Numer. Math. 54, 719–734 (1988)

Numerische
Mathematik

Preconditioning Indefinite Discretization Matrices

Harry Yserentant

Fachbereich Mathematik, Universität Dortmund, Postfach 500500, D-4600 Dortmund 50,
Federal Republic of Germany

Summary. The finite element discretization of many elliptic boundary value problems leads to linear systems with positive definite and symmetric coefficient matrices. Many efficient preconditioners are known for these systems. We show that these preconditioning matrices can also be used for the linear systems arising from boundary value problems which are potentially indefinite due to lower order terms in the partial differential equation. Our main tool is a careful algebraic analysis of the condition numbers and the spectra of perturbed matrices which are preconditioned by the same matrices as in the unperturbed case.

Subject Classifications: AMS(MOS): 65F10, 65N20, 65N30; CR: G1.8.

1. Introduction

The finite element discretization of many elliptic boundary value problems leads to linear algebraic systems

$$Ax=b \tag{1.1}$$

with positive definite and symmetric coefficient matrices A. The most simple example of such a boundary value problem is the Laplace equation

$$-\Delta u=f \tag{1.2}$$

with boundary conditions

$$u=0 \tag{1.3}$$

on a sufficiently regular subdomain of the $\mathbf{R}^2$ or the $\mathbf{R}^3$.

Often systems like (1.1) are solved by applying a conjugate gradient type method implicitly to a preconditioned system

$$B^{-1/2}AB^{-1/2}y=B^{-1/2}b. \tag{1.4}$$

Many efficient preconditioners B are known. Examples are different types of multigrid methods [7, 3], methods based on incomplete factorizations of the coefficient matrix [1] or domain decomposition methods [4, 5, 11].

Typically all these methods get into trouble for equations like

$$-\Delta u+qu=f \tag{1.5}$$

with a selfadjoint lower order term forcing indefiniteness of the corresponding linear system

$$(A+M)x=b. \tag{1.6}$$

For a direct application to (1.5), (1.6) expensive modifications to these iterative methods can be necessary. In addition for parameter dependent equations the matrix M in (1.6) can change very often, whereas A remains fixed. In this paper we show that the resulting problems can be avoided by using the preconditioning matrices B arising from (1.1) also as preconditioners for the linear system (1.6).

Our approach is based on two observations. First, that the spectral condition number of

$$B^{-1/2}(A+M)B^{-1/2} \tag{1.7}$$

can be estimated in terms of the condition number of the unperturbed matrix

$$B^{-1/2}AB^{-1/2}. \tag{1.8}$$

The constant depends only on the stability properties of the boundary value problem and of its discretization, not on the preconditioner B. Secondly, and this is our main argument, the eigenvalues of the matrix (1.7) cluster in the interval bounded by the minimum and the maximum eigenvalue of the matrix (1.8). The number of eigenvalues of the matrix (1.7) outside every fixed small neighborhood of this interval is bounded independently of the choice of the finite element space.

The remainder of this paper is organized as follows: In Sect. 2 the basic algebraic estimates are derived. This section does not refer to the origin of the matrices. In Sect. 3 the connection with finite element discretizations is established; as an illustrative example of application we consider a simple second order boundary value problem. Sect. 4 deals with the consequences for Krylov-space methods.

2. The Basic Algebraic Estimates

We begin with some notations. Let

$$(x, y)=\sum_{i=1}^{n} x_i y_i \tag{2.1}$$

be the Euclidean inner product of two vectors $x, y \in \mathbf{R}^n$ and let

$$|x|=(x, x)^{1/2} \tag{2.2}$$

be the induced Euclidean norm. The associated matrix norm

$$|A| = \max_{|x|=1} |Ax| \tag{2.3}$$

is the spectral norm. The spectral condition number of an invertible square matrix A is

$$\kappa(A) = |A||A^{-1}|. \tag{2.4}$$

I denotes the identity matrix. Let $\lambda_1, \lambda_2 \ldots, \lambda_n > 0$ be the eigenvalues of the symmetric and positive definite $(n \times n)$-matrix A. Assume

$$Ax_i = \lambda_i x_i, \qquad (x_i, x_j) = \delta_{ij}, \tag{2.5}$$

for $i, j = 1, \ldots, n$. Then the symmetric and positive definite $(n \times n)$-matrices A^s are given by

$$A^s x = \sum_{j=1}^{n} \lambda_j^s (x, x_j) x_j. \tag{2.6}$$

In the remainder of this section we fix two symmetric positive definite $(n \times n)$-matrices A and B. In the application that we have in mind and as mentioned above, A is a discretization matrix of an elliptic boundary value problem and B a preconditioner for A. Define

$$\alpha = \min_{x \neq 0} \frac{(x, Ax)}{(x, Bx)}, \qquad \beta = \max_{x \neq 0} \frac{(x, Ax)}{(x, Bx)}. \tag{2.7}$$

Note that α is the minimum and β the maximum eigenvalue of the preconditioned matrix $B^{-1/2} A B^{-1/2}$, and that

$$\frac{\beta}{\alpha} = \kappa(B^{-1/2} A B^{-1/2}). \tag{2.8}$$

Assume that $A + M$ is another symmetric $(n \times n)$-matrix, typically a discretization matrix of a modified boundary value problem with M representing a lower order part of the differential operator. Our question is: What can be said about the spectral condition number and the eigenvalue distribution of

$$B^{-1/2}(A + M) B^{-1/2}, \tag{2.9}$$

quantifying the efficiency of B as a preconditioner for $A + M$?

We remark that for $B^{-1} = HH^T$ the symmetric matrices (2.9) and

$$H^T (A + M) H \tag{2.10}$$

are similar and have the same eigenvalues. This is proved using the orthogonality of $B^{1/2}H$.

We begin our analysis with:

Lemma 1. *Assume that $A+M$ is nonsingular. Then*

$$\kappa(B^{-1/2}(A+M)B^{-1/2})\leqq\kappa(I+A^{-1/2}MA^{-1/2})\kappa(B^{-1/2}AB^{-1/2}) \tag{2.11}$$

Proof. For all nonsingular $(n\times n)$-matrices A_1, A_2 one has

$$\kappa(A_1A_2)\leqq\kappa(A_1)\kappa(A_2).$$

Therefore

$$\begin{aligned}&\kappa(B^{-1/2}(A+M)B^{-1/2})\\&\quad\leqq\kappa(B^{-1/2}A^{1/2})\kappa(I+A^{-1/2}MA^{-1/2})\kappa(A^{1/2}B^{-1/2}).\end{aligned}$$

Using

$$|C|=|C^T|,\qquad |C|^2=|C^TC|$$

one gets

$$\kappa(B^{-1/2}A^{1/2})\kappa(A^{1/2}B^{-1/2})=\kappa(B^{-1/2}AB^{-1/2}),$$

and the proposition follows. □

To state our first result we need the energy norm

$$\|x\|_1=(x,Ax)^{1/2}=|A^{1/2}x| \tag{2.12}$$

and its dual norm

$$\|f\|_{-1}=\max_{\|x\|_1=1}(f,x)=|A^{-1/2}f|. \tag{2.13}$$

Theorem 1. *Let $A+M$ be nonsingular, and assume*

$$(x,(A+M)y)\leqq c_1\|x\|_1\|y\|_1,\qquad x,y\in\mathbf{R}^n, \tag{2.14}$$

and

$$\|(A+M)^{-1}f\|_1\leqq c_2\|f\|_{-1},\qquad f\in\mathbf{R}^n. \tag{2.15}$$

Then

$$\kappa(B^{-1/2}(A+M)B^{-1/2})\leqq c_1c_2\kappa(B^{-1/2}AB^{-1/2}). \tag{2.16}$$

Proof. Because of

$$\begin{aligned}\max_{\|x\|_1=\|y\|_1=1}(x,(A+M)y)&=\max_{|x|=|y|=1}(A^{-1/2}x,(A+M)A^{-1/2}y)\\&=|I+A^{-1/2}MA^{-1/2}|\end{aligned}$$

and

$$\begin{aligned}\max_{\|f\|_{-1}=1}\|(A+M)^{-1}f\|_1&=\max_{|A^{-1/2}f|=1}|A^{1/2}(A+M)^{-1}A^{1/2}A^{-1/2}f|\\&=|(I+A^{-1/2}MA^{-1/2})^{-1}|\end{aligned}$$

we have

$$\kappa(I+A^{-1/2}MA^{-1/2})\leqq c_1 c_2.$$

The theorem now follows from Lemma 1. □

Remark. The symmetry of M does not enter into the proofs of Lemma 1 and Theorem 1.

Now we examine the eigenvalue distribution of the matrix (2.9).

The eigenvalues $\lambda_1 \leqq \lambda_2 \leqq \ldots \leqq \lambda_n$ of an arbitrary symmetric $(n\times n)$-matrix S are given by the min-max characterization

$$\lambda_m = \min_{\dim \mathscr{U}=m} \max_{x\in\mathscr{U},\,|x|=1} (x, Sx) \tag{2.17}$$

and for $m=2, \ldots, n$ by the max-min characterization

$$\lambda_m = \max_{\dim \mathscr{V}=m-1} \min_{x\in\mathscr{V}^{\perp},\,|x|=1} (x, Sx), \tag{2.18}$$

where $\mathscr{U}$ and $\mathscr{V}$, respectively, run over all subspaces of $\mathbf{R}^n$ and $\mathscr{V}^{\perp}$ is the orthogonal complement of $\mathscr{V}$ with respect to the Euclidean inner product. We refer to [6, 10, 12].

Lemma 2. *Let $\mu_1 \leqq \mu_2 \leqq \ldots \leqq \mu_n$ be the eigenvalues of the symmetric matrix*

$$Q=I+A^{-1/2}MA^{-1/2} \tag{2.19}$$

and $\lambda_1 \leqq \lambda_1 \leqq \ldots \leqq \lambda_n$ be the eigenvalues of

$$B^{-1/2}(A+M)B^{-1/2}.$$

Then for all indices m with $\mu_m \geqq 0$

$$\alpha\mu_m \leqq \lambda_m \leqq \beta\mu_m \tag{2.20}$$

where α and β are given by (2.7).

Proof. Let $x_1, x_2, \ldots, x_n$ be an orthonormal basis of the $\mathbf{R}^n$ with

$$Qx_k=\mu_k x_k, \qquad k=1, \ldots, n.$$

For $m=1, \ldots, n$ we set

$$\mathscr{E}_m = \operatorname{span}\{x_1, \ldots, x_m\}.$$

For the proof of the lower estimate we define the space

$$\mathscr{V}_m = \{B^{-1/2}A^{1/2}y \mid y\in\mathscr{E}_m\}$$

and utilize

$$\mathscr{V}_m^{\perp} = \{B^{1/2}A^{-1/2}y \mid y\in\mathscr{E}_m^{\perp}\}.$$

By the max-min characterization (2.18) of λ_m, $m \geqq 2$, we have

$$\begin{aligned}\lambda_m &\geqq \min_{\substack{x\in\mathscr{V}_{m-1}^{\perp}\\ |x|=1}} (x, B^{-1/2}(A+M)B^{-1/2}x)\\ &= \min_{\substack{x\in\mathscr{V}_{m-1}^{\perp}\\ |x|=1}} (A^{1/2}B^{-1/2}x, QA^{1/2}B^{-1/2}x)\\ &= \min_{\substack{y\in\mathscr{E}_{m-1}^{\perp}\\ |B^{1/2}A^{-1/2}y|=1}} (y, Qy).\end{aligned}$$

Assuming $\mu_m \geqq 0$, for all $y \in \mathscr{E}_{m-1}^{\perp}$ with $|B^{1/2}A^{-1/2}y|=1$ one gets

$$(y, Qy) \geqq \mu_m |y|^2 \geqq \mu_m \min_{\substack{z\in\mathbf{R}^n\\ |B^{1/2}A^{-1/2}z|=1}} |z|^2.$$

Therefore

$$\begin{aligned}\lambda_m &\geqq \mu_m \min_{\substack{x\in\mathbf{R}^n\\ |x|=1}} |A^{1/2}B^{-1/2}x|^2\\ &= \mu_m \min_{\substack{x\in\mathbf{R}^n\\ |x|=1}} (x, B^{-1/2}AB^{-1/2}x)\\ &= \mu_m \alpha.\end{aligned}$$

Similarly, for $\mu_1 \geqq 0$, the characterization

$$\lambda_1 = \min_{\substack{x\in\mathbf{R}^n\\ |x|=1}} (x, B^{-1/2}(A+M)B^{-1/2}x)$$

leads to

$$\lambda_1 \geqq \mu_1 \alpha.$$

This proves the first part of the lemma.

For the proof of the second estimate we introduce the spaces

$$\mathscr{U}_m = \{B^{1/2}A^{-1/2}y \mid y \in \mathscr{E}_m\}.$$

The min-max characterization (2.17) of the eigenvalues λ_m gives

$$\begin{aligned}\lambda_m &\leqq \max_{\substack{x\in\mathscr{U}_m\\ |x|=1}} (x, B^{-1/2}(A+M)B^{-1/2}x)\\ &= \max_{\substack{x\in\mathscr{U}_m\\ |x|=1}} (A^{1/2}B^{-1/2}x, QA^{1/2}B^{-1/2}x)\\ &= \max_{\substack{y\in\mathscr{E}_m\\ |B^{1/2}A^{-1/2}y|=1}} (y, Qy).\end{aligned}$$

If $\mu_m \geqq 0$, for all $y \in \mathscr{E}_m$ with $|B^{1/2}A^{-1/2}y| = 1$

$$(y, Qy) \leqq \mu_m |y|^2 \leqq \mu_m \max_{\substack{z \in \mathbf{R}^n \\ |B^{1/2}A^{-1/2}z| = 1}} |z|^2.$$

This estimate leads to the desired upper bound

$$\begin{aligned} \lambda_m &\leqq \mu_m \max_{\substack{x \in \mathbf{R}^n \\ |x| = 1}} |A^{1/2}B^{-1/2}x|^2 \\ &= \mu_m \max_{\substack{x \in \mathbf{R}^n \\ |x| = 1}} (x, B^{-1/2}AB^{-1/2}x) \\ &= \mu_m \beta. \quad \square \end{aligned}$$

Now we can prove our second main theorem.

Theorem 2. *Assume* $0 < \delta \leqq 1$ *and*

$$\sum_{-1/\delta < \lambda < 0} \dim \ker(A - \lambda M) \leqq m_1, \qquad \sum_{0 < \lambda < 1/\delta} \dim \ker(A - \lambda M) \leqq m_2. \quad (2.21)$$

Then at most m_1 *eigenvalues of the matrix*

$$B^{-1/2}(A + M)B^{-1/2}$$

are less than $(1-\delta)\alpha$ *and at most* m_2 *eigenvalues of this matrix greater than* $(1+\delta)\beta$.

Proof. In the notations of Lemma 2 one has, for $\mu \neq 1$,

$$(Q - \mu I)x = 0$$

if and only

$$\left(A - \frac{1}{\mu - 1} M\right) A^{-1/2} x = 0.$$

Therefore

$$\begin{aligned} 1 - \delta \leqq \mu_i, &\quad i = m_1 + 1, \ldots, n, \\ \mu_i \leqq 1 + \delta, &\quad i = 1, \ldots, n - m_2. \end{aligned}$$

By Lemma 2 the proposition follows. □

Remark. For the application of Theorem 2 the matrix $A + M$ does not need to be nonsingular.

Note that by Sylvester's law of inertia the number of eigenvalues less than or equal to zero of the matrix (2.9) is independent of the choice of the preconditioner B.

Our last theorem allows to give upper bounds for the numbers m_1 and m_2 in (2.21).

Theorem 3. *Let N be a symmetric $(n \times n)$-matrix with*

$$\pm(x, Mx) \leqq (n, Nx), \qquad x \in \mathbf{R}^n. \tag{2.22}$$

Assume $\delta > 0$ and

$$\sum_{0<\lambda<1/\delta} \dim \ker(A - \lambda N) \leqq l. \tag{2.23}$$

Then

$$\sum_{-1/\delta<\lambda<0} \dim \ker(A - \lambda M) \leqq l, \qquad \sum_{0<\lambda<1/\delta} \dim \ker(A - \lambda M) \leqq l. \tag{2.24}$$

Proof. Because of

$$(x, A^{-1/2} M A^{-1/2} x) \leqq (x, A^{-1/2} N A^{-1/2} x), \qquad x \in \mathbf{R}^n,$$

and the min-max characterization (2.17), the k-th eigenvalue of $A^{-1/2} N A^{-1/2}$ is greater than or equal to the k-th eigenvalue of $A^{-1/2} M A^{-1/2}$. Therefore

$$\begin{aligned} &\sum_{0<\lambda<1/\delta} \dim \ker(A - \lambda M) \\ &\quad = \sum_{\mu>\delta} \dim \ker(A^{-1/2} M A^{-1/2} - \mu I) \\ &\quad \leqq \sum_{\mu>\delta} \dim \ker(A^{-1/2} N A^{-1/2} - \mu I) \\ &\quad = \sum_{0<\lambda<1/\delta} \dim \ker(A - \lambda N). \end{aligned}$$

Replacing M by $-M$ one gets

$$\begin{aligned} &\sum_{-1/\delta<\lambda<0} \dim \ker(A - \lambda M) \\ &\quad = \sum_{0<\lambda<1/\delta} \dim \ker(A + \lambda M) \\ &\quad \leqq \sum_{0<\lambda<1/\delta} \dim \ker(A - \lambda N). \quad \square \end{aligned}$$

The bounds given in the theorems of this section are sharp. For Lemma 1 and Theorem 1 as for Lemma 2 and Theorem 2 this is shown by the trivial example $B = A$. For Theorem 3 consider the matrices

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad M = \begin{pmatrix} 0 & 2 \\ 2 & 0 \end{pmatrix}, \quad N = \begin{pmatrix} 8 & 0 \\ 0 & 1 \end{pmatrix}.$$

3. Finite Element Equations

Let $\Omega \subseteq \mathbf{R}^d$, $d = 2$ or $d = 3$, be a bounded polygonal domain. As an example of application we consider finite element discretizations of the elliptic boundary value problem

$$-\sum_{i,j=1}^{d} D_j(a_{ij} D_i u) + qu = f \quad \text{on } \Omega, \tag{3.1}$$

$$u = 0 \quad \text{on } \partial\Omega. \tag{3.2}$$

We assume that the coefficient functions a_{ij} and q are bounded and measurable, that

$$a_{ij} = a_{ji}, \qquad i, j = 1, \ldots, d, \tag{3.3}$$

and that

$$\sum_{i,j=1}^{d} a_{ij}(x)\eta_i \eta_j \geqq \mu \sum_{i=1}^{d} \eta_i^2 \tag{3.4}$$

for almost all $x \in \Omega$ and all $\eta \in \mathbf{R}^d$. μ is a given positive constant.

The weak formulation of our boundary value problem is: find a function $u \in W_0^{1,2}(\Omega)$ satisfying

$$B(u, v) = f^*(v), \qquad v \in W_0^{1,2}(\Omega). \tag{3.5}$$

Here f^* is a given bounded linear functional on $W_0^{1,2}(\Omega)$ and the bilinear form B on $W_0^{1,2}(\Omega)$ is defined by

$$B(u, v) = \int_{\Omega} \left\{ \sum_{i,j=1}^{d} a_{ij} D_i u D_j v + q u v \right\} dx. \tag{3.6}$$

Let

$$B_0(u, v) = \int_{\Omega} \sum_{i,j=1}^{d} a_{ij} D_i u D_j v \, dx. \tag{3.7}$$

Under the given assumptions

$$\|u\|_1^2 = B_0(u, u) \tag{3.8}$$

defines a norm on $W_0^{1,2}(\Omega)$ which is equivalent to the usual $W_0^{1,2}(\Omega)$-norm of this space. There exists a constant $C_1 > 0$ with

$$B(u, v) \leqq C_1 \|u\|_1 \|v\|_1 \tag{3.9}$$

for all functions $u, v \in W_0^{1,2}(\Omega)$.

We assume that for all bounded linear functionals f^* on $W_0^{1,2}(\Omega)$ the boundary value problem (3.5) has a unique solution $u \in W_0^{1,2}(\Omega)$ satisfying

$$\|u\|_1 \leqq C_2 \sup_{\|v\|_1 = 1} |f^*(v)| \tag{3.10}$$

with a constant C_2 independent of f^*. For $q(x) \geqq 0$, $x \in \Omega$, one can choose $C_2 = 1$.

To produce an approximate solution of the boundary value problem (3.5) we specify a finite element space $S \subseteq W_0^{1,2}(\Omega)$ and look for a $u \in S$ satisfying

$$B(u, v) = f^*(v), \qquad v \in S. \tag{3.11}$$

We require that the discrete boundary value problem (3.11) is uniquely solvable and that, corresponding to (3.10), there exists a constant c_2 independent of f^* with

$$\|u\|_1 \leqq c_2 \sup_{v\in S,\, \|v\|_1=1} |f^*(v)|. \tag{3.12}$$

This condition holds uniformly for all sufficiently accurate finite element spaces S; see [8] and [14]. For $q(x)\geqq 0$, $x\in\Omega$, (3.12) is satisfied with the constant $c_2=1$, as in the continuous case.

For a given basis $\psi_1, \ldots, \psi_n$ of S the discrete boundary value problem (3.11) is equivalent to the matrix problem

$$(A+M)x=b, \tag{3.13}$$

where the coefficients of the $(n\times n)$-matrices A and M and of the right hand side b are defined by

$$A|_{kl}=B_0(\psi_k, \psi_l) \tag{3.14}$$

$$M|_{kl}=\int_\Omega q\psi_k\psi_l dx \tag{3.15}$$

$$b|_k=f^*(\psi_k), \tag{3.16}$$

and where

$$u=\sum_{k=1}^{n} x|_k\psi_k \tag{3.17}$$

is the solution of (3.11).

For these matrices A and M the assumption (2.15) in Theorem 1 is equivalent to (3.12) and means stability of the finite element discretization. Because of (3.9), (2.14) holds with $c_1=C_1$. Therefore for every symmetric and positive definite matrix B the condition number of the matrix

$$B^{-1/2}(A+M)B^{-1/2} \tag{3.18}$$

is bounded in terms of the condition number of the matrix

$$B^{-1/2}AB^{-1/2}. \tag{3.19}$$

The constant is the same for all (sufficiently accurate) finite element spaces and independent of the preconditioner B.

To examine the spectral behavior of the matrix (3.18) let w: $\Omega\to\mathbf{R}$ be a bounded measurable function satisfying

$$|q(x)|\leqq w(x) \tag{3.20}$$

and

$$w(x)>0 \tag{3.21}$$

for almost all $x \in \Omega$. The linear space of all measurable real-valued functions u on Ω with

$$\|u\|_0^2 = \int_\Omega w(x)|u(x)|^2\,dx \tag{3.22}$$

being finite is a Hilbert-space under the norm (3.22). The solution space $W_0^{1,2}(\Omega)$ is a compactly embedded subspace of this Hilbert-space. Therefore there exists a complete system of eigenfunctions $u_k \in W_0^{1,2}(\Omega)$, $k=1, 2, 3, \ldots$, and eigenvalues $0<\lambda_1 \leqq \lambda_2 \leqq \lambda_3 \leqq \ldots$ with

$$B_0(u_k, v) = \lambda_k \int_\Omega w u_k v\,dx, \qquad v \in W_0^{1,2}(\Omega), \tag{3.23}$$

$$\int_\Omega w u_k u_l\,dx = \delta_{kl} \tag{3.24}$$

and

$$\lim_{k\to\infty} \lambda_k = +\infty. \tag{3.25}$$

Corresponding to this eigensystem there exists a set of discrete eigenfunction $\tilde{u}_1, \ldots, \tilde{u}_n \in S$ and eigenvalues $0<\tilde{\lambda}_1 \leqq \tilde{\lambda}_2 \leqq \ldots \leqq \tilde{\lambda}_n$ with

$$B_0(\tilde{u}_k, v) = \tilde{\lambda}_k \int_\Omega w \tilde{u}_k v\,dx, \qquad v \in S, \tag{3.26}$$

$$\int_\Omega w \tilde{u}_k \tilde{u}_l\,dx = \delta_{kl}. \tag{3.27}$$

It is a well-known fact [10], which can be proved using an appropriate generalization of the min-max principle (2.17), that

$$\lambda_k \leqq \tilde{\lambda}_k, \qquad k=1, \ldots, n. \tag{3.28}$$

If we define the $(n \times n)$-matrix

$$N|_{kl} = \int_\Omega w \psi_k \psi_l\,dx \tag{3.29}$$

with the basis functions ψ_k of S introduced above, this means that for all given $\delta > 0$

$$\sum_{0<\lambda<1/\delta} \dim\ker(A-\lambda N) \tag{3.30}$$

is bounded independently of the finite element space S. By the choice (3.20) of the weight function w (2.22) holds for the matrices (3.15), (3.29). Therefore by Theorem 3 for every given $\delta > 0$ also the numbers

$$\sum_{-1/\delta<\lambda<0} \dim\ker(A-\lambda M), \qquad \sum_{0<\lambda<1/\delta} \dim\ker(A-\lambda M) \tag{3.31}$$

are bounded independently of the choice of the finite element space.

For $q(x)>0$ almost everywhere in Ω (allowing $q(x)=0$ along a curve, for example) the canonical choice of the function w is

$$w(x)=q(x) \tag{3.32}$$

With this choice the number of eigenvalues $\lambda_i<1/\delta$ of the original continuous problem becomes a bound for the corresponding discrete sum (3.31). Obviously this bound cannot be improved independently of the finite element space S. Generally, for $q(x)\neq 0$ almost everywhere in Ω the optimal function w is

$$w(x)=|q(x)|. \tag{3.33}$$

To get a qualitative result it is sufficient to choose

$$w(x)=\sup_{v\in\Omega}|q(y)|, \tag{3.34}$$

making the norm (3.22) to a constant multiple of the $L_2(\Omega)$-norm.

To finish our considerations we apply Theorem 2 and see that the eigenvalues of the preconditioned matrix (3.18) cluster, for every symmetric and positive definite preconditioner B, in the interval bounded by the minimum and maximum eigenvalue of the unperturbed matrix (3.19).

4. Consequences for Krylov-Space Methods

We attempt to solve the linear system

$$(A+M)x=b \tag{4.1}$$

with the $(n\times n)$-coefficient matrix of Sect. 2 (or Sect. 3, respectively) by a residual minimizing Krylov-space method (see [9], for example) using the matrix B as preconditioner. It is well-known that these methods exploit the eigenvalue distribution of the preconditioned matrix. One of their basic properties is the estimate

$$|B^{-1/2}r_j|\leqq\min_{P_j\in\Pi_j}\max_{\lambda_i}|P_j(\lambda_i)|\,|B^{-1/2}r_0| \tag{4.2}$$

for the error of the j-th iterate x_j.

$$r_j=b-(A+M)x_j \tag{4.3}$$

is the residual corresponding to x_j, Π_j denotes the set of all polynomials P_j of a degree less than or equal to j with $P_j(0)=1$, and $\lambda_1\leqq\lambda_2\leqq\ldots\leqq\lambda_n$ are the eigenvalues of the preconditioned matrix (2.9).

To get a rough idea of the performance of these methods we fix a value $\delta\in(0,1)$ and set

$$\alpha'=(1-\delta)\alpha, \qquad \beta'=(1+\delta)\beta, \tag{4.4}$$

where α and β are the constants (2.7). Let

$$P^*(\lambda)=\prod_{\lambda_i<\alpha'}\left(1-\frac{\lambda}{\lambda_i}\right),\quad Q^*(\lambda)=\prod_{\lambda_i>\beta'}\left(1-\frac{\lambda}{\lambda_i}\right) \tag{4.5}$$

be polynomials of orders m_1 and m_2 with $m=m_1+m_2$ and set

$$c^*=\max_{\lambda_i\in[\alpha',\beta']}|P^*(\lambda_i)|. \tag{4.6}$$

Restricting the minimum in (4.2) to all polynomials

$$P_j(\lambda)=P^*(\lambda)\,Q^*(\lambda)\,P_{j-m}(\lambda),\qquad P_{j-m}\in\Pi_{j-m}, \tag{4.7}$$

one obtains, for $j\geqq m$,

$$|B^{-1/2}r_j|\leqq c^*\min_{P_{j-m}\in\Pi_{j-m}}\max_{\lambda_i\in[\alpha',\beta']}|P_{j-m}(\lambda_i)|\,|B^{-1/2}r_0|. \tag{4.8}$$

Provided that m is small, as in the application of Sect. 3, after a short starting phase the speed of convergence of the iteration should be determined by the eigenvalues $\lambda_i\in[\alpha',\beta']$. Defining

$$\kappa'=\frac{\alpha'}{\beta'}=\frac{1+\delta}{1-\delta}\,\kappa(B^{-1/2}AB^{-1/2}) \tag{4.9}$$

and

$$q=\frac{\sqrt{\kappa'}-1}{\sqrt{\kappa'}+1}<1 \tag{4.10}$$

one has

$$\min_{P_k\in\Pi_k}\max_{\alpha'\leqq\lambda\leqq\beta'}|P_k(\lambda)|=\frac{2q^k}{1+q^{2k}}, \tag{4.11}$$

as it is shown in [1], for example. Therefore for $j\geqq m$

$$|B^{-1/2}r_j|\leqq 2c^*q^{j-m}|B^{-1/2}r_0|. \tag{4.12}$$

Up to the factor c^*/q^m and the factor $(1+\delta)/(1-\delta)$ in (4.9) this is the same well-known estimate for the speed of convergence as one gets for the case $M=0$. A bound for the constant c^* in terms of

$$a=\min_{i=1,\ldots,n}|\lambda_i|,\qquad b=\max_{i=1,\ldots,n}|\lambda_i| \tag{4.13}$$

and

$$\frac{b}{a}=\kappa(B^{-1/2}(A+M)B^{-1/2}) \tag{4.14}$$

is

$$c^*\leqq\left(1+\frac{b}{a}\right)^{m_1}. \tag{4.15}$$

For a detailed evaluation of the formula (4.2) and for related questions we refer to [2] and the papers cited therein.

As a simple standard illustration we consider the boundary value problem

$$\begin{aligned} -\Delta u+\omega u&=f \quad \text{on } \Omega,\\ u&=0 \quad \text{on } \partial\Omega, \end{aligned}$$

where $\bar{\Omega}=[0,1]^2$ is the unit square of $\mathbf{R}^2$ and ω is an arbitrary real constant. To discretize the boundary value problem we subdivide the square $\bar{\Omega}$ into small squares of sidelength h. As the discrete solution space we use the space S of all functions being continuous on the unit square and piecewise bilinear on the small subsquares. Using the standard nodal basis of S this discretization leads to the matrices A and M represented by the difference stars

$$\begin{bmatrix} -\frac{1}{3} & -\frac{1}{3} & -\frac{1}{3}\\ -\frac{1}{3} & \frac{8}{3} & -\frac{1}{3}\\ -\frac{1}{3} & -\frac{1}{3} & -\frac{1}{3} \end{bmatrix}$$

and

$$\omega h^2\begin{bmatrix} \frac{1}{36} & \frac{1}{9} & \frac{1}{36}\\ \frac{1}{9} & \frac{4}{9} & \frac{1}{9}\\ \frac{1}{36} & \frac{1}{9} & \frac{1}{36} \end{bmatrix},$$

respectively. As a preconditioning procedure we switched to the hierarchical basis formulation [13, 14] of the linear system to be solved. We counted the number j of iteration steps necessary to reach

$$|B^{-1/2}r_j|\leqq\varepsilon|B^{-1/2}r_0|.$$

Note that for $B^{-1}=SS^T$, because of the orthogonality of $S^TB^{1/2}$,

$$|B^{-1/2}r|=|S^Tr|.$$

The results for some representative values of ω and ε, the gridsize $h=1/80$, a (6×6)-grid as initial grid for the construction of the hierarchical basis and the right-hand side

$$f(x,y)=1$$

Table 1. The number of iteration steps for the given example

ω	$\varepsilon = 10^{-2}$	$\varepsilon = 10^{-4}$	$\varepsilon = 10^{-6}$
90	6	17	25
60	7	17	25
30	8	18	26
0	11	20	27
−10	12	21	28
−20	19	27	35
−30	12	21	29
−40	11	20	28
−50	10	20	28
−60	10	20	31
−70	10	21	31
−80	10	20	32
−90	11	21	33

are listed in Table 1. The condition numbers $C_1 C_2$ of the associated continuous problems differ considerably, and the fact, that the eigenvalue $2\pi^2$ of the Laplace operator is very near to 20, explains the relatively large number of iterations for $\omega = -20$. But as a general observation we can conclude that after a certain starting phase the speed of convergence does not depend very sensitively on the choice of ω.

Acknowledgement. The author gratefully acknowledges the hospitality of Peter Deuflhard, Konrad-Zuse-Zentrum für Informationstechnik Berlin, and the help of his staff during the preparation of this manuscript.

The author has been supported by a Konrad-Zuse-Fellowship.

References

1. Axelsson, O., Barker, V.A.: Finite element solution of boundary value problems: Theory and computation. New York: Academic Press 1984
2. Axelsson, O., Lindskog, G.: On the rate of convergence of the preconditioned conjugate gradient method. Numer. Math. **48**, 499–523 (1986)
3. Bank, R.E., Dupont, T.F., Yserentant, H.: The hierarchical basis multigrid method. Numer. Math. **52**, 427–458 (1988)
4. Bramble, J.H., Pasciak, J.E., Schatz, A.H.: An iterative method for elliptic problems and regions partitioned into substructures. Math. Comput. **46**, 361–369 (1986)
5. Bramble, J.H., Pasciak, J.E., Schatz, A.H.: The construction of preconditioners for elliptic problems by substructuring. I. Math. Comput. **47**, 103–134 (1986)
6. Courant, R., Hilbert, D.: Methoden der Mathematischen Physik. Berlin Heidelberg New York: Springer 1968
7. Hackbusch, W.: Multigrid methods and applications. Berlin Heidelberg New York: Springer 1985
8. Schatz, A.H.: An observation concerning Ritz-Galerkin methods with indefinite bilinear forms. Math. Comput. **28**, 959–962 (1974)
9. Stoer, J.: Solution of large systems of linear equations by conjugate gradient type methods. In: Bachem, A., Grötschel, M., Korte, B. (eds.) Mathematical Programming, the State of the Art. Berlin Heidelberg New York: Springer 1983

10. Weinstein, A., Stenger, W.: Methods of intermediate problems for eigenvalues. New York, London: Academic Press 1972
11. Widlund, O.B.: Iterative substructuring methods: Algorithms and theory for elliptic problems in the plane. (Preprint)
12. Wilkinson, J.H.: The algebraic eigenvalue problem. Oxford: Clarendon Press 1965
13. Yserentant, H.: On the multi-level splitting of finite element spaces. Numer. Math. **49**, 379–412 (1986)
14. Yserentant, H.: On the multi-level splitting of finite element spaces for indefinite elliptic boundary value problems. SIAM J. Numer. Anal. **23**, 581–595 (1986)

Received March 7, 1988/July 11, 1988