3

ITERATIVE METHODS FOR SOLVING INTEGRAL EQUATIONS

R. E. Kleinman and P. M. van den Berg

3.1 Introduction

Acoustic and electromagnetic scattering problems are often formulated as integral equations and it is this form which serves as the starting point for most numerical solutions. Typically the integral operators which occur are boundary integrals when considering scattering by impenetrable objects and domain integrals for penetrable scatterers. These operators are invariably non-selfadjoint which complicates most numerical approaches. In a large number of cases the integral equations may be put in a Hilbert space frame work. In abstract form the equation is

$$Lu = f \tag{3.1}$$

where L is a bounded linear operator, not necessarily selfadjoint, which maps a Hilbert Space H into itself. The space will be equipped with a norm $\|\cdot\|$ and inner product $\langle \cdot, \cdot \rangle$, linear in the first entry. The right-hand side of (3.1), f, is an element of H and is known in terms of a prescribed incident field. Throughout we assume that L is bounded,

i.e., $\|L\| < \infty$ and (3.1) is uniquely solvable for every $f \epsilon H$ which means that L^{-1} exists and $\|L^{-1}\| < \infty$. Furthermore a number of useful properties for the operator L and its adjoint L^* hold:

$$\|L\| = \|L^*\|, \qquad \|L^{-1}\| = \|L^{-1^*}\| \tag{3.2}$$

$$\inf_{\|v\|=1}\|Lv\| = \frac{1}{\|L^{-1}\|}, \quad \inf_{\|v\|=1}\|L^{-1}v\| = \frac{1}{\|L\|} \tag{3.3}$$

If in addition L is selfadjoint and positive (that is $\langle v, Lv\rangle \geq 0$ for every $v \in H$) then L^{-1} is also selfadjoint and positive and (e.g., RIESZ and NAGY, 1955, p. 266)

$$\sup_{\|v\|=1}\langle v, Lv\rangle = \|L\|, \quad \sup_{\|v\|=1}\langle v, L^{-1}v\rangle = \|L^{-1}\| \tag{3.4}$$

$$\inf_{\|v\|=1}\langle v, Lv\rangle = \frac{1}{\|L^{-1}\|}, \quad \inf_{\|v\|=1}\langle v, L^{-1}v\rangle = \frac{1}{\|L\|} \tag{3.5}$$

The prototype Hilbert space is L_2 but of course this is not necessary. Furthermore with the conditions imposed on the operator L another inner product is defined by

$$\langle u, v\rangle_2 = \langle Lu, Lv\rangle \tag{3.6}$$

If L is selfadjoint and positive yet another inner product is defined by

$$\langle u, v\rangle_3 = \langle u, Lv\rangle = \langle Lu, v\rangle \tag{3.7}$$

Each of these inner products generates a norm $\|u\|_i = \sqrt{\langle u, u\rangle_i}$ on H. We will use the subscript 1 to denote the original inner product and norm on H. Note that if L is non-selfadjoint then replacing L by L^*L in the definition of $\langle \cdot, \cdot\rangle_3$ yields the same inner product as $\langle \cdot, \cdot\rangle_2$. Hence in the following we will employ $\langle \cdot, \cdot\rangle_3$ only when L itself is selfadjoint. We denote the spectral radius of any operator A by $\sigma(A)$ where

$$\sigma(A) = \lim_{n\to\infty} \|A^n\|^{\frac{1}{n}} \tag{3.8}$$

In many specific examples the operator L is of the second kind, $L = I - K$, where K is compact.

In this chapter we will consider iterative solutions of (3.1) of the form

$$\begin{aligned} &u_0 \qquad \text{arbitrary,} \\ &u_n = u_{n-1} + \alpha_n v_n, \ n \geq 1 \end{aligned} \tag{3.9}$$

We also introduce the residual

$$r_n = f - Lu_n, \quad n \geq 0 \tag{3.10}$$

in terms of which the functions v_n will subsequently be defined. Note even before explicitly defining v_n we may use (3.9) and (3.10) to obtain the iterative relation

$$r_n = r_{n-1} - \alpha_n L v_n \tag{3.11}$$

Alternatively (3.9) and (3.11) may be employed successively to obtain the following series representations

$$u_n = u_0 + \sum_{m=1}^{n} \alpha_m v_m \tag{3.12}$$

and

$$r_n = r_0 - \sum_{m=1}^{n} \alpha_m L v_m \tag{3.13}$$

While these representations may be useful for some purposes, the expressions in (3.9) and (3.11) are preferable from a numerical view point as they do not require storage of all the α_m and v_m.

Different choices of α_n and v_n give rise to different iterative schemes and we will consider a number of them. Each method will be demonstrated in a numerical example and for this purpose we choose the problem of scattering of a plane wave normally incident on a slab of finite width. This example is the same as that introduced in Chapter 2 (see also Mur and NICIA, 1976; KLEINMAN, ROACH, SCHUETZ, SHIRRON and VAN DEN BERG, 1990; KLEINMAN, ROACH and VAN DEN BERG, 1990). In this case (3.1) has the explicit realization in which

$$f = e^{ikx} \tag{3.14}$$

where $k = \frac{2\pi}{\lambda}$ is the (constant) wave number of the surrounding medium, λ is the wavelength and

$$Lu = u(x) - \frac{ik}{2} \int_0^l e^{ik|x-x'|} \chi(x')\, u(x')\, dx' \tag{3.15}$$

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

$$\chi(x) = \frac{k_s^2(x)}{k^2} - 1 \tag{3.16}$$

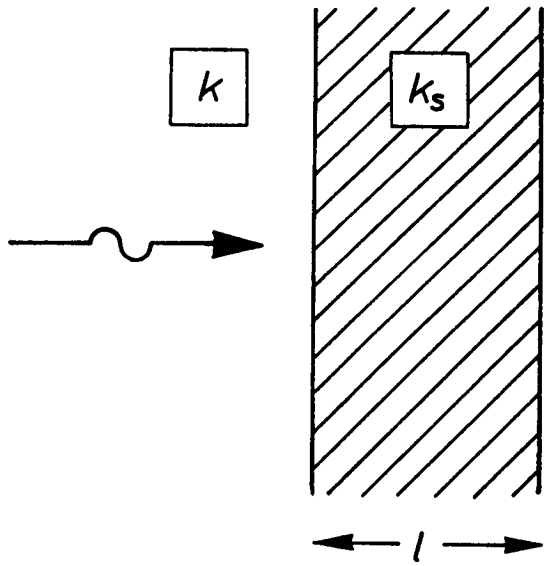

Figure 3.1 Plane-wave scattering by a slab.

in which $k_s(x)$ is the spatially dependent wave number in the slab. In this example we choose the Hilbert space H to be $L_2(0,l)$ and observe that the adjoint of L in this space is given by

$$L^*u = u(x) + \frac{i\bar{k}}{2}\,\overline{\chi}(x)\int_0^l e^{-i\bar{k}|x-x'|}\,u(x')\,dx' \tag{3.17}$$

where the overbar denotes complex conjugate.

In the numerical examples considered in the following sections a number of different iteration schemes will be illustrated. However in order to have a uniform basis for comparison the figures will all plot the normalized residual error, $\frac{\|r_n\|}{\|f\|}$, where $r_n = f - Lu_n$ will have been obtained using different iterative processes. The integrals occurring in the operator expressions and in the inner products of the different iterative schemes are computed numerically with the aid of a trapezoidal integration rule. Numerical results are presented for the homogeneous slab with a normalized width of $\frac{l}{\lambda} = 0.5$. The number of integration points (in our example this ranges from 20 points for χ=0.5 to 400 points for χ=32) is chosen such that the numerical discretization error is less than the error made in the resulting approximation of u, so that further increase in the number of discretization points will not change

the value of the residual error. All computations have been carried out in double precision, while the residual r_n in the equation at each step has been determined by substituting the approximation u_n in the definition (3.10) and not by using the recursive relation (3.11).

3.2 Neumann Iteration

The simplest iterative procedure we consider results from the choice

$$\alpha_n = 1, \quad v_n = r_{n-1} \tag{3.18}$$

in (3.9) and (3.11). This gives rise to the well-known Neumann or Picard-Poincaré-Neumann iteration

$$\begin{aligned} u_n &= u_{n-1} + r_{n-1} = f + (I - L)\, u_{n-1} \\ r_n &= r_{n-1} - L r_{n-1} = (I - L)\, r_{n-1} \end{aligned} \tag{3.19}$$

from which we may deduce that

$$\begin{aligned} u_n &= \sum_{m=0}^{n-1} (I - L)^m f + (I - L)^n u_0 \\ r_n &= (I - L)^n r_0 \end{aligned} \tag{3.20}$$

It is well known that a condition sufficient to ensure convergence of this iterative process is $\sigma(I - L) < 1$. It should be noted that this is not a necessary condition since there exist examples where the Neumann series converges but $\sigma(I - L) = 1$ (PATTERSON, 1974). Nevertheless for most operators of interest in scattering problems the Neumann series does not converge; one notable exception is the Born series for weak scatterers (e.g., KLEINMAN, ROACH and VAN DEN BERG, 1990).

The numerical performance of this method in this case is illustrated for scattering by a slab ((3.14)–(3.16)). Figure 3.2 shows that the method converges for sufficiently low contrasts but diverges when the contrast increases sufficiently. WESTON (1985) has shown that for the homogeneous slab problem of Section 3.1 a sufficient condition for convergence is $kl|\chi| < 2$. From Fig. 3.2, we observe that for the value of $\frac{l}{\lambda} = 0.5$, $\chi = \frac{2}{kl} = \frac{2}{\pi}$ and the Neumann series is still convergent, so that the theoretical bound is seen to be overly restrictive.

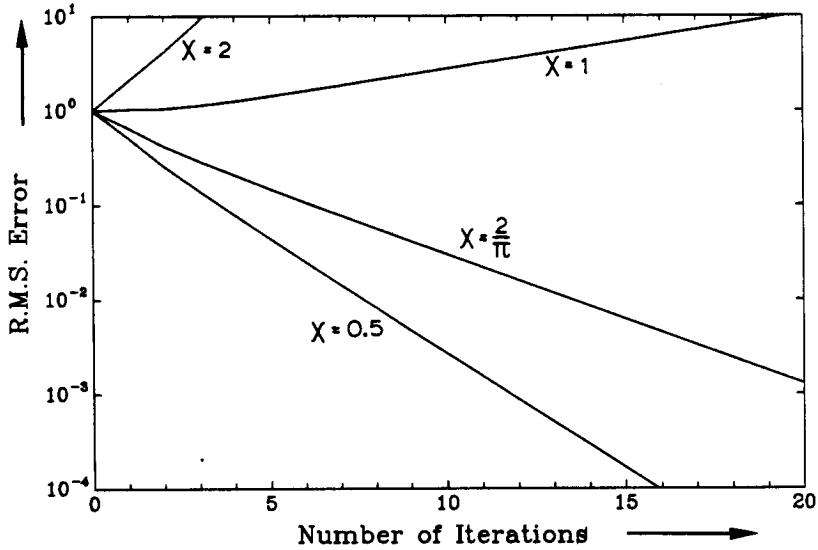

Figure 3.2 Results of the Neumann iteration; homogeneous slab with $\frac{l}{\lambda} = 0.5$.

3.3 The General Iterative Procedure

All of the methods considered subsequently in this chapter may be considered as generalizations of the Neumann iteration based on more elaborate choices of α_n and v_n in (3.9) and (3.11). These choices will be guided by the desire to minimize the error in approximating u by u_n in one of the norms described in Section 3.1. Denoting by u the exact solution of (3.1) the error is then given by $\|u - u_n\|_i$. It is straightforward to verify the following identity

$$\|u-u_n-zw\|_i^2 - \|u-u_n\|_i^2 = \left|\bar{z}\|w\|_i - \frac{\langle u-u_n, w\rangle_i}{\|w\|_i}\right|^2 - \frac{|\langle u-u_n, w\rangle_i|^2}{\|w\|_i^2} \tag{3.21}$$

for every complex number $z \in C$ and $w \in H$. We now prove one of the essential theorems underlying many approximation methods including those considered here. Let M denote a subspace of H. Then we have *Theorem* 4.1: u_n minimizes $\|u-u_n\|_i$ on M if and only if $\langle u-u_n, w\rangle_i = 0$ for every $w \in M$, that is

$$\|u-u_n\|_i \leq \|u-u_n-w\|_i \ \forall w \in M \iff \langle u-u_n, w\rangle_i = 0 \ \forall w \in M$$

Proof. It is easily seen that

$$\|u - u_n - zw\|_i^2 = \|u - u_n\|_i^2 - 2\mathrm{Re}\{\overline{z}\langle u - u_n, w\rangle_i\} + |z|^2\|w\|_i^2 \quad (3.22)$$

Thus if $\langle u - u_n, w\rangle_i = 0$

$$\|u - u_n - zw\|_i^2 = \|u - u_n\|_i^2 + |z|^2\|w\|_i^2 \geq \|u - u_n\|_i^2 \quad (3.23)$$

and this remains true if $z = 1$.

On the other hand if

$$\|u - u_n - w\|_i^2 \geq \|u - u_n\|_i^2 \quad \forall w \in M \quad (3.24)$$

then, replacing w by zw,

$$\|u - u_n - zw\|_i^2 \geq \|u - u_n\|_i^2 \quad \forall w \in M \text{ and } z \in C \quad (3.25)$$

But on using (3.21) it then follows that

$$\left|\overline{z}\|w\|_i - \frac{\langle u - u_n, w\rangle_i}{\|w\|_i}\right|^2 - \frac{|\langle u - u_n, w\rangle_i|^2}{\|w\|_i^2} \geq 0 \quad \forall z \in C \quad (3.26)$$

In particular choosing

$$\overline{z} = \frac{\langle u - u_n, w\rangle_i}{\|w\|_i} \quad (3.27)$$

we have

$$-\left|\frac{\langle u - u_n, w\rangle_i}{\|w\|_i}\right|^2 \geq 0 \quad (3.28)$$

which implies that

$$\langle u - u_n, w\rangle_i = 0 \quad (3.29)$$

This completes the proof.

The iterative processes we will consider consist of minimizing $\|u - u_n\|_i$ on some subspace M for some values of n.

We now specify more explicitly the nature of the functions v_n which occur in our general iteration of (3.9)–(3.11). As in Chapter 2 we define

$$\begin{aligned} v_1 &= Tr_0 \\ v_n &= Tr_{n-1} + \sum_{m=1}^{n-1} \gamma_{nm} v_m, \quad n > 1 \end{aligned} \quad (3.30)$$

where T is a linear operator and γ_{nm} are constants, all yet to be chosen. Choosing v_n in this way is clearly equivalent to

$$v_n = Tr_{n-1} + \sum_{m=1}^{n-1} c_{nm}\, Tr_{m-1} \tag{3.31}$$

where the constants c_{nm} and γ_{nm} are related. The iteration process can then be written

$$\begin{aligned} &u_0 \quad \text{arbitrary} \\ &u_n = u_{n-1} + \sum_{m=1}^{n} \alpha_{nm} Tr_{m-1}, \quad n \geq 1 \end{aligned} \tag{3.32}$$

where $\alpha_{nn} = \alpha_n$, $\alpha_{nm} = \alpha_n c_{nm}$, $m < n$. The residual satisfies the recursion relation

$$r_n = r_{n-1} - \sum_{m=1}^{n} \alpha_{nm}\, LTr_{m-1} \tag{3.33}$$

Various choices of the constants α_{nm} and the operator T will result in a number of iterative procedures.

Yet another representation of v_n may be obtained from (3.30) by observing that there exist constants d_{nm} such that

$$v_n = \sum_{m=1}^{n} d_{nm}\, T(LT)^{m-1} r_0, \quad n \geq 1 \tag{3.34}$$

This may established by mathematical induction as follows. Clearly (3.34) holds for $n = 1$ with $d_{11} = 1$. Now assume that (3.34) holds for $n \leq N$. Then with (3.13) we have, for $n \leq N$,

$$r_n = r_0 - \sum_{m=1}^{n} \alpha_m L \sum_{l=1}^{m} d_{ml}\, T(LT)^{l-1} r_0 \tag{3.35}$$

and since

$$\sum_{m=1}^{n} \sum_{l=1}^{m} a_{ml} = \sum_{l=1}^{n} \sum_{m=l}^{n} a_{ml} \tag{3.36}$$

it follows that for $1 \leq n \leq N$,

$$r_n = r_0 - \sum_{l=1}^{n} \sum_{m=l}^{n} \alpha_m \, d_{ml} \, (LT)^l r_0 \tag{3.37}$$

or

$$r_n = r_0 - \sum_{l=1}^{n} \beta_{nl} \, (LT)^l r_0 \tag{3.38}$$

where

$$\beta_{nl} = \sum_{m=l}^{n} \alpha_m \, d_{ml} \tag{3.39}$$

Substituting (3.34) and (3.38) in (3.30) with $n = N + 1$ we obtain

$$v_{N+1} = Tr_0 - \sum_{l=1}^{N} \beta_{Nl} \, T(LT)^l \, r_0 + \sum_{m=1}^{N} \gamma_{N+1,m} \sum_{l=1}^{m} d_{ml} \, T(LT)^{l-1} \, r_0 \tag{3.40}$$

and with some rearrangement we have the desired result

$$v_{N+1} = \sum_{l=1}^{N+1} d_{N+1,l} \, T(LT)^{l-1} \, r_0 \tag{3.41}$$

where

$$d_{N+1,l} = \begin{cases} 1 + \sum_{m=1}^{N} \gamma_{N+1,m} \, d_{m1}, & l = 1 \\ -\beta_{N,l-1} + \sum_{m=l}^{N} \gamma_{N+1,m} \, d_{ml}, & 2 \leq l \leq N, \\ -\beta_{NN}, & l = N + 1 \end{cases} \tag{3.42}$$

Thus is established the validity, not only of (3.34) but also of (3.38) for $n \geq 1$. Moreover, using (3.34) together with (3.12) and (3.39) enables us to rewrite the n^{th} iterate as

$$u_n = u_0 + \sum_{l=1}^{n} \beta_{nl} \, T(LT)^{l-1} \, r_0 \tag{3.43}$$

Equation (3.38) shows that this general iteration leads to a representation of the residual error at the n^{th} step which lies in the subspace spanned by $\{(LT)^m r_0, \; m = 0, 1, \cdots, n\}$. Such subspaces spanned by

iterates of a particular function are known as Krylov subspaces, e.g., GOLUB and VAN LOAN (1983, p. 324).

We now examine a number of different examples of this general iterative technique.

3.4 A Stationary Over-relaxation Method

The simplest generalization of the Neumann series results from the choice $\alpha_n = \alpha = \text{const}$ in (3.9) and $\gamma_{nm} = 0$ in (3.30) which means $c_{nm} = 0$ in (3.31). Then (3.9) and (3.11) become

$$u_n = u_{n-1} + \alpha T r_{n-1} = \alpha T f + (I - \alpha TL) u_{n-1} \tag{3.44}$$

$$r_n = r_{n-1} - \alpha LT r_{n-1} = (I - \alpha LT) r_{n-1} \tag{3.45}$$

from which we may deduce that

$$u_n = \sum_{m=0}^{n-1} (I - \alpha TL)^m \alpha T f + (I - \alpha TL)^n u_0 \tag{3.46}$$

$$r_n = (I - \alpha LT)^n r_0 \tag{3.47}$$

In this case the constants β_{nl} in (3.38) and (3.43) become

$$\beta_{nl} = (-1)^{l+1} \frac{n! \alpha^l}{l!(n-l)!} \tag{3.48}$$

Convergence of this method is assured if $\sigma(I - \alpha LT) < 1$. The question of whether it is possible to choose α such that this condition holds depends on $\sigma(LT)$, since $\sigma(I - \alpha LT) = \{1 - \alpha\lambda | \lambda \in \sigma(LT)\}$. KLEINMAN, ROACH, SCHUETZ, SHIRRON and VAN DEN BERG (1990) have shown that for L non-selfadjoint and $T = I$ there will exist such an α if $\sigma(L)$ lies in an angular wedge shaped domain in the complex plane with wedge angle less than π excluding a neighborhood of the origin. That this property of $\sigma(L)$ is met depends on the particular operator L. If L is the domain integral operator which occurs in scattering by a penetrable inhomogeneous object then $\sigma(L)$ meets these conditions (KLEINMAN, ROACH and VAN DEN BERG, 1990). Of course even if it can be shown that there exist values of α for which $\sigma(I - \alpha LT) < 1$ holds, there remains the problem of actually determining such values. It has been proposed that α be chosen to minimize $\|u - u_1\|_i$ on the

subspace $M_0 = \text{sp}\{Tr_0\}$. (Recall the definitions of the various norms and inner products in (3.6), (3.7) et seq). Then Theorem 3.1 implies that

$$\langle u - u_1, Tr_0 \rangle_i = 0 \tag{3.49}$$

and with (3.44)

$$\langle u - u_0 - \alpha Tr_0, Tr_0 \rangle_i = 0 \tag{3.50}$$

hence

$$\alpha = \frac{\langle u - u_0, Tr_0 \rangle_i}{\|Tr_0\|_i^2} \tag{3.51}$$

It should be pointed out that because u is not explicitly known since it is the solution we seek, this value of α cannot be computed for all cases, e.g., when $i = 1$ and $T = I$. However explicit results are available in a number of cases as summarized in Table 3.1, where the quantity minimized, $\|u - u_1\|_i$, is exhibited in terms of the original norm and inner product on H.

i	T	Functional Minimized	α	Iteration
1	L^*	$\|u - u_1\|$	$\frac{\|r_0\|^2}{\|L^* r_0\|^2}$	$u_n = u_{n-1} + \alpha L^* r_{n-1}$ $r_n = r_{n-1} - \alpha L L^* r_{n-1}$
2	I	$\|r_1\|$	$\frac{\langle r_0, Lr_0 \rangle}{\|Lr_0\|^2}$	$u_n = u_{n-1} + \alpha r_{n-1}$ $r_n = r_{n-1} - \alpha L r_{n-1}$
2	L^*	$\|r_1\|$	$\frac{\|L^* r_0\|^2}{\|LL^* r_0\|^2}$	$u_n = u_{n-1} + \alpha L^* r_{n-1}$ $r_n = r_{n-1} - \alpha L L^* r_{n-1}$
if L is selfadjoint and positive				
3	I	$\langle u - u_1, L(u - u_1) \rangle$	$\frac{\|r_0\|^2}{\langle r_0, Lr_0 \rangle}$	$u_n = u_{n-1} + \alpha r_{n-1}$ $r_n = r_{n-1} - \alpha L r_{n-1}$

Table 3.1 Stationary over-relaxation methods.

The choice of α given in Table 3.1 for $i = 2$ and $T = I$ was shown to be remarkably effective both in domain scattering, where it is known that α exists for which $\sigma(I - \alpha L) < 1$ (KLEINMAN, ROACH and VAN DEN BERG, 1990), and in rigid body scattering, where the existence of an appropriate α is not yet proven (KLEINMAN, ROACH, SCHUETZ, SHIRRON and VAN DEN BERG, 1990).

The situation is much different if L is selfadjoint and positive. In that case $\sigma(L)$ is real and since (3.1) is always uniquely solvable then

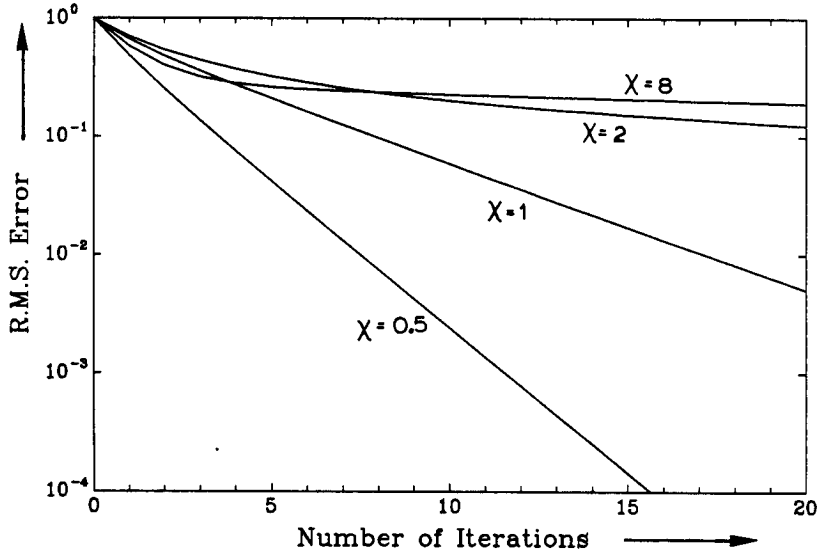

Figure 3.3 Results of the stationary over-relaxation method with $i = 1$ and $T = L^*$; homogeneous slab with $\frac{l}{\lambda} = 0.5$.

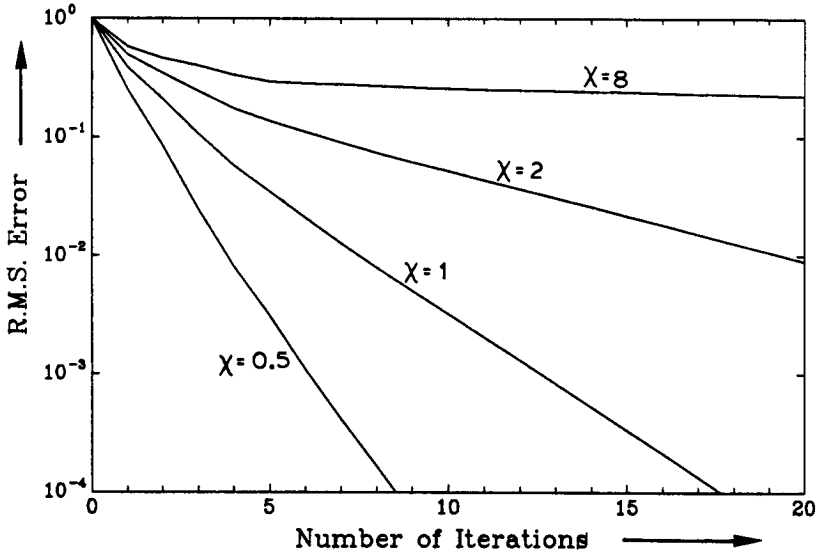

Figure 3.4 Results of the stationary over-relaxation method with $i = 2$ and $T = I$; homogeneous slab with $\frac{l}{\lambda} = 0.5$.

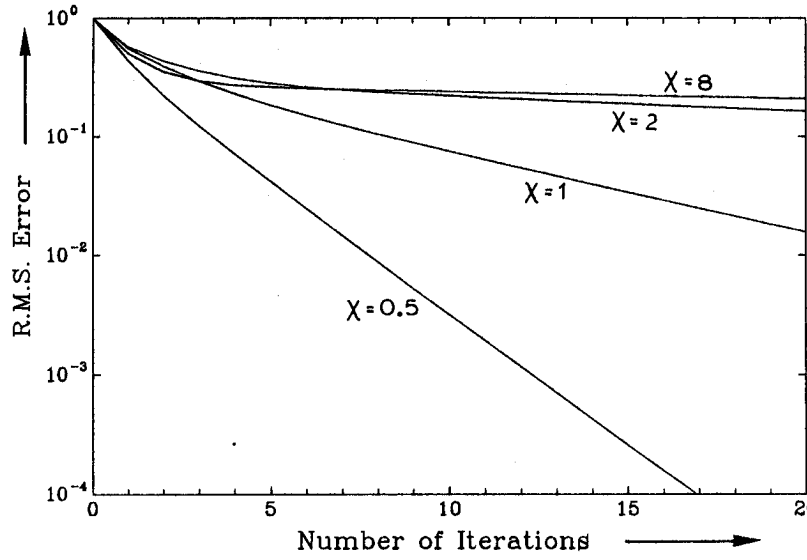

Figure 3.5 Results of the stationary over-relaxation method with $i = 2$ and $T = L^*$; homogeneous slab with $\frac{l}{\lambda} = 0.5$.

there always exists an α $(0 < \alpha < \frac{2}{\|L\|})$ for which the iteration converges for $i = 2$ and $T = I$ (BIALY, 1959; KLEINMAN and ROACH, 1988).

If L is not selfadjoint then the existence of α for which the iteration scheme $i = 2$ and $T = L^*$ indicated in Table 3.1 converges, is assured because of unique solvability and we have convergence if $0 < \alpha < \frac{2}{\|L^*L\|}$. In effect we solve the equation $L^*Lu = L^*f$ rather than (3.1) but they are equivalent since (3.1) is uniquely solvable for all $f \epsilon H$. While it is easier to establish convergence of this iterative scheme than when $T = I$, since less information is needed about $\sigma(L)$, the disadvantage is the appearance of the operators L^*L and LL^* which negatively affect the rate of convergence.

The first three iteration schemes in Table 3.1 are illustrated in the test problem of scattering by a homogeneous slab. Although the same quantity is minimized in different norms, the figures plot the normalized residual error, $\frac{\|r_n\|}{\|f\|}$, in order to have a uniform basis for comparison. The results presented in Figs. 3.3–3.5 show that all three methods represent an improvement over the Neumann iteration of Section 3.2. There the iteration diverged for $\chi = 1$ while all three methods of this section converge for this and even larger contrasts. Of the three

methods, the case when $i = 2$ and $T = I$ gives demonstrably better results as can be seen by comparing the case when $\chi = 2$. For higher contrasts, e.g., $\chi = 8$, all three methods are roughly equivalent for the first 20 iterations. It should be noted however that not only does the method resulting when $i = 2$ and $T = I$ converge more rapidly for low contrasts, but the operation count per iteration is one half that required in the other cases.

If there exists an α for which any of the iteration methods defined in Table 3.1 converges, then in fact there will be a domain of α values which ensure convergence and an optimal value which maximizes the rate of convergence. Failure to be close to optimal can have serious numerical consequences as is illustrated in (KLEINMAN, ROACH, SCHUETZ, SHIRRON and VAN DEN BERG, 1990). One way to avoid the necessity of estimating the optimal α is found in the next method.

3.5 Successive Over-relaxation Method

The next specialization of the general iterative method results from again choosing $\gamma_{nm} = 0$ in (3.30) but letting α_n vary with n in (3.9). Then (3.9) and (3.11) become

$$u_n = u_{n-1} + \alpha_n T r_{n-1} = \alpha_n T f + (I - \alpha_n TL) u_{n-1} \tag{3.52}$$

$$r_n = r_{n-1} - \alpha_n LT r_{n-1} = (I - \alpha_n LT) r_{n-1} \tag{3.53}$$

from which we may deduce that

$$u_n = \alpha_n T f + \sum_{m=1}^{n-1} \left[\prod_{l=m+1}^{n} (I - \alpha_l TL) \right] \alpha_m T f + \left[\prod_{m=1}^{n} (I - \alpha_m TL) \right] u_0 \tag{3.54}$$

$$r_n = \left[\prod_{m=1}^{n} (I - \alpha_m LT) \right] r_0 \tag{3.55}$$

In this case the constants β_{nm} in (3.38) and (3.43) may be shown to be given iteratively by

$$\begin{aligned} \beta_{11} &= \alpha_1 \\ \beta_{n+1,m} &= \begin{cases} \beta_{n1} + \alpha_{n+1}, & m = 1 \\ \beta_{nm} - \alpha_{n+1} \beta_{n,m-1}, & 1 < m \leq n \\ -\alpha_{n+1} \beta_{nn}, & m = n+1 \end{cases} \end{aligned} \tag{3.56}$$

The question of how to choose the constants α_n must be answered before the iteration is fully defined. In keeping with the general procedure outlined in Section 3.3 we choose α_n so that u_n minimizes $\|u - u_n\|_i$ on some subspace. In this case we choose a different subspace for each n, namely,

$$M_n = \text{sp}\{Tr_{n-1}\} \tag{3.57}$$

Then Theorem 3.1 implies that

$$\langle u - u_n, Tr_{n-1}\rangle_i = 0 \tag{3.58}$$

and with (3.52)

$$\langle u - u_{n-1} - \alpha_n Tr_{n-1}, Tr_{n-1}\rangle_i = 0 \tag{3.59}$$

hence

$$\alpha_n = \frac{\langle u - u_{n-1}, Tr_{n-1}\rangle_i}{\|Tr_{n-1}\|_i^2} \tag{3.60}$$

As with the stationary method of the previous section these constants cannot always be calculated since the solution u is not known. However explicit results are available in a number of cases as summarized in Table 3.2 where the quantity minimized, $\|u - u_n\|_i$, is exhibited in terms of the original norm and inner product on H.

i	T	Functional Minimized	α_n	Iteration
1	$L^\star$	$\|u - u_n\|$	$\dfrac{\|r_{n-1}\|^2}{\|L^\star r_{n-1}\|^2}$	$u_n = u_{n-1} + \alpha_n L^\star r_{n-1}$ $r_n = r_{n-1} - \alpha_n L L^\star r_{n-1}$
2	I	$\|r_n\|$	$\dfrac{\langle r_{n-1}, Lr_{n-1}\rangle}{\|Lr_{n-1}\|^2}$	$u_n = u_{n-1} + \alpha_n r_{n-1}$ $r_n = r_{n-1} - \alpha_n Lr_{n-1}$
2	$L^\star$	$\|r_n\|$	$\dfrac{\|L^\star r_{n-1}\|^2}{\|LL^\star r_{n-1}\|^2}$	$u_n = u_{n-1} + \alpha_n L^\star r_{n-1}$ $r_n = r_{n-1} - \alpha_n LL^\star r_{n-1}$
if L is selfadjoint and positive				
3	I	$\langle u - u_n, L(u - u_n)\rangle$	$\dfrac{\|r_{n-1}\|^2}{\langle r_{n-1}, Lr_{n-1}\rangle}$	$u_n = u_{n-1} + \alpha_n r_{n-1}$ $r_n = r_{n-1} - \alpha_n Lr_{n-1}$

Table 3.2 Successive over-relaxation methods.

Convergence of the various iterative schemes illustrated in Table 3.2 is established more easily in some cases than in others. Observe

that with the iteration defined by (3.52) it follows that

$$\begin{aligned}\|u-u_n\|_i^2 &= \|u-u_{n-1}-\alpha_n T r_{n-1}\|_i^2 \\ &= \|u-u_{n-1}\|_i^2 + |\alpha_n|^2 \|Tr_{n-1}\|_i^2 \\ &\quad - 2\mathrm{Re}\{\overline{\alpha}_n \langle u-u_{n-1}, Tr_{n-1}\rangle_i\}\end{aligned} \tag{3.61}$$

and with the choice of α_n in (3.60),

$$\|u-u_n\|_i^2 = \|u-u_{n-1}\|_i^2 \left(1 - \frac{|\langle u-u_{n-1}, Tr_{n-1}\rangle_i|^2}{\|u-u_{n-1}\|_i^2 \|Tr_{n-1}\|_i^2}\right) \tag{3.62}$$

Convergence will then be assured if

$$\frac{|\langle u-u_{n-1}, Tr_{n-1}\rangle_i|^2}{\|u-u_{n-1}\|_i^2 \|Tr_{n-1}\|_i^2} \geq c_i(T) > 0 \quad \forall n \tag{3.63}$$

Then we have

$$\|u-u_n\|_i^2 \leq \|u-u_{n-1}\|_i^2 \,(1-c_i(T)) \leq \|u-u_0\|_i^2 \,(1-c_i(T))^n \tag{3.64}$$

Table 3.3 presents values of $c_i(T)$ that have been obtained for the iterative methods listed in Table 3.2. All of the estimates $c_i(T)$ in Table 3.3 may be obtained by use of the relations in (3.2)–(3.5).

i	T	$\dfrac{\lvert\langle u-u_{n-1}, Tr_{n-1}\rangle_i\rvert^2}{\|u-u_{n-1}\|_i^2 \,\|Tr_{n-1}\|^2}$	$c_i(T)$
1	L^*	$\dfrac{\|r_{n-1}\|^4}{\|L^{-1}r_{n-1}\|^2 \,\|L^*r_{n-1}\|^2}$	$\dfrac{1}{\|L\|^2 \,\|L^{-1}\|^2}$
2	I	$\dfrac{\lvert\langle r_{n-1}, Lr_{n-1}\rangle\rvert^2}{\|r_{n-1}\|^2 \,\|Lr_{n-1}\|^2}$	see below
2	L^*	$\dfrac{\|L^*r_{n-1}\|^4}{\|r_{n-1}\|^2 \,\|LL^*r_{n-1}\|^2}$	$\dfrac{1}{\|L\|^2 \,\|L^{-1}\|^2}$
if L is selfadjoint and positive			
3	I	$\dfrac{\|r_{n-1}\|^4}{\langle L^{-1}r_{n-1}, r_{n-1}\rangle \langle r_{n-1}, Lr_{n-1}\rangle}$	$\dfrac{1}{\|L\| \,\|L^{-1}\|}$

Table 3.3 Convergence factors for successive over-relaxation.

The case $i = 2$ and $T = I$ requires special consideration. If L is selfadjoint and positive we may write $L = L^{\frac{1}{2}} L^{\frac{1}{2}}$, where $L^{\frac{1}{2}}$ also is selfadjoint and positive (RIESZ and NAGY, 1955, p. 265). Then it is straightforward to show

$$\frac{|\langle r_{n-1}, Lr_{n-1}\rangle|^2}{\|r_{n-1}\|^2\|Lr_{n-1}\|^2} = \frac{\|L^{\frac{1}{2}} r_{n-1}\|^4}{\|r_{n-1}\|^2\|L^{\frac{1}{2}} L^{\frac{1}{2}} r_{n-1}\|^2}$$

$$\geq \frac{\| L^{\frac{1}{2}} r_{n-1}\|^2}{\|r_{n-1}\|^2\|L^{\frac{1}{2}}\|^2} \geq \frac{1}{\|L^{\frac{1}{2}}\|^2\|L^{\frac{-1}{2}}\|^2} = \frac{1}{\|L\|\|L^{-1}\|} \tag{3.65}$$

Hence, for this case, the expression of $c_i(T)$ is equal to the one for the case $i = 3, T = I$ and L is selfadjoint and positive.

If L is not positive selfadjoint the situation is not so straightforward. One way to approach the convergence question is through the concept of the numerical range of L defined as (KATO, 1966, p. 267) the set of complex numbers

$$\Theta = \{\langle \overline{v, Lv}\rangle : v \in H,\ \|v\| = 1\} \tag{3.66}$$

Then defining the lower bound

$$m := \inf_{z \in \Theta} \{|z|\} \tag{3.67}$$

it is easy to see that

$$\frac{|\langle r_{n-1}, Lr_{n-1}\rangle|^2}{\|r_{n-1}\|^2\|Lr_{n-1}\|^2} \geq \frac{m^2}{\|L\|^2} = c_2(I) \tag{3.68}$$

If $m > 0$ then convergence is assured. However in order for this convergence proof to apply it must be demonstrated that $m > 0$ for each particular L. Another way to approach convergence in this case is to note that, since

$$r_n = (I - \alpha_n L)\, r_{n-1} \tag{3.69}$$

$$\|u - u_n\|_2 = \|r_n\| = \|(I - \alpha_n L) r_{n-1}\| \tag{3.70}$$

Since α_n was chosen to minimize $\|r_n\|$ it follows that

$$\|u - u_n\|_2 = \|r_n\| \leq \|(I - \alpha L) r_{n-1}\| \leq \|I - \alpha L\|\, \|r_{n-1}\| \tag{3.71}$$

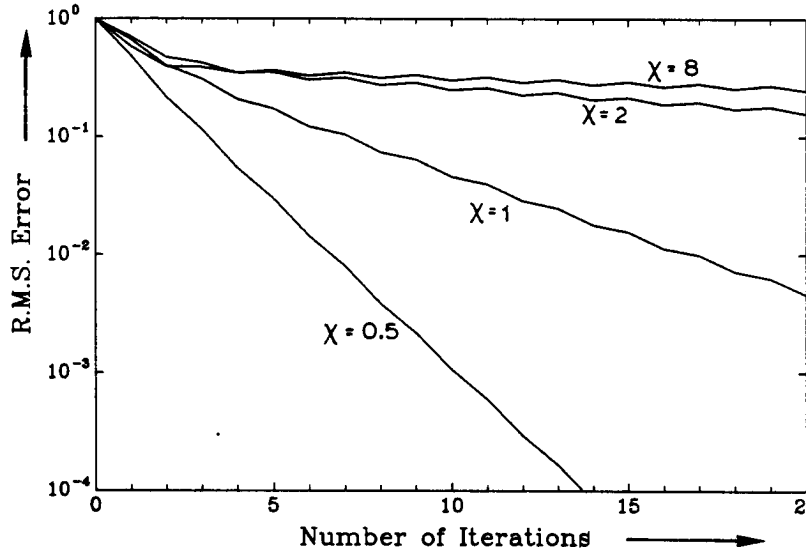

Figure 3.6 Results of the successive over-relaxation method with $i = 1$ and $T = L^*$; homogeneous slab with $\frac{l}{\lambda} = 0.5$.

for every $\alpha \in C$. Repeating this process we find

$$\|r_n\| \leq \|I - \alpha L\|^n \, \|r_0\| \tag{3.72}$$

Convergence will be assured if there exists an α such that

$$\|I - \alpha L\| < 1 \tag{3.73}$$

This is a stronger condition than $\sigma(I - \alpha L) < 1$ which was sufficient for convergence of the stationary method of Section 3.4. (It is true that $\sigma(I - \alpha L) < 1 \Rightarrow \|(I - \alpha L)^n\| < 1$ for some $n \geq 1$, but not necessarily $n = 1$, which is needed for the above convergence proof.) An important difference between the successive and stationary relaxation schemes is that while the *existence* of α such that (3.73) is satisfied will guarantee convergence of the successive iteration, explicit determination of this value of α is not needed in order to actually carry out the iterative procedure. This is in contrast with the stationary scheme where α is needed to implement the process. Again we remark that in order for this convergence proof to apply it must be shown that there exists an α for which (3.73) holds for each particular L.

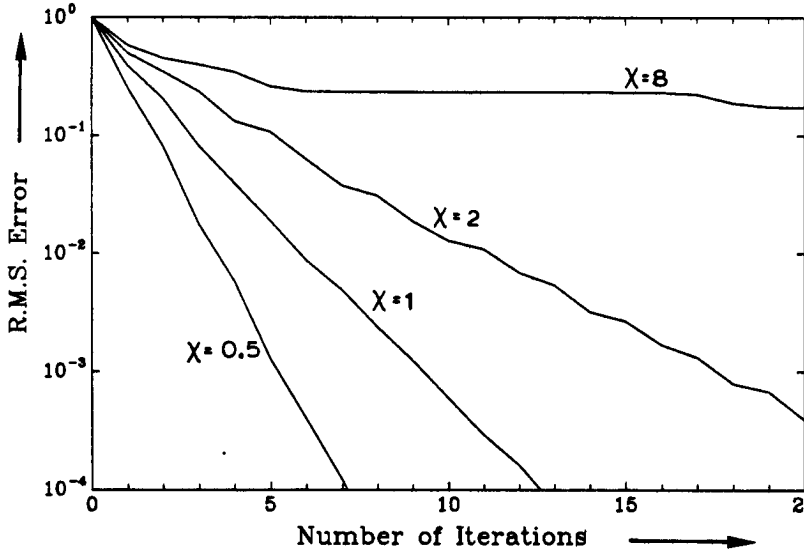

Figure 3.7 Results of the successive over-relaxation method with $i = 2$ and $T = I$; homogeneous slab with $\frac{l}{\lambda} = 0.5$.

The first three iterations schemes in Table 3.2 are illustrated in the test problem of scattering by a homogeneous slab. As before all of the figures plot $\frac{\|r_n\|}{\|f\|}$ as a function of n even though $\|r_n\|$ is not the quantity minimized when $i = 1$. Figures 3.6 - 3.8 show that in all cases the successive over-relaxation method converges faster than the stationary method of Section 3.4 with one exception, when $i = 1$ and $T = I$ at high contrast. Indeed for that case the residual error does not monotonically decrease with the number of iterations. This is attributable to the fact that we have minimized $\|u - u_n\|$ rather than $\|r_n\|$ at each step. The figures also indicate that again the case $i = 2$ and $T = I$ converges most rapidly but now it is also seen that the case $i = 2$ and $T = L^*$ converges faster than $i = 1$ and $T = L^*$. Moreover the results for $i = 2$ and $T = L^*$ are seen to decrease more smoothly though not more rapidly than for $i = 2$ and $T = I$. As before the case $i = 2$ and $T = I$ not only converges fastest but requires only half the operations required by the other two methods at each step. For high contrast, $\chi = 8$, there appears to be no significant improvement in performance over the stationary case.

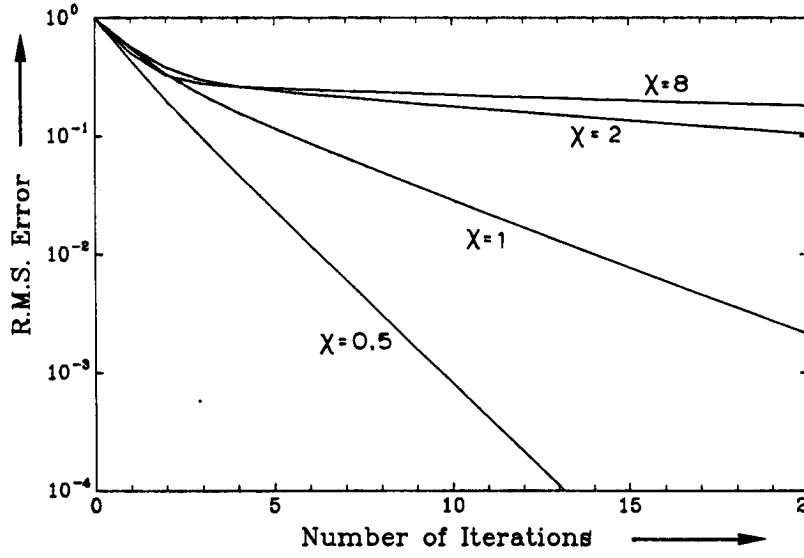

Figure 3.8 Results of the successive over-relaxation method with $i = 2$ and $T = L^*$; homogeneous slab with $\frac{l}{\lambda} = 0.5$.

3.6 Krylov Subspace and Conjugate Gradient Methods

In this section we consider the full Krylov method by which we mean the general iterative method of (3.9) and (3.11) with no prior restrictions on γ_{nm} in (3.30) or α_n in (3.9). Thus, we consider the scheme

$$u_0 \quad \text{arbitrary}$$

$$u_n = u_{n-1} + \alpha_n v_n, \quad n \geq 1 \tag{3.74}$$

$$r_n = r_{n-1} - \alpha_n L v_n, \quad n \geq 1 \tag{3.75}$$

$$v_1 = T r_0 \tag{3.76}$$

$$v_n = T r_{n-1} + \sum_{m=1}^{n-1} \gamma_{nm} v_m, \quad n > 1 \tag{3.77}$$

Again we choose the constants to minimize $\|u - u_n\|_i$ over some subspace. Now however we define the subspace to be the span of all pre-

vious errors. More precisely define

$$H_n = \text{sp}\{v_j\}_{j=1}^n \tag{3.78}$$

Thus H_n is the projection of H onto the span of the first n functions v_j, $j = 1, 2, \cdots$. Equivalently it is easily seen (e.g., (3.31) and (3.34)) that

$$H_n = \text{sp}\{Tr_{j-1}\}_{j=1}^n = \text{sp}\{T(LT)^{j-1}r_0\}_{j=1}^n \tag{3.79}$$

hence H_n is the subspace spanned by the modified residuals $\{Tr_j\}_{j=0}^{n-1}$ as well as the Krylov subspace of modified iterates $\{T(LT)^j r_0\}_{j=0}^{n-1}$. Theorem 3.1 then provides conditions for determining α_n and γ_{nm} so as to minimize $\|u - u_n\|_i$ on H_n, namely

$$\langle u - u_n, v_j \rangle_i = 0, \quad j = 1, \cdots, n \tag{3.80}$$

With (3.74)

$$\langle u - u_{n-1} - \alpha_n v_n, v_j \rangle_i = 0, \quad j = 1, \cdots, n \tag{3.81}$$

which implies that for $j = n$

$$\alpha_n = \frac{\langle u - u_{n-1}, v_n \rangle_i}{\|v_n\|_i^2} \tag{3.82}$$

and for $j < n$

$$\langle v_n, v_j \rangle_i = 0 \tag{3.83}$$

Substituting the expression for v_n of (3.77) in this last equation yields

$$\langle Tr_{n-1} + \sum_{m=1}^{n-1} \gamma_{nm} v_m, v_j \rangle_i = 0, \quad j = 1, \cdots, n-1 \tag{3.84}$$

But with (3.83) this becomes

$$\langle Tr_{n-1} + \gamma_{nj} v_j, v_j \rangle_i = 0 \tag{3.85}$$

hence

$$\gamma_{nj} = -\frac{\langle Tr_{n-1}, v_j \rangle_i}{\|v_j\|_i^2}, \quad j = 1, \cdots, n-1 \tag{3.86}$$

Thus the iteration method is completely defined with (3.74)–(3.77) together with the definitions of α_n and γ_{nj} in (3.82) and (3.86).

An alternate definition of α_n may be obtained with (3.77) and (3.82)

$$\alpha_n = \frac{\langle u - u_{n-1}, Tr_{n-1} + \sum_{m=1}^{n-1} \gamma_{nm} v_m \rangle_i}{\langle v_n, Tr_{n-1} + \sum_{m=1}^{n-1} \gamma_{nm} v_m \rangle_i} = \frac{\langle u - u_{n-1}, Tr_{n-1} \rangle_i}{\langle v_n, Tr_{n-1} \rangle_i} \tag{3.87}$$

where the orthogonality relations of (3.80) and (3.83) have been employed. Note also that the definition of v_n of (3.77) together with the orthogonality relation of (3.80) implies that

$$\langle u - u_n, Tr_{j-1} \rangle_i = 0, \quad j = 1, \cdots, n \tag{3.88}$$

This relation (with n replaced by $n-1$) together with (3.74) leads to the fact that

$$\langle v_n, Tr_{j-1} \rangle_i = 0, \quad j = 1, \cdots, n-1 \tag{3.89}$$

while (3.83) together with the definition of (3.77) implies that

$$\langle v_n, Tr_{n-1} \rangle_i = \|v_n\|_i^2 \tag{3.90}$$

hence we may rewrite α_n of (3.87) as

$$\alpha_n = \frac{\langle u - u_{n-1}, Tr_{n-1} \rangle_i}{\|v_n\|_i^2} = \frac{\langle L^{-1} r_{n-1}, Tr_{n-1} \rangle_i}{\|v_n\|_i^2} \tag{3.91}$$

Explicit examples of the various iterative schemes available using the full Krylov method are exhibited in Table 3.4. As before the quantity minimized, $\|u - u_n\|_i$, is shown in terms of the original norm and inner product on H. All of the methods employ the iteration $u_n = u_{n-1} + \alpha_n v_n$, $r_n = r_{n-1} - \alpha_n L v_n$ with different definitions of α_n and v_n. Krylov subspace methods for matrix equations have been developed by ELMAN (1982) where they are called generalized conjugate residual methods, see SAAD and SCHULTZ (1986) for a discussion of this and related methods.

i	T	Functional Minimized	v_n	α_n
1	$L^\star$	$\|u-u_n\|$	$L^\star r_{n-1} - \sum_{m=1}^{n-1} \frac{\langle L^\star r_{n-1}, v_m\rangle}{\|v_m\|^2} v_m$	$\frac{\|r_{n-1}\|^2}{\|v_n\|^2}$
2	I	$\|r_n\|$	$r_{n-1} - \sum_{m=1}^{n-1} \frac{\langle Lr_{n-1}, Lv_m\rangle}{\|Lv_m\|^2} v_m$	$\frac{\langle r_{n-1}, Lr_{n-1}\rangle}{\|Lv_n\|^2}$
2	$L^\star$	$\|r_n\|$	$L^\star r_{n-1} - \sum_{m=1}^{n-1} \frac{\langle LL^\star r_{n-1}, Lv_m\rangle}{\|Lv_m\|^2} v_m$	$\frac{\|L^\star r_{n-1}\|^2}{\|Lv_n\|^2}$
if L is selfadjoint and positive				
3	I	$\langle u-u_n, L(u-u_n)\rangle$	$r_{n-1} - \sum_{m=1}^{n-1} \frac{\langle r_{n-1}, Lv_m\rangle}{\langle v_m, Lv_m\rangle} v_m$	$\frac{\|r_{n-1}\|^2}{\langle v_n, Lv_n\rangle}$

Table 3.4 Krylov subspace methods.

The iterative form of the Krylov subspace methods is most convenient for numerical application but may mask the fact that they represent solutions of linear systems of equations whose coefficient matrices are the Gram matrices of the subspaces. To see this we may use the form of u_n given in (3.32) and choose the constants α_{nm} to minimize $\|u-u_n\|_i^2$, which is exactly what we did before using the iterative form of v_n rather than an expansion in $\{Tr_m\}_{m=0}^{n-1}$. Then

$$\begin{aligned}\|u-u_n\|_i^2 = \|u-u_{n-1}\|_i^2 + \sum_{m=1}^{n}\sum_{l=1}^{n} \alpha_{nm}\overline{\alpha}_{nl}\langle Tr_{m-1}, Tr_{l-1}\rangle_i \\ - 2\mathrm{Re}\{\sum_{l=1}^{n} \overline{\alpha}_{nl}\langle u-u_{n-1}, Tr_{l-1}\rangle_i\}\end{aligned} \tag{3.92}$$

and conditions on α_{nm} such that this is a minimum are that

$$\sum_{m=1}^{n} \alpha_{nm}\langle Tr_{m-1}, Tr_{l-1}\rangle_i = \langle u-u_{n-1}, Tr_{l-1}\rangle_i, \quad l=1,\cdots,n \tag{3.93}$$

Alternatively using the same error minimization with u_n as given in

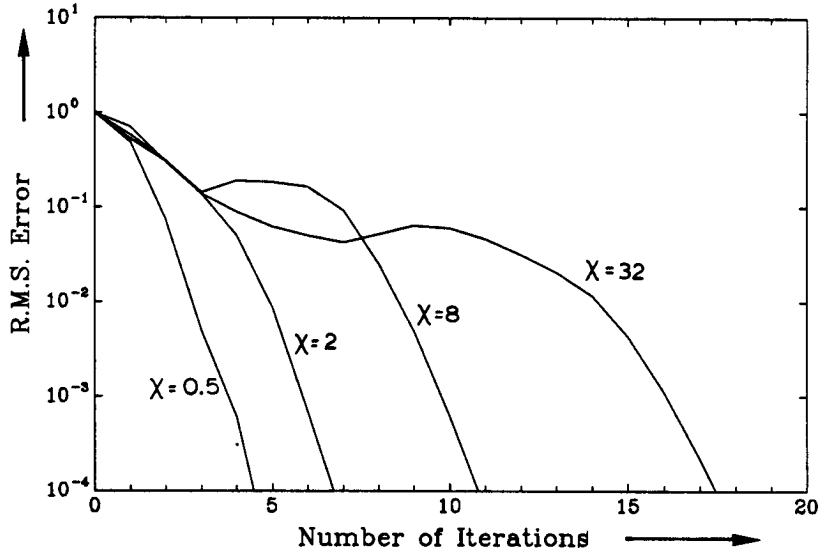

Figure 3.9 Results of the Krylov subspace method with $i = 1$ and $T = L^*$; homogeneous slab with $\frac{l}{\lambda} = 0.5$.

(3.43) we find that the coefficients β_{nm} must satisfy the system

$$\sum_{m=1}^{n} \beta_{nm} \langle T(LT)^{m-1} r_0, T(LT)^{l-1} r_0 \rangle_i = \langle u - u_0, T(LT)^{l-1} r_0 \rangle_i, \qquad l = 1, \cdots, n \tag{3.94}$$

This system is always solvable unless the actual solution u is a linear combination of $u_0 \cup \{T(LT)^{l-1} r_0,\ l = 1, \cdots, n-1\}$ in which case there exist coefficients $\beta_{n-1,m}$ (and hence $\alpha_{n-1,m}$) such that $u = u_{n-1}$ and $r_{n-1} = 0$. It should be noted that though the systems in (3.93) and (3.94) are formally equivalent to the iterative solution, experience has indicated a considerable preference for the iterative solution since the numerical solution of the linear systems exhibits far greater loss of accuracy due to round off than does the iterative solution.

The first three iteration methods in Table 3.4 are illustrated in the test problem of scattering by a homogeneous slab (see Figs. 3.9-3.11). All three methods show a dramatic improvement over the results of the previous sections. Contrasts of $\chi = 32$ can be handled with less than 20 iterations. The case when $i = 2$ and $T = I$ still exhibits the most rapid convergence while still requiring about half the operations

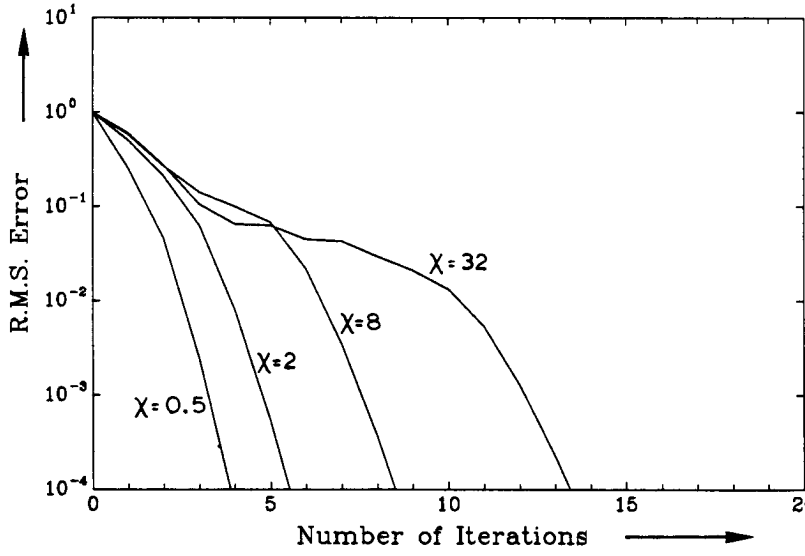

Figure 3.10 Results of the Krylov subspace method with $i = 2$ and $T = I$; homogeneous slab with $\frac{l}{\lambda} = 0.5$.

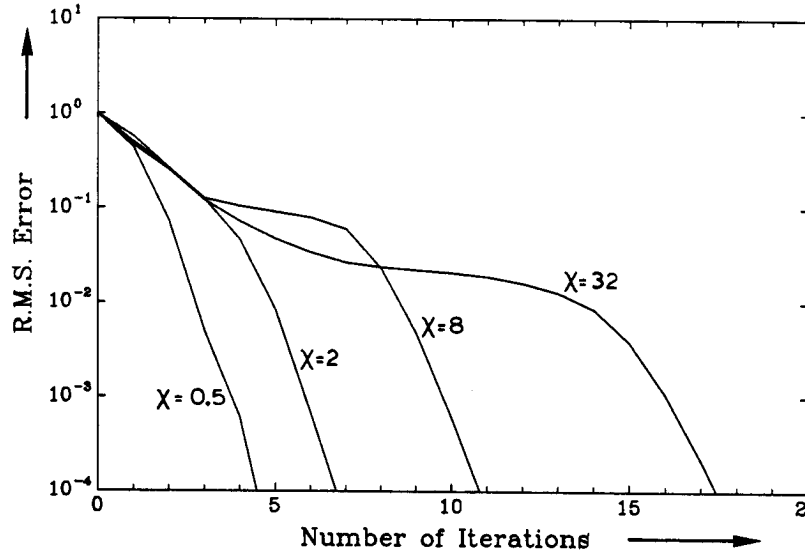

Figure 3.11 Results of the Krylov subspace method with $i = 2$ and $T = L^*$; homogeneous slab with $\frac{l}{\lambda} = 0.5$.

at each step. All three methods require that the directions v_m for all previous m be stored. When $i = 2$ it is also required to store or to compute Lv_m. Note that as before, the scheme for $i = 2$ and $T = L^*$ does not produce a monotonically decreasing residual error.

Considerable simplification occurs if TL is selfadjoint with respect to $\langle \cdot, \cdot \rangle_i$. Selfadjointness of TL with respect to the various inner products carries different meanings when interpreted with respect to $\langle \cdot, \cdot \rangle_1$. In particular: TL selfadjoint with respect to $\langle \cdot, \cdot \rangle_2 \Leftrightarrow LT$ selfadjoint with respect to $\langle \cdot, \cdot \rangle_1, TL$ selfadjoint with respect to $\langle \cdot, \cdot \rangle_3 \Leftrightarrow T$ selfadjoint with respect to $\langle \cdot, \cdot \rangle_1$. Then rewriting (3.88) we have

$$\langle L^{-1}r_n, Tr_{j-1} \rangle_i = \langle L^{-1}r_n, TLL^{-1}r_{j-1} \rangle_i = \langle Tr_n, L^{-1}r_{j-1} \rangle_i = 0, \quad j = 1, \cdots, n \tag{3.95}$$

But with (3.75)

$$\langle Tr_n, L^{-1}(r_{j-2} - \alpha_{j-1}Lv_{j-1}) \rangle_i = 0 \tag{3.96}$$

which implies that

$$\langle Tr_n, v_{j-1} \rangle_i = 0, \quad j = 1, \cdots, n \tag{3.97}$$

where we have again made use of (3.95). Equation (3.97) may be rewritten as

$$\langle Tr_{n-1}, v_j \rangle_i = 0, \quad j = 1, \cdots, n-2 \tag{3.98}$$

Using this fact in the definition of γ_{nj} of (3.86), we find that the expression for v_n simplifies, in this case where TL is selfadjoint, to

$$v_n = Tr_{n-1} - \frac{\langle Tr_{n-1}, v_{n-1} \rangle_i}{\|v_{n-1}\|_i^2} v_{n-1} \tag{3.99}$$

In fact TL is selfadjoint with respect to $\langle \cdot, \cdot \rangle_i$ for all of the cases listed in Table 3.4, except when $i = 2$, $T = I$ and L is not selfadjoint. Then, the Krylov subspace methods become those shown in Table 3.5.

i	T	Functional Minimized	v_n	α_n
1	$L^\star$	$\|u-u_n\|$	$L^\star r_{n-1} - \dfrac{\langle L^\star r_{n-1}, v_{n-1}\rangle}{\|v_{n-1}\|^2} v_{n-1}$	$\dfrac{\|r_{n-1}\|^2}{\|v_n\|^2}$
2	I	$\|r_n\|$	$r_{n-1} - \dfrac{\langle Lr_{n-1}, Lv_{n-1}\rangle}{\|Lv_{n-1}\|^2} v_{n-1}$ only if L is selfadjoint	$\dfrac{\langle r_{n-1}, Lr_{n-1}\rangle}{\|Lv_n\|^2}$
2	$L^\star$	$\|r_n\|$	$L^\star r_{n-1} - \dfrac{\langle LL^\star r_{n-1}, Lv_{n-1}\rangle}{\|Lv_{n-1}\|^2} v_{n-1}$	$\dfrac{\|L^\star r_{n-1}\|^2}{\|Lv_n\|^2}$
if L is selfadjoint and positive				
3	I	$\langle u-u_n, L(u-u_n)\rangle$	$r_{n-1} - \dfrac{\langle r_{n-1}, Lv_{n-1}\rangle}{\langle v_{n-1}, Lv_{n-1}\rangle} v_{n-1}$	$\dfrac{\|r_{n-1}\|^2}{\langle v_n, Lv_n\rangle}$

Table 3.5 Conjugate gradient methods.

An alternate form of the conjugate gradient methods can be obtained by using the fundamental equation (3.74), together with the definition of α_n of (3.82) to yield

$$\begin{aligned}\frac{\langle Tr_{n-1}, v_{n-1}\rangle_i}{\|v_{n-1}\|_i^2} &= \langle Tr_{n-1}, \frac{u_{n-1}-u_{n-2}}{\alpha_{n-1}}\rangle_i \frac{1}{\|v_{n-1}\|_i^2} \\ &= \frac{\langle Tr_{n-1}, u_{n-1}-u_{n-2}\rangle_i}{\langle Tr_{n-2}, u-u_{n-2}\rangle_i}\end{aligned} \tag{3.100}$$

Using the fact $L^{-1}L = I$ and the definition of the residuals we obtain

$$\begin{aligned}\frac{\langle Tr_{n-1}, v_{n-1}\rangle_i}{\|v_{n-1}\|_i^2} &= \frac{\langle Tr_{n-1}, L^{-1}(r_{n-2}-r_{n-1})\rangle_i}{\langle Tr_{n-2}, L^{-1}r_{n-2}\rangle_i} \\ &= -\frac{\langle Tr_{n-1}, L^{-1}r_{n-1}\rangle_i}{\langle Tr_{n-2}, L^{-1}r_{n-2}\rangle_i}\end{aligned} \tag{3.101}$$

where (3.95) has been used. Then in place of (3.99) we have

$$v_n = Tr_{n-1} + \frac{\langle Tr_{n-1}, L^{-1}r_{n-1}\rangle_i}{\langle Tr_{n-2}, L^{-1}r_{n-2}\rangle_i} v_{n-1} \tag{3.102}$$

and Table 3.5 may be replaced by Table 3.6.

i	T	Functional Minimized	v_n	α_n
1	L^*	$\|u-u_n\|$	$L^*r_{n-1}+\frac{\|r_{n-1}\|^2}{\|r_{n-2}\|^2}v_{n-1}$	$\frac{\|r_{n-1}\|^2}{\|v_n\|^2}$
2	I	$\|r_n\|$	$r_{n-1}+\frac{\langle r_{n-1},Lr_{n-1}\rangle}{\langle r_{n-2},Lr_{n-2}\rangle}v_{n-1}$ only if L is selfadjoint	$\frac{\langle r_{n-1},Lr_{n-1}\rangle}{\|Lv_n\|^2}$
2	L^*	$\|r_n\|$	$L^*r_{n-1}+\frac{\|L^*r_{n-1}\|^2}{\|L^*r_{n-2}\|^2}v_{n-1}$	$\frac{\|L^*r_{n-1}\|^2}{\|Lv_n\|^2}$
if L is selfadjoint and positive				
3	I	$\langle u-u_n, L(u-u_n)\rangle$	$r_{n-1}+\frac{\|r_{n-1}\|^2}{\|r_{n-2}\|^2}v_{n-1}$	$\frac{\|r_{n-1}\|^2}{\langle v_n,Lv_n\rangle}$

Table 3.6 Conjugate gradient methods, alternate form.

Note that the expression for v_n when $i=2$ and $T=I$ holds only if L is selfadjoint (not necessarily positive). When L is not selfadjoint then v_n takes the form given in Table 3.4. The algorithm resulting when $i=3$ and $T=I$ is commonly called the conjugate gradient method (see GOLUB and O'LEARY, 1989, for an extensive bibliography). The algorithm resulting when $i=2$ and $T=L^*$ may also be considered the same conjugate gradient method applied to the operator equation $L^*Lu=L^*f$, where the operator L^*L is now selfadjoint. That is replacing L by L^*L and r_n by L^*r_n in the case $i=3$, $T=I$ leads to the case $i=2$, $T=L^*$. In fact if we define the functional $J_i(v)=\|u-v\|_i^2$ on H, where u is the solution of $Lu=f$, then the Gateaux derivative of J_i at u_n is $J_i'(u_n,v)=\lim_{x\to 0}\frac{J_i(u_n+xv)-J_i(u_n)}{x}=2\mathrm{Re}\langle u_n-u,v\rangle_i$ and the gradient of J_i at u_n is u_n-u. In view of (3.80) and (3.83) all of the algorithms in Table 3.5 (or Table 3.6) could be called conjugate gradient methods since the direction of correction v_n at the n^{th} step of the iteration is orthogonal to the gradient of J_i at u_n as well as the previous correction steps. However the term conjugate gradient method is usually meant to also imply that the correction direction, v_n, is defined in terms of the residual and the direction at the previous step and not all of the previous corrections as in the case when $i=2$, $T=I$ and L is not selfadjoint. Note that the first method of Table 3.6 is identical to the one given by SARKAR and ARVAS (1985, case B).

i	T	Functional Minimized	w_n	u_n
1	$L^\star$	$\|u-u_n\|$	$w_{n-1}+\dfrac{L^\star r_{n-1}}{\|r_{n-1}\|^2}$	$u_{n-1}+\dfrac{w_n}{\|w_n\|^2}$
2	I	$\|r_n\|$	$w_{n-1}+\dfrac{r_{n-1}}{\langle r_{n-1}, Lr_{n-1}\rangle}$ only if L is selfadjoint	$u_{n-1}+\dfrac{w_n}{\|Lw_n\|^2}$
2	$L^\star$	$\|r_n\|$	$w_{n-1}+\dfrac{L^\star r_{n-1}}{\|L^\star r_{n-1}\|^2}$	$u_{n-1}+\dfrac{w_n}{\|Lw_n\|^2}$
if L is selfadjoint and positive				
3	I	$\langle u-u_n, L(u-u_n)\rangle$	$w_{n-1}+\dfrac{r_{n-1}}{\|r_{n-1}\|^2}$	$u_{n-1}+\dfrac{w_n}{\langle w_n, Lw_n\rangle}$

Table 3.7 Conjugate gradient methods, simplified form.

The algorithm of the conjugate gradient scheme can be simplified as follows. If we introduce the substitution

$$w_n := \frac{v_n}{\langle Tr_{n-1}, L^{-1}r_{n-1}\rangle_i} \tag{3.103}$$

and substitute it in the recursion relation (3.102), we find

$$w_n = w_{n-1} + \frac{Tr_{n-1}}{\langle Tr_{n-1}, L^{-1}r_{n-1}\rangle_i} \tag{3.104}$$

Using the expression of α_n, Eq. (3.91), we can change the iteration formulas of Eqs. (3.74) and (3.75) as

$$u_n = u_{n-1} + \frac{w_n}{\|w_n\|_i^2} \tag{3.105}$$

$$r_n = r_{n-1} - \frac{Lw_n}{\|w_n\|_i^2} \tag{3.106}$$

The various forms of Table 3.6 may be replaced by the simpler algorithms of Table 3.7.

Convergence of the Krylov and conjugate gradient methods may be established just as in the previous section. It is useful to note first

that with (3.90) and the iterative definition of v_n of (3.77) together with the explicit choice of the constants γ_{nm} of (3.86)

$$\begin{aligned} \|v_n\|_i^2 = \langle v_n, Tr_{n-1}\rangle_i &= \|Tr_{n-1}\|_i^2 + \sum_{m=1}^{n-1} \gamma_{nm}\langle v_m, Tr_{n-1}\rangle_i \\ &= \|Tr_{n-1}\|_i^2 - \sum_{m=1}^{n-1} \frac{|\langle v_m, Tr_{n-1}\rangle_i|^2}{\|v_m\|_i^2} \\ &\leq \|Tr_{n-1}\|_i^2 \end{aligned} \tag{3.107}$$

Now we use the iterative definition of u_n of (3.74) and the explicit choice of α_n given in (3.82) to see that

$$\|u - u_n\|_i^2 = \|u - u_{n-1} - \alpha_n v_n\|_i^2 = \|u - u_{n-1}\|_i^2 - \frac{|\langle u - u_{n-1}, v_n\rangle_i|^2}{\|v_n\|_i^2} \tag{3.108}$$

Next we use the fact that

$$\langle u - u_{n-1}, v_n\rangle_i = \langle u - u_{n-1}, Tr_{n-1}\rangle_i \tag{3.109}$$

which follows from (3.80), together with the inequality derived above, (3.107), to see that

$$\begin{aligned} \|u - u_n\|_i^2 \leq \quad & \|u - u_{n-1}\|_i^2 - \frac{|\langle u - u_{n-1}, Tr_{n-1}\rangle_i|^2}{\|Tr_{n-1}\|_i^2} \\ &= \|u - u_{n-1}\|_i^2 \left(1 - \frac{|\langle u - u_{n-1}, Tr_{n-1}\rangle_i|^2}{\|u - u_{n-1}\|_i^2 \|Tr_{n-1}\|_i^2}\right) \end{aligned} \tag{3.110}$$

which is identical with (3.62). Hence convergence of the Krylov and conjugate gradient methods are assured for all cases, except $i = 2$, $T = I$. In fact we have

$$\|u - u_n\|_i^2 \leq \|u - u_0\|_i^2 \, (1 - c_i(T))^n \tag{3.111}$$

with the same convergence factors $c_i(T)$ as given in Table 3.3

In the special case when $\langle Tr_n, Tr_m\rangle = 0$, $n \neq m$, the full Krylov method reduces to the successive over-relaxation method and thus it will have the same convergence properties. In general however $\langle Tr_n, Tr_m\rangle \neq 0$ in which case it is expected that the full Krylov method will

converge faster than the over-relaxation method, since the correction directions are optimal for the subspace H_n over which we minimize the error. This will not be true for the over-relaxation method.

When $T = I$ and $i = 2$ and L is positive and selfadjoint then (3.65) holds and

$$c_2(I) = \frac{1}{\|L\|\|L^{-1}\|} \tag{3.112}$$

If L is not positive and selfadjoint then just as before ((3.66) et seq) we may obtain

$$c_2(I) = \frac{m^2}{\|L\|^2} \tag{3.113}$$

where m is a lower bound on the magnitude of the numerical range of L. As before, it must be demonstrated that $m > 0$ for each particular L in order that convergence be assured. Another, slightly less restrictive guarantee of convergence whether or not L is positive selfadjoint is provided whenever there exists an α such that $\sigma(I - \alpha L) < 1$. This follows from the fact that by choosing β_{nm} in (3.43) to minimize $\|u - u_n\|_2 = \|r_n\|$, which is an alternate way to view the full Krylov scheme, cf. (3.94), we obtain a value of $\|r_n\|$ that is certainly less than or equal to the value of $\|r_n\|$ that occurs had we chosen β_{mn} to satisfy (3.48), i.e., the choice in the stationary over-relaxation method. But this larger error vanishes as $n \to \infty$ and therefore the Krylov method converges. We remark that it is only necessary to know the existence of an α for which $\sigma(I - \alpha L) < 1$. It is not necessary to actually find α nor is the stronger condition $\|I - \alpha L\| < 1$ required as it was in the successive over-relaxation method.

The cases when $i = 1$, $T = L^\star$ and $i = 2$, $T = L^\star$ in Table 3.6 are illustrated in the test problem of the homogeneous slab. Note that these cases have already been illustrated in Figs. 3.9 and 3.11 using the full Krylov scheme without taking into account the simplifications of (3.99) and (3.102). Thus the results shown in Figs. 3.12 and 3.13 should be identical with Figs. 3.9 and 3.11, if the numerical operations were carried out with infinite precision. However the results are *not* identical as the figures clearly show. It appears that the results are identical for low contrast where the residual error is less than 10^{-4} in less than eight iterations. As the contrast increases the onset of numerical instability is discernible for smaller numbers of iterations. For contrast $\chi = 32$ the results are considerable degraded from those obtained using the full Krylov method. Since L is not selfadjoint there are no results corre-

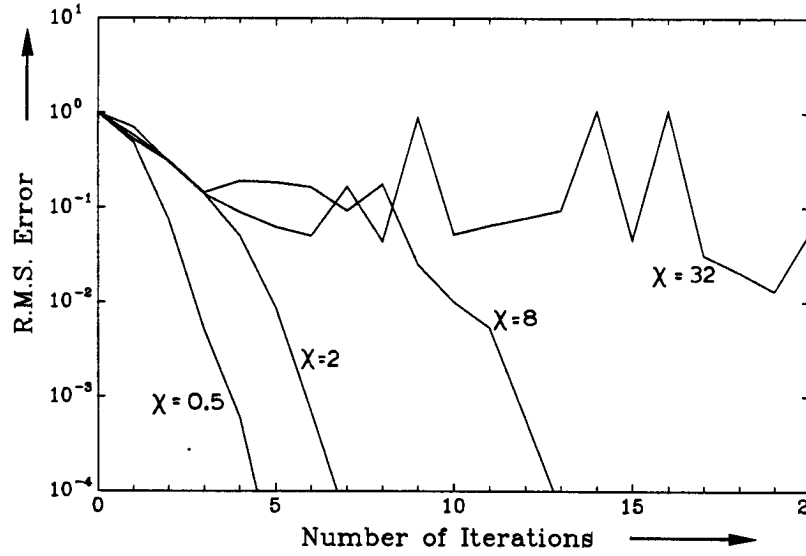

Figure 3.12 Results of the conjugate gradient method with $i = 1$ and $T = L^\star$; homogeneous slab with $\frac{l}{\lambda} = 0.5$.

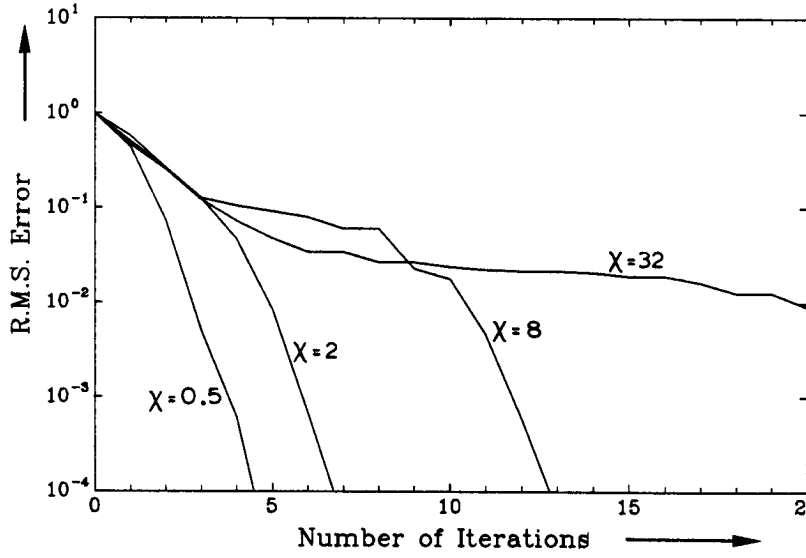

Figure 3.13 Results of the conjugate gradient method with $i = 2$ and $T = L^\star$; homogeneous slab with $\frac{l}{\lambda} = 0.5$.

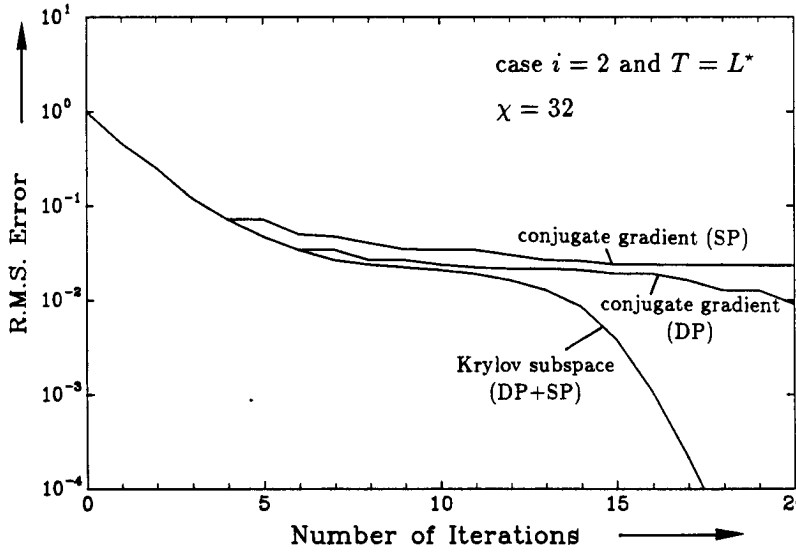

Figure 3.14 Influence of single precision (SP) and double precision (DP) arithmetic in the Krylov subspace method and the conjugate gradient method with $i = 2$ and $T = L^*$; homogeneous slab with $\frac{l}{\lambda} = 0.5$ and $\chi = 32$.

sponding to Fig. 3.10 using a conjugate gradient algorithm. To further investigate the startling degradation in numerical performance when the conjugate gradient scheme is used, we first increased the number of integration points used in our numerical integration. However this had no discernible effect on the results. For the case $i = 2$, $T = L^*$ and $\chi = 32$ we recomputed results for both the Krylov and conjugate gradient schemes using both single and double precision arithmetic. As shown in Fig. 3.14 the results for the Krylov method were almost identical whereas those for the conjugate gradient method display an easily seen difference. This indicates that the theoretical orthogonality used in the conjugate gradient scheme is gradually lost due to loss of significant figures whereas in the Krylov method it is enforced at each step.

3.7 Concluding Remarks

We have described a number of iteration schemes based on error minimization which are gradient methods. They all are generalizations of the Neumann series in which the error is minimized over some subspace. The methods include stationary over-relaxation, in which the relaxation parameter is found by minimizing the residual error in the first iteration step, successive over-relaxation, in which the error at each step is minimized, and Krylov subspace methods, in which the error is minimized over a subspace of all previous errors. Different definitions of the error lead to a number of different schemes including conjugate gradient algorithms.

Numerical examples show that better results come from more sophisticated procedures with the best performance showed by the Krylov method. One surprising result is that if orthogonality which leads to the conjugate gradient method is ignored and the full Krylov scheme is used even when not theoretically necessary, the results obtained are markedly superior to those obtained by the conjugate gradient method when more than a few iterations are employed. This reinforces the widely held belief that in order to be effective the conjugate gradient algorithm must be used together with a good preconditioner which keeps the number of iterations small. However all the methods described benefit from effective preconditioning. Even the simplest iterative scheme would suffice if the preconditioner is sufficient close to the exact inverse. Until now most preconditioners are developed from matrix methods which fail to incorporate asymptotic and approximate solutions. Such solutions have been developed through explicit analysis of the particular equation and take into account the physics of the problem which the equation models and it is suggested that future work in preconditioners use these approximate solutions.

Acknowledgment

This work was supported under NSF Grant No. DMS-8811134, AFOSR Grant-86-0269 and NATO Grant-0230/88.

References

[1] Bialy, H., "Iterative behandlung linearer funktiongleichungen," *Arch. Rational Mech. Anal.*, **4**, 166–176, 1959.

[2] Elman, H. C., *Iterative Methods for Large Sparse Nonsymmetric Systems of Linear Equations*, Ph.D. Thesis, Computer Science Dept., Yale University, 1982.

[3] Golub, G. H., and C. F. Van Loan, *Matrix Computations*, John Hopkins University Press, Baltimore, MD., 1983.

[4] Golub, G. H., and D. P. O'leary, "Some history of the conjugate gradient and Lancos algorithms: 1948-1976," *SIAM Review*, **31**, 50–102, 1989.

[5] Kato, K., *Perturbation Theory for Linear Operators*, Springer-Ver, Berlin, 1966.

[6] Kleinman, R. E., and G. F. Roach, "Iterative solutions of boundary integral equations in acoustics," *Proc. Roy. Soc. London*, **A417**, 45–57, 1998.

[7] Kleinman, R. E., G. F. Roach, I. S. Schuetz, J. Shirron and P. M. van den Berg, "An over-relaxation method for the iterative solution of integral equations in scattering problems," *Wave Motion,*, **12**, 161–170, 1990.

[8] Kleinman, R. E., G. F. Roach, and P. M. van den Berg, "A convergent Born series for large refractive indices," *J. Opt. Soc. Am.*, **A7**, 890–897, 1990.

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

[10] Patterson, W. M., "Iterative methods for the solution of a linear operator equation in Hilbert space — a survey," Lecture Notes in Mathematics 394, Springer-Verlag, Berlin, 1974.

[11] Riesz, F., and B. Sz.-Nagy, *Functional analysis*, Frederick Ungar Publishing Co., New York, 1955.

[12] Saad, Y., and M. H. Schultz, "GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems," *SIAM J. Sci. Stat. Comput.*, **7**, 856–869, 1986.

[13] Sarkar, T. K., and E. Arvas, "On a class of finite step iterative methods (conjugate directions) for the solution of an operator equation arising in electromagnetics," *IEEE Trans. Ant. Propag.*, **AP-33**, 1058–1066, 1985.

[14] Weston, V. H., "On the convergence of the Rytov approximation for the reduced wave equation," *J. Math. Phys.*, **26**, 1979–1985, 1985.