SIAM J. NUMER. ANAL.
Vol. 24, No. 2, April 1987

008

DEFLATION OF CONJUGATE GRADIENTS WITH APPLICATIONS TO BOUNDARY VALUE PROBLEMS*

R. A. NICOLAIDES†

Abstract. A method for improving the convergence of the standard conjugate gradient method is given. This method involves the use of auxiliary subspaces. It is shown how such subspaces may be constructed for boundary value problems and an analysis of convergence for second order problems is presented.

Key words. conjugate gradients, preconditioners, elliptic equations, iterative methods

1. Introduction.

The idea of preconditioning of the conjugate gradient method goes back at least as far as [5]. These authors used various nonstationary iterative methods as preconditioners, a line of development later taken up in [6] and many other references. Meanwhile, [16] introduced the idea of approximate factorization, which was developed in the context of first order iterative methods in [13], [3] and elsewhere. Later, [8] used a slightly different approximate factorization, this time in the conjugate gradient setting. Since then, numerous applications and extensions have been made including cases involving nonsymmetric and even complex matrices. References [1], [2], [4] are recent references for preconditioned conjugate gradients.

In this paper, another way of improving the convergence of conjugate gradients is used. It can be used alone or in conjunction with preconditioners. Used alone, it is at least as efficient as the standard preconditioners on model problems. Used with preconditioning it appears from numerical experiments to give a method considerably better than either used separately—it seems that the approaches are in some sense complementary. However, the main goal of this paper is an analysis of the unpreconditioned case for second order elliptic problems.

The next section summarizes the required facts about conjugate gradients. The following sections deal respectively with the definition of the deflation method, some rigorous analysis for boundary value problems in various settings, questions of practical application and finally some closing remarks are given.

2. Conjugate gradients.

The relevant aspects of the conjugate gradient theory we require are given in this section. We omit the proofs since they are easily obtained by standard methods. It is most convenient to use the second order form of the algorithm [5], and this will be done throughout. To define it, let M be a symmetric positive semidefinite $n\times n$ matrix with range $R(M)$, null space $N(M)$, rank m and pseudoinverse M^+. Let F denote any $n\times(n-m)$-dimensional matrix whose columns are a basis for $N(M)$. Throughout the paper we shall use the notation $\mathscr{C}_k(u, v)$ for the convex combination $\omega_k u+(1-\omega_k)v$, where in all cases $\omega_0=1$. Consider (residual) sequences $\{r_k\}\in\mathbb{R}^n$ generated by

$$r_0\in\mathbb{R}^n,\qquad F^t r_0=0,\tag{2.1}$$

$$r_{k+1}=\mathscr{C}_k(r_k, r_{k-1})-\mu_k M r_k,\qquad k=0,1,\cdots\tag{2.2}$$

where the parameter sequences $\{\mu_k\}$ and $\{\omega_k\}$ are obtained by requiring that r_{k+1} is orthogonal to both r_k and r_{k-1} in the ordinary $\mathbb{R}^n$ sense and $(r_1, r_0)=0$ if $k=0$. The second of the conditions (2.1) ensures that these parameters exist, and it follows that

*Received by the editors May 13, 1985; accepted for publication April 17, 1986. This work was supported by the Air Force Office of Scientific Research, under grant AFOSR-84-0137.

† Department of Mathematics, Carnegie-Mellon University, Pittsburgh, Pennsylvania 15213.

in fact r_{k+1} is orthogonal to r_j, $j=0, 1, \cdots, k$ as is simply proved by induction using $M^t = M$. Also, it is clear using (2.1) and induction that $\{r_k\}$ is orthogonal to $N(M)$; it can be easily shown that for some $l \leqq m$, $r_l = 0$ and (2.2) then terminates.

Bounds on r_k $k \leqq l$ are given by the standard result [7]

$$(r_k, M^+ r_k) \leqq C(k)(r_0, M^+ r_0), \qquad k = 0, 1, \cdots$$

where

$$C(k) = 4\{(1-\gamma)/(1+\gamma)\}^{2k},$$

with γ denoting the square root of the ratio of the smallest positive to greatest eigenvalue of M.

In the case where the equation

$$Mx = b$$

is to be solved, r_j corresponds to the residual $r_j = b - Mx_j$. The iteration for x then becomes

$$r_0 = b - Mx_0, \qquad F^t r_0 = 0, \tag{2.3}$$

$$x_{k+1} = \mathscr{C}_k(x_k, x_{k-1}) + \mu_k r_k, \qquad k = 0, 1 \cdots. \tag{2.3'}$$

Finding the parameters ω_k and μ_k can be easily accomplished by setting $\mu_k = \alpha_k \omega_k$. Then direct calculations give

$$\alpha_k = (r_k, r_k)/(r_k, Mr_k) \tag{2.4}$$

and the recursion for ω_k,

$$\omega_k^{-1} = 1 - \omega_{k-1}^{-1}(\alpha_k(r_k, r_k))/(\alpha_{k-1}(r_{k-1}, r_{k-1})) \tag{2.5}$$

where $\omega_0 = 1$. It is (2.3) which ensures that (2.4) is well defined.

It is usual to include some kind of preconditioning with the conjugate gradient method. In this paper we shall restrict attention to factorized preconditioners and incorporate them into the algorithm (2.2), with s_k denoting $T^{-1} r_k$ where M has the approximate factorization $M \approx TT^t = B$. We shall assume that the latter is nonsingular, which is the usual case, or has the columns of F as a basis for its null space, so that it will be invertible for data r_k. Usually T will be some block diagonal matrix or incomplete Cholesky factor of M. Then the generalized algorithm is

$$x_{k+1} = \mathscr{C}_k(x_k, x_{k-1}) + \alpha_k \omega_k B^{-1} r_k, \tag{2.6}$$

from which it follows that

$$s_{k+1} = \mathscr{C}_k(s_k, s_{k-1}) - \alpha_k \omega_k C s_k \tag{2.7}$$

where $C = T^{-1} M T^{-t}$, which is clearly positive semidefinite. Equation (2.7) is used to choose α_k and ω_k so that s_{k+1} is orthogonal to s_k and s_{k-1}, just as the same idea was used above to determine the parameters for the unpreconditioned case. It then follows that

$$\alpha_k = (s_k, s_k)/(s_k, Cs_k)$$

and that ω_k is given by a formula akin to (2.5) with s_k and s_{k-1}. A similar error estimate to the one already given holds with the eigenvalues of C, rather than M appearing.

3. Deflation of residual. Let the system to be solved be denoted by

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

where L is $n\times n$ symmetric positive definite. For this system denote residuals by $r_k=f-Lu_k$. The unpreconditioned conjugate gradient algorithm for (3.1) is

$$\begin{aligned}&u_0\in\mathbb{R}^n \text{ is arbitrary,}\\ &u_{k+1}=\mathscr{C}_k(u_k,u_{k-1})+\mu_k r_k, \qquad k=0,1,2\cdots.\end{aligned} \tag{3.2}$$

Clearly,

$$r_{k+1}=\mathscr{C}_k(r_k,r_{k-1})-\mu_k Lr_k$$

and the parameters are found as for (2.2), by forcing orthogonality of r_{k+1} with the two previous residuals. The convergence results stated in the previous section are thus immediately applicable.

We now modify (3.2) by "deflating" certain constituents from the residual. Let E denote a fixed $n\times m$ matrix whose columns are a basis for an m-dimensional subspace of $\mathbb{R}^n$ and consider iterations

$$\begin{aligned}&u_0 \text{ is arbitrary,}\\ &u_{k+1}=\mathscr{C}_k(u_k,u_{k-1})+\mu_k(r_k-Ec_k), \qquad k=0,1,2\cdots,\end{aligned} \tag{3.3}$$

in which $r_k=f-Lu_k$. It follows that

$$r_{k+1}=\mathscr{C}_k(r_k,r_{k-1})-\mu_k L(r_k-Ec_k). \tag{3.4}$$

We define c_k by requiring that it minimizes

$$(L(r_k-Ec_k),r_k-Ec_k) \tag{3.5}$$

leading to

$$E^tLEc_k=E^tLr_k, \tag{3.6}$$

which must be solved for c_k. Equation (3.4) now becomes

$$r_{k+1}=\mathscr{C}_k(r_k,r_{k-1})-\mu_k L(I-E\mathbb{L}^{-1}E^tL)r_k \tag{3.7}$$

where $\mathbb{L}=E^tLE$. The bracketed matrix, which is the projector onto span $(LE)^\perp$ along span $\{E\}$, (here and throughout span $\{\ \}$ denotes the column space of the argument) will be denoted by $I-P$. It is important to point out that $L(I-P)$ is symmetric and singular. We shall denote it by M. In view of the singularity of M we must add the additional condition to (3.4) that the initial residual r_0 is orthogonal to $N(M)=$ span (E), i.e., $E^tr_0=0$. It is shown below how to achieve this. Then from (3.7)

$$E^tr_{k+1}=\mathscr{C}_k(E^tr_k,E^tr_{k-1})=0 \tag{3.8}$$

by a simple induction. Thus, $E^tr_j=0$ for $j=0,1,2\cdots$, so that $\{r_j\perp N(M)\}$. To determine the parameters, we proceed as in § 2 and require that $\{r_k, r_{k-1}\perp r_{k+1}\}$. Then with $\mu_k=\alpha_k\omega_k$,

$$\alpha_k=(r_k,r_k)/(r_k,Mr_k),$$

which is well defined since r_k is orthogonal to $N(M)$, or alternatively because

$$\begin{aligned}(r_k,Mr_k)&=(r_k,L(I-P)r_k)\\ &=(r_k,L(r_k-Ec_k))=(r_k,Lr_k)-(c_k,E^tLr_k)\\ &=\min(r_k-Ec,L(r_k-Ec))>0.\end{aligned} \tag{3.9}$$

ω_k is determined by the recurrence (2.5) again with $\omega_0 = 1$. (3.9) gives the most efficient way to calculate (r_k, Mr_k), the second term being computable from (3.6) as an inner product on $\mathbb{R}^m$.

To get $E^t r_0 = 0$ pick v' arbitrarily, let $s' = f - Lv'$ and solve

$$(E^t LE)d = E^t s',$$

then set

$$v_0 = v' + Ed. \tag{3.10}$$

Clearly, with $r_0 = f - L(v' + Ed)$, $E^t r_0 = E^t f - E^t Lv' - E^t LEd = 0$. v_0 can be described equally well as a correction to v' obtained by solving

$$\min\,(L(e' - Ed), (e' - Ed)) \qquad (Le' = s').$$

The algorithm is now fully defined.

Convergence of the algorithm follows immediately from the identity

$$L(I - P) = L^{1/2}(I - L^{1/2}PL^{-1/2})L^{1/2},$$

which shows that $L(I - P)$ is positive semidefinite, since the bracketed matrix is an orthogonal projector and $L^{1/2}$ is positive definite.

The preconditioned case can be dealt with similarly. The best starting point is (2.7), which we shall change to

$$s_{k+1} = \mathscr{C}_k(s_k, s_{k-1}) - \alpha_k \omega_k T^{-1} L(I - P) T^{-t} s_k. \tag{3.11}$$

Here, $I - P$ has its previous meaning. Subject to the usual proviso concerning the null space, the parameters can be determined in the standard way. Let $t_k = T^{-t} s_k$. Then it follows that

$$\alpha_k = (s_k, s_k)/(t_k, Mt_k)$$

where $M = L(I - P)$. The usual recurrence formula for ω_k, analogous to (2.5), can easily be seen to hold. To find the iteration formula for u_k, multiply (3.11) by $L^{-1}T$ to get

$$u_{k+1} = \mathscr{C}_k(u_k, u_{k-1}) + \alpha_k \omega_k (T^{-t} T^{-1} r_k - Ec_k) \tag{3.12}$$

where c_k is the solution of the equation analogous to (3.6),

$$E^t LEc_k = E^t Lt_k.$$

In the same way that (r_k, Mr_k) is expressed as a difference in (3.9), (t_k, Mt_k) can be expressed as the difference

$$(t_k, Lt_k) - (c_k, E^t Lt_k).$$

Since span $\{T^t E\}$ is the null space of $T^{-1} M T^{-t}$, referring to (3.11) we must ensure that s_0 satisfies $E^t T s_0 = E^t r_0 = 0$. The method for doing this was given in (3.10), so the deflated algorithm is now fully defined and as in the undeflated preconditioned case, the extra operations consist of solving linear systems with coefficient matrices T or T^t.

A point which will be of use below is that the equation

$$T^{-1} L T^{-t} y = T^{-1} b, \tag{3.13}$$

when solved by deflated conjugate gradients with deflation matrix $F = T^t E$ and no preconditioning, gives a sequence of iterates which, as is easily verified, is identical with the sequence $\{T^t u_k\}$ where $\{u_k\}$ is the sequence generated by (3.12). Thus the preconditioned algorithm is essentially equivalent to an unpreconditioned one, with a different coefficient matrix and data.

The following lemma, which will be proved initially for the unpreconditioned case, will be used in the next section to estimate the convergence rate of deflated conjugate gradients for a model problem.

LEMMA 3.1. *Let λ_1 and λ_l denote the smallest positive and largest eigenvalue, respectively, of $L(I-P)$. Then the following characterization holds*:

$$\lambda_1 = \min\,(w, w)/(w, L^{-1}w),$$

$$\lambda_l = \max\,(w, w)/(w, L^{-1}w),$$

where the max *and* min *are taken over* span $\{E\}^{\perp}$.

Proof. For each (positive) eigenvalue λ of $L(I-P)$ we have, since the eigenvalues of a product are independent of the order of the factors,

$$(I-P)Lv = \lambda v. \tag{3.14}$$

Let $Q = I - E(E^tE)^{-1}E^t$ be the orthogonal projector onto span $\{E\}^{\perp}$. Multiply (3.14) by Q to get

$$QLv = \lambda Qv \tag{3.15}$$

and by $(I-Q)L$ to get

$$(I-Q)Lv = 0. \tag{3.16}$$

Equations (3.15)-(3.16) now give

$$Lv = \lambda Qv$$

so that

$$v = \lambda L^{-1}Qv$$

and

$$Qv = \lambda QL^{-1}Qv = \lambda QL^{-1}Q(Qv),$$

so that

$$QL^{-1}Qw = \lambda^{-1}w, \qquad w = Qv.$$

The result now follows by the usual Rayleigh quotient argument.

Although we shall not use the following extension of Lemma 3.1 to the preconditioned case in this paper, we record it here for completeness.

COROLLARY 3.1. *Let μ_1 and μ_l denote the smallest positive and greatest eigenvalue, respectively, of the matrix $T^{-1}MT^{-t}$ where $M = L(I-P)$. Then*

$$\mu_1 = \min\,(w, T^{-t}T^{-1}w)/(w, L^{-1}w),$$

$$\mu_l = \max\,(w, T^{-t}T^{-1}w)/(w, L^{-1}w)$$

where the min *and* max *are computed over* span $\{E\}^{\perp}$.

Proof. Based on the observation (3.12) and applying Lemma 3.1 with $E := F = T^tE$ the result follows directly.

A final point worth noticing is that the symmetry of $L(I-P)$ and L taken in conjunction with (3.6) shows that residual deflation for conjugate gradients is equivalent to a preconditioning with a singular matrix, namely $(I-P)$. By itself, however, this does not appear to be a particularly useful interpretation, because the rationale for using this matrix is absent.

4. Analysis for boundary value problems. In applications to boundary value problems, we shall systematically interpret E's columns as being a basis for a subspace of certain slowly varying residual components. (3.5) is used to "deflate" such components from each residual. As an example of the procedure, consider some finite difference or simple finite element discretization of a Poisson problem. Let the discrete system be written as in (3.1), with residuals $r = f - Lu$, where now L is $N \times N$.

Let residual components r_i, $i = 1, 2, \cdots, N$, be partitioned into m disjoint subsets g_k, $k = 1, 2, \cdots, m$, and define column k of E to be the incidence vector for subset g_k, namely

$$E_{ik} = \begin{cases} 1, & r_i \in g_k, \\ 0, & r_i \notin g_k. \end{cases} \tag{4.1}$$

The residual subsets are associated with nodes and corresponding links of the mesh or triangulation. In the simplest case, according to (3.5)-(3.6) and (4.1) a near "piecewise constant" approximation to the residual is being made on the nodes of the subdomain naturally related to the subset g_k. This, together with the need for a relatively simple calculation of the approximation, is what controls the selection of the subsets. In particular, it is necessary that the residual values in any given subset g_k are in some sense well correlated. Moreover, it is reasonable to try to make the subsets such that each of them contributes approximately equally to the approximation error. Such approximations have a long history going back to [14] and, allowing for technical developments, today play a basic role in the multigrid environment. Closer to the spirit of the present paper is work on "coarse mesh rebalancing" and similar ideas [10], [12], [15], [17].

We now analyze a model boundary value problem. Specifically, let Ω in $\mathbb{R}^2$ denote a bounded polygonal domain with boundary $\partial\Omega$ and let τ^h denote a uniformly regular sequence of triangulations of Ω, where h denotes the longest side of all of the triangles of τ^h. $u \in H^1_e(\Omega)$ is sought such that

$$\int_\Omega \operatorname{grad} u \cdot \operatorname{grad} v = \int_\Omega fv \quad \forall v \in H^1_e(\Omega) \tag{4.2}$$

where

$$H^1_e(\Omega) = \{v \in H^1(\Omega) | v|_{\partial\Omega_1} = 0, \partial\Omega = \partial\Omega_1 \cup \partial\Omega_2\}.$$

Boundary values are taken as traces in the usual way, and throughout, all integrals are with respect to the standard Lebesgue measure in $\mathbb{R}^2$. $S^h \subset H^1_e(\Omega)$ is picked as trial space where S^h consists of piecewise linear continuous functions with the standard basis. This yields the linear system

$$LU = F,$$

which will be solved by the deflated conjugate gradient method. L is a symmetric positive definite matrix of order $N \times N$.

Choice of the basis matrix E proceeds as above in (4.1). In this particular case the residual subsets are naturally in a one-to-one correspondence with subsets of nodes of the triangulation. These associated node sets will be denoted by ν_k, consisting of n_k nodes, for $k = 1, 2, \cdots, m$. Observe that E has orthogonal columns. The union of all triangles of τ^h, all of whose vertices are in ν_k, will be denoted by $\bar{\Omega}_k = \Omega_k \cup \partial\Omega_k$. Thus, with strict inclusion except in trivial cases

$$\bigcup \Omega_k \subset \Omega.$$

We shall use Poincaré's inequality in the following form.

LEMMA 4.1. *Let Ω' be a convex domain in $\mathbb{R}^2$, with Lipschitz boundary $\partial\Omega'$. Then a constant C_p exists, such that for all $u \in H^1(\Omega')$ for which*

$$\int_{\Omega'} u = 0$$

the inequality

$$\int_{\Omega'} u^2 \leqq C_p d^2 \int_{\Omega'} |\text{grad } u|^2$$

holds where d denotes the diameter of Ω'.

Proof. See [9, Thm. 3.6.5].

This lemma can be used to provide an error estimate for approximation of a function by its mean.

LEMMA 4.2. *Let $u \in H^1(\Omega')$, and define $\hat{u}$ to be the mean of u over Ω'. Then*

$$\int_{\Omega'} (u - \hat{u})^2 \leqq C_p d^2 \int_{\Omega'} |\text{grad } u|^2.$$

Proof. Apply Poincaré's inequality to $(u - \hat{u})$, noting that this function has zero mean over Ω'.

In the future, elements of S^h will be denoted by lower case letters and their corresponding vectors of nodal values by upper case letters. Some fixed ordering is assumed for this purpose. In the triangulated domain Ω_k, let $S(h, k)$ denote the set of continuous piecewise linear functions on the triangulation inherited from τ_h. Given u, v in $S(h, k)$, with nodal values U, V it follows that

$$\int_{\Omega_k} uv = (U, M_k V)_k \tag{4.3}$$

where M_k is the mass (Gram) matrix of the standard basis for $S(h, k)$ and the subscript on the standard inner product refers to the summation being taken over the set of nodes in ν_k. We shall use the standard estimate

$$(U, M_k U)_k \geqq C_0(k) h^2 (U, U)_k \tag{4.4}$$

and set

$$C_0 = \min_k (C_0(k)).$$

Suppose now that Ω_k are convex, and that

$$\int_{\Omega_k} u = 0, \qquad u \in S(h, k).$$

Then by Lemma 4.1

$$\begin{aligned}\int_{\Omega_k} u^2 &\leqq C_p(k) d^2(k) \int_{\Omega_k} |\text{grad } u|^2 \\ &\leqq C_p d^2 \int_{\Omega_k} |\text{grad } u|^2\end{aligned}$$

where now

$$C_p = \max_k (C_p(k)), \qquad d = \max_k (d(k)).$$

We are now in a position to prove an estimate for λ_1.

THEOREM 4.1. *Let $\Omega_k = 1, 2, \cdots, m$ each be Lipschitz. Then with $C_1 = C_0/C_p$*

$$\lambda_1 \geqq (C_1 h^2/d^2).$$

Proof. Let u be in S^h. By Lemma 4.2, with $\hat{u}$ denoting the mean over Ω_k of u,

$$\int_{\Omega_k} (u-\hat{u})^2 \leqq C_p d^2 \int_{\Omega_k} |\text{grad } u|^2, \tag{4.5}$$

and using (4.3) and (4.4) it then follows with $V = (U - \hat{U})|_k$ that

$$\begin{aligned}(V, V)_k &\leqq C_0^{-1} h^{-2} C_p d^2 \int_{\Omega_k} |\text{grad } u|^2 \\ &= \alpha \int_{\Omega_k} |\text{grad } u|^2 \quad \text{where } \alpha = (C_1 h^2/d^2)^{-1}.\end{aligned}$$

Clearly,

$$\sum_k \int_{\Omega_k} |\text{grad } u|^2 \leqq \int_{\Omega} |\text{grad } u|^2 = (U, LU). \tag{4.6}$$

Suppose now that

$$LU = G, \qquad E^t G = 0. \tag{4.7}$$

Then with W in $\mathbb{R}^m$ denoting the vector of means over the subdomains, and by (4.6) and (4.7),

$$\begin{aligned}(U, LU) &= (U - EW, LU) = (V, LU), \qquad V = U - EW \\ &= \sum_k (V, LU)_k \\ &\leqq \sum_k (V, V)_k^{1/2} (LU, LU)_k^{1/2} \\ &\leqq \sum_k \left\{\alpha \int_{\Omega_k} |\text{grad } u|^2\right\}^{1/2} (LU, LU)_k^{1/2} \\ &\leqq \left\{\sum_k \alpha \int_{\Omega_k} |\text{grad } u|^2\right\}^{1/2} \left\{\sum_k (LU, LU)_k\right\}^{1/2} \\ &\leqq \alpha^{1/2} \left\{\int_{\Omega} |\text{grad } u|^2\right\}^{1/2} (LU, LU)^{1/2} \\ &\leqq \alpha^{1/2} (U, LU)^{1/2} (LU, LU)^{1/2}\end{aligned}$$

yielding the estimate

$$(U, LU) \leqq \alpha (LU, LU),$$

from which it follows immediately that

$$\lambda_1 \geqq \alpha^{-1},$$

so the result is proved.

Also required is an estimate for λ_l. This can easily be obtained by noting that the inequality

$$(U, U) \leqq (U, LU)^{1/2} (U, L^{-1} U)^{1/2}$$

implies, using Lemma 3.1 and the Rayleigh quotient, that

$$\lambda_l \leqq \rho(L),$$

$\rho(\ )$ denoting spectral radius. Gershgorin's theorem yields the estimate

$$\rho(L) \leqq K \tag{4.8}$$

where K is independent of h. Numerical estimates for K can easily be obtained in any particular case.

The estimates for λ_1 and λ_l just proved are for the Poisson equation using linear elements. However, similar estimates with different constants are easily seen to be true in other cases. The first extension is to the generalized equation

$$\operatorname{div}(a(x, y) \operatorname{grad} u) = f$$

with the usual kinds of boundary conditions, on the assumption that $a(x, y)$ is a continuous scalar valued function on $\bar{\Omega}$ satisfying

$$0 < a_0 \leqq a(x, y) \leqq a_1 \quad \forall (x, y) \in \bar{\Omega}$$

with constants a_0, a_1. The effect of this extension is merely to change α of Theorem 4.1 to α/a_0, and to appropriately modify K of (4.8). K remains independent of h. Another extension is to Lagrangian elements of higher order. No changes need to be made for this, except for the trivial one of changing the constant C_0 and the spectral radius estimate, which remains independent of h, to their new values. A case for which the extension is slightly less direct is that of elements with more than one type of unknown, such as one with both function and also derivative values at nodes. The basis matrix E this time may be defined purely in terms of the function values, as for the linear elements case, no deflation of the derivative values being attempted. Then an examination of the above proof shows that the estimates for λ_1 and λ_m retain their validity apart from the obvious changes to the constants. This has the remarkable consequence of yielding, in the case of Hermite cubics, a reduced matrix in (3.6) of about $\frac{1}{3}$ the order of what it would be for a discretization using Lagrangian elements of the same degree and on the same triangulation, and suggests that in the latter case the approximation should perhaps involve only the vertex unknowns. Similar considerations can be used for more complex elements, the general rule being to avoid averaging unlike variables together—function values and derivatives in the above example. Further points are that (i) the proof of the theorem remains valid under the assumption that the system of subdomains $\{\Omega_k\}$ is merely such that the estimate (4.5) is valid uniformly as h approaches 0; (ii) the method of proof is clearly independent of the number of space dimensions of Ω; and (iii) with more work, consisting of proving a finite difference version of Poincaré's inequality [11], similar results can be proved for finite difference approximations.

5. Choice of subdomains. The factors affecting the choice of subdomains Ω_k are best seen in a more specific context. Thus, suppose that Ω is the square

$$\{(x, y) \text{ in } \mathbb{R}^2 | 0 < x, y < 1\}$$

and that τ^h denotes the set of triangles obtained by drawing positive sloping diagonals in each of the mesh squares with vertices (x_i, y_j), $(x_i + h, y_j)$, $(x_i, y_j + h)$, $(x_i + h, y_j + h)$, where $h = 1/n$, $x_i = ih$, $y_j = jh$ for $0 \leqq i, j \leqq n - 1$. The following computations are valid asymptotically as h tends to 0. Suppose we choose each Ω_k to be a square containing $O(n^\beta)$ nodes on each side, for some $0 < \beta < 1$; let $\sigma = 1 - \beta$. Ω thus contains $O(n^{2\sigma})$

disjoint subsquares, and each subsquare contains $O(n^{2\beta})$ nodes of the original triangulation. The deflation calculation requires the formation and factoring of the reduced matrix $\mathbb{L}=E^tLE$, and for each iteration, the substitution work necessary for solution of (3.6). It is clear that $\mathbb{L}$ can be formed in $O(N)$, where $N=n^2$, additions. Moreover it is also easy to see that $\mathbb{L}$ can be ordered to have band-width $O(n^\sigma)$. If we agree to compute the factors of $\mathbb{L}$ by a simple banded matrix scheme then $O(n^{4\sigma})$ operations will be required once and for all, and for the substitutions $O(n^{3\sigma})$ operations will be required at each iteration. To estimate the number of iterations required, recall Theorem 4.1 where we now have $d=O(h^\sigma)$; this gives the reduced condition number $\kappa=O(n^{2-2\sigma})=O(N^\beta)$, and so the number of iterations to reduce the initial error by, say, 10^{-1} is $O(N^{\beta/2})$. Thus, reducing the error by this factor requires $O(N^{\beta/2}(N+N^{3\sigma/2}))$ operations for the iterations themselves and $O(N^{2\sigma})$ operations for the factoring. These magnitudes will be the same when $\beta=2/5$. With these assumptions then, the optimal sidelength for the square subdomains is $O(h^{3/5})$ and $O(N^{1/5})$ iterations will be required to compute each decimal digit of the solution. If the solution of the continuous problem is smooth enough, then provided that reasonable initial approximations can be found, $O(\log N)$ digits will make sense as an approximation to the solution of the differential equation. On this basis we can claim that $O(N^{6/5}\log N)$ operations will suffice to solve the algebraic problems as h tends to zero. Varying the assumptions will give different operation counts. For example, we could use nested dissection to compute the factors and get a further reduction in the count or we could argue that many sets of data are to be analyzed and that the cost of factoring $\mathbb{L}$ is ignorable. The crude nature of the approximations involved suggests that only general guidelines be given for other cases. Thus, generalizing from the special case outlined, it seems that choosing between $N^{1/3}$ and $N^{2/3}$ near convex subdomains of approximately equal areas is best for regular triangulations.

6. Further comments. Numerical results have confirmed the theoretical rates of convergence and will be reported later. For the very simple problem used as a model in the previous section and with $\beta=\frac{1}{2}$ the method compares very favorably in overall computer time with (and has the same theoretical rate of convergence as) standard factorized preconditioning methods. The unpreconditioned deflated scheme has a relatively low cost compared with these schemes, because apart from the back and forward substitutions of the reduced system only additions are used. The data for the reduced equation are already available, being required for computing the α parameter of the conjugate gradient part of the method. The method has something in common with a two level multigrid scheme, although neither smoothing nor subgrids is explicitly used. The fact that the deflation technique can be applied on arbitrary triangulations is a definite advantage not shared by more specialized methods.

No theoretical predictions are available at present on the rate of convergence to be expected with preconditioned versions.

REFERENCES

[1] O. Axelsson and V. A. Barker, *Finite Element Solution of Boundary Value Problems*, Academic Press, New York, 1984.

[2] R. Chandra, *Conjugate gradient methods for partial differential equations*, Ph.D. thesis, Yale Univ., New Haven, CT, 1978.

[3] T. Dupont, R P. Kendall and H. H. Rachford, Jr., *An approximate factorization procedure for solving self-adjoint elliptic difference equations*, this Journal, 5 (1968), pp. 559-573.

[4] H. C. Elman *Iterative methods for large spare nonsymmetric systems of linear equations*, Ph.D. thesis, Yale Univ., New Haven, CT, 1982.

[5] M. ENGELL, T. GINSBURG, H. RUTISHAUSER AND E. STIEFEL, *Refined Iterative Methods for Computation of the Solution and the Eigenvalues of Self-Adjoint Boundary Value Problems*, Birkhauser Verlag, Basel/Stuttgart, 1959.

[6] D. J. EVANS, *The use of preconditioning in iterative methods for solving linear equations with symmetric positive definite matrices*, J. Inst. Math. Applic., 4 (1967), pp. 295-314.

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

[8] J. A. MEIJERINK AND H. A. VAN DER VORST, *An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix*, Math. Comp., 31 (1977), pp. 148-162.

[9] C. B. MORREY, *Multiple Integrals in the Calculus of Variations*, Springer-Verlag, New York, 1966.

[10] S. NAKAMURA, *Computational Methods in Engineering and Science*, John Wiley, New York, 1977.

[11] R. A. NICOLAIDES, *On the observed rate of convergence of an iterative method applied to a model elliptic difference equation*, Math. Comp., 32 (1978), pp. 127-133.

[12] A. SETTARI AND K. AZIZ, *Generalization of the additive correction methods for the iterative solution of matrix problems*, this Journal, 10 (1973) pp. 506-521.

[13] H. L. STONE, *Iterative solution of implicit approximations of multidimensional partial differential equations*, this Journal, 5 (1968), pp. 530-558.

[14] R. V. SOUTHWELL, *Relaxation Methods in Theoretical Physics*, Clarendon Press, Oxford, 1946.

[15] F. DE LA VALLEE POUSSIN, *An accelerated relaxation algorithm for iterative solution of elliptic equations*, this Journal, 5 (1968), pp. 340-351.

[16] R. S. VARGA, *Factorization and normalized iterative methods*, in Boundary Value Problems in Differential Equations, R. E. Langer, ed., Univ. of Wisconsin Press, Madison, 1960, pp. 121-142.

[17] E. L. WACHSPRESS, *Iterative Solution of Elliptic Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1966.