it does not converge well in very rough cases, say, $h/l > 1.0$, i.e., RMS slope angle $> 55°$.

The GCG method converges quickly not only for surfaces of small RMS slope but also for very large slope cases in which other methods can not converge at all. Therefore, we may conclude that the GCG method is the most powerful iterative technique for surface scattering problems. One point of interest is that this method often converges even if it is well beyond the domain implied by the sufficient condition for its convergence, that is to say, the GCG method can sometimes be applied to solving the indefinite linear systems arising from very rough surface scattering problems.

Fig. 4 near here

Fig. 4 shows a comparison of convergence for the four different methods. In Fig. 4 we show the result for the average over 100 surfaces with (a) $h/l = 0.5$ and (b) $h/l = 1.0$, and with $kl = 15.0$ in both cases. It can be seen that the MRR converges very quickly in the small and medium roughness cases, say, $h/l < 1.0$, and the advantage of GCG becomes more significant with increasing roughness. In addition, the variation of convergence rates corresponding to the different surface realizations is quite significant, see Fig. 5 (a). Thus the ensemble average has to be considered to determine the region of basic parameters for which a given method is valid. From Fig. 5 (b), it is found that 100 surfaces is an adequate number to obtain an effective average. It is also been found that the GCG method converges for very rough surfaces, say, $h/l = 1.2$ (i.e., RMS slope angle $\approx 60°$) in Fig. 6. From this figure, a rough upper bound in convergence for the GCG method is $h/l \simeq 1.2$; this can be increased to some extent by using more mesh points. It will be shown later that the second kind linear system treated here is an indefinite system if $h/l > 0.3$ for most kl.

Fig. 5 near here

Fig. 6 near here

Another important point, as mentioned in section 3, is the use of preconditioning techniques. A *Neumann Polynomial* preconditioner is used with the GCG method in this paper unless indicated otherwise. For this method, a truncated Neumann series approximation to $(I+B)^{-1}$ is used. If we rewrite $I+B$ as $I+B = I+D-C$, where D is the diagonal of B, then $I+B = [I - C(I+D)^{-1}](I+D)$ so

$$\begin{aligned}(I+B)^{-1} &= (I+D)^{-1}[I - C(I+D)^{-1}]^{-1} \\ &= (I+D)^{-1}[I + C(I+D)^{-1} + (C(I+D)^{-1})^2 + \cdots]\end{aligned}$$

A truncated form of this series is then used for the preconditioning matrix Q^{-1} and the preconditioned system is given by

$$Q^{-1}(I+B)u = Q^{-1}b.$$

Fig. 7 clearly demonstrates that this is an effective way to improve the convergence.

Fig. 7 near here

Now we present guidelines for determining which method to use in practical situations in scattering problems.

The Small Roughness Case $(h/l < 0.1)$: The weakly non-symmetric matrix or "almost" symmetric matrix is obtained in this case, and usually the standard CG method and its scaled versions (Category 1) can be used and a *quadratic convergence* rate will be attained.

The Moderately Rough Case $(h/l < 0.7)$: In many cases, the BASIC method is more reliable and "cheap"in convergence than CG methods for the mildly non-symmetric linear system. The preconditioned MRR method is to be preferred due to its wider region of validity and the efficiency (CPU time and machine storage) in computation.

The Large Roughness Case $(h/l < 1.2)$: This sort of roughness of the surfaces (say, RMS slope angle approaching $60°$) leads to a highly non-symmetric matrix and the corresponding linear system is usually indefinite. The GCG method (or others from Category 2) is essential to guarantee convergence for roughness which is significantly greater than order one. Of course this method is effective over the whole range of roughness. However, it is quite "expensive" in storage.

The methods can be listed in terms of increasingly broad regions of validity and reliability as: CG method, BASIC-type methods, MRR method, and GCG method. For the case of a general non-symmetric system, a more reliable method based on the normal equations corresponding to $(I+B)$ (i.e., solving $(I+B)^\top(I+B)u = (I+B)^\top b$) may be used. This method is guaranteed to converge in *exact* arithmetic for any nonsingular system; however, its convergence is usually slow and roundoff error may preclude convergence altogether since there is a serious amplification of the spectral condition number, i.e. the ratio of largest to smallest eigenvalues of $(I+B)^{\top}(I+B)$. In ITPACK, there are two implementations of conjugate gradients applied to the normal equations, CGNR (Conjugate Gradient applied to the Normal Equations) and LSQR (Least Squares Algorithm), the latter of which is the most reliable numerically for the first kind linear system treated later in the paper.

4.2 Convergence properties of iterative methods for the scattering problem

In this section, some convergence properties of the GCG algorithm are given. Most of the results given here are also applicable to other iterative schemes. After studying the scattering problem numerically, it is found that the convergence behaviour of the algorithm mainly depends on several parameters used to construct the iterative matrix.

A good approximation to the true solution of the unsymmetric linear system requires convergence of the iterative method, which requires the spectral radius of the symmetric part of the iterative matrix to be less than unity, i.e., $\rho(H(B)) < 1$. Furthermore, the rate of convergence depends not only on the spectral radius but also on the spectrum of the linear system $\sigma(H(B))$, i.e., the distribution of eigenvalues. Clustered eigenvalues give rise to fast convergence because the rows of the iterative matrix are more nearly linearly dependent and hence the rank of the symmetric part of the iterative matrix, $H(B)$, is reduced. It is well known that the eigenvalues of a matrix are continuous functions of the entries, hence, it is natural that the spectrum as well as the spectral radius of the iterative matrix will depend on some crucial parameters. More directly, from a physical viewpoint, the convergence depends on the three factors h/l, kl and $\triangle x$, where $\triangle x$ is the non-dimensional length of the mesh subintervals on the surface. At this stage we define N to be the number of unknowns on the surface, L to be the non-dimensional length of the surface, i.e. the number of correlation lengths, and N_i to be the number of intervals per correlation length. Then $L = N \cdot \triangle x$ and $l = N_i \cdot \triangle x$.

First, the scaled RMS height, $\hat{h} = h/l$ and scaled wavenumber $\hat{k} = kl$ are two major factors in determining convergence. (We use $\hat{h}$, $\hat{k}$ in the analysis for brevity.) The entries of $H(B)$, the symmetric part of the iterative matrix B, in the real linear system (3.14) can be determined from those of A in (2.11); these entries are given by (Kachoyan and Macaskill [2])

$$A_{mn} = \triangle x \cdot K(x_m, x_n) \tag{4.19}$$

where x has been non-dimensionalized by l the correlation length, and where

$$K(x_m, x_n) = \begin{cases} \frac{\hat{k}\hat{h}i}{2} \cdot \frac{H_1^{(1)}(\hat{k}\rho)}{\rho} \cdot [\zeta(x_m) - \zeta(x_n) - (x_m - x_n)\zeta'(x_n)] & m \neq n, \\ \frac{\hat{h}}{2\pi} \cdot \frac{\zeta''(x_m)}{1+[\hat{h}\zeta'(x_m)]^2} & m = n. \end{cases} \tag{4.20}$$

and

$$\rho^2 = (x_m - x_n)^2 + \hat{h}^2[\zeta(x_m) - \zeta(x_n)]^2.$$

So, the entries of $H(B)$, which is obtained from the real and imaginary parts of A, i.e.

$$H(B) = \begin{bmatrix} H(A_1) & -S(A_2) \\ S(A_2) & H(A_1) \end{bmatrix}$$

are given by

$$H(A_{1_{mn}}) = \begin{cases} -\frac{\hat{k}\hat{h}\triangle x}{4\rho} \cdot Y_1(\hat{k}\rho)\Big[(x_m - x_n)\Big(\zeta'(x_m) - \zeta'(x_n)\Big)\Big] & m \neq n, \\ \frac{\triangle x}{2\pi} \cdot \frac{\zeta''(x_m)\hat{h}}{1+[\zeta'(x_m)\hat{h}]^2} & m = n. \end{cases} \tag{4.21}$$

$$S(A_{2_{mn}}) = \begin{cases} \frac{\hat{k}\hat{h}\triangle x}{4\rho} \cdot J_1(\hat{k}\rho)\Big[2\Big(\zeta(x_m) - \zeta(x_n)\Big) \\ \qquad\qquad + (x_n - x_m)\Big(\zeta'(x_m) + \zeta'(x_n)\Big)\Big] & m \neq n, \\ 0 & m = n. \end{cases} \tag{4.22}$$

where $H(A_1)$ is the symmetric part of the real part of A, $S(A_2)$ is the skew-symmetric part of the imaginary part of A and J_1 and Y_1 are Bessel functions.

We consider the entries of $H(B)$ for two cases, $\hat{h}$ small and $\hat{h}$ large.

CASE 1 (for small $\hat{h}$):

Clearly, the diagonal entries of H(B), say $H(B_{mm})$, i.e. those of $H(A_{1_{mm}})$, are independent of $\hat{k}$, and $\zeta''(x_m)\hat{h}/[1 + (\zeta'(x_m)\hat{h})^2] = O(\hat{h})$, hence, $|H(B_{mm})| \propto \hat{h}$.

The magnitudes of the off-diagonal elements of $H(B_{mn})$ are conveniently rewritten in the form

$$|H(A_{1_{mn}})| = a_{mn}\, \hat{k}\rho|Y_1(\hat{k}\rho)|\frac{\hat{h}}{\rho^2},$$

$$|S(A_{2_{mn}})| = b_{mn}\, \hat{k}\rho|J_1(\hat{k}\rho)|\frac{\hat{h}}{\rho^2},$$

where

$$\begin{aligned} a_{mn} &= \frac{\triangle x}{4} \cdot |x_m - x_n||\zeta'(x_m) - \zeta'(x_n)| \\ b_{mn} &= \frac{\triangle x}{4} \cdot \Big|2(\zeta(x_m) - \zeta(x_n)) + (x_n - x_m)(\zeta'(x_m) + \zeta'(x_n))\Big| \end{aligned}$$

If $\hat{k}$ is small, then

$$\hat{k}\rho|Y_1(\hat{k}\rho)| = O(1) \quad \text{and} \quad \hat{k}\rho|J_1(\hat{k}\rho)| = O((\hat{k}\rho)^2),$$

and the off-diagonal elements depend only on $\hat{h}$ and $|H(B_{mn})| \propto \hat{h}$ since $\hat{h}/\rho^2 = O(\hat{h})$.

If $\hat{k}$ is large, then

$$(\hat{k}\rho)|Y_1(\hat{k}\rho)| = O(\sqrt{\hat{k}\rho}) \quad \text{and} \quad (\hat{k}\rho)|J_1(\hat{k}\rho)| = O(\sqrt{\hat{k}\rho}),$$

thus

$$|H(A_{1_{mn}})| \sim a_{mn}\sqrt{\hat{k}} \cdot \hat{h}\rho^{-3/2} \quad \text{and} \quad |S(A_{2_{mn}})| \sim b_{mn}\sqrt{\hat{k}} \cdot \hat{h}\rho^{-3/2}$$

therefore, $|H(B_{mn})| \propto \hat{h}\sqrt{\hat{k}}$ since $\hat{h}\rho^{-3/2} = O(\hat{h})$.

In brief, the magnitude of the diagonal elements of $H(B)$ depend only on $\hat{h}$ and are linearly proportional to $\hat{h}$ for small $\hat{h}$, i.e.,

$$|H(B_{mm})| \propto \hat{h} \qquad \text{if } \hat{h} \text{ is small.} \tag{4.23}$$

The magnitude of the off-diagonal elements of $H(B)$ depends on $\hat{h}$ only if $\hat{k}$ is small, and on $\hat{h}$ as well as $\sqrt{\hat{k}}$ if $\hat{k}$ is large. The relation is given by

$$|H(B_{mn})| \propto \begin{cases} \hat{h} & \text{if } \hat{k} \text{ is small,} \\ \hat{h}\sqrt{\hat{k}} & \text{if } \hat{k} \text{ is large.} \end{cases} \tag{4.24}$$

It is well known, by the *Geršgorin theorem*, that the diagonal entries determine the center of the *Geršgorin discs* while the absolute row sum of the off-diagonal elements determines the radius of these discs, and all the eigenvalues of $H(B)$ are located in the union of the N discs which correspond to the unknowns. Therefore, in the small $\hat{h}$ case, the relation between the spectral radius and the parameters can be given by Eq. (4.23) and (4.24)

$$\rho(H(B)) = \mu\hat{h} + \nu\hat{h}\sqrt{\hat{k}} \tag{4.25}$$

where μ and ν are constants depending in a complicated way on the shape of the surface.

CASE 2 (for $\hat{h}$ not small):

In this case, the situation is more complicated, and we have found no simple analytic relation between the spectral radius and $\hat{h}$. However, the dependence on $\hat{k}$ is unchanged from the small $\hat{h}$ case. Numerical results in Fig. 8 indicate that the spectral radius grows more slowly than $\hat{h}$ for large $\hat{h}$ and that for larger $\hat{h}\sqrt{\hat{k}}$ a simple result holds as discussed below.

Fig. 8 near here

Fig. 9 near here

The result in Fig. 8 is for only 10 surfaces with $L = 8$ and $N_i = 32$ (256 mesh points). This smaller number of realizations was used because of the large amount of computation needed to find the spectral radius for each kl. For adequate resolution kl should be less than 50 (from a requirement of roughly four intervals per acoustic wavelength), so no results are shown for larger kl. The numerical results for small h/l agree with the analysis above, and in particular Fig. 8(b) shows the spectral radius growing with $\sqrt{kl}$ and Fig. 8(d) shows the linear dependence on h/l for small h/l. In Fig. 9 we replot the results for the spectral radius against $h/l \cdot \sqrt{kl}$. Surprisingly, for large $h/l \cdot \sqrt{kl}$ we find that all results collapse on to a single curve and from numerical experiment the spectral radius $\rho(H(B)) \simeq 1.7 + h/l \cdot \sqrt{kl}/6$.

We have not found any analytical explanation for this. Also from Fig. 9, when $h/l \cdot \sqrt{kl} > 1$, $\rho(H(B)) > 1$, and we expect difficulty with the iterative methods as the GCG method is no long to guaranteed to converge. The following table shows a numerical test on one 256-point surface with $kl = 15.0$.

Table 1 indicates that the spectrum of $H(B)$, as well as the spectral radius, increases

h/l	$\sigma(H(B))$	*cond* $(\cdot)$	*No. iter*	*time* (sec.)	$\Vert W^{-1}\Vert_2$, $\Vert W\Vert_2$	*energy*
0.1	$(-0.33, 0.37)$	3	3	8.87	1.01, 1.03	0.990
0.3	$(-0.91, 0.97)$	8	7	14.99	0.93, 1.14	0.999
0.5	$(-1.37, 1.35)$	18	8	16.63	0.89, 1.54	1.062
0.7	$(-1.69, 1.55)$	25	10	19.72	0.83, 1.54	1.050
1.0	$(-2.01, 1.85)$	47	14	25.76	0.72, 1.58	1.008
1.2	$(-2.19, 1.97)$	57	20	35.64	0.66, 1.78	1.034

Table 1: Numerical results for different h/l.

significantly with h/l increasing, and then the number of iterations increases. Here *cond* $(\cdot)$ denotes the condition number of $(I+B)$, $W = Q^{-1}(I+B)$ is the preconditioned matrix and Q^{-1} is defined as mentioned before. Because $\Vert W\Vert_2$ and $\Vert W^{-1}\Vert_2$ are the greatest lower bound and the smallest upper bound for the largest and smallest eigenvalues of the preconditioned matrix separately in the non-symmetric case, it can be seen that the spectrum is compressed efficiently in the preconditioned system. When $h/l = 0.3$ or over, corresponding to $kl = 15.0$, $\rho(H(B)) > 1$, and because the most negative eigenvalue is roughly equal in magnitude to the most positive eigenvalue, then $(I+B)$ has both positive and negative eigenvalues and we have to treat an indefinite system.

Secondly, $\triangle x$, the length of the subinterval depending on the number of mesh points on the surface, is another important factor for convergence. Since the matrix A (similarly for $H(B)$) is the discrete form of an integral operator with continuous kernel K, its *maximum norm*

$$\Vert A\Vert_\infty = \Vert \lim_{\triangle x \to 0} \triangle x \cdot K(x', x)\Vert_\infty = \max_{x \in G} \int |K(x', x)|\, dx'$$

is bounded. For a detailed proof, see [15]. This implies that the maximum norm of A tends to a definite value as $\triangle x \to 0$. Since $\rho(A) = \inf\{\Vert A\Vert : \Vert\cdot\Vert \text{ is a matrix norm}\}$, the spectral radius will not be changed in general whenever the density of mesh points varies and hence $\triangle x$ is not a crucial factor for the spectral radius as $\triangle x$ decreases. However, as $\triangle x$ increases to a certain value, i.e., when the number of mesh points is too few, the true solution may not be obtained since low resolution of the surface leads to an huge error, as is discussed further in the next section. Conversely, too high a density of mesh points leads to needlessly expensive computation. (See

Fig. 10 and Table 2).

Fig. 10 near here

The number of iterations depends on N_i as shown in Fig. 10 if the density of mesh points is too sparse. This fact is demonstrated in Table 2, where the information comes from the ITPACK programme NSPCG with $h/l = 0.5$ and $kl = 25$ by using the GCG on the second kind equation. The *cond* $(\cdot)$ denotes the condition number

N_i	$\sigma(H(B))$	*cond* $(\cdot)$	*No. iter*	*time* (sec.)	$\|W^{-1}\|_2$, $\|W\|_2$	*energy*
128	$(-1.61, 1.51)$	19.76	9	352.0	0.89, 1.41	0.961
64	$(-1.61, 1.51)$	19.72	9	73.2	0.89, 1.41	0.959
32	$(-1.60, 1.50)$	20.00	9	18.0	0.89, 1.41	0.957
16	$(-1.59, 1.49)$	22.02	10	5.1	0.88, 1.40	0.952
8	$(-1.72, 1.76)$	34.60	34	7.4	0.53, 2.54	0.658

Table 2: Numerical results for the different densities of mesh points.

of $(I + B)$ here. The spectra of $H(B)$ are essentially independent of the number of mesh points so long as we have sufficient points. However, if an inadequate density of mesh points is used, say $N_i = 8$ in this case, we may not obtain the true solution of the problem even if the iterative method converges. Physical arguments would indicate that at least four intervals per acoustic wavelength are required. The sudden increases in spectral radius in Fig. 11 correspond to about two intervals per wavelength: for large h/l, these jumps occur at smaller values of kl, so that the physical criterion appears to be appropriate.

Fig. 11 near here

In summary, the convergence of iterative methods for the scattering problem depends on the parameters h/l and kl, in Eq. (4.25). The method will diverge if the spectral radius is large enough and it will converge to a wrong solution if the density of mesh points is inadequate. Hence, choosing an adequate density of mesh points is important, as too sparse a mesh density leads to loss of accuracy, while too high a density leads to an excessively high cost of computation.

4.3 The conservation of energy

It is well known that the total power scattered by a perfectly reflecting rough surface should be conserved. One criterion to check the correctness of the numerical scheme is that the *reflection coefficients* must satisfy energy conservation, i.e., the *reflected* *energy* E must satisfy

$$E = \sum_{r=-\infty}^{\infty} |R_r|^2 \frac{\gamma_r}{\gamma_\circ} = 1 \tag{4.26}$$

where R_r are the reflection coefficients (see [2]), e.g., from which the scattered field can be obtained. In this subsection, we will explore numerically how the energy conservation depends on the basic parameters and on the discretization of the surface.

As discussed above, h/l, kl and $\triangle x$ are the three basic quantities to be considered. Two points should be considered separately; one is the accuracy of energy conservation while the other is the range of the values of the basic parameters where energy is conserved. The reflection coefficients are given by ([2])

$$R_r = \frac{1}{2\gamma_r} \sum_{m=0}^{N-1} \hat{\psi}_m C_{r-m}(kh\gamma_r), \tag{4.27}$$

where

$$C_r(\tau) = \frac{1}{N} \sum_{n=0}^{N-1} \exp[-i\tau\zeta(x_n)] \exp\left(\frac{-2\pi i r n}{N}\right), \tag{4.28}$$

and

$$\gamma_r^2 = \begin{cases} 1 - \left(\frac{2\pi r}{klN\triangle x} + \alpha_0\right)^2 & \text{for } r = 0, 1, \cdots, \frac{N}{2} - 1 \\ 1 - \left(\frac{2\pi(r-N)}{klN\triangle x} + \alpha_0\right)^2 & \text{for } r = \frac{N}{2}, \cdots, N-1 \end{cases} \tag{4.29}$$

where $\hat{\psi}$ is the discrete Fourier transform of ψ.

Obviously, from (4.26) and (4.29), in theory, the error in the total reflected energy $|E - 1.0| \to 0$ as $N \to \infty$ for fixed (and suitably small) $\triangle x$. In other words, for energy conservation to be satisfied all the real γ_r (propagating modes) must be found, and this requires that $N > 2M$ where $M = [1 + kl\ L/\pi]$ is the number of real γ_r. Therefore we expect that E will approch unity as L, the length of the surface, is increased, if N_i is held constant, and this is confirmed in Fig. 12. Conversely, if N is held constant, and $\triangle x$ is reduced, i.e. the mesh density N_i is increased, then L will decrease correspondingly and the error in E will increase.

Fig. 12 near here

Now we describe numerically the range of scaled wavenumbers kl of the incoming wave for which the scattered energy is conserved for different RMS heights of the surfaces. In general, energy conservation is satisfied for a wide range of kl when h/l is small and for a narrower range for larger h/l for some given accuracy. The following graphs are averages over 50 surfaces showing how the reflected energy varies with increasing kl.

Fig. 13 near here

Fig. 13 shows the range of kl for which energy is conserved for fixed h/l. It has been found that the range of kl where energy is conserved decreases rapidly as h/l increases. The reason is that energy conservation is satisfied only if the true solution ψ is obtained by the iterative method, but this is hard to achieve for a wide range of kl if h/l is large. On the other hand, the range of kl where energy is conserved also depends upon the length of the surface L and the resolution of the surfaces N_i. As shown in Fig. 14, the range of kl is reduced as either N_i is reduced while holding L constant or L is decreased while holding N_i fixed. In Fig. 14 (a), the breakdown at large kl is roughly at two intervals per wavelength, consistent with the spectral radius jumping in Fig. 11. In (b), small kl behaviour is improved by increasing L. In other words, for large kl results will be poor if the resolution is inadequate; as $kl \to 0$, by contrast, the limitation comes because of the length of the surface, which must be increased as the acoustic wavelength increases in order to resolve the scattered modes at close to grazing angle and hence again the number of correlation lengths L of the surface must be increased if the resolution is to remain unchanged as kl decreases.

In general, for conservation of the scattered energy, high frequency waves require high resolution of the surface and low frequency waves need long surfaces. These requirements become stricter as the surface roughness is increased.

Fig. 14 near here

4.4 The region of validity of the GCG method

In this section we assess the regions of validity of the GCG method in term of the basic scaled parameters h/l and kl.

The regions of validity for the approximate theories (Kirchhoff and Perturbation) have been established for scattering from surfaces with small roughness and Gaussian correlation by Thorsos and Jackson [4]. In theory, the region of validity for an exact treatment of the scattering problem by integral equation theory should be the whole regime where physical solutions exist. Even with direct methods, however, this region will be reduced by finite resolution. When iterative methods have to be used, there are further limitations imposed by possible non-convergence of the iterative schemes. The regions of validity for the Kirchhoff approximation and for lowest-order perturbation theory are given in terms of the accuracy of the approximation methods. We define the valid regions for the integral equation method to be those where convergence is attained to some given accuracy and where energy is conserved within some bound. We define the region of validity of the GCG method for rough surface scattering, for the discretization used, to be that where convergence of the iteration is attained to five figures and where $|E - 1.0| < \epsilon$, for some given ϵ.

As mentioned above, to generate the region of validity for the GCG method, the procedure divides into two stages. The first stage is to find the solution ψ of Eq. (2.11), which corresponds to the derivatives of the pressure field on the scattered surface, and then compute the reflection coefficients, by which the scattered pressure is performed. The second stage is to determine the region of kl in which the condition of the conservation of energy (4.26) is satisfied for given accuracy. It is clear that the accuracy of the reflection coefficients depends on the accuracy of ψ obtained, hence, in order to guarantee the conservation of energy with the high accuracy, the convergence of the iterative method is essential. Note, however, that convergence of the iterative method says nothing about the accuracy of the results, although inadequate resolution usually implies a large spectral radius so that convergence is less likely to be achieved. For example, in Fig. 15(a), a good approximation for ψ is obtained on a 64-point surface when kl is moderate, and the reflection coefficients are identical to the 256-point case. But in (b) they are not in agreement at all since h/l is too large for the 64-point mesh, leading to a incorrect R_r and so that energy is not conserved.

Fig. 15 near here

Fig. 16 presents the regions of validity for the GCG method on 50 surfaces in terms of convergence and energy conservation for different mesh points densities (the minimum value of kl is 1 here). It can be seen that the regions of validity for the iterative method based on the integral equation are much larger than for perturbation theory and the Kirchhoff approximation [4]. There are two factors that determine the form of the curves in this figure. For intermediate kl when h/l is large then either the spectral radius is too large to allow convergence (see Fig. 8), or the resolution is not high enough to guarantee the true solution ψ to be obtained, and these effects are more significant as kl increases. The upper boundaries of these regions, for large kl, correspond to the requirement that these should be more than 2 intervals per wavelength for energy conservation, as in Fig. 13. For very small kl, by contrast, the iteration works well, but energy is not conserved for the reasons discussed above. Since the range of kl for which energy is conserved increases with increasing resolution, so the region of validity for the GCG method will be expanded as the mesh point density on the surface is increased. All of this is also true for direct inversion techniques.

Fig. 16 near here

However, there is a further possible limitation of the GCG technique that at least in principle is not a difficulty with direct inversion methods. The spectral radius may just be too large. A simple calculation will make this clear. From Fig. 9 we have (for large h/l and kl)

$$\rho(H(B)) \simeq 1.7 + \frac{h/l\sqrt{kl}}{6}. \tag{4.30}$$

Numerical experimentation indicates that if the spectral radius exceeds some maximum value, say $\rho_{max} \simeq 3.5$, then convergence may not be obtained in practice for at least some of the realizations. This puts an upper bound on the practical validity of the iterative method, $\rho < \rho_{max}$, i.e.,

$$\frac{h}{l} \le \frac{6(\rho_{max} - 1.7)}{\sqrt{kl}}. \tag{4.31}$$

This is plotted in Fig. 16 (a1) and (b1) for $\rho_{max} = 3.5$. This puts an upper limit on the GCG method that is not present for direct methods. One solution to this problem would be to find an appropriate preconditioning technique, so as to extend the maximum spectral radius for which the GCG method converges.

4.5 Edge effects and the use of the integral equation of the first kind

As an alternative to using the integral equation of the second kind, the integral equation of the first kind can be developed from (2.3):

$$\int_{-\infty}^{\infty} \psi(x')K(x', x)dx' = p_{inc}(x), \tag{4.32}$$

where

$$\begin{aligned} \psi(x) &= \Big(\frac{i}{k}\frac{\partial p}{\partial n}\frac{ds}{dx}\Big)_{z=\zeta(x)}, \\ K(x', x) &= -\frac{k}{4}H_0^{(1)}(k|\boldsymbol{r} - \boldsymbol{r}'|), \\ p_{inc} &= e^{ik(\alpha_0 x - \gamma_0\zeta(x))}. \end{aligned}$$

Here, α_0 and γ_0 are same as in Eq. (2.6), and x and x' are confined to the surface.

A corresponding expression for the operator equation of the first kind is

$$A\psi = f, \tag{4.33}$$

where $f = p_{inc}$. Similarly, the real linear system is

$$Cu = b. \tag{4.34}$$

Although the integral operator A is complex symmetric for the integral equation of the first kind, no results for the real symmetric case can be applied to it since the complex symmetric matrix need not be diagonalizable, and it is clear that the corresponding matrix B is a general real matrix. This kind of equation has been used for solving rough surface scattering with a Gaussian roughness spectrum in the direct method by Thorsos. We tried to find some applicable iterative methods with the possible preconditioners in ITPACK to deal with this linear system, but only one iterative method, the LSQR, is robust enough to handle the first kind linear system and it is slow (See Fig. 17). The LSQR, the Least Squares Algorithm, is applied to the normal equations of the preconditioned system minimizing $\|Q_L^{-1}(b - Cu^{(n)})\|_2$, the 2-norm of the residual vector of the original preconditioned system, at each iteration (See [26]). Some reasons for the lack of success of iterative methods, even the GCG, for the integral equation of the first kind may be understood from Fig. 17 and the following table.

Table 3 is an numerical example with $h/l = 0.235$, $kl = 5.65$ and a 256-point

Type	$\sigma(H(\cdot))$	*No. iteration*	*cpu* (sec.)
1st kind	$(-4.42, 0.04)$	69	256
2nd kind	$(-2.11, 2.13)$	51	191

Table 3: Numerical results for the integral equations of the first and second kinds using the LSQR method for the same scattering example.

surface with four intervals per correlation length. The LSQR method is used here for the linear system of both kinds. $\sigma(H(\cdot))$ denotes the spectrum of the symmetric parts of B for the second kind and G for the first kind, where G is the iterative matrix of $C = I + G$ in Eq. (4.34). It is found that the convergence of the first kind is slower than that of the second kind due to the larger spectral radius of the former even though both spectra have a similar range. Here $\rho(H(G)) = 4.42$ in the first kind system, while $\rho(H(B)) = 2.13$ in the second kind, therefore, the convergence of the iterative methods for the second kind system is faster than that for the first kind system.

Fig. 17 near here

On the other hand, to avoid edge effects, we must taper the incident field to negligible levels at the ends of the finite surface. It has been found, as shown in Fig. 17, that the convergence of iterative methods does not depend on whether the incident field is tapered. Here we use the Gaussian taper function [3] as

$$p_{inc}(\boldsymbol{r}) = \exp\{i\boldsymbol{k}\cdot\boldsymbol{r}[1 + w(\boldsymbol{r})] - (x - z\cot\theta)^2/g^2\}, \tag{4.35}$$

where

$$w(\boldsymbol{r}) = [2(x - z\cot\theta)^2/g^2 - 1]/(kg\sin\theta)^2.$$

In Fig. 18, numerical tests are presented to show how the treatment of the edges of the computational domain affects the calculated scattering strength. It is found that the edge effect of the first kind linear system is larger than that of the second kind system before the single incident plane wave is tapered, because the kernel is not singular at $\boldsymbol{r} = \boldsymbol{r}'$ for the second kind equation in the 2-D case, whereas the kernel of the first kind has an integrable (weak) singularity at $\boldsymbol{r} = \boldsymbol{r}'$, so, when the Hankel function is evaluated, the discretization errors may be significant at the edge of the surface. After tapering (with $g = L/4$) a significant improvement is attained at the edge of the surface for both kinds of integral equation.

In summary, when iterative methods are considered, using the linear integral equation of the second kind may be preferred over using first kind for a number of reasons. First, the second kind equation can be treated with a wide of range of iterative techniques since its corresponding matrix representation is diagonally dominant or close to diagonally dominant, whereas this is not the case for the first kind equation. The second kind system therefore has a smaller spectral radius than the first kind system, and may be solved by iterative techniques for a wide range of the physical parameters in the problem. Secondly, the diagonal elements of the matrix for the second kind equation have no singularity as the kernel is not singular (in two dimensions) and hence the edge effects are smaller in this case. Of course as discussed in Thorsos [3], the first kind formulation requires less calculation for evaluation of the matrix elements, because of the complex symmetry.

Fig. 18 near here

5 Conclusion

Any kind of iterative method has conditions for its convergence and the rate of convergence depends on the class of iterative matrix. The numerical solution of scattering from a rough surface generates a large complex non-Hermitian matrix. Several iterative methods have been used for solving this problem. It has been found that standard CG methods are tractable only for surfaces with very small roughness, $h/l < 0.1$ say. The BASIC method based on the Neumann series expansion is an effective iterative scheme for mild roughness with $h/l < 0.7$. The MRR method is a very efficient iterative method for non-symmetric systems because it has fast convergence but is also useful for a wide range of roughness, say, h/l close to 1. The GCG method is one of the most effective methods at present for solving the non-Hermitian linear system which arises in discretization of the integral equation of the second kind. The GCG method, furthermore, has applicability over a wider parameter regime up to $h/l \approx 1.2$, i.e., the RMS slope angle of the surface is greater than $60°$. The validity of the GCG method has been assessed in terms of the method achieving convergence and satisfying energy conservation.

Convergence of the iterative methods depends strongly on h/l as well as kl with the spectral radius proportional to h/l for small h/l, independent of kl for small kl and proportional to $\sqrt{kl}$ for large kl. For large h/l there is no simple analytical relation for the spectral radius but numerical results indicate that the spectral radius grows more slowly than h/l. However, for large kl and large h/l we find numerically that the spectral radius is proportional to $h/l \cdot \sqrt{kl}$. The regions in h/l, kl parameter space for which energy is conserved have been delineated. It is shown that the iterative methods used are adequate for h/l of order one so long as sufficient mesh points are taken. Energy may not be conserved if kl is either too large or too small. The large kl case may be dealt with by taking more mesh points; the small kl case may be treated by increasing the surface length which generally again involves an increase in the number of mesh-points.

Finally, the integral equation of the second kind is more efficient computationally than the integral equation of the first kind, when iterative methods are used. Reasons that the second kind system may be preferred over the first kind are given in section 4.

6 Acknowledgment

One of the authors (PC), was supported by OPRS (overseas postgraduate research scholarship) by the Department of Employment, Education and Training, Australia.

A Appendix

That the GCG method converges to the exact solution in at most $2n$ iterations in the absence of rounding error provided is guaranteed by

Theorem A.1 *The symmetric part* $H(I+B) \equiv \frac{1}{2}[(I+B)+(I+B)^\top]$ *of the matrix* $(I+B)$ *is positive definite if*

$$\rho[H(B)] < 1. \tag{1.36}$$

Proof : Let the *spectrum* of A be denoted by $\sigma(A)$ and the *spectral radius* of A be the nonnegative real number $\rho(A) = \max\{|\lambda| : \lambda \in \sigma(A)\}$, then, by using THEOREM 2.4.9 in [23] where $\sigma(A+B) \subseteq \sigma(A) + \sigma(B)$ if A and B commute, therefore, $\sigma(H(I+B)) = \sigma(I+H(B)) \subseteq \sigma(I) + \sigma(H(B)) = \{1\} + \sigma(H(B))$, thus all eigenvalues of $H(I+B)$ are positive if $\rho[H(B)] < 1$.

References

[1] A. K. Fung and M. F. Chen, "Numerical simulation of scattering from simple and composite random surfaces", *J. Opt. Soc. Am.*, A2,pp.2274–2284, (1985).

[2] B. J. Kachoyan and C. Macaskill, "Acoustic scattering from an arbitrarily rough surface", *J. Acoust. Soc. Am.*, 82,pp. 1720–1726, (1987).

[3] E. I. Thorsos, "The validity of the Kirchhoff approximation for rough surface scattering using a Gaussian spectrum", *J. Acoust. Soc. Am.*, 83, pp. 78–92, (1988).

[4] E. I. Thorsos and D. R. Jackson, "Studies of scattering theory using numerical methods", *Waves in Random Media*, 3 S165-S190, (1991).

[5] C. Macaskill and B. J. Kachoyan, "Iterative approach for the numerical simulation of scattering from one- and two- dimensional rough surfaces", *Applied Optics.* Vol. 32, No.15, pp. 2839–2847, (1993).

[6] V. Rokhlin, "Solution of acoustic scattering problems by means of second kind integral equations", *Wave Motion*, 5 pp. 257-272, (1983).

[7] A. F. Peterson and R. Mittra, "Method of conjugate gradient for the numerical solution of large body electromagnetic scattering problems", *J. Opt. Soc. Amer.*, Vol. 2, pp. 971–977, (1985).

[8] A. F. Peterson and R. Mittra, "Convergence of the conjugate gradient method when applied to matrix equations representing electromagnetic scattering problems", *IEEE Trans. Antennas Propagat.*, Vol. AP-34, No. 12, pp. 1447–1454, (1986).

[9] D. J. Wingham and R. H. Devayya, "A note on the use of the Neumann expension in calculating the scatter from rough surfaces", *IEEE Trans. Antennas Propagat.*, Vol. 40, No.5, pp.560–563, May (1992).

[10] T. Sarkar, K. Siarkiewicz and R. F. Stratton, "Survey of numerical methods for solution of large systems of linear equations for electromagnetic field problems", *IEEE Trans. Antennas Propagat.*, Vol.AP-29, No.6, pp. 847–856, (1981).

[11] R. Devayya and D. J. Wingham, "The numerical calculation of rough surface scattering by the conjugate gradient method", *IEEE Trans. Geosci. Remote sensing*, Vol.30 No. 3, pp. 645-648, May (1992).

[12] M. Hestenes, *Conjugate direction methods in optimization*, New York, Springer-Verlag, (1980).

[13] W. C. Meecham, "On the use of the Kirchhoff approximation for the solution of reflection problems", *J. Rat. Mech. Anal.* 5(2), 323–334, (1956).

[14] R. P. Kanwal, *Linear Integral Equations -Theory and Technique*, Academic Press Inc. New York, (1971).

[15] R. Kress, *Linear Integral Equations.* Springer-Verlag Berlin Heidelberg. (1989).

[16] T. Sarkar and S. M. Rao, "The application of the conjugate gradient method for the solution of electromagnetic scattering from arbitrarily oriented wire antennas", *IEEE Trans. Antennas Propagat.*, Vol.AP-32, No.4, pp. 398–403, April (1984).

[17] G. H. Golub and C. F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press. (1989).

[18] H. A. Van der Vorst, "The convergence behaviour of preconditioned CG and CG-S in the presence of rounding errors", in Lecture Notes in Mathmetics, 1457, O. Axelsson and L.Yu. Kolotilina eds., Springer-Verlag, pp. 126–136 Berlin, (1990).

[19] P. Concus and G. H. Golub, "A generalized conjugate gradient method for nonsymmetric systems of linear equations", in Lecture Notes in Economics and Mathematical Systems, 134, R. Glowinski and J.L. Lions, eds., Springer-Verlag, pp. 56–65 Berlin, (1976).

[20] O. Widlund, "A Lanczos method for a class of nonsymmetric systems of linear equations", *SIAM. J. Numer. Anal.*, 15, pp. 801–812, (1978).

[21] O. Axelsson, "Conjugate gradient type methods for unsymmetric and inconsistent systems of linear equations", *Linear Algebra Appl.*, 29, pp. 1–16, (1980).

[22] D. M. Young and K. C. Jea, "Generalized conjugate-gradient Acceleration of nonsymmetrizable iterative methods", *Linear Algebra Appl.*, 34, pp. 159–194, (1980).

[23] R. A. Horn and C. A. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, (1985).

[24] S. C. Eisenstat, H. C. Elman and M. H. Schultz "Variational iterative methods for nonsymmetric systems of linear equations", *SIAM. J. Numer. Anal.*, Vol. 20, No. 2, pp. 345–357, (1983).

[25] D. R. Kincaid and D. M. Young. "A brief review of the ITPACK project", *J. Comp. Appl. Maths.*, 24, pp. 121-127, (1988).

[26] C.C. Paige and M.A. Saunders. "LSQR: An Algorithm for Sparse Linear Equations and Sparse Least Squares", *ACM Transactions on Mathematical Software*, Vol. 8, No. 1, pp. 43-71, March (1982).

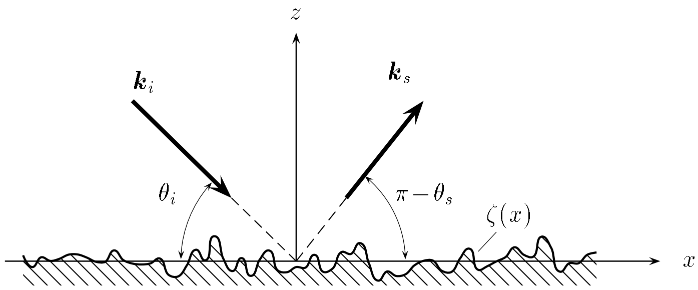

Figure 1: Scattering geometry.

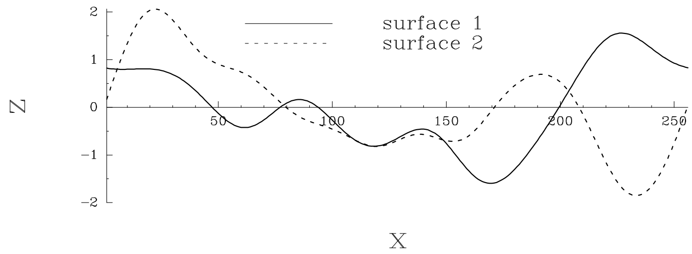

Figure 2: Two 256-point surfaces with a Gaussian correlation function.

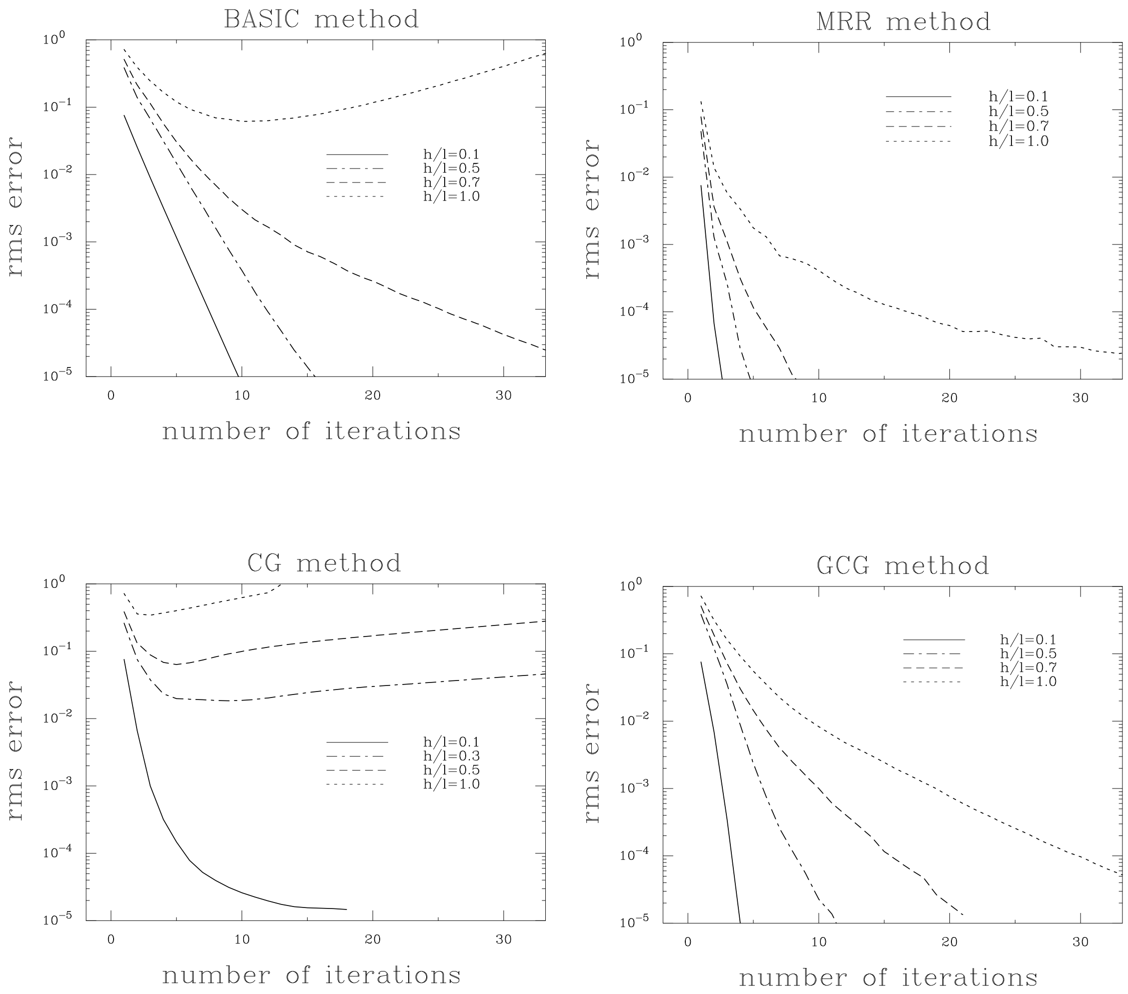

Figure 3: The convergence of four different iterative methods for 100 surface realizations. It shows the decrease of RMS error of the solutions with the number of iterations in different scaled RMS surface height with $kl = 15.0$.

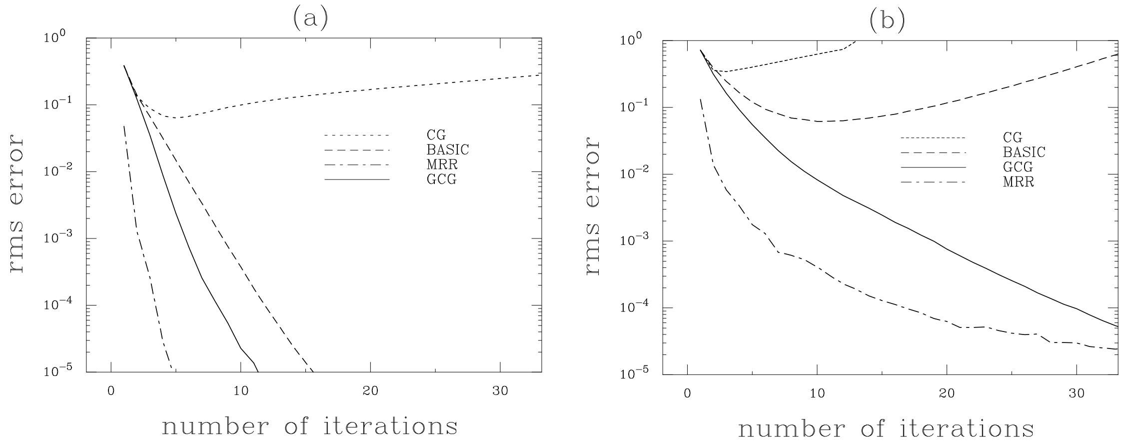

Figure 4: Comparison of convergence by using different methods for the average of 100 different surfaces with (a) $h/l = 0.5$ and (b) $h/l = 1.0$ with $kl = 15.0$ for both cases. Preconditioners were used for all methods here.

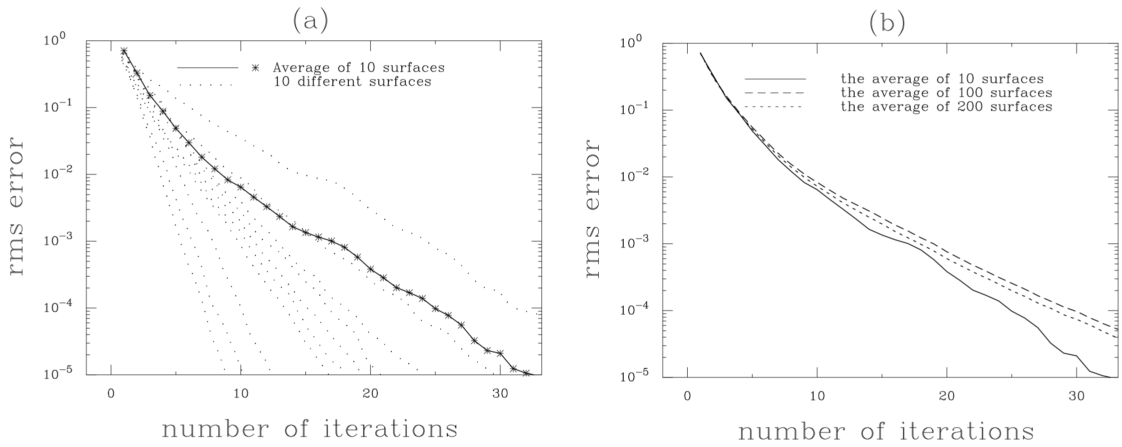

Figure 5: Comparison of convergence by the GCG method for (a) 10 different surfaces; (b) the averages over 10, 100 and 200 surfaces realizations. In both cases $h/l = 1.0$ and $kl = 15.0$.

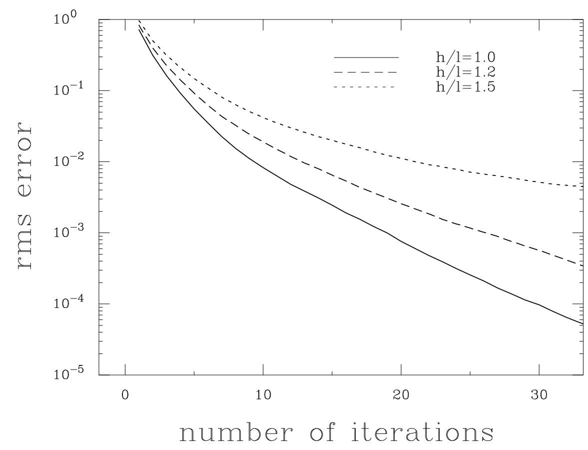

Figure 6: Comparison of convergence by the GCG method for 100 surface realizations with $kl = 15.0$.

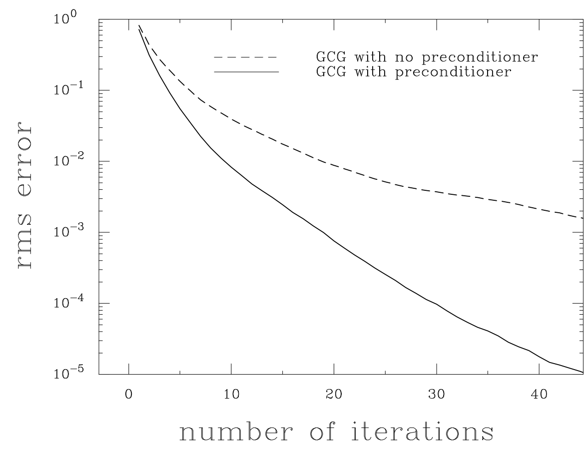

Figure 7: Comparison of using the GCG method with preconditioner and without preconditioner on 100 surface realizations ($h/l = 1.0$ and $kl = 5.0$).

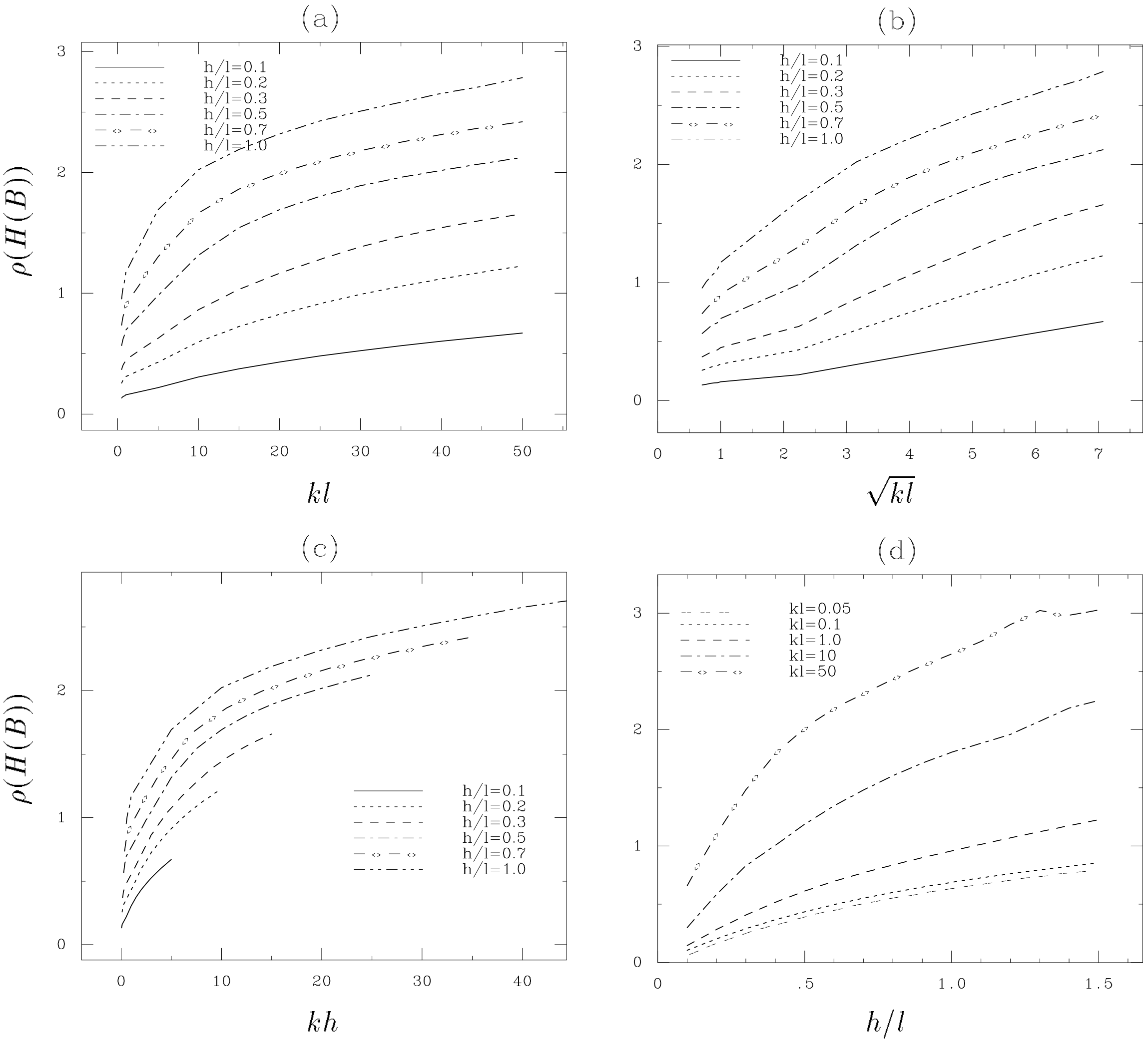

Figure 8: Comparison of spectral radius as function of (a) scaled wave number kl, (b) the square root of the scaled wave number kl, (c) the roughness parameter kh and (d) the RMS surface height h/l.

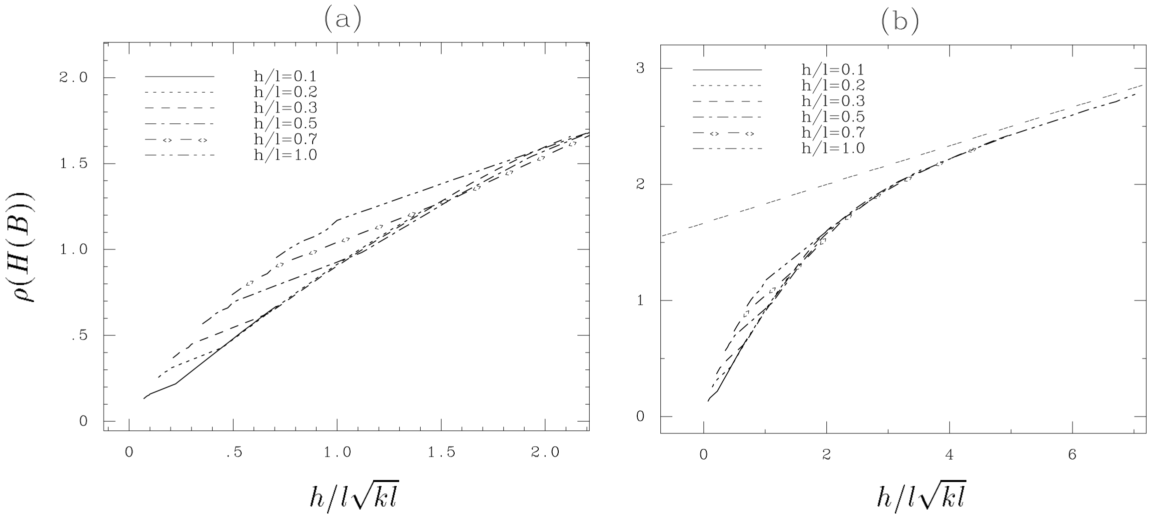

Figure 9:
Plot of the spectral radius against $h/l\sqrt{kl}$.

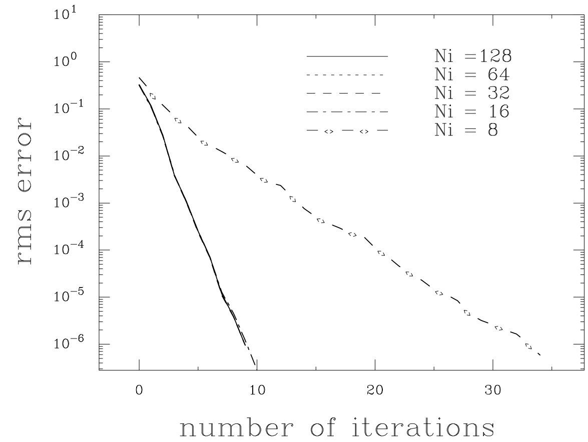

Figure 10: Comparison of convergence for different densities of mesh points on the same surface with $L = 8$, $h/l = 0.5$ and $kl = 25$.

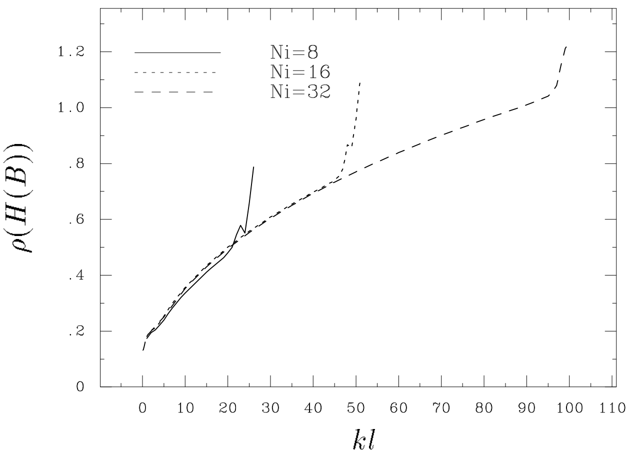

Figure 11: The spectral radius $\rho(H(B))$ increases suddenly at the critical points of kl where the mesh density becomes too small. This is determined for 50 surfaces with $L = 8$ and $h/l = 0.1$ for different mesh densities N_i.

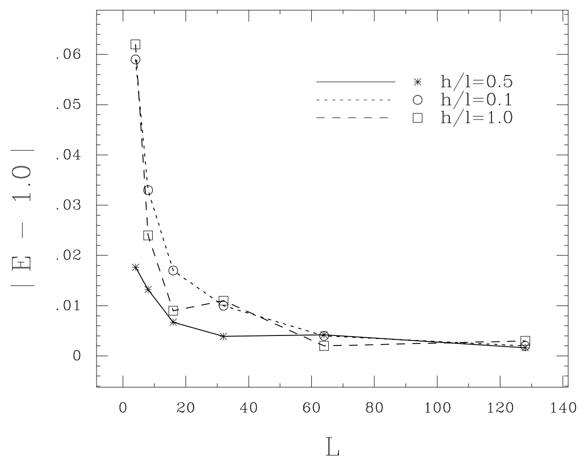

Figure 12: The average error in the reflected energy against the length of the surfaces with $kl = 20$ and $N_i = 32$. The results are averages over 50 surface realizations in each case.

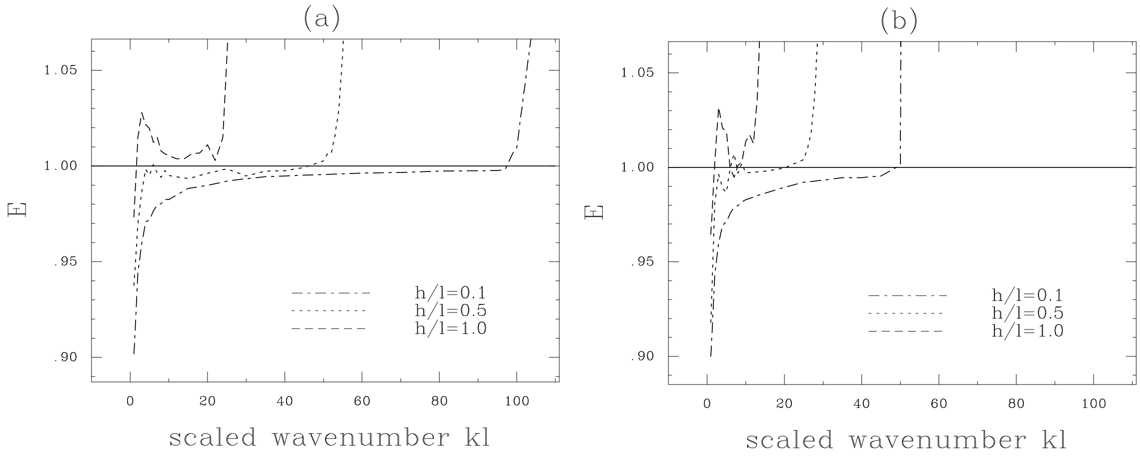

Figure 13: Average reflected energy as a function of the scaled wavenumber kl on 50 surfaces for (a) $L = 32$ and $N_i = 32$; (b) $L = 32$ and $N_i = 16$.

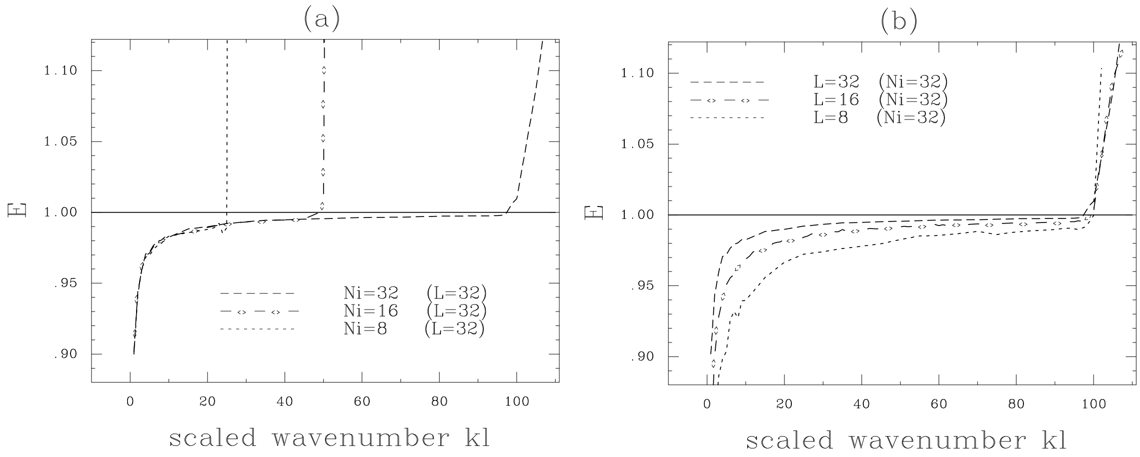

Figure 14: Comparison of the reflected energy on the same 50 surfaces with $h/l = 0.1$ and (a) the same number of correlation lengths but different mesh point density; (b) the same mesh point density but different number surface lengths.

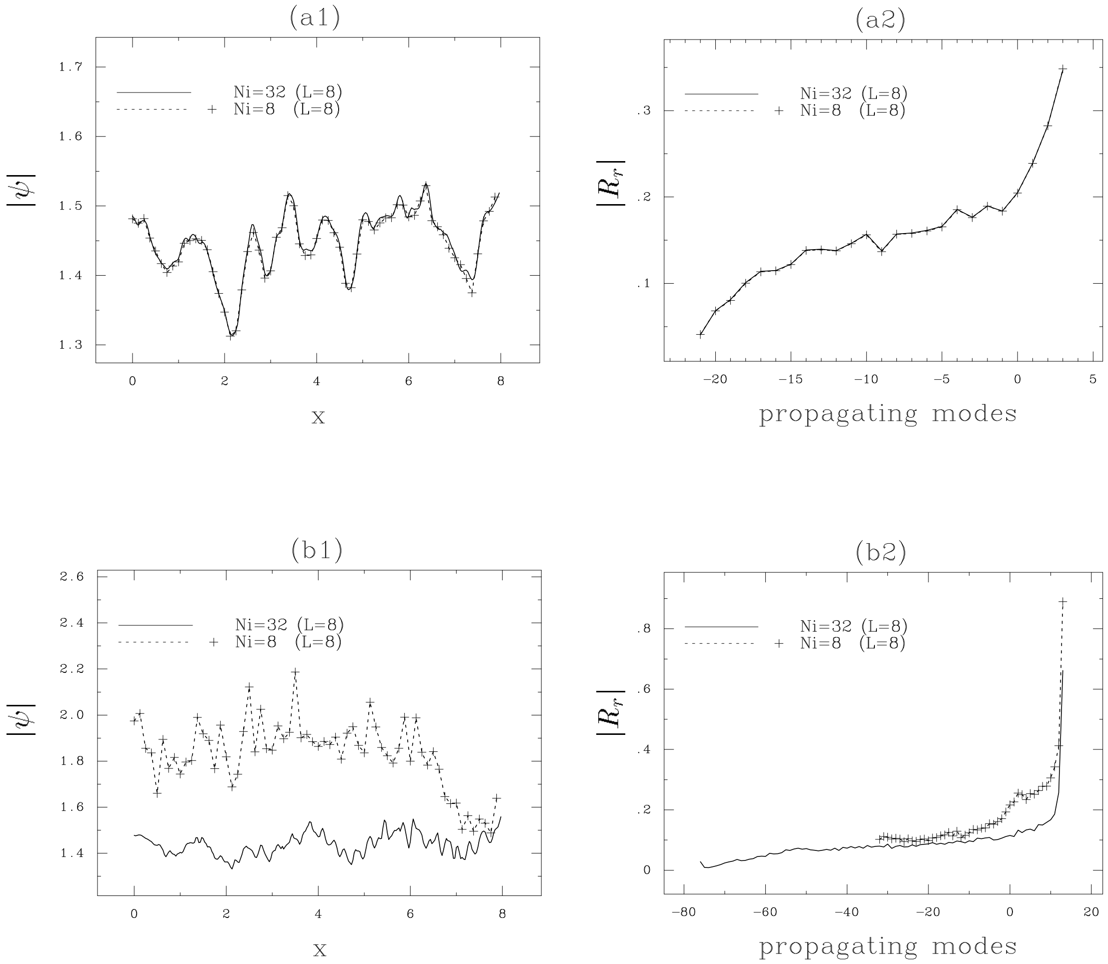

Figure 15: Comparison of $|\psi|$ and $|R_r|$ for different mesh points densities but the same number of correlation length over 50 surfaces with (a) $kl = 10$; (b) $kl = 35$ and both $h/l = 0.3$.

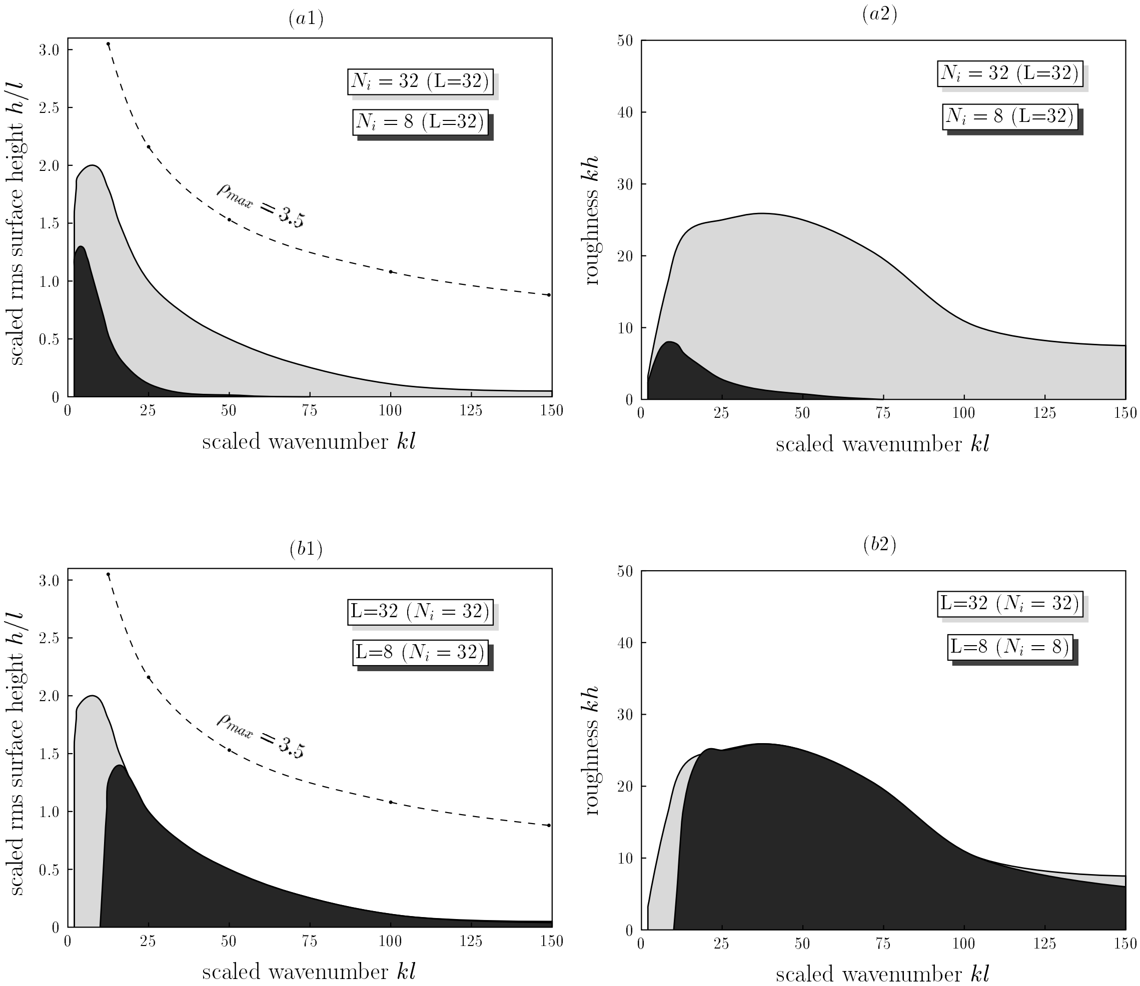

Figure 16: Regions of validity for the GCG method in terms of convergence and energy conservation for cases (a) Same length, different resolution; (b) Same resolution but different length. Here convergence of the iteration means that RMS errors are $< 10^{-7}$. Energy conservation is satisfied in the sense that $|E - 1.0| < 0.05$.

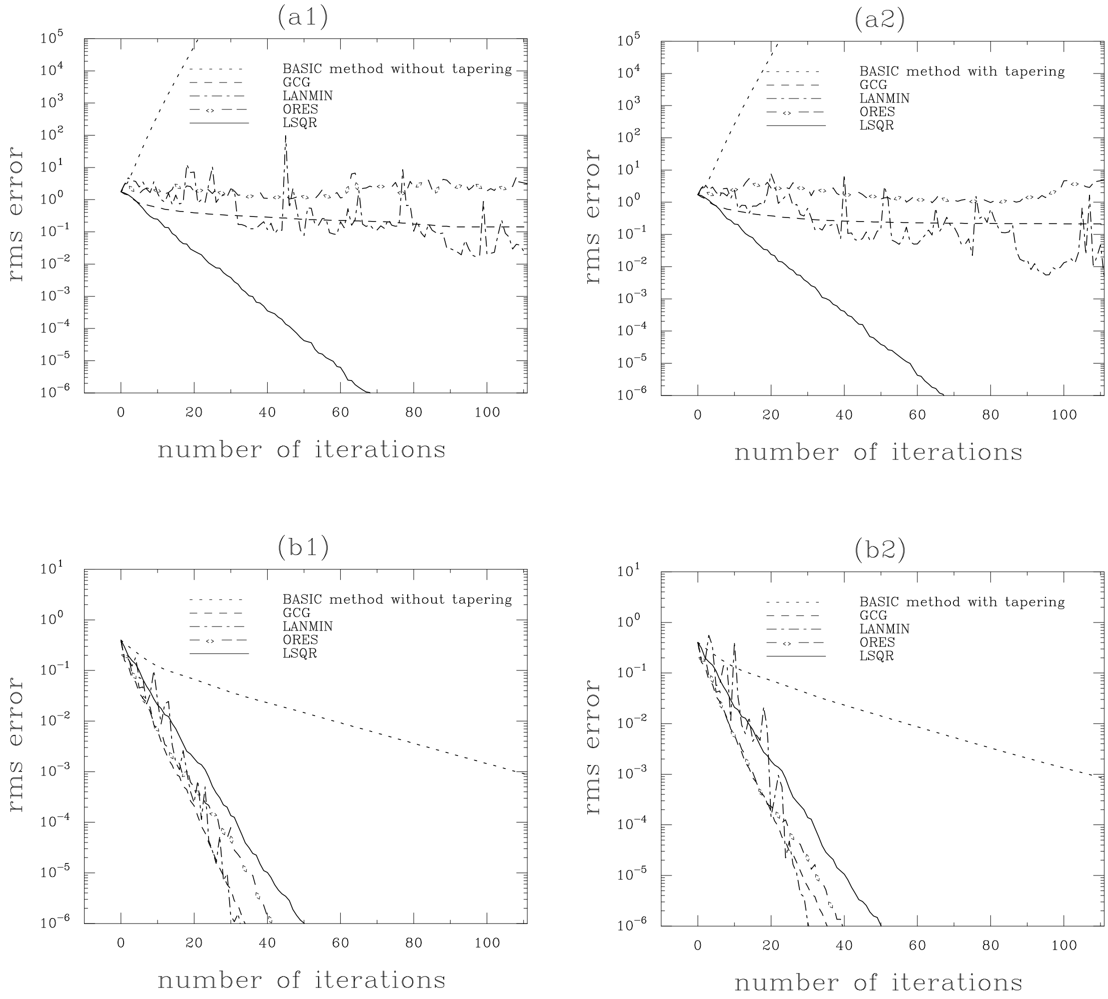

Figure 17: Comparison of the different iterative methods for (a1) the first kind linear system with no tapering and (a2) with tapering; (b1) the second kind without tapering and (b2) with tapering. Both cases have $h/l = 0.235$ and $kl = 5.65$.

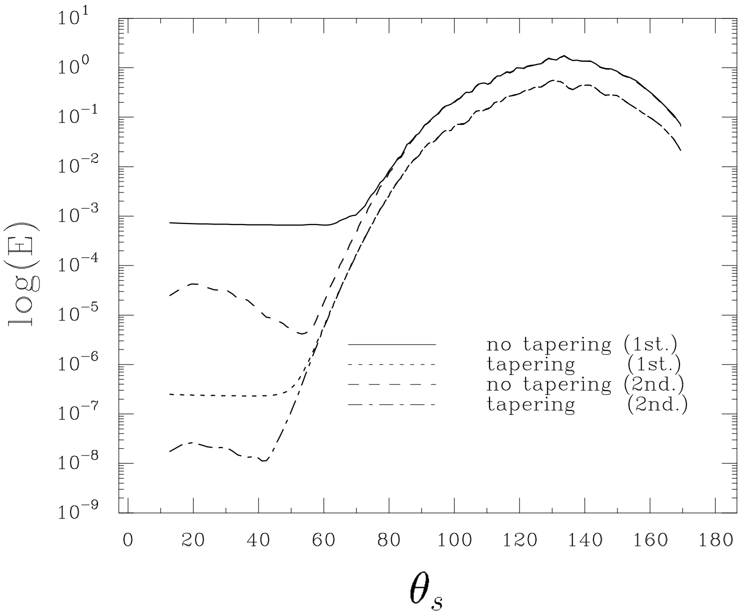

Figure 18: Plot of energy per unit scattered angle against scattered angle θ_s on 50 surface realizations (512-point) with $h/l = 0.1$, $kl = 25$. Comparison is given for the first kind system and the second kind system with no tapering or with tapering. (No attempt has been made to normalize the tapered results.)