2

ITERATIVE SCHEMES BASED ON MINIMIZATION OF A UNIFORM ERROR CRITERION

Peter M. van den Berg

2.1 Introduction
2.2 The Operator Equation
2.3 Direct Minimization of the Error
2.4 Recursive Minimization of the Error
2.5 Selfadjoint Operator LT
2.6 Special Choices for Operator T
2.7 Convergence
2.8 Numerical Results for the Scattering by a Slab
2.9 Numerical Results for the Scattering by a Strip
2.10 Conclusions
References

2.1 Introduction

In the present chapter the iterative solution of the integral equations that are used to formulate electromagnetic, acoustic and elastodynamic scattering problems, is discussed. To have a measure for the accuracy attained, we select the global (i.e., integrated over the domain of the scatterer) root-mean-square error in the equality sign of the integral equation that has to be satisfied by the exact solution. For a given sequence of expansion functions used to represent the unknown field values in the domain of the scatterer, the minimization of the relevant error leads to a particular method of moments (HARRINGTON, 1968). For configurations of realistic size and degree of complexity, this leads to the numerical solution of a large system of linear algebraic equations. In this chapter we develop some iterative techniques, where the intermediate step of the solution of a large system of equations is superfluous. Symmetrization and preconditioning of the integral equation are also discussed. As a preconditioning operator we derive an approximate inverse operator based on our knowledge of the scattering problem at hand. In all these iterative techniques, the integrated square error with respect to the original operator equation is taken as a measure of deviation of the approximate solution from the exact one. Variational techniques are employed to arrive at a minimum error. Our objective of the numerical examples to be presented in this chapter, is to compare the different methods using simple examples with simple numerical discretization techniques rather than employing complex problems with sophisticated numerical discretization techniques. We present numerical results as the tutorial examples of the plane-wave scattering by an (in)homogeneous slab and the plane-wave scattering by a strip. These examples clearly demonstrate the various features of the iterative methods discussed in this chapter.

2.2 The Operator Equation

In this section, we consider the integral equations that arise from the application of the scattering of electromagnetic, acoustic and elastodynamic waves, both in the frequency domain and the time domain. All of these are of the general form

$$\int_{x'\in D} L(x,x')\,u(x')\,dx' = f(x), \quad \text{when } x \in D \tag{2.1}$$

In this equation, u is the unknown field quantity in the relevant contrasting spatial or space-time domain, f is a known field related to the excitation (incident field in scattering problems), and L is the kernel of the integral equation, which is related to the field at x radiated by a source at x'. In general, x and x' stand for the relevant coordinate variables (for example, the Cartesian coordinates $\{x,y,z\}$ in three-dimensional space in the frequency-domain formulation, and $\{x,y,z,t\}$ in the corresponding time-domain formulation); u and f are vector valued, L yields the proper matrix or tensor relationship, and D is the (space or space-time) domain for which (2.1) holds. We further assume that the integral equation has a unique solution, i.e., $u(x)=0$ for all $x\in D$, if and only if $f(x)=0$ for all $x\in D$.

In almost all situations encountered in practice, the integral equation (2.1) can only be solved approximately with the aid of numerical techniques. How good an approximate solution is, can be quantified only after one has chosen a particular quantitative *error*. To discuss this aspect, we introduce an operator formalism together with an inner product of two functions defined on D (with the associated norm). To write (2.1) in an operator form, the (bounded) linear operator L acting on a function $u \in D$ is introduced by

$$Lu = \int_{x' \in D} L(x, x')\, u(x')\, dx' \tag{2.2}$$

Note that the right-hand side of (2.2) is defined for all x. Then, (2.1) is equivalent to

$$Lu = f, \quad \text{when } x \in D \tag{2.3}$$

Further, the inner product of two integrable functions u and v defined on D is taken as the, real or complex, number

$$\langle u, v\rangle = \int_{x \in D} u(x)\, \overline{v}(x)\, dx \tag{2.4}$$

where the overbar denotes complex conjugate. The norm of a function u is, in accordance with (2.4), defined as

$$\|u\| = \langle u, u\rangle^{\frac{1}{2}} \tag{2.5}$$

The (Hermitean) adjoint operator $L^\star$ of L is defined as that one for which

$$\langle Lu, v\rangle = \langle u, L^\star v\rangle \tag{2.6}$$

for all functions u and v defined on D. If $L^\star = L$, the operator is selfadjoint. It is noted that in most scattering problems the operator is, however, not selfadjoint. Combining (2.2) with (2.6) it follows that

$$L^\star v = \int_{x' \in D} L^\star(x', x)\, v(x')\, dx' \tag{2.7}$$

where the * at the matrix kernel $L(x', x)$ denotes the complex conjugate of the transpose. Note that in the kernel of (2.7) the coordinates x and x' have the reverse order of the ones in (2.2).

For any function u^{app} differing from the exact solution u of (2.3) we define the residual as

$$r = f - Lu^{app} \tag{2.8}$$

and the global root-mean-square error in the satisfaction of the equality sign in (2.3) as

$$\text{ERR} = \langle r, r\rangle^{\frac{1}{2}} = \|r\| \tag{2.9}$$

being the norm of r. Note that ERR ≥ 0 and that ERR $= 0$ if and only if $u^{app} = u$. In scattering problems f is related to the exciting field in the domain D; we therefore often normalize the root-mean-square error according to

$$\widehat{\text{ERR}} = \frac{\|r\|}{\|f\|} \tag{2.10}$$

with the properties $\widehat{\text{ERR}} = 0$ if $u^{app} = u$ and $\widehat{\text{ERR}} = 1$ if $u^{app} = 0$. The error defined in (2.8) and (2.9) is used as a measure for the accuracy attained in all the various iterative schemes to be dealt with.

2.3 Direct Minimization of the Error

In this section we first discuss a direct (i.e., non-iterative) approximation to the solution of the operator equation (2.3). To construct an approximate solution, the unknown function u is expanded in terms of a given, appropriately chosen, sequence of linearly independent expansion functions $\{\phi_n;\ n = 1, \cdots, N\}$ that are defined on D and belong to the same vector space to which u belongs. Let, for some $N \geq 1$,

$$u_N = \sum_{n=1}^{N} \alpha_n^{(N)} \phi_n \tag{2.11}$$

and

$$r_N = f - Lu_N \tag{2.12}$$

Then the problem is to determine, for given N, the sequence of expansion coefficients $\{\alpha_n^{(N)};\ n = 1, \cdots, N\}$ such that $\langle r_N, r_N\rangle$ is minimized. The relevant values of $\{\alpha_n^{(N)}\}$ are denoted as the optimum values $\{\alpha_n^{opt}\}$. Assuming that the optimum exists, let

$$\alpha_n^{(N)} = \alpha_n^{opt} + \delta\alpha_n \quad \text{for } n = 1, \cdots, N \tag{2.13}$$

where $\delta\alpha_n$ is arbitrary. Further, let

$$r_N^{opt} = f - \sum_{n=1}^{N} \alpha_n^{opt} L\phi_n \tag{2.14}$$

then

$$\begin{aligned}\langle r_N, r_N\rangle =&\langle r_N^{opt}, r_N^{opt}\rangle - 2\mathrm{Re}\Big(\sum_{m=1}^{N} \overline{\delta\alpha_m}\langle r_N^{opt}, L\phi_m\rangle\Big) \\ &+ \Big\langle\sum_{n=1}^{N} \delta\alpha_n L\phi_n, \sum_{m=1}^{N} \delta\alpha_m L\phi_m\Big\rangle\end{aligned} \tag{2.15}$$

Now, the last term on the right-hand side of this equation is always positive if $\{\delta\alpha_1, \cdots, \delta\alpha_N\} \neq \{0, \cdots, 0\}$. Hence, if

$$\langle r_N^{opt}, L\phi_m\rangle = 0 \quad \text{for } m = 1, \cdots, N \tag{2.16}$$

we have constructed the situation that for $\{\delta\alpha_1, \cdots, \delta\alpha_N\} = \{0, \cdots, 0\}$ the quantity $\langle r_N^{opt}, r_N^{opt}\rangle$ is the absolute minimum of $\langle r_N, r_N\rangle$. Substitution of (2.14) in (2.16) yields the system of linear algebraic equations

$$\sum_{n=1}^{N} \alpha_n^{opt}\langle L\phi_n, L\phi_m\rangle = \langle f, L\phi_m\rangle \quad \text{for } m = 1, \cdots, N \tag{2.17}$$

to be solved for $\{\alpha_n^{opt}\}$. From (2.16) and (2.11) it also follows that

$$\langle r_N^{opt}, Lu_N\rangle = 0 \tag{2.18}$$

The resulting value of the error is

$$\mathrm{ERR}_N = \langle r_N^{opt}, r_N^{opt}\rangle^{\frac{1}{2}} = \langle r_N^{opt}, f\rangle^{\frac{1}{2}} \tag{2.19}$$

If this value does not meet the accuracy requirements set for the solution of the operator equation, it can be reduced either by selecting a more appropriate sequence of expansion functions (which is difficult to realize in practice) or by increasing N. Note that (2.19) would also result from the application of the method of moments (HARRINGTON, 1968), provided that the sequence of testing functions is chosen as $\{\bar{L}\phi_m;\ m = 1, \cdots, N\}$ when the sequence of expansion functions is $\{\phi_n;\ n = 1, \cdots, N\}$. This choice gives the best result in the integrated-square-error sense. For problems of realistic size and complexity the value N soon becomes so large that the storage requirements exceed the capacity of even present-day large computer systems. The problem of excessive computation time and computer storage requirements for a direct numerical solution (e.g., by Gauss elimination) of the system of equations can be circumvented by using suitable iterative techniques. Another argument in favor of solving the pertinent system of equations iteratively is the evident fact that we should not solve the system of equations to a higher degree of accuracy than is needed.

2.4 Recursive Minimization of the Error

In this section we develop a recursive method for calculating the approximate solution to the operator equation (2.3). In a direct procedure for solving such an equation approximately a (finite) sequence of expansion functions is somehow selected beforehand, and the sequence of expansion coefficients is solved from a system of linear algebraic equations that follows from – in our case – minimizing the norm of the error in the residual. In an iterative procedure the elements of the sequence of expansion functions are recursively generated from the operator equation to be solved, one in each iteration step. To achieve this, the successive residuals in the iteration process are at one's disposal. The sequence of expansion coefficients grows with the number of iterations. Since at the N-th step, N presumably linearly independent expansion functions have been generated, N expansion coefficients are available to represent the N-th approximation to the solution of the operator equation. Here, too, the minimization of the norm of the residual at the N-th step will be employed to generate the system of linear algebraic equations that the sequence of expansion coefficients must satisfy.

Let u_N denote the N-th approximation to the solution of the operator equation

$$Lu = f, \quad \text{for } x \in D \tag{2.20}$$

and let $\{\phi_n;\ n = 1, \cdots, N\}$ be the recursively generated sequence of expansion functions. Then, we take

$$u_0 = 0$$

$$u_N = u_{N-1} + u_N^{cor} \quad \text{for } N = 1, \cdots \tag{2.21}$$

where u_N^{cor} is the correction to u_{N-1} to arrive at u_N. The correction is now expressed as

$$u_N^{cor} = \sum_{n=1}^{N} \alpha_n^{(N)} \phi_n \quad \text{for } N = 1, \cdots \tag{2.22}$$

where $\{\alpha_n^{(N)};\ n = 1, \cdots, N\}$ is the sequence of expansion coefficients of u_N^{cor}. The residuals are found as

$$\begin{aligned} r_0 &= f \\ r_N &= f - Lu_N \quad \text{for } N = 1, \cdots \end{aligned} \tag{2.23}$$

From (2.23) it follows that

$$r_N = r_{N-1} - Lu_N^{cor} \tag{2.24}$$

Substitution of (2.21) - (2.22) in (2.24) and taking into account that at the N-th step r_{N-1} is known, minimization of the norm of r_N leads to a system of linear algebraic equations in the N expansion coefficients $\{\alpha_n^{(N)};\ n = 1, \cdots, N\}$. On account of (2.16) this system of equations follows from

$$\langle r_N, L\phi_m \rangle = 0 \quad \text{for } m = 1, \cdots, N \tag{2.25}$$

Substitution of (2.24) and (2.22) in these equations leads to

$$\sum_{n=1}^{N} \alpha_n^{(N)} \langle L\phi_n, L\phi_m \rangle = \langle r_{N-1}, L\phi_m \rangle \quad \text{for } m = 1, \cdots, N \tag{2.26}$$

Now, on account of (2.25) only the right-hand side of (2.26) for $m = N$ may differ from zero. For non-zero values of the coefficients $\{\alpha_n^N;\ n = 1, \cdots, N\}$, this should be the case and hence

$$\langle r_{N-1}, L\phi_N \rangle \neq 0 \tag{2.27}$$

which is denoted as the *improvement condition.* If (2.27) is satisfied, the coefficients $\{\alpha_n^N;\ n = 1, \cdots, N\}$ can be solved from (2.26).

First of all it is observed that the vanishing of the right-hand sides in (2.26) for $m = 1, \cdots, N-1$ entails the property that all $\alpha_n^{(N)}$ for $n = 1, \cdots, N-1$ are proportional to $\alpha_N^{(N)}$. In view of this, we introduce the sequence of functions $\{\psi_N;\ N = 1, \cdots\}$ as

$$\psi_N = \frac{u_N^{cor}}{\alpha_N^{(N)}} \text{ for } N = 1, \cdots \tag{2.28}$$

or (cf. (2.22))

$$\begin{aligned} \psi_1 &= \phi_1 \ , \\ \psi_N &= \phi_N + \sum_{n=1}^{N-1} \frac{\alpha_n^{(N)}}{\alpha_N^{(N)}} \phi_n \text{ for } N = 2, \cdots \end{aligned} \tag{2.29}$$

Evidently, (2.29) expresses ψ_N as a linear combination of $\{\phi_n;\ n = 1, \cdots, N\}$. Since the reverse is also true, ϕ_N can be expressed as a linear combination of $\{\psi_n;\ n = 1, \cdots, N\}$. In view of this, (2.29) can be rewritten as

$$\begin{aligned} \psi_1 &= \phi_1 \ , \\ \psi_N &= \phi_N + \sum_{n=1}^{N-1} \beta_n^{(N)} \psi_n \text{ for } N = 2, \cdots \end{aligned} \tag{2.30}$$

Owing to the orthogonality properties of $\{L\psi_N\}$ to be discussed below, the coefficients $\{\beta_n^{(N)};\ n = 1, \cdots, N-1\}$ can readily be determined. Since any ψ_N is a linear combination of the expansion functions $\{\phi_n;\ n = 1, \cdots, N\}$, (2.25) leads to

$$\langle r_N, L\psi_m \rangle = 0 \text{ for } m = 1, \cdots, N \tag{2.31}$$

Using (2.24) in (2.31), we arrive at the orthogonality relation

$$\langle L\psi_N, L\psi_m \rangle = 0 \text{ for } m = 1, \cdots, N-1 \tag{2.32}$$

Since (2.32) holds for any $N = 2, \cdots$, the sequence of $\{\psi_n\}$ satisfies the orthogonality relationship

$$\langle L\psi_n, L\psi_m \rangle = 0 \, , \quad m \neq n \tag{2.33}$$

Combining (2.33) with (2.30) we obtain

$$\beta_n^{(N)} = -\frac{\langle L\phi_N, L\psi_n\rangle}{||L\psi_n||^2} \text{ for } n = 1, \cdots, N-1 \tag{2.34}$$

Through substitution of (2.34) in (2.30), the sequence of $\{\psi_N;\ N = 1, \cdots\}$ has been constructed.

The value of u_N^{cor} finally follows from the consideration that

$$u_N^{cor} = \alpha_N^{(N)}\psi_N \text{ for } N = 1, \cdots \tag{2.35}$$

which leads to

$$\langle Lu_N^{cor}, L\psi_N\rangle = \alpha_N^{(N)}\langle L\psi_N, L\psi_N\rangle \tag{2.36}$$

However (cf. (2.24)),

$$\langle Lu_N^{cor}, L\psi_N\rangle = \langle r_{N-1} - r_N, L\psi_N\rangle \tag{2.37}$$

Observing that $L\psi_N$ is a linear combination of $\{L\phi_m;\ m = 1, \cdots, N\}$, the application of (2.25) leads to the result

$$\langle Lu_N^{cor}, L\psi_N\rangle = \langle r_{N-1}, L\psi_N\rangle = \langle r_{N-1}, L\phi_N\rangle \tag{2.38}$$

Combining (2.38) with (2.36) we arrive at

$$\alpha_N^{(N)} = \frac{\langle r_{N-1}, L\phi_N\rangle}{||L\psi_N||^2} \tag{2.39}$$

With this, the determination of u_N^{cor} has been completed and the iterative scheme based on error minimization has been defined.

More specifically we consider the case that the function ϕ_N that is generated at the N-th step of iteration is linearly related to the residual r_{N-1} at the previous step. Then,

$$\phi_N = Tr_{N-1} \text{ for } N = 1, \cdots \tag{2.40}$$

where T is a bounded linear operator on D. Then the following computational scheme is arrived at.

Computational Scheme for an Arbitrary Operator T

The scheme starts with the initial values

$$u_0 = 0, \;\; r_0 = f, \;\; \mathrm{ERR}_0 = ||f|| \tag{2.41}$$

Next, the scheme puts

$$\begin{aligned} \psi_1 &= Tr_0 \\ B_1 &= ||L\psi_1||^2 \\ \alpha_1^{(1)} &= \frac{\langle r_0, LTr_0\rangle}{B_1} \\ u_1 &= u_0 + \alpha_1^{(1)}\psi_1 \\ r_1 &= r_0 - \alpha_1^{(1)}L\psi_1 \\ \mathrm{ERR}_1 &= ||r_1|| \end{aligned} \tag{2.42}$$

and computes successively for $N = 2, \cdots$,

$$\begin{aligned} \beta_n^{(N)} &= -\frac{\langle LTr_{N-1}, L\psi_n\rangle}{B_n} \text{ for } n = 1, \cdots, N-1 \\ \psi_N &= Tr_{N-1} + \sum_{n=1}^{N-1} \beta_n^{(N)}\psi_n \\ B_N &= ||L\psi_N||^2 \\ \alpha_N^{(N)} &= \frac{\langle r_{N-1}, LTr_{N-1}\rangle}{B_N} \\ u_N &= u_{N-1} + \alpha_N^{(N)}\psi_N \\ r_N &= f - Lu_N = r_{N-1} - \alpha_N^{(N)}L\psi_N \\ \mathrm{ERR}_N &= ||r_N|| \end{aligned} \tag{2.43}$$

The important orthogonality relations that hold are:

$$\begin{aligned} \langle L\psi_n, L\psi_m\rangle &= 0 \text{ for } m \neq n \\ \langle LTr_n, LTr_m\rangle &= 0 \text{ for } m \neq n \\ \langle r_N, L\psi_m\rangle &= 0 \text{ for } m = 1, \cdots, N \\ \langle r_N, LTr_m\rangle &= 0 \text{ for } m = 0, \cdots, N-1 \end{aligned} \tag{2.44}$$

In this scheme, for each $N = 1, \cdots$, the values of ψ_N, $L\psi_N$ and B_N are stored. This means that at the N-th step of iteration, we need computer storage for the updated values u_N and r_N, as well as some background storage for the values of ψ_m, $L\psi_m$ and B_m, for $m = 1, \cdots, N$. The computation time and computer storage required for each step of iteration increases with an increasing number of iterations.

2.5 Selfadjoint Operator LT

In this section we now investigate the consequences to the scheme of the previous section in case the operator LT is selfadjoint. For such operators the property

$$\langle r_{N-1}, LTr_{N-1}\rangle = \langle LTr_{N-1}, r_{N-1}\rangle \tag{2.45}$$

holds. Then, the last orthogonality relation of (2.44) can be written as

$$\langle r_n, LTr_m\rangle = \langle LTr_n, r_m\rangle = 0 \text{ for } m \neq n \tag{2.46}$$

Since, the quantity in (2.45) is real-valued, it follows that (cf. (2.39) and (2.40))

$$\alpha_N^{(N)} = \frac{\langle r_{N-1}, LTr_{N-1}\rangle}{\|L\psi_N\|^2} \tag{2.47}$$

is also a real quantity. Further, from (2.24) and (2.28) we have

$$L\psi_n = -\frac{r_n - r_{n-1}}{\alpha_n^{(n)}} \text{ for } n = 1, \cdots, N \tag{2.48}$$

Using (2.47), (2.48) and (2.40) in (2.34) the expression for $\beta_n^{(N)}$ becomes

$$\beta_n^{(N)} = \frac{\langle LTr_{N-1}, r_n\rangle - \langle LTr_{N-1}, r_{n-1}\rangle}{\langle r_{n-1}, LTr_{n-1}\rangle} \text{ for } n = 1, \cdots, N-1 \tag{2.49}$$

Taking into account the orthogonality relations of (2.46), we arrive at

$$\beta_n^{(N)} = \begin{cases} 0 & \text{for } n = 1, \cdots, N-2 \\ \frac{\langle r_{N-1}, LTr_{N-1}\rangle}{\langle r_{N-2}, LTr_{N-2}\rangle} & \text{for } n = N-1 \end{cases} \tag{2.50}$$

Hence, only $\beta_{N-1}^{(N)}$ differs from zero and has to be determined.

Computational Scheme for a Selfadjoint Operator LT

The scheme starts with the initial values

$$u_0 = 0, \quad r_0 = f, \quad \mathrm{ERR}_0 = \|f\| \tag{2.51}$$

Next, the scheme puts

$$\begin{aligned} A_1 &= \langle r_0, LTr_0 \rangle \\ \psi_1 &= Tr_0 \\ B_1 &= \|L\psi_1\|^2 \\ u_1 &= u_0 + \frac{A_1}{B_1}\psi_1 \\ r_1 &= r_0 - \frac{A_1}{B_1}L\psi_1 \\ \mathrm{ERR}_1 &= \|r_1\| \end{aligned} \tag{2.52}$$

and computes successively for $N = 2, \cdots$,

$$\begin{aligned} A_N &= \langle r_{N-1}, LTr_{N-1} \rangle \\ \psi_N &= Tr_{N-1} + \frac{A_N}{A_{N-1}}\psi_{N-1} \\ B_N &= \|L\psi_N\|^2 \\ u_N &= u_{N-1} + \frac{A_N}{B_N}\psi_N \\ r_N &= f - Lu_N = r_{N-1} - \frac{A_N}{B_N}L\psi_N \\ \mathrm{ERR}_N &= \|r_N\| \end{aligned} \tag{2.53}$$

In this scheme, we need computer storage for the updated values u_N , ψ_N , A_N and r_N . The computation time and computer storage required for each step of iteration remain the same for all iterations $N = 2, \cdots$ This scheme is equivalent to one of the conjugate-gradient schemes in the literature (DANIEL, 1967).

2.6 Special Choices for Operator T

In this section we now investigate the consequences of some particular choices of the operator T .

Residuals as Expansion Functions $(T = I)$

First, we take the residual of the previous step as a particular choice for the expansion function ϕ_N in the recursive minimization scheme of Section 2.4, i.e.,

$$\phi_N = r_{N-1} \text{ for } n = 1, \cdots \tag{2.54}$$

This is equivalent to setting the operator T introduced in Section 2.4 equal to the identity operator, i.e.,

$$T = I \tag{2.55}$$

The improvement condition of (2.27) is then replaced by

$$\langle r_{N-1}, Lr_{N-1} \rangle \neq 0 \tag{2.56}$$

Whether or not this condition is satisfied, depends on the particular form of the operator L under consideration. The relevant iteration scheme now follows by replacing the operator T by the identity operator I either in (2.41) - (2.43) for non-selfadjoint operators L or in (2.51) - (2.53) for selfadjoint, not necessarily positive, operators L. The latter scheme differs slightly from the standard conjugate-gradient schemes for selfadjoint, positive operators ($\langle Lu, u \rangle > 0$ for all $u \neq 0$ on D) that are given in the literature (HESTENES and STIEFEL, 1952), where the quantity $\langle Lu, u \rangle - \langle f, u \rangle - \langle u, f \rangle$ is minimized.

Preconditioning $(T = P)$

We first observe that, if T is chosen equal to L^{-1}, the inverse of L, then $\phi_1 = \psi_1 = L^{-1}f$ and the recursive scheme of Section 2.4 will terminate in the first iteration; we then have arrived at the exact solution. By this reasoning we take T equal to a suitably chosen preconditioning operator P, where the operator LP more closely resembles the identity operator than L itself does. Thus, the method depends on the availability of an approximate inverse to the operator L. Accordingly, we take

$$\phi_N = Pr_{N-1} \tag{2.57}$$

or

$$T = P \tag{2.58}$$

The relevant iteration scheme now follows by replacing the operator T by the preconditioning operator P either in (2.41) - (2.43) for non-selfadjoint operators LP or in (2.51) - (2.53) for selfadjoint operators LP.

Symmetrization $(T = L^\star)$

When the operator L is not selfadjoint, the previous scheme for $T = I$ leads to a recursive minimization scheme, in which the computer storage of the expansion functions required for each step of iteration increases with an increasing number of iterations. However, when we take

$$\phi_N = L^\star r_{N-1} \tag{2.59}$$

or

$$T = L^\star \tag{2.60}$$

the operator

$$LT = LL^\star = (LL^\star)^\star \tag{2.61}$$

is selfadjoint and we can now use the simple iteration scheme of (2.51) - (2.53) for selfadjoint operators LT with T replaced by $L^\star$. Since

$$\langle r_{N-1}, L\phi_N\rangle = \langle L^\star r_{N-1}, \phi_N\rangle = \langle \phi_N, \phi_N\rangle = \langle L^\star r_{N-1}, L^\star r_{N-1}\rangle \neq 0 \tag{2.62}$$

the improvement condition (cf. (2.27)) is automatically satisfied and the orthogonality relation of (2.46) simplifies to

$$\langle L^\star r_{m-1}, L^\star r_{n-1}\rangle = \langle \phi_m, \phi_n\rangle = 0 \text{ for } m \neq n \tag{2.63}$$

Hence, the expansion functions generated according to (2.59) form an orthogonal sequence.

The scheme of (2.51) - (2.53), with T replaced by $L^\star$ and the expression for A_N replaced by

$$A_N = \|L^\star r_{N-1}\|^2 \text{ for } N = 1, \cdots \tag{2.64}$$

is known as the conjugate-gradient scheme for a non-selfadjoint operator L (VAN DEN BERG, 1984).

Preconditioning and Symmetrization ($T = PP^\star L^\star$)

When the operator LP is not selfadjoint, the scheme for $T = P$ leads to a recursive minimization scheme, in which the computer storage of the expansion functions required for each step of iteration increases with an increasing number of iterations. However, when we take

$$\phi_N = PP^\star L^\star r_{N-1} \tag{2.65}$$

or

$$T = PP^\star L^\star \tag{2.66}$$

the operator

$$LT = LPP^\star L^\star = (LP)(LP)^\star \tag{2.67}$$

is selfadjoint and we can now use the simple iteration scheme of (2.51) - (2.53) for selfadjoint operators LT with T replaced by $PP^\star L^\star$ and the expression for A_N replaced by

$$A_N = \|P^\star L^\star r_{N-1}\|^2 \text{ for } N = 1,\cdots \tag{2.68}$$

is a conjugate-gradient scheme for a preconditioned non-selfadjoint operator L. Note that the error criterion applies to the original operator equation. This differs from standard preconditioned conjugate-gradient schemes, where the error is minimized in the range of the preconditioned operator equation.

2.7 Convergence

In this section, we investigate the convergence properties of the different iterative schemes. For this goal we first define some properties of our operators.

The norm of the operator L is defined as

$$\|L\| = \sup_{u \neq 0} \frac{\|Lu\|}{\|u\|} \text{ for all } u \in D \tag{2.69}$$

with the consequence that

$$\|Lu\| \leq \|L\|\,\|u\| \text{ for all } u \in D \tag{2.70}$$

Using (2.69), the norm of L^{-1}, the (bounded) inverse operator of L, is given by

$$||L^{-1}|| = \sup_{v \neq 0} \frac{||L^{-1}v||}{||v||} \text{ for all } v \in D \tag{2.71}$$

Taking $v = Lu$, it follows that

$$||L^{-1}|| = \sup_{u \neq 0} \frac{||u||}{||Lu||} \text{ for all } u \in D \tag{2.72}$$

with the consequence that

$$||Lu|| \geq \frac{||u||}{||L^{-1}||} \text{ for all } u \in D \tag{2.73}$$

Combining (2.70) and (2.73), we arrive at

$$0 < \frac{||u||}{||L^{-1}||} \leq ||Lu|| \leq ||L|| \, ||u|| < \infty \tag{2.74}$$

for all non-zero $u \in D$. The leftmost inequality is a consequence of the assumption of the boundedness of the operator L^{-1}, while the rightmost inequality is a consequence of the assumption of the boundedness of the operator L. The norm of the operator L^{*} is, using (2.69), found as

$$||L^{*}|| = \sup_{v \neq 0} \frac{||L^{*}v||}{||v||} \text{ for all } v \in D \tag{2.75}$$

It can be shown that (KREYSZIG, 1978, pp. 196 - 200)

$$||L^{*}|| = ||L|| \tag{2.76}$$

and

$$||L^{*}L|| = ||LL^{*}|| = ||L||^2 \tag{2.77}$$

In order to investigate the convergence of our recursive scheme of Section 2.4 we consider the quantity $\mathrm{ERR}_N^2 = \langle r_N, r_N \rangle$. Using (2.24) and (2.35) we obtain

$$\langle r_N, r_N \rangle = \langle r_N, r_{N-1} - \alpha_N^{(N)} L\psi_N \rangle = \langle r_N, r_{N-1} \rangle \tag{2.78}$$

on account of (2.31). Again using (2.24) and (2.35) we obtain

$$\langle r_N, r_N\rangle = \langle r_{N-1} - \alpha_N^{(N)} L\psi_N, r_{N-1}\rangle = \langle r_{N-1}, r_{N-1}\rangle - \alpha_N^{(N)}\langle L\phi_N, r_{N-1}\rangle \tag{2.79}$$

where (2.38) has been used as well. With the expression for $\alpha_N^{(N)}$ of (2.39) we obtain the result

$$\langle r_N, r_N\rangle = \langle r_{N-1}, r_{N-1}\rangle - \frac{|\langle r_{N-1}, L\phi_N\rangle|^2}{\|L\psi_N\|^2} \tag{2.80}$$

which again shows that $\langle r_N, r_N\rangle < \langle r_{N-1}, r_{N-1}\rangle$ and hence $\mathrm{ERR}_N < \mathrm{ERR}_{N-1}$, provided that the improvement condition of (2.37) is satisfied. Let us subsequently consider the expression for the norm of $L\psi_N$. Using (2.30) and the orthogonality relations of (2.33) it follows that

$$\begin{aligned} \|L\psi_N\|^2 &= \langle L\psi_N, L\phi_N\rangle \\ &= \langle L\phi_N, L\phi_N\rangle + \sum_{n=1}^{N-1} \beta_n^{(N)}\langle L\psi_n, L\phi_N\rangle \end{aligned} \tag{2.81}$$

Substitution of the expression of (2.34) for $\beta_n^{(N)}$ yields

$$\begin{aligned} \|L\psi_N\|^2 &= \langle L\phi_N, L\phi_N\rangle - \sum_{n=1}^{N-1} \frac{|\langle L\psi_n, L\phi_N\rangle|^2}{\|L\psi_n\|^2} \\ &\leq \|L\phi_N\|^2 \end{aligned} \tag{2.82}$$

Using this result in (2.80), we obtain the inequality

$$\|r_N\|^2 \leq \|r_{N-1}\|^2 - \frac{|\langle r_{N-1}, L\phi_N\rangle|^2}{\|L\phi_N\|^2} \tag{2.83}$$

More specifically we shall now investigate the case that $\phi_N = Tr_{N-1}$ (cf. (2.40)). The inequality of (2.83) can then be written as

$$\|r_N\|^2 \leq \|r_{N-1}\|^2 - \frac{|\langle r_{N-1}, LTr_{N-1}\rangle|^2}{\|LTr_{N-1}\|^2} \tag{2.84}$$

For a certain class of bounded operators L and T, convergence of the iteration scheme can be proved. The relevant class is characterized by the property that there exists a constant $c \neq 0$ such that

$$|\langle r_{N-1}, LTr_{N-1}\rangle| \geq |c|\,\|r_{N-1}\|^2 \tag{2.85}$$

It is observed that this requirement implies that the improvement condition of (2.27), $\langle L\phi_N, r_{N-1}\rangle = \langle LTr_{N-1}, r_{N-1}\rangle \neq 0$, is satisfied. Since (2.85) is a stronger condition, the reverse is not true. In view of the Cauchy-Schwarz inequality, however,

$$|\langle r_{N-1}, LTr_{N-1}\rangle|^2 \leq ||LTr_{N-1}||^2 ||r_{N-1}||^2 \tag{2.86}$$

Using the definition of the norm of a bounded operator (cf. (2.70)) and applying this definition to the bounded operator LT, the first factor on the right-hand side obeys the inequality

$$||LTr_{N-1}|| \leq ||LT||\, ||r_{N-1}|| \tag{2.87}$$

Using (2.87) in (2.86), we end up with inequality

$$|\langle r_{N-1}, LTr_{N-1}\rangle|^2 \leq ||LT||^2 ||r_{N-1}||^4 \tag{2.88}$$

Comparing (2.88) with (2.85), it follows that the admissible values of c lie in the range $0 < |c| \leq ||LT|| < \infty$. Using (2.85) and (2.87) in (2.84), the inequality can be written as

$$||r_N||^2 \leq (1 - \frac{|c|^2}{||LT||^2})||r_{N-1}||^2 \text{ for } N = 1, \cdots \tag{2.89}$$

results. Repeated application of (2.89) yields

$$||r_N||^2 \leq (1 - \frac{|c|^2}{||LT||^2})^N ||r_0||^2 \text{ with } 0 < |c| \leq ||LT|| < \infty \tag{2.90}$$

From (2.90) it follows that, if there exists some $c \neq 0$ such that (2.85) holds, the error $\mathrm{ERR}_N = ||r_N||$ converges monotonically to zero as $N \to \infty$. The rate of convergence depends on the values of $\frac{|c|}{||LT||}$; the closer this value is to unity, the faster the convergence.

Selfadjoint and Positive Operator LT

If LT is selfadjoint and positive ($\langle u, LTu\rangle > 0$ for all $u \neq 0$ defined on D), then there exists a positive selfadjoint operator $(LT)^{\frac{1}{2}}$ such that (KREYSZIG, 1978, p. 476 - 479)

$$\langle u, LTu\rangle = \langle (LT)^{\frac{1}{2}}u, (LT)^{\frac{1}{2}}u\rangle = ||(LT)^{\frac{1}{2}}u||^2 \tag{2.91}$$

for all $u \in D$. Replacing L in (2.73) by $(LT)^{\frac{1}{2}}$ and using $((LT)^{\frac{1}{2}})^{-1} = (LT)^{-\frac{1}{2}}$, it follows that

$$\|(LT)^{\frac{1}{2}}u\| \le \frac{\|u\|}{\|(LT)^{-\frac{1}{2}}\|} \tag{2.92}$$

for all $u \in D$. Using $\|(LT)^{\frac{1}{2}}\| = \|(LT)^{-1}\|^{\frac{1}{2}}$, (2.91) - (2.92) lead to

$$\langle u, LTu\rangle \ge \frac{\|u\|^2}{\|(LT)^{-1}\|} \tag{2.93}$$

for all $u \in D$. Comparing (2.85) and (2.93), we observe that there indeed exists such a constant c, viz.,

$$|c| = \frac{1}{\|(LT)^{-1}\|} \tag{2.94}$$

In view of the leftmost inequality of (2.74), with L replaced by LT, we have $c \neq 0$. Using this result in (2.90), we arrive at

$$\|r_N\|^2 \le (1 - \frac{1}{\|(LT)^{-1}\|^2\,\|LT\|^2})^N \|r_0\|^2 \tag{2.95}$$

From (2.74), with L replaced by LT, it follows that

$$\frac{1}{\|(LT)^{-1}\|\,\|LT\|} \le 1 \tag{2.96}$$

Equations (2.95) and (2.96) demonstrate the convergence of the scheme. If LT is close to the identity operator, the left-hand side of (2.96) becomes close to 1 and very rapid convergence is expected; for example, in a preconditioning procedure the operator LT has to more closely resemble the identity operator than L itself does.

Symmetrization

In the symmetrization procedure of taking $T = L^*$ (cf. (2.60)), we observe that (2.95) can be rewritten as

$$\|r_N\|^2 \le (1 - \frac{1}{\|L^{-1}\|^4\,\|L\|^4})^N \|r_0\|^2 \tag{2.97}$$

where $\|(LL^\star)^{-1}\| = \|L^{-1}\|^2$ and $\|LL^\star\| = \|L\|^2$ have been used.

Symmetrization and Preconditioning

In the symmetrization and preconditioning procedure of taking $T = PP^\star L^\star$ (cf. (2.66)), we observe that (2.95) can be rewritten as

$$\|r_N\|^2 \leq (1 - \frac{1}{\|(LP)^{-1}\|^4\,\|LP\|^4})^N \|r_0\|^2 \tag{2.98}$$

where $\|(LPP^\star L^\star)^{-1}\| = \|(LP)^{-1}\|^2$ and $\|LPP^\star L^\star\| = \|LP\|^2$ have been used. Comparing (2.97) and (2.98), we observe that the preconditioned scheme under present consideration converges indeed faster than the non-preconditioned scheme as soon as LP is closer to the identity operator than L itself is.

2.8 Numerical Results for the Scattering by a Slab

In this section we consider the numerical solution of the operator equation

$$Lu = f, \quad \text{for } 0 < x < l \tag{2.99}$$

as it arises in the time-harmonic (time factor $\exp(-i\omega t)$) scattering problem (e.g., MUR and NICIA, 1976) of a plane wave normally incident upon an inhomogeneous slab, where l is the width of the slab (Fig. 2.1). In this case, the unknown field quantity u represents the total field in the slab and the known quantity f represents the incident field given by

$$f = \exp(ikx) \tag{2.100}$$

where $k = \frac{2\pi}{\lambda}$ is the angular wave number of the surrounding medium and λ is the wavelength. The operator L acting on u is found to be

$$Lu = (I - K)u \tag{2.101}$$

where I is the identity operator and Ku is given by

$$\begin{aligned} Ku &= \int_{x'=0}^{l} \frac{i}{2k} \exp(ik|x-x'|)\,(k_s^2(x') - k^2)\,u(x')\,dx' \\ &= \frac{ik}{2} \int_{x'=0}^{l} \exp(ik|x-x'|)\,\chi(x')\,u(x')\,dx' \end{aligned} \tag{2.102}$$

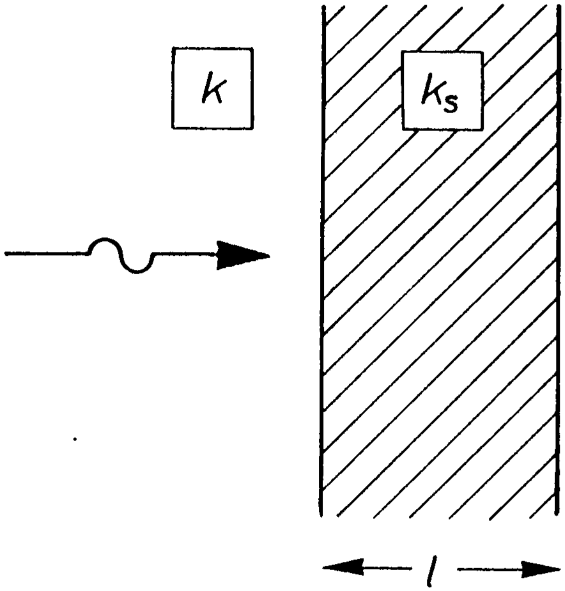

Figure 2.1 **Plane-wave scattering by a slab.**

Here, χ denotes the contrast of the slab with respect to the surrounding medium and is defined as

$$\chi(x) = \frac{k_s^2(x)}{k^2} - 1 \tag{2.103}$$

and $k_s(x)$ denotes the space-dependent angular wave number of the slab. Note that (2.102) represents a source-type integral representation based on the one-dimensional point-source Green's function in free-space with angular wave number k, viz.

$$G(x, x') = \frac{i}{2k}\exp(ik|x - x'|) \tag{2.104}$$

Before actually turning to the numerical results, we first consider a particular preconditioning operator for the case where the slab is homogeneous. Then, the contrast function χ is a constant and (2.101) may be rewritten as

$$Lu = (I - \chi K')u \tag{2.105}$$

where $K'u$ is given as

$$K'u = \frac{ik}{2}\int_{x'=-\infty}^{\infty} \exp(ik|x - x'|)\,\chi_D u(x')\,dx' \tag{2.106}$$

where the characteristic function χ_D of the slab region D $(0 < x < l)$ is defined as

$$\chi_D = \begin{cases} 1, & \text{when } x \in (0,l) \\ 0, & \text{when } x \notin [0,l] \end{cases} \tag{2.107}$$

Let F denote the spatial Fourier transformation according to

$$F\{v\} = \int_{x=-\infty}^{\infty} \exp(-i\alpha x)\, v(x)\, dx, \quad -\infty < \alpha < \infty \tag{2.108}$$

then

$$F\{K'u\} = \frac{k^2}{\alpha^2 - k^2} F\{\chi_D u\} \tag{2.109}$$

where the product rule for the Fourier transformation of a convolution and the result

$$\int_{x=-\infty}^{\infty} \exp(-i\alpha x) \exp(ik|x|)\, dx = -\frac{2ik}{\alpha^2 - k^2} \tag{2.110}$$

have been used. From (2.105) and (2.106) the Fourier transform of Lu is obtained as

$$F\{Lu\} = \frac{\alpha^2 - k_s^2}{\alpha^2 - k^2} F\{\chi_D u\} \tag{2.111}$$

where $k_s^2 = (1+\chi)k^2$ and k_s is now the constant angular wave number of the slab. The equation inverse to (2.111) is

$$F\{\chi_D u\} = \frac{\alpha^2 - k^2}{\alpha^2 - k_s^2} F\{Lu\} \tag{2.112}$$

The value of $\chi_D u$ cannot be obtained by an inverse Fourier transformation of (2.112) into the spatial domain, since Lu is only known for $x \in (0,l)$ and not outside this interval. Nevertheless, (2.112) will be used to construct a preconditioning operator that is under certain circumstances an approximate inverse operator. For any v defined on $x \in (0,l)$, let the operator P defined through

$$F\{Pv\} = \frac{\alpha^2 - k^2}{\alpha^2 - k_s^2} F\{\chi_D v\} \tag{2.113}$$

Inverse Fourier transformation then yields

$$Pv = (I - Q)v \tag{2.114}$$

where

$$
\begin{aligned}
Qv &= \frac{i(k^2 - k_s^2)}{2k_s} \int_{x'=-\infty}^{\infty} \exp(ik_s|x - x'|)\, \chi_D\, v(x')\, dx' \\
&= \int_{x'=0}^{l} \frac{i}{2k_s} \exp(ik_s|x - x'|)\, (k^2 - k_s^2) v(x')\, dx' \qquad (2.115)
\end{aligned}
$$

Note that (2.115) represents a source-type integral representation based on the one-dimensional point-source Green's function in free-space with constant angular wave number k_s, viz.

$$
G_s(x, x') = \frac{i}{2k_s} \exp(ik_s|x - x'|) \qquad (2.116)
$$

Now P is an approximate inverse of L in all cases where Lu is relatively small outside the interval $(0, l)$. In the remainder, P will be employed as a preconditioner operator. Note that the operator P applies to a homogeneous slab.

In the following subsections, we present the numerical results obtained with the different methods discussed earlier in this chapter. The integrals ocurring in the operator expressions and in the inner products of the different iterative schemes are well behaved and can be computed numerically with the aid of a trapezoidal integration rule. The number of integration points (in the order of some tens for low contrasts to several hundreds for high contrasts) is chosen such that the numerical discretization error is less than the error made in the resulting approximation of our pertaining field solution. As soon as the number of iterations grows larger, the danger of loss of significant figures turns up. For this reason, all computations have been carried out in double precision (REAL*8 and COMPLEX*16 in FORTRAN), while the residual in the operator equation has each time been determined by substituting the obtained approximate solution in this equation and not by using the recursive relation for the successive residuals that for a number of cases is available. In those cases where a loss of significant figures was expected, a check has been carried out against the corresponding computation in single precision. In the conjugate-gradient scheme an additional check is provided by the orthogonality relations that must be satisfied. Once a discrepancy in these occurs, the orthogonality relations are enforced by falling back on the scheme defined by (2.41) - (2.43).

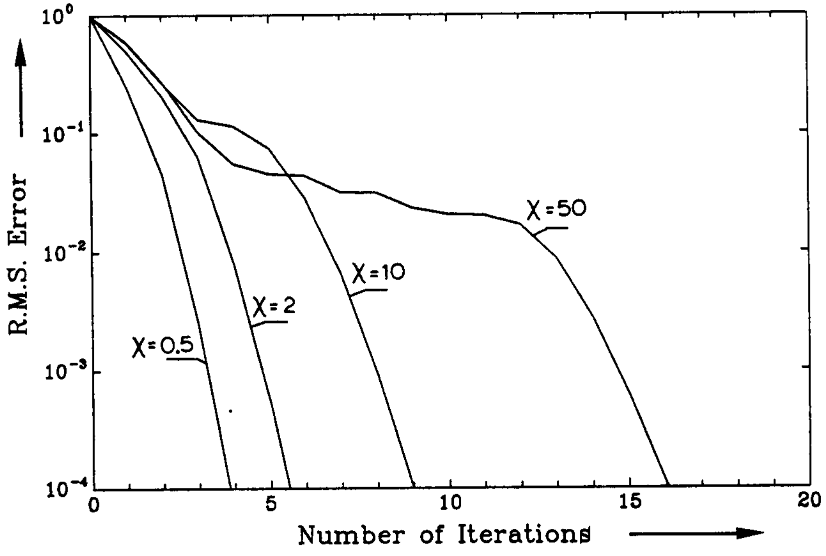

Figure 2.2 Results of the recursive scheme with $T = I$; homogeneous slab with $\frac{l}{\lambda} = 0.5$.

Recursive Solution with $T = I$

We first consider the iterative solution of the recursive scheme of (2.41) - (2.43), in which we take $T = I$. We take $\frac{l}{\lambda} = 0.5$ and consider some constant values of the contrast. In Fig. 2.2 we present the numerical results for the root-mean-square error $\widehat{\mathrm{ERR}}$ (cf. (2.10)) as a function of the number of iterations. We observe a reasonable convergence for all values of χ. The larger this value is, the lower the rate of convergence.

Preconditioned Recursive Solution with $T = P$

Subsequently, we consider the recursive scheme of (2.41) - (2.43), in which we take $T = P$, where P is given by (2.114) - (2.115). We take $\frac{l}{\lambda} = 0.5$ and consider some constant values of the contrast. In Fig. 2.3 we present the numerical results for the root-mean-square error $\widehat{\mathrm{ERR}}$ as a function of the number of iterations. Comparing the results with those of the non-preconditioned recursive scheme shown in Fig. 2.2, we observe that for values of the contrast χ up to 50, convergence ($\widehat{\mathrm{ERR}} < 0.0001$) is obtained within three iterations. Hence, for constant contrast, our preconditioning operator is a very efficient one.

Conjugate-Gradient Method: Recursive Solution with $T = L^{\star}$

We now consider the recursive scheme of (2.41) - (2.43), in which we take $T = L^{\star}$. Since $LT = LL^{\star}$ is a selfadjoint operator we can use a conjugate-gradient scheme (cf. Sections 2.5 and 2.6); in particular, we can employ the scheme of (2.51) - (2.53). We take $\frac{l}{\lambda} = 0.5$ and consider some constant values of the contrast. In Fig. 2.4 we present the numerical results for the root-mean-square error $\widehat{\mathrm{ERR}}$ as a function of the number of iterations. Comparing the results with those of the recursive scheme of Fig. 2.2, we observe that the rate of convergence has been decreased in taking $T = L^{\star}$ in stead of $T = I$. The advantage of the conjugate-gradient scheme is that the orthogonalization of the expansion functions is automatically enforced and storage of these expansion functions of all previous iterations is superfluous. However, after a number of iterations in the conjugate-gradient scheme, loss of significant figures leads to a non-satisfaction of the orthogonality conditions. Then, the convergence slows down for a few iterations. If we enforce the orthogonality by falling back on the recursive scheme of (2.41) - (2.43) the convergence is maintained. In Fig. 2.4 we observe this phenomenon for $\chi \geq 2$. The dashed lines represent the results when the orthogonalization is enforced by using the recursive scheme with full orthogonalization.

Preconditioned Conjugate-Gradient Method: Recursive Solution with $T = PP^{\star}L^{\star}$

Subsequently, we consider the recursive scheme of (2.41) - (2.43), in which we take $T = PP^{\star}L^{\star}$. Since $LT = LPP^{\star}L^{\star}$ is a selfadjoint operator we can use a conjugate-gradient scheme (cf. Sections 2.5 and 2.6); in particular, we employ the scheme of (2.51) - (2.53). We take $\frac{l}{\lambda} = 0.5$ and consider some constant values of the contrast. In Fig. 2.5 we present the numerical results for the root-mean-square error $\widehat{\mathrm{ERR}}$ as a function of the number of iterations. Comparing the results with those of the non-preconditioned conjugate-gradient scheme of Fig. 2.4, we observe that the convergence has been considerably increased, although the results of the preconditioned non-symmetrized recursive scheme of Fig. 2.3 exhibit a much better convergence. Note again that an enforcement of the orthogonality by using the recursive scheme

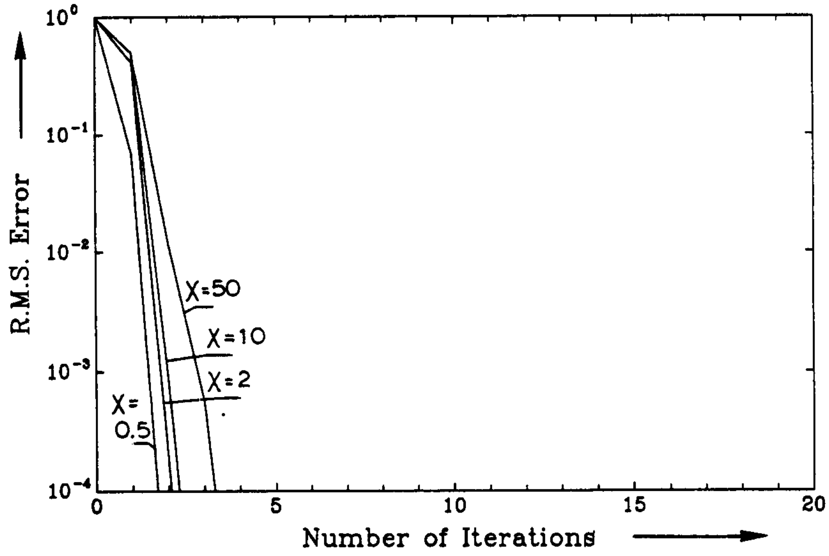

Figure 2.3 **Results of the preconditioned recursive scheme with $T = P$; homogeneous slab with $\frac{l}{\lambda} = 0.5$.**

of (2.41) - (2.43) eliminates the loss of numerical orthogonality. The dashed line in the figure represents the results of the enforcement of the orthogonality.

Before presenting some numerical results for an inhomogeneous slab, we first compare the results of the recursive scheme with $T = I$ and $T = P$ for different values of $\frac{l}{\lambda}$, viz. 0.5 and 1.0, respectively. In Fig. 2.6 we present the root-mean-square-error results for a homogeneous slab with a very high contrast ($\chi = 50$). For larger values of $\frac{l}{\lambda}$ we observe a decrease in the rate of convergence of the non-preconditioned scheme. However, the preconditioned scheme converges in almost the same rate for different values of $\frac{l}{\lambda}$.

Subsequently, we present the numerical results for an inhomogeneous slab. We take a contrast $\chi(x) = 100x$ increasing from zero to hundred in the interval $(0, l)$ (Fig. 2.7) and a contrast $\chi = 100(l - x)$ decreasing from hundred to zero in the interval $(0, l)$ (Fig. 2.8). The average value of χ over the slab domain for both cases is the same (= 50). In the first instance, for the preconditioning operator we use the one of (2.114) - (2.115), derived for a constant $\chi = 50$. From Figs. 2.7

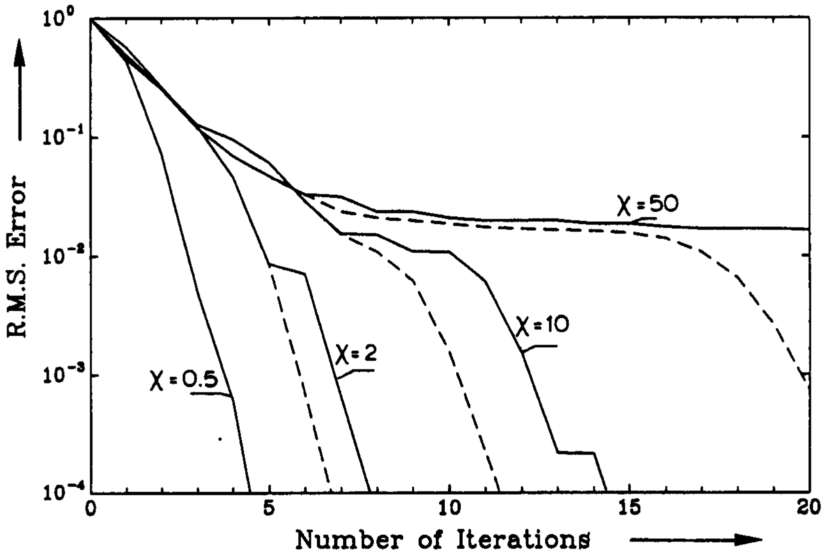

Figure 2.4 **Results of the conjugate-gradient scheme with $T = L^\star$; homogeneous slab with $\frac{l}{\lambda} = 0.5$.**

- 2.8 we observe that this preconditioning operator P does not work very effective in these inhomogeneous cases. However, when we replace the Green's function of (2.116) for the WKB approximation (TIJHUIS, 1987) of the Green's function in an inhomogeneous medium, viz.

$$G_s(x,x') = \frac{i}{2(k_s(x)k_s(x'))^{\frac{1}{2}}} \begin{cases} \exp(i \int_{x''=x'}^{x} k_s(x'')\,dx''), & x > x' \\ \exp(-i \int_{x''=x'}^{x} k_s(x'')\,dx''), & x < x' \end{cases} \tag{2.117}$$

we obtain the preconditioning operator

$$P^{WKB}v = (I - Q^{WKB})v \tag{2.118}$$

where

$$Q^{WKB}v = \int_{x'=0}^{l} G_s(x,x')\,(k^2 - k_s^2(x'))\,v(x')\,dx' \tag{2.119}$$

Note that, if the slab is homogeneous, we arrive at the results of (2.114) - (2.115) and $P = P^{WKB}$. This preconditioning operator P^{WKB} seems

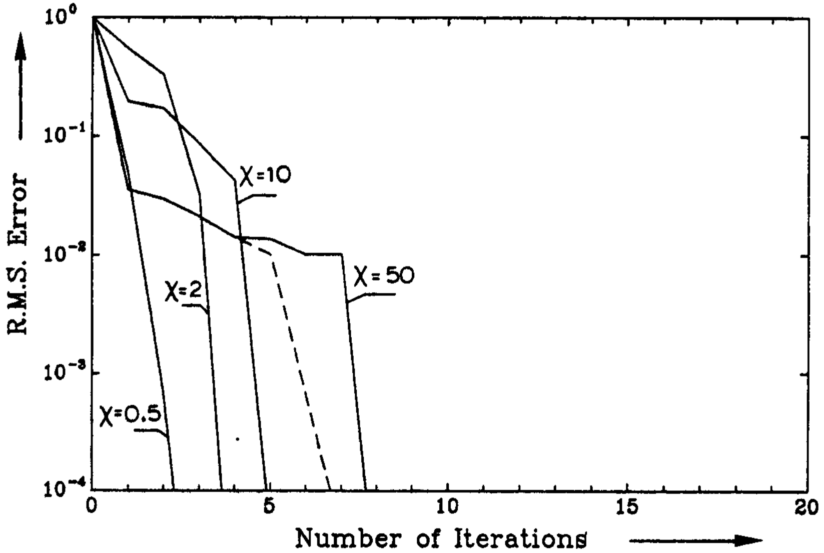

Figure 2.5 Results of the preconditioned conjugate-gradient scheme with $T = PP^{\star}L^{\star}$; homogeneous slab $\frac{l}{\lambda} = 0.5$.

to be very effective: Figs. 2.7 - 2.8 clearly demonstrate that very rapid convergence is arrived at in a few iterations nearly independent of $\frac{l}{\lambda}$.

2.9 Numerical Results for the Scattering by a Strip

In this section we consider the numerical solution of the operator equation

$$Lu = f, \quad \text{for } -a < x < a \tag{2.120}$$

as it arises in the time-harmonic (time factor $\exp(-i\omega t)$) scattering problem (VAN DEN BERG and KLEINMAN, 1988) of a plane wave normally incident upon a strip, where $2a$ is the width of the strip (Fig. 2.9). After some normalization the known quantity f is given by

$$f = 1 \tag{2.121}$$

and the operator L acting on u is found to be

$$Lu = \int_{x'=-a}^{a} \frac{1}{2} H_0^{(1)}(k|x - x'|)\, u(x')\, dx' \tag{2.122}$$

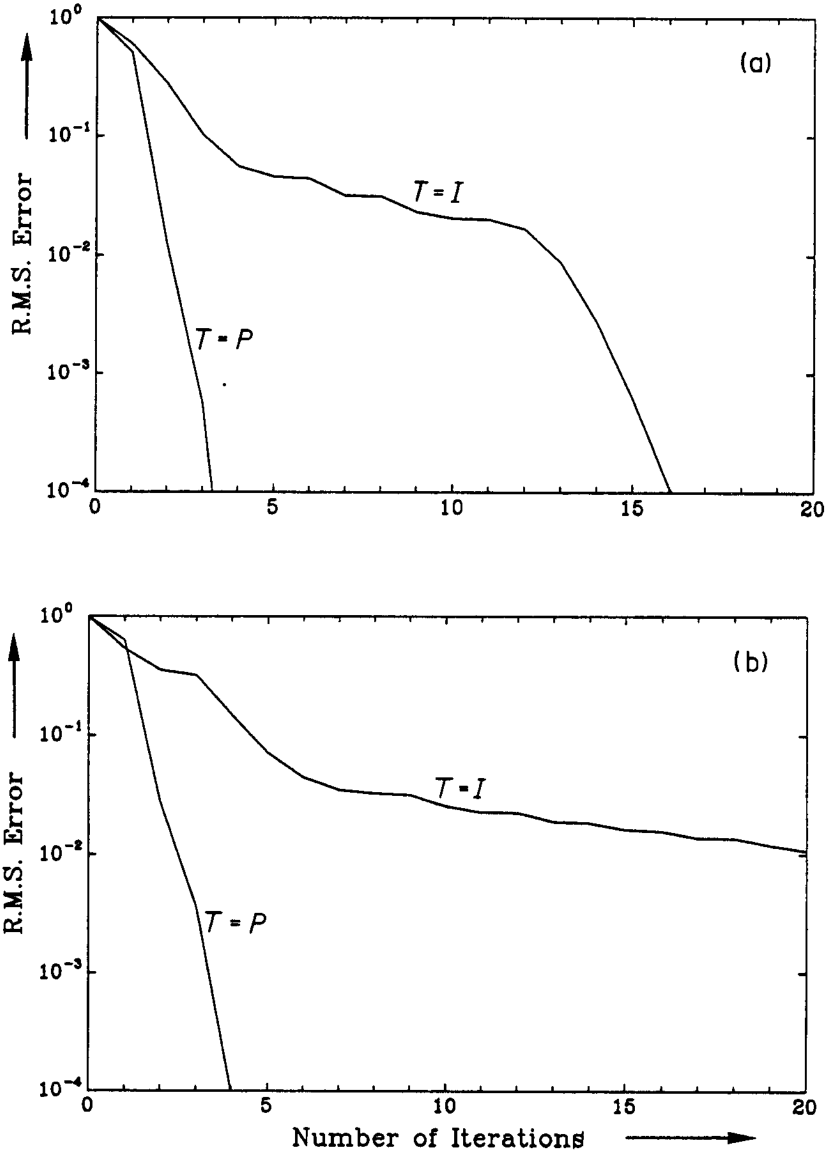

Figure 2.6 Results of the recursive scheme with $T = I$; homogeneous slab with $\chi = 50$; (a) $\frac{l}{\lambda} = 0.5$; (b) $\frac{l}{\lambda} = 1.0$.

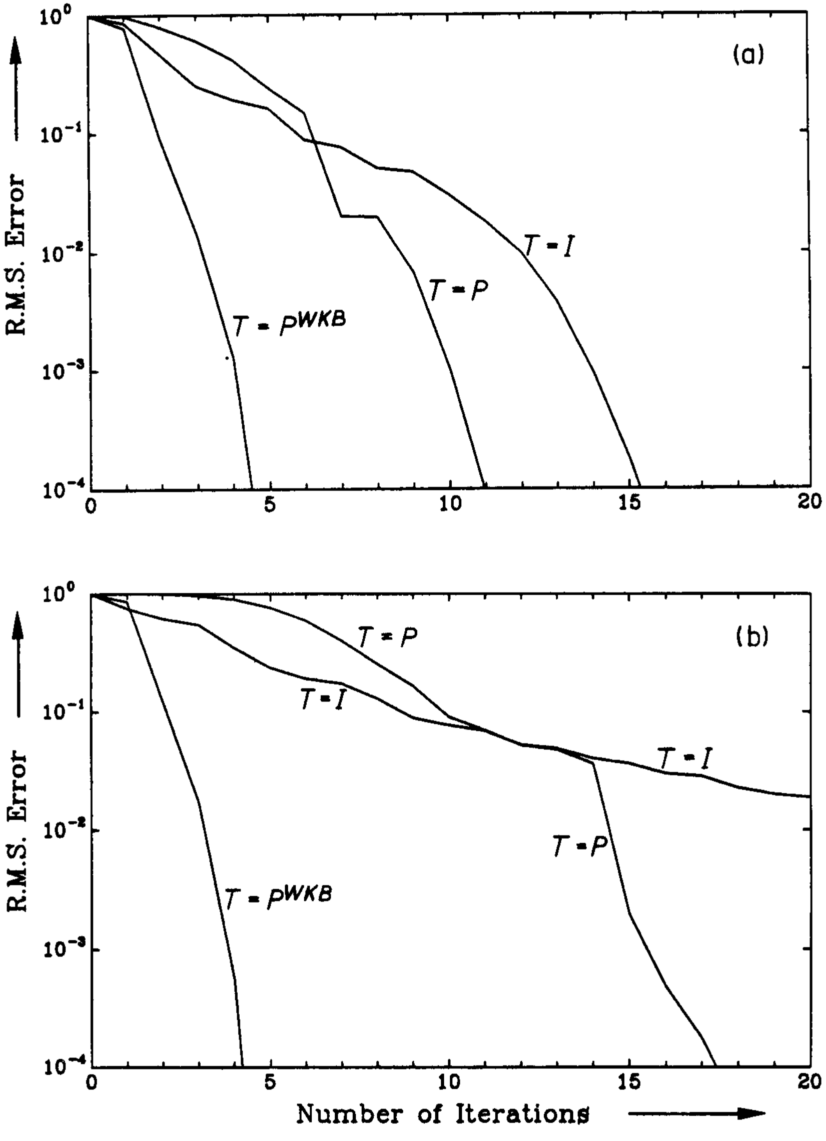

Figure 2.7 **Results of the recursive scheme with $T = I$; inhomogeneous slab with increasing contrast; (a) $\frac{l}{\lambda} = 0.5$; (b) $\frac{l}{\lambda} = 1.0$.**

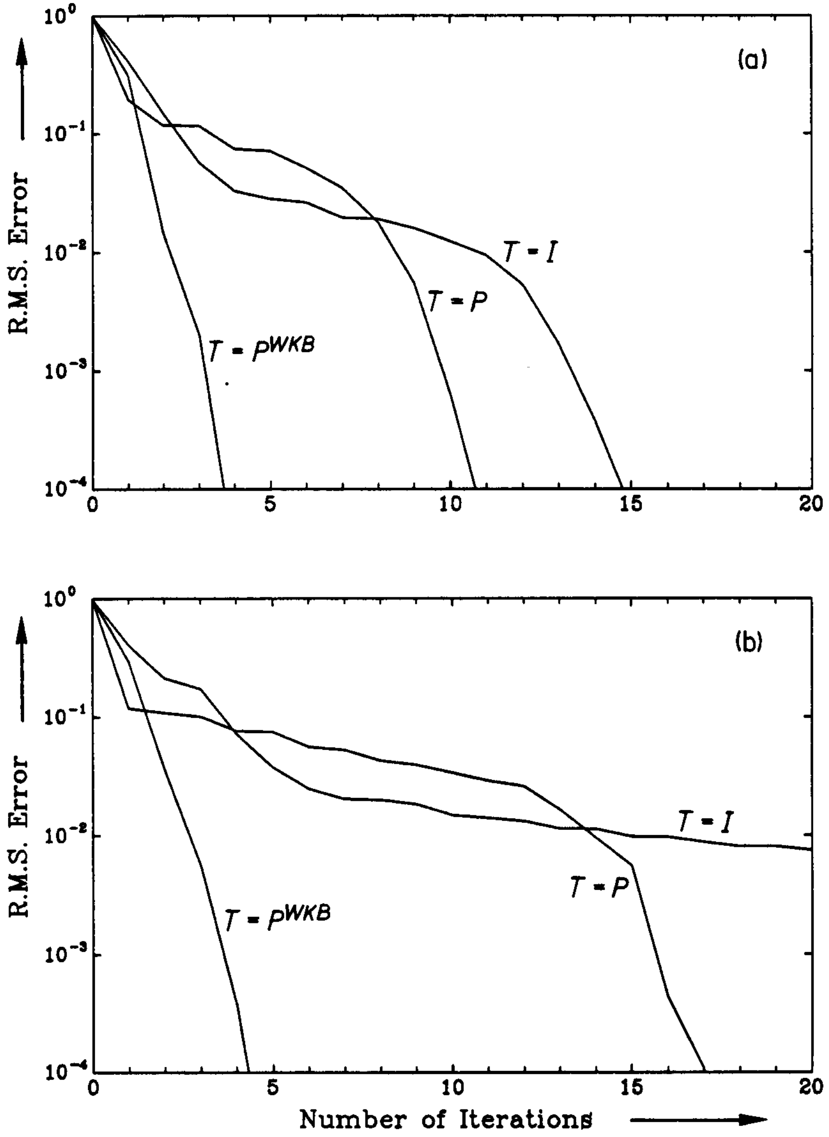

Figure 2.8 Results of the recursive scheme with $T = I$; inhomogeneous slab with decreasing contrast; (a) $\frac{l}{\lambda} = 0.5$; (b) $\frac{l}{\lambda} = 1.0$.

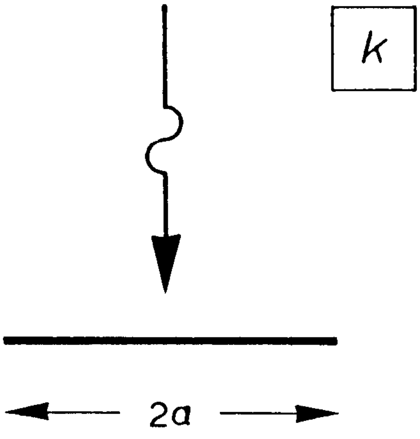

Figure 2.9 Plane-wave scattering by a strip.

where $H_0^{(1)}$ is the zero order Hankel function of the first kind, k is the angular wave number of the surrounding medium. In order to take advantage of the convolution structure in the operator expression, we define the spatial Fourier transformation as

$$F\{v\} = \int_{x=-\infty}^{\infty} \exp(-i\alpha x)\, v(x)\, dx, \quad -\infty < \alpha < \infty \tag{2.123}$$

and the inverse Fourier transformation as

$$F^{-1}\{w\} = \frac{1}{2\pi} \int_{\alpha=-\infty}^{\infty} \exp(i\alpha x)\, w(\alpha)\, d\alpha, \quad -\infty < x < \infty \tag{2.124}$$

Then, the Fourier transformation of (2.122) can be written as

$$F\{Lu\} = (k^2 - \alpha^2)^{-\frac{1}{2}} F\{\chi_D u\} \tag{2.125}$$

where the product rule for the Fourier transformation of a convolution and the result

$$\int_{x=-\infty}^{\infty} \exp(-i\alpha x)\, \frac{1}{2} H_0^{(1)}(k|x|)\, dx = (k^2 - \alpha^2)^{-\frac{1}{2}} \tag{2.126}$$

have been used. In (2.125) the characteristic function χ_D of the strip region D $(-a < x < a)$ is defined as

$$\chi_D = \begin{cases} 1, & \text{when } x \in (-a, a) \\ 0, & \text{when } x \notin [-a, a] \end{cases} \tag{2.127}$$

We observe that the operator expression Lu can efficiently be computed using Fast-Fourier-Transform (FFT) routines to evaluate the forward and inverse Fourier transformations in

$$Lu = F^{-1}\{(k^2 - \alpha^2)^{-\frac{1}{2}} F\{\chi_D u\}\} \tag{2.128}$$

A similar computation can be carried out for the adjoint operator $L^\star$. In order to cope with the branch point in the numerical inverse Fourier transformation in Eq (2.128), we introduce slight lossess in the medium surrounding the strip, in accordance with the condition of causality, by taking the angular wave number to be complex, viz. $k = \frac{2\pi}{\lambda}(1+0.01i)$, where λ is the wavelength.

Before actually turning to the numerical results, we first consider a particular preconditioning operator. The equation inverse to (2.125) is

$$F\{\chi_D u\} = (k^2 - \alpha^2)^{\frac{1}{2}} F\{Lu\} \tag{2.129}$$

The value of $\chi_D u$ cannot be obtained by an inverse Fourier transformation of (2.129) to the spatial domain, since Lu is only known for $x \in (-a, a)$ and not outside this interval. Nevertheless, (2.129) will be used to construct a preconditioning operator that is under certain circumstances an approximate inverse operator. Let for any v defined on $x \in (-a, a)$, the operator P defined through

$$F\{Pv\} = (k^2 - \alpha^2)^{\frac{1}{2}} F\{\chi_D v\} \tag{2.130}$$

Inverse Fourier transformation then yields

$$Pv = F^{-1}\{(k^2 - \alpha^2)^{\frac{1}{2}} F\{\chi_D v\}\} \tag{2.131}$$

Now P is an approximate inverse of L in all cases where Lu is relatively small outside the interval $(-a, a)$. In the remainder, P will be employed as a preconditioner operator. Note that the operator expression Pv can also be computed efficiently by using FFT routines for the forward and inverse Fourier transformations in (2.131). A similar computation can be carried out for the adjoint operator $P^\star$.

In the following subsections, we present the numerical results obtained with the different methods discussed earlier in this chapter. All operator expressions are computed by a 4096-points FFT routine. All integrals ocurring in the inner products of the different iterative

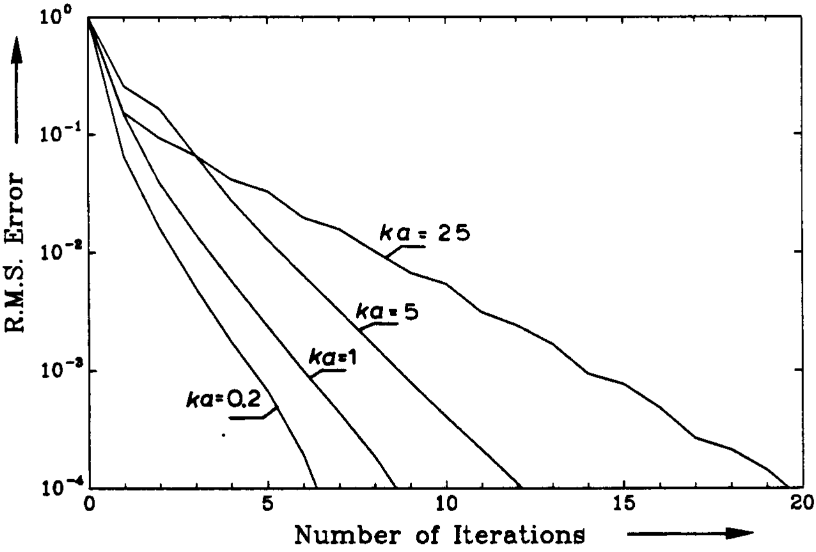

Figure 2.10 **Results of the recursive scheme with $T = I$ in the strip problem.**

schemes are computed numerically with the aid of a simple summation of the function values at the sample points. The number of sample points on the strip (of width $2a$ amounts to 17, 35, 81 and 181, when the real part of ka =0.2, 1, 5 and 25, respectively. These numbers of integration points are chosen such that numerical discretization errors are less than the error made in the resulting approximation of our pertaining field values at the strip. As soon as the number of iterations grows larger, the danger of a loss of significant figures turns up. For this reason, all computations have been carried out in double precision, while the residual in the operator equation has been determined each time by substituting the obtained approximate solution into this equation, and not by using the recursive relation for the successive residuals that is available for a number of cases. In those cases where a loss of significant figures was expected, a check was carried out against the corresponding computation in single precision. In the conjugate-gradient scheme an additional check is provided by the orthogonality relations that have to be satisfied. Once a discrepancy in these occurs, the orthogonality relations are enforced by falling back on the scheme

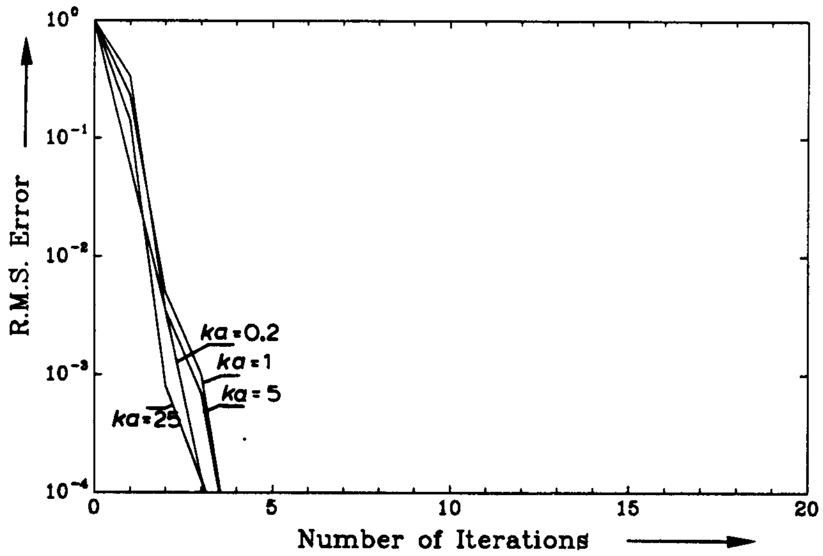

Figure 2.11 Results of the preconditioned recursive scheme with $T = P$ in the strip problem.

defined by (2.41) - (2.43).

Recursive Solution with $T = I$

We first consider the iterative solution of the recursive scheme of (2.41) - (2.43), in which we take $T = I$. In Fig. 2.10 we present the numerical results for the root-mean-square error $\widehat{\mathrm{ERR}}$ (cf. (2.10)) as a function of the number of iterations.

Preconditioned Recursive Solution with $T = P$

Subsequently, we consider the recursive scheme of (2.41) - (2.43), in which we take $T = P$, where P is given by (2.131). In Fig. 2.11 we present the numerical results for the root-mean-square error $\widehat{\mathrm{ERR}}$ as a function of the number of iterations. Comparing the results with those of the non-preconditioned recursive scheme shown in Fig. 2.10, we then notice that it is superior to the non-preconditioned scheme, even when we take into account that the computation time of one iteration is now nearly doubled. For all values of ka considered the preconditioned scheme converges within a very few iterations ($\widehat{\mathrm{ERR}} < 0.0001$). Our

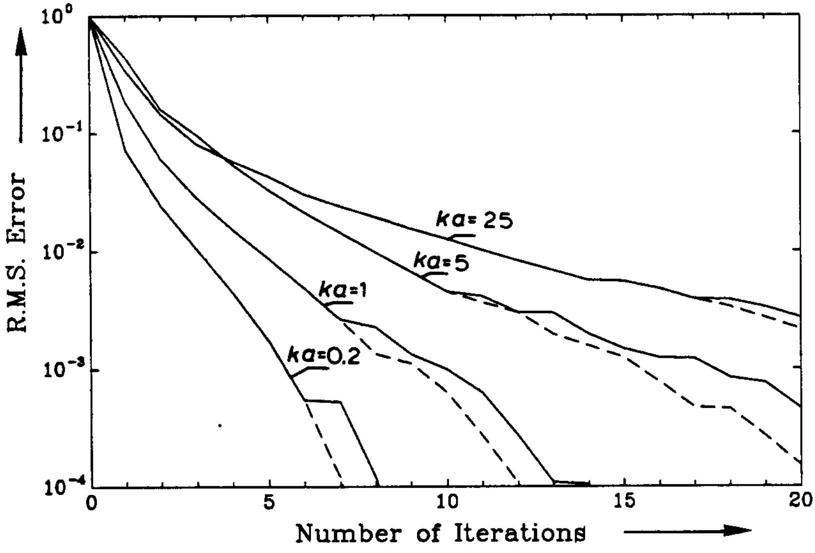

Figure 2.12 **Results of the conjugate-gradient scheme with $T = L^\star$ in the strip problem.**

preconditioning operator seems to be a very efficient one.

Conjugate-Gradient Method: Recursive Solution with $T = L^\star$

We now consider the recursive scheme of (2.41) - (2.43), in which we take $T = L^\star$. Since $LT = LL^\star$ is a selfadjoint operator we can use a conjugate-gradient scheme (cf. Sections 2.5 and 2.6); in particular, we can employ the scheme of (2.51) - (2.53). In Fig. 2.12 we present the numerical results for the root-mean-square error $\widehat{\mathrm{ERR}}$ as a function of the number of iterations. Comparing the results with those of the recursive scheme of Fig. 2.10, we observe that the rate of convergence has been decreased by taking $T = L^\star$ instead of $T = I$. The advantage of the conjugate-gradient scheme is that the orthogonalization of the expansion functions is automatically enforced and storage of these expansion functions of all previous iterations is superfluous. However, after a number of iterations in the conjugate-gradient scheme, loss of significant figures leads to a non-satisfaction of the orthogonality conditions. Then, the convergence slows down for a few iterations. If we enforce the orthogonality by falling back on the recursive scheme of

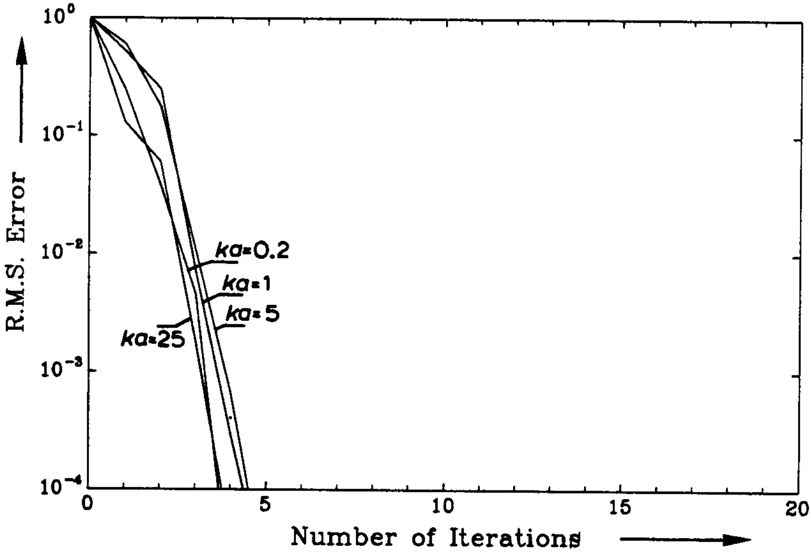

Figure 2.13 Results of the preconditioned conjugate-gradient scheme with $T = PP^\star L^\star$ in the strip problem.

(2.41) - (2.43) the convergence is maintained. In Fig. 2.12 we observe this phenomenon. The dashed lines represent the results when the orthogonalization is enforced by using the recursive scheme with full orthogonalization.

Preconditioned Conjugate-Gradient Method: Recursive Solution with $T = PP^\star L^\star$

Subsequently, we consider the recursive scheme of (2.41) - (2.43), in which we take $T = PP^\star L^\star$. Since $LT = LPP^\star L^\star$ is a selfadjoint operator we can use a conjugate-gradient scheme (cf. Sections 2.5 and 2.6); in particular, we employ the scheme of (2.51) - (2.53). In Fig. 2.13 we present the numerical results for the root-mean-square error $\widehat{\text{ERR}}$ as a function of the number of iterations. Comparing the results with those of the non-preconditioned conjugate-gradient scheme of Fig. 2.12, we observe that the convergence has been considerably increased, although the results of the preconditioned non-symmetrized recursive scheme of Fig. 2.11 exhibit a much better convergence.

2.10 Conclusions

In this chapter we have discussed a general recursive scheme for an iterative minimization of the integrated square error in an operator equation. In this scheme the expansion functions of all previous iterations have to be stored, unless the operator is selfadjoint; in this special case, the scheme becomes a conjugate-gradient scheme with a two-term recursion only. In practice however, the operators occurring in field problems are not selfadjoint and a symmetrization procedure has to be performed to arrive at a conjugate-gradient scheme. This symmetrization negatively influences the rate of convergence. It is noted that loss of significant figures in the computations disturb the convergence in the conjugate-gradient scheme as well. In order to accellerate the convergence of the conjugate-gradient method appropriate preconditioning seems to be a significant tool. When the number of iterations can be kept small enough to accommodate the storage of all expansion functions of previous iterations (e.g., in the background memory of the computer) the complete orthogonalization procedure of the recursive scheme seems to be the most favorable one. Nevertheless, some incomplete orthogonalization procedures (VAN DEN BERG and GHIJSEN, 1988) may yield satisfactory results.

We have formulated the various methods in the continuous operator form which is especially useful in arriving at useful preconditioners in relation to the physics of the problem. It is noted that the numerical discretization of problems more complex than the examples presented in this chapter, should be done with care. The numerical representation of the field quantities should meet the physical and mathematical requirements pertaining to the field problem.

References

[1] Daniel, J. W., "The conjugate gradient method for linear and nonlinear operator equations," *SIAM Journal on Numerical Analysis,* 4, 10–26, 1967.

[2] Harrington, R. F., *Field Computation by Moment Methods,* New York, The Macmillan Company, 1968.

[3] Hestenes, M. R., and E. Stiefel, "Methods of conjugate gradients

for solving linear systems," *Journal of Research of the National Bureau of Standards*, **49**, 409–435, 1952.

[4] Kreyszig, E., *Introductory Functional Analysis with Applications*, New York, John Wiley & Sons, 1978.

[5] Mur, G., and A. J. A. Nicia, "Calculation of reflection and transmission coefficients in one-dimensional wave propagation problems," *Applied Physics*, **47**, 5218–5221, 1976.

[6] Tijhuis, A. G., *Electromagnetic Inverse Profiling: Theory and Numerical Implementation*, VNU Science Press BV, Utrecht, The Netherlands, 66, 1987.

[7] van den Berg, P. M., "Iterative computational techniques in scattering based upon the integrated square error criterion," *IEEE Transactions on Antennas and Propagation*, **AP-32**, 1063–1071, 1984.

[8] van den Berg, P. M., and W. J. Ghijsen, "A spectral iterative technique with Gram-Schmidt orthogonalization," *IEEE Transactions on Microwave Theory and Techniques*, **MTT-36**, 769–772, 1988.

[9] van den Berg, P. M., and R. E. Kleinman, "The conjugate gradient spectral iterative technique for planar structures," *IEEE Transactions on Antennas and Propagation*, **AP-36**, 1418–1423, 1988.