IMA Journal of Numerical Analysis (1997) **17**, 125–149

Multilevel preconditioning for perturbed finite element matrices

O. AXELSSON
Faculty of Mathematics and Informatics, University of Nijmegen, Toernooiveld, NL-6525 ED Nijmegen, The Netherlands

YU. R. HAKOPIAN
Department of Informatics and Numerical Mathematics, Yerevan State University, Yerevan, Armenia

AND

YU. A. KUZNETSOV
Institute of Numerical Mathematics, Russian Academy of Sciences, 32a, Leninskij Prospect, Moscow, 117334 Russia

[Received 4 August 1994 and in revised form 25 April 1996]

Multilevel preconditioning methods for finite element matrices for the approximation of second-order elliptic problems are considered. Using perturbations of the local finite element matrices by zero-order terms it is shown that one can control the smallest eigenvalues. In this way in a multilevel method one can reach a final coarse mesh, where the remaining problem to be solved has a condition number independent of the total degrees of freedom, much earlier than if no perturbations were used. Hence, there is no need in a method of optimal computational complexity to carry out the recursion in the multilevel method to a coarse mesh with a fixed number of degrees of freedom.

1. Introduction

Algebraic multilevel iteration methods can be used for quite general finite element matrices to construct a preconditioner of optimal order of computational complexity, that is, proportional to the degrees of freedom on the finest mesh (see Axelsson & Vassilevski (1989, 1990), for instance). This requires that the recursive decomposition of the grids is carried out until a coarse mesh is reached where the condition number is independent or nearly independent (see Axelsson & Neytcheva (1995)) of the meshsize, so that the problem can be solved with a cost which is proportional to the degrees of freedom on that mesh. However, as it turns out, there can be a significant overhead of data transport and recursive loops associated with methods using many levels. Therefore, it is of interest to consider methods where fewer levels are used but the condition number and arithmetic cost of the preconditioning is still of optimal order and the cost per iteration step is still proportional to the degrees of freedom on the finest mesh.

An efficient method to achieve a condition number of the coarsest used mesh which is independent of the total degrees of freedom is the method of perturbations, using zero-order terms to perturb the finite element matrices. The main idea of the method proposed consists in using on intermediate levels the stiffness matrices which are not the discrete analogues of an initial differential operator. In the case of a Helmholtz-type mesh operator $-\Delta u + \sigma u$ the stiffness

matrices for fine and coarse meshes correspond to different values of the coefficient σ. Namely, the stiffness matrix for the coarse level corresponds to the greater value of the coefficient (here the relationship between the value of σ on the coarse level $k(\sigma_k)$ and on the fine level $k+1(\sigma_{k+1})$ is taken as follows: $\sigma_k = 2^l\sigma_{k+1}$, $l \geqslant 0$). The perturbations depend on a parameter l, which governs the increase of the zero-order terms. Thus, the conditioning of the stiffness matrix in passing from the fine to the coarse mesh is improved not only due to enlarging the meshsize but also because of increasing the value of the coefficient σ. So we can achieve the required conditioning of the stiffness matrix on the coarsest mesh in fewer levels as compared to classical procedures. The latter means that the system on the coarsest mesh has a dimensionality large enough to be efficiently implemented on computers with parallel architecture. The idea just described was applied in Kuznetsov (1992) and Hakopian & Kuznetsov (1991). The algebraic multilevel preconditioners were constructed there using the method of partitioning (decomposing) a mesh into substructures. The rate of increase of the coefficient σ corresponds to the value $l \simeq 1$. Note that even if the original operator is $-\Delta u$ (i.e. with $\sigma = 0$) we can still construct the preconditioner based on $-\Delta u + \sigma u$ where $\sigma > 0$. Call the corresponding finite element matrices A and B, respectively. Then, since the perturbation σu to $-\Delta u$ only affects the eigenvalues by an amount $O(h^2)$, the matrix A_σ corresponding to $-\Delta u + \sigma u$ is spectrally equivalent to A and the optimal order preconditioner to B is therefore also an optimal order preconditioner to A.

Multilevel iteration methods of the type used here have been presented earlier in Axelsson & Vassilevski (1989, 1990), Vassilevski (1989). They are extensions of the two-level method, presented in Bank & Dupont (1980), Axelsson & Gustafsson (1983), Braess (1981). For a survey and presentation of perturbation methods for incomplete factorization methods see Axelsson & Barker (1984). The perturbation method allows us to use greater values of l than in Hakopian & Kuznetsov (1991). Increasing l increases the condition number of the multilevel preconditioned matrix but permits one to stop at an earlier level (less coarse) than with smaller values of l.

The remainder of the paper is organized as follows. In Section 2 the variational formulation of a second-order self-adjoint elliptic boundary value problem is posed and the sequence of perturbed finite element matrices $A^{(k)}$ $(k = 0, 1, \ldots, p)$ is presented. The matrix $A \equiv A^{(p)}$ that corresponds to the finest level p is spectrally equivalent to the original stiffness matrix B. Some basic results are given in Section 3 and in Section 4 the condition numbers of the perturbed matrices are analyzed. The multilevel preconditioning matrices and the associated condition numbers are presented in Section 5. The constructed multilevel preconditioner $M \equiv M^{(p)}$ for matrix A is considered as one for matrix B.

2. The perturbed finite element matrices

Let Ω be a plane polygonal domain which is a union of some number q of triangles G_m, $m = 1, 2, \ldots, q$. Suppose Γ_0 is a closed subset of $\partial\Omega$ consisting of

edges of the triangles G_m. Denote by $H_0^1(\Omega)$ the subspace of the Sobolev space $H^1(\Omega)$ that consists of the functions vanishing on Γ_0.

Consider the variational formulation of a second-order self-adjoint elliptic boundary value problem: for a given function $f \in L_2(\Omega)$ find the function $u \in H_0^1(\Omega)$ such that

$$A(u, v) = \int_\Omega fv \, dx \qquad \forall v \in H_0^1(\Omega), \tag{2.1}$$

where

$$A(u, v) \equiv \int_\Omega \left[\sum_{i,j=1}^{2} a_{ij} \frac{\partial u}{\partial x_j} \frac{\partial v}{\partial x_i} + auv \right] dx \tag{2.2}$$

is a symmetric bilinear form in u and v. As regards a_{ij} and a, we suppose there exist positive constants μ_0, μ_1 and μ such that

$$\mu_0 \sum_{i=1}^{2} \xi_i^2 \leqslant \sum_{i,j=1}^{2} a_{ij}(x)\xi_i\xi_j \leqslant \mu_1 \sum_{i=1}^{2} \xi_i^2 \qquad \forall \xi_1, \xi_2 \in \mathbb{R},$$

$$0 \leqslant a(x) \leqslant \mu$$

for all $x \in \overline{\Omega}$.

We construct a hierarchical sequence of grids $\{\omega_k\}$ by inserting additional nodes at the midedge points of the triangles of ω_k when forming the next finer grid ω_{k+1}, $k = 0, 1, \ldots, p-1$, where ω_0 is an initial coarsest grid with the nodes at the vertices of triangles G_m. Hence the grid ω_k corresponds to the kth level of refinement.

For all values $k = 0, 1, \ldots, p$ we introduce the following notation:

Q_k is the set of nodes of the grid ω_k that belong to $\overline{\Omega}\backslash\Gamma_0$;

n_k is the number of nodes in the set Q_k;

V_k is the space of functions continuous in Ω, linear in each triangle of the grid ω_k and vanishing on Γ_0.

By construction we have

$$Q_{k-1} \subset Q_k, \qquad k = 1, 2, \ldots, p.$$

Therefore at the kth level the partitioning $Q_k\backslash Q_{k-1}$ and Q_{k-1} of the nodes in Q_k can be used. The following ordering of the nodes will be used: the nodes from $Q_k\backslash Q_{k-1}$ are numbered first in some order and then the nodes from Q_{k-1}.

The familiar one-to-one correspondence holds between functions from V_k and coefficient vectors from $\mathbb{R}^{n_k}$. Namely, a function $\tilde{u} \in V_k$ is put in correspondence with a vector $u \in \mathbb{R}^{n_k}$, the ith component of which equals the value of the function $\tilde{u}$ at the ith node of the set Q_k.

In accordance with the rule for numbering the grid nodes, any coefficient vector $u \in \mathbb{R}^{n_k}$ $(k \geqslant 1)$ may be represented in the form

$$u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}, \qquad u_1 \in \mathbb{R}^{n_{k,1}}, \qquad u_2 \in \mathbb{R}^{n_{k,2}},$$

where $n_{k,1} = n_k - n_{k-1}$, $n_{k,2} = n_{k-1}$.

Let us consider the pth level of partitioning, which corresponds to the finest grid. According to the finite element method, in order to find an approximate solution of problem (2.1), we have to solve the system of grid equations

$$Bw = g \tag{2.3}$$

where the matrix B of size $n \times n$ $(n \equiv n_p)$ is such that the following relation is valid

$$(Bw, v) = A(\tilde{w}, \tilde{v})$$

for all $\tilde{w}, \tilde{v} \in V_p$ and the vector $g \in \mathbb{R}^n$ is determined by the relation

$$(g, v) = \int_\Omega f\tilde{v}\, dx,$$

which is valid for any function $\tilde{v} \in V_p$. Here $(\cdot, \cdot)$ is the standard inner product in $\mathbb{R}^n$.

Consider a triangle G_m $(1 \leq m \leq q)$. Let us denote the restriction of the grid ω_k onto the triangle G_m by $\omega_k^{(m)}$ (the nodes of the grid $\omega_k^{(m)}$ are those of the grid ω_k belonging to $\overline{G}_m$). Introduce the following notation:

$Q_k^{(m)}$ is the restriction of the set of nodes Q_k onto the triangle G_m;
$n_k^{(m)}$ is the number of nodes in the set $Q_k^{(m)}$;
$V_k^{(m)}$ is the space of restrictions of functions from V_k onto G_m.

The restriction of an arbitrary vector $u \in \mathbb{R}^{n_k}$ onto the triangle G_m is denoted by $u^{(m)}$. It is clear that $u^{(m)} \in \mathbb{R}^{n_k^{(m)}}$.

Let us take a triangle G_m and perform an isoparametric linear mapping $\mathcal{L}_m$ that transforms it into an equilateral unit triangle G in the plane of variables ξ_1, ξ_2. Under such a mapping the grid $\omega_k^{(m)}$ is transformed into a regular grid $\omega_k(G)$ whose elements are equilateral triangles. For the step size h_k of grid $\omega_k(G)$ the following equality holds:

$$h_k = 2^{-k}, \qquad k = 0, 1, \ldots, p.$$

Further, the linear mapping $\mathcal{L}_m$ transfers the set of nodes $Q_k^{(m)}$ of the grid $\omega_k^{(m)}$ into the set of nodes $Q_k^{(m)}(G)$ of the grid $\omega_k(G)$. Note, that the numbering of nodes remains unchanged under the mapping, i.e. the nodes of the set $Q_k^{(m)}(G)$ have the same numbers as their preimages from $Q_k^{(m)}$. Then, $V_k^{(m)}(G)$ is the space of functions continuous in G, linear in each triangle of the grid $\omega_k(G)$ and vanishing on the image of $\partial G_m \cap \Gamma_0$ under the linear mapping $\mathcal{L}_m$.

We now construct the sequence of perturbed matrices.

Consider then some partitioning level k $(0 \leq k \leq p)$ and an arbitrary triangle G_m $(1 \leq m \leq q)$. In the space $V_k^{(m)}(G)$ let us choose the standard nodal basis. With each node $\xi_i \in Q_k^{(m)}(G)$ we associate a basis function

$$\varphi_i^{(k)} \in V_k^{(m)}(G)$$

such that

$$\varphi_i^{(k)}(\xi_j) = \delta_{ij} \qquad \text{(the Kronecker symbol)}$$

for all nodes $\xi_j \in Q_k^{(m)}(G)$. For any function $\tilde{u}^{(m)} \in V_k^{(m)}(G)$ we have

$$\tilde{u}^{(m)} = \sum_{\xi_i \in Q_k^{(m)}(G)} u_i^{(m)} \varphi_i^{(k)}, \tag{2.4}$$

where $u_i^{(m)} = \tilde{u}^{(m)}(\xi_i)$.

Consider the bilinear forms:

$$l(u, v) \equiv \int_G \nabla u \nabla v \, d\xi$$

and

$$d(u, v) \equiv \int_G uv \, d\xi.$$

Define the matrices, associated with the nodal basis of $V_k^{(m)}(G)$, as follows:

$$L_m^{(k)} = \left[l(\varphi_i^{(k)}, \varphi_j^{(k)}) \right]_{\xi_i, \xi_j \in Q_k^{(m)}(G)} \tag{2.5}$$

and

$$D_m^{(k)} = \left[d(\varphi_i^{(k)}, \varphi_j^{(k)} \right]_{\xi_i, \xi_j \in Q_k^{(m)}(G)}. \tag{2.6}$$

Then, using the operation of assembling (see, for instance, Axelsson & Barker (1984)) construct $n_k \times n_k$ matrices

$$L^{(k)} = \text{assembl}\{\gamma_m L_m^{(k)}\}_{m=1}^q \tag{2.7}$$

and

$$D^{(k)} = \text{assembl}\{\gamma_m D_m^{(k)}\}_{m=1}^q, \tag{2.8}$$

where γ_m are certain positive constants, which will be specified later, in (2.17). The equalities (2.7) and (2.8) mean that the following identities hold for any vectors $u, v \in \mathbb{R}^{n_k}$:

$$(L^{(k)} u, v) = \sum_{m=1}^q \gamma_m (L_m^{(k)} u^{(m)}, v^{(m)}), \tag{2.9}$$

$$(D^{(k)} u, v) = \sum_{m=1}^q \gamma_m (D_m^{(k)} u^{(m)}, v^{(m)}). \tag{2.10}$$

Corresponding to the chosen ordering of the nodes, the matrices $L^{(k)}$ and $D^{(k)}$ for values $k \geq 1$ can be partitioned in two by two block forms

$$L^{(k)} = \begin{bmatrix} L_{11}^{(k)} & L_{12}^{(k)} \\ L_{21}^{(k)} & L_{22}^{(k)} \end{bmatrix}, \qquad D^{(k)} = \begin{bmatrix} D_{11}^{(k)} & D_{12}^{(k)} \\ D_{21}^{(k)} & D_{22}^{(k)} \end{bmatrix} \tag{2.11}$$

with submatrices $L_{ij}^{(k)}$ and $D_{ij}^{(k)}$ of order $n_{k,i} \times n_{k,j}$ $(i, j = 1, 2)$.

Let us now consider a set of parameters

$$\sigma_k > 0, \qquad k = 0, 1, \ldots, p$$

and define the sequence of matrices

$$A^{(k)} = L^{(k)} + \sigma_k D^{(k)}, \qquad k = 0, 1, \ldots, p. \tag{2.12}$$

The choice of the sequence $\{\sigma_k\}$ will be discussed in Section 4. Only note that σ_p is taken to be equal to unity (see (4.31)).

By analogy with (2.11) for $k \geqslant 1$ the matrices $A^{(k)}$ may be represented in the block form

$$A^{(k)} = \begin{bmatrix} A_{11}^{(k)} & A_{12}^{(k)} \\ A_{21}^{(k)} & A_{22}^{(k)} \end{bmatrix}. \tag{2.13}$$

From (2.11) and (2.12) it follows that

$$A_{ij}^{(k)} = L_{ij}^{(k)} + \sigma_k D_{ij}^{(k)}, \qquad i, j = 1, 2.$$

Consider the matrix $A^{(k+1)}$, where $0 \leqslant k \leqslant p-1$. The two by two block representation of the type (2.13) for the matrix $A^{(k+1)}$ can be written in the form

$$A^{(k+1)} = \begin{bmatrix} A_{11}^{(k+1)} & A_{12}^{(k+1)} \\ A_{21}^{(k+1)} & S^{(k+1)} + A_{21}^{(k+1)} A_{11}^{(k+1)^{-1}} A_{12}^{(k+1)} \end{bmatrix}, \tag{2.14}$$

where $S^{(k+1)}$ is the Schur complement defined as follows:

$$S^{(k+1)} = A_{22}^{(k+1)} - A_{21}^{(k+1)} A_{11}^{(k+1)^{-1}} A_{12}^{(k+1)}.$$

We shall consider the preconditioning procedure based on replacing the Schur complement $S^{(k+1)}$ in (2.14) by the matrix

$$\varepsilon_k A^{(k)} \tag{2.15}$$

with some coefficient $\varepsilon_k > 0$. In Section 4 we will establish the bounds of the spectrum of the matrix $(\varepsilon_k A^{(k)})^{-1} S^{(k+1)}$ and there the coefficients ε_k will also be chosen.

The matrix $A (\equiv A^{(p)})$ is spectrally equivalent to the original matrix B from (2.3) (recall that we take $\sigma_p = 1$). The spectral condition number of the matrix $A^{-1}B$ depends on the parameters γ_m involved in the definition of matrix A (see (2.7), (2.8) and (2.12)). It has been shown in Hakopian & Kuznetsov (1991) how these parameters should be chosen to minimize the estimate of $\mathrm{cond}(A^{-1}B)$.

We have

$$B = \mathrm{assembl}\{B_m\}_{m=1}^{q},$$

where the matrices B_m are determined with the help of relations

$$(B_m u^{(m)}, v^{(m)}) = \int_{G_m} \left[\sum_{i,j=1}^{2} a_{ij} \frac{\partial \tilde{u}^{(m)}}{\partial x_j} \frac{\partial \tilde{v}^{(m)}}{\partial x_i} + a \tilde{u}^{(m)} \tilde{v}^{(m)} \right] dx$$

which are valid for any vectors $u^{(m)}, v^{(m)} \in \mathbb{R}^{n_p^{(m)}}$ ($\tilde{u}^{(m)}, \tilde{v}^{(m)} \in V_p^{(m)}$ are piecewise-linear prolongations of vectors $u^{(m)}, v^{(m)}$ respectively). For each triangle G_m there exist positive constants $\delta_m^{(0)}$ and $\delta_m^{(1)}$ such that the equivalence relation

$$\delta_m^{(0)}(A_m u^{(m)}, u^{(m)}) \leqslant (B_m u^{(m)}, u^{(m)}) \leqslant \delta_m^{(1)}(A_m u^{(m)}, u^{(m)}) \tag{2.16}$$

is true for all vectors $u^{(m)} \in \mathbb{R}^{n_p^{(m)}}$. In (2.16) $A_m \equiv A_m^{(p)} = L_m^{(p)} + D_m^{(p)}$. Note that the constants $\delta_m^{(0)}$ and $\delta_m^{(1)}$ depend on the coefficients a_{ij}, a of the bilinear form (2.2) and the geometrical parameters of triangle G_m and do not depend on number of refinement levels (see, for example, Hakopian & Kuznetsov (1991)).

It has been shown in Hakopian & Kuznetsov (1991) that if we take parameters γ_m as follows

$$\gamma_m = \sqrt{\delta_m^{(0)}\delta_m^{(1)}}, \qquad m = 1, 2, \ldots, q \tag{2.17}$$

then we obtain the estimate

$$\operatorname{cond}(A^{-1}B) \leqslant \max_{1 \leqslant m \leqslant q} \frac{\delta_m^{(1)}}{\delta_m^{(0)}}. \tag{2.18}$$

3. Some basic results

In the present section we shall state some basic results which we need for evaluating the bounds of the spectrum of the matrix $(\varepsilon_k A^{(k)})^{-1} S^{(k+1)}$.

Consider first the two-level basis. Since

$$V_{k-1}^{(m)}(G) \subset V_k^{(m)}(G)$$

then the so-called two-level hierarchical basis functions (see Bank & Dupont (1980), Braess (1981), Axelsson & Gustafsson (1983) and Axelsson & Vassilevski (1989, 1990))

$$\psi_i^{(k)} \in V_k^{(m)}(G)$$

can be defined as follows:

$$\psi_i^{(k)} = \varphi_i^{(k)}, \qquad \xi_i \in Q_k^{(m)}(G) \setminus Q_{k-1}^{(m)}(G);$$
$$\psi_i^{(k)} = \varphi_i^{(k-1)}, \qquad \xi_i \in Q_{k-1}^{(m)}(G).$$

Similar to (2.4) any function $\tilde{u}^{(m)} \in V_k^{(m)}(G)$ can be represented in the form

$$\tilde{u}^{(m)} = \sum_{\xi_i \in Q_k^{(m)}(G) \setminus Q_{k-1}^{(m)}(G)} \hat{u}_i^{(m)} \psi_i^{(k)} + \sum_{\xi_i \in Q_{k-1}^{(m)}(G)} u_i^{(m)} \psi_i^{(k)}, \tag{3.1}$$

where $u_i^{(m)} = \tilde{u}^{(m)}(\xi_i)$.

Expressions (2.4) and (3.1) define an $n_k^{(m)} \times n_k^{(m)}$ matrix $J^{(m)}$ ($\equiv J_k^{(m)}$) which transforms any coefficient vector

$$\hat{u} = \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \end{bmatrix}$$

of the representation of $\tilde{u}^{(m)} \in V_k^{(m)}(G)$ in the two-level basis to the coefficient vector

$$u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$$

of the representation of the same function $\hat{u}^{(m)}$ in the nodal basis:

$$u = J^{(m)}\hat{u}.$$

By the ordering of the nodes the matrix $J^{(m)}$ has the following structure (see Axelsson & Vassilevski (1989), Vassilevski (1989)):

$$J^{(m)} = \begin{bmatrix} I & J_{12}^{(m)} \\ 0 & I \end{bmatrix}.$$

Let $\overline{L}_m^{(k)}$ and $\overline{D}_m^{(k)}$ be the matrices computed by the two-level basis functions of $V_k^{(m)}(G)$:

$$\overline{L}_m^{(k)} = [l(\psi_i^{(k)}, \psi_j^{(k)})]_{\xi_i, \xi_j \in Q_k^{(m)}(G)},$$
$$\overline{D}_m^{(k)} = [d(\psi_i^{(k)}, \psi_j^{(k)}]_{\xi_i, \xi_j \in Q_k^{(m)}(G)}.$$

The following simple relations hold (see Vassilevski (1989)):

$$\overline{L}_m^{(k)} = J^{(m)^T} L_m^{(k)} J^{(m)}, \qquad \overline{D}_m^{(k)} = J^{(m)^T} D_m^{(k)} J^{(m)}. \tag{3.2}$$

Having defined the matrices $\overline{L}_m^{(k)}$ and $\overline{D}_m^{(k)}$ for all $m = 1, 2, \ldots, q$ and using the operation of assembling we construct the $n_k \times n_k$ matrices

$$\overline{L}^{(k)} = \text{assembl}\{\gamma_m \overline{L}_m^{(k)}\}_{m=1}^q \tag{3.3}$$

and

$$\overline{D}^{(k)} = \text{assembl}\{\gamma_m \overline{D}_m^{(k)}\}_{m=1}^q, \tag{3.4}$$

where the constants γ_m are those of (2.7) and (2.8). As may be readily shown, there exists an $n_k \times n_k$ matrix J ($\equiv J_k$) such that the relations

$$\overline{L}^{(k)} = J^T L^{(k)} J, \qquad \overline{D}^{(k)} = J^T D^{(k)} J, \tag{3.5}$$

similar to those in (3.2), hold.

Let $k \geqslant 1$. If $\overline{L}^{(k)}$ and $\overline{D}^{(k)}$ are partitioned into two by two block form in the same manner as $L^{(k)}$ and $D^{(k)}$ (see (2.11)):

$$\overline{L}^{(k)} = \begin{bmatrix} \overline{L}_{11}^{(k)} & \overline{L}_{12}^{(k)} \\ \overline{L}_{21}^{(k)} & \overline{L}_{22}^{(k)} \end{bmatrix}, \qquad \overline{D}^{(k)} = \begin{bmatrix} \overline{D}_{11}^{(k)} & \overline{D}_{12}^{(k)} \\ \overline{D}_{21}^{(k)} & \overline{D}_{22}^{(k)} \end{bmatrix}, \tag{3.6}$$

then due to the definition of the two-level basis

$$\overline{L}_{11}^{(k)} = L_{11}^{(k)}, \quad \overline{L}_{22}^{(k)} = L^{(k-1)}; \qquad \overline{D}_{11}^{(k)} = D_{11}^{(k)}, \quad \overline{D}_{22}^{(k)} = D^{(k-1)}. \tag{3.7}$$

FIG. 1. (a) An element $e \in t_{k-1}$ and (b) the corresponding superelement E.

Consider now elements and superelements for the grid $\omega_k(G)$ corresponding to some partitioning level k, $0 \leqslant k \leqslant p$. Any triangular cell of the grid is called an element. By t_k we denote the set of triangular elements.

Let $k \geqslant 1$. Consider a triangular element $e \in t_{k-1}$. At the next stage of partitioning the grid the element e is subdivided into four elements. As a result, the element e turns into a superelement E (see Fig. 1).

For all $k = 1, 2, \ldots, p$ let T_k be the set of superelements of the kth level.

Let G_m $(1 \leqslant m \leqslant q)$ be a triangle in the domain Ω. For the local analysis we will need the restrictions of our coefficient vectors and matrices onto the elements and superelements.

Let the symbol ε denote here either a superelement E or an element e.

By u^ε we denote the restriction of a coefficient vector $u^{(m)} \in \mathbb{R}^{n_k^{(m)}}$ onto the (super) element ε.

Define matrices $L_\varepsilon^{(k)}$ and $D_\varepsilon^{(k)}$ by means of the relations

$$(L_\varepsilon^{(k)} u^\varepsilon, v^\varepsilon) = \int_\varepsilon \nabla \tilde{u}^{(m)} \nabla \tilde{v}^{(m)} \, d\xi \tag{3.8}$$

and

$$(D_\varepsilon^{(k)} u^\varepsilon, v^\varepsilon) = \int_\varepsilon \tilde{u}^{(m)} \tilde{v}^{(m)} \, d\xi \tag{3.9}$$

respectively, which are assumed to hold for all $\tilde{u}^{(m)}, \tilde{v}^{(m)} \in V_k^{(m)}(G)$ (the function $\tilde{u}^{(m)}$ (or $\tilde{v}^{(m)}$) is in one-to-one correspondence with the coefficient vector $u^{(m)}$ (or $v^{(m)}$)). It is easy to verify the identities

$$(L_m^{(k)} u^{(m)}, v^{(m)}) = \sum_{\varepsilon \in T_k \text{ (or } t_k)} (L_\varepsilon^{(k)} u^\varepsilon, v^\varepsilon) \tag{3.10}$$

and

$$(D_m^{(k)} u^{(m)}, v^{(m)}) = \sum_{\varepsilon \in T_k \text{ (or } t_k)} (D_\varepsilon^{(k)} u^\varepsilon, v^\varepsilon) \tag{3.11}$$

for any vectors $u^{(m)}$ and $v^{(m)}$ from $\mathbb{R}^{n_k^{(m)}}$.

Consider a triangular element e with vertices numbered 1, 2 and 3 (Fig. 1(a)). Let u and v be some functions, defined on the set of vertices. We shall consider the bilinear functional

$$\phi_e(u, v) \equiv (u_1+u_2)(v_1+v_2) + (u_2+u_3)(v_2+v_3) + (u_3+u_1)(v_3+v_1),$$

where u_i and v_i are the values of the functions u and v, respectively, at the ith vertex.

Now formulate the following important auxiliary statement (see Hakopian & Kuznetsov (1991)) which will simplify the calculation of integrals. It can be established by a straightforward calculation.

LEMMA 3.1 Let e be an equilateral triangle with side h. Then, for any functions u and v linear on e the following equalities hold

$$\int_e \nabla u \nabla v \, de = \frac{\sqrt{3}h}{6} \int_{\partial e} \frac{du}{ds} \frac{dv}{ds} \, ds \tag{3.12}$$

(here ∂e is the boundary of the triangle e, ds is an element of the boundary),

$$\int_e uv \, de = \frac{\sqrt{3}h^2}{48} \phi_e(u, v). \tag{3.13}$$

Further, the following easily proved statements will be used in the next section.

LEMMA 3.2 Suppose a_i, b_i, $i = 1, 2, \ldots, n$ are nonnegative. Suppose also that at least one b_i is nonzero and that for each i, $a_i = 0$ implies $b_i = 0$. Then

$$\frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i} \geqslant \min_{\substack{1 \leqslant i \leqslant n, \\ b_i \neq 0}} \frac{a_i}{b_i}. \tag{3.14}$$

LEMMA 3.3 Suppose a_i, b_i, $i = 1, 2, \ldots, n$ are nonnegative. Suppose also that at least one b_i is nonzero and that for each i, $b_i = 0$ implies $a_i = 0$. Then

$$\frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i} \leqslant \max_{\substack{1 \leqslant i \leqslant n, \\ b_i \neq 0}} \frac{a_i}{b_i}. \tag{3.15}$$

4. The eigenvalues of $S^{(k+1)}$ with respect to $\varepsilon_k A^{(k)}$

Consider the generalized eigenvalue problem

$$S^{(k+1)} u = \lambda(\varepsilon_k A^{(k)}) u. \tag{4.1}$$

The smallest and the largest eigenvalues of the problem (4.1) are denoted by $\lambda_{\min}^{(k)}$ and $\lambda_{\max}^{(k)}$, respectively. This section deals with finding bounds for these extreme eigenvalues for properly chosen sequences ε_k and σ_k in (2.12).

To estimate the smallest eigenvalue $\lambda_{\min}^{(k)}$, note that there exists a nonzero vector $v_2 \in \mathbb{R}^{n_{k+1,2}}$ such that the equality

$$\lambda_{\min}^{(k)} = \frac{(S^{(k+1)} v_2, v_2)}{\varepsilon_k (A^{(k)} v_2, v_2)} = \min_{u_2 \neq 0} \frac{(S^{k+1)} u_2, u_2)}{\varepsilon_k (A^{(k)} u_2, u_2)} \tag{4.2}$$

holds. Let the vector $v_1 \in \mathbb{R}^{n_{k+1,1}}$ satisfy the equation

$$A_{11}^{(k+1)} v_1 + A_{12}^{(k+1)} v_2 = 0.$$

Then

$$(S^{(k+1)} v_2, v_2) = (A^{(k+1)} v, v), \tag{4.3}$$

where the vector

$$v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \in \mathbb{R}^{n_{k+1}},$$

holds. Thus, from (4.2) and (4.3) we have the equality

$$\lambda_{\min}^{(k)} = \frac{(A^{(k+1)} v, v)}{\varepsilon_k (A^{(k)} v_2, v_2)}. \tag{4.4}$$

In accordance with (2.12) define the matrices

$$A_m^{(k)} = L_m^{(k)} + \sigma_k D_m^{(k)}, \qquad m = 1, 2, \ldots, q.$$

Then, using the identities (2.9), (2.10) and Lemma 3.2, (4.4) show that

$$\lambda_{\min}^{(k)} = \frac{\sum_{m=1}^{q} \gamma_m (A_m^{(k+1)} v^{(m)}, v^{(m)})}{\varepsilon_k \sum_{m=1}^{q} \gamma_m (A_m^{(k)} v_2^{(m)}, v_2^{(m)})} \geqslant \min_{\substack{1 \leqslant m \leqslant q \\ v_2^{(m)} \neq 0}} \frac{(A_m^{(k+1)} v^{(m)}, v^{(m)})}{\varepsilon_k (A_m^{(k)} v_2^{(m)}, v_2^{(m)})}. \tag{4.5}$$

Let m, $1 \leqslant m \leqslant q$, be such that the minimum of the right-hand side of the last inequality is attained. Define the matrices

$$A_E^{(k+1)} = L_E^{(k+1)} + \sigma_{k+1} D_E^{(k+1)}, \qquad E \in T_{k+1} \tag{4.6}$$

and

$$A_e^{(k)} = L_e^{(k)} + \sigma_k D_e^{(k)}, \qquad e \in t_k$$

(see (3.8) and (3.9)). Proceeding from (4.5) the inequality

$$\lambda_{\min}^{(k)} \geqslant \frac{\sum_{E \in T_{k+1}} (A_E^{(k+1)} v^E, v^E)}{\varepsilon_k \sum_{e_E \in t_k} (A_{e_E}^{(k)} v_2^E, v_2^E)} \tag{4.7}$$

is a simple consequence of the identities (3.10) and (3.11); here e_E is the element of kth level which on the next stage of refinement turns into the superelement $E \in T_{k+1}$.

Further, the inequality

$$(A_E^{(k+1)} v^E, v^E) \geqslant (S_E^{(k+1)} v_2^E, v_2^E) \tag{4.7}$$

holds, where

$$S_E^{(k+1)} = A_{22,E}^{(k+1)} - A_{21,E}^{(k+1)} A_{11,E}^{(k+1)^{-1}} A_{12,E}^{(k+1)} \tag{4.9}$$

is the Schur complement of the matrix $A_E^{(k+1)}$ from (4.6), represented in block form

$$A_E^{(k+1)} = \begin{bmatrix} A_{11,E}^{(k+1)} & A_{12,E}^{(k+1)} \\ A_{21,E}^{(k+1)} & A_{22,E}^{(k+1)} \end{bmatrix}$$

according to the chosen ordering of the nodes.

Therefore, from (4.7) and (4.8), using Lemma 3.2, we obtain:

$$\lambda_{\min}^{(k)} \geqslant \frac{\sum_{E \in T_{k+1}} (S_E^{(k+1)} v_2^E, v_2^E)}{\varepsilon_k \sum_{e_E \in t_k} (A_{e_E}^{(k)} v_2^E, v_2^E)} \geqslant \min_{\substack{E \in T_{k+1} \\ v_2^E \neq 0}} \frac{(S_E^{(k+1)} v_2^E, v_2^E)}{\varepsilon_k (A_{e_E}^{(k)} v_2^E, v_2^E)}$$

$$\geqslant \min_{E \in T_{k+1}} \min_{u_2^E \neq 0} \frac{(S_E^{(k+1)} u_2^E, u_2^E)}{\varepsilon_k (A_{e_E}^{(k)} u_2^E, u_2^E)}.$$

It is readily seen that the minimum over $E \in T_{k+1}$ in the right-hand side of the last inequality is attained on any superelement of which all the nodes belong to the set $Q_{k+1}^{(m)}(G)$. Let $E \in T_{k+1}$ be such a superelement. Then

$$\lambda_{\min}^{(k)} \geqslant \min_{u_2^E \neq 0} \frac{(S_E^{(k+1)} u_2^E, u_2^E)}{\varepsilon_k (A_{e_E}^{(k)} u_2^E, u_2^E)}.$$

The last inequality means that

$$\lambda_{\min}^{(k)} \geqslant \mu_k, \tag{4.10}$$

where μ_k is the smallest eigenvalue of the problem

$$S_E^{(k+1)} u = \mu(\varepsilon_k A_{e_E}^{(k)}) u. \tag{4.11}$$

Using relations (3.12), (3.13) and the node numbering given in Fig. 1(b), the matrices $A_E^{(k+1)}$ and $A_{e_E}^{(k)}$ take the following form:

$$A_E^{(k+1)} = \frac{\sqrt{3}}{6} \begin{bmatrix} 3\alpha_{k+1} & -2\beta_{k+1} & -2\beta_{k+1} & 0 & -\beta_{k+1} & -\beta_{k+1} \\ -2\beta_{k+1} & 3\alpha_{k+1} & -2\beta_{k+1} & -\beta_{k+1} & 0 & -\beta_{k+1} \\ -2\beta_{k+1} & -2\beta_{k+1} & 3\alpha_{k+1} & -\beta_{k+1} & -\beta_{k+1} & 0 \\ 0 & -\beta_{k+1} & -\beta_{k+1} & \alpha_{k+1} & 0 & 0 \\ -\beta_{k+1} & 0 & -\beta_{k+1} & 0 & \alpha_{k+1} & 0 \\ -\beta_{k+1} & -\beta_{k+1} & 0 & 0 & 0 & \alpha_{k+1} \end{bmatrix},$$

where

$$\alpha_{k+1} \equiv 2(1 + \tfrac{1}{8} z_{k+1}), \qquad \beta_{k+1} \equiv 1 - \tfrac{1}{8} z_{k+1}, \qquad z_{k+1} \equiv \sigma_{k+1} h_{k+1}^2; \tag{4.12}$$

$$A_{e_E}^{(k)} = \frac{\sqrt{3}}{6} \begin{bmatrix} \alpha_k & -\beta_k & -\beta_k \\ -\beta_k & \alpha_k & -\beta_k \\ -\beta_k & -\beta_k & \alpha_k \end{bmatrix},$$

where

$$\alpha_k \equiv 2(1 + \tfrac{1}{8}z_k), \qquad \beta_k \equiv 1 - \tfrac{1}{8}z_k, \qquad z_k \equiv \sigma_k h_k^2. \tag{4.13}$$

Then the Schur complement $S_E^{(k+1)}$ is calculated according to formula (4.9).

By straightforward calculations the following expressions for the eigenvalues of the problem (4.11) are derived:

$$\mu_1^{(k)} = \frac{1}{\varepsilon_k}\frac{(40+7z_{k+1})(24+z_{k+1})}{4(16+z_{k+1})(24+z_k)}, \qquad \mu_2^{(k)} = \frac{1}{\varepsilon_k}\frac{2(16+z_{k+1})z_{k+1}}{(8+5z_{k+1})z_k}. \tag{4.14}$$

Consequently, from (4.10), (4.14) we arrive at the following estimate for the smallest eigenvalue of the problem (4.1):

$$\lambda_{\min}^{(k)} \geqslant \min(\mu_1^{(k)}, \mu_2^{(k)}). \tag{4.15}$$

To estimate the largest eigenvalue $\lambda_{\max}^{(k)}$, let $w_2 \in \mathbb{R}^{n_{k+1,2}}$ be a nonzero vector such that the equality

$$\lambda_{\max}^{(k)} = \frac{(S^{(k+1)}w_2, w_2)}{\varepsilon_k(A^{(k)}w_2, w_2)} = \max_{u_2 \neq 0} \frac{(S^{(k+1)}u_2, u_2)}{\varepsilon_k(A^{(k)}u_2, u_2)} \tag{4.16}$$

holds.

In addition to the matrix $A^{(k+1)}$ consider the matrix

$$\overline{A}^{(k+1)} = \overline{L}^{(k+1)} + \sigma_{k+1}\overline{D}^{(k+1)} \tag{4.17}$$

(see (3.3) and (3.4)) having the block form

$$\overline{A}^{(k+1)} = \begin{bmatrix} \overline{A}_{11}^{(k+1)} & \overline{A}_{12}^{(k+1)} \\ \overline{A}_{21}^{(k+1)} & \overline{A}_{22}^{(k+1)} \end{bmatrix} \tag{4.18}$$

where, in accordance with (3.6),

$$\overline{A}_{ij}^{(k+1)} = \overline{L}_{ij}^{(k+1)} + \sigma_{k+1}\overline{D}_{ij}^{(k+1)}, \qquad (i, j = 1, 2). \tag{4.19}$$

The identity

$$\overline{A}^{(k+1)} = J^{\mathrm{T}}A^{(k+1)}J \tag{4.20}$$

follows immediately from (4.17) and (3.5).

The Schur complement of the matrix $\overline{A}^{(k+1)}$ takes the form:

$$\overline{S}^{(k+1)} = \overline{A}_{22}^{(k+1)} - \overline{A}_{21}^{(k+1)}\overline{A}_{11}^{(k+1)^{-1}}\overline{A}_{12}^{(k+1)}.$$

Based on the identity (4.20) a straightforward computation (see Vassilevski (1989), Axelsson & Vassilevski (1989)) shows that

$$S^{(k+1)} = \overline{S}^{(k+1)}. \tag{4.21}$$

Using equality (4.21), from (4.16) we have

$$\lambda_{\max}^{(k)} = \frac{(\overline{S}^{(k+1)} w_2, w_2)}{\varepsilon_k (A^{(k)} w_2, w_2)}. \tag{4.22}$$

Furthermore, the obvious inequality

$$(\overline{S}^{(k+1)} w_2, w_2) \leqslant (\overline{A}_{22}^{(k+1)} w_2, w_2)$$

holds. Taking into account (3.7) from (4.19) we get the expression for the block $\overline{A}_{22}^{(k+1)}$ of the block representation (4.18) of the matrix $\overline{A}^{(k+1)}$.

$$\overline{A}_{22}^{(k+1)} = L^{(k)} + \sigma_{k+1} D^{(k)}.$$

Therefore, proceeding from (4.22) and making use of identities (2.9), (2.10) and Lemma 3.3, we obtain:

$$\begin{aligned} \lambda_{\max}^{(k)} &\leqslant \frac{(\overline{A}_{22}^{(k+1)} w_2, w_2)}{\varepsilon_k (A^{(k)} w_2, w_2)} = \frac{((L^{(k)} + \sigma_{k+1} D^{(k)}) w_2, w_2)}{\varepsilon_k ((L^{(k)} + \sigma_k D^{(k)}) w_2, w_2)} \\ &= \frac{\sum_{m=1}^{q} \gamma_m ((L_m^{(k)} + \sigma_{k+1} D_m^{(k)}) w_2^{(m)}, w_2^{(m)})}{\varepsilon_k \sum_{m=1}^{q} \gamma_m ((L_m^{(k)} + \sigma_k D_m^{(k)}) w_2^{(m)}, w_2^{(m)})} \\ &\leqslant \max_{\substack{1 \leqslant m \leqslant q \\ w_2^{(m)} \neq 0}} \frac{((L_m^{(k)} + \sigma_{k+1} D_m^{(k)}) w_2^{(m)}, w_2^{(m)})}{\varepsilon_k ((L_m^{(k)} + \sigma_k D_m^{(k)}) w_2^{(m)}, w_2^{(m)})}. \end{aligned} \tag{4.23}$$

Let $1 \leqslant m \leqslant q$ be some number for which the maximum in the right-hand side of the last inequality is attained. Further, since the identities (3.10) and (3.11) hold, from (4.23) we get:

$$\begin{aligned} \lambda_{\max}^{(k)} &\leqslant \frac{((L_m^{(k)} + \sigma_{k+1} D_m^{(k)}) w_2^{(m)}, w_2^{(m)})}{\varepsilon_k ((L_m^{(k)} + \sigma_k D_m^{(k)}) w_2^{(m)}, w_2^{(m)})} \leqslant \frac{\sum_{e \in t_k} ((L_e^{(k)} + \sigma_{k+1} D_e^{(k)}) w_2^e, w_2^e)}{\varepsilon_k \sum_{e \in t_k} ((L_e^{(k)} + \sigma_k D_e^{(k)}) w_2^e, w_2^e)} \\ &\leqslant \max_{\substack{e \in t_k \\ w_2^e \neq 0}} \frac{((L_e^{(k)} + \sigma_{k+1} D_e^{(k)}) w_2^e, w_2^e)}{\varepsilon_k ((L_e^{(k)} + \sigma_k D_e^{(k)}) w_2^e, w_2^e)} \leqslant \max_{e \in t_k} \max_{u_2^e \neq 0} \frac{((L_e^{(k)} + \sigma_{k+1} D_e^{(k)}) u_2^e, u_2^e)}{\varepsilon_k ((L_e^{(k)} + \sigma_k D_e^{(k)}) u_2^e, u_2^e)}. \end{aligned}$$

A maximum over $e \in t_k$ is attained on any element of which all the nodes belong to the set $Q_k^{(m)}(G)$. Let e be such an element. Then

$$\lambda_{\max}^{(k)} \leqslant \max_{u_2^e \neq 0} \frac{((L_e^{(k)} + \sigma_{k+1} D_e^{(k)}) u_2^e, u_2^e)}{\varepsilon_k ((L_e^{(k)} + \sigma_k D_e^{(k)}) u_2^e, u_2^e)}.$$

This means that

$$\lambda_{\max}^{(k)} \leqslant \overline{\mu}_k, \tag{4.24}$$

where $\overline{\mu}_k$ is the largest eigenvalue of the problem

$$(L_e^{(k)} + \sigma_{k+1} D_e^{(k)})u = \mu(\varepsilon_k (L_e^{(k)} + \sigma_k D_e^{(k)}))u. \tag{4.25}$$

Taking advantage of relations (3.12) and (3.13) let us write the matrix

$$L_e^{(k)} + \sigma D_e^{(k)} = \frac{\sqrt{3}}{6} \begin{bmatrix} \alpha & -\beta & -\beta \\ -\beta & \alpha & -\beta \\ -\beta & -\beta & \alpha \end{bmatrix}, \qquad (\sigma = \sigma_k, \sigma_{k+1})$$

where

$$\alpha \equiv 2(1 + \tfrac{1}{8}\sigma h_k^2), \qquad \beta = 1 - \tfrac{1}{8}\sigma h_k^2.$$

Straightforward calculations give the following expressions for the eigenvalues of the problem (4.25):

$$\overline{\mu}_1^{(k)} = \frac{1}{\varepsilon_k} \frac{24 + 4z_{k+1}}{24 + z_k}, \qquad \overline{\mu}_2^{(k)} = \frac{1}{\varepsilon_k} \frac{4z_{k+1}}{z_k}, \tag{4.26}$$

where z_{k+1} and z_k are defined in (4.12) and (4.13), respectively.

Thus, from (4.24), (4.26) we obtain the estimate for the largest eigenvalue of the problem (4.1):

$$\lambda_{\max}^{(k)} \leqslant \max(\overline{\mu}_1^{(k)}, \overline{\mu}_2^{(k)}). \tag{4.27}$$

Let us take the coefficient ε_k in (2.15) as follows:

$$\varepsilon_k = \frac{24 + 4z_{k+1}}{24 + z_k}. \tag{4.28}$$

Thereby we get the following expressions for the eigenvalues from (4.14):

$$\mu_1^{(k)} = \frac{(40 + 7z_{k+1})(24 + z_{k+1})}{16(16 + z_{k+1})(6 + z_{k+1})}, \quad \mu_2^{(k)} = \frac{(16 + z_{k+1})z_{k+1}}{2(8 + 5z_{k+1})(6 + z_{k+1})} \frac{24 + z_k}{z_k} \tag{4.29}$$

and for the eigenvalues from (4.26):

$$\overline{\mu}_1^{(k)} = 1, \qquad \overline{\mu}_2^{(k)} = \frac{z_{k+1}}{6 + z_{k+1}} \frac{24 + z_k}{z_k}. \tag{4.30}$$

Let the parameters σ_k in (2.12) be defined by recursion as follows:

$$\begin{aligned} &\sigma_p = 1, \\ &\sigma_k = 2^l \sigma_{k+1}, \qquad k = p-1, p-2, \ldots, 0, \end{aligned} \tag{4.31}$$

where $l \geqslant 0$ is some integer.

Thus, on level p, which corresponds to the finest grid, we have a non-perturbed finite element matrix $A^{(p)}$. Then, in passing from the fine grid to the coarse one the coefficient of the mass matrix increases. Therefore the conditioning of the matrix $A^{(k)}$ is improved not only due to enlarging the step size of the grid but also because of increasing the value of the coefficient σ_k. This implies that we

can achieve the required conditioning of the matrix $A^{(k)}$ earlier, on some level $r > 0$, without descending to the coarsest level 0.

The simple relation

$$z_k = 2^{2+l} z_{k+1} \tag{4.32}$$

between quantities z_{k+1} and z_k, defined in (4.12) and (4.13), respectively, follows directly from (4.31). Then the expressions for $\mu_1^{(k)}, \mu_2^{(k)}$ from (4.29) and for $\overline{\mu}_1^{(k)}, \overline{\mu}_2^{(k)}$ from (4.30) take the following form:

$$\mu_1^{(k)} = \frac{(40 + 7z_{k+1})(24 + z_{k+1})}{16(16 + z_{k+1})(6 + z_{k+1})}, \qquad \mu_2^{(k)} = \frac{(16 + z_{k+1})(6 + 2^l z_{k+1})}{2^{1+l}(8 + 5z_{k+1})(6 + z_{k+1})} \tag{4.33}$$

and

$$\overline{\mu}_1^{(k)} = 1, \qquad \frac{1}{2^l} \leqslant \overline{\mu}_2^{(k)} = \frac{6 + 2^l z_{k+1}}{2^l(6 + z_{k+1})} \leqslant 1. \tag{4.34}$$

By combining the above obtained results (see (4.15), (4.33) and (4.27), (4.34)) we arrive at the following statement, which holds for all values $k = 0, 1, \ldots, p-1$.

THEOREM 4.1 Let the coefficient ε_k and parameter σ_k be chosen as shown in (4.28) and (4.31), respectively. Then, regardless of the choice of constants γ_m in (2.7) and (2.8), the eigenvalues of the matrix $(\varepsilon_k A^{(k)})^{-1} S^{(k+1)}$ belong to the interval $[d_k, 1]$, where

$$d_k = \min(\mu_1^{(k)}, \mu_2^{(k)}) \tag{4.35}$$

and $\mu_1^{(k)}$ and $\mu_2^{(k)}$ are given in (4.33).

We remark that the coefficients ε_k satisfy the inequalities

$$\frac{1}{2^l} \leqslant \varepsilon_k \leqslant 1, \qquad k = 0, 1, \ldots, p-1.$$

5. Multilevel perturbed preconditioning matrices

Having defined the sequence (2.12) of the perturbed finite element matrices $A^{(k)}$, in the present section we shall construct the multilevel preconditioning matrices making use of the approach proposed in Axelsson & Vassilevski (1989, 1990). This enables us to derive bounds of the spectral condition number of the preconditioned matrix which holds uniformly with respect to the level number.

Let us choose some level r, $0 \leqslant r < p$. Define the sequence of matrices

$$M^{(r+1)}, M^{(r+2)}, \ldots, M^{(p)}$$

as follows

$$M^{(k+1)} = \begin{bmatrix} A_{11}^{(k+1)} & A_{12}^{(k+1)} \\ A_{21}^{(k+1)} & \widetilde{S}^{(k)} + A_{21}^{(k+1)} A_{11}^{(k+1)^{-1}} A_{12}^{(k+1)} \end{bmatrix} \tag{5.1}$$

where

$$\tilde{S}^{(r)} = \varepsilon_r A^{(r)}, \tag{5.2}$$

$$\tilde{S}^{(k)} = \varepsilon_k A^{(k)}[I - P_\nu(M^{(k)^{-1}} A^{(k)})]^{-1}, \qquad k = r+1, r+2, \ldots, p-1; \tag{5.3}$$

here $P_\nu(x)$ is a polynomial of degree $\nu \geq 1$ which satisfies the following conditions:

$$0 \leq P_\nu(x) < 1, \qquad 0 < x \leq 1; \qquad P_\nu(0) = 1. \tag{5.4}$$

If we compare the two by two block representations of the matrix $A^{(k+1)}$ and the corresponding preconditioner $M^{(k+1)}$ (see (2.14) and (5.1)) then we see that the inverse of the Schur complement $S^{(k+1)}$ in (2.14) is replaced by the matrix polynomial $\tilde{S}^{(k)}$, involving the inverse of the preconditioner on the previous level and the stiffness matrix on the current level. In this way the preconditioner on the finest level is determined recursively, i.e. the definition of the polynomial $\tilde{S}^{(k)}$ permits us to organize multilevel recursion up to level r.

Set

$$A^{(p)} \equiv A, \qquad M^{(p)} \equiv M.$$

The matrix M will be referred to as a multilevel perturbed preconditioner for the matrix A.

For all values $k = r, r+1, \ldots, p-1$ consider the generalized eigenvalue problem

$$A^{(k+1)}u = \lambda M^{(k+1)}u. \tag{5.5}$$

As follows from the block forms (2.14) and (5.1) of the matrices $A^{(k+1)}$ and $M^{(k+1)}$, respectively, $\lambda = 1$ is an eigenvalue of the problem (5.5). The remainder of the eigenvalues are the eigenvalues of the problem

$$S^{(k+1)}u = \lambda \tilde{S}^{(k)}u.$$

Let $k = r$. Due to definition (5.2) of the matrix $\tilde{S}^{(k)}$ we find that the eigenvalues of the matrix $M^{(r+1)^{-1}} A^{(r+1)}$ belong to the interval $[\lambda_{r+1}, 1]$, where $\lambda_{r+1} = d_r$ (Theorem 4.1 has been used).

Now let $r+1 \leq k \leq p-1$. Assume that the eigenvalues of the matrix $M^{(k)^{-1}} A^{(k)}$ belong to the interval $[\lambda_k; 1]$, where $0 < \lambda_k < 1$. By definition (5.3) of the matrix $\tilde{S}^{(k)}$ and the properties (5.4) of the polynomial $P_\nu(x)$ we find that the eigenvalues of the matrix $M^{(k+1)^{-1}} A^{(k+1)}$ are contained in the interval $[\lambda_{k+1}, 1]$, where

$$\lambda_{k+1} = d_k(1 - \max_{[\lambda_k; 1]} P_\nu(x)). \tag{5.6}$$

Define the polynomial $P_\nu^{(k)}(x)$ as follows:

$$P_\nu^{(k)}(x) = \frac{T_\nu\left(\frac{1+\lambda_k-2x}{1-\lambda_k}\right) + 1}{T_\nu\left(\frac{1+\lambda_k}{1-\lambda_k}\right) + 1},$$

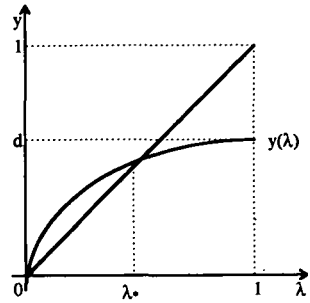

FIG. 2. Graph of the function $y(\lambda)$.

where $T_\nu(x) = \frac{1}{2}((x+\sqrt{x^2-1})^\nu + (x-\sqrt{x^2-1})^\nu)$ is a Chebyshev polynomial of the first-kind of degree ν. The polynomial satisfies conditions (5.4) and has the smallest local maximum in the interval $[\lambda_k; 1]$.

Consider the function

$$\psi^{(\nu)}(\lambda) \equiv \left[\frac{(1+\sqrt{\lambda})^\nu - (1-\sqrt{\lambda})^\nu}{(1+\sqrt{\lambda})^\nu + (1-\sqrt{\lambda})^\nu}\right]^2, \qquad 0 \leqslant \lambda \leqslant 1.$$

Proceeding from the equality (5.6), the following statement can easily be established.

THEOREM 5.1 For all values $k = r, r+1, \ldots, p-1$ the eigenvalues of the matrix $M^{(k+1)^{-1}} A^{(k+1)}$ belong to the interval $[\lambda_{k+1}, 1]$, where

$$\begin{aligned} &\lambda_{r+1} = d_r; \\ &\lambda_{k+1} = d_k \psi^{(\nu)}(\lambda_k), \qquad k = r+1, r+2, \ldots, p-1. \end{aligned} \tag{5.7}$$

Let us consider the function

$$y(\lambda) = d\psi^{(\nu)}(\lambda), \quad (0 < d < 1).$$

The function $y(\lambda)$ is continuous and increases monotonically on the interval $0 < \lambda < 1$. In addition,

$$y(0) = 0, \qquad y(1) = d, \qquad y'(0) = d\nu^2, \qquad y'(1) = 0$$

(see Fig. 2).

Provided the condition

$$d\nu^2 > 1 \tag{5.8}$$

is met, the equation

$$y(\lambda) = \lambda$$

has a positive solution λ_* on the interval $[0, 1]$ (see Fig. 2).

Clearly, since $d < 1$, we must choose $\nu > 1$. In order for the cost of the arithmetic operations of the preconditioner to be proportional to the dimension of the fine-grid system, we confine ourselves to values $\nu = 2$ and $\nu = 3$. We have (Axelsson & Vassilevski (1989))

$$\psi^{(2)}(\lambda) = \frac{4\lambda}{(1+\lambda)^2}, \qquad \lambda_* = 2\sqrt{d} - 1 \qquad (\nu = 2); \tag{5.9}$$

$$\psi^{(3)}(\lambda) = \lambda\left(\frac{\lambda+3}{3\lambda+1}\right)^2, \qquad \lambda_* = \tfrac{8\sqrt{d}-3(1-d)}{9-d} \qquad (\nu = 3). \tag{5.10}$$

We shall now find a lower bound d of d_k. Hence let $d > 0$ be such that

$$d \leqslant d_k, \qquad k = r, r+1, \ldots, p-1.$$

In addition to the sequence $\{\lambda_k\}_{k=r+1}^{p}$ from (5.7) consider the minorizing sequence $\{\widehat{\lambda}_k\}_{k=r+1}^{p}$ which is obtained from the recursion

$$\begin{aligned} &\widehat{\lambda}_{r+1} = d; \\ &\widehat{\lambda}_{k+1} = d\psi^{(\nu)}(\widehat{\lambda}_k), \qquad k = r+1, r+2, \ldots, p-1. \end{aligned} \tag{5.11}$$

Recursions (5.11) and (5.7) may be interpreted geometrically using Fig. 2 in a standard way.

The relations

$$\widehat{\lambda}_k \leqslant \lambda_k, \qquad k = r+1, r+2, \ldots, p \tag{5.12}$$

hold. Besides,

$$\lambda_* < \widehat{\lambda}_p < \cdots < \widehat{\lambda}_{r+2} < \widehat{\lambda}_{r+1}. \tag{5.13}$$

Insert the following functions

$$\varphi_1(z) = \frac{(40+7z)(24+z)}{16(16+z)(6+z)}, \qquad z \geqslant 0,$$

$$\varphi_2^{(l)}(z) = \frac{(16+z)(6+2^l z)}{2^{1+l}(8+5z)(6+z)}, \qquad z \geqslant 0$$

in accordance with expressions (4.33) for the eigenvalues $\mu_1^{(k)}$ and $\mu_2^{(k)}$, respectively.

Let us now discuss the choice of the power l in (4.31), which determines the rate of increase of the parameter σ_k.

5.1 *Case $l = 0$ (no increase)*

Assume that $r = 0$, which means that the descent is carried out upto the coarsest grid $\omega_0(G)$ with step size $h_0 = 1$.

Both the functions $\varphi_1(z)$ and $\varphi_2^{(0)}(z)$ are monotonically decreasing and furthermore,

$$\tfrac{7}{16} \leqslant \varphi_1(z) \leqslant \tfrac{5}{8}, \qquad \tfrac{1}{10} \leqslant \varphi_2^{(0)}(z) \leqslant 1$$

(see the graphs of the functions in Fig. 3).

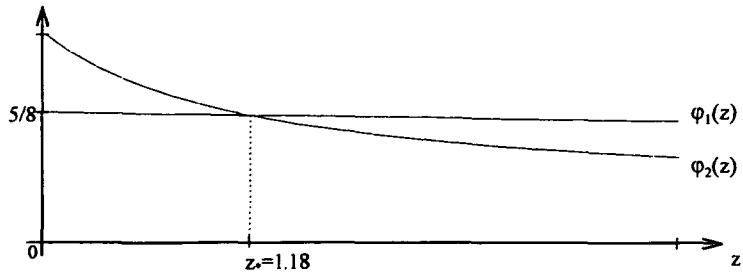

FIG. 3. Graphs of the functions $\varphi_1(z)$ and $\varphi_2^{(0)}(z)$.

Since $z_0 = h_0^2 = 1$, and taking into account the behaviour of functions φ_1 and $\varphi_2^{(0)}$, it follows from (4.35) that

$$d_k = \varphi_1(z_{k+1}), \qquad k = 0, 1, \ldots, p-1.$$

Further, as

$$z_1 = \tfrac{1}{4} z_0 = \tfrac{1}{4},$$

then

$$d_{p-1} > d_{p-2} > \cdots > d_0 = \varphi_1(\tfrac{1}{4}) \equiv d. \tag{5.14}$$

The quantity d takes the value:

$$d = \frac{16199}{26000} \geqslant 0{\cdot}623.$$

Therefore, in accordance with condition (5.8) we can take $\nu = 2$. For λ_* from (5.9), which due to (5.12) and (5.13) is a lower estimate of the eigenvalues of the matrix $M^{-1}A$, we get

$$\lambda_* \geqslant 0{\cdot}578.$$

5.2 *Case* $l \geqslant 1$

Let us choose the coarsest level r, $0 \leqslant r < p$, in the following way: r is taken as the largest integer for which the inequality

$$z_r \geqslant 1 \tag{5.15}$$

holds. Thus,

$$z_{r+1} < 1 \leqslant z_r. \tag{5.16}$$

If we make use of the expression

$$z_r = 2^{lp-(2+l)r}$$

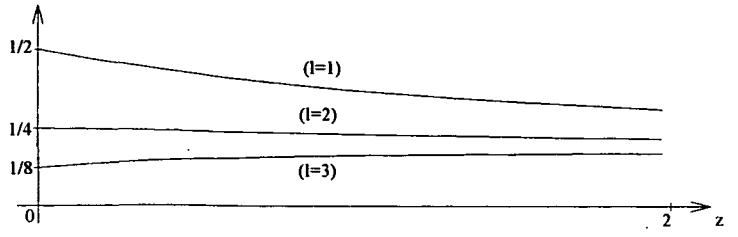

FIG. 4. Graphs of the functions $\varphi_2^{(l)}$ for $l = 1, 2, 3$.

for the quantity z_r, which follows from (4.32), then the criterion (5.15) allows us to get the formula for the level number:

$$r = \left\lfloor \frac{l}{2+l} p \right\rfloor, \tag{5.17}$$

where the symbol $\lfloor \cdots \rfloor$ stands for the integer part of the enclosed number.

Having examined the behaviour of the functions $\varphi_2^{(l)}(z)$ for the values $l = 1, 2$ and 3 (see Fig. 4) we find

$$d_k = \varphi_2^{(l)}(z_{k+1}), \qquad k = r, r+1, \ldots, p-1,$$

where we have used inequality (5.16).

Further,

$$d_{p-1} > d_{p-2} > \cdots > d_r > \varphi_2^{(l)}(1) \equiv d \tag{5.18}$$

for $l = 1, 2$ and

$$d_r > d_{r+1} > \cdots > d_{p-1} > \tfrac{1}{8} \equiv d$$

for $l = 3$.

By calculation, we get

$$d = \frac{34}{91} \quad \text{for } l = 1$$

and

$$d = \frac{85}{364} \quad \text{for } l = 2.$$

For $l = 3$, as mentioned above, $d = \frac{1}{8}$. Hence, (5.8) shows that we may take $\nu = 2$ for $l = 1$ and $\nu = 3$ for $l = 2, 3$.

Calculating the value of λ_* by formulae (5.9) and (5.10), we get:

$$\lambda_* \geqslant 0{\cdot}222 \quad \text{for } l = 1,$$
$$\lambda_* \geqslant 0{\cdot}178 \quad \text{for } l = 2,$$
$$\lambda_* \geqslant 0{\cdot}022 \quad \text{for } l = 3.$$

Recall that the level number r is specified in (5.17).

TABLE 1

l	$r=$	$\nu=$	$\lambda_* \geqslant$	$c^{(l)} \leqslant$
0	0	2	0·578	1·729
1	$\lfloor \frac{1}{3} p \rfloor$	2	0·222	4·495
2	$\lfloor \frac{1}{2} p \rfloor$	3	0·178	5·597
3	$\lfloor \frac{3}{5} p \rfloor$	3	0·022	43·628

For $l = 0, 1, 2, 3$ let

$$c^{(l)} = \frac{1}{\lambda_*}.$$

The results obtained are found in Table 1.

The following statement holds.

THEOREM 5.2 For each $l = 0, 1, 2, 3$, regardless of the choice of constants γ_m in (2.7) and (2.8), the estimate

$$\operatorname{cond}(M^{-1}A) \leqslant c^{(l)} \tag{5.19}$$

is valid, where $c^{(l)}$ is given in Table 1.

In this way we have constructed the multilevel preconditioner M for the matrix A. Let us now regard the matrix M as a multilevel preconditioner for the initial finite element matrix B.

The following inequality holds

$$\operatorname{cond}(M^{-1}B) \leqslant \operatorname{cond}(M^{-1}A)\operatorname{cond}(A^{-1}B).$$

The first condition number entering the right-hand side of the last inequality is estimated in Theorem 5.2 (see (5.19)). The estimate of the second one is given in (2.18). Thus, we arrive at the following statement.

THEOREM 5.3 Let the parameters γ_m be chosen according to (2.17), where $\delta_m^{(0)}$ and $\delta_m^{(1)}$ are defined in (2.16). Then for each value $l = 0, 1, 2, 3$ the estimate

$$\operatorname{cond}(M^{-1}B) \leqslant c^{(l)} \max_{1 \leqslant m \leqslant q} \frac{\delta_m^{(1)}}{\delta_m^{(0)}} \tag{5.20}$$

is valid, where the quantity $c^{(l)}$ is given in Table 1.

We consider now the computational complexity of the preconditioner.

In an iterative method with $M \equiv M^{(p)}$ as a preconditioner, we need to solve linear systems with matrix $M^{(k+1)}$ $(k = r, r+1, \ldots, p-1)$.

Let us discuss the process of solving the system

$$M^{(k+1)}u = f. \tag{5.21}$$

Proceeding from the two by two block structure (5.1) of matrix $M^{(k+1)}$, we find the following algorithm.

ALGORITHM

1° The vector

$$g_2 = f_2 - A_{21}^{(k+1)} A_{11}^{(k+1)^{-1}} f_1$$

is calculated;

2° The system

$$\widetilde{S}^{(k)} u_2 = g_2 \tag{5.22}$$

is solved;

3° The vector

$$u_1 = A_{11}^{(k+1)^{-1}} (f_1 - A_{12}^{(k+1)} u_2)$$

is calculated.

Let us analyze the algorithm in more detail.

Realization of items 1° and 3° requires two solutions with matrix $A_{11}^{(k+1)}$. The superelement approach shows readily that $\operatorname{cond}(A_{11}^{(k+1)}) \leqslant 4$. Therefore, we can solve a system with matrix $A_{11}^{(k+1)}$ to machine number precision by an iterative method in an optimal order of computational complexity (see Axelsson & Vassilevski (1989, 1990)).

We present now the realization of part 2°.

If $k = r$ then, in accordance with the definition (5.2) of matrix $\widetilde{S}^{(r)}$, the problem is reduced to the solution of the system

$$A^{(r)} u_2 = \frac{1}{\varepsilon_r} g_2. \tag{5.23}$$

Now let $k > r$. Define the polynomial

$$Q_{\nu-1}^{(k)}(x) = (1 - P_\nu^{(k)}(x))/x,$$

which may be written as

$$Q_{\nu-1}^{(k)}(x) = q_0^{(k)} + q_1^{(k)} x + \cdots + q_{\nu-1}^{(k)} x^{\nu-1}.$$

Then

$$I - P_\nu^{(k)}(M^{(k)^{-1}} A^{(k)}) = Q_{\nu-1}^{(k)}(M^{(k)^{-1}} A^{(k)}) \cdot (M^{(k)^{-1}} A^{(k)}).$$

Thus, because of the choice of the matrix $\widetilde{S}^{(k)}$ (see (5.3)), finding the solution of system (5.22) is equivalent to performing ν steps of the iterative procedure

$$\begin{aligned}
&\bar{g}_2 = \frac{1}{\varepsilon_k} g_2 \\
&u_2^{(0)} = 0 \\
&\text{for } j = 1 \text{ until } \nu \\
&M^{(k)} u_2^{(j)} = q_{\nu-j}^{(k)} \bar{g}_2 + A^{(k)} u_2^{(j-1)} \\
&\text{then } u_2 = u_2^{(\nu)}.
\end{aligned}$$

Introduce then the following notation:

w_{op} is the number of arithmetic operations required for solving a system with matrix M;

$v(r)$ is the number of arithmetic operations required for solving a system with matrix $A^{(r)}$.

The solution of a system with matrix $A_{11}^{(k+1)}$ requires $O(n_{k+1,1})$ operations. Since $n_{k+1} \leqslant 4n_k$ then $n_{k+1,1} = n_{k+1} - n_k \leqslant 3n_k$. Therefore, solving a system with matrix $A_{11}^{(k+1)}$ requires $C_1 n_k$ arithmetic operations, where $C_1 = \text{const} > 0$ depends on the iterative method used, but does not depend on k and ν.

By elementary considerations we get the inequality (see also Axelsson & Vassilevski (1991))

$$w_{\text{op}} \leqslant \tfrac{1}{4}(15\nu + 11)[n_p + \nu n_{p-1} + \cdots + \nu^{p-r-1} n_{r+1}] + 2C_1[n_{p-1} + \nu n_{p-2} + \cdots + \nu^{p-r-1} n_r] + \nu^{p-r-1} v(r). \tag{5.24}$$

From the above (see criterion (5.15)), we descend down to level r where the condition number of $A^{(r)}$ is $O(1)$. Hence, we may suppose that $v(r) \leqslant C_2 n_r$, where C_2 is some positive constant, which depends on the method of solving the system with matrix $A^{(r)}$, but does not depend on n_r. Taking into account the last fact, it follows that

$$w_{\text{op}} \sim n_p.$$

If a system with matrix $A^{(r)}$ is solved with an iterative method, then as a preconditioner for the matrix $A^{(r)}$ we may choose the diagonal matrix $\tilde{M}^{(r)}$ whose diagonal elements are obtained by summing all the elements of the corresponding row of matrix $A^{(r)}$. In Hakopian & Kuznetsov (1991) it was established that

$$\text{cond}(\tilde{M}^{(r)^{-1}} A^{(r)}) \leqslant \frac{24 + z_r}{4z_r}.$$

Hence, by (5.15),

$$\text{cond}(\tilde{M}^{(r)^{-1}} A^{(r)}) \leqslant 6{\cdot}25.$$

6. Concluding remarks

As has been shown in Axelsson & Eijkhout (1991) and Vassilevski (1989), it can be cost efficient to use polynomials of varying degree (as used in the present paper) instead of having a fixed degree ν. Also, in practice, systems with matrix $A_{11}^{(k)}$ must also be approximated. This has been discussed in Axelsson & Vassilevski (1990). Finally, as has been shown in Axelsson & Neytcheva (1995), for an optimal order method one can stop at an earlier level than the level where the condition number is bounded. What is required is that the total complexity for solving all coarse mesh problems on that level is not larger than the complexity of a matrix–vector multiplication on the finest level. The level one stops at depends on the solution method used for the coarse mesh problem. In order to limit the size of the present paper, we have not included these improvements.

REFERENCES

AXELSSON, O., & BARKER, V. A. 1984 *Finite Element Solution of Boundary Value Problems. Theory and Computation*, Orlando, FL: Academic.

AXELSSON O., & EIJKHOUT, V. 1991 The nested recursive two-level factorization method for nine-point difference matrices. *SIAM J. Sci. Stat. Comput.* **12,** 1373–1400.

AXELSSON, O., & GUSTAFSSON, I. 1983 Preconditioning and two-level multigrid methods of arbitrary degree of approximation. *Math. Comput.* **40,** 219–242.

AXELSSON, O., & NEYTCHEVA, M. 1995 Scalable parallel algorithms in CFD computations *Comput. Fluid Dyn. Rev.* **1,** 837–857 (M. Hafez and K. Oshima, eds; Chichester: Wiley).

AXELSSON, O., & VASSILEVSKI, P. S. 1989 Algebraic multilevel preconditioning methods, I. *Numer. Math.* **56,** 157–177.

AXELSSON, O., & VASSILEVSKI, P. S. 1990 Algebraic multilevel preconditioning methods, II. *SIAM J. Numer. Anal.* **27,** 1569–1590.

AXELSSON, O., & VASSILEVSKI, P. S. 1991 Asymptotic work estimates for AMLI methods. *Appl. Numer. Math.* **7,** 437–451.

BRAESS, D. 1981 The contraction number of a multigrid method for solving the Poisson equation. *Numer. Math.* **37,** 387–404.

BANK, R., & DUPONT, T. 1980 Analysis of a two-level scheme for solving finite element equations. *Report CNA-159*, Center for Numerical Analysis, The University of Texas at Austin.

HAKOPIAN, YU. R., & KUZNETSOV, YU. A. 1991 Algebraic multigrid/substructuring preconditioners on triangular grids. *Sov. J. Numer. Anal. Math. Modelling* **6,** 453–483.

KUZNETSOV, YU. A. 1992 A new parallel algebraic preconditioner. *J. Numer. Linear Algebra Appl.* **1,** 215–225.

VASSILEVSKI, P. S. 1989 Nearly optimal iterative methods for solving finite element elliptic equations based on the multilevel splitting of the matrix. *Report 1989-09*, Enhanced Oil Recovery Institute, The University of Wyoming, Laramie.