SIAM J. NUMER. ANAL.
Vol. 26, No. 6, pp. 1383–1394, December 1989

008

ON THE CONDITIONING OF FINITE ELEMENT EQUATIONS WITH HIGHLY REFINED MESHES *

RANDOLPH E. BANK † AND L. RIDGWAY SCOTT ‡

This paper is dedicated to Jim Douglas, Jr., on the occasion of his 60th birthday.

Abstract. It is proven that the condition number of the linear system representing a finite element discretization of an elliptic boundary value problem does not degrade significantly as the mesh is refined locally, provided the mesh remains *nondegenerate* and a natural scaling of the basis functions is used. Bounds for the Euclidean condition number as a function of the number of degrees of freedom are derived in $n \geq 2$ dimensions. When $n \geq 3$ the bound is the same as for the regular mesh case, but when $n = 2$ a factor appears in the bound for the condition number that is logarithmic in the ratio of the maximum and minimum mesh sizes. Applications of the results to the conjugate-gradient iterative method for solving such linear systems are given.

Key words. condition number, finite element method, refined meshes

AMS(MOS) subject classifications. 65N30, 65F35

0. Introduction.

We prove that the condition number of the linear system representing a finite element discretization need not degenerate as the mesh is refined locally, provided certain restrictions on the mesh are met and a natural scaling of the basis functions is used. The convergence properties of iterative methods, such as the conjugate-gradient method, for solving such linear systems can be estimated (cf. Luenberger [6]) in terms of the condition number of the system. And the sensitivity of the solution to perturbations in the right-hand side can be bounded using the condition number (cf. Isaacson and Keller [5]). Thus the condition number of the system can be of great interest.

A particular setting that we have in mind is the refinement of meshes (perhaps adaptively) to resolve singularities arising at angular points on the domain boundary or at points of discontinuity of the coefficients of the differential equation. It might seem, naïvely, that there would be large ratios of eigenvalues of the linear system (which would imply a large condition number) resulting from large mesh ratios. However, we show that this is not the case if a natural scaling of the finite element basis functions is used and the mesh is *nondegenerate* in a sense that is satisfied by many mesh generation schemes, both ones that adaptively refine the mesh based on intermediate calculations and ones based on a priori information about the boundary value problem (see [2]). We note that the question of optimal scaling of linear systems has been addressed in the past (cf. [7] and the references therein).

We shall consider the case when the finite element method is applied to approximate the solution of a linear, self-adjoint, elliptic boundary value problem in n dimensions. Of special interest will be the situation when the boundary is not smooth or when the coefficients of the partial differential equation are discontinuous. In these

*Received by the editors August 17, 1987; accepted for publication (in revised form) March 28, 1988.

†Department of Mathematics, University of California–San Diego, La Jolla, California 92093.

‡Department of Mathematics, Pennsylvania State University, University Park, Pennsylvania 16802.

cases, it is necessary to refine the mesh appropriately near boundary and coefficient singularities in order to approximate the solution efficiently. However, it has been widely believed that the resulting linear system of equations would be ill-conditioned, leading to slow convergence of iterative methods such as the conjugate-gradient method. It is worth noting that error estimates for direct methods, such as Gaussian elimination, also predict a degradation of performance for an ill-conditioned system. Thus without further justification, it would not be a remedy simply to use a direct method in such a situation. Fortunately, we are able to show that the condition number of such linear systems need not degrade unacceptably as the mesh is refined.

On a regular mesh of size h, the condition number of the finite element equations for a second-order elliptic boundary value problem can easily be seen to be $\mathcal{O}(h^{-2})$ using standard inverse estimates (see subsequent discussion for details). Also, the number, N, of degrees of freedom in this case is $N = \mathcal{O}(h^{-n})$, where n is the dimension of the domain of the boundary value problem. Thus, the condition number can be expressed in terms of the number of degrees of freedom as $\mathcal{O}(N^{2/n})$. In the case that $n \geq 3$, we shall show that the condition number is bounded by $\mathcal{O}(N^{2/n})$ for very general (so-called *nondegenerate*) meshes. In the case $n = 2$, our estimates for the condition number increase slightly from this optimal estimate by a logarithmic factor depending essentially on the ratio of the largest and smallest mesh sizes. (We show that this logarithmic factor can be sharp by an example.)

The condition number need not determine completely either the accuracy of a solution process with a given right-hand side (cf. Rice [8]) or the speed of convergence of an iterative process (cf. Luenberger [6]). For example, if a linear system has a single large eigenvalue, the conjugate-gradient method will not be affected adversely. However, if eigenvalues are distributed over a large range, it is quite conceivable that adverse effects would result. In the type of problems we envisage here, namely ones in which mesh sizes vary over a wide range, having eigenvalues ranging in size correspondingly could be quite detrimental. Thus our prescriptions for avoiding such a spread of eigenvalues is of interest.

1. Notation and preliminary inequalities. Let Ω be a bounded open set in $\mathbb{R}^n$ $(n \geq 2)$. We suppose that we are solving a boundary value problem posed variationally with boundary conditions incorporated in a (closed) subspace, V, of the Sobolev space $H^1(\Omega)$ (cf. Ciarlet [3]). For example, with Neumann boundary conditions, $V = H^1(\Omega)$, whereas for Dirichlet boundary conditions we have $V = H_0^1(\Omega) := \{v \in H^1(\Omega) : v|_{\partial\Omega} = 0\}$. Let $a(\cdot,\cdot)$ denote a symmetric, bilinear form on $H^1(\Omega)$ that it is *continuous* on V,

$$a(v,w) \leq \alpha_0 \|v\|_{H^1(\Omega)} \|w\|_{H^1(\Omega)} \quad \forall v, w \in V, \tag{1.1}$$

and *coercive* on V,

$$\|v\|^2_{H^1(\Omega)} \leq \alpha_1 a(v,v) \quad \forall v \in V. \tag{1.2}$$

For example, with $V = H_0^1(\Omega)$, we might have

$$a(v,w) = \int_\Omega a(x) \nabla v \cdot \nabla w \, dx$$

where $a(x)$ denotes a function in $L^\infty(\Omega)$ such that $a(x) \geq \alpha > 0$ for all $x \in \Omega$. (However, the coefficient a need not be smooth.)

The variational boundary value problem that we wish to approximate takes the following form. Given a continuous linear functional, f, on V, find $u \in V$ such that

$$a(u,v) = f(v) \quad \forall v \in V. \tag{1.3}$$

In view of the assumptions (1.1) and (1.2), this problem has a unique solution (cf. Ciarlet [3]). We approximate this problem via the Galerkin method in the usual way. Let V_N be a subspace of V of dimension N, and let $u_N \in V_N$ be defined by

$$a(u_N,v) = f(v) \quad \forall v \in V_N. \tag{1.4}$$

Existence and uniqueness of the solution, u_N, again follows from the assumptions (1.1) and (1.2) (cf. Ciarlet [3]). However, explicit calculation of the solution frequently requires a constructive approach involving the conversion of the variational equation (1.4) into a matrix equation utilizing a particular basis for V_N. Specifically, suppose that $\{\psi_i : i = 1,\cdots,N\}$ is a given basis for V_N, and define a matrix, $\mathbf{A}$, and a vector, $\mathbf{F}$, via

$$\mathbf{A}_{ij} := a(\psi_i,\psi_j) \quad \text{and} \quad \mathbf{F}_i := f(\psi_i) \quad \forall i,j = 1,\cdots,N.$$

Then (1.4) is equivalent to solving

$$\mathbf{AX} = \mathbf{F}$$

where $u_N = \sum_{i=1}^N x_i\psi_i$ and $\mathbf{X} = (x_i)$. We now give conditions on V_N and the basis $\{\psi_i : i = 1,\cdots,N\}$ that will be used to guarantee that the condition number of $\mathbf{A}$ is well behaved.

To begin with, we suppose that associated with the space V_N is a *subdivision* of Ω, by which we mean a collection, $\mathcal{T}_N$, of nonoverlapping, nonempty open subsets, T, of $\mathbb{R}^n$, such that

$$\overline{\Omega} = \bigcup_{T\in\mathcal{T}_N} \overline{T}.$$

We suppose that $\mathcal{T}_N$ contains at most $\alpha_2^{n/2}N$ members, with α_2 a fixed constant. We assume that there are constants α_3 and α_4 such that the following *inverse estimates* hold:

$$\alpha_3^{-1}\|v\|^2_{H^1(T)} \le h_T^{n-2}\|v\|^2_{L^\infty(T)} \le \alpha_4\|v\|^2_{L^{2n/(n-2)}(T)} \quad \forall T \in \mathcal{T}_N,\ v \in V_N, \tag{1.5}$$

where h_T denotes the diameter of T. In the special case of two dimensions ($n = 2$), the latter inequality is a tautology, and we supplement it with the following assumption:

$$\|v\|_{L^\infty(T)} \le \sqrt{\alpha_4}\, h_T^{-2/p}\|v\|_{L^p(T)} \quad \forall T \in \mathcal{T}_N,\ v \in V_N,\ 1 \le p \le \infty. \tag{1.6}$$

For piecewise polynomials, these properties are standard (cf. Ciarlet [3] and the examples in the next section).

Concerning the basis, $\{\psi_i : i = 1,\cdots,N\}$, of V_N, we make the following assumptions. First of all, we assume that it is a *local* basis:

$$\max_{1\le i\le N} \text{cardinality}\,\{T \in \mathcal{T}_N : \text{supp}(\psi_i) \cap T \ne \emptyset\} \le \alpha_5. \tag{1.7}$$

Finally, we come to the most important assumption, concerning the scaling of the basis. We assume that there are finite constants α_6, α_7 such that for all $T \in \mathcal{T}_N$

$$\alpha_6^{-1} h_T^{n-2} \|v\|^2_{L^\infty(T)} \leq \sum_{\text{supp}(\psi_i) \cap T \neq \emptyset} x_i^2 \leq \alpha_7 h_T^{n-2} \|v\|^2_{L^\infty(T)} \tag{1.8}$$

where $v = \sum_{i=1}^N x_i \psi_i$ and (x_i) is arbitrary.

In order to derive our results, we must make a slight regularity assumption on the domain, Ω, namely that it be Lipschitz in the sense of Stein [10]. (For example, in two dimensions, this rules out "slit" domains.) In this case, there is a continuous extension operator $H^1(\Omega) \to H^1(\mathbb{R}^n)$. Therefore, the Sobolev imbedding

$$H^1(\Omega) \subset L^p(\Omega)$$

follows from its validity (cf. Stein [10]) for the case $\Omega = \mathbb{R}^n$ (or $\Omega =$ a sufficiently large domain with smooth boundary). When $n \geq 3$, we thus have Sobolev's inequality,

$$\|v\|_{L^{2n/n-2}(\Omega)} \leq C_S \|v\|_{H^1(\Omega)} \quad \forall v \in H^1(\Omega).$$

In two dimensions ($n = 2$), since we assume that Ω is bounded, the Sobolev imbedding $H^1(\Omega) \subset L^p(\Omega)$ holds for all $p < \infty$. Moreover, it has a norm, $\sigma(p)$, that is bounded by a constant times the norm of the Sobolev imbedding $H_0^1(B) \subset L^p(B)$ for a sufficiently large ball, B, namely, $\sigma(p) \leq C_S \sqrt{p}$ (cf. Gilbarg and Trudinger [4], especially the proof of Thm. 7.15). Thus for $n = 2$ we have the following Sobolev inequality:

$$\|v\|_{L^p(\Omega)} \leq C_S \sqrt{p}\, \|v\|_{H^1(\Omega)} \quad \forall v \in H^1(\Omega),\ p < \infty.$$

2. Examples satisfying the assumptions. Suppose that Ω has a simplicial (e.g., polygonal if $n = 2$) boundary, $\partial\Omega$. We consider the case when $\mathcal{T}_N$ is a triangulation of Ω, but we make no assumption concerning the relative sizes of simplices in the triangulation.

DEFINITION 2.1. A family, $\mathcal{F}$, of triangulations $\{\mathcal{T}_N : N \in \mathcal{N}\}$ is said to be *nondegenerate* if there is a constant $\rho > 0$ such that for all $N \in \mathcal{N}$ and for all $T \in \mathcal{T}_N$ there is a ball of radius $\rho \operatorname{diam}(T)$ contained in T, where $\operatorname{diam}(T)$ denotes the diameter of T.

In practice we have only a finite number of (finite) triangulations to deal with, and any finite family is nondegenerate. However, all constants discussed below will be bounded in terms of the parameter, ρ, in the definition above. The "chunkiness" of a triangulation can be defined as the largest possible such ρ for a given triangulation.

Example 2.1. Let V_N denote the space of C^0 piecewise polynomials of degree k on the mesh $\mathcal{T}_N$ that are contained in the subspace V. We denote by $\{\phi_i : i = 1, \cdots, N\}$ the standard Lagrangian nodal basis for V_N consisting of functions that equal one at precisely one nodal point in the triangulation (cf. Ciarlet [3]). We also introduce a scaled basis that is of interest in three (and higher) dimensions. For each node (i.e., for each basis function) we may introduce a notion of local mesh size near that node, say h_i. This can be defined as the average diameter of all elements in $\mathcal{T}_N$ whose closure contains that node. (Note that the nondegeneracy assumption on the mesh implies that all such elements will be of comparable size, i.e., a nondegenerate mesh is locally *quasi-uniform*, in two or higher dimensions. This is because neighboring elements are

all connected to each other via a sequence of elements with common faces.) Define a new basis $\{\psi_i : i = 1, \cdots, N\}$ by

$$\psi_i := h_i^{(2-n)/2}\phi_i$$

where n is the dimension of Ω. (Note that this basis does not differ from the original one if $n = 2$.)

Clearly (1.7) holds for the Lagrange space, e.g., in two dimensions, α_5 is a bound on the number of triangles that can meet at a vertex, and this can be bounded in terms of ρ. Standard homogeneity arguments show that (1.5) holds, and our choice of scaling similarly yields (1.8), with α_6 depending only on ρ and k, since $C_\rho^{-1}h_i \leq h_T \leq C_\rho h_i$ for all $T \cap \text{supp}\psi_i \neq \emptyset$ (and $\alpha_7 = 1$). The inverse estimate (1.6) in the two-dimensional case also follows by homogeneity for the case $p = 1$, and then Hölder's inequality implies it holds for the remaining cases with the constant independent of p.

Example 2.2. A commonly used element in two dimensions is the Hermite family (cf. Ciarlet [3]). To obtain (1.8) for Hermite elements in two dimensions, one chooses the basis functions corresponding to derivative nodes to have the corresponding derivative of order $\mathcal{O}(h_i^{-1})$, with the remaining basis functions scaled as in the Lagrangian case. The other assumptions for this element follow as in the Lagrangian case.

Example 2.3. At this point it would be useful to know that Definition 2.1 does not exclude radical mesh refinements, as we have already observed that it does imply local quasi-uniformity. Let Ω_0 denote the square of side 1 centered at the origin, i.e.,

$$\Omega_0 = \left\{(x, y) \in \mathbb{R}^2 : |x| < \frac{1}{2}, |y| < \frac{1}{2}\right\}.$$

Let $\mathcal{T}_{N_0}$ denote the triangulation of Ω_0 generated by its diagonals and the two axes, i.e., consisting of eight isosceles, right triangles (each having two sides of length 1/2). We subdivide to construct $\mathcal{T}_{N_1}$ as shown in Fig. 1 by adding the edges of the square, Ω_1, of side 1/2 centered at the origin together with eight more edges running parallel with the diagonals. We obtain 24 similar triangles in this way. Also note that $\mathcal{T}_{N_1}$ restricted to the square Ω_1 is a triangulation similar to $\mathcal{T}_{N_0}$. Thus we may repeat the process above to this part of the domain alone to define a triangulation $\mathcal{T}_{N_2}$ consisting of isosceles, right triangles. Continuing in this way, we obtain a sequence of triangulations, $\mathcal{T}_{N_i}$, consisting of similar triangles. (Fig. 1 shows the cases $i = 0, 1, 2, 3$.) The ratio of largest to smallest side length is 2^i yet only $16i + 8$ triangles are used. (There are $8i$ interior vertices in $\mathcal{T}_{N_i}$, so $N_i = 8i$ in the case of Lagrange piecewise linear approximation of the Dirichlet problem.) Such a geometric refinement is far more severe than is often used to resolve boundary or interface singularities, but it shows that the assumption of *nondegeneracy* in Definition 2.1 need not restrict mesh refinement.

We now give bounds on the condition number of the matrix $\mathbf{A} := (a(\psi_i, \psi_j))$, where $\{\psi_i : i = 1, \cdots, N\}$ is the (scaled) basis for V_N specified by our assumptions (and defined explicitly in the previous examples). Applications of these results to convergence rates for the conjugate-method for solving $\mathbf{AX} = \mathbf{F}$ will be given in §5.

3. The general case $\mathbf{n} \geq 3$.

THEOREM 3.1. *Suppose that the subspace V_N satisfies assumption* (1.5) *and that the basis $\{\psi_i : i = 1, \cdots, N\}$ satisfies* (1.7) *and* (1.8). *Let* $\mathbf{A}$ *denote the matrix corresponding to the inner product $a(\cdot, \cdot)$, i.e., $\mathbf{A}_{ij} = a(\psi_i, \psi_j)$. Then the l_2-condition number, $\kappa_2(\mathbf{A})$, of* $\mathbf{A}$ *is bounded by*

$$\kappa_2(\mathbf{A}) \leq CN^{2/n}$$

where $C = C_S^2 \prod_{i=0}^{7} \alpha_i$ depends only on the constants, α_i, in the assumptions and the constant in Sobolev's inequality.

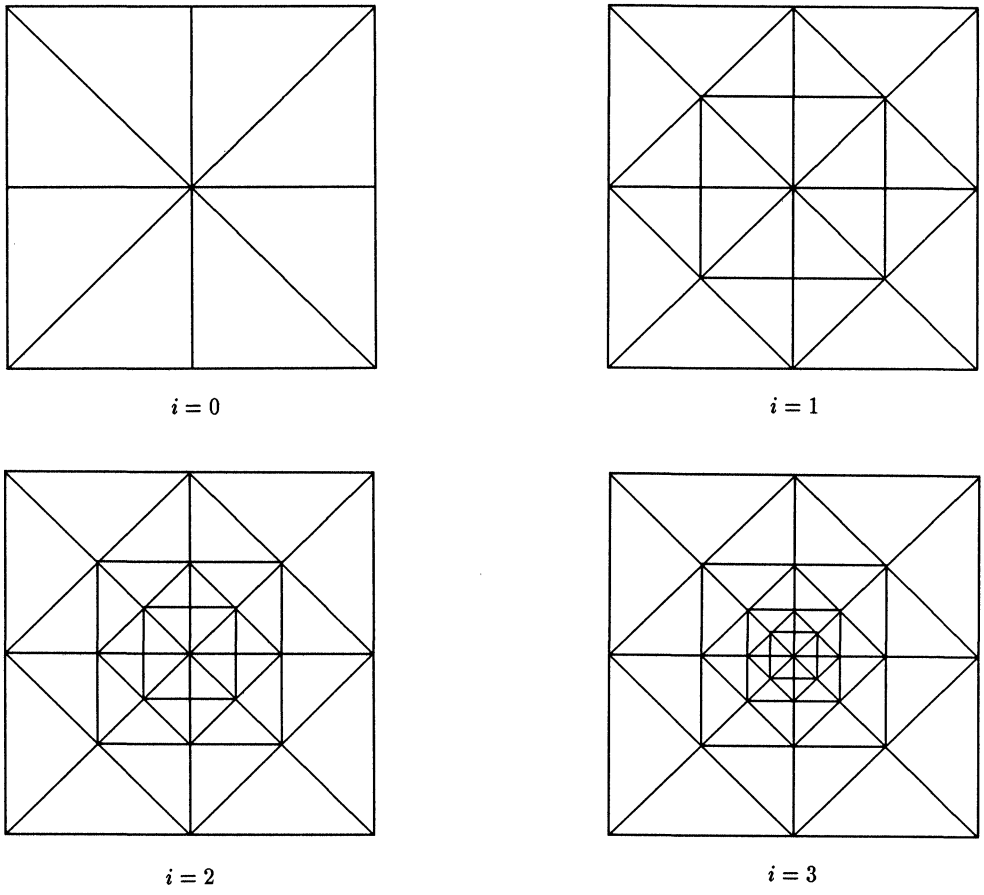

FIG. 1.

Proof. First note that if we set $v = \sum_i x_i \psi_i$ then

$$a(v,v) = \mathbf{X}^t \mathbf{A} \mathbf{X} \tag{3.1}$$

where $\mathbf{X} = (x_i)$, because $a(\cdot,\cdot)$ is bilinear. Observe that

$$\begin{aligned} a(v,v) &\le \alpha_0 \|v\|^2_{H^1(\Omega)} && (1.1)\\ &= \alpha_0 \sum_{T \in \mathcal{T}_N} \|v\|^2_{H^1(T)} && (\mathcal{T}_N \text{ is a subdivision})\\ &\le \alpha_0 \alpha_3 \sum_{T \in \mathcal{T}_N} h_T^{n-2} \|v\|^2_{L^\infty(T)} && (1.5)\\ &\le \alpha_0 \alpha_3 \alpha_6 \sum_{T \in \mathcal{T}_N} \sum_{\text{supp}(\psi_i) \cap T \neq \emptyset} x_i^2 && (1.8)\\ &\le \alpha_0 \alpha_3 \alpha_6 \alpha_5 \, \mathbf{X}^t \mathbf{X} && (1.7). \end{aligned}$$

Here h_T denotes the diameter of T. A complementary inequality can be derived as follows:

$$\mathbf{X}^t \mathbf{X} \le \sum_{T \in \mathcal{T}_N} \sum_{\text{supp}(\psi_i) \cap T \neq \emptyset} x_i^2$$

$$\leq \alpha_7 \sum_{T\in\mathcal{T}_N} h_T^{n-2}\|v\|^2_{L^\infty(T)} \tag{1.8}$$

$$\leq \alpha_7\alpha_4 \sum_{T\in\mathcal{T}_N} \|v\|^2_{L^{2n/n-2}(T)} \tag{1.5}$$

$$\leq \alpha_7\alpha_4 \Big(\sum_{T\in\mathcal{T}_N} 1\Big)^{2/n} \|v\|^2_{L^{2n/n-2}(\Omega)} \qquad \text{(Hölder's } \leq\text{)}$$

$$\leq \alpha_7\alpha_4\alpha_2 N^{2/n}\|v\|^2_{L^{2n/n-2}(\Omega)}$$

$$\leq C_S^2\alpha_7\alpha_4\alpha_2 N^{2/n}\|v\|^2_{H^1(\Omega)} \qquad \text{(Sobolev's } \leq\text{)}$$

$$\leq C_S^2\alpha_7\alpha_4\alpha_2\alpha_1 N^{2/n} a(v,v) \tag{1.2}.$$

Using these estimates we show

$$C_1 N^{-2/n}\mathbf{X}^t\mathbf{X} \leq \mathbf{X}^t\mathbf{A}\mathbf{X} \leq C_2\mathbf{X}^t\mathbf{X}$$

where $C_1 = (C_S^2\alpha_1\alpha_2\alpha_4\alpha_7)^{-1} > 0$ and $C_2 = \alpha_0\alpha_3\alpha_5\alpha_6 < \infty$ depend only on the α_i and the constant in Sobolev's inequality. This proves that

$$C_1 N^{-2/n} \leq \lambda_{\min}(\mathbf{A}) \quad \text{and} \quad \lambda_{\max}(\mathbf{A}) \leq C_2$$

where $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ denote, respectively, the smallest and largest eigenvalues of $\mathbf{A}$. Recall (cf. Isaacson and Keller [5]) that the l_2-condition number, $\kappa_2(\mathbf{A})$, of $\mathbf{A}$ satisfies

$$\kappa_2(\mathbf{A}) = \lambda_{\max}(\mathbf{A})/\lambda_{\min}(\mathbf{A}).$$

Thus the previous two estimates yield the stated result.

As a corollary to this result, we have the following in view of Example 2.1.

THEOREM 3.2. *Suppose that the mesh family, $\{\mathcal{T}_N : N \in \mathcal{N}\}$, is nondegenerate, and let $\mathbf{A}$ denote the matrix corresponding to the inner product $a(\cdot,\cdot)$, i.e., $\mathbf{A}_{ij} = a(\psi_i,\psi_j)$ where $\{\psi_i : i = 1,\cdots,N\}$ is the scaled Lagrange basis for V_N. Then the l_2-condition number, $\kappa_2(\mathbf{A})$, of $\mathbf{A}$ is bounded by*

$$\kappa_2(\mathbf{A}) \leq CN^{2/n}$$

where C depends only on the nondegeneracy constant ρ (cf. Definition 2.1) and the degree k of the piecewise polynomials in V_N.

4. The special case n = 2. A similar result can be given in two dimensions ($n = 2$) as follows.

THEOREM 4.1. *Suppose the subspace V_N satisfies assumptions (1.5) and (1.6), and suppose that the basis $\{\psi_i : i = 1,\cdots,N\}$ satisfies (1.7) and (1.8). Let $\mathbf{A}$ denote the matrix corresponding to the inner product $a(\cdot,\cdot)$, i.e., $\mathbf{A}_{ij} = a(\psi_i,\psi_j)$. Then the l_2-condition number, $\kappa_2(\mathbf{A})$, of $\mathbf{A}$ is bounded by*

$$\kappa_2(\mathbf{A}) \leq CN\,(1 + |\log(N\,h_{\min}(N)^2)|)$$

where $h_{\min} = \min\{h_T : T \in \mathcal{T}_N\}$ and C depends only on the constants α_i in the assumptions and the constant in Sobolev's inequality.

Proof. As in the proof of Theorem 3.1, it is sufficient to prove that

$$C_1\,(N\,(1 + |\log(N\,h_{\min}(N)^2)|))^{-1}\,\mathbf{X}^t\mathbf{X} \leq \mathbf{X}^t\mathbf{A}\mathbf{X} \leq C_2\mathbf{X}^t\mathbf{X}$$

where $C_1 > 0$ and $C_2 < \infty$ depend only on the α_i and C_S. The proof of these inequalities is quite similar to the case $n \geq 3$. For $v \in V_N$, we again write $v = \sum_i x_i \psi_i$ and recall from (3.1) that $a(v,v) = \mathbf{X}^t \mathbf{A} \mathbf{X}$. Then

$$\begin{aligned} a(v,v) &\leq \alpha_0 \|v\|^2_{H^1(\Omega)} && (1.1)\\ &= \alpha_0 \sum_{T \in \mathcal{T}_N} \|v\|^2_{H^1(T)} && (\mathcal{T}_N \text{ is a subdivision})\\ &\leq \alpha_0 \alpha_3 \sum_{T \in \mathcal{T}_N} \|v\|^2_{L^\infty(T)} && (1.5)\\ &\leq \alpha_0 \alpha_3 \alpha_6 \sum_{T \in \mathcal{T}_N} \sum_{\text{supp}(\psi_i) \cap T \neq \emptyset} x_i^2 && (1.8)\\ &\leq \alpha_0 \alpha_3 \alpha_6 \alpha_5 \, \mathbf{X}^t \mathbf{X} && (1.7). \end{aligned}$$

For the remaining inequality, we have (for $p > 2$)

$$\begin{aligned} \mathbf{X}^t \mathbf{X} &\leq \sum_{T \in \mathcal{T}_N} \sum_{\text{supp}(\psi_i) \cap T \neq \emptyset} x_i^2\\ &\leq \alpha_7 \sum_{T \in \mathcal{T}_N} \|v\|^2_{L^\infty(T)} && (1.8)\\ &\leq \alpha_7 \alpha_4 \sum_{T \in \mathcal{T}_N} h_T^{-4/p} \|v\|^2_{L^p(T)} && (1.6)\\ &\leq \alpha_7 \alpha_4 \Big(\sum_{T \in \mathcal{T}_N} h_T^{-4/(p-2)} \Big)^{(p-2)/p} \|v\|^2_{L^p(\Omega)} && (\text{Hölder's } \leq)\\ &\leq C_S^2 \alpha_7 \alpha_4 \Big(\sum_{T \in \mathcal{T}_N} h_T^{-4/(p-2)} \Big)^{(p-2)/p} p \, \|v\|^2_{H^1(\Omega)} && (\text{Sobolev's } \leq)\\ &\leq C_S^2 \alpha_7 \alpha_4 \alpha_1 \Big(\sum_{T \in \mathcal{T}_N} h_T^{-4/(p-2)} \Big)^{(p-2)/p} p \, a(v,v) && (1.2). \end{aligned}$$

A crude estimate yields

$$\begin{aligned} \Big(\sum_{T \in \mathcal{T}_N} h_T^{-4/(p-2)} \Big)^{(p-2)/p} &\leq h_{\min}(N)^{-4/p} \, (\alpha_2 N)^{(p-2)/p}\\ &= \alpha_2^{1-2/p} \, (N \, h_{\min}(N)^2)^{-2/p} \, N. \end{aligned}$$

Thus the estimate above can be simplified to

$$\mathbf{X}^t \mathbf{X} \leq C_S^2 \alpha_7 \alpha_4 \alpha_2^{1-2/p} \Big(p \, (N \, h_{\min}(N)^2)^{-2/p} \Big) N a(v,v).$$

Choosing $p = \max\{2, |\log (N \, h_{\min}(N)^2)|\}$ in this estimate yields the stated result.

As a corollary, we have the following theorem.

THEOREM 4.2. *Suppose that the mesh family, $\{\mathcal{T}_N : N \in \mathcal{N}\}$, is nondegenerate, and let $h_{\min}(N)$ denote the diameter of the smallest triangle in $\mathcal{T}_N$. Let $\mathbf{A}$ denote the matrix corresponding to the inner product $a(\cdot,\cdot)$, i.e., $\mathbf{A} := (a(\psi_i, \psi_j))$ where*

$\{\psi_i : i = 1, \cdots, N\}$ is either the standard Lagrange basis or the scaled Hermite basis. Then the l_2-condition number, $\kappa_2(\mathbf{A})$, of $\mathbf{A}$ is bounded by

$$\kappa_2(\mathbf{A}) \leq CN(1 + |\log(N h_{\min}(N)^2)|)$$

where C depends only on the nondegeneracy constant ρ (cf. Definition 2.1) and the degree k of the piecewise polynomials in V_N.

REMARK 4.3. The result above predicts the correct condition number, $\mathcal{O}(h^{-2})$, in the case of a regular mesh of size h, since $N = \mathcal{O}(h^{-2}) = \mathcal{O}(h_{\min}^{-2})$ in this case. Moreover, if we define $h_{\max}$ to be the diameter of the largest triangle in the case of a general mesh, then $N \geq Ch_{\max}^{-2}$ because

$$\text{measure}(\Omega) = \sum_{T \in \mathcal{T}_N} \text{measure}(T) \leq C_\rho N h_{\max}^2.$$

Thus we conclude that the condition number of $\mathbf{A}$ can be bounded by

$$\kappa_2(\mathbf{A}) \leq CN(1 + |\log(h_{\max}/h_{\min})|).$$

For particular mesh subdivisions, a more precise bound could be attempted for the key term

$$\left(\sum_{T \in \mathcal{T}_N} h_T^{-4/(p-2)} \right)^{(p-2)/p}$$

in the proof of Theorem 4.1. However, for the special mesh introduced in Example 2.3, we can see that the estimate of Theorem 4.2 is sharp, as follows.

Let $\Omega = \Omega_0$, and let N denote N_K for a given K. Suppose that

$$a(u, v) := \int_\Omega \nabla u \cdot \nabla v \, dx.$$

Let V_N be the set of piecewise linear functions vanishing on the boundary of Ω. By choosing $v \in V_N$ equal to one at the origin and zero elsewhere, we find that

$$\lambda_{\max} \geq a(v, v)/\mathbf{X}^t\mathbf{X} = a(v, v) = 4.$$

On the other hand, define

$$u^K(x) := \begin{cases} \left| \log 2|x| \right| & \text{if} \quad 2^{-K-1} \leq |x| \leq 1/2 \\ 0 & \text{if} \quad |x| \geq 1/2 \\ K \log 2 & \text{if} \quad |x| \leq 2^{-K-1}, \end{cases}$$

and let v be the interpolant of u^K. Then it is not hard to see that, on all of the triangles, $T \in \mathcal{T}_N$, except for the eight smallest and eight largest, the integral of $|\nabla v|^2$ is a constant independent of T, namely,

$$\int_T |\nabla v|^2 \, dx = \frac{5}{8}(\log 2)^2.$$

Further, the integral of $|\nabla v|^2$ can be computed easily on the remaining triangles as well, yielding

$$a(v, v) = 2(\log 2)^2 + (16K - 8)\frac{5}{8}(\log 2)^2 = (10K - 3)(\log 2)^2.$$

Similarly, with x_i denoting the nodal values of v,

$$\sum x_i^2 = 4\sum_{j=1}^{K}\left|\log 2^{-j}\right|^2 + 4\sum_{j=1}^{K}\left|\log 2^{-j}\sqrt{2}\right|^2 = 5(\log 2)^2\sum_{j=1}^{K} j^2.$$

Therefore

$$\lambda_{\min} \leq a(v,v)/\mathbf{X}^t\mathbf{X} \leq 3K^{-2}$$

which proves that $\kappa_2(\mathbf{A}) \geq (4/3)K^2$.

Recall that $N \sim K$ for this triangulation, and that $h_{\min} \sim 2^{-K}$ as well. Thus, the bound in Theorem 4.2 reads $\kappa_2(\mathbf{A}) \leq CN(1+|\log N2^{-2K}|) \leq CNK = CK^2$. Therefore the bound in Theorem 4.2 is sharp in this case.

5. Applications to the conjugate-gradient method. The conjugate-gradient method for solving a linear system of the form $\mathbf{AX} = \mathbf{F}$ is an iterative method whose convergence properties can be estimated in terms of the condition number of $\mathbf{A}$ (cf. Luenberger [6]). Specifically, define

$$\|\mathbf{X}\|_A := (\mathbf{X}^t\mathbf{A}\mathbf{X})^{1/2},$$

and let $\mathbf{X}^{(k)}$ denote the sequence of vectors generated by the conjugate-gradient method starting with $\mathbf{X}^{(0)} = \mathbf{0}$. Then

$$\|\mathbf{X} - \mathbf{X}^{(k)}\|_A \leq C\exp\left(-2k/\sqrt{\kappa_2(\mathbf{A})}\right)\|\mathbf{X}\|_A$$

where $\mathbf{X}$ denotes the solution to $\mathbf{AX} = \mathbf{F}$. This can be easily interpreted in terms of norms on V. Define an *energy norm* on V by

$$\|v\|_a := \sqrt{a(v,v)}. \tag{5.1}$$

Then (3.1) implies that, for $v = \sum_i y_i\psi_i$,

$$\|v\|_a = \|\mathbf{Y}\|_A$$

where $\mathbf{Y} = (y_i)$. Let $u_N = \sum_i x_i\psi_i$ and $u_N^{(k)} = \sum_i x_i^{(k)}\psi_i$, where $(x_i) = \mathbf{X}$ and $(x_i^{(k)}) = \mathbf{X}^{(k)}$. Then the above estimate may be written

$$\|u_N - u_N^{(k)}\|_a \leq C\exp\left(-2k/\sqrt{\kappa_2(\mathbf{A})}\right)\|u_N\|_a.$$

(Recall that the continuity and coercivity assumptions (1.1) and (1.2) imply that the energy norm, $\|\cdot\|_a$, is equivalent to the norm on $H^1(\Omega)$.) This estimate says that to reduce the relative error $\|u_N - u_N^{(k)}\|_a/\|u_N\|_a$ to $\mathcal{O}(\epsilon)$ requires at most $k = \mathcal{O}\left(\sqrt{\kappa_2(\mathbf{A})}\,|\log\epsilon|\right)$ iterations. Suppose that we only require $\epsilon = \mathcal{O}(N^{-q})$ for some $q < \infty$. In $n \geq 3$ dimensions, the above estimate says that this order of accuracy will be achieved after only $\mathcal{O}(N^{1/n}\log N)$ iterations. In two dimensions the above estimate becomes slightly more complicated. In typical applications, even with very severe refinements, we have $h_{\min} = \mathcal{O}(h_{\max}^p) = \mathcal{O}(N^{-p/2})$ for some $p < \infty$, as we shall

now assume. In this case, the estimate above says that $\mathcal{O}(\epsilon)$ accuracy will be achieved after only $\mathcal{O}\left(N^{1/2}(\log N)^2\right)$ iterations.

Each conjugate-gradient iteration requires $\mathcal{O}(N)$ operations. Thus the final work estimates for the conjugate-gradient method on refined meshes as described previously would be as shown in Table 1. For the sake of reference, we give the work estimates for banded Gaussian factorization (and the solution process with precomputed factors) assuming a standard (lexicographical) ordering of the basis functions, on a regular mesh, and Gaussian factorization and solution using "nested dissection," as described in Table 7.14 of Axelsson and Barker [1]. The work estimates for nested dissection for $n \geq 3$ follow from arguments similar to those used in the work of Rose and Whitten [9].

TABLE 1. *Order of work estimate (number of operations) as a function of the number, N, of unknowns in the system to achieve an accuracy of* $\mathcal{O}(N^{-r})$. "CG" *refers to the conjugate-gradient algorithm,* "GE" *refers to solution using Gaussian factorization, and "solve" refers to forward- and back-solution using precomputed factors.*

Dimension	$n = 2$	$n = 3$	$n \geq 3$
CG/nondegenerate mesh	$N^{3/2}(\log N)^2$	$N^{4/3} \log N$	$N^{1+1/n} \log N$
GE/lexicographical order	N^2	$N^{7/3}$	$N^{3-2/n}$
solve/lexicographical order	$N^{3/2}$	$N^{5/3}$	$N^{2-1/n}$
GE/nested dissection	$N^{3/2}$	N^2	$N^{3-3/n}$
solve/nested dissection	$N \log N$	$N^{4/3}$	$N^{2-2/n}$

Of course, such work estimates are not competitive with an optimal-order iterative procedure such as a multilevel method, in which an accurate solution is achieved in $\mathcal{O}(N)$ operations. However, such methods typically involve some alternate technique of solution on a "coarse grid." For the latter case we must consider more conventional methods such as conjugate-gradients or Gaussian elimination.

Acknowledgments. We thank Ricardo Duran and Lars Wahlbin for helpful discussions.

REFERENCES

[1] O. AXELSSON AND V. A. BARKER, *Finite Element Solution of Boundary Value Problems*, Academic Press, Orlando, 1984.

[2] I. BABUŠKA, O. C. ZIENKIEWICZ, J. GAGO, E. R. DE A. OLIVEIRA, EDS., *Accuracy Estimates and Adaptive Refinements in Finite Element Computation*, John Wiley, New York, 1986.

[3] P. G. CIARLET, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.

[4] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer-Verlag, Berlin, 1983.

[5] E. ISAACSON AND H. B. KELLER, *Analysis of Numerical Methods*, John Wiley, New York, 1966.

[6] D. G. LUENBERGER, *Introduction to Linear and Nonlinear Programming*, Addison-Wesley, Reading, MA, 1973.

[7] C. MCCARTHY AND G. STRANG, *Optimal conditioning of matrices*, SIAM J. Numer. Anal., 10 (1973), pp. 370–388.

[8] J. R. RICE, *Is the aspect ratio significant for finite element problems?*, Preprint CSD–TR 535, Computer Science Dept., Purdue University, W. Lafayette, IN, 1985.

[9] D. R. ROSE AND G. F. WHITTEN, *A recursive analysis of dissection strategies*, in *Sparse Matrix Computations*, J. R. Bunch and D. R. Rose, eds., Academic Press, New York, 1976, pp. 59–84.

[10] E. M. STEIN, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, 1970.