Numer. Math. 48, 499–523 (1986)

Numerische Mathematik

On the Rate of Convergence of the Preconditioned Conjugate Gradient Method

Owe Axelsson[1] and Gunhild Lindskog[2]

[1] Department of Computer Sciences, Lund University, Lund, Sweden
[2] Department of Computer Sciences, Chalmers University of Technology, Göteborg, Sweden

Summary. We derive new estimates for the rate of convergence of the conjugate gradient method by utilizing isolated eigenvalues of parts of the spectrum. We present a new generalized version of an incomplete factorization method and compare the derived estimates of the number of iterations with the number actually found for some elliptic difference equations and for a similar problem with a model empirical distribution function.

Subject Classifications: AMS(MOS): 65F10; CR: G1.3.

1. Introduction

Let H be a Hilbert space with inner product $(\cdot,\cdot)$. We consider the conjugate gradient method for the solution of $A\hat{x}=\hat{b}$, where $\hat{b}\in\mathscr{R}(A)\subset H$, the range of A and $\hat{x}\in\mathscr{D}(A)=H$, the domain of definition. We assume, that the operator A is compact, selfadjoint and positive semidefinite and that the origin is not a cluster point of the eigenvalues of A.

Since $\hat{b}\in\mathscr{R}(A)$, there exists a solution but if $\mathrm{Ker}(A)$, the nullspace of A, is nontrivial, the solution is not unique. We are then content finding *any* solution.

Let $\|x\|_{A^{1/2}}=(x,Ax)^{1/2}$ be the energy seminorm and let $f(x)=\frac{1}{2}(x,Ax)-(\hat{b},x)$. Then $f(x)=f(\hat{x})+\frac{1}{2}\|\hat{x}-x\|^2_{A^{1/2}}$. Hence the problem is equivalently formulated: Find a $\hat{x}\in\mathscr{D}(A)$ such that $\hat{x}=\arg\min_{x\in\mathscr{D}(A)} f(x)$.

The preconditioned conjugate gradient method to find a minimizer is of the form

$$x^{(l+1)}=x^{(l)}+\tau_l d^{(l)},$$

$$r^{(l+1)}=Ax^{(l+1)}-\hat{b},$$

$$\tilde{r}^{(l+1)}=C^{-1}r^{(l+1)},$$

$$d^{(l+1)}=-\tilde{r}^{(l+1)}+\beta_l d^{(l)}, \qquad l=0,1,\ldots.$$

Here $x^{(0)}$ is arbitrary, $d^{(0)}=-C^{-1}r^{(0)}$, $r^{(0)}=Ax^{(0)}-\hat{b}$ and C (selfadjoint and positive definite) is a preconditioning of A. τ_l and β_l are determined by certain inner products in order to get

$$\|\hat{x}-x^{(l+1)}\|_{A^{1/2}}=\min_{\tau_l}\|\hat{x}-x^{(l)}-\tau_l d^{(l)}\|_{A^{1/2}}$$

and

$$(d^{(l+1)}, Ad^{(j)})=0, \quad j=0,1\ldots l.$$

Clearly, the error $e^{(k)}=\hat{x}-x^{(k)}$ can be written

$$e^{(k)}=P_k(C^{-1}A)e^{(0)}, \quad k=0,1,\ldots,$$

where $P_k\in\pi_k^1$, the set of polynomials of degree k such that $P_k(0)=1$. Furthermore, due to the minimization property, we have

$$\|e^{(k)}\|_{A^{1/2}}=\min_{P_k\in\pi_k^1}\|P_k(C^{-1}A)e^{(0)}\|_{A^{1/2}}.$$

By expanding $e^{(0)}$ in the eigenvectors we readily find

$$\|e^{(k)}\|_{A^{1/2}}\leqq\min_{P_k\in\pi_k^1}\max_{\lambda\in S^1(C^{-1}A)}|P_k(\lambda)|\,\|e^{(0)}\|_{A^{1/2}}, \tag{1.1}$$

where $S^1(C^{-1}A)=\{\lambda_i\}_{i\geqq 1}$ is the subset of positive disjoint eigenvalues of $C^{-1}A$, which we order in increasing order.

The problem we are facing is to find the smallest k such that

$$\min_{P_k\in\pi_k^1}\max_{\lambda\in S^1(C^{-1}A)}|P_k(\lambda)|\leqq\varepsilon, \quad 0<\varepsilon<1. \tag{1.2}$$

By (1.1) this is a (sharp) upper bound for the necessary number of iterations to reach a relative error

$$\|e^{(k)}\|_{A^{1/2}}/\|e^{(0)}\|_{A^{1/2}}\leqq\varepsilon \tag{1.3}$$

for an arbitrary initial vector. Note that any polynomial $P_k\in\pi_k^1$ for which (1.2) is satisfied will give an upper bound of k.

In Sect. 2 we consider estimates of k for various distributions of eigenvalues. These include new estimates in the case of small isolated eigenvalues derived using Chebyshev and Jacobi polynomials. Earlier results related to such estimates are found in [9, 1, 2, 8, 6] and [13]. As is done in [1, 2] and [8], we concentrate here in finding *explicit expressions* for the estimates of k.

In Sect. 3 we consider the iterative solution of linear systems derived from discretized selfadjoint partial differential equations of second order. For the construction of a preconditioning matrix we then consider a generalized incomplete factorization method based on an arbitrary positive vector v for which $Av>0$ and a relaxed form of modification.

In Sect. 4 we report on some numerical tests for the methods, in particular for the problems studied in Sect. 3. As is found in Sect. 4 and as was also discussed in [5] the distribution of eigenvalues for certain preconditioned difference matrices takes one of the forms studied in Sect. 2. We compare the

estimates derived by the methods in Sect. 2 with the actual number of iterations we get for various choices of initial vectors and distribution of eigenvalues.

We also perform a similar study for the case where we have assumed some model distribution of eigenvalues similar to the kind we get with the relaxed forms of modification for the incomplete preconditioned factorization methods.

2. Estimates of the Number of Conjugate Gradient Iterations in the Case of Isolated Eigenvalues

Estimates Based on Chebyshev Polynomials

A classical estimate for the number of iterations in (1.3) is derived if we consider the approximation problem (1.2) for the continuous interval $[a,b]$, where $a=\lambda_1$ and $b=\max_{i\geqq 1}\lambda_i$. The best approximation is then

$$P_k(\lambda)=P_k^*(\lambda;a,b,\varepsilon)\equiv T_k\left(\frac{b+a-2\lambda}{b-a}\right)\Big/T_k\left(\frac{b+a}{b-a}\right),$$

where $T_k(x)=\frac{1}{2}[(x+\sqrt{x^2-1})^k+(x-\sqrt{x^2-1})^k]$ is the Chebyshev polynomial. We then find (see, for instance, [2])

$$\max_{\lambda\in[a,b]}|P_k(\lambda)|=\frac{1}{T_k\left(\frac{b+a}{b-a}\right)}=2\frac{\sigma^k}{1+\sigma^{2k}},$$

where $\sigma=\left(1-\sqrt{\frac{a}{b}}\right)\Big/\left(1+\sqrt{\frac{a}{b}}\right)$. Hence (1.2) is satisfied for any integer k satisfying

$$k\geqq\ln\left(\frac{1}{\varepsilon}+\sqrt{\frac{1}{\varepsilon^2}-1}\right)\Big/\ln\sigma^{-1}. \tag{2.1}$$

Since ε is a small number, we use in practice the following upper bound for k in (1.3):

$$k^*(a,b,\varepsilon)\equiv\left\lceil\ln\frac{2}{\varepsilon}\Big/\ln\sigma^{-1}\right\rceil, \tag{2.2}$$

where $\lceil a\rceil$ indicates the smallest integer $\geqq a$. (Note that if $b/a\gg 1$, then $(\ln\sigma^{-1})^{-1}\approx\frac{1}{2}(b/a)^{1/2}$ can be used in the upper bound. In our applications in this paper we need the accurate number.)

However, when the eigenvalues are not uniformly distributed on $[a,b]$ we can get improved bounds of k. We shall consider three cases where λ_i', $i=1,2,\ldots,q$ denote the q largest eigenvalues in decreasing order:

(2.3) (i) $S^1(C^{-1}A)\in[a,b]\cup\left(\bigcup_{i=1}^{q}\lambda_i'\right)$, $\qquad b<\lambda_q'$

(ii) $S^1(C^{-1}A)\in\left(\bigcup_{i=1}^{p}\lambda_i\right)\cup[a,b], \qquad \lambda_p<a$

(iii) $S^1(C^{-1}A)\in\left(\bigcup_{i=1}^{p}\lambda_i\right)\cup[a,b]\cup\left(\bigcup_{i=1}^{q}\lambda_i'\right), \qquad \lambda_p<a<b<\lambda_q'.$

Consider at first case (i). Then we let

$$P_k(\lambda)=\prod_{i=1}^{q}\left(1-\frac{\lambda}{\lambda_i'}\right)P_{k-q}^*(\lambda;a,b,\varepsilon).$$

Clearly $P_k(0)=1$, $P_k(\lambda_i')=0$, $i=1,2,\ldots,q$. Since $\left|1-\frac{\lambda}{\lambda_i'}\right|<1$, $a\leqq\lambda\leqq b$, we get the upper bound

$$k=q+k^*(a,b,\varepsilon), \tag{2.4}$$

where k^* is defined in (2.2).

Consider now case (ii). We let first

$$P_k(\lambda)=\prod_{i=1}^{p}\left(1-\frac{\lambda}{\lambda_i}\right)P_{k-p}^*(\lambda;a,b,\varepsilon')$$

where $\varepsilon'=\varepsilon\prod_{i=1}^{p}(\lambda_i/b)$. Since $P_k(\lambda_i)=0$, $i=1,2,\ldots,p$ and $\left|1-\frac{\lambda}{\lambda_i}\right|<\frac{b}{\lambda_i}$, $a\leqq\lambda\leqq b$ we have $\max_{\lambda\in\cup\lambda_i\cup[a,b]}|P_k(\lambda)|\leqq\varepsilon$ for

$$k\geqq p+k^*(a,b,\varepsilon'). \tag{2.5}$$

By (2.2) an upper bound is

$$k=\left\lceil\left(\ln\frac{2}{\varepsilon}+\sum_{i=1}^{p}\ln\frac{b}{\lambda_i}\right)\Big/\ln\sigma^{-1}\right\rceil+p. \tag{2.6}$$

This was one of the estimates derived in [1] and [2]. Alternative and frequently more accurate estimates shall be presented in the next subsections.

Consider finally case (iii) where $q=p$. Let then

$$P_k(\lambda)=\prod_{i=1}^{p}(\lambda_i\lambda_i')^{-1}(\lambda_i-\lambda)(\lambda_i'-\lambda)P_{k-2p}^*(\lambda;a,b,\varepsilon'').$$

We have

$$\max_{\lambda\in[a,b]}|(\lambda_i\lambda_i')^{-1}(\lambda_i-\lambda)(\lambda_i'-\lambda)|\leqq\frac{1}{4}\frac{\lambda_i'}{\lambda_i}\left(1-\frac{\lambda_i}{\lambda_i'}\right)^2.$$

Further $P_k(\lambda_i)=P_k(\lambda_i')=0$, $i=1,2,\ldots,p$. Hence with

$$\varepsilon''=\varepsilon\prod_{i=1}^{p}4\frac{\lambda_i}{\lambda_i'}\left(1-\frac{\lambda_i}{\lambda_i'}\right)^{-2}$$

we get

$$\max_{\lambda\in\bigcup_i\lambda_i\cup[a,b]\cup(\bigcup_i\lambda_i')}|P_k(\lambda)|\leqq\varepsilon$$

for

$$k \geqq 2p + k^*(a, b, \varepsilon'') \tag{2.7}$$

and the upper bound is

$$k = \left[\left[\ln \frac{2}{\varepsilon} + \sum_{i=1}^{p} \ln \frac{\lambda_i'}{4\lambda_i} \left(1 - \frac{\lambda_i}{\lambda_i'}\right)^2 \right] \Big/ \ln \sigma^{-1} \right] + 2p. \tag{2.8}$$

An alternative estimate is presented below.

We shall see in Sect. 4 that the estimates of the type (2.4), (2.5) and (2.7) can be applied for a relaxed incomplete factorization preconditioning method for $\omega = 1$, $\omega = 0$ and an optimal value satisfying $0 < \omega < 1$ respectively.

A Generalized Estimate in the Case of Small Isolated Eigenvalues

Consider case (ii) again. We shall now sharpen the estimate (2.6). We write then P_k as a product

$$P_k(\lambda) = R_{k-m-p}(\lambda) \prod_{i=1}^{p} U_{r_i}(\lambda/b) \left(1 - \frac{\lambda}{\lambda_i}\right),$$

where $R_{k-m-p} \in \pi^1_{k-m-p}$, $U_{r_i} \in \pi^1_{r_i}$ and $m = \sum_{i=1}^{p} r_i$. Since $P_k(\lambda_i) = 0$, $i = 1, 2, \ldots, p$, $\max |P_k(\lambda)|$ is considered for $\lambda \in [a, b]$.

Let $l = k - m - p$ and $\eta = \prod_{i=1}^{p} c_i/\lambda_i$, where c_i, $0 < c_i \leqq b$ are constants, to be chosen later to get an optimal result.

In the estimate (2.6), the special case $c_i = b$ and $r_i = 0$, $i = 1, 2, \ldots, p$ was considered. (1.2) is satisfied by choosing R_l and U_{r_i} such that

$$\max_{\lambda \in [a,b]} |R_l(\lambda)| \leqq \varepsilon/\eta \tag{2.9}$$

and

$$\max_{\lambda \in [a,b]} |U_{r_i}(\lambda/b)(1 - \lambda/\lambda_i)| \leqq c_i/\lambda_i, \quad i = 1, 2, \ldots, p. \tag{2.10}$$

Thus given U_{r_i}, we have to find the smallest l such that (2.9) is satisfied.

From (2.2) we have for $R_l = P_l^*$ $\left(\text{if } \frac{2}{\varepsilon}\eta > 1\right)$ that (2.9) is valid for

$$l = \left[\left(\ln \frac{2}{\varepsilon} + \ln \eta \right) \Big/ \ln \sigma^{-1} \right], \quad \sigma = \left(1 - \sqrt{\frac{a}{b}}\right) \Big/ \left(1 + \sqrt{\frac{a}{b}}\right)$$

i.e.

$$k = \left[\left(\ln \frac{2}{\varepsilon} + \ln \eta \right) \Big/ \ln \sigma^{-1} \right] + m + p. \tag{2.11}$$

Since $|1 - \lambda/\lambda_i| \leqq |\lambda/\lambda_i|$, (2.10) is satisfied if

$$\max_{\lambda\in[a,b]} |U_{r_i}(\lambda/b)|\,\lambda \leqq c_i, \qquad i=1,2,\ldots,p,$$

or

$$\max_{x\in[\frac{a}{b},1]} |U_{r_i}(x)|\,x \leqq c_i/b, \qquad i=1,2,\ldots,p, \tag{2.12}$$

where $x=\lambda/b$. In practice, $a/b \ll 1$, so we consider instead the problem: Find the smallest r_i such that

$$\max_{0\leqq x\leqq 1} |U_{r_i}(x)|\,x \leqq c_i/b. \tag{2.13}$$

Let $S_{r_i+1}(x)=U_{r_i}(x)x$. Since $S_{r_i+1}(0)=0$ and $U_{r_i}\in\pi^1_{r_i}$ we have $S'_{r_i+1}(0)=1$.

Let $\alpha<x_0<0$ and let $S_{r+1}\in\pi^{\alpha,-1}_{r+1}$ be the polynomial with the property that

$$\max_{x\in[x_0,1]} |S_{r+1}(x)| = \min_{P_{r+1}\in\pi^{\alpha,-1}_{r+1}} \max_{x\in[x_0,1]} |P_{r+1}(x)|, \tag{2.14}$$

where $\pi^{\alpha,-1}_{r+1}$ denotes the set of polynomials P_{r+1} of degree $r+1$ for which $P_{r+1}(\alpha)=-1$ and where α and x_0 are chosen such that $|x_0|$ is minimal and

$$S_{r+1}(0)=0 \tag{2.15}$$

and

$$S'_{r+1}(0)=1. \tag{2.16}$$

This is sketched in Fig. 2.1.

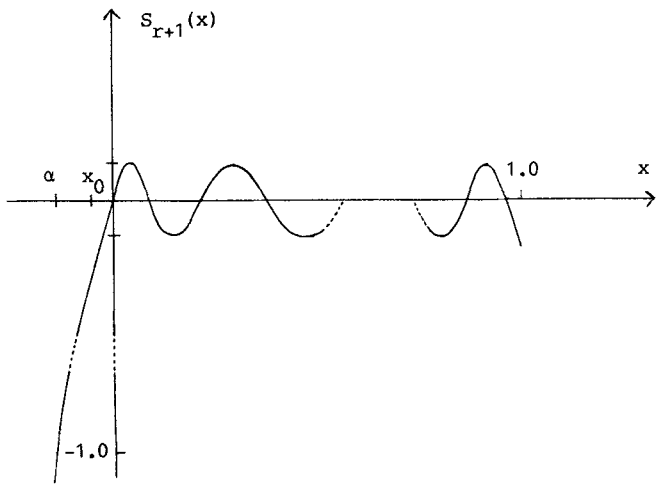

Fig. 2.1. The polynomial $S_{r+1}(x)$

Problem (2.13) can now be replaced by the following: Find the smallest r_i such that

$$\max_{x\in[x_i,1]} |S_{r_i+1}(x)| \leqq c_i/b, \qquad i=1,2,\ldots,p \tag{2.17}$$

with S_{r_i+1} as in (2.14), (2.15) and (2.16) and $x_0=x_i$, $r=r_i$ and $\alpha=\alpha_i$. The solution of problem (2.14) is known to be

$$S_{r_i+1}(x)=-T_{r_i+1}\left(\frac{1+x_i-2x}{1-x_i}\right)\Big/T_{r_i+1}\left(\frac{1+x_i-2\alpha_i}{1-x_i}\right)$$

and

$$\max_{x\in[x_i,1]}|S_{r_i+1}(x)|=1\Big/\left|T_{r_i+1}\left(\frac{1+x_i-2\alpha_i}{1-x_i}\right)\right|.$$

The condition (2.15) implies $T_{r_i+1}\left(\frac{1+x_i}{1-x_i}\right)=0$, i.e.

$$x_i=-\left(\tan\frac{\pi}{4(r_i+1)}\right)^2=-\left(\frac{\pi}{4(r_i+1)}\delta_{r_i}\right)^2, \tag{2.18}$$

where $1<\delta_{r_i}\leqq 4/\pi$ and $\delta_{r_i}\to 1+O(r_i^{-2})$, $r_i\to\infty$.

The condition (2.16) together with (2.18) implies

$$1\Big/\left|T_{r_i+1}\left(\frac{1+x_i-2\alpha_i}{1-x_i}\right)\right|=\{-x_i\}^{1/2}/(r_i+1).$$

Hence the smallest nonnegative integer r_i such that (2.17), and hence (2.13), is valid, which we denote by $r_{i,0}$ is equal to the smallest nonnegative integer r_i such that

$$\tan\frac{\pi}{4(r_i+1)}\Big/(r_i+1)\leqq c_i/b \tag{2.19}$$

(2.11) now gives the upper bound

$$k=k(\underline{c})\equiv\left\lceil\left(\ln\frac{2}{\varepsilon}+\sum_{i=1}^{p}\ln\frac{c_i}{\lambda_i}\right)\Big/\ln\sigma^{-1}\right\rceil+\sum_{i=1}^{p}(r_{i,0}+1) \tag{2.20}$$

where $\underline{c}=(c_1,c_2,\ldots,c_p)$ and $r_{i,0}=r_{i,0}(c_i)$ is defined by (2.19). If we neglect the integer part in (2.20), we see that the minimum of $k(\underline{c})$ is taken when $c_i=c$, $i=1,2,\ldots,p$ for some constant c, $0<c\leqq b$. We shall consider this case now. Then, by (2.19), $r_{i,0}=r_0$, $i=1,2,\ldots,p$ for some nonnegative integer r_0. Since r_0 only takes on integer values, it is better to consider $k=k(r_0)$ as a function of r_0. By (2.19), the smallest value of c corresponding to r_0 is $c(r_0)=b\tan\frac{\pi}{4(r_0+1)}\cdot(r_0+1)$. By writing (2.20) for $r_{i,0}=r_0$ on the form

$$k=k(r_0)=\left\lceil\left(\ln\frac{2}{\varepsilon}+\ln\prod_{i=1}^{p}\frac{b}{\lambda_i}\right)\Big/\ln\sigma^{-1}-p\ln\frac{b}{c}\Big/\ln\sigma^{-1}\right\rceil+p(r_0+1).$$

it is readily seen that due to monotonicity, the optimal value of r_0, r_{opt}, is equal to the smallest integer for which

$$\frac{\ln c(r+1)^{-1}}{\ln\sigma^{-1}}-(r+1)\leqq\frac{\ln c(r)^{-1}}{\ln\sigma^{-1}}-r,$$

i.e. for which

$$c(r)/c(r+1) \leqq \sigma^{-1}. \tag{2.21}$$

The optimal value of c is $c_{\text{opt}} = c(r_{\text{opt}})$. Hence

$$k = p(r_{\text{opt}}+1) + k^*(a, b, \varepsilon'''), \tag{2.22}$$

where $\varepsilon''' = \varepsilon \Big/ \prod_{i=1}^{p} \frac{c_{\text{opt}}}{\lambda_i}$ and k^* is defined in (2.2).

In particular, if $\sigma^{-1} \geqq c(0)/c(1) = 2/\tan\frac{\pi}{8}$, then $r_{\text{opt}} = 0$ and $c_{\text{opt}} = b$. (2.22) takes then the form (2.6). It follows that (2.22) gives an improvement over (2.6) if σ^{-1} is small enough, i.e. if the condition number b/a is large enough. When $b/a \to \infty$, by (2.21), r_{opt} satisfies $\left(\frac{r_{\text{opt}}+1}{r_{\text{opt}}}\right)^2 \sim \sigma^{-1}$, i.e. $r_{\text{opt}} \sim \sqrt{b/a}$. Hence, asymptotically,

$$k(r) \sim \left\lceil \frac{1}{2}\sqrt{\frac{b}{a}}\left(\ln\frac{2}{\varepsilon} + \ln\sum_{i=1}^{p} \beta/\lambda_i\right)\right\rceil \tag{2.23}$$

where $\beta = \frac{\pi}{a}ae^2$, and e is the natural number.

Note that, essentially, b in (2.6) has been replaced by β, and $\beta \ll b$ when $a \ll b$.

This is the formula derived in [8]. Being an asymptotic estimate valid when $b/a \to \infty$, it can give a too small number of iterations in some case. However, in practise it is usually accurate enough.

As for the choice of p we note that the optimal value of p, which gives the smallest upper bound k in (2.22), increases as ε decreases.

We can also consider a combination of the estimates (2.22) and (2.8), thus yielding an estimate for case (iii), utilizing p small and q large isolated eigenvalues, where $p \geqq q$. The estimate uses the q largest and smallest eigenvalues by (2.8), and the $p-q$ next smallest eigenvalues by (2.22). This is illustrated in

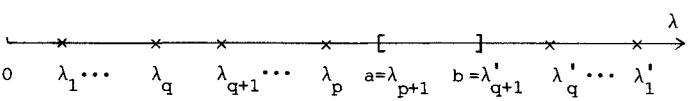

Fig. 2.2. The eigenvalues used in estimate (2.24a)

Fig. 2.2, where we let $a = \lambda_{p+1}$ and $b = \lambda'_{q+1}$. We then get the upper bound

$$k = (p-q)(r_{\text{opt}}+1) + 2q + k^*(a, b, \varepsilon^{\text{IV}}), \quad \text{where } \varepsilon^{\text{IV}} = \varepsilon'' \Big/ \prod_{i=q+1}^{p} \frac{c_{\text{opt}}}{\lambda_i}, \tag{2.24a}$$

where r_{opt} is given by (2.21) and where

$$\varepsilon'' = \varepsilon \prod_{i=1}^{q} 4\frac{\lambda_i}{\lambda'_i}\left(1 - \frac{\lambda_i}{\lambda'_i}\right)^{-2}. \tag{2.24b}$$

We finally comment on a tendency of super-convergence frequently observed for the conjugate gradient method. It follows that for fixed p, as $\varepsilon \to 0$, the term $\ln 2/\varepsilon$ eventually dominates the other terms in the bracket in (2.20).

This means that the estimate is such that it eventually ignores the small eigenvalues and behaves as if A had an "efficient" condition number b/a. This is most easily seen for the case $p=1$ and $b/\lambda_2 \gg 1$, where by (2.23),

$$k \sim \frac{1}{2}\sqrt{\frac{b}{\lambda_2}}\left(\ln\frac{2}{\varepsilon}+\ln\frac{\beta}{\lambda_1}\right), \qquad \beta=\frac{\pi}{4}e^2\lambda_2.$$

As $\beta\varepsilon/\lambda_1$ becomes smaller than 2, the upper bound gradually becomes independent of λ_1. The numerical tests in [5] illustrates also such a phenomenon where the number of iterations needed for each new correct digit decreases. A similar study in [13] is performed by use of so called Ritz values.

An Estimate Based on Jacobi Polynomials

Also here we consider case (ii). Let

$$P_k(\lambda)=\prod_{i=1}^{p}\left(1-\frac{\lambda}{\lambda_i}\right)P_{k-p}^{(0,2p)}\left(2\frac{\lambda-\bar{\lambda}_p}{b-\bar{\lambda}_p}-1\right)\Big/P_{k-p}^{(0,2p)}\left(-\frac{b+\bar{\lambda}_p}{b-\bar{\lambda}_p}\right),$$

where $P_{k-p}^{(0,2p)}$ is the Jacobi polynomial [12] of degree $k-p$. The set of polynomials $\{P_n^{(0,2p)}\}$ are mutually orthogonal on the interval $[-1,1]$ with respect to the weight function $(1+x)^{2p}$. $\bar{\lambda}_p$ is a parameter $0\leqq\bar{\lambda}_p<a$ to be chosen later.

Note that the Jacobi polynomials satisfy the recursion

$$P_0^{(0,2p)}(x)=1, \qquad P_1^{(0,2p)}(x)=(p+1)x-p,$$

$$\begin{aligned} n(n+2p)(n+p-1)P_n^{(0,2p)}(x) &= (2n+2p-1)\{(n+p)(n+p-1)x-p^2\}P_{n-1}^{(0,2p)}(x) \\ &\quad -(n-1)(n+2p-1)(n+p)P_{n-2}^{(0,2p)}(x), \qquad n=2,3,4,\ldots. \end{aligned}$$

We have

$$P_k(\lambda)=\prod_{i=1}^{p}\left\{\left(-\frac{1}{\lambda_i}\right)\frac{\lambda-\lambda_i}{\lambda-\bar{\lambda}_p}\right\}(\lambda-\bar{\lambda}_p)^p P_{k-p}^{(0,2p)}\left(2\frac{\lambda-\bar{\lambda}_p}{b-\bar{\lambda}_p}-1\right)\Big/P_{k-p}^{(0,2p)}\left(-\frac{b+\bar{\lambda}_p}{b-\bar{\lambda}_p}\right)$$

i.e. for $\lambda\in[a,b]$

$$|P_k(\lambda)|\leqq\prod_{i=1}^{p}\frac{b-\bar{\lambda}_p}{\lambda_i}\prod_{i=1}^{p}\frac{(\lambda-\lambda_i)}{(\lambda-\bar{\lambda}_p)}\tilde{\lambda}^p|P_{k-p}^{(0,2p)}(2\tilde{\lambda}-1)|\Big/\left|P_{k-p}^{(0,2p)}\left(-\frac{b+\bar{\lambda}_p}{b-\bar{\lambda}_p}\right)\right|$$

where $\tilde{\lambda}=\dfrac{\lambda-\bar{\lambda}_p}{b-\bar{\lambda}_p}$.

From [11] we have for the Jacobi polynomials

$$\tilde{\lambda}^p|P_{k-p}^{(0,2p)}(2\tilde{\lambda}-1)|\leqq 1, \qquad 0\leqq\tilde{\lambda}\leqq 1.$$

(Further $\int_0^1|\tilde{\lambda}^p P_{k-p}^{(0,2p)}(2\tilde{\lambda}-1)|^2\,d\tilde{\lambda}$ is minimal for all polynomials of degree $k-p$ with the same leading coefficient.)

Thus, for $\bar{\lambda}_p$ fixed we have to find the smallest k such that

$$\left|P_{k-p}^{(0,2p)}\left(-\frac{b+\bar{\lambda}_p}{b-\bar{\lambda}_p}\right)\right| \geqq \frac{1}{\varepsilon}\prod_{i=1}^{p}\frac{\lambda-\lambda_i}{\lambda-\bar{\lambda}_p}\prod_{i=1}^{p}\frac{b-\bar{\lambda}_p}{\lambda_i}\;\forall\lambda; \qquad a\leqq\lambda\leqq b.$$

For a given $\bar{\lambda}_p$ and a set $\{\lambda_i\}_{i=1}^{p}$ let

$$c(\bar{\lambda}_p)=\max_{\lambda\in[a,b]}\prod_{i=1}^{p}\frac{\lambda-\lambda_i}{\lambda-\bar{\lambda}_p} \tag{2.25}$$

where

$$c_1=\min_{\bar{\lambda}_p\in[0,a)} c(\bar{\lambda}_p).$$

Further let $k^*(\bar{\lambda}_p)$ be the smallest k such that

$$\left|P_{k-p}^{(0,2p)}\left(-\frac{b+\bar{\lambda}_p}{b-\bar{\lambda}_p}\right)\right| \geqq \frac{1}{\varepsilon}c(\bar{\lambda}_p)\prod_{i=1}^{p}\frac{b-\bar{\lambda}_p}{\lambda_i}. \tag{2.26}$$

The optimal value, $\bar{\lambda}_p^*$, of $\bar{\lambda}_p$ satisfies

$$k^*(\bar{\lambda}_p^*)=\min_{0\leqq\bar{\lambda}_p<a} k^*(\bar{\lambda}_p).$$

We search for an easily computable choice of $\bar{\lambda}_p$ and $c(\bar{\lambda}_p)$. One such choice of $\bar{\lambda}_p$, for which $\prod_{i=1}^{p}\frac{\lambda-\lambda_i}{\lambda-\bar{\lambda}_p}\leqq 1$, is

$$\hat{\lambda}_p\equiv\min_{i\in S}\tfrac{1}{2}(\lambda_i+\lambda_{p+1-i}) \tag{2.27}$$

where

$$S=\begin{cases}\{i;\, i=1,2,\ldots,p/2\} & \text{for } p \text{ even}\\ \left\{i;\, i=1,2,\ldots,\dfrac{p+1}{2}\right\} & \text{for } p \text{ odd}\end{cases}.$$

Let $\bar{\lambda}_p=\bar{\lambda}_{p,c_0}$ satisfy

(2.28) $c(\bar{\lambda}_{p,c_0})=c_0$, where $c_0\geqq c_1$ is a given constant and c is defined in (2.25).

Clearly $\hat{\lambda}_p\leqq\bar{\lambda}_{p,1}$ i.e. with $k(\hat{\lambda}_p)$, the smallest k such that

$$\left|P_{k-p}^{(0,2p)}\left(-\frac{b+\hat{\lambda}_p}{b-\hat{\lambda}_p}\right)\right| \geqq \frac{1}{\varepsilon}\prod_{i=1}^{p}\frac{b-\hat{\lambda}_p}{\lambda_i}, \tag{2.29}$$

we have $k(\hat{\lambda}_p)\geqq k^*(\bar{\lambda}_{p,1})$. The latter, however, needs the computation of $\bar{\lambda}_{p,1}$, the value of $\bar{\lambda}_p$ for which $c(\bar{\lambda}_p)=1$, see (2.25).

Moreover $k^*(\hat{\lambda}_p)\leqq k(\hat{\lambda}_p)$, where $k^*(\hat{\lambda}_p)$ needs the computation of $c(\hat{\lambda}_p)$.

In the numerical tests, Sect. 4, $\min_{p\geqq 0} k^*(\bar{\lambda}_p^*)$ and $\min_{p\geqq 0} k(\hat{\lambda}_p)$ differ with at most two.

Remark 2.1. The idea above of transforming the Jacobi polynomial to an interval $[\bar{\lambda}_p, b]$, where $\bar{\lambda}_p<a$, can also be applied to the Chebyshev polynomial. By the same analysis as for the Jacobi polynomials we get

$$|P_k(\lambda)| \leqq \prod_{i=1}^{p} \frac{b-\bar{\lambda}_p}{\lambda_i} \Big/ T_{k-p}\left(\frac{b+\bar{\lambda}_p}{b-\bar{\lambda}_p}\right). \tag{2.30}$$

if $\bar{\lambda}_p = \hat{\lambda}_p$ or $\bar{\lambda}_p = \bar{\lambda}_{p,1}$, for instance.

Thus, this is an improvement of (2.6) if the advantage of the smaller value $\prod_{i=1}^{p} \frac{b-\bar{\lambda}_p}{\lambda_i}$, compared to $\prod_{i=1}^{p} \frac{b}{\lambda_i}$, is not eliminated by the smaller value of $T_{k-p}\left(\frac{b+\bar{\lambda}_p}{b-\bar{\lambda}_p}\right)$, compared to $T_{k-p}\left(\frac{b+a}{b-a}\right)$. By use of the Jacobi polynomial, (2.30) is improved for p and $k>p$ for which

$$\left|P_{k-p}^{(0,2p)}\left(-\frac{b+\bar{\lambda}_p}{b-\bar{\lambda}_p}\right)\right| > \left|T_{k-p}\left(-\frac{b+\bar{\lambda}_p}{b-\bar{\lambda}_p}\right)\right|.$$

Remark 2.2. Asymptotically as $k\to\infty$, it follows from the recurrence relation for $P_k^{(0,2p)}$ that it goes over in the recurrence relation for T_k,

$$P_k^{(0,2p)}(x) \approx 2x P_{k-1}^{(0,2p)}(x) - P_{k-2}^{(0,2p)}(x).$$

Hence, asymptotically, $P_k^{(0,2p)}(x) \approx c_1(x-\sqrt{x^2-1})^k + c_2(x+\sqrt{x^2-1})^k$ and

$$\left|P_{k-p}^{(0,2p)}\left(-\frac{b+\bar{\lambda}_p}{b-\bar{\lambda}_p}\right)\right| \approx c_2\left(\frac{b+\bar{\lambda}_p}{b-\bar{\lambda}_p} + \sqrt{\left(\frac{b+\bar{\lambda}_p}{b-\bar{\lambda}_p}\right)^2 - 1}\right)^{k-p}$$
$$= c_2\left(\frac{b+\sqrt{\bar{\lambda}_p}}{b-\sqrt{\bar{\lambda}_p}}\right)^{k-p}, \quad k\to\infty.$$

Hence this leads to about the same estimate as if we use Chebyshev polynomials as in Remark 2.1. This means that the estimates with Jacobi polynomials can only be expected to yield improved estimates when k is not very large.

Analogous to (2.24a) we get an estimate for case (iii) where $p \geqq q$ by considering (2.26) and (2.8). Thus, for a given $\bar{\lambda}_p$, an upper bound is given by the smallest k such that

$$P_{k-(p+q)}^{(0,2(p-q))}\left(-\frac{b+\bar{\lambda}_p}{b-\bar{\lambda}_p}\right) \geqq \frac{1}{\varepsilon''} c(\bar{\lambda}_p) \prod_{i=q+1}^{p} \frac{b-\bar{\lambda}_p}{\lambda_i}, \tag{2.31}$$

where $c(\bar{\lambda}_p)$ is the smallest upper bound such that

$$\prod_{i=q+1}^{p} \frac{\lambda-\lambda_i}{\lambda-\bar{\lambda}_p} \leqq c(\bar{\lambda}_p) \forall \lambda; \quad a \leqq \lambda \leqq b$$

and ε'' is given in (2.24b).

3. Preconditioning by Incomplete Factorizations Based on Generalized Rowsum Criteria

We consider large sparse systems of linear equations $A\hat{x} = \hat{b}$, where $\hat{x}$, $\hat{b} \in R^N$ and A is a symmetric, positive definite matrix of order $N \times N$ derived from

discretization by the finite element method or a finite difference method of a selfadjoint partial differential equation of second order.

The construction of a preconditioning matrix by incomplete factorization of A results in

$$C=A+R,$$

where R is the defect matrix. In the Modified Incomplete Cholesky, MIC, method [7], and the Modified Block Incomplete Cholesky, MBIC, method [4, 3] and [5], the matrix R has the property

$$Rv=0 \tag{3.1}$$

for $v=e=(1,1,\ldots,1)^T$.

We shall here consider general choices of the vector $v>0$, for which $Rv=0$, in the construction of pointwise and blockwise incomplete factorizations. The vector v is assumed to satisfy $Av>0$.

The *pointwise* incomplete factorization based on a general rowsum criterion is denoted the Generalized Modified Incomplete Cholesky, GMIC, method.

For a symmetric, positive definite matrix A the method is defined by the following computations, where J denotes the set of indices (i,j), where fill-ins are permitted in the factorization. It is proven in [3] that the factorization exists if A is an M-matrix. More general cases can be dealt with by a shifted incomplete factorization, see [10].

For $r=1,2,\ldots,N-1$ we compute

$$l_{ir}=\frac{a_{ir}^{(r)}}{\hat{a}_{rr}^{(r)}}$$

where

$$\hat{a}_{11}^{(1)}=a_{11}^{(1)}$$

$$\hat{a}_{rr}^{(r)}=a_{rr}^{(r)}-\delta_r, \quad \text{for } r\geqq 2,$$

$$\delta_r=\frac{\sum_{j=1}^{N} s_{rj}v_j}{v_r},$$

and

$$s_{rr}=\sum_{p=1}^{r-1}\sum_{(r,k)\notin J}(a_{rk}^{(p)}-l_{rp}a_{pk}^{(p)}) \tag{3.2}$$

$$s_{rj}=\begin{cases}0, & (r,j)\in J\\ \sum_{p=1}^{r-1}(-a_{rj}^{(p)}+l_{rp}a_{pj}^{(p)}), & (r,j)\notin J\end{cases} \qquad j\neq r$$

$$a_{ij}^{(r+1)}=\begin{cases}a_{ij}^{(r)}-l_{ir}a_{rj}^{(r)}, & (r+1\leqq j\leqq N)\cap((i,j)\in J)\cap i\neq j\\ 0, & (r+1\leqq j\leqq N)\cap((i,j)\notin J)\\ a_{ii}^{(r)}-l_{ir}a_{ri}^{(r)}+\sum_{\substack{k=r+1\\(i,k)\notin J}}^{N}(a_{ik}^{(r)}-l_{ir}a_{rk}^{(r)}), & j=i\end{cases}$$

where $i=r+1, r+2,\ldots,N$ and $a_{ij}^{(1)}=a_{ij}$.

Remark. When $p>1$, $a_{ij}^{(p)}=0$ for $(i,j)\notin J$.

The matrix L is completely defined when we add

$$l_{ij}=\begin{cases}0, & j>i\\ 1, & j=i.\end{cases}$$

The matrix U is defined by

$$u_{ij}=\begin{cases}\hat{a}_{ii}^{(i)}, & i=1,2,\ldots,N-1;\ i=j\\ a_{ij}^{(i)}, & i=1,2,\ldots,N-1;\ j=i+1,\ldots,N\\ a_{ii}^{(N)}, & i=N\\ 0 & \text{otherwise.}\end{cases}$$

The GMIC factorization of A is then given by

$$C=LU.$$

C can also be written as $C=A+R$, where R satisfies the condition (3.1) for the vector $v>0$.

Note also that $R=R_0-\Delta$, where R_0 is a matrix satisfying the condition (3.1) for $v=e$ and

$$\Delta=\begin{bmatrix}0 & & & & \\ & \delta_2 & & 0 & \\ & & \delta_3 & & \\ & & & \ddots & \\ & & & & \delta_{N-1} \\ & 0 & & & \\ & & & & 0\end{bmatrix}$$

where δ_i, i$=2,3,\ldots,N-1$ is given in (3.2).

In the *blockwise* method we assume that we have a blockpartitioning of A

$$A=\begin{bmatrix}D_1 & U_1 & & & \\ & & \ddots & & 0 \\ L_1 & D_2 & & \ddots & \\ & & \ddots & & U_{M-1} \\ 0 & & & L_{M-1} & D_M\end{bmatrix}.$$

An incomplete block factorization of A can be written as

$$C=\begin{bmatrix}G_1 & & & 0 \\ L_1 & G_2 & & \\ 0 & \ddots & \ddots & \\ & L_{M-1} & G_M\end{bmatrix}\begin{bmatrix}G_1^{-1} & & & 0 \\ & G_2^{-1} & & \\ & & \ddots & \\ 0 & & & G_M^{-1}\end{bmatrix}\begin{bmatrix}G_1 & U_1 & & 0 \\ & G_2 & \ddots & \\ & & \ddots & U_{M-1} \\ 0 & & & G_M\end{bmatrix},$$

where the matrices G_i, $i=1, 2, \ldots, M$ are assumed to be nonsingular and are to be chosen such that the condition (3.1) is satisfied for the vector $v>0$.

Let $(G_i^{-1})^{(p)}$ denote the bandmatrix of G_i^{-1} with bandwidth p located symmetrically about the main diagonal.

We choose G_i, $i=1, 2, \ldots, M$ is

$$\begin{aligned} &G_1 = D_1 \\ &G_i = D_i - L_{i-1}(G_{i-1}^{-1})^{(3)} U_{i-1} - \Delta_{i-1}, \quad i=2, \ldots, M, \end{aligned} \tag{3.3}$$

where Δ_{i-1} is a diagonal matrix formed to get the generalized rowsum criterion satisfied.

With the choice (3.3) we get $C=A+R$ with the defect matrix

$$R = \begin{bmatrix} 0 & & & \\ & \tilde{E}_1 & & 0 \\ & & \tilde{E}_2 & \\ & 0 & & \ddots \\ & & & \tilde{E}_{M-1} \end{bmatrix},$$

where $\tilde{E}_i = L_i G_i^{-1} U_i - L_i (G_i^{-1})^{(3)} U_i - \Delta_i$.

Let $E_i = L_i G_i^{-1} U_i - L_i (G_i^{-1})^{(3)} U_i$ and $v=[v^{(1)}, v^{(2)}, \ldots, v^{(M)}]^T$. Then the condition (3.1) reads

$$\Delta_i v^{(i+1)} = E_i v^{(i+1)}, \quad i=1, 2, \ldots, M-1.$$

Δ_i, $i=1, 2, \ldots, M-1$ is computed by solving the systems

$$G_i z = U_i v^{(i+1)}, \quad i=1, 2, \ldots, M-1,$$

where $z = L_i^{-1} \Delta_i v^{(i+1)} + (G_i^{-1})^{(3)} U_i v^{(i+1)}$, which gives

$$\Delta_i v^{(i+1)} = L_i z - L_i (G_i^{-1})^{(3)} U_i v^{(i+1)}, \quad i=1, 2, \ldots, M-1$$

and hence the matrices Δ_i, $i=1, 2, \ldots, M-1$.

We denote this method by the Generalized Modified Block Incomplete Cholesky, GMBIC, method. The factorization exists if A is an M-matrix, see [3].

In Sect. 4 we study some suitable choices of the vector v.

Relaxed modifications of the GM(B)IC method with the choice $v=e$, the Relaxed (Block) Incomplete Cholesky, R(B)IC, methods are presented in [5].

In the RIC method the computations (3.2) are replaced by the following:

$$l_{ir} = a_{ir}^{(r)} / a_{rr}^{(r)}, \qquad a_{ij}^{(r+1)} = \begin{cases} a_{ij}^{(r)} - l_{ir} a_{rj}^{(r)}, & (r+1 \leqq j \leqq N) \cap ((i,j) \in J) \cap i \neq j \\ 0, & (r+1 \leqq j \leqq N) \cap ((i,j) \notin J) \\ a_{ii}^{(r)} - l_{ir} a_{ri}^{(r)} + \omega \sum\limits_{\substack{k=r+1 \\ (i,k) \notin J}}^{N} (a_{ik}^{(r)} - l_{ir} a_{rk}^{(r)}), & i=j. \end{cases}$$

The RBIC method is defined by $G_1 = D_1$, $G_i = D_i - L_{i-1}(G_{i-1}^{-1})^{(3)} U_{i-1} - \omega \Delta_{i-1}$, $i=2, \ldots, M$. Here ω is the relaxation parameter, $0 \leqq \omega \leqq 1$.

4. Numerical Results

We consider the model problem

$$\begin{aligned} -\Delta u &= f \quad (x, y) \in \Omega \\ u &= 0, \quad (x, y) \in \partial\Omega, \end{aligned} \tag{4.1}$$

where Ω is the unit square with boundary $\partial\Omega$ and $u(x, y) = x(x-1)\,y(y-1)\,e^{xy}$. The problem is discretized by linear finite element approximations over a uniform rightangled triangulation with stepsize h. The number of nodes (unknowns) is then $N = (h^{-1}-1)^2$.

In the Figs. 4.1–4.4 we show graphic representations in the form of empirical distribution functions of the distribution of eigenvalues of $C^{-1}A$ for the methods we compare: the RIC method and the RBIC method for various choices of the relaxation parameter ω, the GMIC method and the GMBIC method for various vectors v. The curves show for $\lambda \geqq 0$ the relative number of eigenvalues less than or equal to λ and are computed for $N = 225$, $h = 1/16$. For the RIC and RBIC methods the curves are plotted for $\omega = 0$, $\omega = 1$, and a value of ω corresponding approximately to the fastest rate of convergence with the initial approximation $x^{(0)} = 0$ and the stopping criterion $\|r^{(k)}\|_2 \leqq 10^{-7}\|r^{(0)}\|_2$. The latter ω is denoted by ω_{opt} and we have found $\omega_{\text{opt}} \approx 0.76$ for RIC method and $\omega_{\text{opt}} \approx 0.7$ for the RBIC method. In the GMIC and GMBIC method the

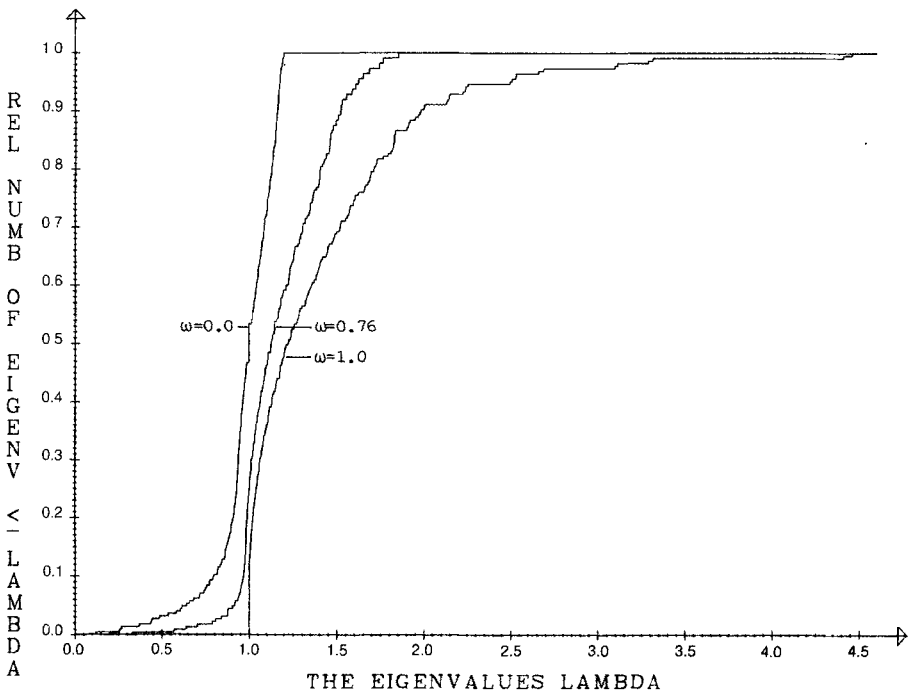

Fig. 4.1. The RIC Method, $N = 225$. The empirical distribution function for the eigenvalues of $C^{-1}A$ and the model problem

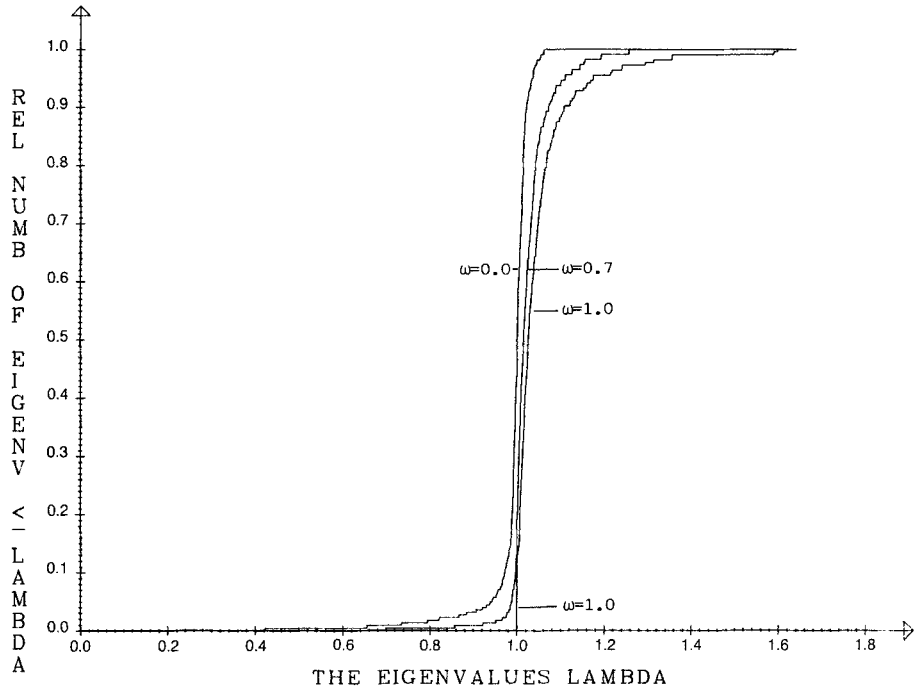

Fig. 4.2. The RBIC Method, $N=225$. The empirical distribution function for the eigenvalues of $C^{-1}A$ and the model problem

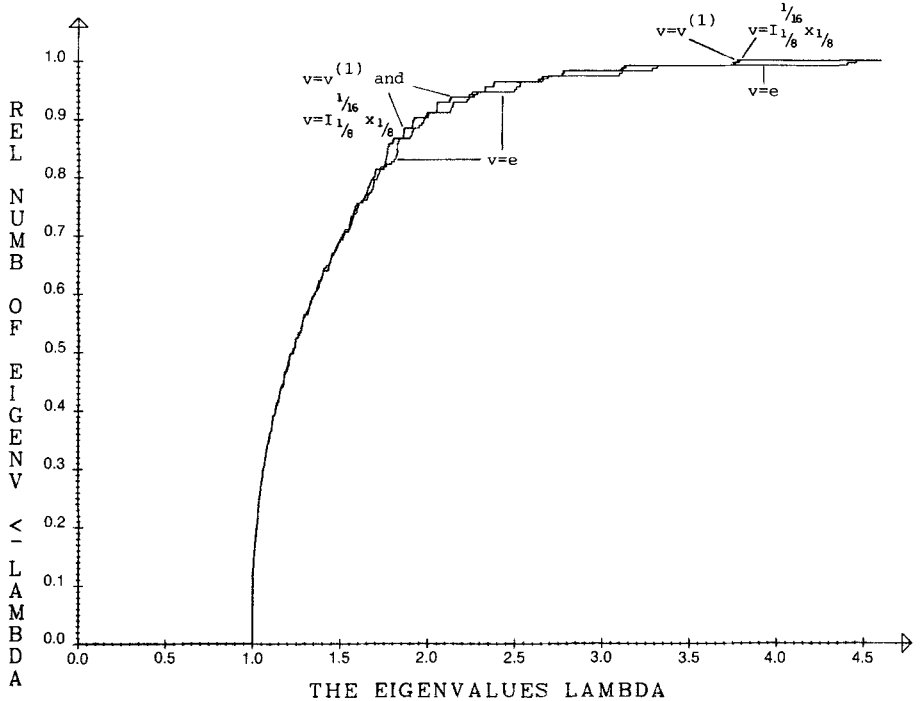

Fig. 4.3. The GMIC Method, $N=225$. The empirical distribution function for the eigenvalues of $C^{-1}A$ and the model problem

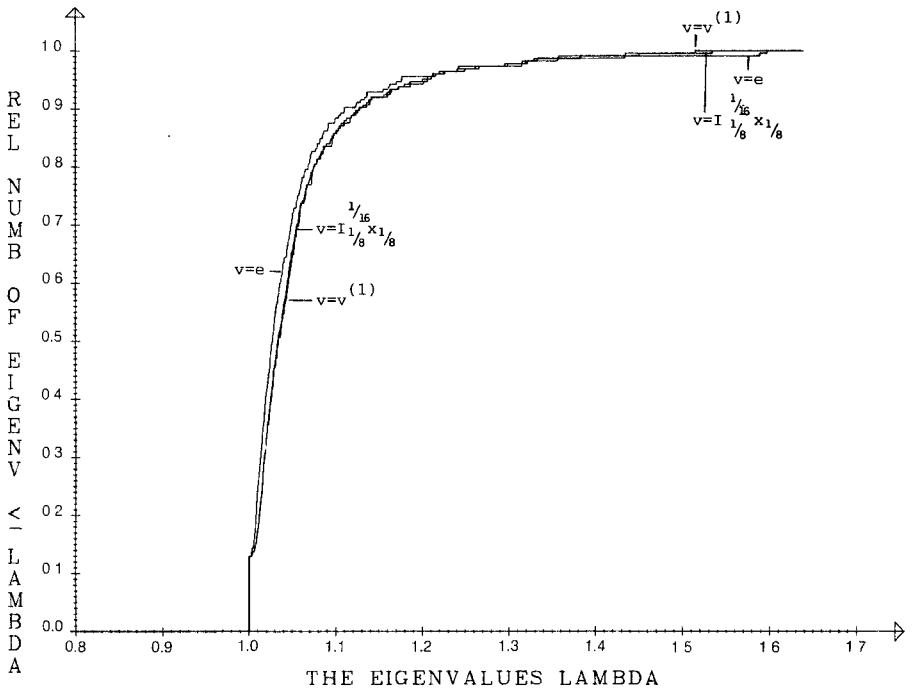

Fig. 4.4. The GMBIC Method, $N=225$. The empirical distribution function for the eigenvalues of $C^{-1}A$ and the model problem

vector v is chosen as $v=e$, $v=v^{(1)}$, where $v^{(1)}(x_i, y_i)=\sin \pi x_i \sin \pi y_i$, for nodes (x_i, y_i) of the triangulation, i.e. the first eigenvector of A, and $v=I_{1/8}^{1/16} x_{1/8}$, the vector which interpolates the approximate solution $x_{1/8}$ for $h=1/8$ onto the mesh $h=1/16$. $x_{1/8}$ is calculated with $x^{(0)}=0$ and the criterion (1.3) with $\varepsilon=10^{-7}$, where $\hat{x}$ is computed with the accuracy $\|r^{(k)}\|_2 \leqq 10^{-10}$. For the construction of C, $h=1/8$, the vector v is here calculated from the solution $u(x, y)$ in appropriate nodes. In most cases however, $u(x, y)$ is unknown and v is chosen for instance as $v=e$.

From Figs. 4.1-4.4 we conclude that case (i) in (2.3) is applicable to the RIC and RBIC methods with $\omega=1$ and the GMIC and GMBIC methods for the three vectors chosen. Further case (ii) in (2.3) is applicable to the RIC and RBIC methods with $\omega=0$ and case (iii) occurs for the RIC and RBIC methods with $\omega \approx \omega_{\text{opt}}$.

We now compare the upper bounds for the necessary number of iterations given in Sect. 2 with the actual number of iterations for problem (4.1) with the criterion (1.3) for $\varepsilon=10^{-7} \cdot \hat{x}$, used in (1.3), is computed with the accuracy $\|r^{(k)}\|_2 \leqq 10^{-10}$. In the GMIC and GMBIC methods we have $v=v^{(1)}$, (described above) and $v=I_{2h}^{h} x_{2h}$, i.e. for $h=h_0$ the vector which interpolates the approximate solution calculated for $h=2h_0$. In the latter case v is calculated from the solution $u(x, y)$ on the initial mesh, where $h=1/8$ (see Tables 4.1 and 4.2).

For each estimate the optimal number of k for $p \geqq 0$ and/or $q \geqq 0$ is given. In most cases we also give the value of p and/or q corresponding to the

optimal value of k found. Since the calculation of the eigenvalues is costly for large N the estimates are computed for $N \leqq 961$.

The actual number of iterations are computed for the following choices of initial approximations:

$$x^{(0)}(i)=0, \quad i^2 \text{ and Random } [-1,1], \quad i=1,2,\ldots,N.$$

Legend Tables 4.1 and 4.2. The various upper bounds corresponding to the cases (i)-(iii) in (2.3) are denoted as follows, where the bounds of the interval $[a,b]$ in (2.3) consists of actual eigenvalues:

(4.2) I: $k=\left\lceil \ln\frac{2}{\varepsilon}\Big/\ln\sigma^{-1}+\mathrm{q}\right\rceil$, $\sigma=\left(1-\sqrt{\frac{\lambda_1}{\lambda'_{q+1}}}\right)\Big/\left(1+\sqrt{\frac{\lambda_1}{\lambda'_{q+1}}}\right)$
(see (2.4)).

IIa: $k=\left\lceil \left(\ln\frac{2}{\varepsilon}+\ln\prod_{i=1}^{p}\frac{c_{\text{opt}}}{\lambda_i}\right)\Big/\ln\sigma^{-1}\right\rceil+p(r_{\text{opt}}+1)$,

where r_{opt} is the smallest nonnegative integer r for which $\frac{c(r)}{c(r+1)}\leqq\sigma^{-1}$, where $c(r)=\lambda'_1\left(\tan\frac{\pi}{4(r+1)}\right)\Big/(r+1)$, $c_{\text{opt}}=c(r_{\text{opt}})$
and

$$\sigma=\left(1-\sqrt{\frac{\lambda_{p+1}}{\lambda'_1}}\right)\Big/\left(1+\sqrt{\frac{\lambda_{p+1}}{\lambda'_1}}\right) \quad \text{(see (2.22))}.$$

IIb: The smallest k such that

$$\left|P_{k-p}^{(0,2p)}\left(-\frac{b+\bar{\lambda}_p}{b-\bar{\lambda}_p}\right)\right|\geqq\frac{1}{\varepsilon}\prod_{i=1}^{p}\frac{b-\bar{\lambda}_p}{\lambda_i}$$

for $\bar{\lambda}_p=\begin{cases}\hat{\lambda}_p & \text{(see (2.27))}\\ \bar{\lambda}_{p,1} & \text{(see (2.28))}.\end{cases}$

IIIa: $k=\left\lceil \left(\ln\frac{2}{\varepsilon''}+\ln\prod_{i=q+1}^{p}\frac{c_{\text{opt}}}{\lambda_i}\right)\Big/\ln\sigma^{-1}\right\rceil+(p-q)(r_{\text{opt}}+1)+2q$,

where r_{opt} is the smallest nonnegative integer r for which $\frac{c(r)}{c(r+1)}\leqq\sigma^{-1}$, $c(r)=\lambda'_{q+1}\left(\tan\frac{\pi}{4(r+1)}\right)\Big/(r+1)$, $c_{\text{opt}}=c(r_{\text{opt}})$,

$$\sigma=\left(1-\sqrt{\frac{\lambda_{p+1}}{\lambda'_{q+1}}}\right)\Big/\left(1+\sqrt{\frac{\lambda_{p+1}}{\lambda'_{q+1}}}\right),$$

$$\varepsilon''=\varepsilon\prod_{i=1}^{q}4\frac{\lambda_i}{\lambda'_i}\left(1-\frac{\lambda_i}{\lambda'_i}\right)^{-2} \quad \text{(see (2.24a), (2.24b))}.$$

IIIb: The smallest k such that

$$\left|P_{k-(p+q)}^{(0,2(p-q))}\left(-\frac{\lambda'_{q+1}+\bar{\lambda}_p}{\lambda'_{q+1}-\bar{\lambda}_p}\right)\right|\geqq\frac{1}{\varepsilon''}\prod_{i=q+1}^{p}\frac{\lambda'_{q+1}-\bar{\lambda}_p}{\lambda_i}$$

for $\bar{\lambda}_p=\bar{\lambda}_{p,1}$ and ε'' as in IIIa (see 2.31)).

Table 4.1. A comparison of the optimal values for $p \geqq 0$ and/or $q \geqq 0$ of the estimated number of iterations k and the actual numbers for the pointwise methods

Method	RIC, $\omega=0$						RIC, $\omega \approx \omega_{opt}$					
N	Estimates			Actual iterations with $x^{(0)}$:			Estimates			Actual iterations with $x^{(0)}$:		
	IIa	IIb $\hat{\lambda}_p$	IIb $\lambda^-_{p,1}$	Rand $[-1,1]$	i^2	0	IIa	IIIa $p \geqq q$	IIIb $p \geqq q$	Rand $[-1,1]$	i^2	0
49	11 ($p=1$)	12	11	10	10	9	11 ($p=1,2$)	11 ($p=q=2$)	11	10	10	8
225	20 ($p=1,3,4$)	21	20	16	17	14	16 ($p=1$)	16 ($p=q=2$)	16	13	14	12
961	38 ($p=1,3,4$)	38	38	29	29	26	25 ($p=1$)	25 ($q=0, p=1$)	26	20	21	17
3,969	-	-	-	49	59	49	-	-	-	25	27	24

Method	RIC, $\omega=1$			
N	Estimate I	Actual iterations with $x^{(0)}$:		
		Rand $[-1,1]$	i^2	0
49	11 ($q=0,2$)	10	9	9
225	16 ($q=2$)	15	13	13
961	24 ($q=2,4$)	21	19	19
3,969	-	30	26	28

Method	GMIC, $v=v^{(1)}$				GMIC, $v=I^h_{2h}x_{2h}$			
N	Estimate I	Actual iterations with $x^{(0)}$:			Estimate I[1]	Actual iterations with $x^{(0)}$:		
		Rand $[-1,1]$	i^2	0		Rand $[-1,1]$	i^2	0
49	10 ($q=0$)	9	9	6	10 ($q=0$)	9	9	8
225	15 ($q=0$)	14	13	9	15 ($q=0$)	14	14	13
961	22 ($q=0,2$)	20	19	13	22 ($q=0,2$)	23	24	18
3,969	-	29	27	18	-	32	-	22

[1] Here $x_{1/8}$ and $x_{1/16}$ are computed with $x(0)=0$

Table 4.2 A comparison of the optimal values for $p \geqq 0$ and/or $q \geqq 0$ of the estimated number of iterations k and the actual numbers for the block methods

Method	RBIC, $\omega = 0$					RBIC, $\omega \approx \omega_{opt}$				
N	Estimates			Actual iterations with $x^{(0)}$:		Estimates			Actual iterations with $x^{(0)}$:	
	IIa	IIb $\hat{\lambda}_p$	IIb $\hat{\lambda}_{p,1}^{-}$	Rand $[-1,1]$	0	IIa	IIIa $p \geqq q$	IIIb $p \geqq q$	Rand $[-1,1]$	0
49	6 ($p=1$)	6	6	5	5	6 ($p=1$)	6 ($p=q=1$)	6	5	5
225	10 ($p=1,2$)	10	10	8	8	9 ($p=1,2$)	9 ($p=q=2$)	9	7	7
961	17 ($p=1$)	18	17	12	14	13 ($p=1,2$)	12 ($p=q=2$)	13	9	10
3,969	–	–	–	22	26	–	–	–	13	14

Method	RBIC, $\omega = 1$		
N	Estimate I	Actual iterations with $x^{(0)}$:	
		Rand $[-1,1]$	0
49	5 ($q=0$)	5	5
225	8 ($q=0$)	7	8
961	13 ($q=0,2$)	11	11
3,969	–	15	16

Method	GMBIC, $v = v^{(1)}$			GMBIC, $v = I_{2h}^{h} x_{2h}$		
N	Estimate I	Actual iterations with $x^{(0)}$:		Estimate I[1]	Actual iterations with $x^{(0)}$:	
		Rand $[-1,1]$	0		Rand $[-1,1]$	0
49	6 ($q=0,1$)	5	5	6 ($q=0,1$)	5	5
225	8 ($q=0$)	7	7	8 ($q=0$)	7	7
961	11 ($q=0$)	10	8	11 ($q=0$)	10	8
3,969	–	14	11	–	14	11

[1] Here $x_{1/8}$ and $x_{1/16}$ are computed with $x(0) = 0$

Table 4.3. The estimate based on Jacobi polynomials for various $\bar{\lambda}_p$ and $p \geqq 0$. The RIC method with $\omega = 0$

N	$\bar{\lambda}_p$	p											
		0	1	2	3	4	5	6	7	8	9	10	11
49	*A*	14	12	11	11	11	11	11	11	12	12	-	-
	B	14	14	12	12	11	11	11	12	12	12	14	-
	C	14	14	12	12	12	12	12	13	13	14	14	-
225	*A*	28	22	21	21	20	19	19	20	20	-	-	-
	B	28	28	23	22	21	20	20	20	20	20	22	-
	C	28	28	23	23	21	21	21	21	21	21	22	-
961	*A*	56	43	41	40	38	37	37	37	37	37	38	-
	B	56	56	44	45	40	38	39	38	38	39	40	39
	C	56	56	44	45	40	40	39	38	39	38	39	-

Table 4.4. The estimate based on Jacobi polynomials for various $\bar{\lambda}_p$ and $p \geqq 0$. The RBIC method with $\omega = 0$

N	$\bar{\lambda}_p$	p									
		0	1	2	3	4	5	6	7	8	9
49	*A*	7	6	6	6	6	6	7	7	-	-
	B	7	7	6	6	6	6	7	7	8	-
	C	7	7	6	6	7	7	7	8	9	-
225	*A*	13	11	10	10	10	10	10	10	10	11
	B	13	13	11	10	10	10	10	10	11	-
	C	13	13	11	11	10	11	11	11	12	-
961	*A*	25	19	19	18	17	17	17	17	17	-
	B	25	25	20	19	18	17	17	17	18	-
	C	25	25	20	20	18	18	18	18	18	19

The eigenvalues have been calculated as described in [5] for the testproblem or are found from the model empirical distribution function.

A comparison for $p \geqq 0$ of the smallest k such that (2.26) is satisfied with A: $\bar{\lambda}_p = \bar{\lambda}_p^*$, i.e. $k^*(\bar{\lambda}_p^*)$ and B: $\bar{\lambda}_p = \bar{\lambda}_{p,1}$ i.e. $k^*(\bar{\lambda}_{p,1})$ and C: the smallest k such that (2.29) is satisfied, i.e. $k(\hat{\lambda}_p)$, is given in Tables 4.3 and 4.4 for the RIC and RBIC methods with $\omega = 0$.

$k^*(\bar{\lambda}_p^*)$ is computed by considering $\bar{\lambda}_p^{(i+1)} = \bar{\lambda}_p^{(i)} + 0.02$, $i = 0, 1, \ldots$, $\bar{\lambda}_p^{(0)} = 0.0$, $\bar{\lambda}_p^{(i)} < \lambda_{p+1}$. In $k^*(\bar{\lambda}_{p,1})$, $\bar{\lambda}_{p,1}$ is approximated by a $\bar{\lambda}_p$ which satisfies $\bar{\lambda}_p \geqq \bar{\lambda}_{p,1} - 0.01$.

Use of Model Empirical Distribution Functions

A more general study of the estimates (4.2) may be performed on some model functions simulating typical (c.f. Figs. 4.1–4.4) empirical distribution functions for the distribution of eigenvalues.

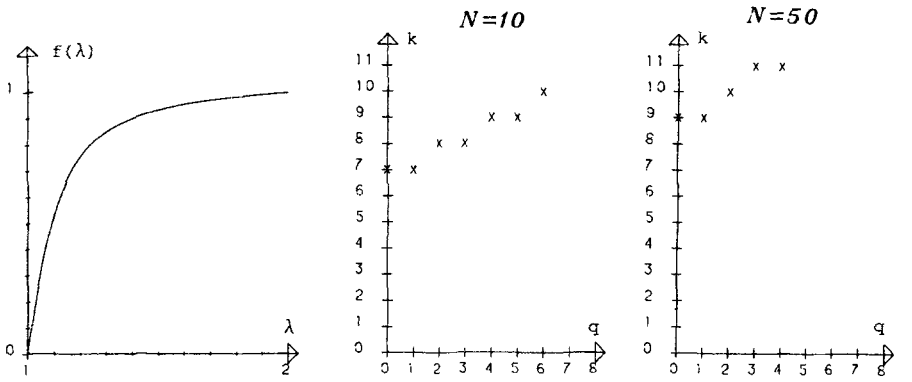

Fig. 4.5. Model distribution for $d=10$. Estimate I for $q \geqq 0$

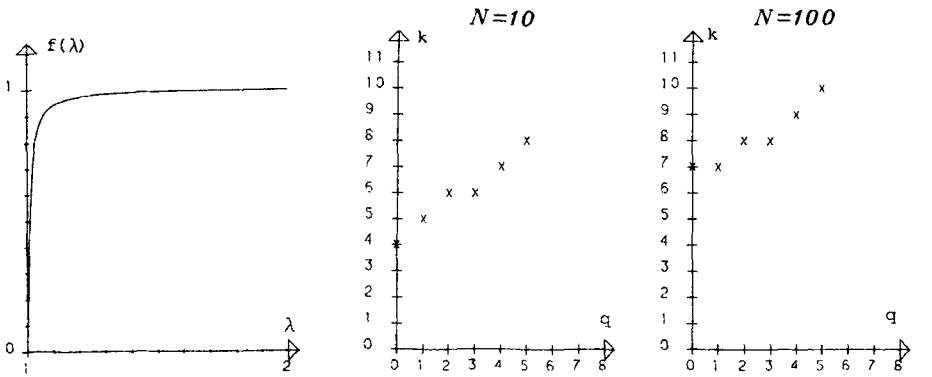

Fig. 4.6. Model distribution for $d=100$. Estimate I for $q \geqq 0$

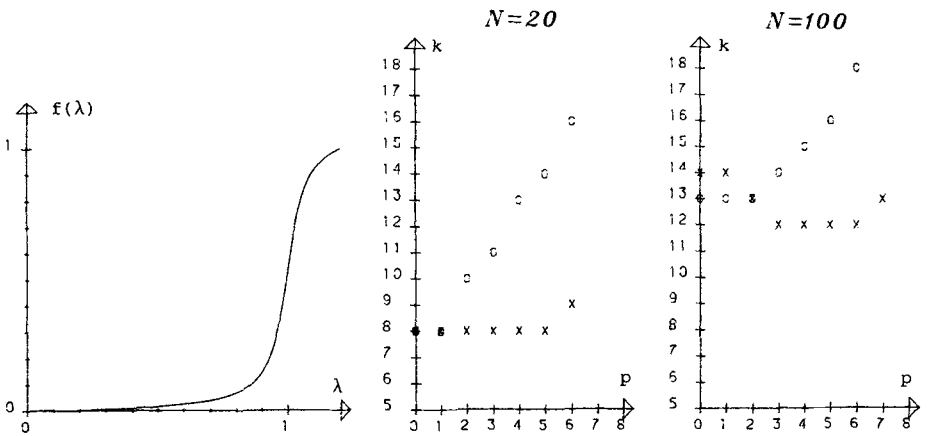

Fig. 4.7. Model distribution for $d=20$. Estimates IIa ($=0$) and IIb ($=x$) for $p \geqq 0$

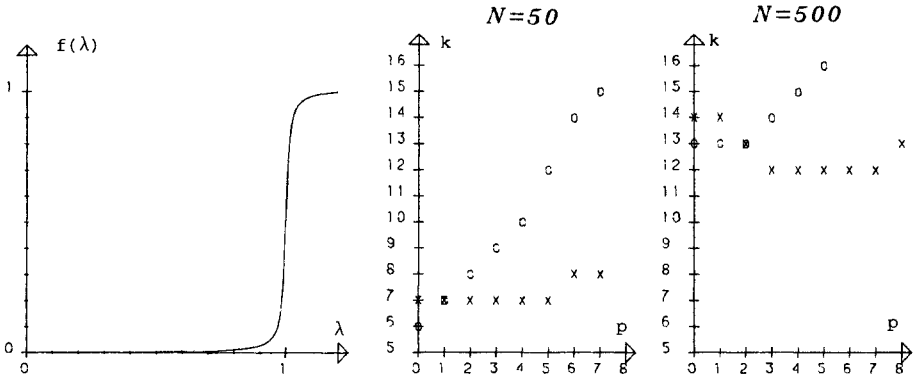

Fig. 4.8. Model distribution for $d=100$. Estimates IIa ($=0$) and IIb ($=x$) for $p \geqq 0$

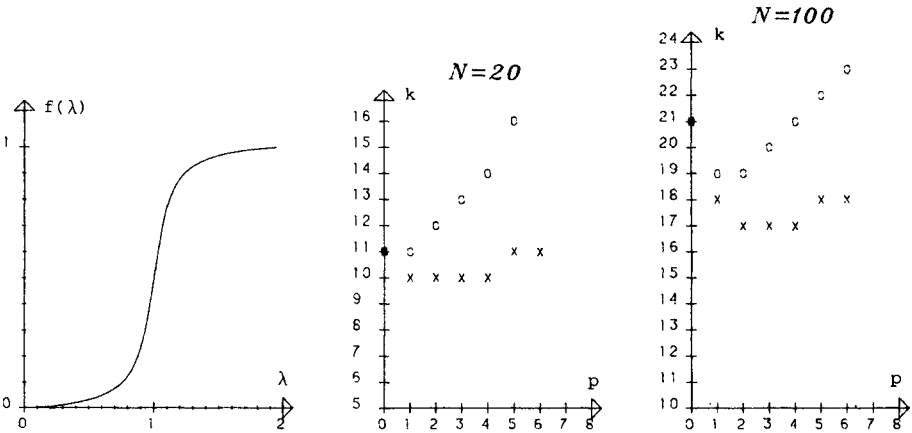

Fig. 4.9. Model distribution for $d=10$. Estimates IIa ($=0$) and IIIa, $p=q$ ($=x$)

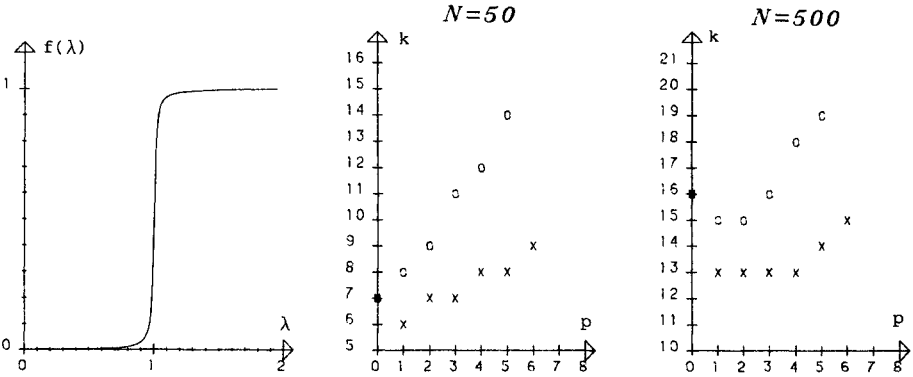

Fig. 4.10. Model distribution for $d=100$. Estimates IIa ($=0$) and IIIa, $p=q$ ($=x$)

Let
$$f(\lambda)=\frac{\arctan d(\lambda-1)+\arctan d(1-a)}{\arctan d(b-1)+\arctan d(1-a)}, \quad d\gg 1,\ 0<a<b.$$

Model distributions corresponding to cases (i)–(iii) in (2.3) are formed by suitable values of a, b and d. A set of N simulated eigenvalues $\{\lambda_i\}_{i=1}^N$ are computed by letting $f(\lambda_i)=\frac{i-1}{N}$ or $f(\lambda_i)=\frac{i}{N+1}$, $i=1,2,\ldots,N$ depending on which case we consider. By varying d and N different degrees of isolated eigenvalues are achieved. In all cases we have $\varepsilon=10^{-7}$.

Case (i). Let $a=1$, $b=2$. For $d=10$ and $d=100$ we have the model distributions in Figs. 4.5 and 4.6.

We let $f(\lambda_i)=\frac{i-1}{N}$, $i=1,2,\ldots,N$. The dependence of k on q for estimate I is shown for $N=10$, $N=50$ ($d=10$) and $N=100$ ($d=100$). In all cases $q=0$ corresponds to the smallest value of k i.e. no eigenvalue isolated enough exists giving an improved value of k.

Case (ii). With $a=0.05$, $b=1.2$ and $d=20$ and 100 we have the distribution functions in Figs. 4.7 and 4.8. Here we let $f(\lambda_i)=\frac{i}{N+1}$, $i=1,2,\ldots,N$. For $d=20$ we have chosen $N=20$ and 100. For $d=100$ we have $N=50$ and 500. In all cases the smallest value of k for $p\geqq 0$ is about the same for IIa and IIb.

Case (iii). Let $a=0.05$, $b=1.95$. The model distributions for $d=10$, 100 are shown in Figs. 4.9 and 4.10. We let $f(\lambda_i)=\frac{i}{N+1}$, $i=1,2,\ldots,N$. Since the small and large eigenvalues are equally separated IIIa, $p=q$ is a better estimate than IIa.

5. Conclusions

It has been demonstrated that the convergence behaviour of the conjugate gradient method for discrete small and large eigenvalues (case $0<\omega<1$) can be described essentially in the following way: The method eliminates a few (q) largest and smallest eigenvalue components (if they are distinctly separated from the remaining part of the spectrum). Moreover the method eliminates a few ($p-q$) of the components corresponding to the next smallest eigenvalues and eventually, when the remaining part $[a, b]$ (see Fig. 2.2) of the spectrum is essentially continuous, eliminates this approximately by a Chebyshev polynomial or a Jacobi polynomial.

In the estimates for the necessary number of iterations to reach a relative error ε the value of the relative error corresponding to the remaining interval $[a, b]$ depends on the relation of the small isolated eigenvalues to the interval $[a, b]$, see (2.5), (2.22) and (2.24a).

The estimates given are for a worst possible initial vector. For various choices of initial vectors we have found that the estimates are quite close to those actually found, in particular for case (i) (large isolated eigenvalues) and for case (iii) (both large and small eigenvalues). For the case with small eigenvalues (case (ii)), the estimates seem in some cases to be less accurate. If this depends on the initial vector or not is unclear.

It has also been found that the generalized (modified) block-incomplete factorization can sometimes improve the distribution of eigenvalues a little for a proper choice of the vector v, compared to the common choice, $v=e$.

Acknowledgements. Professor R. Varga kindly pointed out the relevance of reference [11] for this study. This proved to be of great value. Helpful comments by an anonymous referee are also acknowledged.

References

1. Andersson, L.: SSOR preconditioning of Toeplitz matrices. Thesis, Chalmers University of Technology, Göteborg, Sweden, 1976
2. Axelsson, O.: A class of iterative methods for finite element equations. Comput. Methods Appl. Mech. Eng. **9**, 123-137 (1976)
3. Axelsson, O.: A general incomplete block-matrix factorization method. Linear Algebra Appl. (to appear)
4. Axelsson, O., Brinkkemper, S., Il'in, V.P.: On some versions of incomplete block-matrix factorization iterative methods. Linear Algebra Appl. **58**, 3-15 (1984)
5. Axelsson, O., Lindskog, G.: On the eigenvalue distribution of a class of preconditioning methods. Report 3, Numerical Analysis Group, Department of Computer Sciences, Chalmers University of Technology, Göteborg, Sweden 1985
6. Greenbaum, A.: Comparison of splittings used with the conjugate gradient algorithm. Numer. Math. **33**, 181-194 (1979)
7. Gustafsson, I.: Modified Incomplete Cholesky (MIC) Methods. In: Preconditioning Methods, Theory and Applications (Evans, D.J., ed.), pp. 265-293. New York: Gordon and Breach Science Publishers 1983
8. Jennings, A.: Influence of the Eigenvalue Spectrum on the Convergence Rate of the Conjugate Gradient Method. JIMA **20**, 61-72 (1977)
9. Kaniel, S.: Estimates for some computational techniques in linear algebra. Math. Comput. **20**, 369-378 (1966)
10. Kershaw, D.: The incomplete Cholesky conjugate gradient method for the iterative solution of systems of linear equations. J. Comput. Phys. **26**, 43-65 (1978)
11. Saff, E.B., Varga, R.S.: On incomplete polynomials. In: Numerische Methoden der Approximationstheorie, Band 4. (L. Collatz, G. Meinardus, H. Werner, eds.), pp. 281-298. ISNM 42, Basel: Birkhäuser Verlag 1978
12. Szegö, G.: Orthogonal Polynomials. Americal Mathematical Society Colloquium Publications Volume XXIII, American Mathematical Society, Providence, Rhode Island, 1939
13. van der Vorst, H.A., van der Sluis, A.: The rate of convergence of conjugate gradients. Preprint Nr. 354, Department of Mathematics, University of Utrecht, 1984

Received April 11, 1985 / October 5, 1985