

Jinchao Xu and Colleagues

The Collected Papers of James H. Bramble

September 13, 2021

Springer
Berlin Heidelberg New York
Hong Kong London
Milan Paris Tokyo

Contents

1 Finite Difference Methods	9
1.1 On the formulation of finite difference analogues of the Dirichlet problem for Poisson's equation	9
1.2 Fourth-order finite difference analogues of the Dirichlet problem for Poisson's equation in three and four dimensions	28
1.3 On a finite difference analogue of an elliptic boundary problem which is neither diagonally dominant nor of non-negative type	35
1.4 Approximation of solutions of mixed boundary value problems for Poisson's equation by finite differences	52
1.5 A finite difference analog of the Neumann problem for Poisson's equation	64
1.6 A second order finite difference analog of the first biharmonic boundary value problem	79
1.7 Error estimates for difference methods in forced vibration problems	94
1.8 On the convergence of difference approximations to weak solutions of Dirichlet's problem	107
2 Finite element method	119
2.1 New monotone type approximations for elliptic problems (1964) ...	119
2.2 Bramble-Hilbert Lemma (1970)	139
2.3 Rayleigh-Ritz-Galerkin methods for dirichlet's problem using subspaces without boundary conditions (1970)	153
2.4 Triangular elements in the finite element method (1970)	153
2.5 Higher order local accuracy by averaging in the finite element method (1977)	166
2.6 Single step Galerkin approximations for parabolic problems (1977). ...	185
2.7 Some convergence estimates for semidiscrete Galerkin type approximations for parabolic equations (1977)	185
2.8 Semidiscrete and single step fully discrete approximations for second order hyperbolic equations (1979)	185
2.9 Some estimates for a weighted L^2 projection (1991).....	185

Contents

2.10	A finite element method for interface problems in domains with smooth boundaries and interfaces (1996)	200
2.11	On the stability of the L ₂ projection in $H^1(\Omega)$ (2002)	200
2.12	A proof of the inf-sup condition for the Stokes equations on Lipschitz domains (2003)	211
2.13	Super-convergence	222
3	Multigrid Methods	223
3.1	New convergence estimates for multigrid algorithms	223
3.2	The analysis of multigrid algorithms for nonsymmetric and indefinite elliptic problems	243
3.3	Parallel multilevel preconditioners	270
3.4	The analysis of multigrid algorithms with nonnested spaces or noninherited quadratic forms	293
3.5	Convergence estimates for multigrid algorithms without regularity assumptions	328
3.6	The analysis of smoothers for multigrid algorithms	328
3.7	New estimates for multilevel algorithms including the V-cycle	351
3.8	The analysis of multigrid algorithms for pseudodifferential operators of order minus one	377
3.9	Uniform convergence of multigrid V-cycle iterations for indefinite and nonsymmetric problems	377
3.10	The analysis of multigrid methods	397
3.11	Non-overlapping domain decomposition methods	397
3.12	Overlapping domain decomposition methods	397
4	Domain Decomposition Methods	399
4.1	The construction of preconditioners for elliptic problems by substructuring. I	399
4.2	The construction of preconditioners for elliptic problems by substructuring. II	432
4.3	The construction of preconditioners for elliptic problems by substructuring. III	449
4.4	The construction of preconditioners for elliptic problems by substructuring. IV	466
4.5	An iterative method for elliptic problems on regions partitioned into substructures	491
4.6	A domain decomposition technique for Stokes problems	501
4.7	Convergence estimates for product iterative methods with applications to domain decomposition	513
4.8	Domain decomposition methods for problems with uniform local refinement in two dimensions	535
4.9	Domain decomposition methods for problems with partial refinement	546
4.10	Analysis of non-overlapping domain decomposition algorithms with inexact solves	562

Contents

5 PML Methods	583
5.1 Analysis of a finite PML approximation for the three dimensional time-harmonic Maxwell and acoustic scattering problems	583
6 Others	603
References	605

Preface

In this book, we collect a number of representative papers written by James H. Bramble. Bramble has made fundamental contributions to a number of subjects numerical methods for partial differential equations, including

1

Finite Difference Methods

1.1 On the formulation of finite difference analogues of the Dirichlet problem for Poisson's equation

On the formulation of finite difference analogues of the Dirichlet problem for Poisson's equation [34]

Article

On the Formulation of finite Difference analogues of the Dirichlet Problem for Poisson's Equation.
BRAMBLE, J.H.; HUBBARD, B.E.
in: Numerische Mathematik | Numerische Mathematik - 4
17 Page(s) (313 - 329)



Nutzungsbedingungen

DigiZeitschriften e.V. gewährt ein nicht exklusives, nicht übertragbares, persönliches und beschränktes Recht auf Nutzung dieses Dokuments. Dieses Dokument ist ausschließlich für den persönlichen, nicht kommerziellen Gebrauch bestimmt. Das Copyright bleibt bei den Herausgebern oder sonstigen Rechteinhabern. Als Nutzer sind Sie nicht dazu berechtigt, eine Lizenz zu übertragen, zu transferieren oder an Dritte weiter zu geben.

Die Nutzung stellt keine Übertragung des Eigentumsrechts an diesem Dokument dar und gilt vorbehaltlich der folgenden Einschränkungen:

Sie müssen auf sämtlichen Kopien dieses Dokuments alle Urheberrechtshinweise und sonstigen Hinweise auf gesetzlichen Schutz beibehalten; und Sie dürfen dieses Dokument nicht in irgend einer Weise abändern, noch dürfen Sie dieses Dokument für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen; es sei denn, es liegt Ihnen eine schriftliche Genehmigung von DigiZeitschriften e.V. und vom Herausgeber oder sonstigen Rechteinhaber vor.

Mit dem Gebrauch von DigiZeitschriften e.V. und der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Terms of use

DigiZeitschriften e.V. grants the non-exclusive, non-transferable, personal and restricted right of using this document. This document is intended for the personal, non-commercial use. The copyright belongs to the publisher or to other copyright holders. You do not have the right to transfer a licence or to give it to a third party.

Use does not represent a transfer of the copyright of this document, and the following restrictions apply:

You must abide by all notices of copyright or other legal protection for all copies taken from this document; and You may not change this document in any way, nor may you duplicate, exhibit, display, distribute or use this document for public or commercial reasons unless you have the written permission of DigiZeitschriften e.V. and the publisher or other copyright holders.

By using DigiZeitschriften e.V. and this document you agree to the conditions of use.

Kontakt / Contact

[DigiZeitschriften e.V.](#)
Papendiek 14
37073 Goettingen
[Email: info@digizeitschriften.de](mailto:info@digizeitschriften.de)

On the Formulation of finite Difference analogues of the Dirichlet Problem for Poisson's Equation*

By

J. H. BRAMBLE and B. E. HUBBARD

1. Introduction

In the approximate solution of the Dirichlet problem for Poisson's equation many finite difference methods have been proposed. It is important for the formulation of the approximating problem to have some measure of the deviation of the finite difference solution from the exact solution. The question of convergence itself for a class of problems has been discussed by COURANT, FRIEDRICH and LEWY [3].

In 1930 GERSCHGORIN [5] gave a method for obtaining an estimate of the order of convergence of the finite difference approximation to the solution of the Dirichlet problem for a class of elliptic equations. His method was based on a maximum principle for the finite difference analogue. In a note in 1933 COLLATZ [1] proposed a certain boundary approximation and, using the techniques of GERSCHGORIN, showed that this approximation gives rise to an $O(h^2)$ estimate for the truncation error. The estimates of both GERSCHGORIN and COLLATZ assume the knowledge of bounds for certain higher derivatives of the solution of the Dirichlet problem.

From an analogy to probability theory COURANT, FRIEDRICH, and LEWY [3] give a finite difference Green's function for the Dirichlet problem for Poisson's equation. Using this Green's function they give an analogue of Green's third identity. WASOW [14] studies the asymptotic behavior of the finite difference Green's function and LAASONEN [8] uses an explicit representation of the finite difference Green's function for the rectangle to obtain bounds in that case.

In this paper we obtain some further estimates of the type proposed by GERSCHGORIN. The approach taken here is to define an appropriate related finite difference Green's function for various finite difference analogues. In each case the analogue of Green's third identity is given and used to obtain estimates for the truncation error.

In section 2 the truncation error is studied for a finite difference approximation proposed by SHORTLEY and WELLER [10]. Although at points near the boundary the finite difference operator approximates the Laplace operator only to $O(h)$ it is seen that the resulting contribution to the truncation error is $O(h^3)$.

* This research was supported in part by the United States Air Force through the Air Force Office of Scientific Research of the Air Research and Development Command under Contract No. AF 49(638)-228.

This has been shown to be the case for the homogeneous equation by FORSYTHE and WASOW [4].

In section 3 techniques similar to those of section 2 are used to obtain certain known results. For example the method of COLLATZ is treated from this point of view. From these considerations we are led to formulate finite difference approximations in which the matrix representing the linear system is not of "positive type", (MOTZKIN and WASOW [9]). In this connection we state a general theorem on error estimation for a class of finite difference analogues of the Dirichlet problem for Poisson's equation. This theorem is similar to a special case of a theorem of FORSYTHE and WASOW [4, p. 302] concerning approximations whose associated matrices are of "positive type".

In the final section we apply this theorem to a certain finite difference problem and show that the resulting truncation error is $O(h^4)$. Finally, combining the techniques of the previous two sections we formulate yet another $O(h^4)$ finite difference approximation.

For some other studies of this problem we refer the reader to [1], [4], [7], [11], [12], [13] and references therein.

2. Second Order Estimates

Throughout this paper we shall be concerned with finite difference approximations to the Dirichlet problem for Poisson's equation, i.e.

$$(2.1) \quad \begin{aligned} \Delta u(x, y) &= F(x, y), & (x, y) \in R \\ u(x, y) &= f(x, y), & (x, y) \in C. \end{aligned}$$

We assume that R is a bounded region in the (x, y) plane with boundary C .

We cover the (x, y) plane with a grid made up of two families of lines. Each family consists of lines, a distance h apart, parallel to one of the coordinate axes. The intersections will be called either "grid" or "mesh" points and a function defined at such points will be termed a "mesh" function. If we restrict ourselves to a finite portion of the plane such a function can be considered as a vector in a finite dimensional vector space. Such is the case with the point sets which we shall now define.

Let R_h be the set of those mesh points in R whose nearest neighbors in the x and y directions lie in R . Those grid points in R which do not belong to R_h will make up the set called C_h^* . The points of intersection of the grid with the boundary C form the set C_h .

For any point P belonging to $R_h + C_h^* + C_h$ we define its neighbors $N(P)$ to be those nearest points in $R_h + C_h^* + C_h$, lying along grid lines through P .

If $V(x, y)$ is an arbitrary mesh function defined on $R_h + C_h^* + C_h$ then for such vectors we define the finite difference operator Δ_h . If $(x, y) \in R_h$ then

$$(2.2) \quad \begin{aligned} \Delta_h V(x, y) &\equiv h^{-2} \{V(x+h, y) + V(x, y+h) + \\ &\quad + V(x-h, y) + V(x, y-h) - 4V(x, y)\}. \end{aligned}$$

This is the usual $O(h^2)$ approximation of Δ for functions $v(x, y) \in C^4$ in R . In fact

$$(2.3) \quad |\Delta v(x, y) - \Delta_h v(x, y)| \leq \frac{h^2}{6} M_4, \quad (x, y) \in R_h,$$

where we have used the notation

$$(2.4) \quad M_j = \sup_{P \in R} \left\{ \left| \frac{\partial^j U(P)}{\partial x^i \partial y^{j-i}} \right| \mid i = 0, 1, \dots, j \right\}.$$

At points of C_h^* , Δ_h is defined to be the five point divided difference approximation to Δ . For example if $(\bar{x}, \bar{y}) \in C_h^*$ is the point in Fig. 1 then

$$(2.5) \quad \begin{aligned} \Delta_h V(\bar{x}, \bar{y}) \equiv & 2h^{-2} \left\{ \left(\frac{1}{\alpha+1} \right) V(\bar{x} + h, \bar{y}) + \frac{1}{\alpha(\alpha+1)} V(\bar{x} - \alpha h, \bar{y}) + \right. \\ & \left. + \left(\frac{1}{\beta+1} \right) V(\bar{x}, \bar{y} + h) + \frac{1}{\beta(\beta+1)} V(\bar{x}, \bar{y} - \beta h) - \left(\frac{1}{\alpha} + \frac{1}{\beta} \right) V(\bar{x}, \bar{y}) \right\}. \end{aligned}$$

If $\alpha = \beta = 1$ then Δ_h takes the same form as in (2.2). In fact, either α or β or both may equal 1. We note that Δ_h as defined in (2.5) approximates Δ to $O(h)$ for $v(x, y) \in C^3$ in R , i.e.

$$(2.6) \quad |\Delta v(\bar{x}, \bar{y}) - \Delta_h v(\bar{x}, \bar{y})| \leq \frac{2M_3 h}{3}.$$

As in the formulation of SHORTLEY and WELLER [10] we allow any or all the neighbors of a point $P \in C_h^*$ to lie at a distance less than or equal to h from P . The appropriate analogy of (2.5) is assumed.

We consider the following finite difference analogue of (2.1),

$$(2.7) \quad \begin{aligned} \Delta_h U(x, y) &= F(x, y), \quad (x, y) \in R_h + C_h^*, \\ U(x, y) &= f(x, y), \quad (x, y) \in C_h. \end{aligned}$$

This is just a system of simultaneous linear equations for the determination of the mesh function $U(x, y)$. It is well-known that the associated determinant of such a system does not vanish and hence there exists a unique solution for (2.7); cf. COLLATZ [2, p. 46].

We shall show that the truncation error $\varepsilon(P) \equiv u(P) - U(P)$, $P \in R_h + C_h^* + C_h$, satisfies an inequality of the type

$$(2.8) \quad |\varepsilon|_M \leq K h^2$$

where K is a constant independent of P and h . In (2.8) we have used the notation

$$(2.9) \quad \psi_M = \sup_{P \in S \subset \bar{R}} \psi(P)$$

for any function ψ defined on a subset S of \bar{R} . Before proceeding we shall introduce the finite difference analogue of the Green's function, $G_h(P, Q)$ which is defined by

$$(2.10) \quad \begin{aligned} \Delta_{h,P} G_h(P, Q) &= -\delta(P, Q) h^{-2}, \quad P \in R_h + C_h^* \\ G_h(P, Q) &= \delta(P, Q), \quad P \in C_h \end{aligned}$$

for $Q \in R_h + C_h^* + C_h$.

Here

$$(2.11) \quad \delta(P, Q) = \begin{cases} 1, & P = Q \\ 0, & P \neq Q. \end{cases}$$

We shall now prove the finite difference analogues of some well known theorems in potential theory.

Lemma 2.1 (Maximum Principle). For any mesh function $V(P)$ defined on $R_h + C_h^* + C_h$ if $\Delta_h V(P) \geq 0$ for $P \in R_h + C_h^*$ then $V(P)$ takes on its maximum on C_h .

Such a result was first obtained by GERSCHGORIN [5] and is easily seen to be a special case of more general theorems given by COLLATZ [2].

Lemma 2.2 (Green's Third Identity). Let $V(P)$ be an arbitrary mesh function defined on $R_h + C_h^* + C_h$. Then for any $P \in R_h + C_h^* + C_h$

$$(2.12) \quad V(P) = h^2 \sum_{Q \in R_h + C_h^*} G_h(P, Q) [-\Delta_h V(Q)] + \sum_{Q \in C_h} G_h(P, Q) V(Q).$$

Proof. The relation (2.12) can be proved from the finite difference analogue of Green's second identity. It can be seen more simply from the following considerations. Let $W(P)$ be the right hand side of (2.12). By a direct calculation using the properties of the Green's function $G_h(P, Q)$ it follows that

$$(2.13) \quad \Delta_h W(P) = \Delta_h V(P), \quad P \in R_h + C_h^*$$

$$(2.14) \quad W(P) = V(P), \quad P \in C_h.$$

From the uniqueness of the solution of (2.7) we have $W(P) = V(P)$.

Lemma 2.3.

$$(2.15) \quad G_h(P, Q) \geq 0, \quad Q \in R_h + C_h^* + C_h.$$

Proof. Apply the maximum principle (lemma 2.1) to $-G_h(P, Q)$ for arbitrary but fixed $Q \in R_h + C_h^* + C_h$.

Lemma 2.4.

$$(2.16) \quad \sum_{Q \in C_h^*} G_h(P, Q) \leq 1, \quad P \in R_h + C_h^* + C_h.$$

Proof. Let the mesh function $W(P)$ be given by

$$(2.17) \quad W(Q) = \begin{cases} 1, & Q \in R_h + C_h^*, \\ 0, & Q \in C_h. \end{cases}$$

Then $\Delta_h W(Q) = 0$, $Q \in R_h$. It is easily seen from the definition of Δ_h on C_h^* that $-\Delta_h W(Q) \geq h^{-2}$.

Applying lemma 2.2 it follows that for $P \in R_h + C_h^*$

$$1 = h^2 \sum_{Q \in C_h^*} G_h(P, Q) [-\Delta_h W(Q)] \geq \sum_{Q \in C_h^*} G_h(P, Q).$$

If $P \in C_h$ the inequality (2.16) is trivially satisfied.

Lemma 2.5. If d is the diameter of the smallest circumscribed circle containing R then

$$(2.18) \quad h^2 \sum_{Q \in R_h + C_h^*} G_h(P, Q) \leq \frac{d^2}{16}, \quad P \in R_h + C_h^* + C_h.$$

Proof. Let 0 be the center of the circumscribed circle about R of diameter d . Let $W(P) = \frac{r(P)^2}{4}$ for $P \in R_h + C_h^* + C_h$ where $r(P)$ is the Euclidean distance from 0 to P . Then

$$\Delta_h W(P) = 1, \quad P \in R_h + C_h^*.$$

Now define the mesh function

$$V(P) \equiv h^2 \sum_{Q \in R_h + C_h^*} G_h(P, Q).$$

We see from (2.10) that

$$(2.19) \quad \begin{aligned} \Delta_h V(P) &= -1, & P \in R_h + C_h^* \\ V(P) &= 0, & P \in C_h. \end{aligned}$$

Hence $\Delta_h [V(P) + W(P)] = 0$ for $P \in R_h + C_h^*$ and $V(P) + W(P) \leq \frac{d^2}{16}$ for $P \in C_h$.

By the maximum principle, since $W \geq 0$, it follows that

$$V(P) \leq \frac{d^2}{16}, \quad P \in R_h + C_h^* + C_h$$

which completes the proof of lemma 2.5.

We are now in a position to prove the principal result of section 2.

Theorem 1. Let $u(x, y)$ be the solution of (2.1) and $U(x, y)$ the solution of (2.7). Then the truncation error $\varepsilon(P) \equiv u(P) - U(P)$ satisfies the inequality

$$(2.20) \quad |\varepsilon|_M \leq \frac{M_4 d^2}{96} h^2 + \frac{2M_3}{3} h^3.$$

Proof. Since $\varepsilon(P) = 0$, $P \in C_h$ we see from lemma 2.2 that

$$(2.21) \quad \varepsilon(P) = h^2 \sum_{Q \in R_h + C_h^*} G_h(P, Q) [-\Delta_h \varepsilon(Q)].$$

It follows from (2.1) and (2.7) that

$$(2.22) \quad |-\Delta_h \varepsilon(Q)| = |\Delta_h u(Q) - \Delta u(Q)|.$$

Applying (2.22) to (2.21) and using (2.3) and (2.6) we have

$$|\varepsilon(P)| \leq \left[h^2 \sum_{Q \in R_h} G_h(P, Q) \right] \frac{h^2 M_4}{6} + \left[\sum_{Q \in C_h^*} G_h(P, Q) \right] \frac{2M_3}{3} h^3.$$

Finally using lemmas 2.4 and 2.5 we arrive at (2.20), which completes the proof of theorem 1.

3. Other Boundary Approximations

In this section we derive some known inequalities using the results and techniques of the previous section. In addition we show that in certain cases the requirement of positivity can be removed near the boundary C . In this connection we give an example of a formulation of a finite difference analogue of (2.1) which fails to be of positive type at points of C_h^* . For this problem we derive an over-all $O(h^3)$ estimate for the truncation error and show that the contribution from points of C_h^* is $O(h^3)$ as in Theorem 1.

Finally we give a general theorem on error estimation for finite difference approximations to Poisson's equation. This result, in certain cases, is similar to a general theorem given in G. FORSYTHE and W. WASOW [p. 302, theorem 23.7, 4]. In the theorem stated here the condition of non-negativity is relaxed to admit more general approximations near the boundary.

Let $G_h^*(P, Q)$ be the finite difference Green's function for R_h with boundary C_h^* . This is given by

$$(3.1) \quad \begin{aligned} A_{h,P} G_h^*(P, Q) &= -\delta(P, Q) h^{-2}, & P \in R_h \\ G_h^*(P, Q) &= \delta(P, Q), & P \in C_h^* \end{aligned}$$

for all $Q \in R_h + C_h^*$.

Just as in lemma 2.2 we have the identity

$$(3.2) \quad V(P) = h^2 \sum_{Q \in R_h} G_h^*(P, Q) [-A_h V(Q)] + \sum_{Q \in C_h^*} G_h^*(P, Q) V(Q).$$

In addition all of the other lemmas of section 2 are valid if we make the substitutions

$$(3.3) \quad \begin{aligned} G_h &\rightarrow G_h^*, \\ R_h + C_h^* &\rightarrow R_h \\ C_h &\rightarrow C_h^*. \end{aligned}$$

We shall also need the following lemma.

Lemma 3.1. For $P \in R_h + C_h^*$

$$(3.4) \quad \sum_{Q \in C_h^*} G_h^*(P, Q) = 1.$$

Proof. Apply (3.2) to $V(P) \equiv 1$.

We first define a finite difference analogue of (2.1) and show that the truncation error is $O(h)$. FORSYTHE and WASOW [4] call this "interpolation of order zero". Let $V(P)$ satisfy

$$(3.5) \quad \begin{aligned} A_h U(P) &= F(P), & P \in R_h \\ U(P) &= f(P'), & P \in C_h^* \end{aligned}$$

where P' is one of the neighbors of P in C_h . We have the following theorem.

Theorem 2. Let $u(x, y)$ be the solution of (2.1) and $U(x, y)$ the solution of (3.5). Then the truncation error $\varepsilon(P) = U(P) - u(P)$ satisfies the inequality

$$(3.6) \quad |\varepsilon|_M \leq h M_1 + \frac{M_4 d^2}{96} h^2.$$

Proof. Applying (3.2) to $\varepsilon(P)$ we have

$$(3.7) \quad \varepsilon(P) = h^2 \sum_{Q \in R_h} G_h^*(P, Q) [-A_h \varepsilon(Q)] + \sum_{Q \in C_h^*} G_h^*(P, Q) \varepsilon(Q).$$

We note that for $Q \in C_h^*$

$$(3.8) \quad |\varepsilon(Q)| = |u(Q) - U(Q)| = |u(Q) - U(Q')| \leq h M_1.$$

Now from (2.3) and (3.5) we have also that

$$(3.9) \quad |A_h \varepsilon(Q)| \leq \frac{h^2}{6} M_4, \quad Q \in R_h.$$

Taking absolute values of both sides of (3.7), using (3.8) and (3.9), and applying lemmas 2.5 and 3.1, the inequality (3.6) follows easily.

We next consider the finite difference analogue of (2.1) given by COLLATZ [1], [2]. He defines the following approximation to (2.1).

$$(3.10) \quad \begin{aligned} A_h U(P) &= F(P), & P \in R_h \\ U(P) &= f(P), & P \in C_h. \end{aligned}$$

At a point P of C_h^* he prescribes that $U(P)$ lie on a straight line between the values of U at two neighbors of P , one of which is in R_h , the other in C_h . For example for the point (\bar{x}, \bar{y}) of Fig. 1 we have

$$(3.11) \quad U(\bar{x}, \bar{y}) = \frac{\alpha}{\alpha+1} U(\bar{x} + h, \bar{y}) + \frac{1}{\alpha+1} U(\bar{x} - \alpha h, \bar{y}).$$

Alternatively we could have interpolated in the y direction.

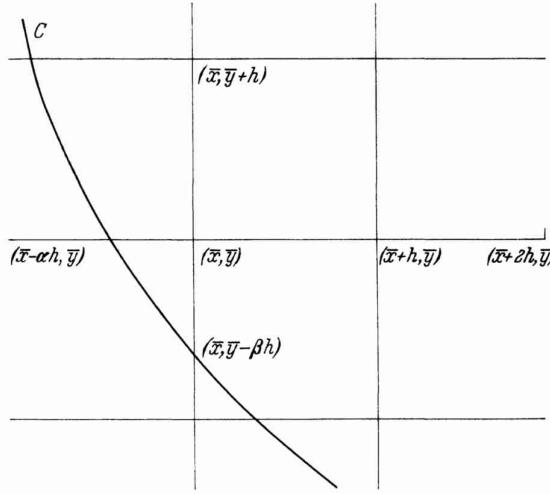


Fig. 1

As COLLATZ has shown [1] this method gives rise to an estimate of the truncation error which is $O(h^2)$. The contribution to the truncation error arising from the points of C_h^* is also $O(h^2)$. The following analysis again yields similar results.

Theorem 3 (COLLATZ). Let $u(x, y)$ be the solution of (2.1) and $U(x, y)$ the solution of (3.10) and (3.11). Then the truncation error $\varepsilon(P) = u(P) - U(P)$ satisfies

$$(3.12) \quad |\varepsilon|_M \leq \left[M_2 + \frac{M_4 d^2}{48} \right] h^2.$$

Proof. As in the proof of theorem 2, $\varepsilon(P)$ satisfies (3.7). For any point $Q \in C_h^*$ we make use of (3.11) to bound $|\varepsilon(Q)|$ as follows.

$$(3.13) \quad \begin{aligned} |\varepsilon(\bar{x}, \bar{y})| &= |u(\bar{x}, \bar{y}) - U(\bar{x}, \bar{y})| \\ &= \left| u(\bar{x}, \bar{y}) - \frac{\alpha}{\alpha+1} U(\bar{x} + h, \bar{y}) - \frac{1}{\alpha+1} u(\bar{x} - \alpha h, \bar{y}) \right| \end{aligned}$$

where we have used the second relation in (3.10). Using the triangle inequality we have

$$(3.14) \quad |\varepsilon(\bar{x}, \bar{y})| \leq \left| u(\bar{x}, \bar{y}) - \frac{\alpha}{\alpha+1} u(\bar{x} + h, \bar{y}) - \frac{1}{\alpha+1} u(\bar{x} - \alpha h, \bar{y}) \right| + \frac{\alpha}{\alpha+1} |\varepsilon|_M.$$

From Taylor's theorem and the fact that $0 < \alpha \leq 1$ it follows that

$$(3.15) \quad |\varepsilon(Q)| \leq \frac{M_2}{2} h^2 + \frac{1}{2} |\varepsilon|_M.$$

Employing inequalities (3.15) and (3.9), as well as lemma 2.5 and (3.1) we see that

$$(3.16) \quad |\varepsilon(Q)| \leq \frac{1}{2} |\varepsilon|_M + \left(\frac{M_2}{2} + \frac{M_4 d^2}{96} \right) h^2.$$

Since the right hand side is independent of Q the inequality (3.12) follows.

We next give an example of a finite difference analogue of (2.1) which fails to be of positive type at points of C_h^* . Let $U(P)$ satisfy the system

$$(3.17) \quad \begin{aligned} A_h U(P) &= F(P), & P \in R_h \\ U(P) &= f(P), & P \in C_h. \end{aligned}$$

At a point P of C_h^* let $U(P)$ lie on a parabola through value of $U(P)$ at a neighboring point of C_h and two points of $R_h + C_h^*$. All four points involved must of course be colinear. In addition we require one of the points of $R_h + C_h^*$ to be a neighbor of P and the other to be taken at a distance $3h$ from P . For example, for the point (\bar{x}, \bar{y}) in Fig. 4

$$(3.18) \quad \begin{aligned} U(\bar{x}, \bar{y}) = \frac{3}{3+\alpha(\alpha+4)} &\left\{ U(\bar{x} - \alpha h, \bar{y}) + \frac{\alpha}{2} (\alpha+3) U(\bar{x} + h, \bar{y}) - \right. \\ &\left. - \frac{\alpha}{6} (\alpha+1) U(\bar{x} + 3h, \bar{y}) \right\}. \end{aligned}$$

From Taylor's theorem it is easy to see that for a sufficiently smooth function $U(P)$ in R we have an inequality of the type

$$(3.19) \quad \left| u(\bar{x}, \bar{y}) - \frac{3}{3+\alpha(\alpha+4)} \left\{ u(\bar{x} - \alpha h, \bar{y}) + \frac{\alpha}{2} (\alpha+3) u(\bar{x} + h, \bar{y}) - \right. \right. \\ \left. \left. - \frac{\alpha}{6} (\alpha+1) u(\bar{x} + 3h, \bar{y}) \right\} \right| \leq \frac{14h^3 M_3}{3},$$

where $(\bar{x}, \bar{y}) \in C_h^*$. In some cases the interpolation will be in the y direction. We thus have the following theorem.

Theorem 4. Let $u(x, y)$ be the solution of (2.1) and $U(x, y)$ the solution of (3.17) and (3.18). Then the truncation error $\varepsilon(P) = u(P) - U(P)$ satisfies

$$(3.20) \quad |\varepsilon|_M \leq \frac{d^2 M_4}{12} h^2 + \frac{112}{3} M_3 h^3.$$

Proof. The proof follows in a manner analogous to that of theorem 3. Instead of (3.14) we have the inequality

$$(3.21) \quad |\varepsilon(\bar{x}, \bar{y})| \leq \left| u(\bar{x}, \bar{y}) - \frac{3}{3+\alpha(\alpha+4)} \left\{ u(\bar{x} - \alpha h, \bar{y}) + \frac{\alpha}{2} (\alpha+3) u(\bar{x} + h, \bar{y}) - \right. \right. \\ \left. \left. - \frac{\alpha}{6} (\alpha+1) u(\bar{x} + 3h, \bar{y}) \right\} \right| + \frac{7}{8} |\varepsilon|_M,$$

for the point (\bar{x}, \bar{y}) of Fig. 1. From (3.21) and (3.19) it follows that

$$(3.22) \quad |\varepsilon(Q)| \leq \frac{14M_3}{3} h^3 + \frac{7}{8} |\varepsilon|_M,$$

where $Q \in C_h^*$. As before we have the analogue of (3.16)

$$(3.23) \quad |\varepsilon(Q)| \leq \frac{d^2 M_4}{96} h^2 + \frac{14M_3}{3} h^3 + \frac{7}{8} |\varepsilon|_M,$$

from which (3.20) easily follows.

We see from the previous theorem that the property of positivity of the matrix representing the linear system is not essential in the problem of obtaining error estimates. It is sufficient to replace this condition with a requirement of "interior positivity", provided a "strict" diagonal dominance is satisfied near the boundary. To be more precise we shall need some further definitions.

Let Δ_h now be defined as some finite difference analogue of Δ at mesh points in R . We define R_h to be the set of those mesh points in R at which the operator Δ_h is defined solely in terms of mesh points in R . Let C_h^* be the set of those mesh points in R which do not belong to R_h . At these points we define a finite difference operator Δ_h^* which involves points of $R_h + C_h^*$ and certain points on C . Those points of C which are involved in Δ_h^* at some point of C_h^* will be called C_h . We define $N(P)$ as those points, other than P involved in Δ_h at P , when $P \in R_h$. If $P \in C_h^*$ the definition is the same with Δ_h replaced by Δ_h^* . We define $\sigma(P, Q)$ to be the coefficient of $V(Q)$ in the expression for the operator applied to V at the point P . In terms of the operators Δ_h and Δ_h^* we define the following finite difference analogue of (2.1).

$$(3.24) \quad \begin{aligned} \Delta_h U(P) &= F(P) + \varphi[F(P)], & P \in R_h \\ \Delta_h^* U(P) &= H(P), & P \in C_h^* \\ U(P) &= f(P), & P \in C_h \end{aligned}$$

where $\varphi[F(P)]$ is an operator which is a linear combination of derivatives of F at P . $H(P)$ is, as yet, arbitrary. We give the following definitions.

Definition 3.1. The matrix representing the linear system (3.24) has the property of *interior positivity* if it is of positive type at each point of R_h , [9]. That is

$$(3.25) \quad \frac{\sigma(P, Q)}{\sigma(P, P)} < 0, \quad P \in R_h, \quad Q \in N(P).$$

Definition 3.2. The matrix representing the linear system (3.24) has the property of *strict diagonal dominance* if

$$(3.26) \quad \sum_{Q \in (R_h + C_h^*) \cap N(P)} |\sigma(P, Q)| \leq |\sigma(P, P)|, \quad P \in R_h$$

and

$$(3.27) \quad \sum_{Q \in (R_h + C_h^*) \cap N(P)} |\sigma(P, Q)| \leq |\sigma(P, P)| \delta, \quad P \in C_h^*$$

where δ is a constant less than 1 and independent of h .

Definition 3.3. Let $\overline{N(P)} = [N(P) \cup P] \cap R_h$. We say that R_h is “connected” if for every set S_h properly contained in R_h the decomposition

$$R_h = \left[\bigcup_{P \in S_h} \overline{N(P)} \right] \cup \left[\bigcup_{Q \in R_h - S_h} \overline{N(Q)} \right]$$

implies that

$$\left[\bigcup_{P \in S_h} \overline{N(P)} \right] \cap \left[\bigcup_{Q \in R_h - S_h} \overline{N(Q)} \right]$$

is not empty.

We now state a theorem pertaining to systems of the type given by (3.24).

Theorem 5. Let $u(x, y)$ be the solution of (2.1) and $U(x, y)$ be the solution of (3.24). The matrix representing (3.24) is assumed to satisfy the properties

- (a) Strict diagonal dominance with constant δ .
- (b) Interior positivity.
- (c) R_h is connected.

The operators Δ_h and Δ_h^* satisfy the inequalities

$$(3.28) \quad |\Delta_h v - [\Delta v + \varphi(\Delta v)]| \leq C_1 M_{n+2} h^n, \quad n \geq 2$$

and

$$(3.29) \quad \left| \frac{\Delta_h^* U(P) - H(P)}{\sigma(P, P)} \right| \leq C_2 M_m h^m, \quad m \geq 1, \quad P \in C_h^*$$

where C_1 and C_2 are constants, and H is a function given on C_h^* . In (3.28), v is any sufficiently smooth function in \bar{R} and M_i is related to v . Then $\varepsilon(P) = u(P) - U(P)$ satisfies the inequality

$$(3.30) \quad |\varepsilon|_M \leq \frac{1}{1-\delta} \left\{ C_2 M_m h^m + C_1 \frac{d^2}{16} M_{n+2} h^n \right\}.$$

Proof. We define the Green's function $G_h^*(P, Q)$ related to the operator Δ_h as defined on R_h with boundary C_h^* . That is

$$(3.31a) \quad \Delta_{h,P} G_h^*(P, Q) = -h^{-2} \delta(P, Q), \quad P \in R_h,$$

$$(3.31b) \quad G_h^*(P, Q) = \delta(P, Q), \quad P \in C_h^*,$$

for all $Q \in R_h + C_h^*$.

The existence of $G_h^*(P, Q)$ is assured by (a) and (c), cf. [2, p. 46]. Because of (a), (b) and (c) any function $V(P)$, $P \in R_h + C_h^*$ for which $\Delta_h V(P) \geq 0$, $P \in R_h$ attains its maximum on C_h^* , cf. COLLATZ [p. 45, 2]. From this it follows easily that

$$(3.32) \quad G_h^*(P, Q) \geq 0, \quad Q \in R_h + C_h^*.$$

As before we have the identity

$$(3.33) \quad \varepsilon(P) = h^2 \sum_{Q \in R_h} G_h^*(P, Q) [-\Delta_h \varepsilon(Q)] + \sum_{Q \in C_h^*} G_h^*(P, Q) \varepsilon(Q).$$

This may be verified as in lemma 2.2. We observe that lemma 3.1 is valid for $G_h^*(P, Q)$ as defined in (3.31). As a consequence of (3.28) we see that

$$(3.34) \quad \Delta_h \left(\frac{r(P)^2}{4} \right) \equiv \Delta \left(\frac{r(P)^2}{4} \right) = 1,$$

since we may take $M_{n+2} \equiv 0$. As in lemma 2.5 we thus have

$$(3.35) \quad h^2 \sum_{Q \in R_h} G_h^*(P, Q) \leq \frac{d^2}{16}.$$

From the definition $\sigma(P, Q)$ and Δ_h^* we have

$$(3.36) \quad \varepsilon(P) = \frac{\Delta_h^* \varepsilon(P)}{\sigma(P, P)} - \frac{\sum_{Q \in N(P)} \sigma(P, Q) \varepsilon(Q)}{\sigma(P, P)}, \quad P \in C_h^*.$$

Taking the absolute value of both sides of (3.36) it follows that

$$(3.37) \quad |\varepsilon(P)| \leq \left| \frac{\Delta_h^* [u(P) - U(P)]}{\sigma(P, P)} \right| + \frac{\sum_{Q \in (R_h \cap C_h^*) \cap N(P)} |\sigma(P, Q)|}{\sigma(P, P)} |\varepsilon|_M.$$

From (3.24), (3.27) and (3.29) we obtain

$$(3.38) \quad |\varepsilon(P)| \leq C_2 M_m h^m + \delta |\varepsilon|_M.$$

If we now take the absolute value of both sides of (3.33) and apply (3.28), (3.35), (3.37) and lemma 3.1 the inequality (3.30) follows.

As a non-trivial application of theorem 5 an $O(h^4)$ analogue of (2.4) will be presented in the next section.

4. Fourth Order Estimates

Let Δ_h be the usual nine point approximation to Δ , defined by

$$(4.1) \quad \begin{aligned} \Delta_h v(x, y) = & \frac{1}{6h^2} \{ 4v(x+h, y) + 4v(x, y+h) + 4v(x-h, y) + \\ & + 4v(x, y-h) + v(x+h, y+h) + v(x-h, y+h) + \\ & + v(x-h, y-h) + v(x+h, y-h) - 20v(x, y) \}, \end{aligned}$$

[cf. KANTOROVICH and KRYLOV, p. 210, 6]. This definition of Δ_h fixes the sets R_h and C_h^* of theorem 5.

At points of C_h^* , the pure second derivatives of Δ are approximated to within terms of order h^2 . Since Δ is invariant under rotations, these derivatives need not be restricted solely to u_{xx} and u_{yy} . That is, if (ξ, η) are new variables resulting from a rotation then

$$(4.2) \quad \Delta v \equiv v_{\xi\xi} + v_{\eta\eta}.$$

Consider the set of lines from $P \in C_h^*$ to the eight nearest grid points. From the definition of C_h^* , C must cut at least one of these lines. In the case where a horizontal line is cut by C we can approximate $v_{xx}(P)$ to $O(h^2)$ by an unbalanced four point formula which involves the value of v at the intersection of the line with C . For example, if P is the point (x, y) of Fig. 2, we have that

$$(4.3) \quad \begin{aligned} & |v_{xx}(x, y) - h^{-2} \left\{ \frac{\alpha-1}{\alpha+2} v(x+2h, y) + \frac{2(2-\alpha)}{\alpha+1} v(x+h, y) + \right. \\ & \left. + \frac{6}{\alpha(\alpha+1)(\alpha+2)} v(x-\alpha h, y) - \left[\frac{6+\alpha(7-\alpha^2)}{\alpha(\alpha+1)(\alpha+2)} \right] v(x, y) \right\}| \\ & \leq \frac{1}{12} M_4 h^2 + \frac{1}{6} M_5 h^3 \end{aligned}$$

and

$$(4.4) \quad |v_{yy}(x, y) - h^{-2}\{v(x, y+h) + v(x, y-h) - 2v(x, y)\}| \leq \frac{1}{12} M_4 h^2.$$

The analogous approximations are made if a vertical line is cut (or both are cut) by C . In this case we take Δ_h^* to be

$$(4.5) \quad \begin{aligned} \Delta_h^* v(x, y) &\equiv h^{-2} \left\{ v(x, y+h) + v(x, y-h) + \frac{\alpha-1}{\alpha+2} v(x+2h, y) + \right. \\ &+ \frac{2(2-\alpha)}{\alpha+1} v(x+h, y) + \frac{6}{\alpha(\alpha+1)(\alpha+2)} v(x-\alpha h, y) - \\ &- \left[2 + \frac{6+\alpha(7-\alpha^2)}{\alpha(\alpha+1)(\alpha+2)} \right] v(x, y) \}. \end{aligned}$$

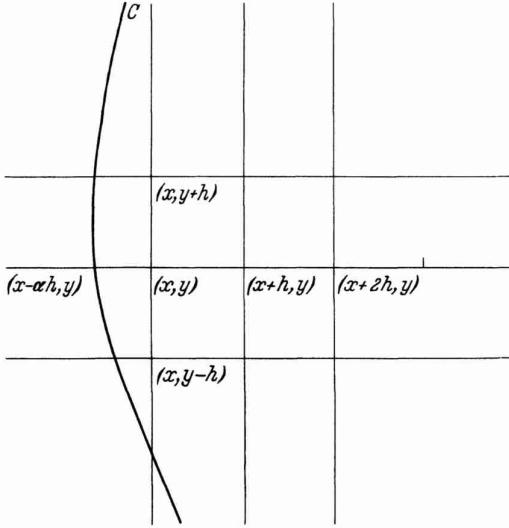


Fig. 2

In certain cases the boundary C may cut only a diagonal from P (see Fig. 3). In such a situation we consider the Laplace operator in the form (4.2) where

$$(4.6) \quad \xi = \frac{x+y}{\sqrt{2}}, \quad \eta = \frac{y-x}{\sqrt{2}},$$

which is just a rotation through 45 degrees. At least one of the terms $v_{\xi\xi}, v_{\eta\eta}$ will be approximated by a four point formula involving a point of C . For example if P is the point (x, y) of Fig. 3 we have

$$(4.7) \quad \begin{aligned} &|v_{\xi\xi}(x, y) - \frac{1}{2h^2} \left\{ \left(\frac{\alpha-1}{\alpha+2} \right) v(x+2h, y+2h) + \right. \\ &+ \frac{2(2-\alpha)}{\alpha+1} v(x+h, y+h) + \frac{6}{\alpha(\alpha+1)(\alpha+2)} v(x-\alpha h, y-\alpha h) - \\ &- \left. \left[\frac{6+\alpha(1-\alpha^2)}{\alpha(\alpha+1)(\alpha+2)} \right] v(x, y) \right\}| \leq \frac{M_4}{6} h^2 + \frac{\sqrt{2} M_5}{3} h^3, \end{aligned}$$

and

$$(4.8) \quad \left| v_{\eta\eta}(x, y) - \frac{1}{2h^2} \{v(x-h, y+h) + v(x+h, y-h) - 2v(x, y)\} \right| \leq \frac{M_4}{6} h^2.$$

Then Δ_h^* is defined by

$$(4.9) \quad \begin{aligned} \Delta_h^* v(x, y) &\equiv \frac{h^{-2}}{2} \left\{ v(x-h, y-h) + v(x+h, y-h) + \frac{\alpha-1}{\alpha+2} v(x+2h, y+2h) + \right. \\ &+ \frac{2(2-\alpha)}{\alpha+1} v(x+h, y+h) + \left[\frac{6}{\alpha(\alpha+1)(\alpha+2)} \right] v(x-\alpha h, y-\alpha h) - \\ &- \left. \left[2 + \frac{6+\alpha(7-\alpha^2)}{\alpha(\alpha+1)(\alpha+2)} \right] v(x, y) \right\}. \end{aligned}$$

Thus having defined Δ_h^* , the set C_h of theorem 5 is fixed.

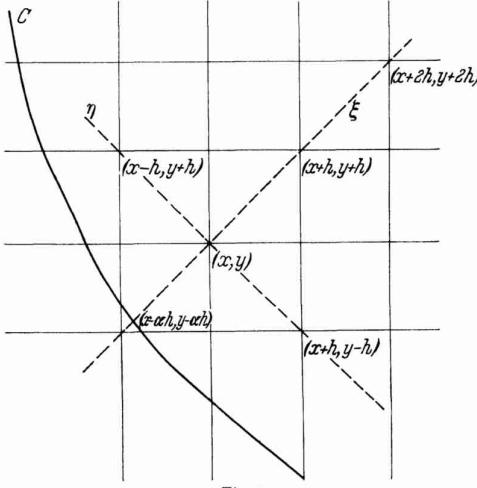


Fig. 3

In the finite difference problem (3.24) let

$$(4.10) \quad \varphi[F(P)] = \frac{h^2}{12} \Delta F(P),$$

and

$$(4.11) \quad H(P) = F(P).$$

For sufficiently small h , the hypotheses (a), (b) and (c) of theorem 5 are easily seen to be satisfied with $\delta = \frac{3}{4}$. The inequality (3.28) takes the form

$$(4.12) \quad \left| \Delta_h v - \left[\Delta v + \frac{h^2}{12} \Delta^2 v \right] \right| \leq \frac{4}{5!} M_6 h^4.$$

Since for $P \in C_h^*$, $\sigma(P, P) \geq \frac{2}{h^2}$ it follows that

$$(4.13) \quad \left| \frac{\Delta_h^* u(P) - F(P)}{\sigma(P, P)} \right| \leq \frac{h^4}{6} M_4 + \frac{\sqrt{2}}{3} h^5 M_5.$$

applying theorem 5 we see that

$$(4.14) \quad |\epsilon|_M \leq \left[\frac{M_6 d^2}{120} + \frac{2M_4}{3} + \frac{4\sqrt{2}}{3} M_5 h \right] h^4.$$

In posing finite difference analogues of (2.11) one can combine the techniques of sections 2 and 3. To illustrate this we define an $O(h^4)$ approximation to (2.1).

Let R_h , C_h^* and C_h be defined as in section 2. The set of points of R_h whose eight nearest neighbors are not all in R will be denoted by C'_h . At points of $R_h - C'_h$ we define Δ_h by (4.1), and at points of C'_h , Δ'_h is defined by (2.2). The operator Δ_h^* is taken to be an unbalanced $O(h^2)$ approximation to Δ , as in (4.5). We note that in this case only the pure second derivatives with respect to x and y are used.

We now pose the finite difference analogue of (2.1)

$$(4.15) \quad \begin{aligned} \Delta_h U(P) &= F(P) + \frac{h^2}{12} \Delta F(P), & P \in R_h - C'_h \\ \Delta'_h U(P) &= F(P), & P \in C'_h \\ \Delta_h^* U(P) &= F(P), & P \in C_h^* \\ U(P) &= f(P), & P \in C_h. \end{aligned}$$

This definition of Δ_h satisfies all hypotheses of theorem 5 except (3.28) with $n=4$. At points of C'_h (3.28) is valid for Δ'_h with $n=2$ and $\varphi(\Delta v) \equiv 0$. Hence theorem 5 is not directly applicable to (4.15). We can however modify the proof of theorem 5 to show that $|\varepsilon|_M = O(h^4)$.

Green's third identity (3.33) is valid for any vector $V(P)$, $P \in R_h + C_h^*$. If we apply (3.33) to the vector $V(P)=1$ for $P \in R_h$, $V(P)=0$ for $P \in C_h^*$, we obtain

$$(4.16) \quad 1 \geq h^2 \sum_{Q \in C'_h} G_h^*(P, Q) [-\Delta'_h V(Q)] + h^2 \sum_{Q \in R_h - C'_h} G_h^*(P, Q) [-\Delta_h V(Q)].$$

Noting that the second term on the right hand side of (4.16) is non-negative and that $-\Delta_h V(Q)=2h^{-2}$, for $Q \in C'_h$, we have that

$$(4.17) \quad \sum_{Q \in C'_h} G_h^*(P, Q) \leq \frac{1}{2}.$$

Equation (3.33) can be written in the form

$$(4.18) \quad \begin{aligned} \varepsilon(P) &= h^2 \sum_{Q \in C'_h} G_h^*(P, Q) [-\Delta'_h \varepsilon(Q)] + h^2 \sum_{Q \in R_h - C'_h} G_h^*(P, Q) [-\Delta_h \varepsilon(Q)] + \\ &\quad + \sum_{Q \in C_h^*} G_h^*(P, Q) \varepsilon(Q). \end{aligned}$$

Thus we see that

$$(4.19) \quad |\varepsilon(P)| \leq \frac{h^2}{2} |\Delta'_h \varepsilon|_M + \frac{d^2}{16} |\Delta_h \varepsilon|_M + \left| \frac{\Delta_h^* \varepsilon(P)}{\sigma(P, P)} \right|_M + \frac{3}{4} |\varepsilon|_M.$$

In this case it is easy to see that

$$(4.20) \quad \left| \frac{\Delta_h^* \varepsilon(P)}{\sigma(P, P)} \right|_M \leq \frac{M_4}{12} h^4 + \frac{M_5}{12} h^5.$$

From (2.3), (4.12) and (4.20) we obtain the estimate

$$(4.21) \quad |\varepsilon|_M \leq \left[\frac{M_6 d^2}{120} + \frac{2M_4}{3} + \frac{M_5}{12} h \right] h^4.$$

Note that the estimates (4.14) and (4.21) differ only in the higher order terms.

References

- [1] COLLATZ, L.: Bemerkungen zur Fehlerabschätzung für das Differenzenverfahren bei partiellen Differentialgleichungen. *Z. angew. Math. Mech.* **13**, 56–57 (1933).
- [2] — Numerical treatment of differential equations, 3rd ed. Berlin-Göttingen-Heidelberg: Springer 1960.
- [3] COURANT, R., K. FRIEDRICH and H. LEWY: Über die partiellen Differenzen-gleichungen der mathematischen Physik. *Math. Ann.* **100**, 32–74 (1928).
- [4] FORSYTHE, G., and W. WASOW: Finite-difference methods for partial differential equations. New York: Wiley 1960.
- [5] GERSCHGORIN, S.: Fehlerabschätzung für das Differenzenverfahren zur Lösung partieller Differentialgleichungen. *Z. angew. Math. Mech.* **10**, 373–382 (1930).
- [6] KANTOROVICH, L., and V. KRYLOV: Approximate Methods of Higher Analysis. Netherlands: Noordhoff Ltd. 1958.
- [7] LAASONEN, P.: On the degree of convergence of discrete approximations for the solutions of the Dirichlet problem. *Ann. Acad. Sci. Fenn. A. I.* **246**, 1–19 (1957).
- [8] — On the solution of Poisson's difference equation. *J. Assoc. Comput. Mach.* **5**, 370–382 (1958).
- [9] MOTZKIN, T., and W. WASOW: On the approximation of linear elliptic differential equations by difference equations with positive coefficients. *J. Math. Phys.* **31**, 253–259 (1953).
- [10] SHORTLEY, G., and R. WELLER: The numerical solution of Laplace's equation. *J. Appl. Phys.* **9**, 334–348 (1938).
- [11] WALSH, J., and D. YOUNG: On the accuracy of the numerical solution of the Dirichlet problem by finite differences. *J. Res. Nat. Bur. Stand.* **51**, 343–363 (1953).
- [12] — — On the degree of convergence of solutions of difference equations to the solution of the Dirichlet problem. *J. Math. Phys.* **33**, 80–93 (1954).
- [13] WASOW, W.: On the truncation error in the solution of Laplace's equation by finite differences. *J. Res. Nat. Bur. Stand.* **48**, 345–348 (1952).
- [14] — The accuracy of difference approximations to plane Dirichlet problems with piecewise analytic boundary values. *Quart. Appl. Math.* **15**, 53–63 (1957).

Institute for Fluid Dynamics
and Applied Mathematics
University of Maryland
College Park, Maryland

(Received September 18, 1961)

d-Minimal Surfaces Spanning Polygons in E_n by Solving a Linear System

By

WALTER L. WILSON, JR.

1.0. Introduction

In [1], p. 368, §4.5, the author described a method involving solution of a linear system for computing approximations to conformal maps of regions bounded by polygons in a plane. Here, we extend the method to compute polyhedral approximations to minimal surfaces spanning a simple closed polygon Γ in Euclidean n -space.

Notation and operators used here are defined in the above cited paper.

Let Γ be a Jordan curve in the form of a polygon in E_n given by the vector function

$$\Gamma: g(t) \quad 0 \leq t \leq l(\Gamma)$$

where $g(0)=g(l(\Gamma))$ is a vertex of Γ , t is arc length on Γ and the σ vertices of Γ are $\{g(t_k) | 0 = t_0 < t_1 < t_2 < \dots < t_\sigma = l(\Gamma)\}$. Specifically, let

$$(1.1) \quad \begin{aligned} g(t) &= g(t_k) + [g(t_{k+1}) - g(t_k)] \frac{t - t_k}{t_{k+1} - t_k} \\ &= g_k + \gamma_k t, \quad t_k \leq t \leq t_{k+1} \end{aligned}$$

where $k=0(1)\sigma-1$. These are σ line segments contained respectively in the lines

$$(1.2) \quad L_k(t) = g_k + \gamma_k t, \quad -\infty < t < +\infty.$$

Definition. $\Gamma_0 : \{b_i(t) | i=1(1)N\}$ is a discrete parametrization of Γ if (1) three elements of Γ_0 , say b_{N-2} , b_{N-1} and b_N , are distinct vertices of Γ and are assumed *fixed*, (2) each b_j not one of the fixed points is *assigned* to a side of Γ and several (or no) elements of Γ_0 may be assigned to any side of Γ . The position of b_j on the side to which it is assigned is *not* prescribed. The orientation of elements of Γ_0 on Γ is prescribed*.

* The functional $A(\bar{\Gamma})$ in (2.1) below was derived assuming Γ is a topological image of a plane simple closed polygon with N vertices — the vertices corresponding to similarly oriented elements of any discrete parametrization of Γ . Therefore if m elements of Γ_0 are between two fixed points then all or any adjacent subset of these may be assigned to any side of Γ between the fixed points assuming the remainder of the set of m points are assigned to similarly oriented sides of Γ .

2.0. The Linear System

In [I] the author derived an analog

$$(2.1) \quad A(\bar{\Gamma}) = \sum_{i,j=1}^N F_{ij}(b_i - b_j)^2$$

to the Douglas functional. (These are equations (4.5) and (1.1) respectively in [I].) $A(\bar{\Gamma})$ is defined on the discrete parametrizations of Γ . If $\Gamma_p : \{b_i | i=1(1)N\}$ is a discrete parametrization of Γ which locally minimizes $A(\bar{\Gamma})$ in the set of all discrete parametrizations containing the same three fixed points and the same number of points on each arc of Γ between fixed points then Γ_p is called minimizing. The unique d -harmonic surface Σ_p spanning the polygon Γ_p is a d -minimal surface. The vertices of Σ_p are defined by the Laplacian operator L in equation (3.10) in [I]. We now define the linear system which is used to determine a minimizing Γ_p . The particular Γ_p obtained depends on the choice of fixed points and assignments in Γ_0 .

Let Γ_0 be a discrete parametrization of Γ as defined above. Then the positive definite quadratic form $A(\Gamma_0) = A(t)$ is minimized only if the non-fixed elements of Γ_0 are chosen so that $\frac{dA}{dt} = \frac{dA}{db_m} \cdot \frac{db_m}{dt} = 0$. That is, if

$$(2.2) \quad \sum_{\beta} F_{m\beta} (b_m - b_{\beta}) \cdot \dot{b}_{\beta} = - \sum_{\alpha=1}^n C_{m\alpha} \sum_{\beta=1}^N F_{m\beta} b_{\beta\alpha} = 0^*, \quad m=1(1)\underline{N-3},$$

where $C_{m\alpha}$ is the α -th component of the vector γ defined in (1.1) for the side of Γ to which b_m was assigned in Γ_0 . These are $N-3$ equations in $n(N-3)$ unknowns — the components of elements of Γ_0 .

If b_{β} is assigned to the side of Γ contained in the line $L_k(t)$ then we make the substitution $b_{\beta\alpha} = (g_k + \gamma_k a_{\beta})_{\alpha} \equiv g_{\beta\alpha} + \gamma_{\beta\alpha} a_{\beta}$ where a_{β} is to be determined. Then, (2.2) may be written in the form

$$(2.3) \quad \sum_{\alpha=1}^n \gamma_{m\alpha} \sum_{\beta=1}^N F_{m\beta} (g_{\beta\alpha} + \gamma_{\beta\alpha} a_{\beta}) = 0, \quad m=1(1)\underline{N-3}$$

where a_{N-2} , a_{N-1} and a_N are fixed.

The system of $N-3$ equations (2.3) is solved for $\{a_{\beta} | \beta=1(1)\underline{N-3}\}$ which determine a set $\bar{\Gamma}_p$ of points some of which may lie on the extensions of the sides to which they were assigned in Γ_0 . If any element of $\bar{\Gamma}_p$ is on the extension of the assigned side then a reassignment of elements giving a new Γ_0 is necessary. When elements of $\bar{\Gamma}_p$ all lie on the assigned sides of Γ then $\bar{\Gamma}_p$ is Γ_p — a minimizing discrete parametrization of Γ . The d -minimal surface is then determined as indicated above.

References

- [1] WILSON JR., W. L.: On Discrete Dirichlet and Plateau Problems. Numerische Mathematik 3, 359—373 (1961).

Department of Mathematics
University of Alabama
University, Alabama

(Received January 11, 1962)

* The first equality is obtained using the definition of scalar product in E_n and Corollary 4.1 in [I].

*1.2. FOURTH-ORDER FINITE DIFFERENCE ANALOGUES OF THE
DIRICHLET PROBLEM FOR POISSON'S EQUATION IN THREE AND FOUR
DIMENSIONS*

**1.2 Fourth-order finite difference analogues of the Dirichlet
problem for Poisson's equation in three and four dimensions**

Fourth-order finite difference analogues of the Dirichlet problem for Poisson's equation in three and four dimensions [6]

Fourth-Order Finite Difference Analogues of the Dirichlet Problem for Poisson's Equation in Three and Four Dimensions

By James H. Bramble

I. Introduction. In a recent paper [1] Bramble and Hubbard formulated finite difference analogues of the Dirichlet problem for Poisson's equation in the plane which were $O(h^4)$, h being the mesh width. Subsequently in [2] they gave a general theorem on error estimation for a class of finite difference analogues to the Dirichlet problem for some general uniformly elliptic equations in N -dimensions. Some examples in the plane are formulated there.

In dealing with Poisson's equation by finite difference methods a very large system of linear equations must be solved. Even with modern high-speed computers this number may be prohibitively large if the desired accuracy is to be obtained. Several methods commonly used in plane problems are $O(h^2)$ and their direct analogues in higher dimensions can also be shown to be $O(h^2)$. Thus, if (for smooth problems) in three dimensions a fourth-order method were used instead of a second-order one, it might be expected that a considerably smaller system would yield comparable accuracy. Consequently, if a higher order method were used some problems might move to within the range of practical feasibility.

In this paper analogues to the Dirichlet problem for Poisson's equation in three and four dimensions are given. These analogues are shown to be $O(h^4)$ as $h \rightarrow 0$.

2. Three Dimensional Case. Let R be a bounded region with boundary C in three dimensions. In the usual manner the space is subdivided into cubes of side h with faces parallel to the (x, y) , (x, z) , and (y, z) planes. The corner points of these cubes will be called mesh points. The set R_h will consist of those mesh points P in R whose 18 nearest neighboring mesh points, and the lines joining them to P , are in R . The set C_h^{**} will denote those mesh points $P \in R - R_h$ whose 6 nearest neighbors and the lines joining them to P are in R . The set of mesh points in $R - R_h - C_h^{**}$ will be called C_h^* . If P is in C_h^* then at least one line joining P to one of its 6 nearest neighbors say $(x - h, y, z)$ is cut by C . Thus for some α , $0 < \alpha \leq 1$, $(x - \alpha h, y, z)$ is on C . Such a point will be said to lie in C_h . Similarly, one of the neighbors of (x, y, z) in the y and z directions may not be in R . These points will also then lie in C_h . The totality of such "neighbors" of points of C_h^* will make up the set C_h . The mesh size is assumed so small that if (x, y, z) is in C_h^* and $(x \pm \alpha h, y, z)$ is in C_h then $(x \pm h, y, z)$ and $(x \pm 2h, y, z)$ are in $R + C$ where either the plus sign is taken at each of the points or the minus sign is taken. Analogous statements are assumed for the y and z directions.

With the preceding sets defined we are in a position to formulate the finite

Received April 30, 1962. This research was supported in part by the United States Air Force through the Air Force Office of Scientific Research of the Air Research and Development Command.

difference problem. The exact problem to be approximated is

$$(2.1) \quad \begin{aligned} \Delta u &= F \quad \text{in } R \\ u &= f \quad \text{on } C, \end{aligned}$$

where Δ is the Laplace operator, $\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$, and F and f are sufficiently smooth functions defined in R and on C respectively.

At a point (x, y, z) of R_h we approximate Δu by

$$(2.2) \quad \begin{aligned} \Delta_h u(x, y, z) &= \frac{1}{6h^2} \{ 2[u(x+h, y, z) + u(x-h, y, z) + u(x, y+h, z) \\ &\quad + u(x, y-h, z) + u(x, y, z+h) + u(x, y, z-h)] + u(x+h, y+h, z) \\ &\quad + u(x+h, y-h, z) + u(x-h, y+h, z) + u(x-h, y-h, z) \\ &\quad + u(x+h, y, z+h) + u(x+h, y, z-h) + u(x-h, y, z+h) \\ &\quad + u(x-h, y, z-h) + u(x, y+h, z+h) + u(x, y+h, z-h) \\ &\quad + u(x, y-h, z+h) + u(x, y-h, z-h) - 24u(x, y, z) \}. \end{aligned}$$

By approximating $\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}$ by means of

$$\Delta_{h(x,y)}^+ u \equiv \frac{1}{h^2} [u(x+h, y, z) + u(x-h, y, z) \\ + u(x, y+h, z) + u(x, y-h, z) - 4u(x, y)]$$

and

$$\Delta_{h(x,y)}^\times u \equiv \frac{1}{2h^2} [u(x+h, y+h, z) + u(x+h, y-h, z) \\ + u(x-h, y+h, z) + u(x-h, y-h, z) - 4u(x, y)]$$

with similar considerations in the (x, z) and (y, z) planes it is easy to see that $\Delta_h u$ given by (2.2) is just

$$\Delta_h u = \frac{1}{2} \{ \Delta_{h(x,y)}^+ u + \Delta_{h(x,z)}^\times u + \Delta_{h(y,z)}^\times u \},$$

where

$$\Delta_{h(x,y)} u = \frac{1}{3} \Delta_{h(x,y)}^+ u + \frac{2}{3} \Delta_{h(x,y)}^\times u.$$

From this structure it is not difficult to see that

$$(2.3) \quad \left| \Delta_h u - \left(\Delta u + \frac{h^2}{6} \Delta^2 u \right) \right| \leq \frac{M_6}{10} h^4,$$

where M_i is a uniform bound for any i th partial derivative of u in $R + C$.

At a point of C_h^{**} we define

$$(2.4) \quad \begin{aligned} \Delta_h^{**} u &= \frac{1}{h^2} [u(x+h, y, z) + u(x-h, y, z) + u(x, y+h, z) \\ &\quad + u(x, y-h, z) + u(x, y, z+h) + u(x, y, z-h) - 6u(x, y, z)]. \end{aligned}$$

The inequality

$$(2.5) \quad |\Delta_h^{**}u - \Delta u| \leq \frac{M_4}{4} h^2$$

holds at points of C_h^{**} .

At a point of C_h^* the pure second partial derivatives are approximated to $O(h^2)$, if necessary by an unbalanced four-point formula which includes a point of C_h . (See [1]). For example, if (x, y, z) is in C_h^* and $(x - \alpha h, y, z) \in C_h$ then we have that

$$(2.6) \quad \left| u_{xx}(x, y, z) - h^{-2} \left\{ \left(\frac{\alpha - 1}{\alpha + 2} \right) u(x + 2h, y, z) + \frac{2(2 - \alpha)}{\alpha + 1} u(x + h, y, z) \right. \right. \\ \left. \left. + \frac{6}{\alpha(\alpha + 1)(\alpha + 2)} u(x - \alpha h, y, z) - \frac{3 - \alpha}{\alpha} u(x, y, z) \right\} \right| \leq \frac{M_4}{6} h^2 + \frac{M_5}{6} h^3.$$

If the neighbors of (x, y, z) in the y and z direction are in R then we define on C_h^*

$$(2.7) \quad \Delta_h^* u = \frac{1}{h^2} \left\{ \frac{\alpha - 1}{\alpha + 2} u(x + 2h, y, z) + \frac{2(2 - \alpha)}{\alpha + 1} u(x + h, y, z) \right. \\ \left. + \frac{6}{\alpha(\alpha + 1)(\alpha + 2)} u(x - \alpha h, y, z) + u(x, y + h, z) + u(x, y - h, z) \right. \\ \left. + u(x, y, z + h) + u(x, y, z - h) - 3 \frac{\alpha + 1}{\alpha} u(x, y, z) \right\}.$$

It is easy to see that

$$(2.8) \quad |\Delta_h^* u - \Delta u| \leq \frac{M_4}{2} h^2 + \frac{M_5}{2} h^3.$$

(See [1]). At each point of C_h^*, Δ_h^* is defined analogously, using the four-point approximation when needed.

As our approximating problem we consider the following linear system

$$(2.9) \quad \begin{aligned} \Delta_h U(P) &= F(P) + \frac{h^2}{6} \Delta F(P), & P \in R_h \\ \Delta_h^{**} U(P) &= F(P), & P \in C_h^{**} \\ \Delta_h^* U(P) &= F(P), & P \in C_h^* \\ U(P) &= f(P), & P \in C_h \end{aligned}$$

for the determination of U at the points of $R_h + C_h^* + C_h^{**}$. The system (2.9) is not of positive type (see e.g., Forsythe and Wasow [3]) however, it does have the properties of "interior positivity" and "strict diagonal dominance". (See [1]). As in the case of the plane [1] these conditions will suffice to show that if

$$\epsilon(P) = u(P) - U(P)$$

then,

$$(2.10) \quad |\epsilon(P)|_M = O(h^4),$$

where the subscript M denotes the maximum over all $P \in R_h + C_h^* + C_h^{**}$.

The method of proof follows closely that given by Bramble and Hubbard in [1].

Let $G_h(P, Q)$ be the “Green’s function” defined by

$$(2.11) \quad \begin{aligned} \Delta_{h,P} G_h(P, Q) &= -h^{-3}\delta(P, Q), & P \in R_h \\ \Delta_{h,P}^{**} G_h(P, Q) &= -h^{-3}\delta(P, Q), & P \in C_h^{**} \\ G_h(P, Q) &= \delta(P, Q), & P \in C_h^*, \end{aligned}$$

for each $Q \in R_h + C_h^{**} + C_h^*$. By the usual means it may be shown that $G_h(P, Q) \geq 0$. It may be easily verified that any mesh function $V(P)$ defined for $P \in R_h + C_h^{**} + C_h^*$ satisfies the identity

$$(2.12) \quad \begin{aligned} V(P) &= h^3 \sum_{Q \in R_h} G_h(P, Q) [-\Delta_h V(Q)] \\ &\quad + h^3 \sum_{Q \in C_h^{**}} G_h(P, Q) [-\Delta_h^{**} V(Q)] + \sum_{Q \in C_h^*} G_h(P, Q) V(Q). \end{aligned}$$

In particular, if we take

$$V(P) = 1, \quad P \in R_h + C_h^{**}$$

and

$$V(P) = 0 \quad \text{for } P \in C_h^*$$

then we obtain

$$(2.13) \quad 1 \geq h \sum_{Q \in C_h^{**}} G_h(P, Q).$$

Because of the interior positivity of (2.9) it can be seen that if

$$(2.14) \quad \begin{aligned} \Delta_h W(P) &\geq 0, & P \in R_h \\ \Delta_h^{**} W(P) &\geq 0, & P \in C_h^{**} \end{aligned}$$

then

$$W(Q) \leq \max_{P \in C_h^*} W(P), \quad Q \in R_h + C_h^{**} + C_h^*.$$

This is just an interior maximum principle. By making use of (2.14) it can be readily shown as was done in [1] that

$$(2.15) \quad h^3 \sum_{Q \in R_h} G_h(P, Q) \leq \frac{d^2}{24},$$

where d is the diameter of R .

Let us now apply (2.12) to $\epsilon(P)$. Making use of the (2.13) and (2.15) and the fact that $G_h(P, Q) \geq 0$ we have that

$$(2.16) \quad \begin{aligned} |\epsilon(P)| &\leq \frac{d^2}{24} [\max_{Q \in R_h} |\Delta_h \epsilon(Q)|] \\ &\quad + h^2 [\max_{Q \in C_h^{**}} |\Delta_h^{**} \epsilon(Q)|] + \sum_{Q \in C_h^*} G_h(P, Q) |\epsilon(Q)|. \end{aligned}$$

From (2.3), (2.5) and (2.9) it follows that

$$(2.17) \quad |\epsilon(P)| \leq \left[\frac{d^2}{240} M_6 + \frac{M_4}{4} \right] h^4 + \sum_{Q \in C_h^*} G_h(P, Q) |\epsilon(Q)|.$$

Now if we use the definition of Δ_h^* ((2.7) or the appropriate analogue of (2.7)) and the fact that $\epsilon(P) = 0$, $P \in C_h$ we find that

$$(2.18) \quad |\epsilon(P)| \leq \frac{5}{6} |\epsilon|_M + \frac{h^2}{6} |\Delta_h^* \epsilon(P)|, \quad P \in C_h^*.$$

But, from (2.9) and (2.8)

$$(2.19) \quad |\Delta_h^* \epsilon(P)| = |\Delta_h^* u(P) - \Delta u(P)| \leq \frac{M_4}{2} h^2 + \frac{M_5}{2} h^3.$$

Hence, combining (2.18) and (2.19), we have

$$(2.20) \quad |\epsilon(P)| \leq \frac{5}{6} |\epsilon|_M + \left(\frac{M_4}{12} + \frac{M_5}{12} h \right) h^4, \quad P \in C_h^*$$

Now, since $\sum_{Q \in C_h^*} G_h(P, Q) \leq 1$, we have from (2.17) and (2.20)

$$(2.21) \quad |\epsilon(P)| \leq \left[\frac{d^2}{240} M_6 + \frac{h}{12} M_5 + \frac{1}{3} M_4 \right] h^4 + \frac{5}{6} |\epsilon|_M.$$

Since the right hand side of (2.21) is independent of P we conclude that

$$(2.22) \quad |\epsilon|_M \leq \left[\frac{d^2}{40} M_6 + \frac{h}{2} M_5 + 2M_4 \right] h^4.$$

This shows that the overall error produced in replacing problem (2.1) by (2.9) is $O(h^4)$.

3. Higher Dimensional Problems. Let us assume that the sets R_h , C_h^{**} , C_h^* , and C_h have been defined in a manner analogous to that of the preceding section.

In formulating $O(h^4)$ analogues to (2.1) in N dimensions we could use the direct analogues of (2.4) and (2.7). The problem reduces to that of finding the analogue of (2.2) at a point $(x_1, \dots, x_N) \in R_h$.

Let us proceed as described after (2.2) and consider various two dimensional planes through P and the Laplace difference operators in these respective planes. It turns out that if we define

$$(3.1) \quad \Delta_h u = \frac{1}{N-1} \left\{ \sum_{\substack{i=1, \dots, N-1 \\ j=2, \dots, N \\ j>i}} \left[\frac{4-N}{3} \Delta_{h(x_i, x_j)}^+ u + \frac{N-1}{3} \Delta_{h(x_i, x_j)}^\times u \right] \right\}$$

then the relation

$$(3.2) \quad \left| \Delta_h u - \left(\Delta u + \frac{N-1}{6} h^2 \Delta^2 u \right) \right| = O(h^4)$$

is valid in R_h . In order that (3.1) be of positive type (see, e.g., [3]) it is necessary that $N \leq 4$. For $N = 2$, (3.1) is the usual nine-point formula in the plane and for $N = 3$, (3.1) reduces to (2.2) of Section II. The case $N = 4$ is interesting in that

$$(3.3) \quad \Delta_h u = \frac{1}{3} \left\{ \sum_{\substack{i=1, \dots, 3 \\ j=2, \dots, 4 \\ j>i}} \Delta_{h(x_i, x_j)}^\times u \right\},$$

and the term involving $\Delta_{h(x_i, x_j)}^+$ drops out. Thus, in three dimensions (2.2) is a 19-point formula while in four dimensions (3.3) is a 25-point formula, both being $O(h^4)$ expressions.

For $N > 4$, although (3.3) is $O(h^4)$ locally, it is not of positive type. Thus, the method of Section III is not applicable and it is not clear that an overall $O(h^4)$ estimate for the truncation error $\epsilon(P)$ would result. It seems that a different approach might be more desirable for $N > 4$.

It should be noted that if, at points of C_h^* , the direct analogue of the Shortley and Weller approximation [4] is used, then an overall $O(h^3)$ estimate for the truncation error could be obtained in two, three or four dimensions.

John Jay Hopkins Laboratory for Pure and Applied Science
General Atomic Division of General Dynamics Corporation
San Diego, California

1. J. H. BRAMBLE, & B. E. HUBBARD, "On the formulation of finite difference analogues of the Dirichlet problem for Poisson's equation," *Numer. Math.*, v. 4, 1962, p. 313-327.
2. J. H. BRAMBLE, & B. E. HUBBARD, "A theorem on error estimation for finite difference analogues of the Dirichlet problem for elliptic equations," (to appear).
3. G. FORSYTHE, & W. WASOW, *Finite Difference Methods for Partial Difference Equations*, New York, Wiley, 1960.
4. G. SHORTLEY & R. WELLER, "The numerical solution of Laplace's equation," *J. Appl. Phys.*, v. 9, 1938, p. 334-348.

1.3 On a finite difference analogue of an elliptic boundary problem which is neither diagonally dominant nor of non-negative type

On a finite difference analogue of an elliptic boundary problem which is neither diagonally dominant nor of non-negative type [12]

ON A FINITE DIFFERENCE ANALOGUE OF AN ELLIPTIC
BOUNDARY PROBLEM WHICH IS NEITHER DIAGONALLY
DOMINANT NOR OF NON-NEGATIVE TYPE

By J. H. BRAMBLE* AND B. E. HUBBARD†

1. Introduction. In the usual study of the discretization error resulting from approximating boundary problems for elliptic equations by finite difference methods the maximum principle plays a central role. In 1930 S. Gershgorin (14) gave a method for estimating the order of convergence of the solution to a certain class of finite difference analogues to the solution of the Dirichlet problem for elliptic equations. The matrix of the resulting system of simultaneous linear equations possesses the property of diagonal dominance, i.e. the sum of the absolute values of the off-diagonal elements in each row does not exceed the magnitude of the diagonal element. Furthermore it satisfies the condition that the diagonal elements are all positive and the off-diagonal elements are non-positive (2.10). Following this, others (3), (8), (9), (12), (20), (21) have extended the results of Gershgorin within the framework of these conditions.

Recently the authors (5), (6) gave a theorem on the formulation of finite difference analogues of the Dirichlet problem for elliptic equations in which the properties (2.10) were relaxed near the boundary. As examples of the theorem certain higher order finite difference analogues were discussed and shown to have convergence properties previously unexpected. The question of a maximum principle for the entire problem was circumvented however and has been dealt with only recently by M. Rockoff (18). Hence it is clear that the sufficient conditions (2.10) are not necessary.

One purpose of this paper is to give a specific example in which conditions (2.10) are violated at every interior point but for which a maximum principle still holds. This is done in section 2 where an $O(h^4)$ approximation to the two point boundary problem (2.1) is studied (h is the mesh constant). The interesting fact that the approximation to the operator near the boundary need be only $O(h^2)$ without destroying the overall accuracy of the problem is shown to be true.

In section 3 we study the convergence of certain common iterative techniques for the solution of the resulting linear systems. It is shown in this case that the method of simultaneous displacements (Jacobi) diverges for sufficiently small h while the symmetric Gauss-Seidel method converges.

For background material and an excellent bibliography the reader is referred to the book by G. Forsythe and W. Wasow (12).

2. An $O(h^4)$ Finite Difference Analogue for the Dirichlet Problem. We shall concern ourselves with the boundary value problem

$$(2.1) \quad \begin{aligned} Ly &\equiv -y''(x) + q(x)y(x) = f(x), & x \in [0, 1] \\ y(0) &= y(1) = 0 \end{aligned}$$

* Supported in part by the National Science Foundation under grant NSF GP-3.

† Supported in part by the National Science Foundation under grant NSF GP-2284

where $q(x) \geq 0$ and both q and f possess four continuous derivatives. A more general uniformly elliptic differential problem can be reduced to one of the type (2.1) by well known techniques e. f. [10, page 292].

Let the interval $[0, 1]$ be divided into N equal parts. The distance $h = 1/N$ between two successive divisions will be called the "mesh size" and the point set $0, h, \dots, Nh = 1$ will be termed "mesh points." We define the following finite difference operators:

$$(2.2) \quad \begin{aligned} \Delta_x V(x) &\equiv h^{-2}\{V(x-h) - 2V(x) + V(x+h)\} \\ \left(1 - \frac{h^2}{12} \Delta_x\right) \Delta_x V(x) &\equiv \frac{h^{-2}}{12} \{-V(x-2h) + 16V(x-h) \right. \\ &\quad \left. - 30V(x) + 16V(x+h) - V(x+2h)\}. \end{aligned}$$

It is easily shown that for any function $V(x)$ with bounded sixth derivatives that

$$(2.3) \quad \begin{aligned} |\Delta_x V(x) - V''(x)| &< \frac{h^2}{12} \left| \frac{d^4 V}{dx^4} \right|_M \\ \left| \left(I - \frac{h^2}{12} \Delta_x \right) \Delta_x V(x) - V''(x) \right| &\leq \frac{h^4}{90} \left| \frac{d^6 V}{dx^6} \right|_M, \end{aligned}$$

where we have adopted the notation

$$(2.4) \quad f_M = \max f.$$

The $O(h^4)$ finite difference analogue of (2.1) which will be considered here is given by

$$(2.5) \quad \begin{aligned} Y(x) &= 0, \quad x = 0, 1 \\ -\Delta_x Y(x) + q(x)Y(x) &= f(x), \quad x = h, (N-1)h \\ -\left(I - \frac{h^2}{12} \Delta_x\right) \Delta_x Y(x) + q(x)Y(x) &= f(x), \quad x = 2h, 3h, \dots, (N-2)h \end{aligned}$$

Let us now define $Y(mh) = y_m$, $m = 0, 1, \dots, N$, the matrix

$$(2.6) \quad A \equiv (a_{ij}) \equiv \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & \cdots & 0 \\ -h^{-2} [2h^{-2} + q(h)] & -h^{-2} & 0 & 0 & 0 & \cdots & 0 \\ \frac{h^{-2}}{12} & \frac{-4}{3} h^{-2} & \left[\frac{5}{2} h^{-2} + q(2h) \right] & \frac{-4}{3} h^{-2} & \frac{h^{-2}}{12} & \cdots & 0 \\ 0 & \frac{h^{-2}}{12} & \frac{-4}{3} h^{-2} & \left[\frac{5}{2} h^{-2} + q(3h) \right] & \frac{-4}{3} h^{-2} & \cdots & 0 \\ \vdots & & \ddots & \ddots & \ddots & \ddots & \vdots \\ & & & & & & \frac{1}{12} h^{-2} \\ & & & & & & -h^{-2} \\ 0 & & & & & & 1 \end{bmatrix}$$

and the vector f with components, $f_0 = f_M = 0, f_m = f(mh), m = 1, \dots, N - 1$. The system of simultaneous linear equations in $N + 1$ unknowns (2.5) can now be written in the form

$$(2.7) \quad \sum_{j=0}^N a_{ij} y_j = f_i, \quad i = 0, 1, \dots, N.$$

Let a^{ij} be the general element of A^{-1} , i.e.

$$(2.8) \quad \sum_k a^{ik} a_{kj} = \sum_k a_{ik} a^{kj} = \delta_{ij} \equiv \begin{cases} 1, & i = j \\ 0, & i \neq j, \end{cases}$$

and if A^{-1} exists then (2.7) can be inverted as

$$(2.9) \quad y_i = \sum_{j,k} a^{ij} a_{kj} y_j = \sum_k a^{ij} f_k.$$

This is sometimes called Poisson's formula in analogy to the corresponding formula in the continuous case. We shall show that if h is taken small enough then $a^{ij} \geq 0$. Such a matrix is said to be 'non-negative' and the notation $A^{-1} \geq 0$ will be used. Before proving this theorem we shall present certain results from matrix theory which aid in the proof.

Definition. A matrix B with elements b_{ij} is said to be monotone if $x \geq 0$ for any vector x such that $Bx \geq 0$.

Theorem 2.1. *The matrix B is monotone if and only if B is non-singular and $B^{-1} \geq 0$.*

This theorem is well known, however, for the sake of completeness, the following proof is included.

If B is monotone and Y belongs to the null space then $B(\pm Y) \geq 0$ and hence $\pm Y \geq 0$. Thus B is non-singular and if z is any column of B^{-1} then $Bz \geq 0$ and hence $z \geq 0$. Conversely if $Bx \geq 0$ then $x = B^{-1}Bx \geq 0$.

Since each of the equivalent properties of monotone matrices is difficult to establish by inspecting the matrix B itself we are interested in sufficient conditions which can be easily verified. For example, each member of the following class of matrices is known to be monotone

Definition. A matrix B is said to be of "positive type" if the following conditions are satisfied

- a) $b_{jj} \leq 0 \quad i \neq j$
- b) $\sum_i b_{ji} \geq 0 \quad$ for all j , and further there exists a non-empty subset $J(B)$ of the integers $0, 1, 2, \dots, N$ such that $\sum_i b_{ji} > 0$ for all $j \in J(B)$
- (2.10) c) for $i \notin J(B)$ there exists $j \in J(B)$ and a sequence of non-zero elements of B the form

$$b_{ik_1}, b_{k_1 k_2}, \dots, b_{k_r j}.$$

We note that a matrix B is of positive type if (c) in (2.10) is replaced by the condition

(c') B is irreducible.

Theorem 2.2 *If B is of positive type then B is monotone*

Remark With theorem 2.2 we see that the class of positive type matrices belongs to the class of M -matrices (i.e. those monotone matrices which satisfy

(2.10(a)) discussed by Ostrowski (16). The so-called Minkowski matrices are just those positive type matrices for which $J = (0, \dots, N)$ and are thus contained in the class of positive type matrices. There are several reasons for introducing this intermediate class of matrices. The first is that there is no simple criterion for a matrix to be an M -matrix. Secondly, the Minkowski matrices are a bit too restrictive although easy to recognize. The positive type matrices include a large number of the interesting cases and are easily recognizable.

We now prove theorem 2.2

Proof. Conditions a) and b) of (2.10) imply that $b_{ii} \geq 0$ for every i . If $b_{ii} = 0$ then $b_{ik} = 0$ for all k and condition c) is violated. Hence $b_{ii} > 0$ for every i . Assume that $Bx \geq 0$. We shall show that $x \geq 0$. Rewriting the above inequality we obtain

$$(2.11) \quad x_i \geq \sum_{k,k \neq i} \left| \frac{b_{ik}}{b_{ii}} \right| x_k$$

where as a consequence of (2.10) we see that

$$(2.12) \quad \sum_{k,k \neq i} \left| \frac{b_{ik}}{b_{ii}} \right| \begin{cases} < 1, & i \in J(B) \\ = 1, & i \notin J(B). \end{cases}$$

Now assume that $x_i = \bar{x}$ is a negative minimum. Then if $i \in J(B)$ we see from (2.11) and (2.12) that

$$(2.13) \quad \bar{x} = x_i \geq \left\{ \sum_{k,k \neq i} \left| \frac{b_{ik}}{b_{ii}} \right| \right\} \bar{x} > \bar{x}$$

which gives a contradiction. On the other hand if $i \notin J(B)$ then $b_{ik} \neq 0$ for some $k \neq i$ by the “connectedness” property (2.10c). Then from (2.11) and (2.12) we see that $\bar{x} = x_k$ for all k for which $b_{ik} \neq 0$. Applying the same considerations to each such k we either arrive at a contradiction or a new set of k 's for which $b_{ik} \neq 0$. Continuing in this manner we construct finite sequences of the type

$$b_{ik_1}, b_{k_1 k_2}, \dots, b_{k_r j}.$$

From condition (2.10c) we must finally arrive at some index $j \in J(B)$ for which (2.13) holds and we have a contradiction. Hence the theorem is proved. The following theorem more fully explains the relationship between positive type matrices and monotone matrices.

Theorem 2.3: *The matrix B is monotone if and only if there exist matrices $P_1 \geq 0$, $P_2 \geq 0$ for which $P_1 B P_2$ is of positive type.*

Proof: If B is monotone then B^{-1} exists by theorem 2.1 and we let $P_1 = I$ (the identity matrix) and $P_2 = B^{-1}$. Hence $P_1 B P_2 = I$ which is of positive type.

On the other hand if there exist matrices $P_1 \geq 0$, $P_2 \geq 0$ for which $P_1 B P_2$ is of positive type we see from theorems 2.1 and 2.2 that $P_1 B P_2$ is non-singular and hence B is also. Furthermore

$$B^{-1} = P_2 (P_1 B P_2)^{-1} P_1 \geq 0$$

and hence B is monotone. Thus the theorem is proved.

The matrices which arise in certain finite difference analogues to elliptic boundary value problems are already of positive type and hence theorem 2.2 tells us that they are monotone. However, a wide class of otherwise acceptable finite difference analogues do not fit in this category. In fact, those which have a local truncation error of higher order quite often are not of positive type. The finite difference analogue (2.7) is a good illustration of this. We shall show, however, that one can construct matrices $P_1 \geq 0$ and $P_2 \geq 0$ for which $P_1 A P_2$ is of positive type and hence, by theorem 2.3, A is monotone. The method employed here involves the known Green's function of a related operator L_h . These results which are embodied in the following theorems are meant to suggest a method of attack in problems with non-positive type matrices which one suspects are monotone.

In order to more clearly illustrate the ideas involved we first consider the case of $q \equiv 0$. The proof of the theorem in the more general case follows the same lines and is included in the appendix.

Theorem 2.4: *The matrix A , defined by (2.6), with $q \equiv 0$ is monotone.*

Proof. Let the matrix G have the elements

$$(2.14) \quad G_{ij} = \begin{cases} h^2 & i = j = 0, N \\ ih^2(1 - jh) & \text{otherwise if } i \leq j \\ jh^2(1 - ih) & \text{otherwise if } j \leq i. \end{cases}$$

Note that G is defined in terms of the Green's function for the continuous problem, restricted to the mesh points.

It is easy to verify that

$$(2.15) \quad AG = \begin{bmatrix} h^2 & 0 & 0 & 0 & \cdots & 0 \\ -1 & 1 & 0 & 0 & \cdots & 0 \\ \frac{1}{12} & -\frac{1}{12} & 1 & -\frac{1}{12} & \cdots & 0 \\ 0 & 0 & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 1 & -\frac{1}{12} \quad \frac{1}{12} \\ \cdot & \cdot & \cdot & 0 & 1 & -1 \\ 0 & 0 & \cdot & \cdot & 0 & 0 \quad h^2 \end{bmatrix}$$

If we were to replace the first (last) column of AG by the sum of the first (last) two columns the resulting matrix would be of positive type. But this is just the matrix $AG(I + W)$ where W has elements $w_{10} = w_{(N-1), N} = 1$, otherwise $w_{ij} = 0$. We now apply Theorem 2.3 with $P_1 = I$ and $P_2 = G(I + W)$. Thus A is monotone. We next state the more general result.

Theorem 2.5: *The matrix A defined by (2.6) is monotone, provided*

$$(2.16) \quad h < h_0$$

where h_0 depends only on q .

Remark. No attempt has been made to obtain a sharp inequality in (2.16),

however the fact that some condition of this type is required in order that the matrix A be monotone is illustrated by the following example. In (2.6) take $N = 4$ (i.e. $h = \frac{1}{4}$) and $q(h) = 400$, $q(2h) = q(3h) = 0$. In this case $a^{31} < 0$.

See the appendix for the proof of Theorem 2.5.

The following lemmas will help us to show that the “discretization error,” $y - Y$, is $O(h^4)$. The first may be found in [15, p. 362].

Lemma 2.1 If A_1 and A_2 are monotone matrices, such that $A_2 \geq A_1 \geq 0$, then $A_1^{-1} \geq A_2^{-1} \geq 0$

Proof

$$A_1^{-1} - A_2^{-1} = A_1^{-1}(A_2 - A_1)A_2^{-1} \geq 0.$$

As a consequence of this lemma we note that for the operator \bar{A} which results from setting $q \equiv 0$ in (2.6) that

$$(2.17) \quad (\bar{A})^{-1} \geq A^{-1}.$$

Lemma 2.2

$$\sum_{j=1}^{N-1} \bar{a}^{ij} \leq \frac{1}{8}$$

Proof Let $v(x) = \frac{1}{2}x(1-x)$. Then

$$-\Delta_x v(x) = -\left(I - \frac{h^2}{12}\Delta_x\right)\Delta_x v(x) = -v''(x) = 1, v(0) = v(1) = 0.$$

Hence

$$\sum_i \bar{a}_{ij}[v(jh) - \sum_{k=1}^{N-1} \bar{a}^{jk}] \geq 0,$$

and by Theorem 2.4 we have

$$\frac{1}{8} \geq v(jh) \geq \sum_{k=1}^{N-1} \bar{a}^{jk}.$$

Lemma 2.3.

$$\bar{a}^{j1}, \bar{a}^{j(N-1)} \leq 3h^2$$

Proof. Define

$$v(x) = \begin{cases} 0, & x = 0 \\ h^2(3 - h/x), & x = h, 2h, \dots \end{cases}$$

It is easily verified that

$$\bar{a}_{ij}v(jh) \geq \delta_{i1}$$

and hence from (2.9)

$$3h^2 \geq v(jh) = \sum_{lk} \bar{a}^{jk} \bar{a}_{kl} v(lh) \geq \bar{a}^{j1}$$

which gives the desired result for a^{j1} . That for $a^{j(N-1)}$ is obtained in a similar manner.

Theorem 2.6. If $\epsilon_i \equiv y(ih) - Y(ih)$, where $y(x)$ and $Y(x)$ are solutions of (2.1) and (2.5) respectively, and if $y \in C^6[0, 1]$, then $\epsilon_i = O(h^4)$.

Proof. From Poisson's formula (2.9)

$$\epsilon_i = \sum_{j,k} a^{ij} a_{kj} \epsilon_j$$

$$|\epsilon_i| \leq \bar{a}^{i1} |\Delta_x y(h) - y''(h)| + \bar{a}^{i(N-1)} |\Delta_x y(Nh - h) - y''(Nh - h)| + \left(\sum_{j=1}^{N-1} \bar{a}^{ij} \right) \max_{k=2, \dots, N-2} \left| \left(I - \frac{h^2}{12} \Delta_x \right) \Delta_x y(kh) - y''(kh) \right|,$$

for h so small that $a^{ij} \geq 0$ (see Theorem 2.5). We find upon substituting from (2.3), lemma 2.2, and lemma 2.3 that

$$|\epsilon_i| \leq h^4 \left\{ \frac{1}{2} \left| \frac{d^4 y}{dx^4} \right|_M + \frac{1}{720} \left| \frac{d^6 y}{dx^6} \right|_M \right\}.$$

Thus the theorem is proved.

An $O(h^4)$ finite difference analogue whose coefficient matrix satisfies the conditions (2.10) was formulated by A. K. Aziz and B. E. Hubbard (2), and the discretization error bounded in terms of data by a different method.

3. Iterative Methods. In this section we shall discuss two iterative methods for solving the linear system (2.7). It will be shown that for h taken sufficiently small the method of simultaneous displacements diverges while the "to and fro" modification of the method of successive displacements converges, c.f. Aitken [1, page 57].

For convenience we shall normalize A so that the diagonal elements are each 1. Let C be the matrix which results, i.e.

$$(3.1) \quad C_{ij} \equiv \frac{a_{ij}}{a_{ii}}.$$

Now we decompose C in the usual manner

$$(3.2) \quad C = I - L - U$$

where I is the identity matrix and U, L are the upper and lower triangular matrices

$$U = \begin{bmatrix} 0 & 0 & 0 & 0 & \cdot & \cdot & \cdot \\ 0 & 0 & C_{23} & 0 & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & C_{34} & C_{35} & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix}$$

$$(3.3) \quad L = \begin{bmatrix} 0 & 0 & 0 & \cdot & \cdot & \cdot \\ C_{21} & 0 & 0 & \cdot & \cdot & \cdot \\ C_{31} & C_{32} & 0 & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix}$$

The linear system corresponding to (27) then takes the form

$$(3.4) \quad (I - L - U)y = \bar{f}$$

where \bar{f} has been properly normalized, i.e.

$$(3.5) \quad \bar{f}_i \equiv \frac{f_i}{a_{ii}}.$$

The simultaneous displacement (point Jacobi) method involves the iteration

$$(3.6) \quad \begin{aligned} y^{(1)} &= (L + U)y^{(0)} + \bar{f} \\ &\vdots \qquad \vdots \qquad \vdots \\ y^{(n)} &= (L + U)y^{(n-1)} + \bar{f}. \end{aligned}$$

This method diverges if the spectral radius $\rho(L + U) > 1$, as will now be proved for N sufficiently large

Theorem 3.1. *For all $h > 0$ satisfying the inequality*

$$\frac{8}{3} \left[\frac{5}{2} + h^2 q_M \right]^{-1} \cos \left(\frac{\pi h}{1 - 2h} \right) > 1$$

it follows that $\rho(L + U) > 1$ and hence the method of simultaneous displacements diverges (Note that the above inequality can always be satisfied for sufficiently small h .)

Proof. The matrix $L + U$ is a matrix whose elements C_{ij} have the same sign as $(-1)^{i+j+1}$. Let $(L + U)^+$ be the non-negative matrix with elements

$$(-1)^{i+j+1} C_{ij}.$$

It is easily verified that the two matrices $(L + U)$ and $(L + U)^+$ have the same eigenvalues but with opposite signs. If x_i is an eigenvector of $(L + U)$ then $(-1)^{i+1} x_i$ is the corresponding eigenvector of $(L + U)^+$. By a corollary of the Perron-Frobenius theorem on non-negative matrices, ρ , (which is the largest eigenvalue of $(L + U)^+$) dominates the absolute value of the eigenvalues of the matrix D , [13, page 69],

$$(3.7) \quad D = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & d & 0 & 0 \\ 0 & 0 & d & 0 & d & 0 \\ 0 & 0 & 0 & d & 0 & d \\ \vdots & & & & & \ddots \\ & & & & & \end{bmatrix}$$

where

$$(3.8) \quad d = \frac{4}{3} \left[\frac{5}{2} + h^2 q_M \right]^{-1}.$$

The eigenvalues of D are known to be zero and [19, page 2]

$$\lambda_k = -2d \cos k\theta, k = 1, 2, \dots, N-3$$

where $\theta = \pi/N - 2$ (Note that $N = 1/h$). Now by hypothesis

$$(3.9) \quad 1 < |\lambda_1| \leq \rho$$

and Theorem 3.1 is proved.

Even though the Jacobi iteration diverges for small values of h , the following “forward-backward” Gauss-Seidel iteration can be shown to converge for small values of h . This iteration involves moving through the mesh points successively from left to right using new values of $y^{(i)}$ as they become available. Following this, one reverses direction and proceeds through the mesh from right to left, again using the new values of $y^{(i)}$ as they become available. This can be written in two stages as

$$(3.10) \quad y^{(n-1)} = Ly^{(n-1)} + Uy^{(n-1)} + \bar{f}$$

followed by

$$(3.11) \quad y^{(n)} = Ly^{(n-1)} + Uy^{(n)} + \bar{f}.$$

These can be combined into the single forward-backward iteration

$$(3.12) \quad y^{(n)} = (I - U)^{-1}(I - L)^{-1}LUy^{(n-1)} + (I - U)^{-1}(I - L)^{-1}\bar{f}.$$

To show that this iterative process converges we shall need the following lemma.

Lemma 3.1. If $h > 0$ satisfies the condition $d^2 > 2/15$, then $(I - U)^{-1} \geq 0$ and $(I - L) \geq 0$, where d is given by (3.8).

Proof: Since U is upper triangular

$$(3.13) \quad (I - U)^{-1} = I + U + U^2 + \dots + U^{N-1}.$$

Let d_{ij} be the elements of $(I - U)^{-1}$. From (3.13) these take the form

$$(3.14) \quad d_{ij} = \delta_{ij} + \sum_{k=1} u_{ik}u_{kj} + \dots + \sum_{k_1, k_2, \dots, k_{N-2}} u_{ik_1}u_{kj}u_{k_1k_2} \dots u_{k_{N-2}j}$$

where $u_{ij} \in U$. We see further from the form of U that d_{ij} satisfies the recurrence

$$(3.15) \quad d_{i(i+n)} = U_{(i+n-1)(i+n)}d_{i(i+n-1)} + U_{(i+n-2)(i+n)}d_{i(i+n-2)}.$$

We observe that $d_{ii}, d_{i(i+1)}, d_{i(i+2)} > 0$ so that we need only show that $d_{i(i+n-1)} > 0$ implies that $d_{i(i+n)} > 0$. Let $y_n = d_{i(i+n)}/d_{i(i+n-1)}$ which we assume exists. Certainly $y_1, y_2 > 0$ and it is clear that if we finally reach the point where $d_{i(i+n-1)} \geq 0 > d_{i(i+n)}$ then $y_n < 0$ for the first time. The set $\{y_n\}$ satisfy the recursion

$$(3.16) \quad y_n = u_{(i+n-1)(i+n)} + \frac{u_{(i+n-2)(i+n)}}{y_{(n-1)}}$$

Now if $\bar{y}_j > 0$ for all j and satisfies the recursion

$$(3.17) \quad \bar{y}_n = 2\bar{a}_n - \frac{\bar{b}_n}{\bar{y}_{n-1}},$$

where

$$(3.18) \quad \bar{a}_n \leq u_{(i+n-1)(i+n)}, \quad \bar{b}_n \geq |u_{(i+n-2)(i+n)}|$$

then $y_i \geq \bar{y}_i$. The proof of this statement, which we omit here, follows easily by an inductive argument.

We make use of the particular case

$$(3.19) \quad \bar{a}_n = d/2, \quad \bar{b}_n = 1/30$$

so that (3.18) is satisfied. Then the recursion (3.17) can be solved explicitly. For, let

$$(3.20) \quad \bar{y}_n = \alpha(\xi_1)^n + \beta(\xi_2)^n$$

where ξ_1, ξ_2 are the roots of the quadratic equation

$$(3.21) \quad \xi^2 - d\xi + 1/30 = 0.$$

We note that by hypothesis

$$(3.22) \quad \xi_1 = d/2 + \sqrt{(d/2)^2 - 1/30} > 0, \quad \xi_2 = d/2 - \sqrt{(d/2)^2 - 1/30} > 0.$$

The real numbers α, β are determined by the initial conditions. If $i \geq 2$, $j \leq N-1$ then

$$(3.23) \quad \alpha + \beta = 1, \quad \alpha\xi_1 + \beta\xi_2 = d$$

which yields

$$(3.24) \quad \alpha = \frac{1}{2} + d[(d/2)^2 - 1/30]^{-\frac{1}{2}}, \quad \beta = \frac{1}{2} - d[(d/2)^2 - 1/30]^{-\frac{1}{2}}.$$

from (3.22) and (3.24) we see that $\bar{y}_n > 0$. In the special case $i = 1$ let

$$(3.25) \quad \bar{y}_n = \bar{\alpha}(\xi_1)^{n-1} + \bar{\beta}(\xi_2)^{n-1},$$

where

$$(3.26) \quad \bar{\alpha} = \alpha[2 + h^2 q(h)]^{-1}, \quad \bar{\beta} = \beta[2 + h^2 q(h)]^{-1}.$$

Again we see that $\bar{y}_n > 0$. For $j = N$ the result is clear. Hence the lemma is proved.

We are now in a position to prove the convergence of the iteration method.

Theorem 3.2: Let h be chosen so small that $d^2 > 2/15$ and that C^{-1} exists (see theorem 2.5). Then

$$\rho[(I - U)^{-1}(I - L)^{-1}LU] < 1$$

and the forward-backward Gauss-Seidel iteration converges.

Proof: Let

$$\epsilon_m \equiv \frac{\epsilon}{12} \left[\frac{5}{12} + h^2 q(mh) \right]^{-1}, \quad 0 \leq \epsilon \leq 1.$$

Let C_ϵ be the matrix formed from C by making the following substitutions in the rows $2 \leq i \leq N - 2$

$$(3.27) \quad \begin{aligned} C_{i(i-2)} &\rightarrow (C_{i(i-2)} - \epsilon_i) \\ C_{i(i-1)} &\rightarrow (C_{i(i-1)} + \epsilon_i) \\ C_{i(i+1)} &\rightarrow (C_{i(i+1)} + \epsilon_i) \\ C_{i(i+2)} &\rightarrow (C_{i(i+2)} - \epsilon_i). \end{aligned}$$

It is clear that the matrix C_1 satisfies the conditions (2.10). Furthermore since $(I - L_1)^{-1} \geq 0$, $(I - U_1)^{-1} \geq 0$ it follows that

$$(3.28) \quad I - (I - U_1)^{-1}(I - L_1)^{-1}L_1U_1 \equiv (I - U_1)^{-1}(I - L_1)^{-1}C_1$$

also satisfies the conditions (2.10). From this we infer

$$(3.29) \quad \rho[(I - U_1)^{-1}(I - L_1)^{-1}L_1U_1] < 1.$$

Now the two matrices $(I - L_\epsilon)^{-1}L_\epsilon U_\epsilon(I - U_\epsilon)^{-1}$ and $(I - U_\epsilon)^{-1}(I - L_\epsilon)^{-1}L_\epsilon U_\epsilon$ are similar and consequently have the same eigenvalues. As in lemma 3.1 it is easily shown that $(I - L_\epsilon)^{-1} \geq 0$ and $(I - U_\epsilon)^{-1} \geq 0$. Hence

$$(3.30) \quad \begin{aligned} U_\epsilon(I - U_\epsilon)^{-1} &\equiv (I - U_\epsilon)^{-1} - I \geq 0, \\ (I - L_\epsilon)^{-1}L_\epsilon &\equiv (I - L_\epsilon)^{-1} - I \geq 0. \end{aligned}$$

From this we have

$$(3.31) \quad (I - L_\epsilon)^{-1}L_\epsilon U_\epsilon(I - U_\epsilon)^{-1} \geq 0.$$

Moreover by the same reasoning as was used in lemma 2.2 it can be shown that C_ϵ^{-1} exists. Therefore we infer the existence of

$$(3.32) \quad [I - (I - L_\epsilon)^{-1}L_\epsilon U_\epsilon(I - U_\epsilon)^{-1}]^{-1} \equiv (I - U_\epsilon)^{-1}C_\epsilon^{-1}(I - L_\epsilon)^{-1},$$

and consequently the spectral radius (which is the largest eigenvalue) depends continuously on ϵ and hence satisfies the condition

$$(3.33) \quad \rho[(I - L_\epsilon)^{-1}L_\epsilon U_\epsilon(I - U_\epsilon)^{-1}] < 1, \quad 0 \leq \epsilon \leq 1.$$

In particular this is true for $\epsilon = 0$ and the theorem is proved.

IV. Appendix: (Proof of Theorem 2.5). In order to prove this theorem we shall construct non-negative matrices P_1 and P_2 and apply theorem 2.3

For this purpose let C be a positive real number such that $C^2 > 2q_M$. Define $K(x, \xi)$ to be the Green's function of the Dirichlet problem for the operator

$$(4.1) \quad \bar{L}u \equiv -u'' + C^2u.$$

It is well known [10, page 353] that

$$(4.2) \quad \frac{dK(x, \xi)}{dx} \Big|_{\substack{x = \xi + 0 \\ x = \xi - 0}} = -1$$

and that K considered as a function of x satisfies the differential equation

$$(4.3) \quad \bar{L}K = 0$$

throughout $(0, 1)$ except at the point $x = \xi$. Further $K(x, \xi)$ is symmetric and satisfies the boundary conditions

$$(4.4) \quad K(0, \xi) = K(1, \xi) = 0.$$

In fact K is seen to be

$$(4.5) \quad K(x, \xi) = [2C \sinh C]^{-1} \{ \cosh C[|\xi - x| - 1] - \cosh C[\xi + x - 1]\}.$$

Let \bar{B} be the matrix defined by

$$(4.6) \quad \bar{B}_{ij} = \begin{cases} h^2, & i = j = 0 \\ h^2, & i = j = N \\ hK(ih, jh), & \text{otherwise.} \end{cases}$$

Let \bar{D} be the diagonal matrix defined by

$$(4.7) \quad \bar{D}_{00} = \bar{D}_{NN} = h^{-2}, \quad D_{11} = \cdots = D_{(N-1)(N-1)} = 1.$$

The matrix $\bar{D}A\bar{B}$ has the same first and last columns as $h^2\bar{D}A$. We establish by a Taylor expansion that for $h \leq \xi \leq (N-1)h$, $x = mh$, $\xi = nh$

$$(4.8) \quad (\bar{D}A\bar{B})_{mn} = L_{h,x}[hK(x, \xi)] < 0; \\ x = h, \dots, \xi - 2h, \xi + 2h, \dots, (N-1)h,$$

for $h < h_1$, where h_1 is a constant depending only on C . (For $\xi = h$ (4.8) holds for $x = 3h, \dots, (N-1)h$ and similarly for $\xi = (N-1)h$). At the point $x = h$ we see that

$$(4.9) \quad \begin{aligned} L_{h,x}(Kx, \xi) &= -\frac{\partial^2 K(x, \xi)}{\partial x^2} + q(x)K(x, \xi) - \frac{h^2}{24} \left[\frac{\partial^4 K(\bar{x}, \xi)}{\partial x^4} + \frac{\partial^4 K(\tilde{x}, \xi)}{\partial x^4} \right] \\ &= [q(x) - C^2]K(x, \xi) - \frac{C^4 h^2}{24} [K(\bar{x}, \xi) + K(\tilde{x}, \xi)] < 0 \end{aligned}$$

where $x - h < \bar{x} < x < \tilde{x} < x + h$. A similar equation holds for $x = (N-1)h$.

At the points $2h \leq x \leq (N-2)h$ we have

$$(4.10) \quad \begin{aligned} L_{h,x}K(x, \xi) &= -\frac{\partial^2 K(x, \xi)}{\partial x^2} + q(x)K(x, \xi) \\ &\quad + \frac{h^{-2}}{6} K(x, \xi) \sum_{n=3}^{\infty} \left[\frac{(2h)^{2n} - 16h^{2n}}{2n!} \right] C^{2n} \\ &= \left\{ q(x) - C^2 + \frac{h^{-2}}{6} \sum_{n=3}^{\infty} \left[\frac{(2h)^{2n} - 16h^{2n}}{2n!} \right] C^{2n} \right\} K(x, \xi) < 0 \end{aligned}$$

when $h \leq h_1$ (h_1 chosen sufficiently small). Hence (4.8) is verified. We observe further that for $2h < \xi < (N - 2)h$ and $x = \xi - h$

$$\begin{aligned}
(4.11) \quad L_{h,x}[hK(x, \xi)] &\equiv \frac{1}{12} \Delta_x[hK(\xi - 2h, \xi)] - \frac{14}{12} \Delta_x[hK(\xi - h, \xi)] \\
&\quad + \frac{1}{12} \Delta_x[hK(\xi, \xi)] + hq(\xi - h)K(\xi - h, \xi) \\
&= -\frac{1}{12} - [C^2 - q(\xi - h)]hK(\xi - h, \xi) - \frac{h^2 C^2}{72} \\
&\quad + \frac{h^3 C^4}{2(12)^2} \{K(x_1, \xi) + K(x_2, \xi) - 14K(x_3, \xi) - 14K(x_4, \xi) \\
&\quad + K(\bar{x}, \xi) + K(\tilde{x}, \xi) + 12K(x_5, \xi) + 12K(x_6, \xi)\} < 0
\end{aligned}$$

for $h < h_2$ (h_2 a constant depending only on C) where the indicated intermediate points lie near $\xi - h$. A similar expression holds for $x = \xi + h$. At the point $x = \xi$ we have

$$\begin{aligned}
(4.12) \quad L_{h,x}[hK(x, \xi)] &= \frac{14}{12} - [C^2 - q(\xi)]hK(\xi, \xi) + \frac{h^2 C^2}{9} \\
&\quad + \frac{h^3 C^4}{2(12)^2} \{K(x_1, \xi) + K(x_2, \xi) - 14K(\bar{x}, \xi) - 14K(\tilde{x}, \xi) \\
&\quad + K(x_3, \xi) + K(x_4, \xi) + 12K(x_5, \xi) + 12K(x_6, \xi)\}
\end{aligned}$$

For ξ in the range indicated we see that

$$\begin{aligned}
(4.13) \quad \sum_{x=h}^{(N-1)h} L_{h,x}[hK(x, \xi)] &\geq 1 - h \sum_{m=1}^{N-1} [C^2 - q(mh)]K(mh, \xi) + \frac{h^2 C^2}{12} - \frac{7}{72} h^2 C^4 K_M.
\end{aligned}$$

Since K is a convex function of x we have

$$(4.14) \quad hK(x, \xi) \leq \int_{x-h/2}^{x+h/2} K(t, \xi) dt$$

and hence

$$(4.15) \quad C^2 h \sum_{m=1}^{N-1} K(mh, \xi) \leq 1 + C^2 h K_M - 2 \left[\frac{\sinh C/2}{\sinh C} \right]$$

We see then that

$$(4.16) \quad \sum_{x=h}^{(N-1)h} L_{h,x}[hK(x, \xi)] > 0$$

if $h \leq h_3$ (h_3 a constant depending only on C).

We note that the same considerations can be applied where $\xi = 2h$ (or $(N - 2)h$) with inequalities (4.11) through (4.16) holding. Inequality (4.11) is replaced for $x = h$ ($x = (N - 1)h$) by (4.8)

When $\xi = h$ (the case $\xi = (N - 1)h$ is similar) we see that (4.11) applies only to $x = 2h$ and that (4.12) is replaced by

$$(4.17) \quad \begin{aligned} L_{h,x}[hK(h, h)] &= 1 + h[q(h) - C^2]K(h, h) \\ &\quad + \frac{C^2h^2}{6} - \frac{C^4h^3}{24} \{K(\bar{x}, h) + K(\tilde{x}, h)\}. \end{aligned}$$

We further note that

$$(4.18) \quad \sum_{x=h}^{(N-1)h} L_{h,x}[hK(x, h)] > 0$$

for $h < h_4$ by the same considerations as before.

Interpreting inequalities (4.8) through (4.18) in terms of the matrix $\bar{D}A\bar{B}$ we see that

$$(4.19) \quad \begin{aligned} (\bar{D}A\bar{B})_{ij} &< 0, \quad i \neq j, \quad j = 1, \dots, N - 1 \\ \sum_i (\bar{D}A\bar{B})_{i,j} &> 0, \quad j = 1, \dots, N - 1. \end{aligned}$$

Unfortunately these properties do not hold for the first and last columns which remain unchanged. We shall show that

$$(4.20) \quad \begin{aligned} a) \quad &(\bar{D}A\bar{B})_{i0} + (\bar{D}A\bar{B})_{i1} < 0, \quad i \neq 0 \\ b) \quad &\sum_i \{(\bar{D}A\bar{B})_{i0} + (\bar{D}A\bar{B})_{i1}\} > 0 \end{aligned}$$

for h chosen sufficiently small.

From (4.17) we see that

$$(4.21) \quad (\bar{D}A\bar{B})_{10} + (\bar{D}A\bar{B})_{11} < -[C^2 - q(h)]hK(h, h) + C^2h^2/6.$$

A lower bound for $K(h, h)$ is given by

$$(4.22) \quad K(h, h) \geq 3h/4$$

Inequality (4.20 a) now follows from (4.21) and (4.22) for $h < h_5$ and $i = 1$. For $i = 2$ we have

$$(4.23) \quad (\bar{D}A\bar{B})_{20} - (\bar{D}A\bar{B})_{21} < -[C^2 - q_M]hK(h, 2h) - \frac{h^2C^2}{72} + \frac{7}{72}h^3C^4K_M < 0$$

for $h < h_6$. Finally we see that the remaining terms are negative from (4.9) and (4.10) and the fact that $(\bar{D}A\bar{B})_{i0} = 0$, $i > 2$. This completes the proof of (4.20 a).

To establish (4.20 b) we note that $K(x, h) = O(h)$. It then follows from (4.9) and (4.10) that all terms of the sum in (4.20 b) for $i \geq 3$ are $O(h^2)$. Since the total number of terms is $O(h^{-1})$ the contribution from these terms is $O(h)$. On the other hand, the first term is 1 and we see from (4.11) and (4.17) that the second and third terms are also $O(h^2)$. Hence for $h < h_7$ (4.20 b) holds. Similarly, if we add the elements of column $N - 1$ to column N , like inequalities hold.

Consequently we use the matrix W defined after (2.15). We then define the non-negative matrices P_1, P_2 to be

$$(4.24) \quad P_1 = \bar{D}, \quad P_2 = \bar{B}(I + W).$$

We see that P_1AP_2 is of positive type and hence by theorem 2.3 the matrix A is monotone. The theorem is thus proved.

BIBLIOGRAPHY

- 1 AITKEN, A C, *On the Iterative Solution of a System of Linear Equations*, Proc Royal Soc Edinburgh, **63**, pp. 52-60, (1950)
- 2 AZIZ, A K, HUBBARD, B. E, *Bounds for the Solutions of the Sturm-Liouville Problem with Application to Finite Difference Methods*, Georgetown University report, (1961)
- 3 BATSCHELET, E., *Über die Numerische Auflösung von Randwertproblemen bei Elliptischen Differentialgleichungen*, A. Angew Math Physik, **3**, pp 165-193, (1952)
- 4 BRAMBLE, J H, *Fourth Order Finite Difference Analogues of the Dirichlet Problem for Poisson's Equation in Three and Four Dimensions*, (to appear in Mathematics of Computation)
- 5 BRAMBLE, J H, HUBBARD, B E, *On the Formulation of Finite Difference Analogues of the Dirichlet Problem for Poisson's Equation*, Numerische Mathematik **4**, pp 313-327 (1962)
- 6 BRAMBLE, J H, HUBBARD, B E, *A Theorem on Error Estimation for Finite Difference Analogues of the Dirichlet Problem for Elliptic Equations*, Tech Note BN-281, University of Maryland, (1962)
- 7 BRAMBLE, J H., HUBBARD, B. E., *A Priori Bounds on the Discretization Error in the Numerical Solution of the Dirichlet Problem*, (to appear in the Contributions to Differential Equations)
- 8 COLLATZ, L, *Bemerkungen zur Fehlerabschätzung für das Differenzenverfahren bei Partiellen Differentialgleichungen*, Z Angew Math. Mech., **13**, pp 56-57, (1933)
- 9 COLLATZ, L., *Numerical Treatment of Differential Equations*, 3rd ed Berlin, Springer-Verlag, (1960)
- 10 COURANT, R, HILBERT, D, *Methods of Mathematical Physics*, Vol I, New York, Interscience, (1953)
- 11 COURANT, R, HILBERT, D, *Methods of Mathematical Physics*, Vol II, New York, Wiley and Sons, (1962)
- 12 FORSYTHE, G, WASOW, W, *Finite Difference Methods for Partial Differential Equations*, New York, Wiley, (1960)
- 13 GANTMACHER, F R, *Applications of the Theory of Matrices*, New York, Interscience, (1959)
- 14 GERSCHGORIN, S, *Fehlerabschätzung für das Differenzenverfahren zur Lösung Partieller Differentialgleichungen*, Z Angew Math Mech., **10**, pp. 373-382, (1930).
- 15 HENRICI, P, *Discrete Variable Methods in Ordinary Differential Equations*, John Wiley and Sons, New York, (1962)
- 16 OSTROWSKI, A M, *Determination mit überwiegender Hauptdiagonale und die absolute Konvergenz von linearen Iterationsprozessen*, Comm Math. Helv **30**, 175-210, (1955)
- 17 PHILLIPS, H B, WIENER, N, *Nets and the Dirichlet Problem*, J Math. Phys., **2**, pp 105-124, (1923).
- 18 ROCKOFF, M, *On the Numerical Solution of Finite Difference Approximations which Are Not of Positive Type*, Abs Submitted for presentation to Am Math Soc, Notices, January 1963
- 19 TODD, J, *The Condition of Certain Matrices III*, J Res Nat Bur. Standards, **60**, pp 1-7, (1958)

- 20 UHLMANN, N , *Differenzenverfahren fur die 1ste Randwertaufgabe mit krummflachigen Randern bei $u(x, y, z) = r(x, y, z, u)$* , Z Angew Math Mech , **38**, pp. 130-139, (1958)
- 21 WASOW, W , *On the Truncation Error in the Solution of Laplace's Equation by Finite Differences*, J Res Nat Bur Standards, **48**, pp 345-348, (1952)

INSTITUTE FOR FLUID DYNAMICS AND APPLIED MATHEMATICS
UNIVERSITY OF MARYLAND
COLLEGE PARK, MARYLAND

(Received July 4, 1963)

*1.4. APPROXIMATION OF SOLUTIONS OF MIXED BOUNDARY VALUE
PROBLEMS FOR POISSON'S EQUATION BY FINITE DIFFERENCES*

**1.4 Approximation of solutions of mixed boundary value
problems for Poisson's equation by finite differences**

Approximation of solutions of mixed boundary value problems for Poisson's equation by finite differences [35]

Approximation of Solutions of Mixed Boundary Value Problems for Poisson's Equation by Finite Differences

J. H. BRAMBLE AND B. E. HUBBARD

The University of Maryland, College Park, Maryland†

Abstract. This paper is concerned with the formulation of finite-difference analogues of mixed boundary value problems for Poisson's equation. The normal derivative is approximated in such a way that the matrix of the resulting system is of positive type. The discretization error is shown to be $O(h^2)$, where h is the mesh constant.

I. Introduction

In this paper we are concerned with a finite-difference approximation to the solution of the boundary value problem

$$\begin{aligned} -\Delta u &= f, \quad \text{in } R \\ \frac{\partial u}{\partial n} + \alpha u &= g, \quad \text{on } C_1 \\ u &= H, \quad \text{on } C_2. \end{aligned} \tag{1.1}$$

The region R is a bounded connected open set in the (x, y) plane whose boundary C consists of the two parts C_1 , and C_2 . The symbol Δ is the Laplace operator $\Delta = (\partial^2/\partial x^2) + (\partial^2/\partial y^2)$, and $\partial/\partial n$ denotes differentiation with respect to the outward-directed normal on C_1 . The functions f , g and H are defined to be sufficiently smooth functions on R , C_1 and C_2 respectively. The function α is required to satisfy the following conditions on C_1 : (a) piecewise continuity with a finite number of discontinuities, (b) piecewise differentiability, (c) at all points of continuity, either $\alpha = 0$ (the set $C_1^{(1)}$) or $\alpha \geq \alpha_m > 0$, where α_m is a constant (the set $C_1^{(2)}$).

We restrict our considerations to the cases in which either the set C_2 or $C_1^{(2)}$ contains a nonempty open subset of C . In these cases (1.1) has a unique solution provided the data and boundary are sufficiently smooth. The case in which C_2 is all of C is just the Dirichlet problem. Results for this special case are contained in [8].

We are interested in formulating a finite-difference analogue of (1.1) which has the following properties: (a) The boundary approximations involve at most three interior points (and one boundary point) (b) The matrix of the system is

This research was supported in part by a grant with the National Science Foundation NSF Grant GP-3 and with the Air Force Office of Scientific Research, Air Research and Development Command AFOSR 62-454.

† Institute for Fluid Dynamics & Applied Mathematics.

of "positive type" (cf. [4]) (c) The truncation error tends to zero quadratically (as $O(h^2)$, where h is the mesh size).

Section 2 is concerned with the construction of the boundary operator at smooth points of C_1 . The construction given there is intended to show that appropriate points can always be chosen. Practically speaking, the choice would simply be made by examining a finite number of possibilities (in as clever a fashion as possible).

In the last section the approximating finite difference problem is defined and the tools with which to study the truncation error are developed. The error is shown to be $O(h^2)$. We wish to emphasize the interesting fact that at points near the boundary the differential operator is approximated only to $O(h)$, while the boundary condition itself is approximated to $O(h^2)$. A direct application of the method of Gershgorin [9] would show only that our method converges as $O(h)$. Our treatment shows that the convergence is, in fact, $O(h^2)$.

Also in the last section a finite-difference analogue involving an $O(h)$ approximation to $\partial u / \partial n$ using two interior points, (cf [11], p. 213) is seen to lead to $O(h)$ convergence.

Among the authors who have studied problems of this type, Batschelet [2] seems to have gone the furthest. He gives an $O(h)$ approximation to the mixed problem and proves convergence by a direct extension of Gershgorin's technique [9].

Methods for setting up boundary operators (to approximate the normal derivative) are given by Kantorovich and Krylov [11], Shaw [13], Allen [1], Viswanathan [16], Forsythe and Wasow [8], Uhlmann [15], and Greenspan [10]. The local boundary approximation of Viswanathan resembles ours in that the differential equation and boundary condition are taken into account. Uhlmann and Greenspan derive higher order formulas by simply involving more points. Only local properties are discussed and no convergence proofs are given in any but the paper of Batschelet.

II. An $O(h^2)$ Approximation of $\partial u / \partial n$

We consider an arbitrary point 0 on C at which C is smooth (see Figure 1). Choose 0 to be the origin of a Cartesian coordinate system (x, y) such that the x -axis is tangent to C at 0 . The positive y -direction is taken along the inward normal. It can be shown, for any smooth enough function v , that $v_{xy} = -v_{ns} + Kv_s$ at the origin (cf. [14] for the use of geodesic normal coordinates). The subscripts denote the indicated partial differentiation, n being the outward normal direction, s arc length and K the curvature of C . Thus we have

$$v_{xy} = -\frac{\partial}{\partial s}(v_n + \alpha v) + (\alpha + K)v_s + \alpha_s v \quad (2.1)$$

at 0 . Also, of course

$$v_{yy} = \Delta v - v_{xx}. \quad (2.2)$$

Now consider the Taylor expansion of v about 0; i.e.

$$\begin{aligned} v(P) &= v(0) + xv_x(0) + yv_y(0) \\ &\quad + \frac{1}{2}\{x^2v_{xx}(0) + 2xyv_{xy}(0) + y^2v_{yy}(0)\} + O(x^3 + y^3). \end{aligned} \quad (2)$$

We note that $v_x(0) = v_s(0)$ and $v_y(0) = -v_n(0)$. Thus, using (2.1) and (2.2)

$$\begin{aligned} v(P) &= [1 + xy\alpha_s(0)]v(0) + x[1 + y(\alpha(0) + K(0))]v_s(0) \\ &\quad - yv_n(0) + \frac{1}{2}[x^2 - y^2]v_{xx}(0) + \frac{y^2}{2}\Delta v(0) \\ &\quad - xy[v_n + \alpha v]_s + O(x^3 + y^3). \end{aligned}$$

Let $P_i = (x_i, y_i)$, $i = 1, 2, 3$. We wish to determine three numbers a_i , $i = 1, 2, 3$ such that

$$\begin{aligned} \sum_{i=1}^3 a_i\{v(P_i) - [1 + x_i y_i \alpha_s(0)]v(0)\} \\ &= -v_n(0) + \sum_{i=1}^3 a_i \left\{ \frac{y_i^2}{2} \Delta v(0) - x_i y_i [v_n + \alpha v]_s(0) \right\} \\ &\quad + O\left(\sum_{i=1}^3 a_i[x_i^3 + y_i^3]\right). \end{aligned} \quad (2.4)$$

For (2.4) to hold for any v , a_i must satisfy

$$\begin{aligned} \sum_{i=1}^3 a_i y_i &= 1 \\ \sum_{i=1}^3 a_i x_i [1 + y_i(\alpha(0) + K(0))] &= 0 \\ \sum_{i=1}^3 a_i [x_i^2 - y_i^2] &= 0. \end{aligned} \quad (2.5)$$

We will show further that the points P_i may be chosen so that $a_i \geq 0$. This will be useful in the later applications.

To show that we can get a non-negative solution of (2.5), we consider the system

$$\begin{aligned} \sum_{i=1}^3 \bar{a}_i y_i &= 1 \\ \sum_{i=1}^3 \bar{a}_i x_i &= 0 \\ \sum_{i=1}^3 \bar{a}_i (x_i^2 - y_i^2) &= 0. \end{aligned} \quad (2.6)$$

Since we are interested in small values of x and y , \bar{a}_i will be close to a_i . The system (2.6) in matrix notation takes the form

Now consider the Taylor expansion of v about 0; i.e.

$$\begin{aligned} v(P) &= v(0) + xv_x(0) + yv_y(0) \\ &\quad + \frac{1}{2}\{x^2v_{xx}(0) + 2xyv_{xy}(0) + y^2v_{yy}(0)\} + O(x^3 + y^3). \end{aligned} \quad (2.3)$$

We note that $v_x(0) = v_s(0)$ and $v_y(0) = -v_n(0)$. Thus, using (2.1) and (2.2),

$$\begin{aligned} v(P) &= [1 + xy\alpha_s(0)]v(0) + x[1 + y(\alpha(0) + K(0))]v_s(0) \\ &\quad - yv_n(0) + \frac{1}{2}[x^2 - y^2]v_{xx}(0) + \frac{y^2}{2}\Delta v(0) \\ &\quad - xy[v_n + \alpha v]_s + O(x^3 + y^3). \end{aligned}$$

Let $P_i = (x_i, y_i)$, $i = 1, 2, 3$. We wish to determine three numbers a_i , $i = 1, 2, 3$ such that

$$\begin{aligned} \sum_{i=1}^3 a_i\{v(P_i) - [1 + x_i y_i \alpha_s(0)]v(0)\} \\ &= -v_n(0) + \sum_{i=1}^3 a_i \left\{ \frac{y_i^2}{2} \Delta v(0) - x_i y_i [v_n + \alpha v]_s(0) \right\} \\ &\quad + O\left(\sum_{i=1}^3 a_i[x_i^3 + y_i^3]\right). \end{aligned} \quad (2.4)$$

For (2.4) to hold for any v , a_i must satisfy

$$\begin{aligned} \sum_{i=1}^3 a_i y_i &= 1 \\ \sum_{i=1}^3 a_i x_i [1 + y_i(\alpha(0) + K(0))] &= 0 \\ \sum_{i=1}^3 a_i [x_i^2 - y_i^2] &= 0. \end{aligned} \quad (2.5)$$

We will show further that the points P_i may be chosen so that $a_i \geq 0$. This will be useful in the later applications.

To show that we can get a non-negative solution of (2.5), we consider the system

$$\begin{aligned} \sum_{i=1}^3 \bar{a}_i y_i &= 1 \\ \sum_{i=1}^3 \bar{a}_i x_i &= 0 \\ \sum_{i=1}^3 \bar{a}_i (x_i^2 - y_i^2) &= 0. \end{aligned} \quad (2.6)$$

Since we are interested in small values of x and y , \bar{a}_i will be close to a_i . The system (2.6) in matrix notation takes the form

$$\begin{bmatrix} y_1 & y_2 & y_3 \\ x_1 & x_2 & x_3 \\ x_1^2 - y_1^2 & x_2^2 - y_2^2 & x_3^2 - y_3^2 \end{bmatrix} \begin{bmatrix} \bar{a}_1 \\ \bar{a}_2 \\ \bar{a}_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}. \quad (2.7)$$

The solution of (2.7) is

$$\begin{aligned} \bar{a}_1 &\equiv \frac{1}{\bar{D}} [x_2(x_3^2 - y_3^2) - x_3(x_2^2 - y_2^2)] = \frac{\bar{D}_1}{\bar{D}} \\ \bar{a}_2 &\equiv \frac{1}{\bar{D}} [x_3(x_1^2 - y_1^2) - x_1(x_3^2 - y_3^2)] = \frac{\bar{D}_2}{\bar{D}} \\ \bar{a}_3 &\equiv \frac{1}{\bar{D}} [x_1(x_2^2 - y_2^2) - x_2(x_1^2 - y_1^2)] = \frac{\bar{D}_3}{\bar{D}} \end{aligned} \quad (2.8)$$

where \bar{D} is the determinant of the system. If the \bar{D}_i 's in (2.8) are chosen positive, then clearly the condition that $y_i > 0$ insures that $\bar{D} > 0$ (since $\bar{D} = \sum_{i=1}^3 y_i \bar{D}_i$) and hence the \bar{a}_i 's will be positive. The condition that $y_i > 0$ means essentially that the points P_i are to be taken from R .

Now we need to restrict our attention to only certain points of R . In particular we show that we may always select three *nearby* interior mesh points for which (2.5) is satisfied with $\bar{a}_i > 0$. By nearby we mean that the points (x_i, y_i) lie within a circle about 0 with radius Mh whenever the mesh size, h , is taken sufficiently small. We give now one possible construction to show that this can be done.

In the usual manner we put a square mesh of size h on R and call the crossings *mesh points*. Now let ϵ be a given positive number (which will depend on h). Choose

$$\begin{aligned} 4\epsilon &> x_1 > y_1 + \epsilon > 2\epsilon \\ 4\epsilon &> -x_2 > y_2 + \epsilon > 2\epsilon \\ 6\epsilon &\geq y_3 > |x_3| + 5\epsilon. \end{aligned} \quad (2.9)$$

Geometrically this means that P_1 lies in I, P_2 in II and P_3 in III of Figure 1. The number \bar{K} denotes the maximum positive curvature of C . It follows that $\bar{D}_i > 12\epsilon^3$, $i = 1, 2, 3$. The triangular regions I, II, III will lie in R provided $\epsilon < 2/17\bar{K}$ and the region is "wide" enough. If on the other hand $\epsilon = (\frac{3}{2})h$, then at least one point of the mesh will lie in each of the regions I, II and III. Hence $h < 4/51\bar{K}$ is a sufficient condition for the existence of $(x_i, y_i) \in R$ for which a solution $\bar{a}_i \geq 0$ of (2.7) exists. It is also easy to see that the points in question always lie in a region of radius $10h$ so that we have only a finite number of points (independent of h) to consider. Now it is easy to see that $\bar{D} < 768\epsilon^4$. We want finally to relate this to the solution of (2.5). Let $a_i = D_i/D$, where D is the determinant of the system and the D_i 's are the appropriate cofactors. Comparing the two systems and using the inequalities (2.9) we have

$$-672\epsilon^4 |\alpha + K|_M + \bar{D}_i \leq D_i \leq \bar{D}_i + 672\epsilon^4 |\alpha + K|_M, \quad (2.10)$$

where $|\alpha + K|_M = \text{Max}_{P \in C} |\alpha(P) + K(P)|$. From (2.10) it follows imme-

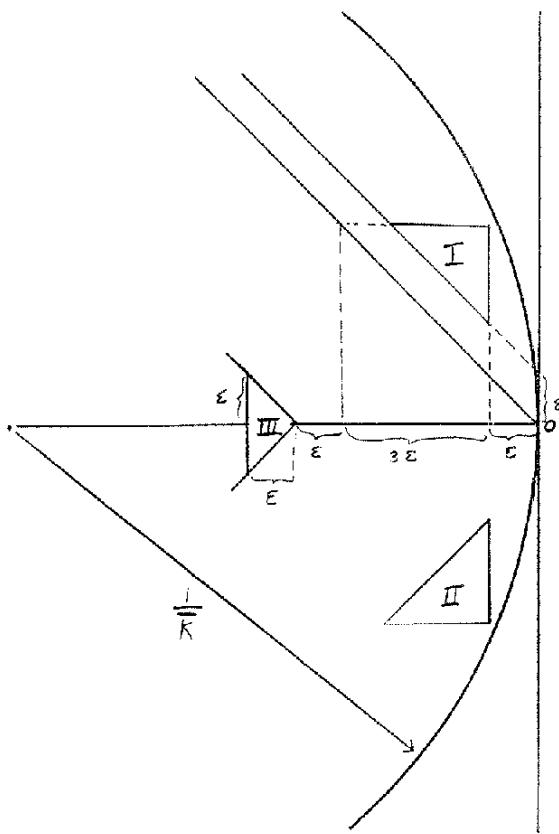


FIG. 1

diately that

$$a_i > h^{-1} \left[\frac{1 - (84 |\alpha| + K |_M)h}{96 + (756 |\alpha| + K |_M)h} \right].$$

Clearly, for sufficiently small h , $a_i > 0$ provided $|\alpha| + K$ is bounded. In like manner it can be shown that each a_i is bounded above by a term which is $O(h^{-1})$.

Thus we have shown, under these assumptions, that it is always possible to choose three points P_i from a given set of mesh points such that $a_i > 0$. Furthermore the points can be found within a sphere of radius βh where β is a constant independent of h .

Thus we define the boundary operator

$$\delta_n V(P) = \sum_{i=1}^3 a_i \{ [1 + x_i y_i \alpha_s(P)] V(P) - V(P_i) \},$$

where α is a given function on C and the P_i are chosen as mesh points in R such that $a_i \geq 0$, $i = 1, 2, 3$. The equations (2.5) are assumed to be satisfied. From (2.4) and (2.10) it is easy to see that u of (1.1) satisfies

$$\begin{aligned} \left| \delta_n u(P) + \alpha(P)u(P) - \left\{ g(P) + \sum_{i=1}^3 a_i \left[\frac{y_i^2}{2} f(P) \right. \right. \right. \\ \left. \left. \left. + x_i y_i \frac{\partial g(P)}{\partial s} \right] \right\} \right| \leq k_1 h^2, \end{aligned} \quad (2.11)$$

where k_1 is a constant independent of h .

We remark here that an $O(h)$ approximation to $\partial u / \partial n$ which is of positive type is easily obtained (cf. [11]). Choosing only two interior points P_1 and P_2 in (2.3) we obtain

$$-\tilde{\delta}_n V(0) \equiv \sum_{i=1}^2 b_i [V(P_i) - V(0)] = -V_n(0) + O\left[\sum_{i=1}^2 b_i (x_i^2 + y_i^2)\right], \quad (2.12)$$

provided

$$\begin{bmatrix} y_1 & y_2 \\ x_1 & x_2 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \quad (2.13)$$

Clearly $y_1, y_2 > 0, x_1 > 0, x_2 < 0$ will guarantee the existence of non-negative numbers b_1 and b_2 . Thus analogous to (2.11) we are led to

$$|\tilde{\delta}_n u(P) + \alpha(P)u(P) - g(P)| \leq \bar{k}_1 h. \quad (2.14)$$

III. Finite-Difference Analogue

As mentioned in the last section we place a square mesh of width h on the region R and call the mesh crossings *mesh points*. The set R_h will consist of those mesh points of R whose four nearest neighbors are in R . The intersection of the mesh with C_i will make up the set C_{ih} , $i = 1, 2$. The sets C_{ih}^* will denote those mesh points of R which are at a distance less than or equal to h (along the horizontal or vertical) from C_{ih} , $i = 1, 2$.

We define the following operators. At a point (x, y) of R_h , $\Delta_h V(x, y) = h^{-2} \{V(x+h, y) + V(x-h, y) + V(x, y+h) + V(x, y-h) - 4V(x, y)\}$. It is well known that for $u \in C^{(4)}(\bar{R})$

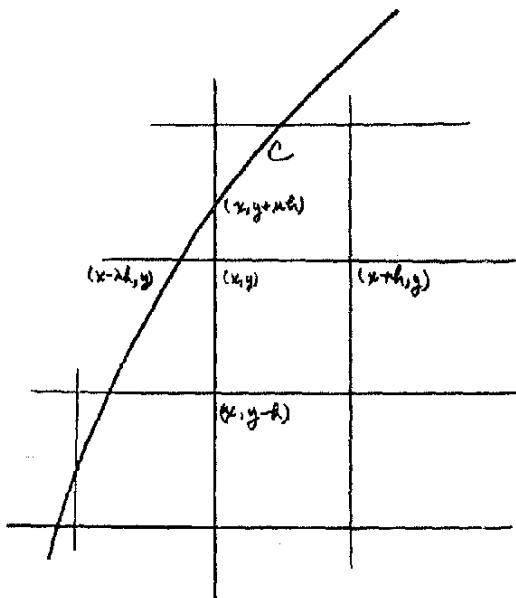


FIG. 2

$$|\Delta u(P) - \Delta_h u(P)| \leq M_4 h^2, \quad P \in R_h, \quad (3.1)$$

where M_4 is a constant depending on the fourth derivatives of u . On C_{ih}^* we use the operator of Shortley and Weller [12],

$$\begin{aligned} \Delta_h V(x, y) = & 2h^{-2} \left\{ \frac{1}{\lambda(1+\lambda)} V(x - \lambda h, y) + \frac{1}{1+\lambda} V(x + h, y) \right. \\ & + \frac{1}{\mu(1+\mu)} V(x, y + \mu h) + \frac{1}{1+\mu} V(x, y - h) \\ & \left. - [(1/\lambda) + (1/\mu)] V(x, y) \right\} \end{aligned}$$

For example, if (x, y) is the point (x, y) in Figure 2, then $0 < \lambda, \mu \leq 1$. The inequality

$$|\Delta u(P) - \Delta_h u(P)| \leq M_3 h, \quad P \in C_{ih}^*, \quad (3.2)$$

where M_3 depends on the third derivatives of u , is easily verified.

We now pose the following finite-difference analogue of (1.1):

$$\begin{aligned} -\Delta_h U(P) &= f(P), \quad P \in R_h + C_{ih}^*, \quad i = 1, 2 \\ \delta_n U(P) + \alpha(P)U(P) &= g(P) + \sum_{i=1}^3 a_i(P) \left[\frac{y_i^2}{2} f(P) + x_i y_i \frac{\partial g}{\partial s}(P) \right], \\ P &\in C_{ih} \end{aligned} \quad (3.3)$$

$$U(P) = H(P), \quad P \in C_{2h}.$$

In Section 2 it was shown that the operator δ_n , with the desired properties could always be constructed if P is a point where C_1 is smooth and h is small enough. We emphasize, however, that we do not require that the points P_i be chosen as in Section 2. They must simply be chosen so that the corresponding a_i 's are non-negative and of course so that (2.11) is satisfied. We assume now that this has been done.

In all cases the matrix of the system (3.3) is of positive type for h sufficiently small that $\sum_{i=1}^3 a_i x_i y_i \alpha_s + \alpha \geq 0$ and therefore possesses a non-negative inverse (cf [4]). In the subsequent discussion we assume that this condition on h is satisfied. We thus may introduce Green's function corresponding to (3.3). Let $G_h(P, Q)$ be defined as

$$\begin{aligned} -\Delta_{h,i} G_h(P, Q) &= h^{-2} \delta(P, Q), \quad P \in R_h + C_{ih}^*, \quad i = 1, \\ \delta_n G_h(P, Q) + \alpha(P)G_h(P, Q) &= \delta(P, Q), \quad P \in C_{ih}, \\ G_h(P, Q) &= \delta(P, Q), \quad P \in C_{2h}, \end{aligned}$$

for $Q \in R_h + C_{ih}^* + C_{2h} + C_{1h} + C_{2h}$. Since $G_h(P, Q)$ is just the inverse of (3.3) multiplied by a diagonal matrix with positive diagonal elements, it follows also that $G_h(P, Q) \geq 0$. Now, using $G_h(P, Q)$ we have the relation

$$\begin{aligned}\Gamma(P) &= h^2 \sum_{Q \in R_h + C_{1h}^* + C_{2h}^*} G_h(P, Q) [-\Delta_h V(Q)] \\ &\quad + \sum_{Q \in C_{1h}} G_h(P, Q) [\delta_n V(Q) + \alpha V(Q)] + \sum_{Q \in C_{2h}} G_h(P, Q) V(Q).\end{aligned}\quad (3.4)$$

This follows immediately from the uniqueness of solutions of (3.3) for arbitrary right-hand side. Letting $V(P) = 1$ in (3.4), it follows that $\sum_{Q \in C_{2h}} G_h(P, Q) \leq 1$.

Now suppose R is such that there exists a function $\phi \in C^3(\bar{R})$ (ϕ has continuous third derivatives in the closure of R) satisfying

$$\begin{aligned}-\Delta\phi &\geq 1 \quad \text{in } R \\ \frac{\partial\phi}{\partial n} + \alpha\phi &\geq 1 \quad \text{on } C_1.\end{aligned}\quad (3.5)$$

Then for small enough h , $-\Delta_h\phi(P) \geq \frac{1}{2}$, $P \in R_h + C_{1h}^* + C_{2h}^*$, and $\delta_n\phi(P) + \alpha(P)\phi(P) \geq \frac{1}{2}$, $P \in C_{1h}$. Taking $V(P) = \phi(P)$ in (3.4) we have

$$\sum_{Q \in C_{1h}} G_h(P, Q) + h^2 \sum_{Q \in R_h + C_{1h}^* + C_{2h}^*} G_h(P, Q) \leq 4 |\phi|_M. \quad (3.6)$$

To estimate $\sum_{Q \in C_{1h}^* + C_{2h}^*} G_h(P, Q)$ we observe as in [4] that if W is the function which is zero on C and one in R then $-\Delta_h W(P) \geq h^{-2}$, $P \in C_{1h}^* + C_{2h}^*$ and $-\Delta_h W(P) = 0$, $P \in R_h$. Letting $V(P) = W(P)$ in (3.4) we have

$$\sum_{Q \in C_{1h}^* + C_{2h}^*} G_h(P, Q) \leq 1 + \max_{\tilde{Q} \in C_{1h}} \left[\sum_{i=1}^3 a_i(\tilde{Q}) \right] \sum_{Q \in C_{1h}} G_h(P, Q).$$

Now $1 = \sum_{i=1}^3 a_i y_i \geq [\sum_{i=1}^3 a_i] \min y_i \geq [\sum_{i=1}^3 a_i] 3h$ for any $P \in C_{1h}$. Thus $\sum_{i=1}^3 a_i \leq 1/3h$. Hence, using (3.6),

$$h \sum_{Q \in C_{1h}^* + C_{2h}^*} G_h(P, Q) \leq h + \frac{4}{3} |\phi|_M. \quad (3.7)$$

Actually we can obtain a sharper estimate for that part of the sum in (3.7) for $Q \in C_{2h}^*$. To do this let $\tilde{W}(P)$ be the function which has the values zero on C_{2h} and one otherwise. Then $-\Delta_h \tilde{W}(P) \geq h^{-2}$, $P \in C_{2h}^*$ and $-\Delta_h \tilde{W}(P) = 0$, $P \in R_h + C_{1h}^*$. Letting $V(P) = \tilde{W}(P)$ in (3.4) it follows immediately that

$$\sum_{Q \in C_{2h}^*} G_h(P, Q) \leq 1. \quad (3.8)$$

We are now in a position to prove the following theorem.

THEOREM 1. *Let $u \in C^4(\bar{R})$ be the solution of (1.1). Suppose that the function ϕ of (3.5) exists. Then $\epsilon(P) \equiv u(P) - U(P)$, $P \in R_h + C_{1h}^* + C_{2h}^* + C_{1h}$, where $U(P)$ is the solution of (3.3), satisfies the inequality*

$$\max_P |\epsilon(P)| \leq kh^2. \quad (3.9)$$

In (3.9) k is a constant which depends on u and ϕ but not on h .

PROOF. Let $V(P) = \epsilon(P)$ in (3.4). Thus

$$\begin{aligned}\epsilon(P) &= h^2 \sum_{Q \in R_h + C_{1h}^* + C_{2h}^*} G_h(P, Q) [-\Delta_h \epsilon(Q)] \\ &\quad + \sum_{Q \in C_{1h}} G_h(P, Q) [\delta_n \epsilon(Q) + \alpha(Q) \epsilon(Q)].\end{aligned}$$

Now since $G_h(P, Q) \geq 0$ we have

$$\begin{aligned} |\epsilon(P)| &\leq [h^2 \sum_{Q \in R_h} G_h(P, Q)] \max_{Q \in R_h} |\Delta_h \epsilon(Q)| \\ &+ [h \sum_{Q \in C_{1h}^* + C_2} G_h(P, Q)] \max_{Q \in C_{1h}^* + C_{2h}^*} |h|\Delta_h \epsilon(Q) | \\ &[\sum_{Q \in C_{1h}} G_h(P, Q)] \max_{Q \in C_{1h}} |\delta_n \epsilon(Q) + \alpha(Q)\epsilon(Q)|. \end{aligned} \quad (3.10)$$

From (3.1), (3.2) and (2.11) we have the estimates

$$\begin{aligned} |\Delta_h \epsilon(P)| &\leq M_4 h^2, \quad P \in R_h \\ |h\Delta_h \epsilon(P)| &\leq M_3 h^2, \quad P \in C_{1h}^* + C_{2h}^* \\ |\delta_n \epsilon(P) + \alpha(P)\epsilon(P)| &\leq k_1 h^2, \quad P \in C_{1h}. \end{aligned} \quad (3.11)$$

The result (3.9) now follows by inserting (3.6), (3.7) and (3.11) into (3.10). By reasoning quite parallel to that leading up to, and in, the proof of Theorem 1 it is possible to prove

THEOREM 2. *Let $u \in C^3(\bar{R})$ be the solution of (1.1). Suppose that the function ϕ of (3.5) exists. Then $\epsilon(P) = u(P) - U(P)$, $P \in R_h + C_{1h}^* + C_{2h}^* + C_{1h} + C_{2h}$, where $U(P)$ is the solution of (3.3) with the second equation replaced by $\delta_n U(P) + \alpha(P) U(P) = g(P)$ (see (2.12)), satisfies*

$$\max_P |\epsilon(P)| \leq \bar{k}h. \quad (3.12)$$

In (3.12) \bar{k} is a constant which depends on u and ϕ but not on h .

It is clear that if one is interested only in convergence proofs for either of the two difference methods given above, then the assumption that the second derivatives of u are continuous is a sufficient smoothness requirement.

We note that in the hypothesis of Theorem 1 the local truncation error (i.e. the error in approximating the equations (1.1)) is $O(h^2)$ on the sets C_{1h} and R_h . It was assumed that no error occurred on C_{2h} although it is clear from the development that an error of not worse than $O(h^2)$ would not have changed the final result. On the other hand the error committed on the sets C_{1h}^* and C_{2h}^* was $O(h)$. In spite of this apparent defect the truncation error itself was shown to be $O(h^2)$. Making use of the more refined estimate (3.8) it is easily seen that the contribution to the error from C_{2h}^* was in fact $O(h^3)$, hence on C_{2h}^* crude approximations could have been made.

It is the opinion of the present authors that under quite general conditions if a truncation error of $O(h^n)$ is desired, it is sufficient to have a local approximation of $O(h^{n-1})$ on C_{1h}^* and $O(h^{n-2})$ on C_{2h}^* , while at most of the other points the defect should be $O(h^n)$.

RECEIVED SEPTEMBER, 1963.

REFERENCES

1. ALLEN, D. N. DEG. *Relaxation Methods*. McGraw-Hill, New York, 1954.
2. BATSCHELET. Über die numerische Anlösung von Randwertproblemen bei elliptischen partiellen Differentialgleichungen. *Z. Angew. Math. Phys.* 3, 1952.

3. BRAMBLE, J. H. Fourth order finite-difference analogues of the Dirichlet problem for Poisson's equation in three and four dimensions. *Math. Comput.* 17 (1963), 217-222.
4. BRAMBLE, J. H. AND HUBBARD, B. E. On the formulation of finite-difference analogues of the Dirichlet problem for Poisson's equation. *Numer. Math.* 4 (1962), 313-327.
5. —— AND ——. A theorem on error estimation for finite-difference analogues of the Dirichlet problem for elliptic equations. *Contrib. Diff. Eq.* 2 (1963), 319-340.
6. —— AND ——. A priori bounds on the discretization error in the numerical solution of the Dirichlet problem. *Contrib. Diff. Eq.* 2 (1963), 229-252.
7. —— AND ——. On a finite-difference analogue of an elliptic boundary problem which is neither diagonally dominant nor of non-negative type. *J. Math. Phys.* 43 (1964), 117-135.
8. FORSYTHE, G., AND WASOW, W. *Finite Difference Methods for Partial Differential Equations*. Wiley, New York, 1960.
9. GERSCHGORIN, S. Fehlerabschätzung für das Differenzenverfahren zur Lösung Partieller Differentialgleichungen. *Z. Angew. Math. Mech.* 10 (1930), 373-382.
10. GREENSPAN, D. On the Numerical Solution of Problems Allowing Mixed Boundary Conditions. *Notices Am. Math. Soc.* 10 (1963), 92.
11. KANTOROVICH, L. AND KRYLOV, V. *Approximate Methods of Higher Analysis*. Noordhoff Ltd., Netherlands, 1958.
12. SHORTLEY, G. AND WELLER, R. The numerical solution of Laplace's equation. *J. Appl. Phys.* 9 (1938), 331-348.
13. SHAW, F. S. *An Introduction to Relaxation Methods*. Dover, New York, 1950.
14. SYNGE, J. L. AND SCHILD, A. *Tensor Calculus*. U. of Toronto Press, Toronto, 1952.
15. UHLMANN. Differenzenverfahren für die 2. und 3. Randwertaufgabe mit Rändern bei $\Delta u(x, y) = r(x, y, u)$. *Z. Angew. Math. Mech.* 38 (1958).
16. VISWANATHAN, R. V. Solution of Poisson's equation by relaxation method-normal gradient specified on curved boundaries. *Math. Tables Aids Comput.* 11 (1957).

*1.5. A FINITE DIFFERENCE ANALOG OF THE NEUMANN PROBLEM FOR
POISSON'S EQUATION*

**1.5 A finite difference analog of the Neumann problem for
Poisson's equation**

A finite difference analog of the Neumann problem for Poisson's equation [36]

J. SIAM NUMER. ANAL.
Ser. B, Vol. 2, No. 1
Printed in U.S.A., 1964

A FINITE DIFFERENCE ANALOG OF THE NEUMANN PROBLEM FOR POISSON'S EQUATION*

J. H. BRAMBLE AND B. E. HUBBARD†

1. Introduction. This paper is concerned with estimates for the order of convergence of certain discrete analogues of the Neumann problem for Poisson's equation. Compared with the Dirichlet problem this subject has received little attention in the literature. Three papers [12], [15], and [25] are of particular importance.

The paper by J. Geise [15] gives two finite difference analogues of the Neumann problem for Poisson's equation on a rectangle with error in approximating $\partial u / \partial n$ which are $O(h^2)$ and $O(h)$ respectively. The method of Fourier series is used to obtain estimates for the total error which are $O(h^2 |\log h|)$ and $O(h |\log h|)$. K. O. Friedrichs [12] formulates a finite difference analog for the Neumann problem for elliptic systems of second order and shows convergence in the mean. The method of approach used there is to apply a variational principle for the differential equation to piecewise linear functions. Volkov [25] derives estimates of the type $O(h^2 \log^2 h)$.

In this paper a finite difference analog is formulated with $O(h^2)$ local truncation error in such a manner that the matrix of the resulting linear system is of "positive type". A maximum principle is then applied to yield estimates for the order of convergence which are $O(h^2 |\log h|)$. An $O(h |\log h|)$ analog is also treated. In §3 examples are given which show that these estimates are sharp.

We note that by [9] the estimate (2.40) on $e(p)$ arising from (2.17) can be used to show that the various differences of $e(p)$ have the same order of convergence, i.e., $O(h^2 |\log h|)$ on any compact subset of R . Similar results hold for differences of $e(p)$ arising from (2.43).

2. The finite difference analog. Consider the Neumann problem for Poisson's equation:

$$(2.1) \quad \begin{aligned} -\Delta u &= f && \text{in } R, \\ \frac{\partial u}{\partial n} &= g && \text{on } C, \end{aligned}$$

where R is a bounded connected region with sufficiently smooth boundary C and $\partial/\partial n$ denotes the normal derivative taken in the outward direction.

* Received by the editors November 15, 1963, and in revised form July 13, 1964.

† Institute for Fluid Dynamics and Applied Mathematics, University of Maryland, College Park, Maryland. The work of the first named author was supported in part by a grant from the National Science Foundation, GP-3. The work of the second named author was supported in part by the Air Force Office of Scientific Research, Grant 62-454.

It follows immediately from Green's first identity that f and g cannot be chosen independently, but that they must satisfy the relation

$$(2.2) \quad \int_R f \, dv = \int_C g \, ds.$$

Furthermore it is clear that the solution of (2.1) is unique only up to an additive constant. This constant is normally determined by a normalization, e.g.,

$$(2.3) \quad \int_C u \, ds = 0.$$

Hence we shall consider the problem solved once any solution of (2.1) is obtained.

The authors [2], [3], [4], [7], [8] have previously posed finite difference analogues of the Dirichlet and mixed boundary value problems for elliptic operators. In each of these cases the matrix of the resulting system of linear equations was shown to be nonsingular. In fact, either the matrix of the original system or that of a "reduced" system was shown to have an inverse with nonnegative elements.

However, in the present situation, if one formulates a finite difference analog following the usual rules, then the matrix of the resulting system will be singular with rank one less than its order and a condition of the type (2.2) would need to be imposed on the right side of the linear system to insure consistency. Since the dimension of the null space of the matrix is one, we see that the solution would be unique only up to a vector in the null space, thus imitating the situation in (2.1).

We shall now define certain finite difference analogues of the differential operators involved in (2.1). First we place a square mesh of width h on the region R and call the intersections as well as the boundary crossings "mesh points". The set R_h will consist of those mesh points in R whose four nearest neighbors are also in R . The remaining mesh points in R will make up the set C_h^* . We note that each point of C_h^* has at least one adjacent point which is a boundary crossing. The boundary crossings themselves will be denoted by C_h . If we desire a uniform discretization error then previous experience (e.g., [3], [4], [7], [8]) indicates that the local truncation error, i.e., the error committed by substituting finite difference operators for differential operators, need not be uniform. In fact if we select the local truncation error on R_h and C_h to be $O(h^2)$ then experience indicates that we need only choose an $O(h)$ approximation at points of C_h^* . Hence, we define the following operators. At a point (x, y) of R_h let

$$(2.4) \quad \begin{aligned} \Delta_h V(x, y) &\equiv h^{-2} \{ V(x + h, y) + V(x - h, y) \\ &\quad + V(x, y + h) + V(x, y - h) - 4V(x, y) \}. \end{aligned}$$

It is well known that if $u \in C^4(\bar{R})$, then

$$(2.5) \quad |\Delta u(p) - \Delta_h u(p)| \leq M_4 h^2, \quad p \in R_h,$$

where M_4 is a constant depending on the fourth derivatives of u . On C_h^* we use the operator of Shortley and Weller [19]. For example, if (x, y) is a point of C_h^* with $(x - \lambda h, y), (x, y - \mu h) \in C_h$, where $0 < \lambda, \mu < 1$, then

$$(2.6) \quad \begin{aligned} \Delta_h V(x, y) \equiv 2h^{-2} \left\{ & \frac{1}{\lambda(1 + \lambda)} V(x - \lambda h, y) + \frac{1}{1 + \lambda} V(x + h, y) \right. \\ & + \frac{1}{\mu(1 + \mu)} V(x, y - \mu h) + \frac{1}{1 + \mu} V(x, y + h) \\ & \left. - \left(\frac{1}{\lambda} + \frac{1}{\mu} \right) V(x, y) \right\}. \end{aligned}$$

This is simply the five point divided difference analog of Δ and it is easily verified that

$$(2.7) \quad |\Delta_h u(p) - \Delta u(p)| \leq M_3 h, \quad p \in C_h^*,$$

where M_3 depends on the third derivatives of u .

The construction of an $O(h^2)$ approximation to the normal derivative at points of C_h is more complicated, particularly if we wish the resulting difference operator to be of "positive type" [11]. Many different authors have proposed analogues to $\partial u / \partial n$ among whom we might mention Batschelet [1], Friedrichs [12], Fox [13], Shaw [20], Uhlmann [22], and Viswanathan [24]. For a good account of this work the reader is referred to the recent book by L. Fox [13]. We shall use still another analog proposed by the authors [8]. In the above paper the authors prove that within a circle of radius βh about each point of C_h one can find three points of $R_h + C_h^*$ such that the resulting four point finite difference analog of $\partial u / \partial n$ has an $O(h^2)$ local truncation error. Furthermore the coefficients corresponding to the three points of $R_h + C_h^*$ will be nonpositive and that of the point of C_h will be positive, thus giving rise to an equation which is of positive type. As we shall see, this property is useful when showing that the inverse matrix of the finite difference problem exists and is nonnegative.

We shall consider an arbitrary point 0 on C at which C is smooth. Choose 0 to be the origin of a Cartesian coordinate system (x, y) such that the x -axis is tangent to C at 0. The positive y direction is taken along the inward normal. It can be shown, for any smooth enough function v , that

$$(2.8) \quad v_{xy} = -v_{ns} + Kv_s$$

at the origin (cf. Synge and Schild [21] for the use of geodesic normal coordinates). The subscripts denote the indicated partial differentiation, n being the outward normal direction, s arc length, and K the curvature of

C . Thus we have

$$(2.9) \quad v_{xy} = -\frac{\partial}{\partial s} (v_n) + Kv_s$$

at 0. Also, of course,

$$(2.10) \quad v_{yy} = \Delta v - v_{xx}.$$

Now consider the Taylor expansion of v about 0; i.e.,

$$(2.11) \quad \begin{aligned} v(P) &= v(0) + xv_x(0) + yv_y(0) \\ &+ \frac{1}{2}\{x^2v_{xx}(0) + 2xyv_{xy}(0) + y^2v_{yy}(0)\} + O(x^3 + y^3). \end{aligned}$$

We note that $v_x(0) = v_s(0)$ and $v_y(0) = -v_n(0)$. Thus, using (2.9) and (2.10),

$$(2.12) \quad \begin{aligned} v(P) &= v(0) + x[1 + yK(0)]v_s(0) - yv_n(0) \\ &+ \frac{1}{2}[x^2 - y^2]v_{xx}(0) + \frac{y^2}{2}\Delta v(0) - xyv_{ns} + O(x^3 + y^3) \end{aligned}$$

Let $P_i = (x_i, y_i) \in R_h + C_h^*$, $i = 1, 2, 3$. We wish to determine three numbers A_i , $i = 1, 2, 3$, such that

$$(2.13) \quad \begin{aligned} \sum_{i=1}^3 A_i\{v(p_i) - v(0)\} &= -v_n(0) \\ &+ \sum_{i=1}^3 A_i\left\{\frac{y_i^2}{2}\Delta v(0) - x_i y_i v_{ns}(0)\right\} + O\left(\sum_{i=1}^3 A_i[x_i^3 + y_i^3]\right). \end{aligned}$$

For (2.13) to hold for any v , A_i must satisfy

$$(2.14) \quad \begin{aligned} \sum_{i=1}^3 A_i y_i &= 1, \\ \sum_{i=1}^3 A_i x_i [1 + y_i K(0)] &= 0, \\ \sum_{i=1}^3 A_i [x_i^2 - y_i^2] &= 0. \end{aligned}$$

In [8] it is shown that the points P_i may be chosen so that $A_i \geq 0$, if h is sufficiently small. Thus we define the boundary operator

$$(2.15) \quad \delta_n V(p) = \sum_{i=1}^3 A_i\{V(p) - V(p_i)\}.$$

Furthermore as a consequence of (2.13) and the fact that $A_i = O(h^{-1})$ as is proved in [8], it can be shown that

$$(2.16) \quad \left| \delta_n u(p) - \left\{ g(p) + \sum_{i=1}^3 \left[A_i \frac{y_i^2}{2} f(p) + x_i y_i \frac{\partial g(p)}{\partial s} \right] \right\} \right| \leq k_0 h^2,$$

where k_0 is a constant independent of h . The construction of δ_n in general requires that the region have no acute corners. In the case of a rectilinear region however special considerations can be made as we shall see in §3.

Let $o \in R_h$ be a mesh point in the interior of R and define R_h' to be the set $R_h - o$. A finite difference analog of (2.1) which gives a consistent system of linear equations is

$$(2.17) \quad \begin{aligned} (a) \quad & -\Delta_h U(p) = f(p), & p \in R_h' + C_h^*, \\ (b) \quad & \delta_n U(p) = g(p) + \sum_{i=1}^3 A_i(p) \left[\frac{y_i^2}{2} f(p) + x_i y_i \frac{\partial g(p)}{\partial s} \right], \\ (c) \quad & U(o) = \text{a given value}. & p \in C_h, \end{aligned}$$

It is seen that the matrix of the system (2.17) satisfies the following definition [7].

DEFINITION 2.1. An $N \times N$ matrix B with elements b_{ij} is said to be of *positive type* if the following conditions are satisfied.

- (a) $b_{ij} \leq 0$, $i \neq j$,
- (b) $\sum_k b_{jk} \geq 0$ for all j , with $\sum_k b_{jk} > 0$ for $j \in J(B) \neq \emptyset$, $J(B) \subset \{1, 2, \dots, N\}$,
- (c) for $i \notin J(B)$ there exists a finite sequence of nonzero elements of the form $b_{ik_1}, b_{k_1 k_2}, \dots, b_{k_r j}$, where $j \in J(B)$. Such a sequence is called a *connection* in B from i to $J(B)$.

This definition is a modification of well known sufficient conditions for a matrix to be an M -matrix [23].

If B were the matrix of (2.17) then we see that $J(B) =$ (the index of o) and that every point of $R_h' + C_h^* + C_h$ is connected to $J(B)$.

We note that if (2.17c) were replaced by the condition

$$(2.18) \quad -\Delta_h U(o) = f(o),$$

then the resulting system would not be of positive type since $J(B)$ would be empty. In fact the matrix B would be singular with the vector $\eta(p)$,

$$(2.19) \quad \eta(p) = 1, \quad p \in R_h + C_h^* + C_h,$$

spanning the null space, i.e.,

$$(2.20) \quad B\eta = 0.$$

If $\bar{\eta}$ lies in the null space of B^T we see that the consistency condition in this case is

$$(2.21) \quad \sum_{p \in R_h + C_h^*} \bar{\eta}(p)f(p) + \sum_{p \in C_h} \bar{\eta}(p)g(p) = 0,$$

which is the analog of (2.2). In general (2.21) will not be satisfied and, since $\bar{\eta}$ is difficult to find, the proper modification of the right side of the linear system is not easy to achieve except for special geometries. The formulation of the finite difference analog (2.17), which we shall now consider, provides a consistent system, without altering the right side, by eliminating one equation and replacing it by the normalization (2.17c). As was observed by J. H. Geise [15] in the case of the rectangle, the resulting error estimate for a slightly different analog is $O(h^2 |\log h|)$ which for all practical purposes is as good as $O(h^2)$.

The factor $\log h$ arises from the fundamental solution of Δ_h in two dimensions which is discussed in the Appendix. Consequently we would expect that the same process in higher dimensions would yield materially lower rates of convergence as indeed can be shown to be the case. The finite difference analog of Giese differs from ours in two respects. First, in Giese's formulation, the condition (2.18) is added to (2.17) and the resulting problem is turned into a consistent system by adding an $O(h^2)$ term at each boundary equation. Secondly, equations of the type (2.17a) are given on C_h which introduces another band of mesh points outside of R . Centered difference approximations to $\partial u / \partial n$ are also given at points of C_h which completes the formulation. Each pair of equations at a given boundary point could be combined to give a limiting case of (2.17b).

It is easy to prove (cf [6]) that positive-type matrices are nonsingular and have inverses with nonnegative elements. Hence, we define $N(p, q)$, $p, q \in R_h + C_h^* + C_h$, by

$$(2.22) \quad \begin{aligned} -\Delta_{h,p} N(p, q) &= h^{-2} \delta(p, q), & p \in R_h' + C_h^*, \\ \delta_{n,p} N(p, q) &= h^{-1} \delta(p, q), & p \in C_h, \\ N(o, q) &= \delta(o, q), \end{aligned}$$

where the Kronecker delta is defined by

$$(2.23) \quad \delta(p, q) = \begin{cases} 1 & \text{if } p = q, \\ 0 & \text{if } p \neq q. \end{cases}$$

It follows that

$$(2.24) \quad N(p, q) \geq 0.$$

For any mesh function $V(p)$ it is clear that

$$(2.25) \quad \begin{aligned} V(p) &= h^2 \sum_{q \in R_h' + C_h^*} N(p, q) [-\Delta_h V(q)] \\ &\quad + h \sum_{q \in C_h} N(p, q) [\delta_n V(q)] + N(p, o) V(o). \end{aligned}$$

Substituting the function $V(p) \equiv 1$ gives

$$(2.26) \quad N(p, o) = 1, \quad p \in R_h + C_h^* + C_h.$$

Consequently we may rewrite (2.25) as

$$(2.27) \quad \begin{aligned} V(p) - V(o) &= h^2 \sum_{q \in R_h' + C_h^*} N(p, q) \{-\Delta_h V(q)\} \\ &\quad + h \sum_{q \in C_h} N(p, q) [\delta_n V(q)]. \end{aligned}$$

To obtain the desired bound on the discretization error we shall first show that

$$(2.28) \quad h^2 \sum_{q \in R_h' + C_h^*} N(p, q) = O(|\log h|).$$

This is accomplished by considering the finite difference Green's function of a corresponding Dirichlet problem. We define $G(p, q)$ by

$$(2.29) \quad \begin{aligned} -\Delta_{h,p} G(p, q) &= \delta(p, q) h^{-2}, \quad p \in R_h + C_h^*, \\ G(p, q) &= \delta(p, q), \quad p \in C_h. \end{aligned}$$

It is well known [3] that $G(p, q)$ exists and is nonnegative. If $g(p, q)$ is the Green's function of the corresponding continuous problem then it is shown in the Appendix that for C sufficiently smooth,

$$(2.30) \quad |G(p, o) - g(p, o)| = O(h^2), \quad p \in C_h,$$

and consequently

$$(2.31) \quad |\delta_{n,p} G(p, o) - \delta_{n,p} g(p, o)| = O(h), \quad p \in C_h.$$

Moreover, since o is far from C , we see that on smooth portions of C ,

$$(2.32) \quad \left| \frac{\partial}{\partial n_p} g(p, o) - \delta_{n,p} g(p, o) \right| = O(h^2).$$

If we assume for the moment that C has no corners then it follows from a lemma of Hopf [10, p. 327] and the fact that g has bounded derivatives in R that

$$(2.33) \quad \beta^{-1} \geq -\frac{\partial g}{\partial n_p}(p, o) \geq \beta > 0.$$

A consequence of the lemma is that the normal derivative of a function harmonic on a region with smooth boundary cannot vanish at points of the boundary where the function takes on either a maximum or a minimum.

If R_ϵ is the region R with a small circle about o deleted then we see that $-g(p, o)$ takes on a maximum (zero) at each point of C , from which we infer (2.33). If C has corners then the above argument may fail with the

result that $-\partial g(p, o)/\partial n_p = 0$ at these corners. Since (2.30) and (2.32) may be violated as well, we shall assume, for the sake of simplicity, that C is a smooth curve and later treat the case where C has corners. It now follows from (2.31), (2.32), and (2.33) that for h chosen sufficiently small,

$$(2.34) \quad \delta^{-1} \geq -\delta_{n,p} G(p, o) \geq \delta > 0, \quad p \in C_h.$$

We now substitute $V(p) = -G(p, o)$ into (2.27) and use (2.34) to obtain

$$(2.35) \quad \delta^{-1} h \sum_{q \in C_h} N(p, q) \geq -G(p, o) + G(o, o) \geq \delta h \sum_{q \in C_h} N(p, q).$$

Hence it follows from (2.35) and (4.8) that

$$(2.36) \quad h \sum_{q \in C_h} N(p, q) \leq k_1 |\log h|.$$

Now if $r(p)$ is defined to be the distance from o to p and $V(p) = -r^2(p)$ in (2.27), we infer from (2.36) that

$$(2.37) \quad h^2 \sum_{q \in R'_h + C_h^*} N(p, q) \leq k_2 |\log h|.$$

Further if we let $V(p)$ in (2.27) be given by

$$(2.38) \quad V(p) = \begin{cases} 1 & \text{if } p \in R_h + C_h^*, \\ 0 & \text{if } p \in C_h, \end{cases}$$

then (2.36) yields

$$(2.39) \quad h^2 \sum_{q \in C_h^*} N(p, q) \leq k_3 h |\log h|.$$

Let $u(p)$, $U(p)$ be the solutions of (2.1) and (2.17) respectively, and let $e(p) \equiv u(p) - U(p)$ be the discretization error; then from (2.24) and (2.27) we see that

$$(2.40) \quad \begin{aligned} |e(p)| &\leq \left\{ \max_{q \in R'_h} |\Delta_h e(q)| \right\} h^2 \sum_{q \in R'_h} N(p, q) \\ &+ \left\{ \max_{q \in C_h^*} |\Delta_h e(q)| \right\} h^2 \sum_{q \in C_h^*} N(p, q) \\ &+ \left\{ \max_{q \in C_h} |\delta_n e(q)| \right\} h \sum_{q \in C_h} N(p, q), \end{aligned}$$

where we assume that $u(o) = U(o)$. It now follows from (2.5), (2.7), (2.16), (2.36), (2.37), and (2.39) that

$$(2.41) \quad |e(p)| \leq kh^2 |\log h|,$$

provided that $u \in C^4$ in \bar{R} .

In the case where C has a finite number of corners the preceding analysis requires some modification. We assume that for any given mesh, δ_n can be defined at every point of C_h so that the resulting matrix is of positive type.

In order to establish the inequality (2.36) we may no longer use simply $G(p, o)$ in (2.34) since that inequality will not in general be satisfied on all of C_h . To avoid this apparent difficulty let C_1 be a smooth arc of nonzero length, whose end points are not corners. Let R_1 be a region contained in R which has o as an interior point and whose boundary is smooth and contains C_1 . We further insist that R_1 be so chosen that δ_n on C_1 involves only mesh points of R_1 . The preceding analysis holds for $G_1(p, o)$. Hence there is a constant δ independent of h (for sufficiently small h) such that (2.34) holds for $G_1(p, o)$, i.e.,

$$(2.42) \quad -\delta_n G_1(p, o) \geq \delta > 0$$

for any p on the mesh boundary of R_1 and in particular for $p \in C_{1h}$. We note from (2.15) that

$$(2.43) \quad \sum_{i=1}^3 A_i G_1(p_i, o) = \delta_n G(p, o) \geq \delta > 0.$$

On the other hand it was shown in [5] that

$$(2.44) \quad G(p, o) \geq G_1(p, o)$$

for points at which G and G_1 are commonly defined. It follows then that

$$(2.45) \quad -\delta_n G(p, o) \geq \delta > 0$$

for $p \in C_{1h}$.

Let ϕ be a smooth function satisfying

$$(2.44) \quad \begin{aligned} -\Delta\phi &\geq 1 \quad \text{in } R, \\ \frac{\partial\phi}{\partial n} &\geq 1 \quad \text{on } C - C_1, \\ \left| \frac{\partial\phi}{\partial n} \right| &< \delta_1 \quad \text{on } C_1, \end{aligned}$$

where δ_1 is some constant. Thus we note that

$$(2.47) \quad \begin{aligned} -\Delta_h \left[-G(p, o) + \frac{\delta}{2\delta_1} \phi(p) \right] &\geq \frac{\delta}{2\delta_1} + O(h) \geq \lambda, \quad p \in R_h' + C_h^*, \\ \delta_n \left[-G(p, o) + \frac{\delta}{2\delta_1} \phi(p) \right] &\geq \frac{\delta}{2} + O(h^2) \geq \lambda, \quad p \in C_{1h}, \\ \delta_n \left[-G(p, o) + \frac{\delta}{2\delta_1} \phi(p) \right] &\geq \frac{\delta}{2\delta_1} + O(h^2) \geq \lambda, \quad p \in C_h - C_{1h}, \end{aligned}$$

where λ is a positive constant, if h is sufficiently small. Inequalities (2.36) and (2.37) follow immediately if we take

$$V(p) = -G(p, o) + \frac{\delta}{2\delta_1} \phi(p)$$

in (2.27).

We remark here than an $O(h)$ approximation to $\partial u / \partial n$ which is of positive type is easily obtained (cf. Kantorovich and Krylov [16]). Choosing only two interior points P_1 and P_2 in (2.11) we have

$$\begin{aligned} -\bar{\delta}_n V(0) &\equiv \sum_{i=1}^2 b_i[V(p_i) - V(0)] \\ (2.48) \quad &= -V_n(0) + O\left[\sum_{i=1}^2 b_i(x_i^2 + y_i^2)\right], \end{aligned}$$

provided

$$(2.49) \quad \begin{bmatrix} y_1 & y_2 \\ x_1 & x_2 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

Clearly $y_1, y_2 > 0, x_1 > 0, x_2 < 0$ will guarantee the existence of non-negative numbers b_1 and b_2 . Thus, analogous to (2.16), we are led to

$$(2.50) \quad |\bar{\delta}_n u(p) - g(p)| \leq \bar{k}_1 h.$$

The finite difference problem then becomes

$$\begin{aligned} -\Delta_h U(p) &= f(p), \quad p \in R_h' + C_h^*, \\ (2.51) \quad \bar{\delta}_n U(p) &= g(p), \quad p \in C_h, \\ U(0) &= u(0). \end{aligned}$$

Following the same method of proof it can be shown that for this analog,

$$(2.52) \quad |e| \leq \bar{k}h |\log h|,$$

if $u \in C^3$ in \bar{R} .

In the next section examples will be given which show that the error estimates (2.41) and (2.46) are, in fact, sharp.

3. Sharpness of the estimates. We first give an example using (2.17) for which the order of convergence is no better than $h^2 |\log h|$. Let R be the rectangle with vertices $(0, 0), (0, 1), (1, 1), (1, 0)$. In this case δ_n degenerates into the three point formula. At a typical point, say $(1, y)$,

$$\begin{aligned} (3.1) \quad \delta_n V(1, y) \\ &\equiv h^{-1}\{V(1, y) - \frac{1}{2}[V(1-h, y+h) + V(1+h, y-h)]\}. \end{aligned}$$

We see that

$$(3.2) \quad \delta_n u = \frac{\partial u}{\partial n} - \frac{h}{2} \Delta u + \frac{h^2}{6} F(u) + O(h^3),$$

where

$$(3.3) \quad F(u) = \begin{cases} u_{xxx} + 3u_{xyy}, & x = 1, \\ -u_{xxx} - 3u_{xyy}, & x = 0, \\ u_{yyy} + 3u_{xxy}, & y = 1, \\ -u_{yyy} - 3u_{xxy}, & y = 0. \end{cases}$$

Consider the function

$$(3.4) \quad u = x^2y^2 + (x - 1)^2(y - 1)^2,$$

and define the finite difference analog based on (2.17), i.e.,

$$(3.5) \quad \begin{aligned} -\Delta_h U(p) &= -\Delta u, \quad p \in R_h' + C_h^*, \\ \delta_n U(p) &= \frac{\partial u}{\partial n} - \frac{h}{2} \Delta u, \quad p \in C_h, \\ U(o) &= u(o). \end{aligned}$$

We note that for $e(p) \equiv u(p) - U(p)$,

$$(3.6) \quad \begin{aligned} \Delta_h e(p) &= \frac{h^2}{12} [u_{xxxx}(p) + u_{yyyy}(p)] = 0, \\ \delta_n e(p) &= \frac{h^2}{6} F(u) + O(h^3) = [2 + O(h)]h^2. \end{aligned}$$

Consequently we see that (2.27) becomes

$$(3.7) \quad e(p) = h \left\{ \sum_{q \in C_h} N(p, q) [2 + O(h)]h^2 \right\}.$$

By considerations similar to those which produced (2.36) we see that for p bounded away from o ,

$$(3.8) \quad h \sum_{q \in C_h} N(p, q) \geq k_4 |\log h|.$$

It now follows from (3.7) and (3.8) that

$$(3.9) \quad e(p) \geq k_5 h^2 |\log h|.$$

In the same manner we can exhibit an example of convergence of the solution of (2.51) which is no better than $h |\log h|$.

For $\bar{\delta}_n$ we need only the two point operator, e.g.,

$$(3.10) \quad \bar{\delta}_n V(1, y) \equiv h^{-1}[V(1, y) - V(1, y - h)].$$

Clearly,

$$(3.11) \quad \bar{\delta}_n u = \frac{\partial u}{\partial n} + \frac{h}{2} \bar{F}(u),$$

where

$$(3.12) \quad \bar{F}(u) = \begin{cases} u_{xx}, & x = 0, 1, \\ u_{yy}, & y = 0, 1. \end{cases}$$

Let u be the function

$$(3.13) \quad u = x^2 + y^2.$$

We see that $e(p)$ satisfies the identity

$$(3.14) \quad e(p) = h^2 \sum_{q \in C_h} N(p, q).$$

If p is bounded away from o , then (3.14) shows that

$$(3.15) \quad e(p) \geq k_4 h |\log h|.$$

The overdetermined system (3.5) with R_h' replaced by R_h in this example is clearly incompatible. Indeed, if it were compatible then (3.6) shows that $\Delta_h e(o) = 0$. We note that since $\Delta_h G(o, o) = h^{-2}$, the analog of (2.35) for (2.51) yields for p_i , the four neighbors of o ,

$$(3.16) \quad \sum_{i=1}^4 h \sum_{q \in C_h} N(p_i, q) \geq -\delta h^2 \Delta_h G(o, o) = \delta.$$

Hence, since $e(o) = 0$, we see that $\Delta_h e(o) = O(h^{-1})$, which clearly is a contradiction.

4. Appendix. In this section the estimate (2.30) is developed. Let $\bar{\Delta}_h$ be the five point operator (2.4) at each grid point in the (x, y) -plane. We know from the work of McCrea and Whipple (cf. [11, p. 317]) that there exists a fundamental solution $\Gamma(p, o)$ of

$$(4.1) \quad \bar{\Delta}_{h,p} \Gamma(p, o) = -h^2 \delta(p, o),$$

with the property that

$$(4.2) \quad \left| \Gamma(p, o) + \frac{1}{2\pi} \log r_{po} \right| = O(h^2)$$

for $\overline{po} \geq \delta > 0$. Define $\Phi(p)$ and $\phi(p)$ by the relations

$$(4.3) \quad \begin{aligned} G(p, o) &= \Gamma(p, o) + \Phi(p), \\ g(p, o) &= -\frac{1}{2\pi} \log r_{po} + \phi(p). \end{aligned}$$

We note that $\Phi(p) - \phi(p)$ has the following properties:

$$(4.4) \quad \begin{aligned} \Phi(p) - \phi(p) &= 0, \quad p \in C_h, \\ \Delta_h[\Phi(p) - \phi(p)] &= \begin{cases} O(h^2) & \text{if } p \in R_h, \\ O(1) & \text{if } p \in C_h^*, \end{cases} \end{aligned}$$

where the last statement holds for regions with sufficiently smooth boundaries (cf. Laasonen [17]). As was pointed out in [3], $\Phi(p) - \phi(p)$ satisfies the finite difference Poisson formula

$$(4.5) \quad \Phi(p) - \phi(p) = h^2 \sum_{q \in R_h + C_h^*} G(p, q) [-\Delta_h(\Phi(q) - \phi(q))].$$

It is further shown in that paper that

$$(4.6) \quad \begin{aligned} h^2 \sum_{q \in R_h} G(p, q) &= O(1), \\ \sum_{q \in C_h^*} G(p, q) &= O(1). \end{aligned}$$

Hence, from (4.4), (4.5), and (4.6) we see that

$$(4.7) \quad |\Phi(p) - \phi(p)| = O(h^2).$$

The relation (2.30) now follows from (4.2), (4.3), and (4.7). We note the close connection between (2.28) and Theorem 23.8 of Forsythe and Wasow [11].

In [18] P. Laasonen shows that the finite difference Green's function for a rectangle behaves like $\log h$ at the singular point. It was shown in [5] that $G(p, q)$ is a monotone function of region and hence it is true that

$$(4.8) \quad |G(p, q)| \leq \alpha |\log h|.$$

REFERENCES

- [1] E. BATSCHET, *Über die numerische Auflösung von Randwertproblemen bei elliptischen partiellen Differentialgleichungen*, Z. Angew. Math. Phys., 3 (1952), pp. 165–193.
- [2] J. H. BRAMBLE, *Fourth order finite difference analogues of the Dirichlet problem for Poisson's equation in three and four dimensions*, Math. Comp., 17 (1963), pp. 217–222.
- [3] J. H. BRAMBLE AND B. E. HUBBARD, *On the formulation of finite difference analogues of the Dirichlet problem for Poisson's equation*, Numer. Math., 4 (1962), pp. 313–327.
- [4] ———, *A theorem on error estimation for finite difference analogues of the Dirichlet problem for elliptic equations*, Contributions to Differential Equations, 2 (1963), pp. 319–340.
- [5] ———, *A priori bounds on the discretization error in the numerical solution of the Dirichlet problem*, Ibid., 2 (1963), pp. 229–252.
- [6] ———, *On a finite difference analogue of an elliptic boundary problem which is*

- neither diagonally dominant nor of non-negative type, *J. Math. and Phys.*, 43 (1964), pp. 117–132.
- [7] ———, New monotone type approximations for elliptic problems, *Math. Comp.*, 18 (1964), pp. 349–367.
- [8] ———, Approximation of solutions of mixed boundary value problems for Poisson's equation by finite differences, (to appear).
- [9] ———, Approximation of derivatives by finite difference methods in elliptic boundary value problems, *Contributions to Differential Equations*, 2 (1964), pp. 399–410.
- [10] COURANT AND HILBERT, *Methods of Mathematical Physics*, vol. II., Interscience, New York, 1961.
- [11] G. FORSYTHE AND W. WASOW, *Finite Difference Methods for Partial Differential Equations*, Wiley, New York, 1960.
- [12] K. O. FRIEDRICHHS, A finite difference scheme for the Neumann and the Dirichlet problem, N.Y.O.-9760, 1962.
- [13] L. FOX, *Numerical Solution of Ordinary and Partial Differential Equations*, Pergamon Press, Oxford, 1962.
- [14] S. GERSCHGORIN, Fehlerabschätzung für das Differenzverfahren zur Lösung Partieller Differentialgleichungen, *Z. Angew. Math. Mech.*, 10 (1930), pp. 373–382.
- [15] J. H. GIENSE, On the truncation error in a numerical solution of the Neumann problem for a rectangle, *J. Math. Phys.*, 37 (1958), pp. 169–177.
- [16] L. KANTOROVICH AND V. KRYLOV, *Approximate Methods of Higher Analysis*, Noordhoff Ltd., Netherlands, 1958.
- [17] P. LAASONEN, On the behavior of the solution of the Dirichlet problem for analytic corners, *Ann. Acad. Sci. Fenn. Ser. A I.*, 241 (1957).
- [18] ———, On the solution of Poisson's difference equations, *J. Assoc. Comput. Mach.*, 5 (1958), pp. 370–382.
- [19] G. SHORTLEY AND R. WELLER, The numerical solution of Laplace's equation, *Appl. Phys.*, 9 (1938), pp. 334–348.
- [20] F. S. SHAW, *An Introduction to Relaxation Methods*, Dover, New York, 1950.
- [21] J. L. SYNGE AND A. SCHILD, *Tensor Calculus*, University of Toronto Press, Toronto, 1952.
- [22] W. UHLMANN, Differenzenverfahren für die 2. und 3. Randwertaufgabe mit krummlinigen Rändern bei $\Delta u(x, y) = r(x, y, u)$, *Z. Angew. Math. Mech.*, 38 (1958), pp. 236–251.
- [23] R. VARGA, *Matrix Iterative Analysis*, Prentice Hall, Englewood Cliffs, New Jersey, 1962.
- [24] R. V. VISWANATHAN, Solution of Poisson's equation by relaxation method—Normal gradient specified on curved boundaries, *MTAC*, 11 (1957), pp. 67–78.
- [25] YE. A. VOLKOV, On the method of nets for a boundary problem with an oblique and normal derivative, *U.S.S.R. Comput. Math. and Math. Phys.*, 2 (1961) pp. 705–722.

1.6 A second order finite difference analog of the first biharmonic boundary value problem

A second order finite difference analog of the first biharmonic boundary value problem [7]

A Second Order Finite Difference Analog of the First Biharmonic Boundary Value Problem*

J. H. BRAMBLE

Received December 10, 1965

1. Introduction

In recent years there has been much work on difference methods for approximating solutions to boundary value problems for elliptic partial differential equations. Most of this work has centered on second order equations and strong use has been made of an analogous maximum principle for the difference equations in order to obtain asymptotic estimates for the error (c.f. [1–8]).

As early as 1928 COURANT, FRIEDRICH and LEWY [9] posed a difference analog for the first boundary value problem for the biharmonic equation and proved that the approximate solutions converge to the exact solution as the mesh is refined. They gave, however, no estimates for the error. Recently, THOMÉE [12] treated the Dirichlet problem for a class of elliptic equations of order $2m$ with constant coefficients. Among the problems treated by THOMÉE was that of [9]. THOMÉE, however, proved that in a certain norm the error is $O(h^4)$ where h is the mesh size.

Still more recently ZLÁMAL [13] posed a different difference analog for a fourth order elliptic operator with variable coefficients which includes the biharmonic operator in two dimensions. He proved that for his problem the error is $O(h^4)$.

This paper is concerned with the first boundary value problem for the biharmonic operator in the plane:

$$(1.1) \quad \begin{aligned} \Delta^2 u &= F \quad \text{in } R \\ u &= \frac{\partial u}{\partial n} = 0 \quad \text{on } \dot{R} \end{aligned}$$

where Δ is the Laplace operator, R is a bounded region with boundary \dot{R} , F is a given function in R and $\partial u / \partial n$ is the outward normal derivative on \dot{R} . The boundary conditions are taken to be homogeneous for convenience, this restriction being removed in the appendix.

Section 2 simply gives some notation and definitions needed for the later sections.

In Section 3 some basic lemmas are proved which provide some a priori estimates needed later. The third of these lemmas, Lemma 3.3, is proved by using difference inequalities closely related to inequalities used by MIRANDA [11] in proving his biharmonic maximum principle in the plane. This inequality did not, however, lead to a discrete analog of MIRANDA's maximum principle but

* This research was supported in part by the National Science Foundation under Grant NSF GP-3666.

was used in conjunction with a method related to one of FICHERA, [10], in order to deduce inequalities for certain discrete L_2 norms.

Section 4 makes use of the lemmas of the previous section to show that in the case of the biharmonic operator in the plane the results of THOMÉE [12] and ZLÁMAL [13] can be improved.

In the final section a difference analog of problem (1.1) is constructed for regions with boundaries of arbitrary shape (only piecewise smooth). It is shown that the error is $O(h^2)$ in certain norms and $O(h^2|\log 1/h|^\frac{1}{2})$ in maximum norm. Strong use is made of the lemmas of Section 3. Furthermore, the matrix of the resulting linear system is symmetric and positive definite and a general formula is given for the construction of the modified matrix near the boundary. It should be pointed out that, while in many second order problems difference methods were often formulated and later estimates given, this represents the first time an $O(h^2)$ method has even been formulated for the first boundary problem for any higher order elliptic partial differential equation in a general domain.

Although the results of this paper are special in that only the biharmonic operator in two dimensions is treated it is hoped that some of the ideas will lead to similar results for more general elliptic operators in regions of general shape.

2. Notation and Definitions

Let R be a bounded, open, connected set in the (x, y) plane. We denote by \dot{R} the boundary of R and $\bar{R} = R \cup \dot{R}$.

We shall be concerned with difference approximations of problem (1.1). In order to study such problems we cover the (x, y) plane with a square mesh of width h , whose lines are parallel to the x and y axes. The intersection of these lines will be called mesh (or grid) points and the set of all such mesh points will be denoted by S_h .

For any function $V(x, y)$ defined at the mesh points we define in the usual way the following difference operators:

$$(2.1) \quad \begin{aligned} V_x(x, y) &= \frac{1}{h} [V(x + h, y) - V(x, y)] \\ V_{\bar{x}}(x, y) &= \frac{1}{h} [V(x, y) - V(x - h, y)] \end{aligned}$$

and analogously for y . Further

$$(2.2) \quad \Delta_h V(x, y) = V_{\bar{x}\bar{y}}(x, y) + V_{y\bar{y}}(x, y)$$

where

$$V_{\bar{x}\bar{y}}(x, y) = (V_x)_{\bar{x}}(x, y).$$

The analog of the biharmonic operator is the usual 13 point operator

$$(2.3) \quad \Delta_h^2 V(P) = \Delta_h (\Delta_h V(P))$$

with $P = (x, y)$.

We also need to define certain norms. Let Q_h be an arbitrary bounded subset of S_h defined for each h and let V be any mesh function such that $V(P) = 0$,

$P \notin Q_h$. Then we define

$$(2.4) \quad \|V\| = \left(h^2 \sum_{P \in S_h} V^2(P) \right)^{\frac{1}{2}}$$

and for any integer p

$$(2.5) \quad \|V\|_{(2p)} = \left(h^{2p} \sum_{P \in S_h} V^{2p}(P) \right)^{1/2p}.$$

Now if $V(P)$ is defined on Q_h (otherwise arbitrary) we define

$$(2.6) \quad \|V\|_{Q_h} = \left(h^N \sum_{P \in Q_h} V^2(P) \right)^{\frac{1}{2}}$$

where N is the smallest integer such that the number of points in Q_h is $O(h^{-N})$ for $h \rightarrow 0$. Again if $V(P) = 0$, $P \notin Q_h$ we define

$$(2.7) \quad \|\delta V\| = \left(h^2 \sum_{P \in S_h} [V_x^2(P) + V_y^2(P)] \right)^{\frac{1}{2}}.$$

Finally the maximum norm is given by

$$(2.8) \quad |V|_{Q_h} = \max_{P \in Q_h} |V(P)|$$

for any V defined on Q_h .

We shall need some names for neighborhoods of a point P relative to the operators Δ_h and Δ_h^2 . Thus let $\varphi(P) = \delta(P, P_0)$ (Kronecker's delta) and define

$$(2.9) \quad N_1(P_0) = \{P \mid \Delta_h \varphi(P) \neq 0\}$$

and

$$(2.10) \quad N_2(P_0) = \{P \mid \Delta_h^2 \varphi(P) \neq 0\}.$$

If we take Q_h to be an arbitrary subset of S_h then

$$(2.11) \quad N_i(Q_h) = \bigcup_{P \in Q_h} N_i(P), \quad i = 1, 2.$$

Throughout this paper, we shall use the symbol, C , to denote a generic constant which does not depend on h . In two different places C will not necessarily refer to the same constant.

3. Some Discrete a Priori Inequalities

In this section we shall prove some lemmas which will be used in obtaining our error estimates in the later sections.

The following lemma is the discrete analog of a well known inequality valid in two dimensions.

Lemma 3.1. Let $V(P)$ be any function defined at the mesh points, which vanishes outside a bounded set of mesh points, R_h . Then for any integer $p \geq 1$,

$$(3.1) \quad \|V\|_{(2p)} \leq C_p \|\delta V\|,$$

where C_p is a constant which depends on p and R , but not on h .

Proof. For any mesh point $(\bar{x}, \bar{y}) \in R_h$

$$(3.2) \quad \begin{aligned} |V^p(\bar{x}, \bar{y})| &\leq \frac{1}{2} h \sum_x |(V^p)_x|, \\ |V^p(\bar{x}, \bar{y})| &\leq \frac{1}{2} h \sum_y |(V^p)_y| \end{aligned}$$

where the \sum_x is taken over the mesh points along the line $y = \bar{y}$ and similarly for \sum_y .

Thus

$$(3.3) \quad \|V\|_{(2,p)}^{2p} \leq \frac{1}{4} \left[h^2 \sum_{S_h} |(V^p)_x| \right] \left[h^2 \sum_{S_h} |(V^p)_y| \right].$$

By factoring the differences $(V^p)_x$ and using Schwarz's inequality we see that there is a constant C depending only on p and R such that

$$(3.4) \quad h^2 \sum_{S_h} |(V^p)_x| \leq C \|V_x\| \|V\|_{(2(p-1))}^{p-1},$$

with a similar inequality for the other factor in (3.3). Thus

$$(3.5) \quad \|V\|_{(2,p)}^{2p} \leq C \|\delta V\|^2 \|V\|_{(2(p-1))}^{2(p-1)}.$$

Iterating this inequality p times we obtain (3.1).

We note that the constant C_p in (3.1) tends to infinity as $p \rightarrow \infty$ so that we do not obtain a maximum norm estimate. The next lemma, however, enables us to obtain an estimate for the maximum norm.

Lemma 3.2. Let $V(p)$ be any function defined at the mesh points which vanishes outside R_h . Then

$$(3.6) \quad |V|_{R_h} \leq C |\log 1/h|^{\frac{1}{2}} \|\delta V\|.$$

Proof. Let $G(P, Q)$ be the Green's function defined by

$$\begin{aligned} \Delta_{h,p} G(P, Q) &= -h^{-2} \delta(P, Q), & P \in R_h \\ G(P, Q) &= 0, & P \notin R_h \end{aligned}$$

for $Q \in S_h$. Here again $\delta(P, Q)$ is the Kronecker delta. Now since $V(P) = 0$, $P \notin R_h$ we have the well known relation

$$(3.7) \quad \begin{aligned} V(P) &= -h^2 \sum_{Q \in S_h} G(P, Q) \Delta_h V(Q) \\ &= h^2 \sum_{Q \in S_h} [G_x(P, Q) V_x(Q) + G_y(P, Q) V_y(Q)]. \end{aligned}$$

Using Schwarz's inequality we have

$$(3.8) \quad |V(P)| \leq \left(h^2 \sum_{Q \in S_h} [G_x^2(P, Q) + G_y^2(P, Q)] \right)^{\frac{1}{2}} \|\delta V\|.$$

If we note that $G(P, Q) = G(Q, P)$ and set $V(S) = G(S, P)$ in (3.7) we see that

$$(3.9) \quad G(P, P) = h^2 \sum_{Q \in S_h} [G_x^2(P, Q) + G_y^2(P, Q)].$$

But it was shown in [3] that there is a constant C independent of h such that

$$(3.10) \quad G(P, P) \leq C |\log 1/h|.$$

Inequality (3.6) follows now from (3.8)–(3.10).

The next lemma is an inequality specifically involving the discrete biharmonic operator Δ_h^2 . We shall need to define some sets of mesh points. Let \dot{R}_h be a subset of those mesh points whose distance to the boundary is less than h . The set R_h will denote those mesh points of R not in \dot{R}_h and R'_h is the set of points $P \in R_h$ such that $N_2(P) \subset R_h$. Finally $R_h^* = R_h - R'_h$ with these sets defined we shall prove

Lemma 3.3. Let $V(P)$ be any mesh function vanishing for $P \notin R_h$. Suppose that the function Φ , defined by $\Delta_h \Phi(P) = -1$, $P \in R_h$; $\Phi(P) = 0$, $P \notin R_h$, satisfies $\Phi(P) \leq K h$ for $P \in R_h^*$, ($K = \text{constant}$).

Then there exists a constant C independent of h , for h sufficiently small, such that

$$(3.11) \quad \|V\| + \|\delta V\| \leq C\{h^{-1}\|V\|_{R_h^*} + \|\Delta_h^2 V\|_{R'_h}\}.$$

Proof. By a direct calculation we have

$$(3.12) \quad \begin{aligned} \Delta_h [\frac{1}{2}(V_x^2 + V_{\bar{x}}^2 + V_y^2 + V_{\bar{y}}^2) - V \Delta_h V] \\ = -V \Delta_h^2 V + \frac{1}{2}[V_{xx}^2 + 2V_{x\bar{x}}^2 + V_{\bar{x}\bar{x}}^2] + \frac{1}{2}[V_{yy}^2 + 2V_{y\bar{y}}^2 + V_{\bar{y}\bar{y}}^2] - \\ - (\Delta_h V)^2 + V_{xy}^2 + V_{\bar{x}y}^2 + V_{x\bar{y}}^2 + V_{\bar{x}\bar{y}}^2. \end{aligned}$$

Since

$$(3.13) \quad -(\Delta_h V)^2 \geq -2[V_{x\bar{x}}^2 + V_{y\bar{y}}^2]$$

it follows that

$$(3.14) \quad \begin{aligned} \Delta_h [\frac{1}{2}(V_x^2 + V_{\bar{x}}^2 + V_y^2 + V_{\bar{y}}^2) - V \Delta_h V] \\ \geq -V \Delta_h^2 V + \frac{1}{2}[V_{xx}^2 - 2V_{x\bar{x}}^2 + V_{\bar{x}\bar{x}}^2] + \frac{1}{2}[V_{yy}^2 - 2V_{y\bar{y}}^2 + V_{\bar{y}\bar{y}}^2] \\ = -V \Delta_h^2 V + h^2/2((V_{x\bar{x}}^2)_{x\bar{x}} + (V_{y\bar{y}}^2)_{y\bar{y}}). \end{aligned}$$

Now let

$$(3.15) \quad \chi = \frac{1}{2}V_x^2 + V_{\bar{x}}^2 + V_y^2 + V_{\bar{y}}^2 - V \Delta_h V$$

so that (3.14) is simply

$$(3.16) \quad -\Delta_h \chi \leq V \Delta_h^2 V - h^2/2((V_{x\bar{x}}^2)_{x\bar{x}} + (V_{y\bar{y}}^2)_{y\bar{y}}).$$

Now by hypothesis the mesh function Φ defined by

$$(3.17) \quad \begin{aligned} \Delta_h \Phi(P) &= -1, & P \in R_h \\ \Phi(P) &= 0, & P \notin R_h, \end{aligned}$$

satisfies

$$(3.18) \quad \Phi(P) \leq K h, \quad P \in R_h^*.$$

(We note here that it is easily shown that if R has a piecewise smooth boundary with no reentrant corners such a mesh function will exist. The next lemma is applicable in the more general case, allowing reentrant corners. The behaviour near the boundary of the discrete "torsion function" in this case is discussed in [8].)

Now we have

$$(3.19) \quad -h^2 \sum_{S_h} \chi \Delta_h \Phi = -h^2 \sum_{S_h} \Phi \Delta_h \chi.$$

Using (3.16)

$$(3.20) \quad \begin{aligned} -h^2 \sum_{S_h} \chi \Delta_h \Phi &\leq h^2 \sum_{R_h} (\Phi V \Delta_h^2 V) - \frac{h^4}{2} \sum_{S_h} [\Phi \{(V_{x\bar{x}}^2)_{x\bar{x}} + (V_{y\bar{y}}^2)_{y\bar{y}}\}] \\ &= h^2 \sum_{R_h} [\Phi V \Delta_h^2 V] + \frac{h^4}{2} \sum_{S_h} [\Phi_x (V_{x\bar{x}}^2)_x + \Phi_y (V_{y\bar{y}}^2)_y]. \end{aligned}$$

In order to obtain our results from (3.20) we need to show that the difference quotients Φ_x and Φ_y are uniformly bounded. This is clear since any first difference quotient, say Φ_x , satisfies

$$(3.21) \quad \begin{aligned} \Delta_h \Phi_x(p) &= 0, & P \in R'_h \\ |\Delta_h \Phi_x(p)| &\leq C/h^2, & P \in [N_1(R_h^*) - R'_h] \\ \Phi_x(p) &= 0, & P \notin N_1(\bar{R}_h) \end{aligned}$$

and hence by the results of [2]

$$(3.22) \quad |\Phi_x|_{S_h} \leq C, \quad |\Phi_y|_{S_h} \leq C.$$

We note now that,

$$(3.23) \quad h^2 \sum_{S_h} \chi = 2 \|\delta V\|^2.$$

Returning to (3.20) and using (3.23) we have

$$(3.24) \quad \begin{aligned} 2 \|\delta V\|^2 &= h^2 \sum_{S_h} \chi \leq \frac{h^2}{2} \sum_{S_h - R_h} (1 + \Delta_h \Phi) (V_x^2 + V_{\bar{x}}^2 + V_y^2 + V_{\bar{y}}^2) + \\ &\quad + \frac{h^4}{2} \sum_{S_h} [\Phi_x (V_{x\bar{x}}^2)_x + \Phi_y (V_{y\bar{y}}^2)_y] + \\ &\quad + h^2 \sum_{R_h} [\Phi V \Delta_h^2 V]. \end{aligned}$$

Using (3.22) we have, since $\sum_{S_h} |(V_{x\bar{x}}^2)_x| \leq \frac{2}{h} \sum_{S_h} V_{x\bar{x}}^2$ etc.,

$$(3.25) \quad \begin{aligned} h^4 \sum_{S_h} [\Phi_x (V_{x\bar{x}}^2)_x + \Phi_y (V_{y\bar{y}}^2)_y] &\leq C h h^2 \sum_{S_h} [V_{x\bar{x}}^2 + V_{y\bar{y}}^2] \\ &\leq C h h^2 \sum_{S_h} [V_{x\bar{x}}^2 + 2V_{x\bar{y}}^2 + V_{y\bar{y}}^2] \\ &= C h h^2 \sum_{R_h} V \Delta_h^2 V. \end{aligned}$$

Also from (3.18) the first term on the right of (3.24) is bounded in terms of $h^{-2} \|V\|_{N_1(R_h^*)}^2$. Thus we have

$$(3.26) \quad \begin{aligned} \|\delta V\|^2 &\leq C \left\{ h^{-2} \|V\|_{N_1(R_h^*)}^2 + h h^2 \sum_{R_h} |V \Delta_h^2 V| + h^2 \sum_{R_h} |V \Delta_h^2 V| \right\} \\ &\leq C \left\{ h^{-2} \|V\|_{N_1(R_h^*)}^2 + h^2 \sum_{R_h} |V \Delta_h^2 V| \right\}. \end{aligned}$$

From the well known inequality

$$(3.27) \quad \|V\| \leq C \|\delta V\|$$

we have, using the arithmetic-geometric mean inequality

$$(3.28) \quad \|\delta V\| \leq C\{h^{-1}\|V\|_{N_h(R_h^*)} + \|\Delta_h^2 V\|_{R_h'}\}$$

for some constant C independent of h . The estimate of THOMÉE

$$\bar{C} h^{-2} \|V\|_{N_h(R_h^*)} \leq \|\Delta_h V\| \leq C (\|\Delta_h^2 V\|_{R_h'} + h^{-2} \|V\|_{R_h^*})$$

together with (3.27) and (3.28) yields the desired result.

Lemma 3.3 is not quite general enough for domains with reentrant corners in that the inequality $\Phi(P) \leq Kh$ in the hypothesis will not be satisfied. However, if we weaken this hypothesis we can include such regions. It is clear, following the proof of Lemma 3.3, that the following lemma, which includes lemma 3.3, is true.

Lemma 3.4. Let $V(P)$ be any mesh function vanishing for $P \notin R_h$. Suppose that the function Φ defined by $\Delta_h \Phi(P) = -1$, $P \in R_h$, $\Phi(P) = 0$, $P \notin R_h$ satisfies $\Phi(P) \leq Kh^\alpha$, for $0 < \alpha \leq 1$, $P \in R_h^*$. Then there exists a constant C independent of h , for h sufficiently small, such that

$$(3.29) \quad \|V\| + \|\delta V\| \leq C h^{\frac{\alpha-1}{2}} \{h^{-1}\|V\|_{R_h^*} + \|\Delta_h^2 V\|_{R_h'}\}.$$

4. Application of Lemmas to the Results of Thomée and Zlámal

A particular case of those problems studied by THOMÉE [12] was problem (1.1). The difference problem in that case which he posed was that studied by COURANT, FRIEDRICH and LEWY [9] who showed convergence only. THOMÉE essentially gave the following result. Let $U(P)$ satisfy

$$(4.1) \quad \begin{aligned} \Delta_h^2 U(P) &= F(P), & P \in R_h \\ U(P) &= 0, & P \notin R_h. \end{aligned}$$

Then if $e(P) = u(P) - U(P)$, $P \in R_h$, $e(P) = 0$, $P \notin R_h$ and if R and u are sufficiently smooth, e satisfies

$$(4.2) \quad \|\Delta_h e\| \leq Ch^{\frac{1}{2}},$$

where C is independent of h . Now it follows at once from (4.2) and the fact that $e(P) = 0$, $P \notin R_h$, that

$$(4.3) \quad \|e\|_{R_h^*} \leq Ch^2.$$

Thus in the case that Lemma 3.3 is applicable (R piecewise smooth with no reentrant corners) we have

$$(4.4) \quad \|e\| + \|\delta e\| \leq Ch$$

and from Lemmas 3.1 and 3.2

$$(4.5) \quad \|e\|_{2p} \leq Ch$$

and

$$(4.6) \quad |e|_{R_h} \leq Ch |\log 1/h|^{\frac{1}{2}}.$$

In case Lemma 3.4 holds but not Lemma 3.3 the factor h on the right hand sides of (4.4) – (4.6) will simply be replaced by $h^{\frac{\alpha+1}{2}}$.

Recently ZLÁMAL [13] has posed a difference analog of (1.1) (and more general fourth order equations) in the case that R is composed of a finite number of

rectangles and \dot{R} lies on mesh lines for a sequence of meshes with mesh width h_n , $h_n \rightarrow 0$, $n \rightarrow \infty$.

In that the more general formulation in the next section includes his formulation as a special case, it will not be explicitly given here. We state, however his result and show how it may be extended by means of the lemmas.

Again let e be the error in ZLÁMAL's problem and take $e(P) = 0$ $P \notin R_h$. ZLÁMAL showed that

$$(4.7) \quad \|\Delta_h e\| \leq C h^{\frac{3}{2}}$$

and

$$(4.8) \quad |e|_{R_h} \leq C h^{\frac{3}{2}}.$$

Now in the case of a rectangle Lemma 3.3 applies and we obtain

$$(4.9) \quad \|e\| + \|\delta e\| \leq C h^2$$

and from Lemmas 3.1 and 3.2

$$(4.10) \quad \|e\|_{2,p} \leq C h^2$$

and

$$(4.11) \quad |e|_{R_h}^{\frac{1}{2}} \leq C h^2 |\log 1/h|^{\frac{1}{2}}.$$

Clearly (4.11) is a sharper result than (4.8). Now if R has reentrant corners then the interior angles will be $\vartheta = \frac{3\pi}{2}$, and it can be shown that we may take $\alpha = \frac{3}{2} - \varepsilon$ for any $\varepsilon > 0$, in Lemma 3.4. We then obtain instead of (4.9)–(4.11)

$$(4.12) \quad \|e\| + \|\delta e\| \leq C h^{11/6 - \varepsilon},$$

$$(4.13) \quad \|e\|_{2,p} \leq C h^{11/6 - \varepsilon},$$

$$(4.14) \quad |e|_{R_h} \leq C h^{11/6 - \varepsilon}$$

for any fixed $\varepsilon > 0$. Hence for any $\varepsilon < \frac{1}{3}$ (4.14) is a sharper estimate than (4.8).

5. Second Order Approximation

In order to simplify the presentation and proof we shall treat in detail only the case of simply connected regions with smooth boundaries. The modifications needed to deal with regions whose boundaries have piecewise continuous curvature (possibly corners) are technical and will not be of concern here. It will be evident from the development that the method may, in fact, be applied equally well to this more general class of domains.

We start by defining a set of mesh points which will be analogous to the boundary \dot{R} . Let $\dot{R}_{1,h}$ be the set of grid points not in R whose horizontal or vertical distance to \dot{R} is less than or equal to $2h/3$, and let $\dot{R}_{2,h}$ be those mesh points of R whose horizontal or vertical distance to \dot{R} is less than $h/3$. We set $\dot{R}_h = \dot{R}_{1,h} \cup \dot{R}_{2,h}$. The set of mesh points of R but not in \dot{R}_h will be called R_h . Further let us denote by R_h^* the subset of points $P \in R_h$ such that $\dot{R}_h \cap N_2(P)$ is not empty. Finally let $R'_h = R_h - R_h^*$ and $\bar{R}_h = R_h \cup \dot{R}_h$.

Now on R'_h we take the usual thirteen point operator

$$(5.1) \quad \Delta_h^2 V(P) = \Delta_h \Delta_h V(P)$$

for any V defined in R_h .

We want yet to define a difference operator $\bar{\Delta}_h^2$ on the set R_h^* . This operator should have the property that

$$(5.2) \quad \bar{\Delta}_h^2 v(P) = \Delta^2 v(P) + O(h^{-1}), \quad P \in R_h^*$$

for any $v \in C^{(4)}(\bar{R})$ which vanishes with its gradient on \dot{R} . As will be seen, this is a property which will be used in obtaining $O(h^2)$ estimates for the error in our problem. For simplicity we are treating only the case of homogeneous boundary conditions, however the necessary modification for inhomogeneous boundary conditions will be stated in the appendix.

In order to define the difference operator at a point $P \in R_h^*$ it is convenient to introduce the following sets:

$$(5.3) \quad \begin{aligned} J_0(P) &= N_1(P) \cap R_h - P \\ J_1(P) &= N_1(P) \cap \dot{R}_h \\ J_2(P) &= N_1(J_1(P)) \cap R_h - P. \end{aligned}$$

Now let V be any function, defined at the mesh points, which vanishes outside R_h . Further let $\alpha(P)h$ be the distance from any point P to the boundary \dot{R} . At an arbitrary point $P \in R_h^*$ we define a mesh function $U_P(Q)$ for each point $Q \in N_1(P)$ as

$$(5.4) \quad \begin{aligned} U_P(Q) &= \Delta_h V(Q) + h^{-2} \sum_{S \in J_1(Q)} \left(\frac{\alpha(S)}{\alpha(P)} \right)^2 V(P), \quad Q \in J_0(P) \cup P \\ U_P(Q) &= \frac{2}{\alpha^2(P)h^2} V(P), \quad Q \in J_1(P). \end{aligned}$$

In terms of $U_P(Q)$ we define the difference operator

$$(5.5) \quad \bar{\Delta}_h^2 V(P) = \Delta_h U_P(P),$$

for each $P \in R_h^*$. By looking at the Taylor expansion it may be directly verified that the difference operator defined by (5.4), (5.5) satisfies (5.2) in case the curvature of \dot{R} is piecewise continuous and \dot{R} has only "convex" corners. The extension of the results of this section to non-convex corners is technical and is omitted here for simplicity.

Intuitively, however, the construction is based on the following considerations. If W is a sufficiently smooth function defined in the whole plane, which vanishes with its gradient on \dot{R} , then it is easily verified that for $V(P) = W(P)$, $P \in R_h$,

$$(5.6) \quad U_P(Q) = \Delta W(Q) + O(h)$$

for $Q \in N_1(P)$. Thus (5.2) will obviously be satisfied for $V = W$.

It will be useful to compare this operator with the operator Δ_h^2 for functions $V(P) = 0$, $P \in R_h$. The relationship is easily seen to be

$$(5.7) \quad \bar{\Delta}_h^2 V(P) = \Delta_h^2 V(P) + h^{-4} \gamma(P) V(P) - h^{-4} \sum_{S \in J_1(P)} V(S)$$

where

$$(5.8) \quad \gamma(P) = \sum_{Q \in J_0(P)} \left\{ \sum_{S \in J_1(Q)} \left(\frac{\alpha(S)}{\alpha(P)} \right)^2 \right\} + \sum_{Q \in J_1(P)} \left\{ 2 \left[\frac{1}{\alpha^2(P)} - 2 \left(\frac{\alpha(Q)}{\alpha(P)} \right)^2 \right] - 1 \right\}.$$

Thus the modification of $\Delta_h^2 V(P)$ is simply a change in the coefficient of $V(P)$ itself and possibly a change of some other coefficients from 2 to 1. The number $\gamma(P)$ is easily calculated from the α 's.

The difference analog of (1.1) which we take is

$$(5.9) \quad \begin{aligned} \Delta_h^2 U(P) &= F(P), & P \in R'_h \\ \Delta_h^2 U(P) &= F(P), & P \in R_h^* \\ U(P) &= 0, & P \notin R_h. \end{aligned}$$

We have then the following theorem.

Theorem 5.1. There exists a unique solution of (5.9). Furthermore if $u \in C^{(6)}(\bar{R})$ is the solution to (1.1) and if $e(P) = U(P) - u(P)$, $P \in R_h$, $e(P) = 0$, $P \notin R_h$ then for h sufficiently small

$$(5.10) \quad \|e\| + \|\delta e\| \leq Ch^2$$

where C is a constant independent of h .

Proof. In order to simplify the argument we shall assume that the set \dot{R}_h has the property that each of its points has at least one horizontal and vertical neighbor not in R_h . We also suppose that h is chosen small enough that for each pair of points P and Q such that $P \in J_2(Q)$, the set $J_2(P) \cap J_2(Q)$ is a single point.

Let V be any mesh function which vanishes outside R_h . Then

$$h^2 \sum_{S_h} (\Delta_h V)^2 = h^2 \sum_{R_h} V \Delta_h^2 V.$$

By (5.7) we have

$$(5.11) \quad \begin{aligned} h^2 \sum_{S_h} (\Delta_h V)^2 + h^{-2} \sum_{P \in R_h^*} \left[\gamma(P) V(P) - \sum_{S \in J_2(P)} V(S) \right] V(P) \\ = h^2 \sum_{R_h} V \Delta_h^2 V + h^2 \sum_{R_h^*} V \Delta_h^2 V. \end{aligned}$$

Clearly from (5.7) and the definition of $J_2(P)$ the matrix of the system (5.9) is symmetric. By examining the left hand side of (5.11) we shall show that it is also positive definite.

Now let R_{0h}^* be the subset of points $P \in R_h^*$ for which $J_1(P)$ is empty. Then for $P \in R_{0h}^*$

$$(5.12) \quad \gamma(P) = \sum_{Q \in J_0(P)} \left\{ \sum_{S \in J_1(Q)} \left(\frac{\alpha(S)}{\alpha(P)} \right)^2 \right\} \geq 0$$

and $J_2(P)$ is also empty. Further let R_{1h}^* be that subset of R_h^* where $J_1(P)$ is not empty but $J_2(P)$ is empty and finally let R_{2h}^* be those points of R_h^* where neither $J_1(P)$ nor $J_2(P)$ is empty. We have

$$(5.13) \quad \begin{aligned} h^2 \sum_{S_h} (\Delta_h V)^2 + h^{-2} \sum_{P \in R_h^*} \left[\gamma(P) V(P) - \sum_{S \in J_2(P)} V(S) \right] V(P) \\ \geq h^2 \sum_{S_h} (\Delta_h V)^2 + h^{-2} \sum_{P \in R_{1h}^*} \gamma(P) V^2(P) + \\ + h^{-2} \sum_{P \in R_{2h}^*} \left[\gamma(P) V(P) - \sum_{S \in J_1(P)} V(S) \right] V(P). \end{aligned}$$

Now since $V(P)=0$, $P \notin R_h$

$$(5.14) \quad h^2 \sum_{S_h} (\Delta_h V)^2 = h^2 \sum_{S_h} [V_{xx}^2 + 2V_{xy}^2 + V_{yy}^2]$$

and hence because of the assumption on \dot{R}_h

$$(5.15) \quad h^2 \sum_{S_h} (\Delta_h V)^2 \geq h^2 \sum_{S_h - [R_{1h}^* \cup R_{2h}^*]} (\Delta_h V)^2 + h^{-2} \sum_{P \in R_{1h}^* \cup R_{2h}^*} j(P) V^2(P),$$

where $j(P)$ is the number of points in $J_1(P)$. Thus combining (5.13) and (5.15) we have

$$(5.16) \quad \begin{aligned} & h^2 \sum_{S_h} (\Delta_h V)^2 + h^{-2} \sum_{P \in R_h^*} [\gamma(P)V(P) - \sum_{S \in J_1(P)} V(S)] V(P) \\ & \geq h^2 \sum_{S_h - [R_{1h}^* \cup R_{2h}^*]} (\Delta_h V)^2 + h^{-2} \sum_{P \in R_{1h}^*} (\gamma(P) + j(P)) V^2(P) + \\ & \quad + h^{-2} \sum_{P \in R_{2h}^*} [(\gamma(P) + j(P)) v(P) - \sum_{S \in J_1(P)} V(S)] V(P). \end{aligned}$$

Now for any $P \in R_{1h}^* \cup R_{2h}^*$ it is easy to see from the definition of \dot{R}_h and $\gamma(P)$ that

$$(5.17) \quad \gamma(P) + j(P) \geq \frac{7}{8} j(P) \geq \frac{7}{8}.$$

This estimate will suffice for R_{1h}^* however we must examine R_{2h}^* more closely. From the definition of $\gamma(P)$ we have

$$(5.18) \quad \gamma(P) + j(P) \geq \sum_{Q \in J_1(P)} \gamma(P, Q)$$

where

$$\gamma(P, Q) = 2 \left[\frac{1}{\alpha^2(P)} - 2 \left(\frac{\alpha(Q)}{\alpha(P)} \right)^2 \right].$$

By the definition of $J_2(P)$ we have that $S \in J_2(P)$ if and only if $P \in J_2(S)$ and with each such pair P, S there is a unique $Q = J_1(P) \cap J_1(S)$. Thus we have

$$(5.19) \quad \begin{aligned} & h^{-2} \sum_{P \in R_{2h}^*} [\gamma(P) + j(P) V(P) - \sum_{S \in J_1(P)} V(S)] V(P) \\ & \geq h^{-2} \sum_{(P, Q, S) \in T} [\gamma(P, Q) V^2(P) - 2V(P) V(S) + \gamma(S, Q) V^2(S)] \end{aligned}$$

where T is the set of triples satisfying $P, S \in R_{1h}^*$, $S \in J_2(P)$, $Q = J_1(P) \cap J_1(S)$. Now if $Q \in \dot{R}_{1h}$ then it is easily verified that $\gamma(P, Q) \geq 2$. In case $Q \in \dot{R}_{2h}$ we could have at worst

$$\gamma(P, Q) \geq 2 \left[\frac{1-2\alpha^2}{(1+\alpha)^2} \right] \quad \text{for } \alpha < \frac{1}{3}.$$

But in this case we have

$$\gamma(S, Q) \geq 2 \left[\frac{1-2\alpha^2}{1+\alpha^2} \right], \quad \alpha < \frac{1}{3}$$

(see Fig. 1). In any case a simple calculation shows that

$$(5.20) \quad \begin{aligned} & h^{-2} \sum_{(P, Q, S) \in T} [\gamma(P, Q) V^2(P) - 2V(P) V(S) + \gamma(S, Q) V^2(S)] \\ & \geq \frac{1}{10} h^{-2} \sum_{P \in R_{2h}^*} V^2(P). \end{aligned}$$

Combining (5.16), (5.17), (5.19) and (5.20) we have

$$(5.21) \quad \begin{aligned} h^2 \sum_{S_h} (\Delta_h V)^2 + h^{-2} \sum_{P \in R_h^*} \left[\gamma(P) V(P) - \sum_{S \in J_h(P)} V(S) \right] V(P) \\ \geq \frac{1}{10} h^2 \sum_{S_h} (\Delta_h V)^2. \end{aligned}$$

The case in which the assumption on \dot{R}_h at the beginning of the proof is not satisfied can be dealt with again by examining terms of $h^2 \sum_{S_h} (\Delta_h V)^2$ and (5.21) can be shown to hold more generally.

Thus combining (5.11) and (5.21) we have

$$(5.22) \quad h^2 \sum_{S_h} (\Delta_h V)^2 \leq 10 \left\{ h^2 \sum_{R_h'} V \Delta_h^2 V + h^2 \sum_{R_h^*} V \bar{\Delta}_h^2 V \right\}.$$

In view of uniqueness in the discrete Dirichlet problem (5.22) tells us immediately that the solution of (5.9) is unique. But for linear systems uniqueness implies existence for any given F .

In order to obtain the estimate (5.10) note that the number of points in $N_2(R_h^*)$ is $O(h^{-1})$ and by the definition of the norm $\|V\|_{N_2(R_h^*)}$ ($\|V\|_{N_2(R_h^*)} = \left(h \sum_{P \in N_2(R_h^*)} V^2(P) \right)^{\frac{1}{2}}$) and the fact that $V(P) = 0$, $P \notin R_h$, we have

$$(5.23) \quad h^{-\frac{1}{2}} \|V\|_{R_h^*} \leq C \|\Delta_h V\|$$

where C is a constant which does not depend on h . In addition to this we need the well known inequality

$$(5.24) \quad \|V\| \leq C \|\Delta_h V\|.$$

Now we have for e

$$(5.25) \quad \begin{aligned} |\Delta_h^2 e|_{R_h'} &\leq C h^2 \\ |\bar{\Delta}_h^2 e|_{R_h^*} &\leq C h^{-1} \\ e(P) &= 0, \quad P \notin R_h. \end{aligned}$$

If we set $V = e$ in (5.22)–(5.24) and apply the Schwarz inequality to (5.22) it follows, in view of (5.25), that

$$(5.26) \quad \|e\|_{R_h^*} \leq C h^3.$$

But by Lemma 3.3 the estimate (5.10) follows. This completes the proof of the theorem.

Corollary 1. There exists a constant C independent of h such that

$$(5.27) \quad \|e\|_{(2,p)} \leq C h^2$$

for any integer $p \geq 1$ and h sufficiently small.

Proof. Set $V = e$ in Lemma 3.1 and apply Theorem 5.1.

Corollary 2. There exists a constant C independent of h such that

$$(5.28) \quad |e|_{R_h} \leq C h^2 |\log 1/h|^{\frac{1}{2}}$$

for h sufficiently small.

Proof. Set $V = e$ in Lemma 3.2 and apply Theorem 5.1.

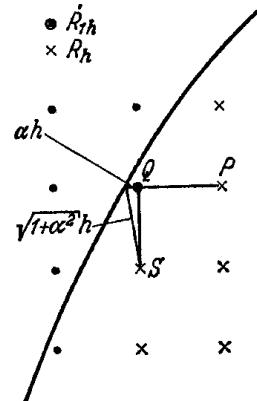


Fig. 1

Appendix

Formulation of Difference Problem for Inhomogeneous Boundary Data

We suppose for simplicity that the boundary is smooth and that u is smooth in \bar{R} .

At an arbitrary point $P \in R_h^*$ we define

$$W_P(Q; u) = h^{-2} \sum_{S \in J_1(Q)} \left\{ \left(\frac{\alpha(S)}{\alpha(P)} \right)^2 u(\bar{P}) - u(\bar{S}) + \alpha(S) h \left[\frac{\alpha(S)}{\alpha(P)} u_n(\bar{P}) - u_n(\bar{S}) \right] \right\} \quad \text{if } Q \in J_0(P) \cap P$$

and

$$W_P(Q; u) = \frac{2}{\alpha^2(P) h^2} \left\{ u(\bar{P}) + \alpha(P) h u_n(\bar{P}) - \frac{\alpha^2(P) h^2}{2} [u_{ss}(\bar{P}) + k(\bar{P}) u_n(\bar{P})] \right\} \quad \text{if } Q \in J_1(P).$$

We have used the notation:

- a) \bar{P} is the point of \dot{R} closest to P (same for S).
- b) $k(\bar{P})$ is the curvature of \dot{R} at P .
- c) $u_n(\bar{P})$ is the outward normal derivative of u at \bar{P} on \dot{R} .
- d) $u_{ss}(\bar{P})$ is the second derivative of u with respect to arc length on \dot{R} at \bar{P} .

If, instead of (1.1), we have u and u_n as given functions of arc length on R then the difference problem (5.9) is replaced by

$$\begin{aligned} \Delta_h^2 U(P) &= F(P), & P \in R'_h \\ \Delta_h^2 U(P) &= F(P) + \Delta_h W_P(P; U), & P \in R_h^* \\ U(P) &= 0, & P \notin R_h. \end{aligned}$$

Now for $e(P)$, as defined in Theorem 1, it then follows that we have

$$\begin{aligned} \Delta_h^2 e(P) &= O(h^2), & P \in R'_h \\ \Delta_h^2 e(P) &= O(h^{-1}), & P \in R_h^* \\ e(P) &= 0, & P \notin R_h. \end{aligned}$$

All the previous results now follow for the inhomogeneous problem.

References

- [1] BRAMBLE, J. H.: Error estimates for difference methods in forced vibration problems. SIAM Series B **3**, No. 1, 1–12 (1966).
- [2] —, and B. E. HUBBARD: On the formulation of finite difference analogues of the Dirichlet problem for Poisson's equation. Num. Math. **4**, 313–327 (1962).
- [3] — — A priori bounds on the discretization error in the numerical solution of the Dirichlet problem. Contributions to Diff. Eq. V **2**, 229–252 (1963).
- [4] — — New monotone type approximations for elliptic problems. Math. Comp. V **18**, 349–367 (1964).
- [5] — — Approximations of derivatives by finite difference methods in elliptic boundary value problems. Contributions to Diff. Eq. V **3**, No. 4, 399–410 (1964).
- [6] — — Approximation of solutions of mixed boundary value problems for Poisson's equation by finite differences. J. ACM, V **12**, No. 1, 114–123 (1965).

- [7] BRAMBLE, J. H., and B. E. HUBBARD: A finite difference analog of the Neumann problem for Poisson's equation. SIAM Series B **2**, No. 1, 1–14 (1965).
- [8] — — Eigenvalue and eigenvector estimates for the fixed membrane problem by difference methods. (To appear.)
- [9] COURANT, R., K. O. FRIEDRICHS u. H. LEWY: Über die partiellen Differenzen-gleichungen der mathematischen Physik. Math. Ann. **100**, 32–74 (1928).
- [10] FICHERA, G.: On a unified theory of boundary value problems for elliptic-parabolic equations of second order. Boundary problems in differential equations, pp. 97–120. U. Wis. Press 1960.
- [11] MIRANDA, C.: Formule di maggiorazione e teorema di esistenza per le funzioni biarmoniche di due variabili. Giornale de Mat. di Battaglini **78**, 1–22 (1948/49).
- [12] THOMÉE, V.: Elliptic difference operators and Dirichlet's problem. Contributions to Diff. Eq. V **3**, 301–324 (1964).
- [13] ZLÁMAL, M.: Asymptotic error estimates in solving elliptic equations of the fourth order by the method of finite differences. SIAM Series B **2**, 337–344 (1965).

University of Maryland
College Park, Maryland, USA

*1.7. ERROR ESTIMATES FOR DIFFERENCE METHODS IN FORCED
VIBRATION PROBLEMS*

1.7 Error estimates for difference methods in forced vibration problems

Error estimates for difference methods in forced vibration problems [32]

J. SIAM NUMER. ANAL.
Vol. 3, No. 1, 1966
Printed in U.S.A.

ERROR ESTIMATES FOR DIFFERENCE METHODS IN FORCED VIBRATION PROBLEMS*

J. H. BRAMBLE†

1. Introduction. In some recent papers [1], [2], [3], [4], [5], Bramble and Hubbard have studied the error in finite difference approximations to solutions of various boundary value problems for second order elliptic partial differential equations. In addition to examining various standard difference operators they analyzed some operators which failed to possess standard properties such as diagonal dominance (cf. [7]).

This paper is concerned with obtaining error estimates for some difference methods for another type of problem which leads, in general, to difference approximations that are not diagonally dominant. The class of problems considered is sometimes referred to as "forced vibration problems."

Specifically we shall consider the problem

$$(1.1) \quad \begin{aligned} \Delta u + ku &= F \quad \text{in } R, \\ u &= f \quad \text{on } C, \end{aligned}$$

where R is an N -dimensional bounded region with sufficiently smooth boundary C , and Δ denotes the Laplace operator. The function k is only required to be bounded and such that the solution u is unique. If, in addition, the function k were restricted to be nonpositive, then quite standard techniques apply in the study of errors in related difference problems. Our concern will be with the more general situation and hence some techniques, other than those most often used for such problems, will be employed. This method was suggested by the work of Payne and the author [6] in which a priori inequalities relating to forced vibration problems were derived.

For the sake of clarity we shall confine our attention to two dimensions. The N -dimensional case is discussed however in §7 as is the extension to more general operators.

2. Notation. As mentioned we shall discuss in detail the case $N = 2$. Thus C is a smooth, simple closed curve in the (x, y) -plane bounding the region R . The plane will be covered with a square grid with mesh width h . The mesh points are the intersection of the grid lines. The set R_h is defined to consist of those mesh points in R whose four nearest neighbors also

* Received by the editors March 4, 1965.

† Institute for Fluid Dynamics and Applied Mathematics, University of Maryland, College Park, Maryland. This research was supported in part by the National Science Foundation under Grant NSF-GP-2284.

belong to R . Those mesh points in R which do not belong to R_h will make up the set C_h^* . The points of intersection of the grid with C form the set C_h . Finally $\bar{R}_h = R_h \cup C_h^* \cup C_h$.

The following difference operators will be needed. Let $V(x, y)$ be any function defined on \bar{R}_h . Then

$$(2.1) \quad \begin{aligned} \Delta_h V(x, y) &= h^{-2} \{ V(x + h, y) + V(x - h, y) \\ &\quad + V(x, y + h) + V(x, y - h) - 4V(x, y) \}, \quad (x, y) \in R_h. \end{aligned}$$

This is the usual five point approximation to Δ . If $v \in C^4$ in \bar{R} then

$$(2.2) \quad |\Delta v(x, y) - \Delta_h V(x, y)| \leq \frac{h^2}{6} M_4, \quad (x, y) \in R_h,$$

where we have used the notation

$$(2.3) \quad M_j = \sup_{p \in R} \left\{ \left| \frac{\partial^j v(p)}{\partial x^i \partial y^{j-i}} \right|, i = 0, \dots, j \right\}.$$

At points of C_h^* , Δ_h is defined to be the five point divided difference approximation to Δ . For example, if $(x, y) \in C_h^*$ and if $(x - \alpha h, y)$ and $(x, y - \beta h) \in C_h$ while $(x + h, y)$ and $(x, y + h) \in R_h$, then Δ_h would take the form

$$(2.4) \quad \begin{aligned} \Delta_h^{(1)} V(x, y) &= 2h^{-2} \left\{ \left(\frac{1}{\alpha + 1} \right) V(x + h, y) + \frac{1}{\alpha(\alpha + 1)} V(x - \alpha h, y) \right. \\ &\quad + \left(\frac{1}{\beta + 1} \right) V(x, y + h) + \frac{1}{\beta(\beta + 1)} V(x, y - \beta h) \\ &\quad \left. - \left(\frac{1}{\alpha} + \frac{1}{\beta} \right) V(x, y) \right\}. \end{aligned}$$

Because of the definition of C_h^* , $0 < \alpha \leq 1$ and $0 < \beta \leq 1$. At each point of C_h^* the appropriate analog of (2.4) is assumed. In this case

$$(2.5) \quad |\Delta v(x, y) - \Delta_h^{(1)} v(x, y)| \leq \frac{2h}{3} M_3, \quad (x, y) \in C_h^*,$$

for $v \in C^3$.

We shall consider also a slightly different operator at points of C_h^* . Let

$$(2.6) \quad \begin{aligned} \Delta_h^{(2)} V(x, y) &= h^{-2} \left\{ V(x + h, y) + \frac{1}{\alpha} V(x - \alpha h, y) \right. \\ &\quad + V(x, y + h) + \frac{1}{\beta} V(x, y - \beta h) - \left(\frac{\alpha + 1}{\alpha} + \frac{\beta + 1}{\beta} \right) V(x, y) \left. \right\}. \end{aligned}$$

It is easy to see that

$$|\Delta_h^{(2)}v(x, y)| \leq 2M_2, \quad (x, y) \in C_h^*,$$

if $v \in C^2$. As is easily seen, the operators (2.1) and (2.6) give rise to a difference analog of (1.1) whose associated matrix is symmetric. For some purposes this seems to be an advantage.

3. Lemmas and preliminaries. In this section we will present some lemmas and results which will be needed in establishing the results in the subsequent sections. The difference analogs of (1.1) which we shall study are

$$(3.1) \quad \begin{aligned} \Delta_h U(P) + kU(P) &= F(P), & P \in R_h \cup C_h^*, \\ U(P) &= f(P), & P \in C_h. \end{aligned}$$

In (3.1) the operator Δ_h on C_h^* is taken to be *either* everywhere of the type $\Delta_h^{(1)}$ or everywhere of the type $\Delta_h^{(2)}$.

The first lemma is the well-known maximum principle.

LEMMA 1. *Let $V(P)$ and $q(P)$ be defined on \bar{R}_h and satisfy*

$$(3.2) \quad \Delta_h V(P) - q(P)V(P) \geq 0, \quad P \in R_h \cup C_h^*,$$

with $q \geq 0$. Then either $V(P) \leq 0$, $P \in \bar{R}_h$, or

$$(3.3) \quad \max_{P \in \bar{R}_h} V(P) \leq \max_{P \in C_h} V(P).$$

For a detailed proof, cf. [3].

LEMMA 2. *The solution of (3.1), with $k \equiv -q \leq 0$, exists and is unique for any given F and f .*

Proof. Uniqueness follows immediately from the maximum principle. But for linear systems uniqueness implies existence.

Thus, with Lemma 2, ($q \equiv 0$) we can introduce the discrete analog of the Green's function. Let $G(P, Q)$ satisfy

$$(3.4) \quad \begin{aligned} \Delta_{h,P} G(P, Q) &= -h^{-2}\delta(P, Q), & P \in R_h \cup C_h^*, \\ G(P, Q) &= \delta(P, Q), & P \in C_h, \end{aligned}$$

for $Q \in \bar{R}_h$. The symbol $\Delta_{h,P}$ means that the operator Δ_h operates on variable P and $\delta(P, Q)$ is the Kronecker delta

$$\delta(P, Q) = \begin{cases} 1 & \text{if } P = Q, \\ 0 & \text{if } P \neq Q. \end{cases}$$

By applying Lemma 1 to $-G(P, Q)$ for fixed but arbitrary Q we obtain

$$\max_{P \in \bar{R}_h} -G(P, Q) \leq \max_{P \in C_h} -G(P, Q) = 0.$$

This yields immediately the following.

LEMMA 3.

$$(3.5) \quad G(P, Q) \geq 0, \quad P, Q \in \bar{R}_h.$$

The next lemma gives the useful representation formula for an arbitrary mesh function V in terms of G .

LEMMA 4. *For any $V(P)$ defined on \bar{R}_h ,*

$$(3.6) \quad V(P) = h^2 \sum_{Q \in R_h \cup C_h^*} G(P, Q) [-\Delta_h V(Q)] + \sum_{Q \in C_h} G(P, Q) V(Q).$$

Proof. Let the right-hand side of (3.6) be denoted by $W(P)$. Using (3.4) we see that

$$\begin{aligned} \Delta_h W(P) &= \Delta_h V(P), & P \in R_h \cup C_h^*, \\ W(P) &= V(P), & P \in C_h. \end{aligned}$$

It follows from Lemma 2 that $V(P) \equiv W(P)$, $P \in \bar{R}_h$.

If we take $V(P) \equiv 1$, $P \in \bar{R}_h$, we obtain:

LEMMA 5.

$$(3.7) \quad \sum_{Q \in C_h} G(P, Q) = 1, \quad P \in \bar{R}_h.$$

The next lemma was given by Bramble and Hubbard in [1]. It is this result which is crucial in showing that less accurate approximations to Δ may be made on C_h^* than on R_h without affecting the overall accuracy.

LEMMA 6.

$$(3.8) \quad \sum_{Q \in C_h^*} G(P, Q) \leq 1, \quad P \in R_h.$$

Proof. In (3.6) take $V(P) = 1$ for $P \in R_h \cup C_h^*$, and $V(P) = 0$ for $P \in C_h$. Examination of $\Delta_h^{(1)}$ and $\Delta_h^{(2)}$ shows that

$$(3.9) \quad -\Delta_h V(P) \geq h^{-2}, \quad P \in C_h^*.$$

Clearly $\Delta_h V(P) = 0$, $P \in R_h$. Using (3.5) we obtain (3.8).

In obtaining results, even in the case $k \equiv 0$, an additional inequality is required.

LEMMA 7.

$$(3.10) \quad h^2 \sum_{Q \in R_h \cup C_h^*} G(P, Q) \leq C, \quad P \in \bar{R}_h,$$

where C is a constant independent of h .

Proof. Let $r(P)$ be the distance from an arbitrary but fixed point. Then $\Delta_h r^2 = \Delta r^2 = 4$. Let $V(P) = -r(P)^2$ in (3.6). The result then follows from Lemmas 3 and 5 and the fact that R is bounded.

We shall use the notation C for a generic constant, independent of h

(and the particular difference problem), but not necessarily always the same.

A simple consequence of Lemmas 4, 5, 6 and 7 is that

$$(3.11) \quad |V|_{\bar{R}_h} \leq C |\Delta_h V|_{R_h} + h^2 |\Delta_h V|_{C_h^*} + |V|_{C_h}$$

for any V , where we have used the notation

$$(3.12) \quad |W|_S = \sup_{P \in S} |W(P)|$$

for a function W defined on a set S . This is the result given in [1].

In order to treat the case with general k we shall need two additional lemmas giving information about G .

LEMMA 8.

$$(3.13) \quad h^2 \sum_{Q \in R_h \cup C_h^*} G^2(P, Q) \leq C, \quad P \in \bar{R}_h.$$

This lemma is proved in [2] for the operator $\Delta_h^{(1)}$ on C_h^* . The argument for $\Delta_h^{(2)}$ on C_h^* follows in precisely the same manner.

Finally we need a bit more information “near the boundary.”

LEMMA 9.

$$(3.14) \quad h^2 \sum_{Q \in R_h \cup C_h^*} G(P, Q) \leq Ch, \quad P \in C_h^*.$$

Proof. We have assumed that the boundary of R is sufficiently smooth. In this instance we need the function ϕ defined by

$$(3.15) \quad \begin{aligned} \Delta\phi &= -1 && \text{in } R, \\ \phi &= 0 && \text{on } C, \end{aligned}$$

to exist and possess uniformly bounded third derivatives. Now the mesh function

$$(3.16) \quad \phi_h(P) = h^2 \sum_{Q \in R_h \cup C_h^*} G(P, Q)$$

satisfies

$$(3.17) \quad \begin{aligned} \Delta_h \phi(P) &= -1, & P \in R_h \cup C_h^*, \\ \phi(P) &= 0, & P \in C_h. \end{aligned}$$

Taking $V(P) = \phi(P) - \phi_h(P)$ in (3.11), we find that

$$(3.18) \quad |\phi(P) - \phi_h(P)| \leq Ch, \quad P \in \bar{R}_h.$$

Hence

$$(3.19) \quad |\phi_h(P)| \leq |\phi(P)| + Ch.$$

But since $\phi = 0$ on the boundary it follows that

$$(3.20) \quad |\phi_h(P)| \leq Ch, \quad P \in C_h^*,$$

which is the same as (3.14).

4. Associated estimates for the discrete L_2 norm. Let V be any mesh function defined in R_h . We define

$$(4.1) \quad \|V\|^2 = h^2 \sum_{Q \in R_h} V(Q)^2.$$

In this section we shall prove the inequality

$$(4.2) \quad \|V\| \leq C(\|\Delta_h V + kV\| + |V|_{C_h^*}),$$

where k is a bounded function,

$$(4.3) \quad |k|_R \leq \bar{k},$$

with bound \bar{k} .

The uniqueness condition for (1.1) we take in the following form. Consider the eigenvalue problem

$$(4.4) \quad \begin{aligned} \Delta W + (k - \bar{k})W + \lambda W &= 0 \quad \text{in } R, \\ W &= 0 \quad \text{on } C. \end{aligned}$$

Thus we assume that $\lambda = \bar{k}$ is not an eigenvalue of (4.4). It is well known that the eigenvalues form an increasing sequence $\lambda_1, \lambda_2, \dots$ of positive real numbers. Analogously we have the discrete problem

$$(4.5) \quad \begin{aligned} \Delta_h U + (k - \bar{k})U + \mu U &= 0 \quad \text{in } R_h, \\ U &= 0 \quad \text{on } C_h^*. \end{aligned}$$

This is just a symmetric positive definite matrix eigenvalue problem and thus the real positive eigenvalues μ_i with eigenvectors U_i satisfy

$$(4.6) \quad \begin{aligned} \Delta_h U_i + (k - \bar{k})U_i + \mu_i U_i &= 0 \quad \text{in } R_h, \\ U_i &= 0 \quad \text{on } C_h^*, \end{aligned}$$

and the eigenvectors span the m -dimensional space of vectors indexed by the m points of R_h . Furthermore it is also known (cf. [10]) that if $i = \alpha$, where α is fixed, then

$$(4.7) \quad \mu_\alpha \rightarrow \lambda_\alpha \quad \text{as } h \rightarrow 0.$$

Clearly we have, then, for any V ,

$$(4.8) \quad \|V\|^2 = \sum_{i=1}^m (h^2 \sum_{P \in R_h} V(P) U_i(P))^2,$$

provided the U_i have been chosen so that

$$(4.9) \quad h^2 \sum_{P \in R_h} U_i(P) U_j(P) = \delta_{ij},$$

where δ_{ij} is the Kronecker delta.

Now consider for any fixed i ,

$$(4.10) \quad \mu_i h^2 \sum_{P \in R_h} V U_i = -h^2 \sum_{P \in R_h} V [\Delta_h U_i + (k - \bar{k}) U_i],$$

where the argument P in the mesh functions will be omitted for simplicity. By virtue of Lemma 2 we may introduce the mesh function H as the solution of

$$(4.11) \quad \begin{aligned} \Delta_h H + (k - \bar{k}) H &= 0 \text{ in } R_h, \\ H &= V \text{ on } C_h^*. \end{aligned}$$

Applying (4.10) to both V and H and combining we have

$$(4.12) \quad \begin{aligned} \mu_i h^2 \sum_{P \in R_h} V U_i \\ = -h^2 \sum_{P \in R_h} (V - H) [\Delta_h U_i + (k - \bar{k}) U_i] + \mu_i h^2 \sum_{P \in R_h} H U_i. \end{aligned}$$

But since $V - H$ and U_i both vanish on C_h^* we have

$$(4.13) \quad \begin{aligned} h^2 \sum_{P \in R_h} (H - V) [\Delta_h U_i + (k - \bar{k}) U_i] \\ = -h^2 \sum_{P \in R_h} U_i [\Delta_h V + (k - \bar{k}) V]. \end{aligned}$$

Combining (4.12) and (4.13) yields

$$(4.14) \quad (\mu_i - \bar{k}) h^2 \sum_{P \in R_h} V U_i = -h^2 \sum_{P \in R_h} [\Delta_h V + kV] U_i + \mu_i h^2 \sum_{P \in R_h} H U_i.$$

Now since $\lambda \neq \bar{k}$ and $\mu_i \rightarrow \lambda_i$ as $h \rightarrow 0$, we may choose h_0 sufficiently small that $\mu_i - \bar{k} \neq 0$ for $h < h_0$. Hence

$$(4.15) \quad \begin{aligned} \sum_{i=1}^m (h^2 \sum_{P \in R_h} V U_i)^2 &= \sum_{i=1}^m \left[\left(\frac{-1}{\mu_i - \bar{k}} \right) h^2 \sum_{P \in R_h} [\Delta_h V + kV] U_i \right. \\ &\quad \left. + \left(\frac{\mu_i}{\mu_i - \bar{k}} \right) h^2 \sum_{P \in R_h} H U_i \right]^2 \\ &\leq 2 \sum_{i=1}^m \left(\frac{1}{\bar{k} - \mu_i} \right)^2 (h^2 \sum_{P \in R_h} [\Delta_h V + kV] U_i)^2 \\ &\quad + 2 \sum_{i=1}^m \left(\frac{\mu_i}{\bar{k} - \mu_i} \right)^2 (h^2 \sum_{P \in R_h} H U_i)^2 \\ &\leq 2C^2 (\| \Delta_h V + kV \|_h^2 + \| H \|_h^2), \end{aligned}$$

where $C = \max(1/\rho, 1 + \bar{k}/\rho)$, $\rho = \min_i |\mu_i - \bar{k}|$. But by Lemma 1 applied to R_h ,

$$(4.16) \quad |H|_{R_h} \leq |H|_{c_h^*} = |V|_{c_h^*}.$$

Combining (4.8), (4.15), and (4.16) there is another constant C such that

$$(4.17) \quad \|V\| \leq C(\|\Delta_h V + kV\| + |V|_{c_h^*}).$$

5. Estimates for the maximum norm. We start with the representation (3.6) of Lemma 4, adding and subtracting the appropriate terms

$$(5.1) \quad \begin{aligned} V(P) &= -h^2 \sum_{Q \in R_h \cup C_h^*} G(P, Q) [\Delta_h V(Q) + kV(Q)] \\ &\quad + h^2 \sum_{Q \in R_h \cup C_h^*} G(P, Q) kV(Q) + \sum_{Q \in C_h} G(P, Q) V(Q). \end{aligned}$$

Breaking up the sums, using Schwarz's inequality, and the bound for k we have

$$(5.2) \quad \begin{aligned} |V(P)| &\leq (h^2 \sum_{Q \in R_h} G^2(P, Q))^{1/2} \{ \|\Delta_h V + kV\| + \bar{k} \|V\| \} \\ &\quad + (h^2 \sum_{Q \in C_h^*} G(P, Q)) \{ |\Delta_h V + kV|_{c_h^*} + \bar{k} |V|_{c_h^*} \} \\ &\quad + (\sum_{Q \in C_h} G(P, Q)) |V|_{c_h}, \end{aligned}$$

where we have used the fact that $G(P, Q) \geq 0$ (Lemma 3). By virtue of Lemmas 6, 7, and 8 we obtain

$$(5.3) \quad \begin{aligned} |V|_{R_h} &\leq C \{ |\Delta_h V + kV|_{R_h} + \|V\| \\ &\quad + h^2 |\Delta_h V + kV|_{c_h^*} + h^2 |V|_{c_h^*} + |V|_{c_h} \}. \end{aligned}$$

Now (5.3) together with (4.17) yields

$$(5.4) \quad |V|_{R_h} \leq \bar{C} \{ |\Delta_h V + kV|_{R_h} + h^2 |\Delta_h V + kV|_{c_h^*} + |V|_{c_h} + |V|_{c_h^*} \},$$

for some \bar{C} . To obtain a bound in terms of only "data" we must still bound $|V|_{c_h^*}$. This may be accomplished by considering again the representation (5.1) with P restricted to lie on C_h^* . Then Lemma 9 yields

$$(5.5) \quad \begin{aligned} |V|_{c_h^*} &\leq \bar{\bar{C}} \{ h |\Delta_h V + kV|_{R_h} + h^2 |\Delta_h V + kV|_{c_h^*} + |V|_{c_h} + h |V|_{R_h} \}, \end{aligned}$$

with $\bar{\bar{C}}$ an appropriate constant. Combining (5.4) and (5.5) we obtain

$$(5.6) \quad |V|_{R_h} \leq C \{ |\Delta_h V + kV|_{R_h} + h^2 |\Delta_h V + kV|_{c_h^*} + |V|_{c_h} \},$$

provided h is taken strictly less than $\min(h_0, h_1)$, where $h_1 \leq 1/\bar{C}\bar{\bar{C}}$.

6. Error estimates. Let $|k| \leq \bar{k}$ and suppose \bar{k} is not an eigenvalue of (4.4). Denote by u the unique solution of (1.1) and suppose that u has continuous fourth partial derivatives in the closure of R . Further define U to be the solution of (3.1) for small h . If $e(P) = u(P) - U(P)$, $P \in \bar{R}_h$, then there is an \bar{h} such that if $h \leq \bar{h}$,

$$(6.1) \quad |e|_{\bar{R}_h} \leq Ch^2.$$

This is easily seen by taking $V = e$ in (5.6) and \bar{h} such that (5.6) holds for $h \leq \bar{h}$. The bound on the first term on the right-hand side is obtained from (2.2):

$$(6.2) \quad \begin{aligned} |\Delta_h e + ke|_{R_h} &= |(\Delta_h u + ku) - (\Delta_h U + kU)|_{R_h} \\ &= |\Delta_h u + ku - F|_{R_h} = |\Delta_h u - \Delta u|_{R_h} \leq \frac{h^2}{6} M_4. \end{aligned}$$

A similar consideration for the second term yields

$$(6.3) \quad |\Delta_h e + ke|_{C_h^*} \leq \frac{2}{3} h M_3$$

in case we are using $\Delta_h^{(1)}$, and

$$(6.4) \quad |\Delta e + ke|_{C_h^*} \leq 4M_2$$

if we use $\Delta_h^{(1)}$. In either case the contribution of this term to the error is no worse than $O(h^2)$. The last term on the right vanishes.

Bounds on difference quotients of e can be obtained by exactly those considerations of [5], making use of (5.6). We only state the result here. If P and P^1 are neighboring points of \bar{R}_h let $\delta e(P) = [e(P) - e(P^1)]/h$. Then if $\Delta_h = \Delta_h^{(1)}$ on C_h^* ,

$$(6.5) \quad |\delta e|_{\bar{R}_h} \leq Ch^2.$$

Thus the difference quotients are uniformly $O(h^2)$ if the “better” approximation $\Delta_h^{(1)}$ is used on C_h^* .

7. Extension of the preceding results to higher dimensions, more general second order elliptic operators, and higher order elliptic equations.

(a) N -dimensional case.

If $N \leq 3$ then all of the preceding considerations are valid, defining, of course, the mesh and operators analogously. For more than three dimensions the only lemma of §3 which is not a straightforward generalization is Lemma 8. However, it is tedious but not difficult to show that

$$(7.1) \quad \begin{aligned} h^N \sum_{Q \in R_h \cup C_h^*} r_{PQ}^\alpha G^2(P, Q) \\ + h^{N-2} G(P, P) \leq C, \quad P \in \bar{R}_h, \quad N - 2 < \alpha. \end{aligned}$$

The failure of Lemma 8 to hold corresponds to the fact that the Green's function is not square integrable for $N > 3$.

Now in §4 all steps are valid for any N . Since we may no longer use (5.2) to derive (5.4) we must modify the proof. We look now at the analog of (5.4):

$$(7.2) \quad V(P) = h^N \sum_{Q \in R_h} G(P, Q) kV(Q) + T(P),$$

where $T(P)$ stands for terms which can be dealt with as before. Now

$$(7.3) \quad \begin{aligned} |V(P)| &\leq \bar{k}h^N \sum_{\substack{Q \in R_h \\ Q \neq P}} G(P, Q) |V(Q)| \\ &\quad + \bar{k}h^N G(P, P) |V(P)| + |T(P)|. \end{aligned}$$

Actually we can obtain

$$(7.4) \quad \begin{aligned} |T|_{R_h} &\leq C \{ |\Delta_h V + kV|_{R_h} + h^2 |\Delta_h V + kV|_{C_h^*} + |V|_{C_h^*} + |V|_{C_h} \} \\ &\equiv T. \end{aligned}$$

Thus by virtue of (7.1), with h small enough that $\bar{k}h^N G(P, P) \leq \bar{k}Ch^2 < 1$, we obtain

$$(7.5) \quad |V(P)| \leq C \left\{ h^N \sum_{\substack{Q \in R_h \\ Q \neq P}} G(P, Q) |V(Q)| + T \right\}.$$

Again using (7.1) we obtain

$$(7.6) \quad V(P)^2 \leq C \left\{ h^N \sum_{\substack{Q \in R_h \\ Q \neq P}} \frac{V(Q)^2}{r_{PQ}^\alpha} + T^2 \right\}$$

for $N - 2 < \alpha < N$.

We now extend the definition of $V(P)$ to be piecewise constant on hypercubes of side h centered at P , and zero for $P \notin R_h$.

Now since $r_{PQ}^{-\alpha}$ is a subharmonic function of Q for $P \neq Q$ we have that

$$(7.7) \quad r_{PQ}^{-\alpha} \leq \frac{1}{\omega_N h^N} \int_{S_h(Q)} r_{PR}^{-\alpha} dv_R,$$

where $S_h(Q)$ is the sphere of radius $h/2$ centered at Q and ω_N is the surface area of the N -dimensional unit sphere. Thus since $V(Q)$ is constant on $S_h(Q)$,

$$(7.8) \quad \begin{aligned} V^2(P) &\leq C \left\{ h^N \sum_{\substack{Q \in R_h \\ Q \neq P}} \int_{S_h(Q)} \frac{V^2(R)}{r_{PR}^\alpha} dv_R + T^2 \right\} \\ &\leq C \left\{ \int_{E_N} \frac{V^2(R)}{r_{PR}^\alpha} dv_R + T^2 \right\}, \end{aligned}$$

for $P \in R_h$, where E_N is N -dimensional euclidean space. It is easy to see that this inequality can be extended to hold for all points P of E_N . Now it is well known (see [8]) that for $\alpha < N$ such an inequality implies

$$(7.9) \quad V^2(P) \leq C \left\{ \int_{E_N} V^2(Q) dv_Q + T^2 \right\}.$$

(This can be seen by iterating inequality (7.8) once and using the identity of Riesz and Hölder's inequality.) But since V is piecewise constant on cubes and vanishes outside R , (7.9) may be written as

$$(7.10) \quad |V(P)| \leq C\{\|V\| + T\}, \quad P \in R_h,$$

or

$$(7.11) \quad |V|_{R_h} \leq C\{\|V\| + |\Delta_h V + kV|_{R_h} + h^2 |\Delta_h V + kV|_{C_h^*} + |V|_{C_h}\}.$$

This inequality combined with the other appropriate inequalities yields (5.6) for any N .

(b) The previous results carry over to second order self-adjoint operators with variable coefficients provided the difference analog chosen is symmetric in R_h and has a positive discrete Green's function and in addition Lemma 8 is obtainable. For example, for $N = 2, 3$, if no mixed partial derivatives are present a symmetric difference analog with positive Green's function is easily derivable. The question of obtaining Lemma 8 for such difference operators should be studied.

(c) In his paper [9] Thomée studies a general class of difference methods for the Dirichlet problem for elliptic equations with constant coefficients of order $2m$ which contain only the highest derivatives. If L is one of the operators studied by Thomée then the methods used in §4 can easily be used to obtain convergence estimates for problems corresponding to $L + k$, provided k is such that a condition analogous to (3.1) is satisfied and estimates for the eigenvalues in the discrete eigenvalue problem are known. In fact if $|k|$ is less than the first eigenvalue in the continuous problem then the result is complete since Thomée has shown convergence for the first discrete eigenvalue.

REFERENCES

- [1] J. H. BRAMBLE AND B. E. HUBBARD, *On the formulation of finite difference analogues of the Dirichlet problem for Poisson's equation*, Numer. Math., 4 (1962), pp. 313–327.
- [2] ———, *A priori bounds on the discretization error in the numerical solution of the Dirichlet problem*, Contributions to Differential Equations, 2 (1963), pp. 229–252.
- [3] ———, *On a finite difference analogue of an elliptic boundary problem which is*

- neither diagonally dominant nor of non-negative type*, J. Math. and Phys., 43 (1964), pp. 117–132.
- [4] ———, *New monotone type approximations for elliptic problems*, Math. Comp., 18 (1964), pp. 349–367.
- [5] ———, *Approximations of derivatives by finite difference methods in elliptic boundary value problems*, to appear.
- [6] J. H. BRAMBLE AND L. E. PAYNE, *Upper and lower bounds in equations of forced vibration type*, Arch. Rational Mech. Anal., 14 (1963), pp. 153–170.
- [7] G. FORSYTHE AND W. WASOW, *Finite Difference Methods for Partial Differential Equations*, John Wiley, New York, 1960.
- [8] M. RIESZ, *L'intégrale de Riemann-Liouville et le problème de Cauchy*, Acta Math., 81 (1949), pp. 1–223.
- [9] V. THOMÉE, *Elliptic difference operators and Dirichlet's problème*, Contributions to Differential Equations, 3 (1964), pp. 301–324.
- [10] H. F. WEINBERGER, *Lower bounds for higher eigenvalues by finite difference methods*, Pacific J. Math., 8 (1958), pp. 339–368.

1.8 On the convergence of difference approximations to weak solutions of Dirichlet's problem

On the convergence of difference approximations to weak solutions of Dirichlet's problem [33]

On the Convergence of Difference Approximations to Weak Solutions of Dirichlet's Problem

J. H. BRAMBLE*

Received June 20, 1968

I. Introduction

This paper is concerned with the Dirichlet problem for Poisson's equation, generalized in a certain way, and the convergence properties of related finite difference approximations. The problem considered may be written formally as

$$(1.1) \quad \begin{aligned} \Delta u &= F \quad \text{in } R \\ u &= 0 \quad \text{on } \partial R, \end{aligned}$$

where R is a bounded open subset of the N -dimensional euclidian space E_N , ∂R is its boundary, Δ is the Laplace operator and F is an element of $(L_\infty)'$. ($(L_\infty)'$ is the dual space of L_∞ , the space of Lebegue measurable, essentially bounded functions on R .) The precise meaning of (1.1) will be given in the next section.

A convergence theorem for difference approximations for the classical problem (1.1) (i.e. F and ∂R sufficiently regular) was given by COURANT, FRIEDRICH and LEWY [6] and estimates for rates of convergence by GERSHGORIN [7] and others. Recently some rather refined estimates have been given by BRAMBLE, HUBBARD and THOMÉE [2]. CÉA [4] has considered general second order self adjoint elliptic operators, more general boundary conditions but with $F \in L_2$.

The third section contains some definitions and discrete a priori inequalities and in the following section is given an existence and uniqueness theorem for our problem which itself seems to be new. The proof of existence is done by means of a finite difference method since, almost as a biproduct, we can make assertions as to convergence of related sequences of difference approximations. Several convergence theorems are given.

The fifth section contains a study of the generalized Green's function and from the convergence properties of the sequence of "discrete Green's functions" is proved pointwise convergence when $F \in L_q$, $q > N/2$, and uniform convergence if, in addition, the boundary is of class C^2 .

II. Formulation of the Problem

In order to formulate precisely our problem we need the following definitions. Let $C_0^\infty(R)$ be the class of real functions, each of which is infinitely differentiable and has its support contained in R . The Hilbert space \mathring{H}_1 is defined as the

* National Science Foundation Senior Postdoctoral Fellow for the academic year 1966–67 at the University of Rome, Rome, Italy. Also supported in part by the National Science Foundation under NSF-GP-6631.

completion of $C_0^\infty(R)$ with respect to the norm

$$(2.1) \quad \|v\|_{H_1}^2 = \int_R v^2 dx + \sum_{i=1}^N \int_R \left(\frac{\partial v}{\partial x_i} \right)^2 dx.$$

The class V is defined as follows:

$$(2.2) \quad V = \{ \phi | \phi \in \dot{H}_1; \Delta \phi \in C_0^\infty(R) \}.$$

If B is a linear topological space and B' is its dual then if $v \in B$ and $f \in B'$ the notation $\langle v, f \rangle$ means the value of f at v .

We now may state our problem.

Problem (D). Given $F \in (L_\infty)'$, find $u \in L_p$, $1 \leq p < \frac{N}{N-2}$ such that

$$(2.3) \quad \int_R u \Delta \phi dx = \langle \phi, F \rangle$$

for all $\phi \in V$.

We remark that it will be shown that $V \subset L_\infty$ so that (2.3) is well defined. Also note that when F is a sufficiently smooth function and ∂R is sufficiently regular, problem (D) is just the classical Dirichlet problem.

III. Definitions, Notations and Lemmas

Let E_{Nh} be those points of E_N of the form $(i_1 h, \dots, i_N h)$ where $h > 0$ and i_1, \dots, i_N are integers, i.e. the "mesh" points. Let $\bar{R}_h = R \cap E_{Nh}$ and $S_\varrho(x)$ be the sphere with center $x = (x_1, \dots, x_N)$ and radius ϱ ; i.e. $S_\varrho(x) = \{y | |x - y| \leq \varrho\}$.

Let

$$R_h = \bar{R}_h \cap \{x | S_{\sqrt{Nh}}(x) \subset R\}.$$

For any function v defined on E_{Nh} the $2N+1$ point discrete Laplace operator is defined as

$$(3.1) \quad \Delta_h v(x) = h^{-2} \sum_{i=1}^N [v(x + h e_i) + v(x - h e_i) - 2v(x)],$$

where e_i is the vector with 1 in the i -th position and 0 in the others. We will need two extensions of v to all of E_N . First we define \tilde{v} to be the following:

$$(3.2) \quad \tilde{v}(y) = v(x) \quad \text{for } y \in C_h(x)$$

where $C_h(x) = \{y | x_i - h/2 \leq y_i < x_i + h/2, i = 1, \dots, N\}$. Let ∇_i be the forward difference operator defined by $\nabla_i v(x) = \frac{v(x + e_i h) - v(x)}{h}$. Then for each point y of the cube $C'_h(x) = \{y | x_i \leq y_i < x_i + h, i = 1, \dots, N\}$, we define $v'(y)$, $x \in E_{Nh}$ as follows:

$$(3.3) \quad v'(y) = \left[\prod_{i=1}^N (1 + (y_i - x_i) \nabla_i) \right] v(x).$$

This function is continuous in E_N , linear in each of its variables in each cube $C'_h(x)$ and $v'(x) = v(x)$ for $x \in E_{Nh}$ (c.f. COMINCIOLI [5]). It is also obvious that there is a constant C independent of h and v such that

$$(3.4) \quad \|v'\|_{H_1}^2 \leq C \left\{ \left(h^N \sum_{E_{Nh}} v^2 \right) + \sum_{i=1}^N \left(h^N \sum_{E_{Nh}} (\nabla_i v)^2 \right) \right\}.$$

Thus if v has bounded support then $v' \in H_1$. (H_1 is the space obtained by completing the space of infinitely differentiable functions with respect to the norm (2.4).)

Let us now consider the discrete Green's function defined by

$$\begin{aligned} A_{h,x} G_h(x, y) &= -h^{-N} \delta(x, y), & x \in R_h, \\ G_h(x, y) &= 0, & x \in E_{Nh} - R_h \end{aligned}$$

for $y \in E_{Nh}$. It is well known that G_h exists and is unique. It was shown in [3] that if we define $w(x)$ as

$$(3.5) \quad w(x) = \begin{cases} 1/\gamma_2 \ln \left[\frac{d_0^2 + \alpha h^2}{|x|^2 + \alpha h^2} \right], & N = 2 \\ \frac{1}{(N-2)\gamma_N} [|x|^2 + \alpha h^2]^{\frac{2-N}{2}}, & N \geq 3 \end{cases}$$

then

$$(3.6) \quad 0 \leq G_h(x, y) \leq w(x - y),$$

for a certain choice of γ_N , α and d_0 which is independent of h . From this it is easily seen that the following lemma is true.

Lemma 1. For $1 \leq p < \frac{N}{N-2}$ there is a constant C which depends only on ϕ and the diameter of R such that

$$(3.7) \quad \left(h^N \sum_{y \in R_h} |G_h(x, y)|^p \right)^{1/p} \leq C$$

for all x .

This lemma follows immediately by applying (3.6) and estimating the result by making comparisons with the analogous integrals.

Now let us consider the function $\phi_h(x)$ defined on E_{Nh} as the solution of

$$(3.8) \quad \begin{aligned} A_h \phi_h(x) &= (\Delta \phi)_h(x) = \frac{1}{h^N} \int_{C_h(x)} \Delta \phi(y) dy, & x \in R_h \\ \phi_h(x) &= 0, & x \in E_{Nh} - R_h, \end{aligned}$$

where ϕ is an arbitrary function in the class V . It is well known that

$$(3.9) \quad \phi_h(x) = -h^N \sum_{y \in R_h} G_h(x, y) (\Delta \phi)_h(y)$$

and it follows immediately from Hölder's inequality and Lemma 1 that

$$(3.10) \quad \|\tilde{\phi}_h\|_{L_\infty} \leq C \left(h^N \sum_{R_h} |(\Delta \phi)_h|^q \right)^{1/q}, \quad q > \frac{N}{2},$$

But from (3.8) we have the following:

Lemma 2. For any $\phi \in V$

$$\|\tilde{\phi}_h\|_{L_\infty} \leq C \|\Delta \phi\|_{L_q}, \quad q > \frac{N}{2}.$$

where C does not depend on h or ϕ .

Here and in the sequel we use C for a generic constant not necessarily the same in any two places.

Let us consider $v = \phi_h$ in (3.4). Since, from the definition of ϕ_h , ϕ'_h has support contained in R we have

$$(3.11) \quad \|\phi'_h\|_{H_1}^2 \leq C \left\{ h^N \sum_{R_h} \phi_h^2 + \sum_{i=1}^N \left(h^N \sum_{E_{Nh}} (\nabla_i \phi_h)^2 \right) \right\},$$

and hence $\phi'_h \in \dot{H}_1$. But it is well known that there is a constant C independent of h and ϕ'_h such that

$$(3.12) \quad h^N \sum_{R_h} \phi_h^2 \leq C h^N \sum_{E_{Nh}} (\nabla_i \phi_h)^2.$$

Also from partial summation we have

$$(3.13) \quad h^N \sum_{R_h} (\nabla_i \phi_h)^2 = -h^N \sum_{R_h} \phi_h \Delta_h \phi_h.$$

(The notation \sum with nothing written below will always mean summation over E_{Nh} .) An immediate consequence of (3.11)–(3.12), Schwarz's inequality and the definition of ϕ_h is the following lemma.

Lemma 3. There exists a constant C , independent of ϕ and h such that

$$(3.14) \quad \|\phi'_h\|_{H_1} \leq C \|\Delta \phi\|_{L_\infty}$$

for all $\phi \in V$.

IV. Existence, Uniqueness and Convergence

We are now in a position to prove the following.

Theorem 1. There exists one and only one solution of problem (D).

Proof. Uniqueness. Let u_1 and u_2 be any two solutions. Then if $v = u_1 - u_2$ we have that $\int_R v \Delta \phi dx = 0$, $\forall \phi \in V$. We want to show that $\int_R v \psi dx = 0$, $\forall \psi \in C_0^\infty(R)$, i.e. that $\Delta \phi = \psi$ has a solution in V for each $\psi \in C_0^\infty(R)$. Consider the equation

$$(4.1) \quad - \sum_{i=1}^N \int_R \frac{\partial \bar{\phi}}{\partial x_i} \frac{\partial \chi}{\partial x_i} dx = \int_R \chi \psi dx, \quad \forall \chi \in \dot{H}_1.$$

It is well known that (4.1) has a unique solution $\bar{\phi} \in \dot{H}_1$. But by Weyl's lemma there exists ϕ such that $\Delta \phi = \psi$ and $\phi = \bar{\phi}$ almost everywhere. Thus $\phi \in V$ and we have $\int_R v \psi dx = 0$, $\forall \psi \in C_0^\infty(R)$. Since $v \in L_p$ for some $p \geq 1$ it follows that $v = 0$ in L_p .

Existence. In order to prove the existence it is sufficient to prove that there exists a constant C depending only on q and R such that

$$(4.2) \quad \|\phi\|_{L_\infty} \leq C \|\Delta \phi\|_{L_q}, \quad q > \frac{N}{2}$$

for all $\phi \in V$. For, let us consider the linear functional

$$(4.3) \quad T(\Delta \phi) = \langle \phi, F \rangle$$

defined on the linear subspace of L_q , $q < \infty$, defined by $\{\psi | \psi = \Delta\phi, \phi \in V\}$. It follows from (4.2) that T is well defined and in fact continuous. Thus from the Hahn-Banach theorem T may be extended as a continuous linear functional to the whole space L_q . Since the linear functionals on L_q can be represented as integrals it follows that a function u exists in L_p , $1 < p < \frac{N}{N-2}$ such that

$$\int u \Delta\phi \, dx = \langle \phi, F \rangle, \quad \forall \phi \in V.$$

To prove (4.2) we use the difference method. Let $\phi \in V$. Applying Lemma 2 and using the weak compactness of bounded sets in L_p , $1 < p < \infty$ or more generally, reflexive Banach spaces, it follows that there exists a sequence $\{h_n\}$ such that $h_n \rightarrow 0$ as $n \rightarrow \infty$ and $\tilde{\phi}_{h_n} \rightarrow \tilde{\phi}$ weakly in L_p for $1 < p < \infty$. Clearly

$$(4.4) \quad \|\tilde{\phi}\|_{L_\infty} \leq C \|\Delta\phi\|_{L_q}, \quad q > \frac{N}{2},$$

i.e. the inequality of Lemma 2 holds in the limit. Applying Lemma 3 and noting that $\phi'_h \in \dot{H}_1$, it follows again from the weak compactness of bounded subsets of \dot{H}_1 that there exists $\phi' \in \dot{H}_1$ such that a subsequence of $\{\phi'_{h_n}\}$ (call it again $\{\phi'_{h_n}\}$) converges weakly in \dot{H}_1 to ϕ' . We will show that $\phi' = \phi$ and that $\tilde{\phi} = \phi'$.

Now for h sufficiently small we have, for any $\psi \in C_0^\infty(R)$,

$$(4.5) \quad h^N \sum \phi_h \Delta_h \psi = h^N \sum \psi (\Delta\phi)_h.$$

From the uniform boundedness of ϕ_h and $(\Delta\phi)_h$ it follows easily that

$$(4.6) \quad \int_R \phi'_{h_n} \Delta \psi \, dx = \int_R \psi \Delta \phi \, dx + o(1)$$

as $h_n \rightarrow 0$. But from the weak convergence of ϕ'_{h_n} to ϕ' we have

$$(4.7) \quad \int_R \phi' \Delta \psi \, dx = \int_R \psi \Delta \phi \, dx.$$

Clearly for all $\psi \in C_0^\infty(R)$ and $\phi \in V$

$$(4.8) \quad \int_R \phi \Delta \psi \, dx = \int_R \psi \Delta \phi \, dx$$

and hence

$$(4.9) \quad \int_R (\phi - \phi') \Delta \psi \, dx = 0, \quad \forall \psi \in C_0^\infty(R).$$

Now $\phi - \phi' \in \dot{H}_1$ and applying Weyl's lemma there exists $W = \phi - \phi'$ almost everywhere in R and such that $\Delta W = 0$. But this implies that $W = 0$. Hence $\phi' = \phi$ almost everywhere.

To show that $\tilde{\phi} = \phi' = \phi$, almost everywhere, we again consider an arbitrary $\psi \in C_0^\infty(R)$. Now

$$(4.10) \quad \int_R (\tilde{\phi} - \phi) \psi \, dx = \int_R (\tilde{\phi} - \tilde{\phi}_{h_n}) \psi \, dx + \int_R (\tilde{\phi}_{h_n} - \phi) \psi \, dx + \int_R (\tilde{\phi}_{h_n} - \phi'_{h_n}) \psi \, dx.$$

The first two terms clearly tend to zero as $n \rightarrow \infty$ since $\tilde{\phi}$ and ϕ are the respective weak limits of $\{\tilde{\phi}_{h_n}\}$ and $\{\phi'_{h_n}\}$. Because of the smoothness of ψ the last term is

easily seen to satisfy

$$(4.11) \quad \left| \int_R (\tilde{\phi}_{h_n} - \phi'_{h_n}) \psi \, dx \right| \leq C h \|A\phi\|_{L_1}.$$

Thus $\int_R (\phi - \tilde{\phi}) \psi \, dx = 0$, $\forall \psi \in C_0^\infty(R)$, and hence $\tilde{\phi} = \phi$ almost everywhere from which (4.2) follows. This completes the proof.

We shall consider now the case in which $F \in L_1$ so that we may write the Eq. (3.2) in the form

$$(4.12) \quad \int_R u A\phi \, dx = \int_R \phi F \, dx.$$

To define an approximating difference problem we take

$$(4.13) \quad (F)_h(x) = h^{-N} \int_{C_h(x)} F(y) \, dy, \quad x \in R_h$$

and consider u_h as the solution of

$$(4.14) \quad \begin{aligned} A u_h(x) &= (F)_h(x), & x \in R_h \\ u_h(x) &= 0, & x \in E_{Nh} - R_h. \end{aligned}$$

Now

$$(4.15) \quad u_h(x) = -h^N \sum_{y \in R_h} G_h(x, y) (F)_h(y).$$

For any p such that $1 \leq p < \frac{N}{N-2}$ we obtain

$$(4.16) \quad |u_h(x)|^p \leq h^N \sum_{y \in R_h} G_h(x, y) |u_h(x)|^{p-1} |(F)_h(y)|.$$

Summing both sides of (4.16) with respect to x we have

$$(4.17) \quad h^N \sum |u_h|^p \leq h^N \sum_{y \in R_h} |(F)_h(y)| h^N \sum_{x \in R_h} G_h(x, y) |u_h(x)|^{p-1}.$$

Using Hölder's inequality, the symmetry of G_h and Lemma 1 we obtain

$$(4.18) \quad (h^N \sum |u_h|^p)^{1/p} \leq C h^N \sum_{y \in R_h} |(F)_h(y)|$$

from which follows immediately

$$(4.19) \quad \|\tilde{u}_h\|_{L_p} \leq C \|F\|_{L_1}, \quad 1 \leq p < \frac{N}{N-2}.$$

Again we can extract a sequence, (call it again $\{h_n\}$) such that $\tilde{u}_{h_n} \rightarrow u^*$ weakly in L_p , $1 < p < \frac{N}{N-2}$. But

$$(4.20) \quad \int_R \tilde{u}_{h_n} A\phi \, dx = h_n^N \sum u_{h_n} (A\phi)_{h_n} = h_n^N \sum \phi_{h_n} (F)_{h_n} = \int_R \tilde{\phi}_{h_n} F \, dx.$$

Now the left hand side tends to $\int_R u^* A\phi \, dx$ as $n \rightarrow \infty$. Since $F \in L_1$ we can approximate F by a sequence $\{F_m\}$ such that $F_m \in C_0^\infty(R)$ for all m and $\lim_{m \rightarrow \infty} \int_R |F - F_m| \, dx = 0$. Hence

$$(4.21) \quad \left| \int_R (\tilde{\phi}_{h_n} - \phi) F \, dx \right| \leq \left| \int_R (\tilde{\phi}_{h_n} - \phi) F_m \, dx \right| + \left| \int_R (\tilde{\phi}_{h_n} - \phi) (F - F_m) \, dx \right|.$$

Now since $\tilde{\phi}_{h_n}$ and ϕ are bounded we can choose m so large that the last term of (4.21) is as small as we like, say $\varepsilon/2$. Then we may choose n so large that the first term is less than $\varepsilon/2$. Hence it follows that

$$(4.22) \quad \int_R \tilde{\phi}_{h_n} F dx \rightarrow \int_R \phi F dx \quad \text{as } h_n \rightarrow 0.$$

Thus

$$\int_R u^* \Delta \phi dx = \int_R \phi F dx$$

and because of the uniqueness $u^* = u$. Also from the uniqueness it follows that every subsequence converges weakly to u . Thus we have proved

Theorem 2. Let $F \in L_1$. Then $\{\tilde{u}_h\}$ defined by (4.14) converges weakly to u in L_p , for $1 \leq p < \frac{N}{N-2}$, as $h \rightarrow 0$.

In order to show that in fact $\{u_h\}$ converges strongly to u we again approximate F by $\{F_n\}$ in such a way that $F_n \in C_0^\infty(R)$ for each n and

$$\lim_{n \rightarrow \infty} \int_R |F - F_n| dx = 0.$$

Now let u_n be the solution of problem (D) with F replaced by F_n and u_{nh} be the solution of (4.14) with F replaced by F_n .

Now by the triangle inequality

$$(4.23) \quad \|\tilde{u}_h - u\|_{L_p} \leq \|\tilde{u}_h - \tilde{u}_{nh}\|_{L_p} + \|\tilde{u}_{nh} - u_n\|_{L_p} + \|u_n - u\|_{L_p}.$$

From (4.19) it certainly follows that

$$\|\tilde{u}_h - \tilde{u}_{nh}\|_{L_p} \leq C \|F - F_n\|_{L_1}$$

and since by Theorem 2, $u - u_n$ is the weak limit of $\tilde{u}_h - \tilde{u}_{nh}$ as $h \rightarrow 0$ we also have

$$\|u_n - u\|_{L_p} \leq C \|F - F_n\|_{L_1}, \quad 1 \leq p < \frac{N}{N-2}.$$

Hence given $\varepsilon > 0$ we can choose n such that $\|F - F_n\|_{L_1} \leq \varepsilon/4C$ and hence

$$(4.24) \quad \|\tilde{u}_h - u\|_{L_p} \leq \|\tilde{u}_{nh} - u_n\|_{L_p} + \varepsilon/2.$$

Now since $F_n \in L_2$ it follows from the results of CÉA [4] that

$$\lim_{h \rightarrow 0} \|\tilde{u}_{nh} - u_n\|_{L_2} = 0.$$

But $u_n \in V$ so that from Lemma 2 and (4.2) $\tilde{u}_{nh} - u_n$ is bounded. Hence we may choose h so small that

$$\|\tilde{u}_{nh} - u_n\|_{L_p} < \varepsilon/2 \quad \text{for any } p \geq 1.$$

Thus $\|\tilde{u}_h - u\|_{L_p} < \varepsilon$ and we have

Theorem 3. Let $F \in L_1$. Then $\{\tilde{u}_h\}$ defined by (4.14) converges strongly to u in L_p for $1 \leq p < \frac{N}{N-2}$ as $h \rightarrow 0$.

We want to consider now the more general case in which $F \in (L_\infty)'$. Let $M_h(x)$ be the function of y for each $h > 0$ and each $x \in R_h$ defined as

$$M_h(x) = \begin{cases} h^{-N}, & y \in C_h(x) \\ 0, & y \notin C_h(x). \end{cases}$$

In this case we define $u_h(x)$ to be the solution of

$$(4.25) \quad \begin{aligned} \Delta_h u_h(x) &= \langle M_h(x), F \rangle, & x \in R_h \\ u_h(x) &= 0, & x \notin R_h. \end{aligned}$$

By the same techniques as before it is easily seen that

$$(4.26) \quad \|\tilde{u}_h\|_{L_p} \leq C \|F\|_{(L_\infty)'}, \quad 1 \leq p < \frac{N}{N-2}.$$

As before we obtain a subsequence $\{\tilde{u}_{h_n}\}$ such that $\tilde{u}_{h_n} \rightarrow u^*$ weakly in L_p , $1 \leq p < \frac{N}{N-2}$. In analogy with (4.20) one easily verifies that

$$(4.27) \quad \int_R \tilde{u}_{h_n} \Delta \phi \, dx = h_n^N \sum u_{h_n}(\Delta \phi)_{h_n} = h_n^N \sum \phi_{h_n} \langle M_{h_n}(x), F \rangle = \langle \tilde{\phi}_{h_n}, F \rangle.$$

As before

$$\lim_{n \rightarrow \infty} \int_R \tilde{u}_{h_n} \Delta \phi \, dx = \int_R u^* \Delta \phi \, dx.$$

The question is: When does $\langle \tilde{\phi}_{h_n}, F \rangle \rightarrow \langle \phi, F \rangle$ as $n \rightarrow \infty$? The next two theorems give sufficient conditions.

Theorem 4. Let $F \in (L_\infty)'$ and suppose that F has the property that, for some compact subset Ω of R , $\langle v, F \rangle = 0$ for all $v \in L_\infty$ which vanish on Ω . Then $\{\tilde{u}_h\}$ converges to u weakly in L_p , $1 \leq p < \frac{N}{N-2}$ as $h \rightarrow 0$.

In order to complete the proof we need to show that if Ω is any compact subset of R then $\tilde{\phi}_{h_n} \rightarrow \phi$ uniformly on Ω , at least for a subsequence of $\{h_n\}$. But this is not difficult since it is shown in [1] that for any open subset Ω' whose closure is contained in R there is a constant $K(\Omega')$ such that

$$(4.28) \quad \max_{\Omega'} |\nabla_i \phi_{h_n}| \leq K(\Omega') (\|\tilde{\phi}_{h_n}\|_{L_\infty} + \max_{i, R} |\nabla_i \Delta \phi|).$$

The right hand side is clearly bounded so that if we take $\Omega' \supset \Omega$ it follows from the definition of ϕ'_{h_n} that ϕ'_{h_n} has first difference quotients which are uniformly bounded on Ω' , the bound not depending on h_n for h_n sufficiently small. Thus by the Ascoli-Arzelà theorem there is a subsequence, call it again $\{\phi'_{h_n}\}$, which converges uniformly (to ϕ) on Ω . But by the definitions of $\tilde{\phi}_h$ and ϕ'_h and (4.28) it follows that $\sup_{\Omega} |\tilde{\phi}_{h_n} - \phi'_{h_n}| \rightarrow 0$ and hence $\tilde{\phi}_{h_n} \rightarrow \phi$ uniformly on Ω as $n \rightarrow \infty$.

Thus, as before, we see that $\langle \tilde{\phi}_{h_n}, F \rangle \rightarrow \langle \phi, F \rangle$ as $n \rightarrow \infty$, that $u^* = u$ and that every subsequence converges to u .

In the next theorem, instead, we impose some regularity on ∂R .

Theorem 5. Let $F \in (L_\infty)'$ and $\partial R \in C^2$. Then $\{\tilde{u}_h\}$ converges to u weakly in L_p , $1 \leq p < \frac{N}{N-2}$, as $h \rightarrow 0$.

To complete the proof it suffices to remark that it was shown recently in a paper by BRAMBLE, HUBBARD and THOMÉE [2] that if $\partial R \in C^2$ the sequence of difference approximations, defined slightly differently from that of (3.8), in fact converges uniformly, with the error tending to zero quadratically. By exactly the same considerations it can be shown that the solution of (3.8) satisfies

$$(4.29) \quad \sup_R |\tilde{\phi}_h - \phi| \leq Ch.$$

Clearly we can state, from the above considerations, the following result.

Theorem 6. Let $F \in (L_\infty)'$ and $\partial R \in C^2$. Then for any $\psi \in C_0^\infty(R)$

$$\left| \int_R (\tilde{u}_h - u) \psi \, dx \right| \leq Ch \|F\|_{(L_\infty)'},$$

where C depends on ψ but not on h .

V. Further Convergence Results

In order to study further convergence properties we introduce the generalized Green's function. For any point x in R , $G(x, y)$ is defined as the solution of

$$(5.1) \quad \phi(x) = - \int_R G(x, y) \Delta \phi(y) \, dy$$

for all $\phi \in V$. From Theorem 1, $G(x, y)$ exists and is unique and as a function of y belongs to L_p , $1 \leq p < \frac{N}{N-2}$. At this point we note the interesting special case of Theorem 4.

Corollary of Theorem 4. $\tilde{G}_h(x, y)$ converges weakly to $G(x, y)$ as a function of y in L_p , $1 \leq p < \frac{N}{N-2}$ as $h \rightarrow 0$, for each fixed x in R . By $\tilde{G}_h(x, y)$ we mean that $\tilde{G}_h(x, y) = G_h(x_0, y_0)$ if $(x, y) \in C_h(x_0) \times C_h(y_0)$, $(x_0, y_0) \in E_{Nh} \times E_{Nh}$.

The functional F in this case is the so called "Dirac delta function", i.e. the linear functional on V defined by $\langle \phi, F \rangle = \phi(x)$ and extended to L_∞ by the Hahn-Banach theorem. Clearly there is an extension F which has properties required in Theorem 4 so the corollary follows. We want to show the symmetry of G . In order to prove this we need the following lemma.

Lemma 4. The Green's function $G(x, y)$ belongs to $L_p(R \times R)$, $1 \leq p < \frac{N}{N-2}$.

Proof. For $h > 0$, $G_h(x, y)$, defined in Section III satisfies

$$(5.2) \quad \|\tilde{G}_h\|_{L_p(R \times R)} \leq C, \quad 1 \leq p < \frac{N}{N-2},$$

where C does not depend on h . This follows from (3.6). Thus again since for $1 < p < \frac{N}{N-2}$, $L_p(R \times R)$ is a reflexive Banach space we can extract a subsequence $\{\tilde{G}_{h_n}\}$ such that $\tilde{G}_{h_n} \rightarrow G^*$ weakly in $L_p(R \times R)$. We shall show that $G^* = G$ (in $L_p(R \times R)$). Let ψ_1 and ψ_2 belong to $C_0^\infty(R)$. We consider

$$(5.3) \quad \begin{aligned} & \int_R \psi_1(x) \left[\int_R (G(x, y) - G^*(x, y)) \psi_2(y) \, dy \right] dx \\ &= \int_R \psi_1(x) \left[\int_R (G(x, y) - \tilde{G}_{h_n}(x, y)) \psi_2(y) \, dy \right] dx \\ & \quad + \int_{R \times R} (\tilde{G}_{h_n}(x, y) - G^*(x, y)) \psi_1(x) \psi_2(y) \, dx \, dy, \end{aligned}$$

the last integral being written as an integral over $R \times R$. That this is permissible follows from the theorem of FUBINI-TONELLI (c.f. [8], p. 18) since $\tilde{G}_h - G^* \in L_p(R \times R)$. It is clear from the definition of G^* that the last term on the right of (5.3) tends to zero as $n \rightarrow \infty$. The first term has the form

$$\int_R \psi_1(x) (\phi(x) - \tilde{\phi}_{h_n}(x)) dx$$

where $\phi \in V$ and $\tilde{\phi}_{h_n}$ is defined as in (3.8) with $\Delta\phi = \psi_2$. But from the weak convergence in L_p of $\tilde{\phi}_{h_n}$ to ϕ , as was shown in the proof of Theorem 1, it follows that this term also tends to zero as $n \rightarrow \infty$. Thus

$$\int_R \psi_1(x) \left[\int_R (G(x, y) - G^*(x, y)) \psi_2(y) dy \right] dx = 0$$

for all ψ_1 and ψ_2 in $C_0^\infty(R)$, which implies that $G(x, y) = G^*(x, y)$ for almost all (x, y) in $R \times R$. This completes the proof of Lemma 4. We now can prove the symmetry relation

Lemma 5. $G(x, y) = G(y, x)$.

Proof. $G(x, y)$ is the weak limit in $L_p(R \times R)$, $1 < p < \frac{N}{N-2}$, of the sequence $\{\tilde{G}_{h_n}\}$ as $n \rightarrow \infty$ and for each n , $\tilde{G}_{h_n}(x, y) = \tilde{G}_{h_n}(y, x)$.

We can now prove the following representation.

Lemma 6. Let $F \in L_1$. Then

$$(5.4) \quad u(x) = - \int_R G(x, y) F(y) dy$$

is the solution of problem (D).

(We shall only use this lemma here when $F \in L_q$, $q > N/2$ but since it is true for $q = 1$ and $L_q \subset L_1$ we prove it in that case.)

Proof. Since $G(x, y)$ is integrable in $R \times R$ it follows from the Fubini-Tonelli theorem that for $\phi \in V$, $\int_R |G(x, y)| |\Delta\phi(x)| dx$ is integrable, since $G(x, y)$ is the weak limit in L_p , $1 \leq p < \frac{N}{N-2}$ of $G_h(x, y)$ we conclude from (3.7) and the symmetry of G that its L_p norm as a function of x is also bounded. Hence

$$\int_R |F(y)| \left[\int_R |G(x, y)| |\Delta\phi(x)| dx \right] dy$$

is finite and again, using the Fubini-Tonelli theorem, (5.4) and Lemma 5, we have

$$\begin{aligned} \int_R \phi F dx &= - \int_R F(y) \left[\int_R G(x, y) \Delta\phi(x) dx \right] dy \\ &= \int_R \Delta\phi(x) \left[- \int_R G(x, y) F(y) dy \right] dx = \int_R u \Delta\phi dx, \end{aligned}$$

which proves the lemma.

With the preceding lemma we can prove the following convergence theorem.

Theorem 7. Let $F \in L_q$, $q > N/2$. Then $\{\tilde{u}_h\}$ converges pointwise to u in R as $h \rightarrow 0$.

Proof. We can write, using Lemma 6 and the definition of u_h

$$u_h(x) - u(x) = \int_R (G(x, y) - \tilde{G}_h(x, y)) F(y) dy.$$

Now for each fixed x in R , $\tilde{G}_h(x, \cdot)$ converges weakly to $G(x, \cdot)$ in L_p for $1 < p < \frac{N}{N-2}$. The theorem follows immediately.

In case the boundary is somewhat regular we can prove

Theorem 8. Let $F \in L_q$, $q > N/2$ and $\partial R \in C^2$. Then $\{\tilde{u}_h\}$ converges uniformly to u as $h \rightarrow 0$.

Proof. Let $F_n \in C_0^\infty(R)$ for $n = 1, 2, \dots$ and $\{F_n\}$ converge strongly to F in L_q , $q > N/2$. Then

$$\begin{aligned} \tilde{u}_h(x) - u(x) &= \int_R (G(x, y) - \tilde{G}_h(x, y)) F_n(y) dy \\ &\quad + \int_R (G(x, y) - \tilde{G}_h(x, y)) (F(y) - F_n(y)) dy. \end{aligned}$$

Hence

$$(5.5) \quad |\tilde{u}_h(x) - u(x)| \leq |\tilde{\phi}_{nh}(x) - \phi_n(x)| + \|\tilde{G}_h(x, \cdot) - G(x, \cdot)\|_{L_p} \|F - F_n\|_{L_q}$$

$q > \frac{N}{2}$, $\frac{1}{p} + \frac{1}{q} = 1$. In (5.5) $\phi_n(x)$ is defined as $-\int_R G(x, y) F_n(y) dy$ and ϕ_{nh} correspondingly. Since $F_n \in C_0^\infty(R)$, $\phi_n \in V$ and as is discussed in the proof of Theorem 5, $\phi_{nh} \rightarrow \phi_n$ uniformly as $h \rightarrow 0$ for each fixed n . Now from (3.7) it follows that $\|\tilde{G}_h(x, \cdot) - G(x, \cdot)\|_{L_p}$, $1 \leq p < \frac{N}{N-2}$ is uniformly bounded in x and h . Clearly if we take n sufficiently large and then h small we can make the right hand side less than any preassigned $\varepsilon > 0$. This proves Theorem 8.

References

1. BRAMBLE, J. H., and B. E. HUBBARD: Approximation of derivatives by finite difference methods in elliptic boundary value problems. Contributions to Diff. Eq. **3**, 399–410 (1964).
2. —, and V. THOMÉE: Convergence estimates for essentially positive type discrete Dirichlet problems. Math. of Comp. (to appear).
3. —, and M. ZLÁMAL: Discrete analogs of the Dirichlet problem with isolated singularities. SIAM J. Numer. Anal. **5**, No. 1, 1–25 (1968).
4. CÉA, J.: Approximation variationnelle des problèmes aux limites. Ann. Inst. Fourier **14**, 2, 345–444 (1964).
5. COMINCIOLI, V.: Analisi numerica di alcuni problemi ai limiti per l'operatore di Laplace iterato. Rend. Sem. Mat. Padova (1965).
6. COURANT, R., D. FRIEDRICHS u. H. LEWY: Über die partiellen Differenzengleichungen der mathematischen Physik. Math. Ann. **100**, 32–74 (1928).
7. GERSHGORIN, S.: Fehlerabschätzung für das Differenzenverfahren zur Lösung partieller Differentialgleichungen. ZAMM **10**, 373–382 (1930).
8. YOSIDA, K.: Functional analysis. New York: Academic Press; Berlin-Heidelberg-New York: Springer 1965.

Dr. J. H. BRAMBLE
Department of Mathematics
Cornell University
Ithaca, New York, USA

2

Finite element method

2.1 New monotone type approximations for elliptic problems (1964)

New monotone type approximations for elliptic problems[[11](#)]

New Monotone Type Approximations for Elliptic Problems

By James H. Bramble and Bert E. Hubbard

I. Introduction. In the usual study of the discretization error resulting from approximating boundary problems for elliptic equations by finite difference methods the maximum principle plays a central role. In 1930 S. Gershgorin [11] gave a method for estimating the order of convergence of the solution to a certain class of finite difference analogues to the solution of the Dirichlet problem for elliptic equations. The matrix of the resulting system of simultaneous linear equations belongs to a special class for which one can easily prove that the inverse exists and has only non-negative elements. Interpreted in the language of analysis this means that the finite difference Green's function shares the property of non-negativity with its continuous counterpart.

Armed with this knowledge the final step in relating the discretization error (the difference between the solutions of the continuous and discrete problems) to the local truncation error (the error produced in approximating the differential equation in the region R and boundary conditions on the boundary C) now only involves bounding the maximum row sum of the inverse matrix. This can be accomplished easily, either directly or by using a function which bounds the corresponding integrals in the continuous problem. The bound thus produced is independent of the mesh size, h . In fact, using this approach, we are able to isolate the contribution to the discretization error arising from the local truncation error in different parts of the region. In particular it can be shown that under certain conditions if one desires the discretization error to be $O(h^n)$ then it is sufficient that the local truncation error be

- (a) $O(h^n)$ in the interior of the region R ,
- (b) $O(h^n)$ on the boundary C ,
- (c) $O(h^{n-1})$ at points in R adjacent to C_1 (that portion of C where a mixed or Neumann condition is given),
- (d) $O(h^{n-2})$ at points in R adjacent to C_2 (that portion of C where Dirichlet data are given).

A discussion of these questions for the Dirichlet problem for Poisson's equation can be found in Bramble and Hubbard [3].

As was mentioned earlier the matrix of the finite difference analogue formulated by S. Gershgorin belongs to a special class. In particular it is of "positive type" and thus is easily shown to possess a non-negative inverse. Matrices which arise very naturally in connection with elliptic boundary value problems may or may not be of positive type. The main purpose of this paper is to show that in many of these cases it is possible to prove that the inverse matrix exists and is non-negative

Received August 13, 1963. Revised January 6, 1964. This research was supported in part by a grant with the National Science Foundation NSF Grant GP-3, and by a grant with the Air Force Office of Scientific Research, Air Research and Development Command AFOSR 62-454.

even though the matrices are not of positive type. In each case we then derive the associated error estimates.

In Section 2 certain known theorems on monotone matrices are presented to lay the groundwork for two new theorems, 2.6 and 2.7, which give sufficient conditions for monotonicity. Theorem 2.6 in fact gives a necessary and sufficient condition for a matrix to be monotone.

The remaining sections give applications of Theorem 2.7 to estimates of the order of convergence of the solution of certain finite difference problems to the solutions of various elliptic boundary value problems. In each case the matrix of the resulting linear system violates the sufficient conditions for monotonicity given in the classical theorems.

In Section 3 an $O(h^4)$ finite difference analogue of the Dirichlet problem for Poisson's equation in a rectangle is given. The finite difference Laplace operator is the nine point cross. The usual five point $O(h^2)$ operator is used near the boundary, yet the discretization error is shown to be $O(h^4)$. In Section 4 the results of Section 3 are extended to general regions. In Section 5 a very high order ($O(h^9)$) approximation for the rectangle is given where the finite difference Laplace operator is a thirteen point $O(h^{10})$ operator.

Finally we conclude with an application of Theorem 2.7 to the "seemingly most natural" finite difference analogue of the Dirichlet problem for the elliptic operator $U_{xx} + U_{xy} + U_{yy}$.

For a further introduction to this problem cf. Forsythe and Wasow [10] and the references contained therein.

II. Matrix Preliminaries. As a prelude to the study of the discretization error we shall classify the matrices involved and discuss their properties.

It is well known that if $v(x, y)$ is a sufficiently smooth function for which $-\Delta v \geq 0$ in R and $v \geq 0$ on C then $v \geq 0$ in R also. This property of the Laplace operator is sometimes called the "maximum principle." Since this is true in the limiting case, we should expect that for sufficiently small mesh size our finite difference analogue would possess the same property.

Definition 2.1. A matrix A is said to be "monotone" if $Ax \geq 0$ implies $x \geq 0$ for any vector x . (The inequality is understood to be element-wise.)

Another characterization of monotone matrices is given by the following well-known theorem, cf. Collatz [8].

THEOREM 2.1. *A is monotone if and only if A is non-singular and $A^{-1} \geq 0$ (i.e. each element of A^{-1} is non-negative).*

This property of monotone matrices corresponds to the non-negativity of the Green's function in the continuous problem. It is not easy, in general, to discover by inspection that a given matrix A is monotone, although this is a property which is useful in studying the order of convergence. However, many common finite difference analogues do belong to the following easily identifiable subclass of monotone matrices.

Definition 2.2. An $N \times N$ matrix B with elements b_{ij} is said to be of "positive type" if the following conditions are satisfied:

- (a) $b_{ji} \leq 0$, $i \neq j$
- (2.1) (b) $\sum_k b_{jk} \geq 0$ for all j , with $\sum_k b_{jk} > 0$ for $j \in J(B) \neq 0$,
- (c) for $i \notin J(B)$ there exists a finite sequence of non-zero elements of the form $b_{ik_1}, b_{k_1 k_2}, \dots, b_{k_r j}$ where $j \in J(B)$. Such a sequence is called a “connection” in B from i to $J(B)$.

THEOREM 2.2. *If B is of positive type, then B is monotone.*

This theorem has been proved for classes of matrices closely related to those of positive type cf. L. Collatz [8, p. 45], and for this case in [6].

A particular subclass of matrices of positive type are the Minkowski matrices considered by Ostrowski [15], [16], and [17]. A positive type matrix is a Minkowski matrix if $J(B) = \{1, 2, \dots, N\}$. Ostrowski also defines an intermediate class of matrices between those of positive type and those which are monotone, which he calls “ M -matrices.”

Definition 2.3. A monotone matrix B is an M -matrix if

$$(2.2) \quad b_{\alpha\beta} \leq 0, \quad \alpha \neq \beta.$$

This implies $b_{\alpha\alpha} > 0$, since if $b_{\alpha\alpha} \leq 0$ for some α then $1 = \sum_j b_{\alpha j} (b^{-1})_{j\alpha} \leq 0$.

THEOREM 2.3 (OSTROWSKI). *Let B satisfy (2.2); then B is an M -matrix if and only if all of the principle minors of B are positive.*

THEOREM 2.4 (OSTROWSKI). *Let B satisfy (2.2); then B is an M -matrix if there exists a vector $x \geq 0$ such that $Ax > 0$.*

We see from the above theorems, particularly the latter, that M -matrices form a somewhat more easily identified class of monotone matrices. The following theorem of Ostrowski, is of particular interest in this connection.

THEOREM 2.5. *An M -matrix B is characterized by the property (2.2) and the existence of a positive diagonal matrix D such that $D^{-1}BD$ is a Minkowski matrix.*

The monotone matrices which arise in certain finite difference analogues to elliptic boundary value problems are M -matrices (in most of these cases they are even of positive type). However, a wide class of otherwise acceptable finite difference analogues do not fit in this category. For example, the five point $O(h^4)$ approximation to U_{xx} violates the condition (2.2) since the coefficients alternate in sign. In general, those finite difference analogues with higher order local truncation error will not be M -matrices. From the heuristic argument given above, we might expect them to lead to monotone matrices, at least for sufficiently small mesh size. That this is indeed the case for a broad class, which includes positive type matrices, is the main point of our discussion.

The following generalization of Theorem 2.5 gives a characterization of the entire class of monotone matrices.

THEOREM 2.6. *B is a monotone matrix if and only if there exist non-negative matrices P_1 and P_2 such that P_1BP_2 is of positive type.*

For a proof of this theorem cf. Bramble and Hubbard [6]. This theorem was applied in that paper to yield higher order estimates for a finite difference analogue to the one-dimensional boundary value problem based on the $O(h^4)$, five-point approximation to d^2/dx^2 . The approach used there was analytical in nature, using

the properties of the Green's function in the continuous problem. In this paper we take an entirely algebraic approach which seems to be both simpler and to yield sharper results.

We note first that if B admits the factorization $B = B_1 \cdot B_2 \cdot \dots \cdot B_r$, where B_i , $i = 1, \dots, r$, are M -matrices then B is monotone. The following theorem suggests a factorization into M -matrices, which applies to the matrices of many common finite difference analogues of elliptic boundary value problems which are not themselves M -matrices.

THEOREM 2.7. *Let B have unit diagonal with $\sum_j b_{ij} \geq 0$, $J(B) \neq 0$. Let B be written as the matrix sum $B = I - H_1 - H_2$ where*

- (a) $(H_1)_{\alpha\alpha} = 0$,
- (2.3) (b) $I - H_1$ is of positive type,
- (c) $(I - H_1)^{-1}H_2 \geq 0$,
- (d) for each $k \notin J(B)$ there exists a "connection" in H_1 from k to $J(B)$.

Then the factorization

$$(2.4) \quad B = (I - H_1)[I - (I - H_1)^{-1}H_2]$$

is such that each factor on the right is of positive type and hence B is monotone.

Proof. Since $I - H_1$ is of positive type it is an M -matrix. Hence by Theorem 2.5 there exists a diagonal matrix D with positive diagonal elements such that $I - D^{-1}H_1D$ is a Minkowski matrix. Thus $\rho(H_1) = \rho(D^{-1}H_1D) < 1$. Hence the Neumann expansion converges; i.e.

$$(2.5) \quad (I - H_1)^{-1} = I + H_1 + (H_1)^2 + \dots.$$

We now show that

$$(2.6) \quad [I - (I - H_1)^{-1}H_2] = (I - H_1)^{-1}B$$

is a Minkowski matrix. By assumption, (2.2) is satisfied. It remains to be shown that the row sums of $(I - H_1)^{-1}B$ are positive. If $i \in J(B)$ we see from the first term on the right side of (2.5) that the corresponding row sum is positive. On the other hand, if $i \notin J(B)$ then by (2.3d) there is an integer r and an element in $(H_1)^r$ with $j \in J(B)$ such that

$$(2.7) \quad (H_1)_{ik_1} \cdot (H_1)_{k_1 k_2} \cdot \dots \cdot (H_1)_{k_r j} > 0.$$

Hence

$$\sum_k [(H_1)^r B]_{ik} = \sum_l (H_1)_{il}^r [\sum_k B_{lk}] \geq (H_1)_{ij}^r [\sum_k B_{jk}] > 0.$$

Now from (2.5) we see that the row sum of (2.6) corresponding to any row $i \notin J(B)$ is positive. Thus B is the product of two matrices of positive type and is therefore monotone.

III. Dirichlet Problem for Poisson's Equation in a Rectangle, $\epsilon = O(h^4)$. For simplicity assume that R is a rectangle in two dimensions, with a square mesh

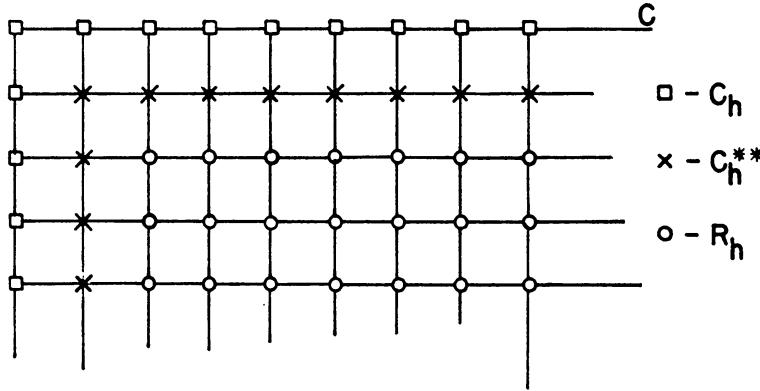


FIGURE 1

(size h) which fits R exactly. The necessary modifications which yield the corresponding estimate for the discretization error for general regions are treated separately in the next section. Consider the Dirichlet problem for Poisson's equation for the region R :

$$(3.1) \quad -\Delta u \equiv -\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) = f \quad \text{in } R, \quad u = g \text{ in } C.$$

Let the points of R_h , C_h^{**} , C_h be as indicated in Figure 1. We wish to formulate a finite difference analogue of (3.1) in such a manner that the discretization error is $O(h^4)$. We shall follow the general rules laid down in the introduction. For $(x, y) \in C_h^{**}$ let

$$(3.2) \quad \begin{aligned} \Delta_h V(x, y) \equiv h^{-2} & \{ V(x+h, y) + V(x-h, y) \\ & + V(x, y+h) + V(x, y-h) - 4V(x, y) \}. \end{aligned}$$

Clearly if $u(x, y) \in C^6$ in $R + C$ then

$$(3.3) \quad [\Delta_h u(x, y) - \Delta u(x, y)] = O(h^2).$$

For $(\xi, \eta) \in R_h$ let

$$(3.4) \quad \begin{aligned} \Delta_h V(\xi, \eta) \equiv h^{-2} & \{ -\frac{1}{12}[V(\xi+2h, \eta) \\ & + V(\xi-2h, \eta) + V(\xi, \eta+2h) + V(\xi, \eta-2h)] \\ & + \frac{4}{3}[V(\xi+h, \eta) + V(\xi-h, \eta) + V(\xi, \eta+h) + V(\xi, \eta-h)] \\ & - 5V(\xi, \eta) \}. \end{aligned}$$

Again we have

$$(3.5) \quad [\Delta_h u(\xi, \eta) - \Delta u(\xi, \eta)] = O(h^4).$$

The difference operator defined in (3.4) is seen to be the nine point "cross" which clearly violates the sign condition (2.2) and hence the resulting matrix will not be an M -matrix. A different $O(h^4)$ approximation is the nine point "box" operator which does satisfy (2.2) and has been considered previously [3]. The finite difference analogue of (3.1) is then given by

$$(3.6) \quad -\Delta_h V(p) = f(p), \quad p \in R_h + C_h^{**}, \quad V(p) = g(p), \quad p \in C_h.$$

Let \bar{A} be the matrix of the system (3.6). Define $A = D\bar{A}$ where D is the diagonal matrix defined by

$$(3.7) \quad d_{\alpha\alpha} = \begin{cases} 1, & \alpha \in C_h, \\ h^2/4, & \alpha \in C_h^{**}, \\ h^2/5, & \alpha \in R_h. \end{cases}$$

The matrix A has unit diagonal. The operator at the points (x, y) and (ξ, η) then has the coefficients given in Figure 2.

Because of the difficulty of visualizing the matrix A arising from the two dimensional problem we shall use the set of mesh points instead. A row of A corresponds to a point (like (x, y) or (ξ, η) in Figure 2) with which the finite difference operator is associated. The columns of A represent the points which are involved in the finite difference operator. For example the element $\frac{1}{6}$ will appear in the row associated with (ξ, η) , and columns corresponding to $(\xi \pm 2h, \eta)$, $(\xi, \eta \pm 2h)$.

LEMMA 3.1. *A is a monotone matrix.*

Proof. We shall decompose $A = I - H_1 - H_2$ and apply Theorem 2.7. Let H_1 be the matrix with zero rows corresponding to points of C_h and patterns at typical points $(x, y) \in C_h^{**}$ and $(\xi, \eta) \in R_h$ where $0 \leq \epsilon, \bar{\epsilon} \leq \frac{1}{4}$ and are otherwise arbitrary (Figure 3). The set $J(A)$ corresponds to points of C_h and clearly any point of $R_h + C_h^{**}$ is connected to C_h through elements of H_1 . $I - H_1$ is clearly of posi-

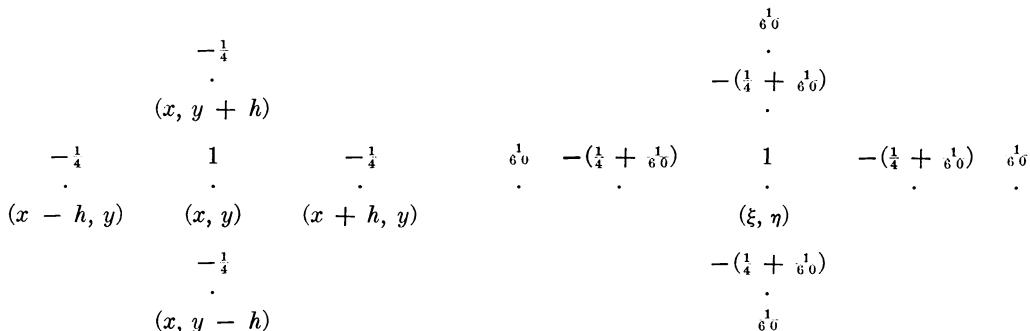


FIGURE 2

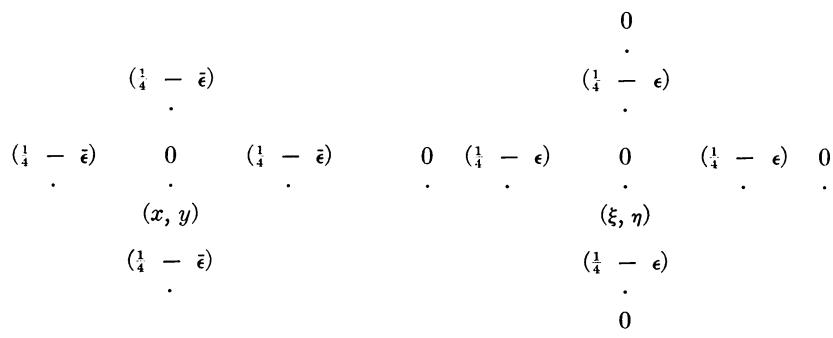


FIGURE 3

tive type. The matrix H_2 is represented by the patterns, shown in Figure 4, at points $(x, y) \in C_h^{**}$ and $(\xi, \eta) \in R_h$.

$$\begin{array}{ccccccccc}
 & & & -\frac{1}{60} & & & & & \\
 & & & \cdot & & & & & \\
 & \bar{\epsilon} & & & (\frac{1}{60} + \epsilon) & & & & \\
 & \cdot & & & \cdot & & & & \\
 & \bar{\epsilon} & 0 & \bar{\epsilon} & -\frac{1}{60} & (\frac{1}{60} + \epsilon) & 0 & (\frac{1}{60} + \epsilon) & -\frac{1}{60} \\
 & \cdot \\
 & (x, y) & & & & & (\xi, \eta) & & \\
 & \bar{\epsilon} & & & & & (\frac{1}{60} + \epsilon) & & \\
 & \cdot & & & & & \cdot & & \\
 & & & & & & \cdot & & \\
 & & & & & & -\frac{1}{60} & &
 \end{array}$$

FIGURE 4

We need only verify that $(I - H_1)^{-1}H_2 \geq 0$. Now

$$(3.8) \quad (I - H_1)^{-1}H_2 = H_2 + H_1H_2 + H_1^2H_2 + \dots$$

Let us examine this series term by term. The negative terms in H_2 are $-\frac{1}{60}$ and arise for example from a connection in H_2 from (ξ, η) to $(\xi + 2h, \eta)$ (Figure 5) or if $(\xi + h, \eta) \in C_h^*$ (Figure 6). We wish to determine $\epsilon, \bar{\epsilon}$ so that the indicated element in H_1H_2 is larger than $\frac{1}{60}$, i.e.

$$\begin{aligned}
 (3.9) \quad & (\frac{1}{4} - \epsilon)(\frac{1}{60} + \epsilon) \geq \frac{1}{60}, \\
 & (\frac{1}{4} - \epsilon)\bar{\epsilon} \geq \frac{1}{60}.
 \end{aligned}$$

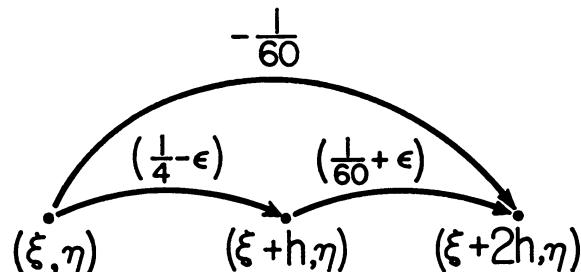


FIGURE 5

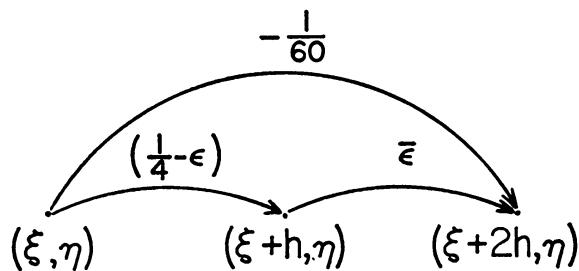


FIGURE 6

Let $\epsilon = \frac{1}{60}$ and $\tilde{\epsilon} = \frac{1}{8}$ and we see that the inequality is satisfied. Hence the negative elements in H_2 are cancelled out by terms in $H_1 H_2$. Similar considerations apply to any two successive terms since

$$(3.10) \quad H_1' H_2 + H_1'^{+1} H_2 = H_1'[H_2 + H_1 H_2].$$

In words, (3.10) tells us that the negative contribution toward the element of $(I - H_1)^{-1} H_2$ made by a product of the form

$$(H_1)_{ii_1}(H_1)_{i_1 i_2} \cdots (H_1)_{i_{r-1} i_r}(H_2)_{i_r j} < 0$$

is cancelled by adding a term of the type

$$(H_1)_{ii_1}, (H_1)_{i_1 i_2} \cdots (H_1)_{i_{r-1} i_r}(H_1)_{i_r k}(H_2)_{kj} > 0.$$

We note that this last term is needed nowhere else to overcome negative terms, a fact which is of crucial importance to the proof. This is the reason for all of the negative elements appearing in H_2 . The hypotheses of Theorem 2.7 are all satisfied and we conclude that A is a monotone matrix. Since $D \geq 0$ then $\bar{A} = D^{-1}A$ is also monotone. For $\epsilon, \tilde{\epsilon}$ as chosen above we can show by similar reasoning that $H_2(I - H_1)^{-1} \geq 0$. Since $H_2(I - H_1)^{-1}$ is similar to $(I - H_1)^{-1} H_2$ we see that they have the same spectral radius $\rho < 1$. Hence $[I - H_2(I - H_1)^{-1}]^{-1}$ exists and is non-negative. The matrix $I - H_2(I - H_1)^{-1}$ belongs to the more general class of matrices called M -matrices which includes those of positive type as was pointed out in Section 2.

Let the elements of the finite difference Green's function (\bar{A}^{-1} renormalized) be $g(p, q)$ defined by

$$(3.11) \quad \begin{aligned} -\Delta_{h,p} g(p, q) &= h^{-2} \delta(p, q), & p \in R_h + C_h^{**} \\ g(p, q) &= \delta(p, q), & p \in C_h \\ \delta(p, q) &= \begin{cases} 1, & p = q \\ 0, & p \neq q, \end{cases} \end{aligned}$$

where $q \in R_h + C_h^{**} + C_h$. Poisson's formula (which is just a restatement of the fact that $\bar{A}x = y \Rightarrow x = \bar{A}^{-1}y$) becomes

$$(3.12) \quad W(p) = h^2 \sum_{q \in R_h + C_h^{**}} g(p, q) [-\Delta_h W(q)] + \sum_{q \in C_h} g(p, q) W(q),$$

where $W(p)$ is an arbitrary mesh function. We note that $g(p, q) \geq 0$ by Lemma 3.1.

LEMMA 3.2. $h^2 \sum_{q \in R_h} g(p, q) \leq d^2/16$ where d is the diameter of the smallest circumscribed circle about R .

Proof. Let $W(p) \equiv d^2/16 - r^2/4 - h^2 \sum_{q \in R_h + C_h^{**}} g(p, q)$ where r is the distance from the center of the circle to p . Clearly $\bar{A}W \geq 0$ implies $W \geq 0$ and the conclusion follows.

LEMMA 3.3.

$$\sum_{q \in C_h^{**}} g(p, q) \leq 2.$$

Proof. Let \bar{D} be the diagonal matrix

$$(\bar{d}_{\alpha\alpha})^{-1} = \sum_j (I - H_1)_{\alpha j} = \begin{cases} 1, & \alpha \in C_h \\ 4\bar{\epsilon}, & \alpha \in C_h^{**} \\ 4\epsilon, & \alpha \in R_h \end{cases}$$

so that

$$(3.13) \quad \sum_j [\bar{D}(I - H_1)]_{\alpha j} = 1.$$

Consider the factorization

$$(3.14) \quad A = \bar{D}^{-1}[I - \bar{D}H_2(I - H_1)^{-1}\bar{D}^{-1}]\bar{D}(I - H_1)$$

and let

$$H = \bar{D}H_2(I - H_1)^{-1}\bar{D}^{-1}$$

so that

$$(3.15) \quad A^{-1} = [\bar{D}(I - H_1)]^{-1}(I - H)^{-1}\bar{D}.$$

Now

$$(3.16) \quad \begin{aligned} \sum_{q \in C_h^{**}} g(p, q) &= h^{-2} \sum_{q \in C_h^{**}} (A^{-1}D)_{pq} \\ &= \frac{1}{4} \sum_{q \in C_h^{**}} \{[\bar{D}(I - H_1)]^{-1}(I - H)^{-1}\bar{D}\}_{pq}. \end{aligned}$$

We see from (3.13) that

$$(3.17) \quad \begin{aligned} 1 &= \sum_k \sum_j \{[\bar{D}(I - H_1)]^{-1}\}_{pj} [\bar{D}(I - H_1)]_{jk} \\ &= \sum_j \{[\bar{D}(I - H_1)]^{-1}\}_{pj}. \end{aligned}$$

Hence

$$(3.18) \quad \sum_{q \in C_h^{**}} g(p, q) \leq \frac{1}{4} \left\{ \max_{\beta \in R_h + C_h^{**} + C_h} \sum_{q \in C_h^{**}} [(I - H)^{-1}\bar{D}]_{\beta q} \right\}.$$

From (3.3) we see that

$$(3.19) \quad 0 \leq \sum_j a_{\alpha j} = \sum_j [\bar{D}^{-1}(I - H)]_{\alpha j}.$$

Furthermore since H and $(I - H_1)^{-1}H_2$ are similar we have $\rho(H) = \rho[(I - H_1)^{-1}H_2] < 1$. Hence $(I - H)$ is nonsingular and $(I - H)^{-1} \geq 0$. We recall that

$$(3.20) \quad \bar{d}_{\alpha\alpha} = 2, \quad \alpha \in C_h^{**}; \quad \bar{d}_{\alpha\alpha} = 1, \quad \alpha \in C_h.$$

Now if $c^* \in C_h^{**}$ and $c \in C_h$ is such that $a_{c^*c} \neq 0$ then

$$(3.21) \quad [H_2(I - H_1)^{-1}]_{c^*c} \geq (H_2)_{c^*c} = \bar{\epsilon} = \frac{1}{8}$$

since the negative terms in the expansion are cancelled out by the remaining terms. In view of this we see that

$$(3.22) \quad \bar{d}_{c^*c}[H_2(I - H_1)^{-1}]_{c^*c}(\bar{d}^{-1})_{cc} \geq 2\bar{\epsilon} = \frac{1}{4}.$$

Now by defining y such that

$$(3.23) \quad y_\alpha = \begin{cases} 1, & \alpha \in R_h + C_h^{**}, \\ 0, & \alpha \in C_h, \end{cases}$$

we conclude, using (3.19) and (3.20), that

$$(3.24) \quad \{\tilde{D}^{-1}(I - H)y\}_{c^*} \geq \frac{1}{8}.$$

Hence

$$(3.25) \quad \begin{aligned} 1 &\geq \{(I - H)^{-1}\tilde{D}][\tilde{D}^{-1}(I - H)y]\}_\alpha \\ &> \sum_{j \in C_h^{**}} [(I - H)^{-1}\tilde{D}]_{\alpha j} [\tilde{D}^{-1}(I - H)y]_j \\ &\geq \frac{1}{8} \sum_{j \in C_h^{**}} [(I - H)^{-1}\tilde{D}]_{\alpha j}. \end{aligned}$$

Finally upon substituting (3.25) into (3.18) we prove the lemma.

The following theorem which relates the discretization error to the local truncation error now follows immediately.

THEOREM 3.1. *Let $\epsilon \equiv u - v$, where u, v are defined by (3.1) and (3.6) be the discretization error. If u has bounded sixth derivatives in $R + C$ then $\epsilon = O(h^4)$.*

Proof. Substituting ϵ into (3.12) and noting that $\epsilon = 0$ on C_h we see that

$$(3.26) \quad \begin{aligned} |\epsilon(p)| &\leq \left[h^2 \sum_{q \in R_h} g(p, q) \right] \{ \max_{R_h} |\Delta_h u - \Delta u| \} \\ &\quad + h^2 \left[\sum_{q \in C_h^{**}} g(p, q) \right] \{ \max_{C_h^{**}} |\Delta_h u - \Delta u| \}. \end{aligned}$$

Substituting (3.3), (3.5) into (3.26) and applying the result of Lemmas 3.2 and 3.3, yields the desired estimate. Here again as in [3] we note that the local truncation error at points near the boundary is only $O(h^2)$. Having demonstrated the ideas involved by considering for R a rectangle we now treat the problem for a general region.

IV. The Dirichlet Problem for Poisson's Equation in a General Region, $\epsilon = O(h^4)$. We again consider the problem (3.1) but for a region R with boundary C . A square mesh with mesh size h is placed on R . We define three disjoint sets of mesh points in R . Let C_h^* be made up of those points in R each of which has at least one of its four nearest neighbors lying in the complement of R . Let C_h^{**} be the set of mesh points in R with one or more neighbors in C_h^* . Let R_h be the remaining mesh points in R . We note that this implies that the four nearest neighbors of each point in R_h are in C_h^{**} , a fact of crucial importance in the development of the preceding section. Let the set of boundary crossings make up the set C_h .

Define the operator Δ_h by (3.2) and (3.4) on the sets C_h^{**} and R_h respectively. If $p = (x, y) \in C_h^*$ then u_{xx} and u_{yy} are each approximated to within $O(h^2)$ even though this will usually involve the use of unbalanced four point formulas in each case. For example if both $(x - \lambda h, y), (x, y - \mu h) \in C_h$ with $0 < \lambda, \mu \leq 1$ near the boundary (we assume that h is chosen so small compared to the radius of curvature of C that at most two neighbors of p will belong to C_h) then we define

$$\begin{aligned}
(4.1) \quad \Delta_x v(x, y) &\equiv h^{-2} \left[\frac{\lambda - 1}{\lambda + 2} v(x + 2h, y) + \frac{2(2 - \lambda)}{\lambda + 1} v(x + h, y) \right. \\
&\quad \left. + \frac{6}{\lambda(\lambda + 1)(\lambda + 2)} v(x - \lambda h, y) - \left(\frac{3 - \lambda}{\lambda} \right) v(x, y) \right], \\
\Delta_y v(x, y) &\equiv h^{-2} \left[\frac{\mu - 1}{\mu + 2} v(x, y + 2h) + \frac{2(2 - \mu)}{\mu + 1} v(x, y + h) \right. \\
&\quad \left. + \frac{6}{\mu(\mu + 1)(\mu + 2)} v(x, y - \mu h) - \left(\frac{3 - \mu}{\mu} \right) v(x, y) \right].
\end{aligned}$$

Of course if $(x - h, y)$ and $(x + h, y) \in C$ then each reduces to a three point operator. The assumption that p has at most two neighbors outside of R may eliminate from consideration certain regions having corners with acute angles.

We now define the operator Δ_h at the point $p \in C_h^*$ to be

$$(4.2) \quad \Delta_h v(p) \equiv \Delta_x v(p) + \Delta_y v(p),$$

and note that the local truncation error is

$$(4.3) \quad |\Delta_h u(p) - \Delta u(p)| = O(h^2).$$

The finite difference analogue of (3.1) is given by

$$\begin{aligned}
(4.4) \quad -\Delta_h V(p) &= f(p), \quad p \in R_h + C_h^{**} + C_h^*, \\
V(p) &= g(p), \quad p \in C_h.
\end{aligned}$$

It is not clear that the inverse matrix (Green's function) for the total problem is non-negative although it can be easily established as in [3] that the inverse matrix exists. This question has been considered in [20] for an $O(h^4)$ analogue based on the usual nine point approximation to Δ at points of R_h given in [3]. Such knowledge is not required to establish the order of the discretization error as was pointed out in [10, p. 288] and utilized in [3]. The same technique can be used here. We define an interior finite difference Green's function $g(p, q)$ as the solution of the following problem for each value of the parameter $q \in R_h + C_h^{**} + C_h^*$

$$\begin{aligned}
(4.5) \quad -\Delta_{h,p} g(p, q) &= \delta(p, q) h^{-2}, \quad p \in R_h + C_h^{**}, \\
g(p, q) &= \delta(p, q), \quad p \in C_h^*.
\end{aligned}$$

We note that the considerations of the preceding section, while derived only for a rectangle, are equally valid for a rectilinear region whose sides lie along mesh lines. The mesh region $R_h + C_h^{**}$ with boundary points C_h^* are of the same type as would arise in such a case. Hence the considerations of the preceding section apply directly to $g(p, q)$ as defined by (4.5) including its existence, non-negativity, and the inequalities given in Lemmas 3.2 and 3.3.

The Poisson formula in this case is given by

$$(4.6) \quad W(p) = h^2 \sum_{q \in R_h + C_h^{**}} g(p, q) [-\Delta_h W(q)] + \sum_{q \in C_h^*} g(p, q) W(q),$$

where $W(q)$ is any mesh function defined on the point set $R_h + C_h^{**} + C_h^*$.

Substituting the function $W \equiv 1$ into (4.6) yields the relation

$$(4.7) \quad \sum_{q \in C_h^*} g(p, q) = 1.$$

We note further that if \bar{A} is the matrix of the system (4.4) then for any $i \in C_h^*$ and any function W which vanishes at points of C_h we have the equation

$$(4.8) \quad W_i = \frac{\sum_j \bar{a}_{ij} W_j}{\bar{a}_{ii}} - \frac{\sum_{j \neq i} \bar{a}_{ij} W_j}{\bar{a}_{ii}}$$

and hence by an easy calculation the inequality

$$(4.9) \quad |W_i| \leq \frac{h^2}{2} |\Delta_h W_i| + \frac{3}{4} \max_j |W_j|.$$

The order of the discretization error is now given by the following theorem.

THEOREM 4.1. *Let $u \in C^6(\bar{R})$ and V be the solutions of (3.1) and (4.4) respectively. Then the discretization error $\epsilon \equiv u - V$ is $O(h^4)$.*

Proof. Substituting ϵ into (4.6) and using (4.9) we arrive at the inequality

$$(4.10) \quad \begin{aligned} |\epsilon(p)| &\leq \left[h^2 \sum_{q \in R_h} g(p, q) \right] \max_{t \in R_h} |\Delta_h u(t) - \Delta u(t)| \\ &+ h^2 \left[\sum_{q \in C_h^{**}} g(p, q) \right] \max_{t \in C_h^{**}} |\Delta_h u(t) - \Delta u(t)| \\ &+ \left[\sum_{q \in C_h^*} g(p, q) \right] \left\{ \frac{h^2}{2} \max_{t \in C_h^*} |\Delta_h u(t) - \Delta u(t)| + \frac{3}{4} \max_j |\epsilon| \right\}. \end{aligned}$$

Substituting (3.3), (3.5), (4.3), (4.7) and the results of Lemmas 3.2 and 3.3 yields the desired result.

We comment at this point that another finite difference analogue for this problem which involves the $O(h^4)$ operator for the point and its eight nearest neighbors has been proposed by the authors in [3] and shown to have an $O(h^4)$ discretization error. Moreover the reduced matrix in that case is an M -matrix and hence certain theorems can be applied there to show the convergence of various iterative methods, cf. [20] and [22]. We note further, however, that the right hand side of the finite difference equation is more complicated in the problem defined in that paper.

V. Dirichlet Problem for Laplace's Equation in a Rectangle, $\epsilon = O(h^9)$. Another interesting application of this theory is the formulation of very high order approximations to the solution of the Dirichlet problem for Poisson's equation in a rectangle.

Since the nine point "box" approximation to Δ gives rise to a positive type matrix and since the local truncation error in this case is $O(h^6)$, cf. Kantorovich and Kryloff [13, p. 190] we can apply the technique of Gershgorin and show that the discretization error is $O(h^6)$, as has been pointed out in various places [21], [23]. We shall formulate an $O(h^9)$ finite difference analogue which is of nonpositive type and apply Theorem 2.7 to show that the resulting matrix is monotone.

Again we consider the rectangle in Figure 1 with the sets R_h , C_h^{**} , and C_h as described there. We consider the problem

$$(5.1) \quad \begin{aligned} \Delta u &= 0 \quad \text{in } R, \\ u &= g \quad \text{on } C \end{aligned}$$

where u was chosen to be harmonic only as a matter of convenience. Let 0 be the

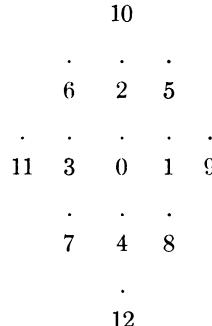


FIGURE 7

point in Figure 7 and we define Δ_h to be the usual nine point "box" operator there, i.e.

$$(5.2) \quad \Delta_h W_0 \equiv \frac{1}{6h^2} \left\{ 4 \sum_{i=1}^4 W_i + \sum_{i=5}^8 W_i - 20 W_0 \right\}.$$

It is well known, cf. Kantorovich and Kryloff [13, p. 190], that for the harmonic function $u(x, y)$

$$(5.3) \quad \Delta_h u - \Delta u \equiv \Delta_h u = \frac{40}{3(8!)} \frac{\partial^8 u}{\partial x^4 \partial y^4} + O(h^{10}).$$

The h^8 term contains the factor $(\partial^{10} u / \partial x^{10} + \partial^{10} u / \partial y^{10})$ which is zero for harmonic functions. If (x, y) is a point of C_h^{**} e.g. on the bottom, (x, h) , and $u \in C^9(\bar{R})$ then

$$(5.4) \quad \frac{\partial^8 u}{\partial x^4 \partial y^4}(x, h) = \frac{\partial^8 u}{\partial x^8}(x, h) = \frac{\partial^8 u}{\partial x^8}(x, 0) + h \frac{\partial^9 u}{\partial x^9}(x, \eta).$$

Hence at such a point we pose the following finite difference analogue whose local truncation error is $O(h^7)$

$$(5.5) \quad \Delta_h V(x, h) = \frac{40h^6}{3(8!)} \frac{\partial^8 g}{\partial x^8}(x, 0).$$

The corresponding difference equation is prescribed at the remaining points of C_h^{**} .

On the other hand we can define a thirteen point difference operator at points of R_h whose local truncation error is $O(h^{10})$ in the following manner. Define the operator Δ_h^* at the point 0 of Figure 7 to be

$$(5.6) \quad \Delta_h^* W_0 \equiv \frac{1}{12h^2} \left\{ 4 \sum_{i=5}^8 W_i + \sum_{i=9}^{12} W_i - 20 W_0 \right\}.$$

If (ξ, η) represent the rotated coordinate system

$$(5.7) \quad \begin{aligned} x &= \frac{1}{\sqrt{2}} \xi - \frac{1}{\sqrt{2}} \eta, \\ y &= \frac{1}{\sqrt{2}} \xi + \frac{1}{\sqrt{2}} \eta \end{aligned}$$

then the error is given by

$$(5.8) \quad \begin{aligned} \Delta_h^* u - \Delta u &= \frac{40h^6}{3(7!)} \frac{\partial^8 u}{\partial \xi^4 \partial \eta^4} + O(h^{10}) \\ &= \frac{40h^6}{3(7!)} \frac{\partial^8 u}{\partial x^4 \partial y^4} + O(h^{10}). \end{aligned}$$

We now define the operator L_h as the linear combination of Δ_h and Δ_h^* which eliminates the mixed derivative, i.e.

$$(5.9) \quad \begin{aligned} L_h W_0 &\equiv \frac{1}{7} [8\Delta_h W_0 - \Delta_h^* W_0] \\ &= \frac{1}{84 h^2} \left\{ 64 \sum_{i=1}^4 W_i + 12 \sum_{i=5}^8 W_i - \sum_{i=9}^{12} W_i - 300 W_0 \right\}. \end{aligned}$$

From (5.3) and (5.8) we see that

$$(5.10) \quad L_h u - \Delta u = L_h u = O(h^{10}).$$

We pose the following finite difference analogue of (5.1)

$$\begin{aligned} -L_h V(p) &= 0, \quad p \in R_h, \\ -\Delta_h V(p) &= l(g), \quad p \in C_h^{**}, \\ V(p) &= g, \quad p \in C_h, \end{aligned}$$

where $l(g)$ is the appropriate eighth tangential derivative of g . We see from (5.9) that the matrix \bar{A} of the system (5.11) is not of positive type. We shall now show how Theorem 2.7 can be applied to prove that matrix \bar{A} is monotone. As before we normalize \bar{A} through multiplication by a positive diagonal matrix D to yield A . The patterns of A are given in Figure 8.

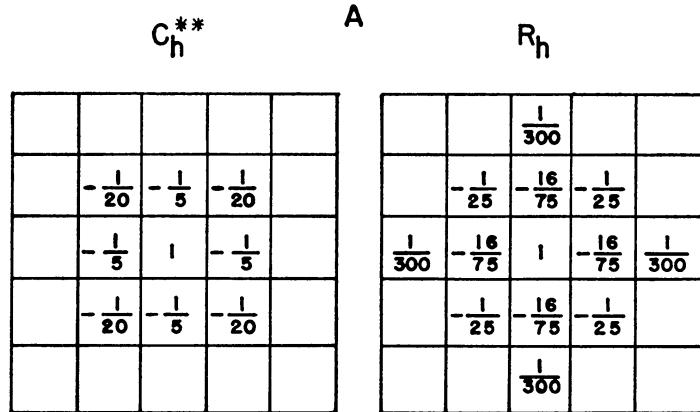


FIGURE 8

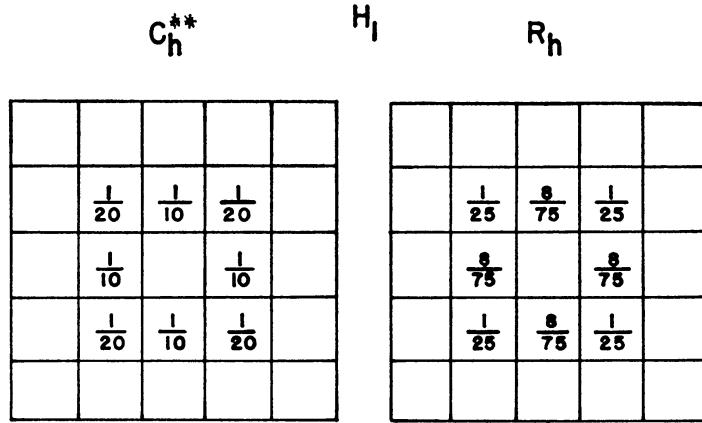


FIGURE 9

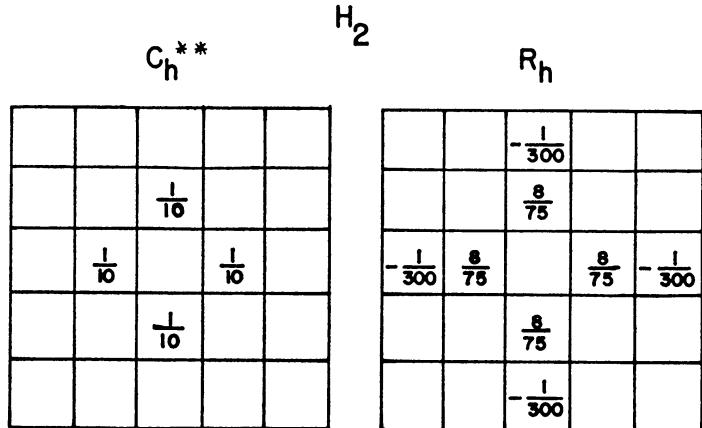


FIGURE 10

We decompose $A = I - H_1 - H_2$ as required in the hypothesis of Theorem 2.7 so that H_1 and H_2 are the matrices corresponding to the patterns in Figures 9 and 10 respectively. By the technique used in Section 3 we can establish the inequalities

$$(5.12) \quad \begin{aligned} (I - H_1)^{-1}H_2 &\geq 0, \\ H_2(I - H_1)^{-1} &\geq 0. \end{aligned}$$

Hence by Theorem 2.7, A and therefore also \bar{A} are monotone.

To obtain estimates of the discretization error we once again define the finite difference Green's function $g(p, q)$ by the equations

$$(5.13) \quad \begin{aligned} -L_{h,p}g(p, q) &= \delta(p, q)h^{-2}, & p \in R_h, \\ -\Delta_{h,p}g(p, q) &= \delta(p, q)h^{-2}, & p \in C_h^{**}, \\ g(p, q) &= \delta(p, q), & p \in C_h. \end{aligned}$$

Since A is monotone we see that $g(p, q)$ exists and is non-negative.

We have the Poisson formula

$$(5.14) \quad \begin{aligned} W(p) &= h^2 \sum_{q \in R_h} g(p, q)[-L_h W(q)] + h^2 \sum_{q \in C_h^{**}} g(p, q)[- \Delta_h W(q)] \\ &\quad + \sum_{q \in C_h} g(p, q)W(q). \end{aligned}$$

The inequality of Lemma 3.2 is again valid as is an inequality of the type given in Lemma 3.3. The proofs in both cases follow in the same manner as those given in Section 3 and hence are not reproduced here.

Finally we have the error estimate:

THEOREM 5.1. *Let $\epsilon = u - V$ where u and V are defined by 5.1 and 5.11 respectively. If $u \in C^{12}(\bar{R})$ then $\epsilon = O(h^9)$.*

The proof follows from (3.26) and the estimates on the local truncation errors derived above.

We note in passing that in extending these results to Poisson's equation the right hand sides of (5.11) will be a function $H(\Delta u)$ which involves Δu and its derivatives through order eight.

VI. Dirichlet Problem for an Elliptic Equation, $\epsilon = O(h^2)$. We consider here the problem

$$(6.1) \quad \begin{aligned} Lu &\equiv -[u_{xx} + u_{xy} + u_{yy}] = f && \text{in } R, \\ u &= g && \text{in } C \end{aligned}$$

where R is the rectangle of Section 3. The results of this section can be extended to more general regions by the same technique used in Section 4. The choice of a particular uniformly elliptic operator L is to a great extent arbitrary. Our choice is used as an illustration in a discussion of the maximum principle in a paper by Diaz and Roberts [9].

Define the sets R_h and C_h^{**} for R as in Section 3. The "seemingly most natural" $O(h^2)$ finite difference analogue to L at a point of $R_h + C_h^{**}$ involves the eight nearest neighbors, cf. [10, p. 190]

$$(6.2) \quad \begin{aligned} -L_h v(x, y) &\equiv h^{-2}\{v(x+h, y) + v(x-h, y) + v(x, y+h) + v(x, y-h) \\ &\quad + \frac{1}{4}[v(x+h, y-h) - v(x-h, y+h) - v(x+h, y-h) \\ &\quad + v(x-h, y-h)] - 4v(x, y)\}. \end{aligned}$$

Clearly

$$(6.3) \quad |L_h u - Lu| = O(h^2).$$

We define the finite difference problem

$$(6.4) \quad \begin{aligned} L_h V(p) &= f(p), & p \in R_h + C_h^{**}, \\ V(p) &= g(p), & p \in C_h. \end{aligned}$$

As was pointed out in [9] the matrix of the linear system (6.4) is not monotone. This is easily seen by considering the square with one interior point (see Figure 11) and the mesh function

.	.	.
1	8	7
.	.	.
2	0	6
.	.	.
3	4	5

FIGURE 11

$$(6.5) \quad V(p) = \begin{cases} -1, & p = 0 \\ 16, & p = 1 \\ 0, & p = 2, \dots, 8. \end{cases}$$

Clearly

$$(6.6) \quad \begin{aligned} L_h V(p) &\geq 0, & p \in R_h + C_h^{**}, \\ V(p) &\geq 0, & p \in C_h, \end{aligned}$$

and yet $V < 0$ at the interior point.

A maximum principle is valid, however, for mesh functions which vanish on C_h . That is

$$(6.7) \quad \begin{aligned} L_h V(p) &\geq 0, & p \in R_h + C_h^{**}, \\ V(p) &= 0, & p \in C_h, \end{aligned}$$

implies that

$$(6.8) \quad V(p) \geq 0, \quad p \in R_h + C_h^{**}.$$

Equivalently it is true that the finite difference Green's function for the reduced problem, obtained by substituting the boundary values and solving the resulting system, is non-negative. This function is given by

$$(6.9) \quad L_{h,p} g(p, q) = \delta(p, q) h^{-2}, \quad p \in R_h + C_h^{**}$$

where L_h at points of C_h^{**} now involves only points of $R_h + C_h^{**}$. Poisson's formula for an arbitrary mesh function V defined on $R_h + C_h^{**}$ is given by

$$(6.10) \quad V(p) = h^2 \sum_{q \in R_h + C_h^{**}} g(p, q) [L_h V(q)].$$

The existence and non-negativity of $g(p, q)$ will now be established using Theorem 2.7.

Let \bar{A} be the matrix of the reduced system. Define $A = D\bar{A}$ where D is the diagonal matrix defined by

$$(6.11) \quad d_{ij} = \delta_{ij} \left(\frac{h^2}{4} \right),$$

so that A has a unit diagonal. We now write A as

$$(6.12) \quad A = I - H_1 - H_2$$

where H_1 and H_2 are matrices corresponding to the patterns in Figure 12 at the point $(\xi, \eta) \in R_h$ and if $(x, y) \in C_h^{**}$, for example, is a typical point on the bottom

$$\begin{array}{ccccc}
 & & H_2 & & \\
 & H_1 & -\frac{1}{16} & \frac{1}{4} & \frac{1}{16} \\
 & . & . & . & . \\
 & \frac{1}{4} & 0 & \frac{1}{4} & 0 & 0 \\
 & . & . & . & . & . \\
 & (\xi, \eta) & & & (\xi, \eta) & \\
 & & \frac{1}{16} & \frac{1}{4} & -\frac{1}{16} \\
 & & . & . & .
 \end{array}$$

FIGURE 12

$$\begin{array}{ccccc}
 & H_1 & & H_2 & \\
 & -\frac{1}{16} & & \frac{1}{4} & \frac{1}{16} \\
 & . & & . & . \\
 & \frac{1}{4} & 0 & \frac{1}{4} & 0 & 0 \\
 & . & . & . & . & . \\
 & (x, y) & & & (x, y) & \\
 & & & & & .
 \end{array}$$

FIGURE 13

we have Figure 13. We note that the various hypotheses of Theorem 2.7 are satisfied. The verification of $(I - H_1)^{-1}H_2 \geq 0$ follows from the same argument used in Section 3. Hence $g(p, q)$ exists and $g(p, q) \geq 0$. A bound on $h^2 \sum_{q \in R_h + C_h^{**}} g(p, q)$ is given by Lemma 3.2.

Let $\epsilon = u - V$ as before. We see that the discretization error satisfies

$$(6.13) \quad \epsilon(p) = h^2 \sum_{q \in R_h + C_h^{**}} g(p, q) [L_h \epsilon(q)]$$

and hence

$$(6.14) \quad |\epsilon| \leq \frac{d^2}{16} \max_{q \in R_h + C_h^{**}} |L_h u(q) - L u(q)| = O(h^2).$$

We note that substituting $V \equiv 1$ in (6.10) gives

$$(6.15) \quad \sum_{q \in C_h^{**}} g(p, q) \leq 1$$

and hence the contribution of the C_h^{**} points to the error are $O(h^4)$ as we would expect. If we were to replace the equations in (6.4) at corner points of C_h^{**} by the equations

$$(6.16) \quad -\Delta_h V(p) = 0$$

where Δ_h is defined by (3.2) then the above argument can be carried through directly for the matrix of the total problem. The local truncation error at these points is $O(1)$ and in the light of the above remark we see that the discretization error is still $O(h^2)$.

If R were a general region then positive type approximations to L at points of C_h^* could be formulated with a local truncation error of $O(h^0)$ or $O(h)$ so that the same argument can be used to establish the monotonicity of the total matrix. Of course we could formulate an $O(h^2)$ finite difference analogue which gives rise to a positive type matrix, cf. Bramble and Hubbard [5] and the references contained therein.

The examples given are meant to illustrate the use of Theorem 2.7 and are in no sense exhaustive. A further interesting class of applications comes from the second and third boundary value problems and will be reported separately in a subsequent paper.

Institute for Fluid Dynamics and Applied Mathematics
 University of Maryland
 College Park, Maryland

1. A. K. AZIZ & B. E. HUBBARD, "Bounds for the solutions of the Sturm-Liouville problem with application to finite difference methods," *J. SIAM*, v. 12, 1964, p. 163-178.
2. J. H. BRAMBLE, "Fourth order finite difference analogues of the Dirichlet problem for Poisson's equation in three and four dimensions," *Math. Comp.*, v. 17, 1963, p. 217-222.
3. J. H. BRAMBLE & B. E. HUBBARD, "On the formulation of finite difference analogues of the Dirichlet problem for Poisson's equation," *Numer. Math.*, v. 4, 1963, p. 313-327.
4. J. H. BRAMBLE & B. E. HUBBARD, "A priori bounds on the discretization error in the numerical solution of the Dirichlet problem," *Contributions to Differential Equations*, v. 2, 1963, p. 229-252.
5. J. H. BRAMBLE & B. E. HUBBARD, "A theorem on error estimation for finite difference analogues of the Dirichlet problem for elliptic equations," *Contributions to Differential Equations*, v. 2, 1963, p. 319-340.
6. J. H. BRAMBLE & B. E. HUBBARD, "On a finite difference analogue of an elliptic boundary problem which is neither diagonally dominant nor of non-negative type." (To appear.)
7. L. COLLATZ, "Bemerkungen zur Fehlerabschätzung für das Differenzenverfahren bei partiellen Differentialgleichungen," *Z. Angew. Math. Mech.*, v. 13, 1933, p. 56-57.
8. L. COLLATZ, *Numerical Treatment of Differential Equations*, 3rd ed., Springer, Berlin, 1960.
9. J. B. DIAZ & R. C. ROBERTS, "On the numerical solution of the Dirichlet problem for Laplace's difference equation," *Quart. Appl. Math.*, v. 9, 1952, p. 355-360.
10. G. FORSYTHE & W. WASOW, *Finite-difference Methods for Partial Differential Equations*, Wiley, New York, 1960.
11. S. GERSCHGORIN, "Fehlerabschätzung für das Differenzenverfahren zur Lösung partieller Differentialgleichungen," *Z. Angew. Math. Mech.*, v. 10, 1930, p. 373-382.
12. E. HOPF, "Elementare Betrachtungen über die Lösungen partieller Differentialgleichungen Zweiter Ordnung vom Elliptischen Typus," *Sitz. Preuss. Akad. Wiss.*, v. 19, 1927, p. 147-152.
13. L. KANTOROVICH & V. KRYLOFF, *Approximate Methods of Higher Analysis*, Noordhoff Ltd., Netherlands, 1958.
14. T. MOTZKIN & W. WASOW, "On the approximation of linear elliptic differential equations by difference equations with positive coefficients," *J. Math. Phys.*, v. 31, 1953, p. 253-259.
15. A. M. OSTROWSKI, "Über die Determinanten mit überwiegender Hauptdiagonale," *Comment. Math. Helv.*, v. 10, 1937, p. 69-96.
16. A. M. OSTROWSKI, "Determinanten mit überwiegender Hauptdiagonale und die absolute Konvergenz von Linearen Iterationsprozessen," *Comment. Math. Helv.*, v. 30, 1955, p. 175-210.
17. A. M. OSTROWSKI, "On some metrical properties of operator matrices and matrices partitioned into blocks," *J. Math. Anal. Appl.*, v. 2, 1961, p. 161-209.
18. H. B. PHILLIPS & N. WIENER, "Nets and Dirichlet problem," *J. Math. Phys.*, v. 2, 1923, p. 105-124.
19. C. PUCCI, *Some Topics in Parabolic and Elliptic Equations*, Institute for Fluid Dynamics and Applied Mathematics, lecture series, 36, Feb.-May, 1958.
20. M. ROCKOFF, "On the numerical solution of finite difference approximations which are not of positive type," *Notices Amer. Math. Soc.*, v. 10, 1963, p. 108.
21. N. UHLMANN, "Differenzenverfahren für die 1. Randwertaufgabe mit Krümmflächigen Rädern bei $\Delta u(x, y, z) = r(x, y, z, u)$," *Z. Angew. Math. Mech.*, v. 38, 1958, p. 130-139.
22. R. VARGA, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, N. J., 1962.
23. E. A. VOLKOV, "Solution of boundary value problems for Poisson's equation in a rectangle," *Dokl. Akad. Nauk SSSR*, v. 147, 1962, p. 13-16 = *Soviet Math. Dokl.*, v. 3, 1962, p. 1524-1527.
24. W. WASOW, "On the truncation error in the solution of Laplace's equation by finite differences," *J. Res. Nat. Bur. Standards*, v. 48, 1952, p. 345-348.

2.2 Bramble-Hilbert Lemma (1970)

Bramble-Hilbert Lemma[10]

ESTIMATION OF LINEAR FUNCTIONALS ON SOBOLEV SPACES WITH APPLICATION TO FOURIER TRANSFORMS AND SPLINE INTERPOLATION*

J. H. BRAMBLE† AND S. R. HILBERT‡

1. Introduction. In this paper some general theorems on estimation for classes of linear functionals on Sobolev spaces are given. These are applied to the study of convergence properties of discrete Fourier transforms in N -dimensional Euclidean space, E^N . In addition, a class of spline functions on uniform meshes in E^N is considered.

Specifically, in § 2, definitions and notation are introduced.

Section 3 is devoted to the estimation of bounded linear functionals on Sobolev spaces. The particular functionals of interest are those which annihilate polynomials of a certain degree (or less). Such functionals are of central importance in the study of errors in approximation and interpolation of functions. Our estimates can often be used to replace standard Taylor series approaches to the estimation of local errors such as in the comparison of difference quotients with derivatives (in E^N) or the estimation of the remainder term in the Taylor series itself. In addition, our results can frequently serve as a substitute for estimates based on an ad hoc use of Peano kernel theorems (c.f. Sard [5, p. 25]). In such estimates, the particular form of the kernel must be utilized whereas for our theorems only properties which are easily verified are required. For example, the use of kernel theorems by Birkhoff, Schultz and Varga [1] in the study of errors in Hermite interpolation could now be avoided by applying our theorems. This would seem to be of particular importance in more than one dimension where the kernel representations are a bit cumbersome.

Section 4 is devoted to the study of the behavior of the difference between the discrete and continuous Fourier transforms in E^N as the mesh size tends to zero. Our approximation theorems are applied to obtain these estimates via certain lemmas which are also employed in § 5.

The last section deals with a class of spline interpolants of order k on Sobolev spaces. We investigate the error in interpolation by such splines as the mesh size tends to zero. We also obtain a connection between the discrete Fourier transform and the N -dimensional analogue of the so-called *cardinal series*. It is shown finally that this series is obtained as a limiting case of splines of order k as $k \rightarrow \infty$. In this connection, Schoenberg [6] has considered this problem in one dimension but for a somewhat more general class of functions and for splines of even order (piecewise polynomials of odd degree). We also want to mention the interesting paper of Golomb [3] in which he uses Fourier methods to study periodic splines on uniform meshes in one dimension.

* Received by the editors May 29, 1969. This research was supported in part by the National Science Foundation under Grant NSF-GP-9467.

† Department of Mathematics, Cornell University, Ithaca, New York 14850.

‡ Department of Mathematics, Ithaca College, Ithaca, New York 14850.

2. Notation and preliminaries. Let R with boundary ∂R be a bounded domain in Euclidean N -space, E^N . Let ρ be the diameter of R . We shall assume that R satisfies a strong cone property; that is, there exists a finite open covering $\{O_i\}$, $i = 1, \dots, n$, of ∂R and corresponding cones $\{C_i\}$ with vertices at the origin such that $x + C_i$ is contained in R for any $x \in R \cap O_i$.

We shall consider complex-valued functions defined on R . As usual we denote by $L_p(R)$ the completion of the space of complex-valued functions defined on R such that

$$\left(\frac{1}{\rho^N} \int_R |f(x)|^p dx \right)^{1/p} = \|f\|_{p,R}$$

is finite. We shall need the following seminorms:

$$(2.1) \quad |u|_{p,k,R} = \sum_{|\alpha|=k} \|D^\alpha u\|_{p,R}$$

and

$$(2.2) \quad |u|_{k,R} = \sum_{|\alpha|=k} |D^\alpha u|_R,$$

where $|u|_R = \sup_{x \in R} |u(x)|$.

In (2.1), (2.2) and the sequel, α is a multi-index;

$$\alpha = (\alpha_1, \dots, \alpha_N) \quad \text{and} \quad |\alpha| = \sum_{i=1}^N \alpha_i, \quad D^\alpha = \left(\frac{\partial}{\partial x_1} \right)^{\alpha_1} \cdots \left(\frac{\partial}{\partial x_N} \right)^{\alpha_N}.$$

Now for $1 \leq p < \infty$ and m a nonnegative integer let $H_p^m(R)$ be the set of all functions in L_p with distributional (weak) derivatives of order j for $0 \leq j \leq m$ in L_p . In this paper we take the norm on $H_p^m(R)$ to be

$$(2.3) \quad \|u\|_{p,m,R}^p = \sum_{k=0}^m \rho^{kp} |u|_{p,k,R}^p.$$

It is trivial that this is equivalent to the usual norm for $H_p^m(R)$.

We shall also consider the space of functions which have continuous derivatives of order up to and including m in R ; this space will be denoted by $C^m(R)$. For the purposes of this paper we take the norm on $C^m(R)$ to be:

$$(2.4) \quad \|u\|_{m,R} = \sum_{k=0}^m \rho^k |u|_{k,R}.$$

Again, the usual norm on $C^m(R)$ is equivalent to (2.4).

We shall denote by P_k the set of polynomials of degree less than or equal to k , restricted to R .

Let h be a (small) positive parameter and define the set of mesh points E_h^N as $E_h^N = \{x|x = (n_1 h, \dots, n_N h), n_j \text{ an integer}, j = 1, \dots, N\}$.

Throughout this paper we shall use C to denote a generic constant not necessarily the same in any two places.

We shall also use Sobolev norms on E^N . As usual these are given by $\|u\|_{H_p^m}^p = \sum_{|\alpha| \leq m} \|D^\alpha u\|_p^p$, where $\|D^\alpha u\|_p = \left(\int_{E^N} |D^\alpha u(x)|^p dx \right)^{1/p}$ and we denote by $\|u\|_0$ the L_2 -norm, $\|u\|_0 = \left(\int_{E^N} |u(x)|^2 dx \right)^{1/2}$. By $\|u\|_m$ we shall mean the Sobolev

norm of $u \in H_2^m(E^N)$. The notation $|u|_m$ will be used to denote the seminorm $\sum_{|\alpha|=m} \|D^\alpha u\|_0$.

3. Estimation of linear functionals. Consider B a Banach space with norm $\|\cdot\|_B$ and let B_1 be a closed linear subspace of B . We define Q to be the quotient or factor space of B with respect to B_1 , denoted by B/B_1 . The elements of Q are equivalence classes $[u]$, where $[u]$ is the class containing u . The equivalence relation is given by \sim where for $u, v \in B$, $u \sim v$ if and only if $u - v \in B_1$. The usual norm on Q is given by $\|[u]\|_Q = \inf_{v \in [u]} \|v\|_B$. It is easy to show that $\|[u]\|_Q = \inf_{v \in B_1} \|u + v\|_B$. Under the assumptions we have made for B and B_1 , it is well known that Q is a Banach space with norm $\|\cdot\|_Q$.

Now consider the (closed) finite-dimensional subspace of $H_p^k(R)$ given by P_{k-1} . Therefore $p(x) \in P_{k-1}$ if and only if $p(x) = \sum_{|\gamma| \leq k-1} a_\gamma x^\gamma$ for $x \in R$, where the a_γ are complex numbers and γ is a multi-index.

THEOREM 1. *Let $Q = H_p^k(R)/P_{k-1}$. Then $|u|_{k,p,R}$ is a norm on Q equivalent to $\|[u]\|_Q$. Further, there exists C independent of ρ and u such that for any $u \in H_p^k(R)$*

$$(3.1) \quad \rho^k |u|_{k,p,R} \leq \|[u]\|_Q \leq C \rho^k |u|_{k,p,R}.$$

Proof. We shall make use of two lemmas which can be found in Morrey [4, p. 85].

LEMMA 1. *For any $u \in H_p^k(R)$ there is a unique polynomial p of degree less than or equal to $k-1$ (or 0) such that $\int_R D^\alpha(u + p) = 0$ for all α with $0 \leq |\alpha| \leq k-1$.*

LEMMA 2. *Let R satisfy a strong cone condition. Then (since R is contained in a sphere of radius ρ) $|u|_{j,p,R} \leq C \rho^{k-j} |u|_{k,p,R}$ for $0 \leq j \leq k-1$ for all $u \in H_p^k(R)$ such that the average over R of each $D^\alpha u$ is 0 for $0 \leq |\alpha| \leq k-1$, where C is a constant independent of ρ and u .*

Note. Morrey assumes that his domain is strongly Lipschitz, but the proof is exactly the same if the domain satisfies a strong cone condition.

We shall now prove the right-hand inequality in Theorem 1. By Lemma 1 we can choose $\bar{p} \in P_{k-1}$ such that $\int_R D^\gamma(u + \bar{p}) = 0$ for $|\gamma| \leq k-1$. Hence using Lemma 2 it follows that $\|u + \bar{p}\|_{k,p,R} \leq C \rho^k |u + \bar{p}|_{k,p,R} = C \rho^k |u|_{k,p,R}$. However, since $\bar{p} \in P_{k-1}$ we have that $\|[u]\|_Q \leq \|u + \bar{p}\|_{k,p,R}$. Hence $\|[u]\|_Q \leq C \rho^k |u|_{k,p,R}$ for $u \in H_p^k(R)$.

The other inequality is easily seen from the observation that $\rho^k |u + p|_{k,p,R} = \rho^k |u|_{k,p,R}$ for any $p \in P_{k-1}$ from which we immediately obtain

$$\rho^k |u|_{k,p,R} \leq \inf_{p \in P_{k-1}} \|u + p\|_{k,p,R} = \|[u]\|_Q.$$

We shall now use this theorem to obtain error estimates for linear functionals. The main result of this section is the following theorem.

THEOREM 2. *Let F be a linear functional on $H_p^k(R)$ which satisfies*

- (i) $|F(u)| \leq C \|u\|_{k,p,R}$ for all $u \in H_p^k(R)$ with C independent of ρ and u
- (ii) $F(p) = 0$ for all $p \in P_{k-1}$.

Then $|F(u)| \leq C_1 \rho^k |u|_{k,p,R}$ for any $u \in H_p^k(R)$ with C_1 independent of ρ and u .

Proof. Since F is linear and satisfies condition (ii),

$$(3.2) \quad |F(u)| = |F(u + p)| \quad \text{for all } p \in P_{k-1}.$$

By condition (i) and (3.2) we have

$$(3.3) \quad |F(u)| \leq C\|u + p\|_{k,p,R}.$$

Taking the infimum over P_{k-1} in (3.3) we have

$$(3.4) \quad |F(u)| \leq C\|[u]\|_Q.$$

The result now follows from Theorem 1.

THEOREM 3. *Let F be a linear functional satisfying*

- (i) $|F(u)| \leq C\|u\|_{j,R}$ for all $u \in C^j(R)$, where C is independent of ρ and u and
- (ii) $F(p) = 0$ for all $p \in P_{k-1}$.

Then $|F(u)| \leq C_1\rho^k|u|_{k,p,R}$ for $p > N/(k-j)$, where C_1 does not depend on ρ or u .

Proof. Since R satisfies a strong cone condition it follows easily from Sobolev's lemma (c.f. [4, p. 78]) that $\|u\|_{j,R} \leq C\|u\|_{k,p,R}$ with C independent of ρ and u provided $p > N/(k-j)$. Clearly F satisfies the hypotheses of Theorem 2.

For the final result of this section we define the usual Lipschitz spaces. Let s be any positive real number with $s = S + \sigma$, $0 < \sigma \leq 1$, S a nonnegative integer. We denote by $C^s(R)$ those elements of $C^0(R)$ such that

$$\sup_{x,y \in R} \sum_{|\alpha|=S} \frac{|D^\alpha u(x) - D^\alpha u(y)|}{|x-y|^\sigma}$$

is bounded.

THEOREM 4. *Let $u \in C^s(R)$ and let F be a linear functional on $C^0(R)$ which satisfies*

- (i) $|F(u)| \leq C|u|_{0,R}$ for all $u \in C^0(R)$ with C independent of ρ and u and
- (ii) $F(q) = 0$ for all $q \in P_{k-1}$.

Then

$$|F(u)| \leq C_1\rho^s \sup_{x,y \in R} \sum_{|\alpha|=S} \frac{|D^\alpha u(x) - D^\alpha u(y)|}{|x-y|^\alpha}, \quad 0 \leq s < k,$$

where C_1 does not depend on ρ or u .

Proof. Since R is bounded, $|u|_{k,p,R} \leq |u|_{k,R}$ for all $p \geq 1$. Hence it follows directly from Theorem 3 that

$$(3.5) \quad |F(u)| \leq C\rho^k|u|_k \quad \text{for any } u \in C^k(R)$$

with C independent of ρ and u . Interpolating between the spaces $C^0(R)$ and $C^k(R)$ we obtain, for $s < k$,

$$(3.6) \quad |F(u)| \leq C\rho^s \left(|u|_0 + \sup_{x,y \in R} \sum_{|\alpha|=S} \frac{|D^\alpha u(x) - D^\alpha u(y)|}{|x-y|^\sigma} \right),$$

where C is independent of ρ and u (c.f. [2] Bramble, Hubbard, Thomée, Lemma 4.1 and 4.2).

Now $F(u) = F(u + q)$ for any $q \in P_S$ since $S \leq k-1$. Choosing $q_0 \in P_S$ such that $D^\alpha(u + q_0)(x_0) = 0$ for some $x_0 \in R$ and all $|\alpha| \leq S$, we may easily obtain

$$|u + q_0|_{0,R} \leq C \sup_{x,y \in R} \sum_{|\alpha|=S} \frac{|D^\alpha u(x) - D^\alpha u(y)|}{|x-y|^\sigma},$$

where C is independent of ρ and u . Hence

$$|F(u)| = |F(u + q_0)| \leq C\rho^s \left(\sup_{x,y \in R} \sum_{|\alpha|=s} \frac{|D^\alpha u(x) - D^\alpha u(y)|}{|x-y|^\sigma} \right),$$

where C is independent of ρ and u .

4. Discrete and continuous Fourier transforms. In this section we shall use the results of the last section to compare the continuous and discrete Fourier transforms.

Let \mathcal{S} be the space of complex-valued infinitely differentiable rapidly decreasing functions on E^N . We remark that $C_0^\infty(R) \subset \mathcal{S}$ for any domain $R \subset E^N$. Now for any function f in \mathcal{S} we define the Fourier transform of f which will be denoted by \hat{f} as $\hat{f}(\xi) = \int_{E^N} f(x) e^{-i\langle \xi, x \rangle} dx$, where $\langle \xi, x \rangle = \sum_{i=1}^N \xi_i x_i$. The Fourier transform is defined for a function in $L_2(E^N)$ or $L_1(E^N)$ by using the density of \mathcal{S} in $L_2(E^N)$ or $L_1(E^N)$. It is well known that the Fourier transform is a one-to-one map of $L_2(E^N)$ onto $L_2(E^N)$ and that the Parseval–Plancheral formulas $\|f\|_0$

$$\begin{aligned} &= (2\pi)^{-N/2} \|\hat{f}\|_0 \quad \text{and} \quad \int_{E^N} f(x) \overline{g(x)} dx = (2\pi)^{-N} \int_{E^N} \hat{f}(x) \overline{\hat{g}(x)} dx \quad \text{hold for any} \\ &g, f \in L_2(E^N). \quad \text{We define an inverse Fourier transform denoted by } \check{f} \text{ as} \\ &\check{f}(x) = (2\pi)^{-N} \int_{E^N} f(\xi) e^{i\langle x, \xi \rangle} d\xi. \quad \text{For any function in } \mathcal{S}, \text{ we know that } \widehat{D^\alpha f}(\xi) \\ &= (i\xi)^\alpha \hat{f}(\xi) \text{ for any multi-index } \alpha, \text{ so we can express } \|f\|_k \text{ as } \int_{E^N} (1 + |\xi|^2)^k |\hat{f}(\xi)|^2 d\xi. \end{aligned}$$

Finally if $f \in L_2(E^N)$ and $g \in L_1(E^N)$ then $(f * g)(x) = \int_{E^N} f(x-y) g(y) dy \in L_2(E^N)$ and $\widehat{f * g}(\xi) = \widehat{f}(\xi) \cdot \widehat{g}(\xi)$ in $L_2(E^N)$.

We can define a discrete Fourier transform for any function which has bounded support and is defined on all the mesh points E_h^N by $\tilde{u}(\theta) = h^N \sum_{x \in E_h^N} u(x) e^{-i\langle x, \theta \rangle}$. We remark that \tilde{u} is a periodic function of period $2\pi/h$. We shall later show that we can define \tilde{u} for any $u \in H_2^m$ for $m > N/2$.

Let χ_h be the characteristic function of the cube S_h where $S_h = \{\xi | \xi \in E^N, |\xi_j| \leq \pi/h \text{ for } j = 1, \dots, N\}$.

Our main aim in this section is to study $\chi_h \tilde{u} - \hat{u}$, as $h \rightarrow 0$. We shall first prove the following lemma.

LEMMA 3. *Let $u \in C_0^\infty(E^N)$. Then there exists a constant C independent of h and u such that for $m > N/2$,*

$$(4.1) \quad \|(\chi_h \tilde{u})^\vee - u\|_j \leq Ch^{m-j} |u|_m$$

for any integer j with $0 \leq j \leq m$.

Remark. Our main theorem in this section is the same as this lemma but with $u \in H_2^m$. However, we shall use this lemma to define \tilde{u} for $u \in H_2^m$.

In order to prove Lemma 3 we shall prove two other lemmas which will also be needed in § 5. We need to introduce first some notation.

Let $I_h(y)$ for any $y \in E_h^N$ be the cube given by $I_h(y) = \{x | x \in E^N, y_j - h/2 < x_j \leq y_j + h/2 \text{ for } j = 1, \dots, N\}$. Define an extension operator by $P_h u(x) = u(y)$

for $x \in I_h(y)$. Let

$$\psi(x) = \begin{cases} 1/h^N & \text{for } x \in I_h(0), \\ 0 & \text{for } x \notin I_h(0) \end{cases}$$

and

$$\psi_j(x_j) = \begin{cases} 1/h & \text{for } -h/2 < x_j \leq h/2, \\ 0 & \text{otherwise} \end{cases}$$

so that $\psi(x) = \prod_{j=1}^N \psi_j(x_j)$ and $\hat{\psi}(\xi) = \prod_{j=1}^N (\sin \xi_j h/2)/(\xi_j h/2)$.

It is easy to see that for any u such that \tilde{u} exists,

$$(4.2) \quad \widehat{P_h u}(\xi) = \hat{\psi}(\xi) \tilde{u}(\xi).$$

For fixed $x \in E^N$ we define a linear functional on C_0^∞ by

$$\begin{aligned} F_m(x; u) &= (\psi^{(m-1)} * P_h u)(x) - (\psi^{(m)} * u)(x) \\ &= h^N \sum_{z \in E_h^N} \psi^{(m)}(x - z) u(z) - \int_{E^N} \psi^{(m)}(x - z) u(z) dz, \end{aligned}$$

where $\psi^{(k)} = \psi * \dots * \psi$, k times.

We have the following lemma.

LEMMA 4. *There exist constants C , C_1 and K independent of x , h and u such that*

$$|F_m(x; u)| \leq C |u|_{I_{Kh}(x)} \leq C_1 \|u\|_{m, 2, I_{Kh}(x)}$$

for $m > N/2$.

Proof. Choose K such that $I_{Kh}(0) = \text{supp } \psi^{(m)}$. Then the first inequality is obvious and the second an immediate application of Sobolev's lemma.

The next lemma is more difficult to prove.

LEMMA 5. *For any polynomial $p \in P_{m-1}$,*

$$F_m(x, p) \equiv 0.$$

Proof. Let J be any set of mesh points which are translates of E_h^N , i.e., $J = \{x|x + a = y, \text{ where } y \in E_h^N \text{ for fixed } a \in E^N\}$. Now define $F_{m,J}(x; u)$ for u continuous by $h^N \sum_{z \in J} \psi^{(m)}(x - z) u(z) - \int_{E^N} \psi^{(m)}(x - z) u(z) dz$ and we shall prove the following proposition which contains our lemma.

PROPOSITION. *For any J and any $x \in E^N$, $F_{m,J}(x, p) = 0$ for all $p \in P_{m-1}$.*

We prove this by induction. It is easy to see that the result is true for $m = 1$ and 2 and if m is an integer greater than two, then $\psi^{(m)}$ belongs to C^{m-2} in E^N . Now assume the lemma is true for m ; we want to show that $F_{m+1,J}(x, p) = 0$ for all $p \in P_m$ and any J or x . Consider $\partial/\partial x_j(F_{m+1,J}(x, p))$ for $j = 1, \dots, N$. (For $m \geq 3$ we know that $F_{m+1,J}$ is at least a continuously differentiable function of x .) Now

$$\begin{aligned} \frac{\partial}{\partial x_j} \psi^{(m+1)}(x) &= \frac{\partial}{\partial x_j} \int_E \psi_j(x_j - y_j) \psi_j^{(m)}(y_j) dy_j \prod_{l \neq j} \psi_l^{(m+1)}(x_l) \\ &= \frac{1}{h} \left[\psi_j^{(m)} \left(x_j + \frac{h}{2} \right) - \psi_j^{(m)} \left(x_j - \frac{h}{2} \right) \right] \prod_{l \neq j} \psi_l^{(m+1)}(x_l). \end{aligned}$$

We define $\partial_j f(x) = [f(x_1, \dots, x_j + h/2, x_{j+1}, \dots, x_N) - f(x_1, \dots, x_j - h/2, x_{j+1}, \dots, x_N)]/h$ and easily obtain

$$\begin{aligned} \frac{\partial}{\partial x_j} F_{m+1,J}(x, p) &= h^N \sum_{y \in J} \prod_{l \neq j} \psi_l^{(m+1)}(x_l - y_l) \partial_j \psi_j^{(m)}(x_j - y_j) p(y) \\ &\quad - \int_{E^N} \prod_{l \neq j} \psi_l^{(m+1)}(x_l - y_l) \partial_j \psi_j^{(m)}(x_j - y_j) p(y) dy. \end{aligned}$$

Now defining a new set of mesh points $\bar{J} = \{z|z_l = y_l, l \neq j, z_j = y_j - h/2, y \in J\}$ we have

$$\begin{aligned} \frac{\partial}{\partial x_j} F_{m+1,J}(x, p) &= h^N \sum_{z \in \bar{J}} \prod_{l \neq j} \psi_l^{(m+1)}(x_l - z_l) \psi_j^{(m)}(x_j - z_j) \partial_j p(z) \\ &\quad - \int_{E^N} \prod_{l \neq j} \psi_l^{(m+1)}(x_l - z_l) \psi_j^{(m)}(x_j - z_j) \partial_j p(z) dz \\ &= \left[\left(\delta_j \prod_{l \neq j} \psi_l \right) * F_{m,\bar{J}}(\cdot, \partial_j p) \right](x), \end{aligned}$$

where δ_j is the one-dimensional Dirac measure with respect to x_j . However, this is zero since $\partial_j p$ is in P_{m-1} if p is in P_m . Since $\partial/(\partial x_j) F_{m+1,J}(x, p) = 0$ for any x, J and $p \in P_m$ and $j = 1, \dots, N$, then $F_{m+1,J}(x, p) = C$, where C does not depend on x . Hence

$$C = h^N \sum_{y \in J} \psi^{(m+1)}(x - y) p(y) - \int_{E^N} \psi^{(m+1)}(x - y) p(y) dy.$$

Using the fact that $F_{m+1,J}$ annihilates polynomials in P_{m-1} we can replace $p(y)$ by $p(y - x)$, and noting that $\psi^{(m+1)}$ is even, by a change of variable we obtain

$$(4.3) \quad C = h^N \sum_{y \in J} \psi^{(m+1)}(x + y) p(x + y) - \int_{E^N} \psi^{(m+1)}(t) p(t) dt.$$

Averaging both sides of (4.3) over $I_h(0)$ we find that

$$\begin{aligned} C &= \frac{1}{h^N} \int_{I_h(0)} (h^N \sum_{y \in J} \psi^{(m+1)}(x + y) p(x + y)) dx - \int_{E^N} \psi^{(m+1)}(x) p(x) dx \\ &= \sum_{y \in J} \int_{I_h(y)} \psi^{(m+1)}(x) p(x) dx - \int_{E^N} \psi^{(m+1)}(x) p(x) dx = 0. \end{aligned}$$

This proves the proposition.

We can now complete the proof of Lemma 3. Consider $j = 0$. Now by Parseval's identity $\|(\chi_h \tilde{u})^\vee - u\|_0 = (2\pi)^{-N/2} \|\chi_h \tilde{u} - \hat{u}\|_0$, and

$$\|\chi_h \tilde{u} - \hat{u}\|_0^2 = \int_{S_h} |\tilde{u} - \hat{u}|^2 d\xi + \int_{|\xi_j| > \pi/h} |\hat{u}(\xi)|^2 d\xi.$$

The second integral is easily estimated since for $\xi \notin S_h$ there is a constant C_m such

that $C_m h^{2m} |\xi|^{2m} \geq 1$ for all $\xi \notin S_h$. Hence

$$(4.4) \quad \begin{aligned} \int_{|\xi| > \pi/h} |\hat{u}(\xi)|^2 d\xi &\leq C_m h^{2m} \int_{|\xi_j| > \pi/h} |\xi|^{2m} |\hat{u}(\xi)|^2 d\xi \\ &\leq C_m h^{2m} |u|_m^2. \end{aligned}$$

Now consider the first integral. Since $\hat{\psi}(\xi) = \prod_{j=1}^N (\sin \xi_j h/2)/(\xi_j h/2)$, there are positive constants C_1 and C_2 independent of h such that $0 < C_1 \leq \hat{\psi}(\xi) \leq C_2$ for any $\xi \in S_h$. Thus in S_h , $\tilde{u}(\xi) = \hat{\psi}^{-1}(\xi) \widehat{P_h u}(\xi)$. We have

$$\begin{aligned} \int_{S_h} |\tilde{u} - \hat{u}|^2 d\xi &= \int_{S_h} |\hat{\psi}^{-1} \widehat{P_h u} - \hat{u}|^2 d\xi = \int_{S_h} |\hat{\psi}^{(m-1)} \widehat{P_h u} - \widehat{\psi^{(m)}} \hat{u}|^2 d\xi \\ &\leq C \int_{S_h} |\hat{\psi}^{(m-1)} \widehat{P_h u} - \widehat{\psi^{(m)}} \hat{u}|^2 d\xi \\ &\leq C \|\hat{\psi}^{(m-1)} \widehat{P_h u} - \widehat{\psi^{(m)}} \hat{u}\|_0^2, \end{aligned}$$

and so, by Parseval's formula we obtain

$$(4.5) \quad \begin{aligned} \int_{S_h} |\tilde{u} - \hat{u}|^2 d\xi &\leq C \|\psi^{(m-1)} * P_h u - \psi^{(m)} * u\|_0^2 \\ &= C \|F_m(\cdot, u)\|_0^2. \end{aligned}$$

Clearly by Lemma 4 we can extend $F_m(x; u)$ to $H_2^m(I_{Kh}(x))$ by continuity. Now by Lemmas 4 and 5, F_m satisfies the hypotheses of Theorem 2 with $R = I_{Kh}(x)$. Hence $|F_m(x; u)| \leq Ch^m |u|_{m,2,I_{Kh}}(x)$, where C is independent of x, h and u . Explicitly

$$|F_m(x; u)|^2 \leq Ch^{2m} \left(1/\text{meas } I_{Kh}(x) \int_{I_{Kh}(x)} \sum_{|\alpha|=m} |D^\alpha u(z)|^2 dz \right).$$

Let

$$\varphi_m(y) = \begin{cases} 1/\text{meas } I_{Kh}(0) & \text{if } y \in I_{Kh}(0), \\ 0 & \text{if } y \notin I_{Kh}(0) \end{cases}$$

so that $\|F_m(\cdot; u)\|_0 \leq Ch^m \left(\int_{E^N} (\varphi_m * \sum_{|\alpha|=m} |D^\alpha u|^2)(x) dx \right)^{1/2}$. However, since

$\int_{E^N} \varphi_m(x) dx = 1$, by interchanging the integration in the convolution with the integration with respect to x we have

$$(4.6) \quad \|F_m(\cdot, u)\|_0 \leq Ch^m |u|_m.$$

Thus (4.4) and (4.6) prove Lemma 3 for $j = 0$. Now, for $0 < j \leq m$, following the same steps as before we are easily led to

$$\|(\chi_h \tilde{u})^\vee - u\|_j \leq C(\|F_m(\cdot, u)\|_j + h^{m-j} |u|_m).$$

The estimate for the first term on the right is obtained by applying Theorem 2 to the functional $G_{m,\alpha}(x; u) = h^{|\alpha|} D^\alpha F_m(x; u)$ for each α with $|\alpha| \leq j$, which clearly satisfies the hypotheses. The proof is completed as before. Thus we have proved Lemma 3.

We now wish to extend the definition of the discrete Fourier transform to any function in H_2^m where $m > N/2$. Define l_2 as the set of all square summable functions defined on E_h^N , with norm $\|\varphi\|_{l_2}^2 = h^N \sum_{x \in E_h^N} |\varphi(x)|^2$. It is easy to see that for any function $\varphi \in C_0^\infty$ we have

$$(4.7) \quad h^N \sum_{x \in E_h^N} |\varphi(x)|^2 = \frac{1}{(2\pi)^N} \int_{S_h} |\varphi(\xi)|^2 d\xi.$$

For any $u \in H_2^m$ there exists a sequence $\{\varphi_j\} \in C_0^\infty$ such that $\varphi_j \rightarrow u$ in H_2^m . Now since

$$\int_{S_h} |\tilde{\varphi}_j - \tilde{\varphi}_k|^2 d\xi \leq \int_{S_h} |\tilde{\varphi}_j - \tilde{\varphi}_k - \hat{\varphi}_j + \hat{\varphi}_k|^2 d\xi + \int_{E^N} |\hat{\varphi}_j - \hat{\varphi}_k|^2 d\xi,$$

using Lemma 1 and the fact that $\{\varphi_j\}$ is a Cauchy sequence in H_2^m , we have $\{\tilde{\varphi}_j\}$ is a Cauchy sequence in $L_2(S_h)$. We define \tilde{u} as the limit in $L_2(S_h)$ of $\tilde{\varphi}_j$ and extend it periodically to all of E^N . It is easy to see that \tilde{u} is independent of the choice of the sequence $\varphi_j \rightarrow u$ in H_2^m , so \tilde{u} is well-defined.

We now have the main result of this section.

THEOREM 5. *Let $u \in H_2^m$. Then there exists a constant C independent of h and u such that for $m > N/2$,*

$$\|(\chi_h \tilde{u})^\vee - u\|_j \leq Ch^{m-j}|u|_m$$

for any integer j with $0 \leq j \leq m$.

Proof. Let $\{\varphi_n\}$ be a sequence such that $\varphi_n \in C_0^\infty$, $n = 1, 2, \dots$, and $\varphi_n \rightarrow u$ in H_2^m as $n \rightarrow \infty$. Now

$$\|(\chi_h \tilde{u})^\vee - u\|_j \leq \|(\chi_h(\tilde{u} - \tilde{\varphi}_n))^\vee\|_j + \|(\chi_h \tilde{\varphi}_n)^\vee - \varphi_n\|_j + \|\varphi_n - u\|_j.$$

By Lemma 3 and the definition of χ_h and $\|\cdot\|_j$ it follows that

$$\|(\chi_h \tilde{u})^\vee - u\|_j \leq Ch^{-j}\|\chi_h(\tilde{u} - \tilde{\varphi}_n)\|_0 + Ch^{m-j}|\varphi_n|_j + \|\varphi_n - u\|_j.$$

By letting $n \rightarrow \infty$ the theorem follows.

We shall also give a version of the Poisson summation formula relating the discrete and continuous Fourier transforms. This will be needed in the next section.

THEOREM 6. *Let $u \in H_2^m$ with $m > N/2$; then*

$$\tilde{u}(\xi) = \sum_{\beta \in E_1^N} \hat{u}(\xi + 2\pi\beta/h) \quad a.e.$$

We shall not give a proof of this formula here since it is essentially a well-known result.

Finally we wish to remark that a proof of Theorem 5 can be based on Theorem 6, but since Lemmas 4 and 5 are required in the next section it seemed preferable to present a self-contained proof based on these lemmas.

5. Splines in E^N . In this section we shall apply the results of § 3 and § 4 to certain types of splines. If v is defined on E_h^N , then a function of the form $h^N \sum_{y \in E_h^N} \psi^{(k)}(x - y)v(y)$ is called a *spline of order k*. It is easy to see that regarded

as a function of x , a spline of order k has continuous partial derivatives of order $k - 2$ and is a piecewise polynomial of degree $k - 1$. Now let u be defined on E_h^N . Then we call $S_k(x; u)$ a *spline interpolant of order k* for u , provided $S_k(x; u) = h^N \sum_{y \in E_h^N} \psi^{(k)}(x - y)v(y)$ and $S_k(x; u) = u(x)$ for all $x \in E_h^N$. We remark that the functions $\psi_j^{(k)}(x - z)$, $j = 1, 2, \dots, N$, are all the so-called *B-splines* of Schoenberg [6].

We have the following existence and uniqueness theorem.

THEOREM 7. *Let $u \in l_2$. Then there exists a unique $v \in l_2$ such that $S_k(x, u) = h^N \sum_{y \in E_h^N} \psi^{(k)}(x - y)v(y)$ is a spline interpolant of order k for u , $k = 1, 2, \dots$.*

Proof. In order to prove this theorem we need the following lemma whose proof will be deferred.

LEMMA 6. *For each $k = 1, 2, \dots$ there exists a constant C_k such that*

$$\tilde{\psi}^{(k)}(\xi) \geq C_k > 0 \quad \text{for all } \xi \in E^N.$$

Let L be an operator defined on l_2 by

$$(Lv)(x) = h^N \sum_{y \in E_h^N} \psi^{(k)}(x - y)v(y), \quad x \in E_h^N.$$

For any $\varphi \in l_2$ let $\{\varphi_n\}$ be a sequence such that $\varphi_n \in l_2$ has bounded support for each n and $\varphi_n \rightarrow \varphi$ as $n \rightarrow \infty$ in l_2 . Now clearly $\widetilde{L\varphi_n} = \widetilde{\psi^{(k)}}\tilde{\varphi}_n$, and by the Parseval identity (4.7) and Lemma 6 it follows that

$$\begin{aligned} \|L\varphi_n\|_{l_2}^2 &= \frac{1}{(2\pi)^N} \int_{S_h} |\widetilde{L\varphi_n}|^2 d\theta = \frac{1}{(2\pi)^N} \int_{S_h} |\widetilde{\psi^{(k)}}|^2 |\tilde{\varphi}_n|^2 d\theta \\ &\geq \frac{C_k^2}{(2\pi)^N} \|\varphi_n\|_{l_2}^2. \end{aligned}$$

The operator $L : l_2 \rightarrow l_2$ is obviously continuous. Hence we obtain

$$\|\varphi\|_{l_2} \leq \frac{C_k^{-1}}{(2\pi)^{-N/2}} \|L\varphi\|_{l_2}$$

for all $\varphi \in l_2$. Denote by $R(L)$ the range of L , regarded as a subspace of l_2 .

Define a functional F on the range of L by $F(L\varphi) = (u, \varphi)$ for all $\varphi \in l_2$. Now $|F(L\varphi)| \leq \|\varphi\|_{l_2} \|u\|_{l_2} \leq C \|L\varphi\|_{l_2} \|u\|_{l_2}$ and hence F is well-defined and bounded on $R(L)$. By the Hahn-Banach theorem there is an extension \bar{F} of F to all of l_2 . Now by the Riesz-Fréchet theorem, there exists a unique $v \in l_2$ such that $\bar{F}(w) = (v, w)$ for all $w \in l_2$. Thus we have, for all $\varphi \in l_2$,

$$(v, L\varphi) = \bar{F}(L\varphi) = F(L\varphi) = (u, \varphi).$$

Clearly L is symmetric so that

$$(Lv, \varphi) = (u, \varphi) \quad \text{for all } \varphi \in l_2,$$

which proves the theorem.

Proof of Lemma 6. It is sufficient to prove the result in one dimension since $\psi^{(k)}(\xi) = \prod_{j=1}^N \psi_j^{(k)}(\xi_j)$. Now by the Poisson summation formula $\widetilde{\psi_j^{(k)}}(\xi_j)$

$= \sum_{l=-\infty}^{\infty} \widehat{\psi_j^{(k)}}(\xi_j + 2\pi l/h)$ a.e. Hence, we have

$$\begin{aligned} \widetilde{\psi_j^{(k)}}(\xi_j) &= \sum_{l=-\infty}^{\infty} \left(\frac{\sin(\xi_j h/2 + \pi l)}{\xi_j h/2 + \pi l} \right)^k \\ &= \left(\frac{\sin \xi_j h/2}{\xi_j h/2} \right)^k + (\sin \xi_j h/2)^k \sum_{l=1}^{\infty} \left\{ (-1)^{lk} \left[\left(\frac{1}{\xi_j h/2 + l\pi} \right)^k \right. \right. \\ &\quad \left. \left. + \left(\frac{-1}{l\pi - \xi_j h/2} \right)^k \right] \right\} \quad \text{a.e.} \end{aligned}$$

Now since $\widetilde{\psi_j^{(k)}}$ has period $2\pi/h$ and is even, we have for any positive integer k , $\widetilde{\psi_j^{(k)}}(\xi_j) \geq [(\sin \xi_j h/2)/(\xi_j h/2)]^k$ a.e. for $0 \leq \xi_j \leq \pi/h$ and therefore we have $\widetilde{\psi_j^{(k)}}(\xi_j) \geq (\cos \xi_0)^k$ a.e. for $0 \leq \xi_j \leq \pi/h$, when $\xi_0 = \tan \xi_0$, $0 < \xi_0 < \pi/2$. Then by the continuity, the periodicity and the evenness of ψ_j^k we obtain

$$\psi^{(k)}(\xi) \geq \prod_{j=1}^N (\cos \xi_0)^k = (\cos \xi_0)^{kN} > 0 \quad \text{for all } \xi \in E^N.$$

We shall define $S_k(x; u)$, $k = 1, 2, \dots$, for $u \in H_2^m$ with $m > N/2$. By Sobolev's lemma for any $u \in H_2^m$ there exists $u_1 \in H_2^m$ such that $u - u_1 = 0$ a.e. and $u_1 \in C^0$. Again by Sobolev's lemma, the restriction of u_1 to E_h^N belongs to l_2 . Hence $S_k(x; u_1)$ exists and is unique. We define $S_k(x; u)$ to be equal to $S_k(x; u_1)$. We remark that it follows from the definition of \tilde{u} that $\tilde{u} = \tilde{u}_1$.

We shall now obtain a representation for $S_k(x; u)$ in terms of u .

THEOREM 8. *Let $u \in H_2^m$ with $m > N/2$. Then*

$$(5.1) \quad S_k(x; u) = ((\widehat{\psi^{(k)}}/\widetilde{\psi^{(k)}})\tilde{u})^\vee.$$

Proof. From the definition of $S_k(x; u)$ for $u \in H_2^m$ and Theorem 7 there exists $v \in l_2$ such that

$$S_k(x; u) = h^N \sum_{y \in E_h^N} \psi^{(k)}(x - y)v(y).$$

Hence

$$\widehat{S}_k = \widehat{\psi^{(k)}}\tilde{v}$$

and

$$\tilde{u} = \tilde{u}_1 = \tilde{S}_k = \widetilde{\psi^{(k)}}\tilde{v}.$$

By Lemma 6 it follows that

$$\widehat{S}_k = \widehat{\psi^{(k)}}/\widetilde{\psi^{(k)}}\tilde{u}.$$

Taking the inverse transform we obtain (5.1). Using this representation we obtain the following error estimate.

THEOREM 9. *Let $u \in H_2^k$, $k > N/2$. Then there exists a constant C independent of h and u such that*

$$\|S_k - u\|_j \leq Ch^{k-j}\|u\|_k, \quad 0 \leq j \leq k.$$

Proof. Now

$$\|S_k - u\|_j^2 \leq C \int_{E^N} (1 + |\theta|^2)^j |\widehat{S}_k - \hat{u}|^2 d\theta.$$

Since we can express $\widehat{S}_k - \hat{u}$ as

$$(\widetilde{\psi}^{(k)})^{-1}[\widehat{\psi}^{(k)}\tilde{u} - \widetilde{\psi}^{(k)}\hat{u}] + (\widetilde{\psi}^{(k)})^{-1}[\widehat{\psi}^{(k)}\hat{u} - \widetilde{\psi}^{(k)}\hat{u}]$$

we have by Lemma 6

$$(5.2) \quad \|S_k - u\|_j \leq C_k^{-1} \left[\|F_k(\cdot, u)\|_j + \left(\int_{E^N} (1 + |\theta|^2)^j |\widehat{\psi}^{(k)} - \widetilde{\psi}^{(k)}|^2 |\hat{u}|^2 d\theta \right)^{1/2} \right].$$

However, we have

$$\begin{aligned} |\widehat{\psi}^{(k)}(\theta) - \widetilde{\psi}^{(k)}(\theta)| &= |h^N \sum_{x \in E_h^N} \psi^{(k)}(x) e^{-i\langle x, \theta \rangle} - \int_{E^N} \psi^{(k)}(x) e^{-i\langle x, \theta \rangle} dx| \\ &= |F_k(0, e^{-i\langle \cdot, \theta \rangle})| \leq Ch^l |\theta|^l \quad \text{for any } 0 \leq l \leq k, \end{aligned}$$

by Theorem 4. Take $l = k - j$. Then it is clear that the second term on the right of (5.2) is bounded by $Ch^{k-j} \|u\|_k$. We have already seen in the proof of Theorem 5 that $\|F_k(\cdot, u)\|_j \leq Ch^{k-j} \|u\|_k$. Thus the proof is complete.

Now in § 4, we showed that $\|(\chi_h \tilde{u})^\vee - u\|_j \leq Ch^{m-j} \|u\|_m$. We shall now show that we can regard $(\chi_h \tilde{u})^\vee$ as a limiting case of $S_k(x; u)$. Consider $(\chi_h \tilde{u})^\vee$ for u in C_0^∞ ; then

$$\chi_h \tilde{u} = \begin{cases} h^N \sum_{x \in E_h^N} u(x) e^{-i\langle x, \theta \rangle}, & \theta \in S_h, \\ 0 & \text{for } |\theta| > \pi/h. \end{cases}$$

Hence

$$\begin{aligned} (5.3) \quad (\chi_h \tilde{u})^\vee(\xi) &= \frac{h^N}{(2\pi)^N} \int_{S_h} \sum u(x) e^{-i\langle x, \theta \rangle} e^{i\langle \theta, \xi \rangle} d\theta \\ &= \frac{h^N}{(2\pi)^N} \sum_{x \in E_h^N} u(x) \int_{S_h} e^{i\langle \theta, \xi - x \rangle} d\theta \\ &= \sum_{x \in E_h^N} u(x) \prod_{j=1}^N \left(\frac{\sin(\xi_j - x_j)\pi/h}{(\xi_j - x_j)\pi/h} \right). \end{aligned}$$

Note that $(\chi_h \tilde{u})^\vee$ is just the N -dimensional cardinal series of Whittaker [7]. We shall denote $(\chi_h \tilde{u})^\vee$ by $S_\infty(x; u)$.

The behavior of S_k as $k \rightarrow \infty$ is studied in the following theorem.

THEOREM 10. *Let $u \in H_2^m$ with $m > N/2$. Then S_k converges uniformly to S_∞ on E^N .*

Proof. By Theorem 8 we have

$$S_k(x; u) - S_\infty(x; u) = \left(\frac{1}{2\pi} \right)^N \int_{E^N} \left(\frac{\widetilde{\psi}^{(k)}}{\widehat{\psi}^{(k)}} - \chi_h \right) \tilde{u} e^{i\langle x, \theta \rangle} d\theta,$$

and hence

$$\begin{aligned} |S_k(x; u) - S_\infty(x; u)| &\leq \left(\frac{1}{2\pi} \right)^N \left[\int_{S_h} |\widehat{\psi}^{(k)} / \widetilde{\psi}^{(k)} - 1| |\tilde{u}| d\theta \right. \\ &\quad \left. + \int_{|\theta_j| > \pi/h} |\widehat{\psi}^{(k)} / \widetilde{\psi}^{(k)}| |\tilde{u}| d\theta \right]. \end{aligned}$$

By an application of the monotone convergence theorem and the Poisson summation formula we can write the first integral as

$$\sum_{\beta \neq 0} \int_{S_h} \frac{\widehat{\psi^{(k)}}(\theta + 2\pi\beta/h)}{\widehat{\psi^{(k)}}\theta} |\tilde{u}(\theta)| d\theta.$$

Using the periodicity of \tilde{u} and $\widehat{\psi}$ and regarding the integral over $|\theta_j| > \pi/h$ as the sum of the integrals over $S_h + 2\beta\pi/h$ for all $\beta \neq 0$, after making a change of variable for each β , we can express the second integral as

$$\sum_{\beta \neq 0} \int_{S_h} \frac{\widehat{\psi^{(k)}}(\theta + 2\beta\pi/h)}{\widehat{\psi^{(k)}}\theta} |\tilde{u}(\theta)| d\theta.$$

Now since $\widetilde{\psi^{(k)}} \geq \widehat{\psi^{(k)}}$ in S_h , and since $|\sin(x + \pi)| = |\sin x|$, we obtain

$$|S_k(x; u) - S_\infty(x; u)| \leq \frac{2}{(2\pi)^N} \sum_{\beta \neq 0} \int_{S_h} \prod_{j=1}^N \left| \frac{\theta_j h}{\theta_j h + 2\pi\beta_j} \right|^k |\tilde{u}(\theta)| d\theta.$$

Using the Cauchy-Schwarz inequality we have

$$|S_k(x; u) - S_\infty(x; u)| \leq \frac{2}{(2\pi)^N} \|\tilde{u}\|_{0, S_h} \sum_{\beta \neq 0} \left(\int_{S_h} \prod_{j=1}^N \left| \frac{\theta_j h}{\theta_j h + 2\pi\beta_j} \right|^{2k} d\theta \right)^{1/2}.$$

By an elementary estimate it follows easily that there is a constant C independent of k such that

$$2 \sum_{\beta \neq 0} \left(\int_{S_h} \prod_{j=1}^N \left| \frac{\theta_j h}{\theta_j h + 2\pi\beta_j} \right|^{2k} d\theta \right)^{1/2} \leq C(2k-1)^{-N/2}.$$

Thus we have $|S_k(x; u) - S_\infty(x; u)| \leq C(2k-1)^{-N/2}$, where C is independent of χ ; by letting $k \rightarrow \infty$ the theorem follows.

REFERENCES

- [1] G. BIRKHOFF, M. SCHULTZ AND R. VARGA, *Piecewise Hermite interpolation in one and two variables with applications to partial differential equations*, Numer Math., 11 (1968), pp. 232-256.
- [2] J. BRAMBLE, B. HUBBARD AND V. THOMÉE, *Convergence estimates for essentially positive type discrete Dirichlet problems*, Math. Comp., to appear.
- [3] M. GOLOMB, *Approximation by periodic spline interpolants on uniform meshes*, J. Approx. Theor., 1 (1968), pp. 26-65.
- [4] C. MORREY, *Multiple Integrals in the Calculus of Variations*, Springer-Verlag, New York, 1966.
- [5] A. SARD, *Linear Approximation*, Math. Surveys, No. 9, American Mathematical Society, Providence, 1963.
- [6] I. J. SCHOENBERG, *On cardinal spline interpolation and a summability method for the cardinal series*, Lecture Notes, Symposium on Approximation, University of Cincinnati, Cincinnati, Ohio, 1969.
- [7] J. M. WHITTAKER, *Interpolatory Function Theory*, Cambridge University Press, Cambridge, 1935.

2.3 Rayleigh-Ritz-Galerkin methods for dirichlet's problem using subspaces without boundary conditions (1970)

Rayleigh-Ritz-Galerkin methods for dirichlet's problem using subspaces without boundary conditions[28]

2.4 Triangular elements in the finite element method (1970)

Triangular elements in the finite element method[31]

Triangular Elements in the Finite Element Method

By James H. Bramble and Miloš Zlámal

Abstract. For a plane polygonal domain Ω and a corresponding (general) triangulation we define classes of functions $p_m(x, y)$ which are polynomials on each triangle and which are in $C^{(m)}(\Omega)$ and also belong to the Sobolev space $W_2^{(m+1)}(\Omega)$. Approximation theoretic properties are proved concerning these functions. These results are then applied to the approximate solution of arbitrary-order elliptic boundary value problems by the Galerkin method. Estimates for the error are given. The case of second-order problems is discussed in conjunction with special choices of approximating polynomials.

1. Introduction. The classical Ritz and Galerkin method has several advantages over the finite-difference method. Nevertheless two things have prevented its more extensive use: 1. The practical construction of the basic functions in more dimensions was possible only for some simple domains. 2. Even for these domains the procedures can be highly unstable.

The finite element method is nothing else than the Ritz or Galerkin method using special trial functions. The first idea goes back to Courant [9] who suggested triangulating the given domain and using functions which are linear on each triangle as trial functions for solving boundary value problems of the second order. This idea was rediscovered by the engineers and developed, originally as a concept of structural analysis, into a method called the finite element method (see Turner, Clough, Martin and Topp [15] and the references in Zienkiewicz [18]). Practical experience, the large amount of numerical results and the first theoretical results show that the finite element method removes the above mentioned shortcomings of the classical Ritz and Galerkin method.

One feature of the procedures described by the engineers consists in introducing higher degree polynomials for interpolation of the solution on the given element. Some procedures of this kind for triangular elements were proposed and justified by the second of the authors [19]. For fourth-order equations the trial functions used are polynomials of the fifth degree.¹ The results and the method of [19] were generalized by Ženíšek [17]. He proposed to use polynomials of the degree $4m + 1$ introduced later in this paper² and he justified the method for $m = 2, 3$ (the case $m = 1$ being justified in [19]).

The method of this paper differs completely from the method of [19]. A lemma about linear functionals on $W_p^{(k)}$ by Bramble and Hilbert [7] allows us to get general results for any m . We prove a general interpolation theorem and apply it to V -elliptic

Received January 30, 1970.

AMS 1969 subject classifications. Primary 6565; Secondary 6520.

Key words and phrases. Finite element method, Ritz method, Galerkin method, piecewise polynomial subspaces, approximation of solution, elliptic boundary problems.

¹ Almost simultaneously this procedure was described and applied to bending of plates by Bell [3] and [4], Visser [16], Bosshard [6] and Argyris, Fried, Scharpf [2].

² As a matter of fact, he also introduces polynomials of the degree $4m + 2$, $4m + 3$ and $4(m + 1)$. We restrict ourselves to the case of polynomials of the degree $4m + 1$. The others are easy to deal with in the same way.

boundary value problems of arbitrary order. The seminorm used in this paper for the discretization error is more appropriate than that used in [19].

2. Interpolation Polynomials on Triangles. To define the interpolation polynomials introduced by Ženíšek [17] we denote by P_j ($j = 1, 2, 3$) the vertices of a triangle T ,³ by (x_i, y_i) the coordinates of P_j , by P_0 the center of gravity of T , by l_i the sides of T , by ν_i the normals to l_i . We divide every side l_i in $r + 1$ equal parts ($r = 1, 2, \dots$) by the points $Q_i^{(\rho, r)}$ ($j = 1, 2, 3, \rho = 1, \dots, r$).

Now a polynomial $p_m(x, y)$ in two variables of the degree $4m + 1$ ($m = 0, 1, \dots$) has $(2m + 1)(4m + 3)$ coefficients. Hence we cannot prescribe more than $(2m + 1) \cdot (4m + 3)$ conditions for such a polynomial. Let us prescribe the following values:

$$(1) \quad D^i p_m(P_j),^4 \quad j = 1, 2, 3, \quad |i| \leq 2m,$$

$$(2) \quad \frac{\partial^r p_m(Q_i^{(\rho, r)})}{\partial \nu_i^r}, \quad j = 1, 2, 3, \quad \rho = 1, \dots, r, \quad r = 1, \dots, m,$$

$$(3) \quad D^i p_m(P_0), \quad |i| \leq m - 2.$$

We must add that we leave out the values (2) and (3) if $m = 0$ and $m = 0, 1$, respectively. Thus, $p_0(x, y)$ is a linear polynomial determined by the values of $u(x, y)$ at the vertices of T and $p_1(x, y)$ is the polynomial introduced in [19, p. 404] and in the papers quoted in footnote 1.

The importance of the polynomials $p_m(x, y)$ follows from the property proved in [17] which we formulate in this way: Suppose the values of the form (1), (2), (3) determine uniquely a polynomial $p_m(x, y)$ of the degree not greater than $4m + 1$. Let Ω be a polygonal domain triangulated by triangles $\{T_k\}_{k=1}^M$ and let values of the form (1), (2), (3) be prescribed at every vertex of the triangulation, at every point $Q_i^{(\rho, r)}$ and at every center of gravity. Then the function $v(x, y)$ which on every T_k is equal to a polynomial $p_m^k(x, y)$ defined in the way just described belongs to $C^{(m)}(\bar{\Omega})$. Later we shall construct trial functions for the Galerkin method by means of the polynomials $p_m(x, y)$. First, we must, of course, prove the existence and uniqueness of $p_m(x, y)$.

THEOREM 1. *There exists exactly one polynomial $p_m(x, y)$ of the degree not greater than $4m + 1$ assuming the values (1), (2), (3).*

Proof. The assertion is trivial for $m = 0$, hence we consider $m \geq 1$. It is sufficient to prove that if

$$(4) \quad D^i p_m(P_j) = 0, \quad j = 1, 2, 3, \quad |i| \leq 2m,$$

$$(5) \quad \frac{\partial^r p_m(Q_i^{(\rho, r)})}{\partial \nu_i^r} = 0, \quad j = 1, 2, 3, \quad \rho = 1, \dots, r, \quad r = 1, \dots, m,$$

$$(6) \quad D^i p_m(P_0) = 0, \quad |i| \leq m - 2,$$

and $p_m(x, y)$ is a polynomial of a degree not greater than $4m + 1$ then $p_m(x, y) \equiv 0$. (That is, the linearity of (1), (2) and (3) permits the uniqueness of their solution to imply existence of a solution.)

³ At the same time T means the interior of T ; it will always be clear what meaning of T is necessary to be taken.

⁴ Here $i = (i_1, i_2)$, $|i| = i_1 + i_2$, $D^i u = \partial^{i_1 i_2} u / \partial x^{i_1} \partial y^{i_2}$.

The derivatives $\partial^r p_m / \partial \nu_j^r$ ($r = 0, \dots, m$, $j = 1, 2, 3$) are Hermite polynomials (see, for instance, [5]) in one variable on the corresponding sides of the triangle T which, with respect to (4) and (5), assume homogeneous boundary values. Therefore they are identically equal to zero on the sides of T . Using the reasoning of the proof of Theorem 1 in [17] we find out that

$$(7) \quad D^i p_m(x, y) |_{\delta T} = 0, \quad |i| \leq m.$$

Now let us consider the transformation

$$(8) \quad \begin{aligned} x &= x(\xi, \eta) \equiv x_1 + (x_2 - x_1)\xi + (x_3 - x_1)\eta, \\ y &= y(\xi, \eta) \equiv y_1 + (y_2 - y_1)\xi + (y_3 - y_1)\eta \end{aligned}$$

and the polynomial $\tilde{p}_m(\xi, \eta) = p_m[x(\xi, \eta), y(\xi, \eta)]$. The equations (8) map T onto the triangle T_1 with vertices $\tilde{P}_1(0, 0)$, $\tilde{P}_2(1, 0)$, $\tilde{P}_3(0, 1)$. The points $Q_i^{(\rho, r)}$ are mapped on the points $\tilde{Q}_i^{(\rho, r)}$ which again divide the new sides \tilde{l}_i into $r+1$ equal parts and P_0 is mapped on the center of gravity $\tilde{P}_0(\frac{1}{3}, \frac{1}{3})$ of the triangle T_1 . From (7) and (6) it follows that

$$(9) \quad D^i \tilde{p}_m(\xi, \eta) |_{\delta T_1} = 0, \quad |i| \leq m,$$

$$(10) \quad D^i \tilde{p}(P_0) = 0, \quad |i| \leq m-2,$$

(if we use the symbol D applied to functions of ξ and η we always mean a derivative with respect to ξ and η ; thus $D^i \tilde{p}_m(\xi, \eta) = \partial^{i_1} \tilde{p}_m(\xi, \eta) / (\partial \xi^{i_1} \partial \eta^{i_2})$). A consequence of (9) is that $\partial^r \tilde{p}_m(\xi, 0) / \partial \eta^r = 0$ for $0 \leq \xi \leq 1$, $r = 0, \dots, m$. Therefore $\tilde{p}_m(\xi, \eta)$ is divisible by η^{m+1} . Similarly, one can show that $\tilde{p}_m(\xi, \eta)$ is divisible by $(1 - \xi - \eta)^{m+1}$ and by ξ^{m+1} . Hence, if $m = 1$ it must be that $\tilde{p}_1(\xi, \eta) \equiv 0$, and consequently $p_1(x, y) \equiv 0$, and if $m \geq 2$

$$\tilde{p}_m(\xi, \eta) = [\xi \eta (1 - \xi - \eta)]^{m+1} Q(\xi, \eta),$$

where $Q(\xi, \eta)$ is a polynomial of the degree not greater than $m-2$. Now it is sufficient to use (10). Since $[\xi \eta (1 - \xi - \eta)]_{\xi=\eta=1/3} \neq 0$ we get

$$D^i Q(\tilde{P}_0) = 0, \quad |i| \leq m-2,$$

and since $Q(\xi, \eta)$ is a polynomial of the degree not greater than $m-2$ it follows that $Q(\xi, \eta) \equiv 0$, hence $\tilde{p}_m(\xi, \eta) \equiv 0$ and $p_m(x, y) \equiv 0$.

Next what we need is some estimate of the error arising when we approximate a function $u(x, y) \in C^{(2m)}(\bar{T})$ by a polynomial $p_m(x, y)$. We will say that $p_m(x, y)$ is the interpolation polynomial corresponding to $u(x, y)$ if

$$(11) \quad D^i p_m(P_j) = D^i u(P_j), \quad j = 1, 2, 3, \quad |i| \leq 2m,$$

$$(12) \quad \partial^r p_m(Q_i^{(\rho, r)}) / \partial \nu_i^r = \partial^r u(Q_i^{(\rho, r)}) / \partial \nu_i^r, \quad j = 1, 2, 3, \quad \rho = 1, \dots, r, \quad r = 1, \dots, m,$$

$$(13) \quad D^i p_m(P_0) = D^i u(P_0), \quad |i| \leq m-2.$$

To get the estimate we make use of a lemma by Bramble and Hilbert [7]. First we introduce some notation. By $W_2^{(k)}(\Omega)$ we denote the Hilbert space of all functions which together with their generalized derivatives up to the k th order belong to $L_2(\Omega)$.

The norm is given by

$$\|u\|_{k,\Omega}^2 = \sum_{i=1}^k \|u\|_{i,\Omega}^2, \quad \text{where} \quad \|u\|_{i,\Omega}^2 = \sum_{|\alpha|=i} \|D^\alpha u\|_{L_2(\Omega)}^2.$$

LEMMA.⁵ Let Ω be a bounded domain in E_N with $\text{diam } (\Omega) = 1$. Assume that Ω satisfies the ordinary cone condition (see [1]). Let $F(u)$ be a bounded linear functional on $W_2^{(k)}(\Omega)$,

$$|F(u)| \leq C_1 \|u\|_{k,\Omega},$$

and let $F(q) = 0$ for every polynomial q of the degree less than k . Then there exists a constant C_2 depending on the cone condition only such that

$$(14) \quad |F(u)| \leq C_1 C_2 \|u\|_{k,\Omega}$$

for all $u \in W_2^{(k)}(\Omega)$.

THEOREM 2. Let $u(x, y) \in W_2^{(k)}(T)$ where $2m + 2 \leq k \leq 4m + 2$. Let $p_m(x, y)$ be the interpolation polynomial corresponding to $u(x, y)$. Then, for $0 \leq n \leq k$,

$$(15) \quad \|u - p_m\|_{n,T} \leq \frac{K}{(\sin \alpha)^{m+n}} c^{k-n} \|u\|_{k,T},$$

where the constant K does not depend on the triangle T and the function u and where α is the smallest angle and c is the length of the greatest side of T .

Proof. We denote by $\alpha \leq \beta \leq \gamma$ the angles of the triangle T and we choose the notation of the vertices such that α lies at P_1 , β at P_2 and γ at P_3 . The lengths of the sides are denoted by a, b, c , a being the smallest and c the greatest. As $a + b > c$ we have $b > \frac{1}{2}c$. The area of T is equal to one half of $|J|$ where J is the Jacobian of the transformation (8) so that

$$|J|^{-1} = \frac{1}{bc \sin \alpha} < \frac{2}{c^2 \sin \alpha}.$$

For the inverse transformation to (8) we easily find out that

$$(16) \quad \left| \frac{\partial \xi}{\partial x} \right|, \left| \frac{\partial \xi}{\partial y} \right|, \left| \frac{\partial \eta}{\partial x} \right|, \left| \frac{\partial \eta}{\partial y} \right| \leq \frac{2}{c \sin \alpha}.$$

Let us denote $w(x, y) = u(x, y) - p_m(x, y)$ and consider the function $\tilde{w}(\xi, \eta) = w[x(\xi, \eta), y(\xi, \eta)]$. The derivatives $D^\alpha \tilde{w}(\xi, \eta)$ are linear combinations of the derivatives $D^\alpha w(x, y)$ and using (16) we easily obtain

$$(17) \quad \|w\|_{n,T} \leq \frac{K_1}{(c \sin \alpha)^n} |J|^{1/2} \|\tilde{w}\|_{n,T_1}.$$

Here K_1 is a constant which does not depend on T and the functions considered (in the sequel we shall denote such constants by K_1, K_2, \dots).

Now to get an estimate for $\|\tilde{w}\|_{n,T_1}$ we apply the Lemma. Let us consider the linear functional $F(\tilde{u}) = (\tilde{u} - \tilde{p}_m, v)_{n,T_1}$ on $W_2^{(k)}(T_1)$ where $(\tilde{w}, v)_{n,T_1}$ means the scalar product in $W_2^{(n)}(T_1)$ and v is an arbitrary function from $W_2^{(n)}(T_1)$. If $\tilde{u}(\xi, \eta)$ is a polynomial of the degree less than k then $u(x, y)$ is also a polynomial of the degree less

⁵ Actually, it is true for more general spaces $W_p^{(k)}(\Omega)$ and the formulation introduced here differs a little from the formulation introduced in [7].

than k . For $k \leq 4m + 2$ it follows by Theorem 1 that $u(x, y) - p_m(x, y) \equiv 0$, hence $\tilde{u}(\xi, \eta) - \tilde{p}_m(\xi, \eta) \equiv 0$ and $F(\tilde{u}) = 0$. Further,

$$|F(\tilde{u})| \leq \|v\|_{n, T_1} \|\tilde{u} - \tilde{p}_m\|_{n, T_1} \leq \|v\|_{n, T_1} \{ \|\tilde{u}\|_{k, T_1} + \|\tilde{p}_m\|_{k, T_1} \}.$$

Assume we succeed in proving

$$(18) \quad \|\tilde{p}_m\|_{k, T_1} \leq \frac{K_2}{(\sin \alpha)^m} \|\tilde{u}\|_{k, T_1}.$$

Then

$$|F(\tilde{u})| \leq \frac{K_3}{(\sin \alpha)^m} \|v\|_{n, T_1} \|\tilde{u}\|_{k, T_1}$$

and applying the Lemma (actually in our case $\text{diam } (\Omega) = \text{diam } (T_1) = \sqrt{2}$; however obviously (14) is also true with C_2 being an absolute constant) we have

$$|F(\tilde{u})| \leq \frac{K_4}{(\sin \alpha)^m} \|v\|_{n, T_1} |\tilde{u}|_{k, T_1}.$$

Choosing $v = \tilde{u} - \tilde{p}_m$ we get

$$\|\tilde{u} - \tilde{p}_m\|_{n, T_1} \leq \frac{K_4}{(\sin \alpha)^m} |\tilde{u}|_{k, T_1}.$$

From (17) it follows

$$\|u - p_m\|_{n, T} \leq \frac{K_5}{(\sin \alpha)^{m+n}} c^{-n} |J|^{1/2} |\tilde{u}|_{k, T},$$

and since

$$|\tilde{u}|_{k, T} \leq K_6 c^k |J|^{-1/2} |u|_{k, T}$$

the final result is the estimate (15).

To prove (18) we remark that the polynomial $\tilde{p}_m(\xi, \eta)$ is, according to Theorem 1, uniquely determined by the values

$$\begin{aligned} D^i \tilde{p}_m(\tilde{P}_i), & \quad j = 1, 2, 3, \quad |i| \leq 2m, \\ \partial^r \tilde{p}_m(\tilde{Q}_i^{(\rho, r)}) / \partial \tilde{r}_i^r, & \quad j = 1, 2, 3, \quad \rho = 1, \dots, r, \quad r = 1, \dots, m, \\ D^i(\tilde{P}_0), & \quad |i| \leq m - 2. \end{aligned}$$

If we order these values in some way and denote by a_i ($j = 1, \dots, N_0 = (2m + 1) \cdot (4m + 3)$), it obviously holds that $\tilde{p}_m(\xi, \eta) = \sum_{i=1}^{N_0} a_i r_i(\xi, \eta)$, where $r_i(\xi, \eta)$ are polynomials such that from the above-mentioned N_0 values one of their values is equal to 1 and the others are zero. Hence, the polynomials $r_i(\xi, \eta)$ as well as their derivatives of an arbitrary order are bounded by absolute constants and it is sufficient to prove that

$$(19) \quad |a_i| \leq \frac{K_7}{(\sin \alpha)^m} \|\tilde{u}\|_{k, T_1}.$$

Now from (11) and (13) it follows immediately

$$(20) \quad \begin{aligned} D^i \tilde{p}_m(\tilde{P}_i) &= D^i \tilde{u}(\tilde{P}_i), \quad j = 1, 2, 3, \quad |i| \leq 2m, \\ D^i \tilde{p}_m(\tilde{P}_0) &= D^i \tilde{u}(\tilde{P}_0), \quad |i| \leq m - 2, \end{aligned}$$

so that for these values we get (19) by means of the Sobolev lemma (actually, (19) is true even without the factor $1/(\sin \alpha)^m$).

To prove (19) for the remaining values we first notice the following formula. Let the direction l make an angle ϕ with the positive ξ -axis and let ν be the direction perpendicular to l . Let τ be a direction making an angle ω with the positive ξ -axis and let $\partial f(P)/\partial \tau = 0$. Then

$$(21) \quad \partial f(P)/\partial \nu = -\sigma \partial f(P)/\partial l, \quad \sigma = \cot g(\omega - \phi)$$

(the Eq. (21) follows from the formula $\partial f/\partial \tau = \cos(\omega - \phi) \partial f/\partial l + \sin(\omega - \phi) \partial f/\partial \nu$). Further we notice that the condition (12) is just the condition

$$\partial^r \tilde{p}_m(\tilde{Q}_i^{(\rho, r)})/\partial \tau_i^r = \partial^r \tilde{u}(\tilde{Q}_i^{(\rho, r)})/\partial \tau_i^r,$$

where τ_i ($i = 1, 2, 3$) are certain directions which are easy to find. Apply now (21) to $f = \tilde{p}_m - \tilde{u}$ and to $l = l_i$, $\tau = \tau_i$. By elementary calculations, which we leave out, we get

$$(22) \quad \frac{\partial \tilde{p}_m(\tilde{Q}_i^{(1,1)})}{\partial \tilde{\nu}_i} = \frac{\partial \tilde{u}(\tilde{Q}_i^{(1,1)})}{\partial \tilde{\nu}_i} + \sigma_1 \frac{\partial \tilde{u}(\tilde{Q}_i^{(1,1)})}{\partial l_i} - \sigma_2 \frac{\partial \tilde{p}_m(\tilde{Q}_i^{(1,1)})}{\partial l_i},$$

where

$$|\sigma_1| = \frac{|(x_2 - x_1)(x_3 - x_1) + (y_2 - y_1)(y_3 - y_1)|}{c^2}, \quad |\sigma_2| = \frac{c^2 - b^2}{a^2},$$

$$|\sigma_3| = \frac{|(x_2 - x_1)(x_3 - x_1) + (y_2 - y_1)(y_3 - y_1)|}{b^2}.$$

σ_1 and σ_3 are bounded by absolute constants:

$$|\sigma_1| \leq bc/c^2 \leq 1, \quad |\sigma_3| \leq bc/b^2 < 2,$$

whereas σ_2 is not bounded by an absolute constant.⁶ However,

$$|\sigma_2| = (c - b)(c + b)/a^2 < 2ac/a^2 \leq 2/\sin \alpha.$$

Since \tilde{p}_m is a Hermite interpolation polynomial in one variable on the sides of T_1 which is determined by the values (20) it follows from (22) by means of the Sobolev lemma that

$$\left| \frac{\partial \tilde{p}_m(\tilde{Q}_i^{(1,1)})}{\partial \tilde{\nu}_i} \right| \leq \frac{K_s}{\sin \alpha} \|\tilde{u}\|_{k, T_1}.$$

In general, we get

$$\left| \frac{\partial^r \tilde{p}_m(\tilde{Q}_i^{(\rho, r)})}{\partial \tilde{\nu}_i^r} \right| \leq \frac{K_s}{(\sin \alpha)^r} \|\tilde{u}\|_{k, T_1}, \quad r = 1, \dots, m,$$

if we proceed by induction and use the formula

$$\frac{\partial^r f(P)}{\partial \nu^r} = - \sum_{i=1}^r \binom{r}{j} \sigma^i \frac{\partial^r f(P)}{\partial l^i \partial \nu^{r-i}}$$

which holds if $\partial^r f(P)/\partial \tau^r = 0$.

⁶ There are triangles for which $(c^2 - b^2)/a^2 \geq 1/2\sin \alpha$.

Remark. In a similar way one can prove

$$\max_{\overline{\tau}} |D^i(u - p_m)| \leq \frac{K}{(\sin \alpha)^{m+|i|}} |J|^{-1/2} c^{k-|i|} |u|_{k,\tau}$$

if

$$2m + 2 \leq k \leq 4m + 2, \quad |i| \leq k - 2.$$

3. Application to V -Elliptic Boundary Value Problems. Let Ω be a bounded simply or multiply connected domain in E_2 with a boundary Γ consisting of a finite number of polygons Γ_j ($j = 0, 1, \dots, s$); $\Gamma_1, \dots, \Gamma_s$ lie inside of Γ_0 and do not intersect. This assumption enables one to triangulate Ω . Let V be a Hilbert space such that

$$\overset{\circ}{W}_2^{(n)}(\Omega) \subset V \subset W_2^{(n)}(\Omega),$$

with the norm induced by $W_2^{(n)}(\Omega)$. Here, $\overset{\circ}{W}_2^{(n)}(\Omega)$ is the completion with respect to the norm $\|\cdot\|_n$ ⁷ of functions from $C^{(\infty)}(\Omega)$ with compact support in Ω . Let $a(u, v)$ be a bilinear form continuous on $V \times V$ and V -elliptic, i.e., a mapping $(u, v) \rightarrow a(u, v)$ from $V \times V$ into the field of complex numbers which is linear in u , antilinear in v and satisfies the conditions of boundedness and coerciveness

$$(23) \quad |a(u, v)| \leq M \|u\|_n \|v\|_n, \quad \forall u, v \in V, \quad M = \text{const} > 0,$$

$$(24) \quad \operatorname{Re} a(v, v) \geq \alpha \|v\|_n^2, \quad \forall v \in V, \quad \alpha = \text{const} > 0.$$

Finally, let $L(v)$ be an antilinear functional continuous on V . Under these conditions there exists just one $u \in V$ such that

$$(25) \quad a(u, v) = L(v), \quad \forall v \in V,$$

(see Lions and Magenes [13]).

We shall approximate the problem (25) by the Galerkin method (see Céa [8]) using the following finite-dimensional subspaces V_h^m of V . We triangulate Ω , i.e., we cover Ω by a finite number of arbitrary triangles such that any two triangles are either disjoint or have a common vertex or a common side. To every triangulation we associate two parameters: h , ϑ . h is the largest side and ϑ the smallest angle of all triangles of the given triangulation. In the sequel we assume that as $h \rightarrow 0$, ϑ remains bounded away from zero,

$$(26) \quad \vartheta \geq \vartheta_0 > 0.$$

Now V_h^m is the finite-dimensional subspace of V consisting of all functions which on the triangles of the given triangulation are equal to polynomials $p_m(x, y)$ introduced in the preceding section. Every function from V_h^m belongs to $C^{(m)}(\bar{\Omega})$ and, at the same time, to $W_2^{(m+1)}(\Omega)$.

Let us consider the problem of finding u_h^m such that

$$(27) \quad a(u_h^m, v) = L(v), \quad \forall v \in V_h^m.$$

THEOREM 3. *Let $n \leq m + 1$. Under the assumptions (23), (24) and (26) there exists*

⁷ In this and the last section we write $\|\cdot\|_n$ instead of $\|\cdot\|_{n,\Omega}$ and $|\cdot|_n$ instead of $|\cdot|_{n,\Omega}$.

just one $u_h^m \in V_h^m$ satisfying (27) and

$$(28) \quad \|u - u_h^m\|_n \rightarrow 0 \quad \text{as } h \rightarrow 0.$$

Proof. It is an immediate consequence of the theorem by Céa about the Galerkin method (see [8, p. 363, Théorème 3.1]) and of Theorem 2 proved in Section 2. We must show that the subspaces V_h^m have the following property of density: there is a subspace $\mathcal{V} \subset V$ which is dense in V and a family of linear operators r_h^m from \mathcal{V} into V_h^m such that

$$(29) \quad \|v - r_h^m v\|_n \rightarrow 0, \quad \forall v \in \mathcal{V} \quad \text{as } h \rightarrow 0.$$

For \mathcal{V} we choose functions from V belonging to $W_2^{(k)}(\Omega)$ with $2m + 2 \leq k \leq 4m + 2$. As $k \geq 2m + 2$ it follows by Sobolev's lemma that $\mathcal{V} \subset C^{(2m)}(\bar{\Omega})$. $r_h^m v$ is then the function which on every triangle of the corresponding triangulation is equal to the interpolation polynomial $p_m(x, y)$ corresponding to $v(x, y)$. According to (15) and (26) we have

$$\|v - r_h^m v\|_{n,T}^2 \leq K_{10} h^{2(k-n)} |v|_{k,T}^2.$$

Hence,

$$(30) \quad \|v - r_h^m v\|_n \leq K_{11} |v|_k h^{k-n}$$

and (29) follows.

Theorem 3 proves only the convergence of the finite element method. Of course, we did not ask more than that the solution u of the boundary value problem (25) of the $2n$ th order belongs to $W_2^{(n)}(\Omega)$. If we suppose more about the smoothness of u we get an asymptotic estimate of the rate of convergence:

THEOREM 4. *Suppose that the form $a(u, v)$ is Hermitian. Let the assumptions of Theorem 3 hold and let*

$$u(x, y) \in W_2^{(k)}(\Omega), \quad 2m + 2 \leq k \leq 4m + 2.$$

Then

$$(31) \quad \|u - u_h^m\|_n \leq K |u|_k h^{k-n},$$

where the constant K does not depend on the triangulation and on the solution u .

Proof. We use a lemma by Céa [8, p. 365, Proposition 3.1]. According to the inequality 3.14 of this lemma

$$\|u - u_h^m\|_n \leq (M/\alpha)^{1/2} \|u - r_h^m u\|_n$$

holds. As $u \in W_2^{(k)}(\Omega)$ we can set $v = u$ in (30) and the proof is finished.

In case $n = 2, m = 1$, (31) gives

$$\|u - u_h^1\|_2 \leq K |u|_k h^{k-2}, \quad 4 \leq k \leq 6,$$

for $u \in W_2^{(k)}(\Omega)$. The highest order of accuracy is attained for $k = 6$,

$$\|u - u_h^1\|_2 \leq K |u|_6 h^4.$$

This result is a generalization of the result of [19] where instead of $|u|_6$ the seminorm $M_6 = \sup_{\Omega} |D^i u|$, $|i| = 6$, is used. In the same way we get for $n = 3, m = 2$ and $n = 4, m = 3$ the generalization of the results of [17].

4. Some Special Cases. 1. To get the asymptotic estimate (31) we had to assume a greater smoothness of the solution $u(x, y)$ than that guaranteed by the conditions (23) and (24) which, on the other hand, are sufficient for the uniqueness and existence of $u(x, y)$. In one case we need not impose any additional condition on the smoothness of the solution and yet we obtain an asymptotic error estimate, even in terms of data only. Consider, namely, the Dirichlet problem

$$(32) \quad Lu \equiv - \sum_{i,k=1}^2 \frac{\partial}{\partial x_i} \left(a_{ik} \frac{\partial u}{\partial x_k} \right) + au = f$$

on a convex polygon Ω . Let us assume that

$$(33) \quad a_{ik}(x, y) \in C^{(0,1)}(\Omega), \quad a(x, y), f(x, y) \in L_2(\Omega) \quad (j, k = 1, 2),$$

that the operator Lu is uniformly elliptic,

$$\sum_{i,k=1}^2 a_{ik}(x, y) \xi_i \xi_k \geq \alpha_0 \sum_{i=1}^2 \xi_i^2, \quad \alpha_0 > \text{const} > 0,$$

and that $a(x, y) \geq 0$. Then the form $a(u, v)$ corresponding to the above Dirichlet problem,

$$a(u, v) = \int_{\Omega} \left[\sum_{i,k=1}^2 a_{ik} \frac{\partial u}{\partial x_k} \frac{\partial v}{\partial x_i} + auv \right] dx dy,$$

is $\dot{W}_2^{(1)}(\Omega)$ -elliptic. According to a theorem of Kadlec [11] the solution $u(x, y)$ belongs to $W_2^{(2)}(\Omega)$ and

$$(34) \quad \|u\|_2 \leq C \|f\|_{L_2(\Omega)},$$

where the constant C depends only on the coefficients of the operator Lu and on the domain Ω . Actually, the result is stated in [11] for the equation

$$L_0 u \equiv - \sum_{i,k=1}^2 \frac{\partial}{\partial x_i} \left(a_{ik} \frac{\partial u}{\partial x_k} \right) = f.$$

However, if we write (32) in the form $L_0 u = -au + f$ we see that the right-hand side belongs to $L_2(\Omega)$. By the theorem of Kadlec

$$\|u\|_2 \leq C \| -au + f \|_{L_2(\Omega)} \leq C(K_{12}) \|u\|_1 + \|f\|_{L_2(\Omega)}.$$

As $a(u, u) = (Lu, u)_L$, for $u \in W_2^{(2)}(\Omega) \cap \dot{W}_2^{(1)}(\Omega)$ it follows from the $\dot{W}_2^{(1)}(\Omega)$ -ellipticity of $a(u, v)$ that $\|u\|_1 \leq (1/\alpha) \|f\|_{L_2(\Omega)}$. Hence (34) is true. Now the assumptions of Theorem 4 are satisfied ($n = 1, m = 0, k = 2$) and we have the following.

THEOREM 5. *Let Ω be a convex polygon and suppose that the real coefficients and the right-hand side of Eq. (32) satisfy (33). Further, let Lu be uniformly elliptic and let $a(x, y) \geq 0$. Then*

$$(35) \quad \|u - u_h^0\|_1 \leq C \|f\|_{L_2(\Omega)} h,$$

where the constant C depends on the coefficients of Lu and on the domain Ω only.

Using an argument similar to that of Nitsche [14] we can obtain the additional result

$$(36) \quad \|u - u_h^0\|_{L_2(\Omega)} \leq Ch^2 \|f\|_{L_2(\Omega)}.$$

Remark. The estimate of the form (36) is also given in the paper by L. A. Oganesjan, P. A. Ruchovc: "Investigation of the convergence rate of variational-difference schemes for elliptic second order equations in a two-dimensional domain with a smooth boundary," *Z. Vyčisl. Mat. i Mat. Fiz.*, v. 9, 1969, 1102–1120. (Russian)

The proof is as follows: Write

$$\|u - u_h^0\|_0 = \sup_{\psi \in L_2(\Omega)} \frac{|(u - u_h^0, \psi)_0|}{\|\psi\|_0}.$$

Now let ϕ satisfy

$$a(v, \phi) = (v, \psi)_0, \quad \forall v \in \overset{\circ}{W}_2^{(1)}(\Omega).$$

Then, as in (34), we have $\|\phi\|_2 \leq C\|\psi\|_0$. Hence,

$$(37) \quad \|u - u_h^0\|_0 \leq C \sup |a(u - u_h^0, \phi)| / \|\phi\|_2.$$

But

$$a(u - u_h^0, \phi) = a(u - u_h^0, \phi - \bar{\phi}), \quad \forall \bar{\phi} \in V_h^0.$$

Hence,

$$(38) \quad |a(u - u_h^0, \phi)| \leq M \|u - u_h^0\|_1 \|\phi - \bar{\phi}\|_1.$$

Choose $\bar{\phi}$ such that

$$(39) \quad \|\phi - \bar{\phi}\|_1 \leq K_{13} \|\phi\|_2 h.$$

Then (37), (38) and (39) imply

$$\|u - u_h^0\|_0 \leq K_{14} h \|u - u_h^0\|_1.$$

This together with Theorem 5 yields the result.

2. In [19] there was also introduced a cubic polynomial $p(x, y)$ determined by ten values

$$p(P_j), \quad \partial p(P_j)/\partial x, \quad \partial p(P_j)/\partial y, \quad p(P_0), \quad j = 1, 2, 3.$$

This element can be used for solving second-order boundary value problems. It is easy to show, in the same way as Theorem 2 was proved, that

$$\|u - p\|_{n, T} \leq \frac{K}{(\sin \alpha)^n} c^{k-n} |u|_{k, T}, \quad k = 3, 4, \quad n \leq k,$$

if $u \in W_2^{(k)}(T)$. For the corresponding finite element procedure (again under the assumptions (23), (24) and (26)) it follows first that it converges in the $\|\cdot\|_1$ norm, and secondly that

$$\|u - u_h\|_1 \leq K |u|_k h^{k-1}, \quad k = 3, 4,$$

if $u \in W_2^{(k)}(\Omega)$. For $k = 4$ this result is a generalization of the estimate (13) in [19].

3. The polynomial $p_1(x, y)$ is a 21-degree-of-freedom element. However, the values $\partial p_1(Q_j^{(1,1)})/\partial \nu_j$ ($j = 1, 2, 3$) are not necessary in applications. Bell proposed in [3] (also Goël in [10]) an 18-degree-of-freedom element and applied it to bending of thin plates. We get it from $p_1(x, y)$ if we eliminate the three above mentioned values by imposing on $p_1(x, y)$ the condition that $\partial p_1/\partial \nu_j$ ($j = 1, 2, 3$) be cubic

polynomials on the corresponding sides of T . In general, $\partial p_1/\partial \nu_i$ is a polynomial of the fourth degree in one variable on the side l_i of T and it is easy to see that the above condition determines uniquely the values $\partial p_1(Q_i^{(1,1)})/\partial \nu_i$ as linear combinations of the remaining 18 values

$$D^i p_1(P_i), \quad j = 1, 2, 3, \quad |i| \leq 2.$$

We denote this 18-degree-of-freedom element by $q(x, y)$. If we inspect the proof of Theorem 2, we easily find out that an estimate corresponding to (18) is again true in case of the element $q(x, y)$ and that the only change is that the functional F vanishes for polynomials of the degree less than 5, whereas, in case of $p_1(x, y)$ it vanishes for polynomials of the degree less than 6. We have

$$\|u - q\|_{n,T} \leq \frac{K}{(\sin \alpha)^n} c^{k-n} |u|_{k,T}, \quad n = 1, 2, \quad k = 4, 5,$$

if $u \in W_2^{(k)}(T)$. For the corresponding finite element procedure (again under the assumptions (23), (24) and (26)), it follows first that it converges in the norm $\|\cdot\|_1$ and $\|\cdot\|_2$, respectively, and secondly that

$$\|u - u_h\|_n \leq K |u|_k h^{k-n}, \quad n = 1, 2, \quad k = 4, 5,$$

if $u \in W_2^{(k)}(\Omega)$. Thus, for bending of thin plates the highest order of accuracy is the third order.

Similarly one can generalize the results of [20] where, by eliminating the value $p(P_0)$ from the cubic element $p(x, y)$, there was constructed a 9-degree-of-freedom element.

4. For practical applications it is desirable (see [20, p. 395]) that as many parameters determining the polynomials as possible are prescribed at the vertices only. In [12] it is remarked that in the case of polynomials of degree $4m + 1$ and $4m + 3$ (see footnote 2) the parameters prescribed on the sides of the triangle can be eliminated by imposing on the polynomials the condition that the normal derivatives of the k th order be polynomials of degree $n - 2k$ along the sides of the triangle. For the corresponding finite element procedure one can easily prove that

$$\|u - u_h\|_n \leq K |u|_k h^{k-n}$$

for $2m + 2 \leq k \leq 3m + 2$ and $2m + 3 \leq k \leq 3m + 4$, respectively, if $n \leq m + 1$ and $u \in W_2^{(k)}(\Omega)$.

It is also possible to eliminate the parameters prescribed at the center of gravity by imposing some restrictions on the polynomials. However, in this case a better practical way is to retain them and to use the method of condensation of internal parameters (see [21] or [22]).

Department of Mathematics
Cornell University
Ithaca, New York 14850

Technical University
Brno
Czechoslovakia

1. S. AGMON, *Lectures on Elliptic Boundary Value Problems*, Van Nostrand Math. Studies, no. 2, Van Nostrand, Princeton, N. J., 1965. MR 31 #2504.
2. J. H. ARGYRIS, I. FRIED & D. W. SCHARPF, "The tuba family of plate elements for the matrix displacement method," *Aeronautical J. Roy. Aeronautical Soc.*, v. 72, 1968, pp. 618–623.
3. K. BELL, *Analysis of Thin Plates in Bending Using Triangular Finite Elements*, The Technical University of Norway, Trondheim, 1968.
4. K. BELL, "A refined triangular plate bending finite element," *Internat. J. Numer. Methods in Engrg.*, v. 1, 1969, pp. 101–122.
5. I. S. BEREZIN & N. P. ŽIDKOV, *Computing Methods*. Vol. I, 2nd ed., Fizmatgiz, Moscow, 1962; English transl., of 1st ed., Pergamon Press, New York, 1965. MR 30 #4372; MR 31 #1756.
6. W. BOSSHARD, "Ein neues, vollverträgliches endliches Element für Plattenbiegung," *Abh. Int. Verein. Brückenbau und Hochbau, Zürich*, v. 28, 1968, pp. 27–40.
7. J. H. BRAMBLE & S. R. HILBERT, "Estimation of linear functionals on Sobolev spaces with application to Fourier transforms and spline interpolation," *Siam. J. Numer. Anal.*, v. 7, 1970, pp. 113–124.
8. J. CÉA, "Approximation variationnelle des problèmes aux limites," *Ann. Inst. Fourier (Grenoble)*, v. 14, 1964, pp. 345–444. MR 30 #5037.
9. R. COURANT, "Variational methods for the solution of problems of equilibrium and vibrations," *Bull. Amer. Math. Soc.*, v. 49, 1943, pp. 1–23. MR 4, 200.
10. J. J. GOËL, *List of Basic Functions for Numerical Utilisation of Ritz's Method. Application to the Problem of the Plate*, École Polytechnique Fédérale, Lausanne, 1969.
11. J. KADLEC, "The regularity of the solution of the Poisson problem in a domain whose boundary is similar to that of a convex domain," *Czechoslovak Math. J.*, v. 14(89), 1964, pp. 386–393. (Russian) MR 30 #329.
12. J. KRATOCHVÍL, A. ŽENÍŠEK & M. ZLÁMAL, "A simple algorithm for the stiffness matrix of triangular plate bending finite elements," *Numer. Methods in Engineering*. (To appear.)
13. J. L. LIONS & E. MAGENES, *Problèmes aux Limites non Homogènes et Applications*. Vol. 1, Travaux Recherches Math., no. 17, Dunod, Paris, 1968.
14. J. NITSCHE, "Lineare Spline-Funktionen und die Methode von Ritz für elliptische Randwertprobleme." (To appear.)
15. M. J. TURNER, R. W. CLOUGH, H. C. MARTIN & L. J. TOPP, "Stiffness and deflection analysis of complex structures," *J. Aeronautical Sci.*, v. 23, 1956, pp. 805–823.
16. M. VISSER, *The Finite Element Method in Deformation and Heat Conduction Problems*, Delft, 1968.
17. A. ŽENÍŠEK, "Interpolation polynomials on the triangle," *Numer. Math.*, v. 15, 1970, pp. 283–296.
18. O. C. ZIENKIEWICZ, *The Finite Element Method in Structural and Continuum Mechanics*, McGraw-Hill, New York, 1967.
19. M. ZLÁMAL, "On the finite element method," *Numer. Math.*, v. 12, 1968, pp. 394–409. MR 39 #5074.
20. M. ZLÁMAL, "A finite element procedure of the second order of accuracy," *Numer. Math.*, v. 16, 1970, pp. 394–402.
21. A. C. FELIPPA, *Refined Finite Element Analysis of Linear and Nonlinear Two-Dimensional Structures*, SESM Report No. 66–22, University of California, Berkeley, Calif., 1967.
22. E. ANDERHEGGEN, *Programme zur Methode der finiten Elemente*, Institut für Baustatik, Eidgenössische Technische Hochschule, Zürich, 1969.
23. G. R. COWPER, E. KOSKO, G. M. LINDBERG & M. D. OLSON, "Formulation of a new triangular plate bending element," *C.A.S.I. Trans.*, v. 1, 1968, pp. 86–90.
24. G. R. COWPER, E. KOSKO, G. M. LINDBERG & M. D. OLSON, "Static and dynamic applications of a high-precision triangular plate bending element," *AIAA J.*, v. 7, 1969, pp. 1957–1965.

2.5. HIGHER ORDER LOCAL ACCURACY BY AVERAGING IN THE FINITE ELEMENT METHOD (1977)

2.5 Higher order local accuracy by averaging in the finite element method (1977)

Higher order local accuracy by averaging in the finite element method[29]

Higher Order Local Accuracy by Averaging in the Finite Element Method

By J. H. Bramble and A. H. Schatz*

Abstract. Let u_h be a Ritz-Galerkin approximation, corresponding to the solution u of an elliptic boundary value problem, which is based on a uniform subdivision in the interior of the domain. In this paper we show that by “averaging” the values of u_h in the neighborhood of a point x we may (for a wide class of problems) construct an approximation to $u(x)$ which is often a better approximation than $u_h(x)$ itself. The “averaging” operator does not depend on the specific elliptic operator involved and is easily constructed.

1. Introduction. In this paper we shall discuss some local “superconvergence” results for a large class of Ritz-Galerkin methods, which are used to approximate solutions of elliptic boundary value problems. Briefly, let Ω be a bounded domain in \mathbb{R}^N with boundary $\partial\Omega$ and for simplicity consider the boundary value problem

$$(1.1) \quad Lu = - \sum_{i,j=1}^N \frac{\partial}{\partial x_j} \left(a_{ij}(x) \frac{\partial u}{\partial x_i} \right) + \sum_{i=1}^N b_i(x) \frac{\partial u}{\partial x_i} + c(x)u = f \quad \text{in } \Omega,$$

$$(1.2) \quad bu = g \quad \text{on } \partial\Omega,$$

where L is uniformly elliptic on Ω and b is some boundary operator such that the problem (1.1), (1.2) has a unique solution.

Suppose that for each $0 < h < 1$ we are given a linear space of finite elements S_h (which we roughly think of as being a space of piecewise polynomials defined on some partition of Ω) and an approximate solution $u_h \in S_h$ of (1.1), (1.2) determined by one of the many Ritz-Galerkin methods which have been proposed. For a survey of some of these we refer the reader to [2]. Let $\Omega_1 \subset\subset \Omega$. For many of these methods, u_h satisfies the interior equations

$$\begin{aligned} B(u_h, \varphi) &= \int_{\Omega} \left(\sum_{i,j=1}^N a_{ij} \frac{\partial u_h}{\partial x_i} \frac{\partial \varphi}{\partial x_j} + \sum_{i=1}^N b_i \frac{\partial u_h}{\partial x_i} \varphi + cu_h \varphi \right) dx \\ &= B(u, \varphi) = \int_{\Omega} f\varphi dx \end{aligned}$$

for all $\varphi \in \overset{\circ}{S}_h(\Omega_1)$, where $\overset{\circ}{S}_h(\Omega_1) = \{\varphi | \varphi \in S_h(\Omega), \text{supp}(\varphi) \subset \Omega_1\}$. Roughly speaking, suppose that S_h has the property (and some others described in Section 3) that for each u belonging to the Sobolev space H^r ,

$$\inf_{\psi \in S_h} \|u - \psi\|_{1, \Omega_1} \leq Ch^{r-1} \|u\|_{r, \Omega_1}.$$

Received July 15, 1976.

AMS (MOS) subject classifications (1970). Primary 65N30.

*This work was supported in part by a grant from the National Science Foundation.

Copyright © 1977, American Mathematical Society

Here $r \geq 2$ is a given integer and for $s \geq 0$, $\|\cdot\|_{s, \Omega_1}$ denotes the norm on $H^s(\Omega_1)$. It was shown in [15] that for $\Omega_0 \subset\subset \Omega_1$ the interior estimate

$$(1.4) \quad \|u - u_h\|_{0, \Omega_0} \leq C\{h^r\|u\|_{r, \Omega_1} + \|u - u_h\|_{-p, \Omega_1}\}$$

holds. Here p is a nonnegative integer and $\|\cdot\|_{-p, \Omega_1}$ is the norm on the dual space of $\overset{\circ}{H}{}^p(\Omega_1)$ (the completion of $C_0^\infty(\Omega_1)$ under $\|\cdot\|_p$). Under the further assumption that the restriction of elements of S_h to Ω_1 are piecewise polynomials defined on a uniform mesh, it was shown in [4] that

$$(1.5) \quad |u - u_h|_{0, \Omega_0} \leq C\{h^r\|u\|_{r+[N/2]+1, \Omega_1} + \|u - u_h\|_{-p, \Omega_1}\}.$$

Here $|\cdot|_{0, \Omega_0}$ denotes the maximum norm on $\bar{\Omega}_0$ and $[N/2]$ is the integral part of $N/2$.

For many methods, estimates for $\|e\|_{-p, \Omega_1}$ already exist in the literature. Noting that $\|e\|_{-p, \Omega_1} \leq \|e\|_{-p, \Omega}$ and taking $p = r - 2$, one often finds that $\|u - u_h\|_{2-r, \Omega} \leq Ch^{2r-2}\|u\|_{r, \Omega}$. Hence in these cases, if $u \in H^r(\Omega)$ and $r \geq 3$, this latter term is of higher order in h than the first terms on the right in (1.4) and (1.5).

Instead of considering u_h , in this paper we shall consider certain “averages” of u_h as approximations to u . More precisely, the averages are formed by computing $K_h * u_h$, where K_h is a fixed function and $*$ denotes convolution. We shall see that the function K_h has the following properties:

- (i) K_h has small support,
- (ii) K_h is independent of the specific choice of S_h or the operator L ,
- (iii) $K_h * u_h$ is easily computable from u_h .

Furthermore we shall prove the following estimates (where again we assume S_h to be defined on a uniform mesh on Ω_1):

$$\|u - K_h * u_h\|_{0, \Omega_0} \leq C\{h^{2r-2}\|u\|_{2r-2, \Omega_1} + \|u - u_h\|_{-p, \Omega_1}\}$$

and

$$|u - K_h * u_h|_{0, \Omega_0} \leq C\{h^{2r-2}\|u\|_{2r-2+[N/2]+1, \Omega_1} + \|u - u_h\|_{-p, \Omega_1}\}.$$

Hence in the cases in which $r \geq 3$ and $\|u - u_h\|_{2-r, \Omega_1} \leq Ch^{2r-2}\|u\|_{r, \Omega}$, $K_h * u$ is of order h^{2r-2} locally and therefore, for h sufficiently small, better approximates u than does u_h .

It is important to note that $K_h * u_h$ may be computed at each point in Ω_0 . However if we restrict our attention to specific points, say, for example, mesh points, then it very often takes on a very simple form; this will be discussed in Section 5.

An announcement of some of the results obtained in this paper was given in [7]. For some superconvergence type results for two point boundary value problems the reader is referred to [9], [10] and [11]. For some interior superconvergence type results for multi-dimensional elliptic problems see [6]. For parabolic problems we refer the reader to [17] and [18].

An outline of this paper is as follows: In Section 2 we collect some notation.

In Section 3 we discuss the necessary properties of the subspaces and in Section 4 we discuss the interior equations and collect some further preliminaries. In Section 5 the averaging operator is defined and its computation is discussed. Section 6 contains our main superconvergence results. In Section 7 we give some applications of the results and in one example show how to obtain superconvergence up to the boundary when Ω is the unit square, $Lu = -\Delta u$ in Ω and $bu \equiv u$ on $\partial\Omega$.

2. Notation and Some Preliminaries. Let Ω be a bounded open set in \mathbf{R}^N . For m a nonnegative integer, $C^m(\Omega)$ will denote the space of real-valued functions having continuous derivatives up to order m on $\bar{\Omega}$ with the norm

$$(2.1) \quad |u|_m = \sum_{|\alpha| \leq m} \sup_{x \in \bar{\Omega}} |D^\alpha u(x)|.$$

$C_0^\infty(\Omega)$ will denote the infinitely differentiable functions on Ω whose support is contained in Ω . $H^m(\Omega)$ (resp. $\overset{\circ}{H}{}^m(\Omega)$) is the completion of $C^\infty(\Omega)$ (resp. $C_0^\infty(\Omega)$) with respect to the norm

$$(2.2) \quad \|u\|_{m,\Omega} = \left(\sum_{|\alpha| \leq m} \int_\Omega |D^\alpha u|^2 dx \right)^{1/2}.$$

Note that $H^0(\Omega) = \overset{\circ}{H}{}^0(\Omega) = L_2(\Omega)$.

For m a nonnegative integer $H^{-m}(\Omega)$ is the completion of $C_0^\infty(\Omega)$ with respect to the norm

$$(2.3) \quad \|u\|_{-m,\Omega} = \sup_{v \in C_0^\infty(\Omega)} \frac{\int_\Omega uv dx}{\|v\|_{m,\Omega}} = \sup_{v \in C_0^\infty(\Omega)} \frac{(u, v)}{\|v\|_{m,\Omega}}.$$

For β a multi-integer define the translation operator $T_h^\beta u(x) = u(x + \beta h)$ and the difference quotients,

$$\partial_{h,j} u = h^{-1} (T_{h/2}^{e_j} - T_{h/2}^{-e_j}) u.$$

Here e_j is the multi-index whose j th component is 1 and all others 0. For any multi-index α we set

$$\partial_h^\alpha u = (\partial_{h,1}^{\alpha_1} \cdots \partial_{h,N}^{\alpha_N}) u.$$

We shall also need the following results:

LEMMA 2.1 (cf. e.g. [12]). *Let $\Omega_0 \subset\subset \Omega_1$. If $u \in H^{[N/2]+1}(\Omega_1)$, then (after possible modification on a set of measure zero) $u \in C^0(\Omega_0)$ and*

$$(2.4) \quad |u|_{0,\Omega_0} \leq C \|u\|_{[N/2]+1,\Omega_1}$$

where $C = C(\Omega_0, \Omega_1)$ and $[N/2]$ is the integral part of $N/2$.

LEMMA 4.2. *Let $\Omega_0 \subset\subset \Omega_1$ and s be an arbitrary but fixed nonnegative integer. Then there is a constant C such that*

$$(2.5) \quad \|u\|_{0,\Omega_0} \leq C \sum_{|\alpha| \leq s} \|D^\alpha u\|_{-s,\Omega_1}.$$

Proof. Without loss of generality we may assume that Ω_1 is smooth and $\Omega_0 \subset\subset \Omega_1 \subset\subset \mathbf{R}^N$. First suppose that $\text{supp } u \subset\subset \Omega_1$. In this case the proof proceeds by induction. For $s = 0$ (2.5) is trivial. Assume (2.5) for $s \geq 0$, then for any α with $|\alpha| \leq s$

$$(2.6) \quad \|D^\alpha u\|_{-s, \Omega_1} = \sup_{v \in C_0^\infty(\Omega_1)} \frac{(D^\alpha u, v)}{\|v\|_{s, \Omega_1}}.$$

Let g be a solution of $-\Delta g = v$ in Ω_1 , $v = 0$ on $\partial\Omega_1$. Since u has compact support, integration by parts and Schwarz's inequality yields

$$(2.7) \quad \begin{aligned} (D^\alpha u, v) &= \sum_{i=1}^N \left(\frac{\partial}{\partial x_i} D^\alpha u, \frac{\partial g}{\partial x_i} \right) \\ &\leq C \left(\sum_{|\beta|=|\alpha|+1} \|D^\beta u\|_{-s-1, \Omega_1} \right) \|g\|_{s+2}. \end{aligned}$$

But by elliptic regularity (cf. [13])

$$\|g\|_{s+2, \Omega_1} \leq C \|v\|_{s, \Omega_1}.$$

In view of this, (2.7) and (2.6), it follows that

$$\sum_{|\alpha| \leq s} \|D^\alpha u\|_{-s, \Omega_1} \leq \sum_{|\beta| \leq s+1} \|D^\beta u\|_{-s-1, \Omega_1},$$

which completes the proof for u with compact support in Ω_1 . More generally let $w \in C^\infty(\Omega_1)$ with $w = 1$ on Ω_0 and $\text{supp } w \subset\subset \Omega_1$. Applying (2.5) to the function wu we obtain

$$\|u\|_{0, \Omega_0} \leq C \sum_{|\alpha| \leq s} \|D^\alpha(wu)\|_{-s, \Omega_1}.$$

Using Leibniz's rule and the definition of the $H^{-s}(\Omega_1)$ norm it is easily seen that

$$\sum_{|\alpha| \leq s} \|D^\alpha(wu)\|_{-s, \Omega_1} \leq C \sum_{|\alpha| \leq s} \|D^\alpha u\|_{-s, \Omega_1},$$

which completes the proof.

3. Piecewise Polynomial Subspaces. In the following sections we shall be concerned with Ritz-Galerkin methods in which the functions used for approximation are piecewise polynomials on a uniform mesh. More precisely, let $\{Q_\beta\}$, $\beta \in Z^N$ (the multi-integers) denote a family of disjoint open sets which partition \mathbf{R}^N . We shall assume that the partition is invariant under translations by ν , for any $\nu \in Z^N$. Let $r \geq 2$ be an integer. S'_1 will denote a linear space of functions which are polynomials on each Q_β and belong to $H^1(\Omega)$ for any $\Omega \subset\subset \mathbf{R}^N$. Furthermore we assume that S'_1 is translation invariant in the sense that if $\varphi \in S'_1$, then $T_1^\alpha \varphi \in S'_1$, for all $\alpha \in Z^N$.

For each $0 < h \leq 1$, let $Q_\beta^h = hQ_\beta$. Then the family $\{Q_\beta^h\}$ forms a partition of \mathbf{R}^N which is invariant under translations by $h\nu$, $\nu \in Z^N$. For $\Omega_1 \subset\subset \mathbf{R}^N$, $S'_h(\Omega_1)$ will denote the finite-dimensional subspace of $H^1(\Omega_1)$ consisting of functions which are the restrictions to Ω_1 of functions of the form $\varphi(x/h)$, where $\varphi(x) \in S'_1$. Let $\overset{\circ}{S}'_h(\Omega) =$

$\{\varphi \mid \varphi \in S_h^r(\Omega), \text{supp } \varphi \subseteq \Omega\}$. We note that if $\Omega_0 \subset\subset \Omega_1$ and $\alpha \in Z^N$, then for h sufficiently small (depending on α), $\partial_h^\alpha \varphi \in \overset{\circ}{S}_h^r(\Omega_1)$ for all $\varphi \in \overset{\circ}{S}_h^r(\Omega_0)$. We shall also make the following approximability assumptions:

A.1. Let $\Omega_0 \subset\subset \Omega_1$. There exists an $h_0 > 0$ such that for all $h \in (0, h_0]$ and $u \in \overset{\circ}{H}^j(\Omega_0)$

$$(3.1) \quad \inf_{\eta \in \overset{\circ}{S}_h^r(\Omega_0)} \|u - \eta\|_{1, \Omega_0} \leq Ch^{j-1} \|u\|_{j, \Omega_1}, \quad 1 \leq j \leq r,$$

where C is independent of u and h .

A.2. If $\varphi \in S_h^r(\Omega_1)$, $h \in (0, h_0]$, $\Omega_0 \subset\subset \Omega_1$, $\omega \in C_0^\infty(\Omega_0)$, then there exists an $\eta \in \overset{\circ}{S}_h^r(\Omega_1)$ such that

$$(3.2) \quad \|\omega\varphi - \eta\|_{1, \Omega_0} \leq Ch\|\varphi\|_{1, \Omega_1},$$

where C is independent of h and φ .

Often in practice subspaces which satisfy the above conditions can be described in the following manner: There exist functions $\varphi_1, \dots, \varphi_m$, in H^1 , which are piecewise polynomials with compact support, such that $\varphi \in S_h^r(\Omega)$ is of the form

$$(3.3) \quad \varphi(x) = \sum_{j=1}^m \sum_{\beta \in Z^N} a_\beta^j \varphi_j(h^{-1}x - \beta) \quad \text{for } x \in \Omega.$$

Here the coefficients a_β^j are real.

Example 1. An example of subspaces of this type are those generated by the *B-splines* of Schoenberg [16]. These will play a central role in what follows. Let us describe them more precisely.

For t real, define

$$\chi(t) = \begin{cases} 1, & |t| \leq \frac{1}{2}, \\ 0, & |t| > \frac{1}{2}, \end{cases}$$

and for l an integer, set

$$(3.4) \quad \psi_1^{(l)}(t) = \chi * \chi * \cdots * \chi, \quad \text{convolution } l-1 \text{ times.}$$

The function $\psi_1^{(l)}$ is the one-dimensional *B-spline basis function* of order l . Then

$$(3.5) \quad \psi^{(l)}(x) = \prod_{j=1}^N \psi_1^{(l)}(x_j)$$

is the N -dimensional *B-spline basis function* of order l . In this case every element of S_h on Ω is of the form

$$(3.6) \quad \varphi(x) = \sum_{\alpha \in Z^N} a_\alpha \psi^{(l)}(h^{-1}x - \alpha), \quad x \in \Omega.$$

Example 2. A more general class of splines than (3.6) which satisfy our assumptions are the tensor products of one-dimensional splines. They may be described as follows: Let us subdivide \mathbf{R}^1 into intervals I_j of length h . Let $M_{k,r}^h(\mathbf{R}^1) = \{\varphi \mid \varphi \in P_{r-1}(I_j), \varphi \in C^k(\Omega) \text{ for any } \Omega \subset\subset \mathbf{R}^1\}$, where $P_{r-1}(I_j)$ are the polynomials of degree

$\leq r - 1$ restricted to I_j . $M_{k,r}^h(\mathbf{R}^N)$ is then defined to be the N -fold tensor product of $M_{k,r}^h(\mathbf{R}^1)$ and for any $\Omega \subset\subset \mathbf{R}^N$, $M_{k,r}^h(\Omega)$ is the restriction of $M_{k,r}^h(\mathbf{R}^N)$ to Ω . We note that if $k = r - 2$ then these are the B -splines defined above. If m is a positive integer, $r = 2m$ and $k = m - 1$, then these are the so-called piecewise Hermite polynomials (cf. [4]).

Example 3. It was shown in [15] that the triangular elements of Bramble and Zlámal [8] satisfy our conditions. Here we take them on a uniform triangulation of \mathbf{R}^N .

4. Interior Equations for Ritz-Galerkin Methods and Some Preliminaries. Let Ω_1 be a bounded open set in \mathbf{R}^N and $B(u, v)$ a bilinear form defined on $H^1(\Omega_1) \times H^1(\Omega_1)$ given by

$$(4.1) \quad B(u, v) = \int_{\Omega_1} \left(\sum_{i,j=1}^N a_{ij}(x) D_i u D_j v + \sum_{i=1}^N b_i(x) (D_i u) v + c(x) u v \right) dx,$$

where for simplicity the coefficients a_{ij}, b_i , are assumed to be of class $C^\infty(\Omega_1)$ and $a_{ij} = a_{ji}$. We note that in general $B(u, v)$ may not be symmetric. We assume throughout that $B(u, v)$ is uniformly elliptic on Ω_1 ; i.e., there exists a constant $C > 0$ such that for all $x \in \bar{\Omega}_1$ and real vectors $\xi = (\xi_1, \dots, \xi_N) \neq 0$,

$$C \sum_{i=1}^N \xi_i^2 \leq \sum_{i,j=1}^N a_{ij} \xi_i \xi_j.$$

Let $u \in H^1(\Omega_1)$ and $u_h \in S_h^r(\Omega_1)$ and suppose that $u - u_h$ satisfies the “interior” Ritz-Galerkin equations

$$(4.2) \quad B(u - u_h, \varphi) = 0$$

for all $\varphi \in \overset{\circ}{S}_h^r(\Omega_1)$. Here u_h may be thought of as an approximation to u on Ω_1 obtained by using some Ritz-Galerkin method on a larger set Ω . Instead of investigating the properties of the error $u - u_h$ we shall be interested in investigating the properties of the difference between u and certain averages of u_h . In order to do so we shall need some results concerning $u - u_h$ derived in [15] and [4].

LEMMA 4.1. *Let $\Omega_0 \subset\subset \Omega_1 \subset\subset \mathbf{R}^N$, $u \in H^{r+|\alpha|}(\Omega_1)$, α a multi-index, $u_h \in S_h^r(\Omega_1)$, $r \geq 2$ and p be a nonnegative integer, arbitrary but fixed. Suppose that A.1 and A.2 are satisfied and that $u - u_h$ satisfies (4.2). Then for h sufficiently small*

$$(4.3) \quad \|\partial_h^\alpha (u - u_h)\|_{2-r, \Omega_0} \leq C(h^{2r-2} \|u\|_{r+|\alpha|, \Omega_1} + \|u - u_h\|_{-p, \Omega_1}).$$

Here C is independent of u , u_h and h but in general depends on p , α , Ω_0 , Ω_1 and the coefficients a_{ij} , b_i and c .

Proof. Clearly it is sufficient to prove (4.3) for ∂_h^α replaced by the forward difference operator $\tilde{\partial}_h^\alpha = T_{h/2}^\alpha \partial_h^\alpha$, and Ω_0 replaced by Ω'_0 with $\Omega_0 \subset\subset \Omega'_0$. Now let $\Omega_0 \subset\subset \Omega'_0 \subset\subset \Omega'_1 \subset\subset \Omega_1$. It follows from (6.8) of [15], that for h sufficiently small and $e = u - u_h$

$$(4.4) \quad B(\tilde{\partial}_h^\alpha e, \varphi) = \int_{\Omega_0} \sum_{i,j=1}^N \sum_{\beta < \alpha} \binom{\alpha}{\alpha - \beta} \tilde{\partial}_h^{\alpha-\beta} a_{ij} T_h^{\alpha-\beta} \tilde{\partial}_h^\beta D_i e D_j \varphi \, dx,$$

$$\forall \varphi \in \mathring{S}^h(\Omega'_0).$$

Hence, referring to [15] in particular, from (5.6) of Theorem 5.2, (4.3) of Lemma 4.2, (6.5) following Theorem 6.1 and (4.4) above we have that

$$(4.5) \quad \|\tilde{\partial}_h^\alpha e\|_{2-r, \Omega'_0} \leq C \left(h^{2r-2} \|u\|_{r-2+|\alpha|, \Omega'_1} + \|e\|_{-p, \Omega'_1} + \sum_{\beta < \alpha} \|\tilde{\partial}_h^\beta e\|_{2-r, \Omega'_1} \right).$$

The proof now proceeds by induction on $|\alpha|$. If $\alpha = 0$ then (4.3) follows from (4.5). Assume now that (4.3) holds for $|\alpha| \leq m$. If $|\alpha| = m + 1$, then applying (4.3) to the third term on the right of (4.5) on the domains Ω'_1 and Ω_1 the inequality (4.3) for $|\alpha| = m + 1$ is obtained, which completes the proof.

LEMMA 4.2 [4]. *Under the conditions of Lemma 4.1 there exists a constant $C > 0$ such that*

$$(4.6) \quad |\partial_h^\alpha (u - u_h)|_{0, \Omega_0} \leq C(h^r \|u\|_{r+|\alpha|+[N/2]+1, \Omega_1} + \|u - u_h\|_{-p, \Omega_1}),$$

where C is independent of u , u_h and h but in general depends on p , α , Ω_0 , Ω_1 and the coefficients a_{ij} , b_i and c .

5. A Class of Convolution Operators. In this section we shall introduce a particular class of convolution operators and discuss some of their properties. In the next section we shall use these to show (under certain conditions) that a certain convolution of the Ritz-Galerkin solution u_h is closer to u than u_h is itself.

We shall begin by defining the kernels of the above-mentioned convolution operators. In each case it is an N -dimensional smooth spline which is the product of a single suitably chosen one-dimensional smooth spline. More precisely let $\psi_1^{(l)}$ be the one-dimensional smooth spline of order l defined by (3.4). For $l = r - 2$, $r \geq 2$ given, and $x \in \mathbf{R}^N$ set

$$(5.1) \quad K_h(x) = \prod_{m=1}^N \left(\sum_{j=-(r-2)}^{r-2} h^{-1} k'_j \psi_1^{(r-2)}(h^{-1} x_m - j) \right).$$

Here the constants K'_j are defined as follows:

- (i) $k'_{-j} = k'_j$, $j = 0, \dots, r - 2$.
- (ii) $k'_0 = k_0$ and $k'_j = k_{j/2}$, $j = 1, \dots, r - 2$,

where the k_j , $j = 0, \dots, r - 2$, are determined as the unique solution (see Lemma 5.1 below) of the linear system of algebraic equations.

$$(5.2) \quad \sum_{j=0}^{r-2} k_j \int_{R_1} \psi_1^{(r-2)}(y)(y+j)^{2m} dy = \begin{cases} 1 & \text{if } m = 0, \\ 0 & \text{if } m = 1, \dots, r - 2. \end{cases}$$

We note that the constants k_j can be easily computed and depend only on the choice of r .

The convolution operators in which we are interested are of the form

$$(5.3) \quad (K_h * u)(x) = h^{-N} \int_{\mathbf{R}^N} \left[\prod_{m=1}^N \sum_{j=2-r}^{r-2} k'_j \psi_1^{(r-2)}(h^{-1}(x_m - y_m) - j) \right] u(y) dy.$$

If we set $k'_j = 0$ for $|j| \geq r-1$ and for β a multi-integer define $k'_\beta = \prod_{j=1}^N k'_{\beta_j}$, then (5.2) and (5.3) may be rewritten as

$$(5.4) \quad K_h(x) = h^{-N} \sum_{\beta} k'_\beta \psi^{(r-2)}(h^{-1}x - \beta)$$

and

$$(5.5) \quad (K_h * u)(x) = h^{-N} \int_{\mathbf{R}^N} \sum_{\beta} k'_\beta \psi^{(r-2)}(h^{-1}(x - y) - \beta) u(y) dy,$$

respectively.

In what follows we shall be interested in $K_h * u_h$ as an approximation to u . Before investigating the properties of K_h , let us first consider the computation of $K_h * u_h$. If, in the region to be studied, u_h is of the form (3.3), that is

$$(5.6) \quad u_h = \sum_{j=1}^K \sum_{\alpha \in Z^N} a_\alpha^j \varphi_j(h^{-1}x - \alpha),$$

then using (5.5) and making a change in variables we have

$$(5.7) \quad \begin{aligned} & (K_h * u_h)(x) \\ &= \sum_{j=1}^m \sum_{\alpha \in Z^N} a_\alpha^j \left(\sum_{\beta \in Z^N} k'_\beta \int_{\mathbf{R}^N} \psi^{r-2}(h^{-1}x - \eta - \beta - \alpha) \varphi_j(\eta) d\eta \right). \end{aligned}$$

Although the terms $k'_\beta \int_{\mathbf{R}^N} \psi^l(h^{-1}x - \eta - \beta - \alpha) \varphi_j(\eta) d\eta$ and hence $K_h * u_h$ may be calculated at any point, we shall, for simplicity restrict our attention to points of the form $x = h\gamma$, $\gamma \in Z^N$. In this case

$$(K_h * u_h)(h\gamma) = \sum_{j=1}^m \sum_{\alpha \in Z^N} a_\alpha^j \left(\sum_{\beta \in Z^N} k'_\beta \int_{\mathbf{R}^N} \psi^{r-2}(\gamma - \beta - \alpha - \eta) \varphi_j(\eta) d\eta \right)$$

or

$$(5.8) \quad (K_h * u_h)(h\gamma) = \sum_{j=1}^m \sum_{\gamma, \delta \in Z^N} a_{\gamma-\delta}^j d_\delta^j,$$

where

$$d_\delta^j = \sum_{\beta \in Z^N} k'_\beta \int_{\mathbf{R}^N} \psi^{r-2}(\delta - \beta - \eta) \varphi_j(\eta) d\eta.$$

Thus the values of $K_h * u_h$ at mesh points are determined by taking a fixed finite linear combination of the values of the coefficients a_α^j . There are only a finite number of nonzero coefficients d_δ^j . They may be computed a priori and are independent of h and the particular mesh point.

As an example, let us take the case where u_h itself is a smooth spline. Then

$k = 1$, $\varphi_1 = \psi^{(r)}$ and u_h is of the form

$$(5.9) \quad u_h = \sum_{\alpha \in \mathbb{Z}^N} a_\alpha \psi^{(r)}(h^{-1}x - \alpha).$$

Setting $d_\delta^j = d_\delta$ in this case, we have

$$(5.10) \quad d_\delta = \sum_{\beta \in \mathbb{Z}^N} k'_\beta \psi^{2r-2}(\delta - \beta) = \prod_{l=1}^N d'_{\delta_l} \equiv \prod_{l=1}^N \left(\sum_{\eta=-(t-1)}^{t-1} k'_\eta \psi_1^{2r-2}(\delta_l - \eta) \right).$$

Here the d'_l correspond to the weights d_δ in the one-dimensional case. A table of the one-dimensional coefficients d'_l from which the N -dimensional weights d_δ in (5.10) may be formed will be given in the appendix. We remark that the computation of the d_δ^j may also be reduced to the product of the one-dimensional case if we take S_h to be the more general class of tensor products of one-dimensional splines given in Example 2 of Section 3.

For technical reasons in what follows we shall make use of a slightly more general class of convolution operators than those defined by (5.3). We shall now introduce these and discuss some of their properties. The following result will be needed.

LEMMA 5.1 (cf. [6, Lemma 8.1]). *Let $t \geq 1$ and $l \geq 1$ be arbitrary but fixed integers. There exist uniquely determined real constants k_j , $j = 0, \dots, t-1$, which satisfy the linear system of algebraic equations.*

$$(5.11) \quad \sum_{j=0}^{t-1} k_j \int_{\mathbb{R}^N} \psi_1^{(l)}(y) (y + j)^{2m} dy = \begin{cases} 1 & \text{if } m = 0, \\ 0 & \text{if } m = 1, \dots, t-1. \end{cases}$$

Here, with a slight abuse of notation, we have suppressed the dependence of the solutions of (5.11) on t and l . We note that (5.2) and (5.11) coincide in the case $l = r-2$ and $t = r-1$.

As before let $k'_0 = k_0$, $k'_j = k_j/2$, $k'_{-j} = k'_j$ for $j = 1, \dots, t-1$ and $k_j = 0$ for $|j| \geq t$. For $\beta \in \mathbb{Z}^N$ set $k'_\beta = \prod_{j=1}^N k'_{\beta_j}$ and define

$$(5.12) \quad \begin{aligned} K_{h,l}^{2t}(x) &= h^{-N} \sum_{\beta \in \mathbb{Z}^N} k'_\beta \psi^l(h^{-1}x - \beta) \\ &= \prod_{m=1}^N \left[h^{-1} \sum_{j=-(t-1)}^{t-1} k'_j \psi_1^{(l)}(h^{-1}x_m - j) \right]. \end{aligned}$$

We remark that $\text{supp}(K_{h,l}^{2t}(x))$ is the cube with side of length $(2t+l)h$ centered at the origin. Note also that $K_{h,r-2}^{2r-2}(x) = K_h(x)$.

The function $K_{h,l}^{2t}(x)$ was constructed in such a way that $K_{h,l}^{2t} * u$, where $*$ denotes convolution, is an approximation of order h^{2t} to u . More precisely we have the following:

LEMMA 5.2. *Let $t \geq 1$ and $l \geq 1$ be fixed integers and $\Omega_0 \subset\subset \Omega_1$. Then for h sufficiently small*

$$(5.13) \quad |u - K_{h,l}^{2t} * u|_{0,\Omega_0} \leq Ch^s |u|_{s,\Omega_1}, \quad 0 \leq s \leq 2t,$$

and

$$(5.14) \quad \|u - K_{h,l}^{2t} * u\|_{0,\Omega_0} \leq Ch^s \|u\|_{s,\Omega_1}, \quad 0 \leq s \leq 2t.$$

The proofs will not be given. We only remark that (5.13) and (5.14) follow as a simple consequence of the Bramble-Hilbert Lemma (cf. [3]), using the fact that Lemma 5.1 implies that $K_{h,l}^{2t} * u$ reproduces polynomials of degree not exceeding $2t - 1$ in each variable (cf. [6, Lemma 8.1]). We also note that (5.13) may also be proved using this latter fact and Taylor's Theorem.

We shall need some other properties of $K_{h,l}^{2t} * u$.

LEMMA 5.3. *Let t, l, h, Ω_0 and Ω_1 be as above. Then*

(i) *For any multi-index α and $u \in L_2(\Omega_1)$*

$$(5.15) \quad \partial_h^\alpha (K_{h,l}^{2t} * u)(x) = K_{h,l}^{2t} * \partial_h^\alpha u(x), \quad x \in \Omega_0.$$

(ii) *If s is any fixed integer (positive or negative) and α is any multi-index with $\alpha_j \leq l, j = 1, \dots, N$,*

$$(5.16) \quad \|D^\alpha(K_{h,l}^{2t} * u)\|_{s,\Omega_0} \leq C \|\partial_h^\alpha u\|_{s,\Omega_1} \quad \text{for all } u \in H^s(\Omega_1)$$

and

$$(5.17) \quad |D^\alpha(K_{h,l}^{2t} * u)|_{0,\Omega_0} \leq C |\partial_h^\alpha u|_{0,\Omega_1} \quad \text{for all } u \in C^0(\Omega_1).$$

Here C is a constant which is independent of h and u .

Proof. The identity (5.15) follows very simply using a change of variables.

For simplicity in notation we shall prove (5.16) and (5.17) in the case $N = 1$.

The multi-dimensional case follows in essentially the same manner. For $N = 1$ we have

$$(K_{h,l}^{2t} * u)(x) = h^{-1} \int_{\mathbb{R}^1} \sum_{j=-(t-1)}^{t-1} k'_j \psi_1^{(l)}(h^{-1}(x-y) - j) u(y) dy.$$

If $0 \leq \alpha \leq l$ define

$$\mathcal{V}_{h,l-\alpha}^{2t}(x) = h^{-1} \sum_{j=-(t-1)}^{t-1} k'_j \psi_1^{(l-\alpha)}(h^{-1}x - j).$$

Then it is easy to see from (5.15) and (3.4) that

$$D^\alpha(K_{h,l}^{2t} * u) = \mathcal{V}_{h,l-\alpha}^{2t} * \partial_h^\alpha u.$$

Now since $\psi_1^{(l-\alpha)} \geq 0$, $h^{-1} \int_{\mathbb{R}^1} \psi_1^{(l-\alpha)}(h^{-1}x - j) dx = 1$ and $\text{supp}(\psi_1^{(l-\alpha)}(h^{-1}x))$ is an interval of length $(l + 1 - \alpha)h$, it follows that for h sufficiently small

$$(5.18) \quad |\mathcal{V}_{h,l-\alpha}^{2t} * \partial_h^\alpha u|_{0,\Omega_0} \leq C |\partial_h^\alpha u|_{0,\Omega_1}$$

and

$$(5.19) \quad \|\mathcal{V}_{h,l-\alpha}^{2t} * \partial_h^\alpha u\|_{0,\Omega_0} \leq C \|\partial_h^\alpha u\|_{0,\Omega_1}.$$

The inequality (5.8) is just the inequality (5.17) and (5.19) is the inequality (5.16) in the case that $s = 0$. In the case that s is nonnegative, the inequality (5.16) follows from (5.9) on observing that for any nonnegative integer m ,

$$D^m(D^\alpha(K_{h,l}^{2t} * u)) = V_{h,l-2}^{2t} * D^m \partial_h^\alpha u.$$

For s a negative integer we have

$$\begin{aligned} \|D^\alpha K_{h,l}^{2t} * u\|_{s,\Omega_0} &= \sup_{v \in C_0^\infty(\Omega_0)} \frac{(D^\alpha(K_{h,l}^{2t} * u), v)}{\|v\|_{-s,\Omega_0}} \\ &= \sup_{v \in C_0^\infty(\Omega_0)} \frac{(V_{h,l-\alpha}^{2t} * \partial_h^\alpha u, v)}{\|v\|_{-s,\Omega_0}} = \sup_{v \in C_0^\infty(\Omega_0)} \frac{(\partial_h^\alpha u, V_{h,l-\alpha}^{2t} * v)}{\|v\|_{-s,\Omega_0}} \\ &\leq \sup_{v \in C_0^\infty(\Omega_0)} \frac{\|\partial_h^\alpha u\|_{s,\Omega_1} \|V_{h,l-\alpha}^{2t} * v\|_{-s,\Omega_1}}{\|v\|_{-s,\Omega_0}}. \end{aligned}$$

It is easily seen that

$$\|V_{h,l-\alpha}^{2t} * v\|_{-s,\Omega_1} \leq C \|v\|_{-s,\Omega_0}$$

where C is independent of h and $v \in C_0^\infty(\Omega_0)$. This completes the proof.

6. Superconvergence Estimates. Let u and $u_h \in S_h^r$ satisfy the interior equations (4.2). The inequalities (1.4) and (1.5) provide estimates for the error in the L_2 and maximum norm respectively. It is often the case that for specific problems the term $\|e\|_{-p,\Omega_1}$ is $O(h^r)$ and hence the estimates (1.4) and (1.5) yield $O(h^r)$ convergence locally. For many important problems $\|e\|_{-p,\Omega_1}$ is in fact $O(h^{2r-2})$. We shall show that if one considers $K_h * u_h$ instead of u_h as an approximation to u , and if $\|e\|_{-p,\Omega_1}$ is $O(h^{2r-2})$, then $u - K_h * u_h$ is locally of order h^{2r-2} in both L_2 and maximum norms. Hence in these cases processing the solution u_h yields higher order accuracy for $r \geq 3$.

We emphasize that $u - K_h * u_h$ will be $O(h^{2r-2})$ at every point of the region in which the error is measured. In practice one usually computes the solution at specific points, for example at mesh points. In this case $K_h * u_h$ takes a particularly simple form. This will also be discussed. We shall start with finding estimates for $u - K_h * u$ in L_2 norm (Theorem 1) and then in the maximum norm (Theorem 2).

THEOREM 1. *Let $\Omega_0 \subset\subset \Omega_1$, $u \in H^{2r-2}(\Omega_1)$, $u_h \in S_h^r(\Omega_1)$, $r \geq 3$, where S_h^r satisfies A.1 and A.2 and $u - u_h$ satisfies (4.2). Let K_h be chosen as in (5.1) and let p be an arbitrary but fixed nonnegative integer. Then for all h sufficiently small*

$$(6.1) \quad \|u - K_h * u_h\|_{0,\Omega_0} \leq C(h^{2r-2} \|u\|_{2r-2,\Omega_1} + \|u - u_h\|_{-p,\Omega_1}),$$

where C is independent of u and h .

Proof. By the triangle inequality

$$(6.2) \quad \|u - K_h * u_h\|_{0,\Omega_0} \leq \|u - K_h * u\|_{0,\Omega_0} + \|K_h * (u - u_h)\|_{0,\Omega_0}.$$

Let $\Omega_0 \subset\subset \Omega'_0 \subset\subset \Omega''_0 \subset\subset \Omega_1$. Since $K_h(x) = K_{h,r-2}^{2r-2}(x)$ it follows from (5.14) that

$$(6.3) \quad \|u - K_h * u\|_{0,\Omega_0} \leq Ch^{2r-2} \|u\|_{2r-2,\Omega_1}.$$

Now using (2.5), we have that

$$(6.4) \quad \|K_h * (u - u_h)\|_{0, \Omega_0} \leq \sum_{|\beta| \leq r-2} \|D^\beta K_h * (u - u_h)\|_{2-r, \Omega'_0}.$$

The inequality (6.1) now follows from (5.16), (4.3) and (6.3).

In order to prove a maximum norm error estimate we shall make use of the following.

LEMMA 6.1. *Let $\Omega_0 \subset\subset \Omega_1$, $v \in C^0(\Omega)$ and $N_0 = [N/2] + 1$. Then for all h sufficiently small*

$$(6.5) \quad |K_h * v|_{0, \Omega_0} \leq C \left[\sum_{|\alpha| \leq N_0 + r-2} \|\partial_h^\alpha v\|_{2-r, \Omega_1} + h^{r-2} \sum_{|\alpha| \leq r-2} |\partial_h^\alpha v|_{0, \Omega_1} \right]$$

where C is independent of v and h .

Proof. By the triangle inequality

$$(6.6) \quad |K_h * v|_{0, \Omega_0} \leq |K_{h, N_0}^{2r-2} * K_h * v|_{0, \Omega_0} + |K_h * v - K_{h, N_0}^{2r-2} * K_h * v|_{0, \Omega_0}$$

where $K_{h, N_0}^{2r-2} * u$ is defined by (5.12). Let $\Omega_0 \subset\subset \Omega'_0 \subset\subset \Omega_1$. Then using (2.4), (5.16), (5.15) and (2.5) we have for the first term on the right that

$$(6.7) \quad \begin{aligned} |K_{h, N_0}^{2r-2} * K_h * v|_{0, \Omega_0} &\leq C \sum_{|\alpha| \leq N_0} \|D^\alpha K_{h, N_0}^{2r-2} * K_h * v\|_{0, \Omega'_0} \\ &\leq C \left(\sum_{|\alpha| \leq N_0} \|K_h * \partial_h^\alpha v\|_{0, \Omega'_0} \right) \leq C \left(\sum_{|\alpha| \leq N_0 + r-2} \|\partial_h^\alpha v\|_{2-r, \Omega_1} \right). \end{aligned}$$

For the second term on the right of (6.6) we have

$$(6.8) \quad \begin{aligned} |K_h * v - K_{h, N_0}^{2r-2} * K_h * v|_{0, \Omega_0} &\leq Ch^{r-2} |K_h * v|_{r-2, \Omega'_0} \\ &\leq Ch^{r-2} \left(\sum_{|\alpha| \leq r-2} |\partial_h^\alpha v|_{0, \Omega'} \right), \end{aligned}$$

where we have used (5.13) and (5.17). The inequality (6.5) now follows from (6.6), (6.7) and (6.8).

THEOREM 2. *Suppose that the conditions of Theorem 1 are satisfied and that $u \in H^{2r-2+N_0}$ ($N_0 = [N/2] + 1$). Then*

$$(6.9) \quad |u - K_h * u_h|_{0, \Omega_0} \leq C(h^{2r-2} \|u\|_{2r-2+N_0, \Omega_1} + \|u - u_h\|_{-p, \Omega_1}),$$

where C is independent of u and h .

Proof. Using the triangle inequality and setting $u - u_h = e$,

$$|u - K_h * u_h|_{0, \Omega_0} \leq |u - K_h * u|_{0, \Omega_0} + |K_h * e|_{0, \Omega_0}.$$

Let $\Omega_0 \subset\subset \Omega'_0 \subset\subset \Omega_1$. Then using (5.13) and (2.4) we have

$$|u - K_h * u|_0 \leq Ch^{2r-2} |u|_{2r-2, \Omega'_0} \leq Ch^{2r-2} \|u\|_{2r-2+N_0, \Omega_1}.$$

Applying Lemma 6.1, (4.3) and (4.4)

$$\begin{aligned}
|K_h * e|_{0,\Omega_0} &\leq C \left(\sum_{|\alpha| \leq r-2+N_0} \|\partial_h^\alpha e\|_{2-r,\Omega'_0} + h^{r-2} \sum_{|\alpha| \leq r-2} |\partial_h^\alpha e|_{0,\Omega'_0} \right) \\
&\leq C(h^{2r-2} \|u\|_{2r-2+N_0,\Omega_1} + \|u - u_h\|_{-p,\Omega_1})
\end{aligned}$$

which completes the proof.

Although $K_{h,r-2}^{2r-2} * u_h$ can be calculated at arbitrary points in Ω_0 , Theorem 2 takes a particularly simple form if u_h is of the form (3.3) and if, for example, we are interested in points in Ω_0 of the form $h\gamma$, $\gamma \in \mathbf{Z}^N$.

COROLLARY. *Suppose the conditions of Theorem 2 are satisfied and u_h is of the form (3.3). Then at points $h\gamma \in \bar{\Omega}_0$, $\gamma \in \mathbf{Z}^N$*

$$\begin{aligned}
(6.10) \quad & \sup_{h\gamma \in \bar{\Omega}_0; \gamma \in \mathbf{Z}^N} \left| u(h\gamma) - \sum_{j=1}^m \sum_{\alpha} a_{\gamma-\alpha}^j d_{\alpha}^j \right| \\
&\leq C(h^{2r-2} \|u\|_{2r-2+N_0,\Omega_1} + \|u - u_h\|_{-p,\Omega_1}),
\end{aligned}$$

where the d_{α}^j are given by (5.10).

7. Examples. Here the theory given in Section 6 will be exemplified, where for simplicity we shall restrict ourselves to discussing Dirichlet's problem. In Example 1 interior superconvergence estimates for two different methods will be discussed. In Example 2 a way of obtaining superconvergence up to the boundary, when the domain Ω is the unit square and the subspaces are taken to satisfy the boundary conditions, will be given.

Let Ω be a bounded domain in \mathbf{R}^N with boundary $\partial\Omega$. For simplicity consider

$$\begin{aligned}
(7.1) \quad & -\Delta u = f \quad \text{in } \Omega, \\
& u = 0 \quad \text{on } \partial\Omega.
\end{aligned}$$

For the purposes of the applications given here an additional assumption on the subspace $S_h^r(\Omega)$ will be made. Namely we suppose that there exists a constant C independent of u and h such that for all $u \in H^t(\Omega)$; $1 \leq t \leq r$,

$$(7.2) \quad \inf_{x \in S_h^r(\Omega)} \|u - x\|_{1,\Omega} \leq Ch^{t-1} \|u\|_{t,\Omega}.$$

In Example 2 the elements of $S_h^r(\Omega)$ are required to vanish on $\partial\Omega$. In this case (7.2) is assumed to hold only for $u \in H^1(\Omega) \cap H^r(\Omega)$.

In what follows we shall assume that the hypothesis of Theorems 1 and 2 are satisfied, where $\Omega_0 \subset\subset \Omega_1 \subset\subset \Omega$.

Example 1. Dirichlet's Problem on a Smooth Domain. Here we shall assume for simplicity that $\partial\Omega \in C^\infty$. In Babuška [1] and Nitsche [14], methods were introduced for approximating the solution of (7.1) in which the approximating subspaces need not satisfy the boundary condition. These methods have the same interior equations; i.e. if u_h is the approximate solution determined by either of these methods, then

$$(7.3) \quad \begin{aligned} \int_{\Omega} \sum_{i=1}^N \frac{\partial u_h}{\partial x_i} \frac{\partial \varphi}{\partial x_i} dx &= \int_{\Omega} f \varphi dx \\ &= \int_{\Omega} \sum_{i=1}^N \frac{\partial u}{\partial x_i} \frac{\partial \varphi}{\partial x_i} dx, \quad \forall \varphi \in \mathring{S}_h^r(\Omega_1). \end{aligned}$$

What is important here is that we may choose $S_h^r(\Omega)$ to satisfy all of our conditions (cf. [4, Example 2], for more details). Now it was shown in [5] that the estimate

$$(7.4) \quad \|u - u_h\|_{2-r, \Omega} \leq h^{r-1-t} \|u\|_{t, \Omega}$$

is valid for $1 \leq t \leq r$. Using this inequality and the fact that $\|u - u_h\|_{2-r, \Omega_1} \leq \|u - u_h\|_{2-r, \Omega}$, we obtain from Theorems 1 and 2 the error estimates

$$(7.5) \quad \|u - K_h * u_h\|_{0, \Omega_0} \leq Ch^{2r-2} (\|u\|_{2r-2, \Omega_1} + \|u\|_{r, \Omega}),$$

and in the maximum norm

$$(7.6) \quad \begin{aligned} |u - K_h * u_h|_{0, \Omega_0} \\ \leq Ch^{2r-2} (\|u\|_{2r-2+\lceil N/2 \rceil+1, \Omega_1} + \|u\|_{r, \Omega}). \end{aligned}$$

If we take $N = 2$ and for example S_h^r to be piecewise cubic Hermite polynomials or cubic smooth splines, then $r = 4$ and (7.5) and (7.6) become

$$\|u - K_h * u_h\|_{0, \Omega_0} \leq Ch^6 (\|u\|_{6, \Omega_1} + \|u\|_{4, \Omega}),$$

$$|u - K_h * u_h|_{0, \Omega_0} \leq Ch^6 (\|u\|_{8, \Omega_1} + \|u\|_{4, \Omega}).$$

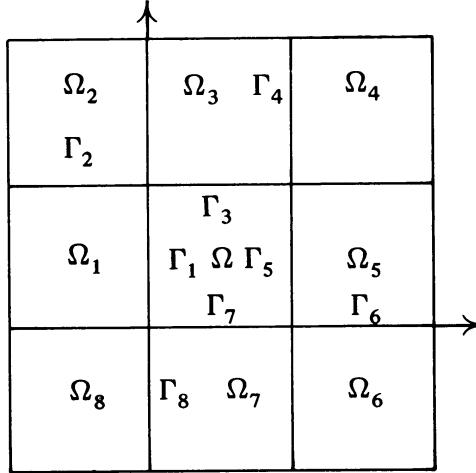
Example 2. Superconvergence Up to the Boundary for Dirichlet's Problem on a Square. Let Ω be the unit square in \mathbf{R}^2 (i.e. $\Omega = \{x \mid 0 < x_i < 1, i = 1, 2\}$), with boundary $\partial\Omega$ and let u be the solution of (7.1).

Let $h = 1/n$, $n = 1, 2, 3, \dots$, and let l be either 0 or 1. We shall approximate u taking our subspace $S_h^r(\Omega) = \bar{M}_{k,r}^h(\Omega) = \{\varphi \mid \varphi \in M_{l,r}^h(\Omega), \varphi|_{\partial\Omega} = 0, k = 0, 1\}$. That is, $\bar{M}_{k,r}^h(\Omega)$ is the subspace of the space of tensor products of one-dimensional C^0 or C^1 splines (discussed in Section 3) consisting of elements which vanish on $\partial\Omega$.

Let $u_h \in S_h^r(\Omega)$ be the approximate solution of (7.1) defined by

$$(7.7) \quad \begin{aligned} \sum_{i=1}^2 \int_{\Omega} \frac{\partial u_h}{\partial x_i} \frac{\partial \varphi}{\partial x_i} dx &= \int_{\Omega} f \varphi dx \\ &= \sum_{i=1}^2 \int_{\Omega} \frac{\partial u}{\partial x_i} \frac{\partial \varphi}{\partial x_i} dx, \quad \forall S_h^r(\Omega). \end{aligned}$$

We shall show that u_h can be extended, in a simple manner, from Ω to a larger domain so that $u - K_h * u_h = O(h^{2r-2})$ on $\bar{\Omega}$ in both the L_2 and maximum norms. The extension u_h can be described as follows:



- (i) Extend u_h from Ω to $\Omega_1, \Omega_3, \Omega_5$ and Ω_7 as an odd function by reflection across the faces $\Gamma_1, \Gamma_3, \Gamma_5$ and Γ_7 respectively.
- (ii) Then extend u_h to $\Omega_2, \Omega_4, \Omega_6$ and Ω_8 as an odd function by reflection across the faces $\Gamma_2, \Gamma_4, \Gamma_6$ and Γ_8 respectively.

Notice that since k is either 0 or 1, $U_h \in \bar{M}_{k,r}^h(\Omega^E)$, where $\Omega^E = \{x \mid -1 < x_i < 2\}$.

THEOREM 3. *Let S_h^r and U_h be defined as above with $r \geq 3$. If h is sufficiently small and*

- (i) $f \in \overset{\circ}{H}^{2r-4}(\Omega)$, then

$$(7.8) \quad \|u - K_h * U_h\|_{0,\Omega} \leq C(h^{2r-2} \|f\|_{2r-4,\Omega}).$$

- (ii) If $f \in \overset{\circ}{H}^{2r-2}(\Omega)$, then

$$(7.9) \quad |u - K_h * U_h|_{0,\Omega} \leq C(h^{2r-2} \|f\|_{2r-2,\Omega}),$$

where C is independent of h and f .

Proof. Extend u to Ω^E in the same manner as U_h and call its extension U . Let Λ_i , $i = 1, 2, 3, 4$, be the squares

$$\Lambda_1 = \{x \mid -1 < x_1 < 1, 0 < x_2 < 2\}, \quad \Lambda_2 = \{x \mid 0 < x_1 < 2, 0 < x_2 < 2\},$$

$$\Lambda_3 = \{x \mid -1 < x_1 < 1, -1 < x_2 < 1\}, \quad \Lambda_4 = \{x \mid 0 < x_1 < 2, -1 < x_2 < 1\}.$$

We first claim that

$$(7.10) \quad \sum_{j=1}^2 \int_{\Lambda_i} \frac{\partial(U - U_h)}{\partial x_j} \frac{\partial \varphi}{\partial x_j} dx = 0, \quad \forall \varphi \in \bar{M}_k^h(\Lambda_i), i = 1, 2, 3, 4.$$

Note that $U_h \in \bar{M}_k^h(\Lambda_i)$ and if $f \in \overset{\circ}{H}^s(\Omega)$ ($s \geq 0$ an integer) then $U \in H^{s+2}(\Lambda_i) \cap \overset{\circ}{H}^1(\Lambda_i)$, $i = 1, 2, 3, 4$. Furthermore

$$(7.11) \quad \|U\|_{s+2,\Lambda_i} \leq C \|f\|_{s,\Omega}, \quad i = 1, 2, 3, 4.$$

Before proving (7.10), let us show how (7.8) and (7.9) follow from it. Equation (7.10) says that U_h is the Ritz-Galerkin approximation in $\bar{M}_k^h(\Lambda_i)$ of U on each of the

domains Λ_i , $i = 1, 2, 3, 4$. Hence on each of these domains the interior estimates obtained in Theorems 1 and 2 are applicable. Let $\Omega_0^i \subset\subset \Lambda_i$, $i = 1, 2, 3, 4$, be chosen such that $\overline{\bigcup_{i=1}^4 \Omega_0^i}$ cover Ω . From (6.1) and (6.5) we have for h sufficiently small and $i = 1, 2, 3, 4$

$$(7.12) \quad \|U - K_h * U_h\|_{0, \Omega_0^i} \leq C(h^{2r-2} \|U\|_{2r-2, \Lambda_i} + \|U - U_h\|_{2-r, \Lambda_i})$$

and

$$(7.13) \quad |U - K_h * U_h|_{0, \Omega_0^i} \leq C(h^{2r-2} \|U\|_{2r, \Lambda_i} + \|U - U_h\|_{2-r, \Omega^E}).$$

It was shown in [15] that

$$(7.14) \quad \|U - U_h\|_{2-r, \Lambda_i} \leq Ch^{2r-2} \|U\|_{r, \Lambda_i}.$$

The inequalities (7.8) and (7.9) now easily follow from (7.12), (7.13), (7.14), (7.11) and the fact that $\overline{\bigcup_{i=1}^4 \Omega_0^i}$ covers Ω .

It remains to prove (7.10). We shall do so in the case $i = 1$, the other cases follow in the same manner. Let $\Lambda = \Omega \cup \Omega_1 \cup \Gamma_1$. We shall first show that

$$(7.15) \quad \sum_{i=1}^2 \int_{\Lambda} \frac{\partial(U - U_h)}{\partial x_i} \frac{\partial \varphi}{\partial x_i} dx = 0, \quad \forall \varphi \in \bar{M}_{k,r}^h(\Lambda).$$

If $\varphi \in \bar{M}_{k,r}^h(\Lambda)$, then let $\varphi = \varphi^0 + \varphi^e$ where

$$\varphi^0(x_1, x_2) = \frac{\varphi(x_1, x_2) - \varphi(-x_1, x_2)}{2}$$

and

$$\varphi^e(x_1, x_2) = \frac{\varphi(x_1, x_2) + \varphi(-x_1, x_2)}{2}.$$

Now $\varphi^0 \in \bar{M}_{k,r}^h(\Omega_1)$ and $\bar{M}_{k,r}^h(\Omega)$, hence using (7.7)

$$\sum_{i=1}^2 \int_{\Lambda} \frac{\partial(U - U_h)}{\partial x_i} \frac{\partial \varphi^0}{\partial x_i} dx = 2 \sum_{i=1}^2 \int_{\Omega} \frac{\partial(U - U_h)}{\partial x_i} \frac{\partial \varphi^0}{\partial x_i} dx = 0.$$

Since φ^e is an even function of x_1 and $U - U_h$ is an odd function of x_1 it follows that

$$\sum_{i=1}^2 \int_{\Lambda} \frac{\partial(U - U_h)}{\partial x_i} \frac{\partial \varphi^e}{\partial x_i} dx = 0,$$

which proves (7.15).

Since U and U_h on Ω_2 and Ω_3 may be obtained from U and U_h on Λ by an odd reflection about the line $y = 1$, the argument given to prove (7.15) also shows that (7.10) holds which completes the proof.

Appendix. Here we present two tables. Table 1 gives values for the weights k'_j for various values of r . For a given choice of basis functions these may be used to calculate $K_h * u_h(x)$ at each point x or just at mesh points using (5.8). In Table 2 we give values of the weights d'_j for various values of r in the case where S_h^r is chosen to be the splines generated by the B -spline basis $\psi^{(r)}$.

TABLE 1. $k'_j, l = r - 2, t = r - 1$

$j \setminus r$	3	4	5	6
0	13/12	37/30	346517/241920	76691/45360
1	-1/24	-23/180	-81329/322560	-48061/113400
2		1/90	6337/161280	20701/226800
3			-3229/967680	-1573/113400
4				479/453600

TABLE 2. d'_j

$j \setminus r$	3	4	5	6
0	51/72	673/1080	$\frac{33055739}{58060800}$	$\frac{967, 356, 037}{1, 828, 915, 200}$
1	11/72	4283/21600	$\frac{3589969}{16257024}$	$\frac{3, 841, 481, 473}{16, 460, 236, 800}$
2	-1/144	-61/5400	$\frac{-12162977}{1625702400}$	$\frac{31, 253, 191}{82, 301, 184, 000}$
3		29/21600	$\frac{4795283}{2438553600}$	$\frac{48, 179, 483}{27, 433, 728, 000}$
4		1/10800	$\frac{26273}{270950400}$	$\frac{89711}{514, 382, 400}$
5			$\frac{-58243}{812851200}$	$\frac{-2905789}{16, 460, 236, 800}$
6			$\frac{-3229}{4877107200}$	$\frac{25867}{1, 097, 349, 120}$
7				$\frac{117083}{82, 301, 184, 000}$
8				479/164, 602, 368, 000

Department of Mathematics
Cornell University
Ithaca, New York 14853

1. I. BABUŠKA, "The finite element method with Lagrangian multipliers," *Numer. Math.*, v. 20, 1973, pp. 179–192. MR 50 #11806.
2. J. BRAMBLE, "A survey of some finite element methods proposed for treating the Dirichlet problem," *Advances in Math.*, v. 16, 1975, pp. 187–196.
3. J. H. BRAMBLE & S. HILBERT, "Bounds for a class of linear functionals with applications to Hermite interpolation," *Numer. Math.*, v. 16, 1971, pp. 362–369.
4. J. H. BRAMBLE, J. A. NITSCHE & A. H. SCHATZ, "Maximum-norm interior estimates for Ritz-Galerkin methods," *Math. Comp.*, v. 29, 1975, pp. 677–688.

5. J. H. BRAMBLE & J. E. OSBORNE, "Rate of convergence estimates for nonselfadjoint eigenvalue approximations," *Math. Comp.*, v. 27, 1973, pp. 525–549. MR 51 #2280.
6. J. H. BRAMBLE & A. H. SCHATZ, "Estimates for spline projection," *Rev. Française Automat. Informat. Recherche Opérationnelle Analyse Numérique*, v. 10, n° 8, août 1976, pp. 5–37.
7. J. H. BRAMBLE & A. H. SCHATZ, "Higher order local accuracy by averaging in the finite element method," *Mathematical Aspects of Finite Elements in Partial Differential Equations* (Proc. Sympos., Math. Res. Center, Univ. of Wisconsin, Madison, 1974), edited by Carl de Boor, Academic Press, New York, 1974, pp. 1–14. MR 50 #1525.
8. J. H. BRAMBLE & M. ZLAMAL, "Triangular elements in the finite element method," *Math. Comp.*, v. 24, 1970, pp. 809–820. MR 43 #8250.
9. C. de BOOR & B. SWARTZ, "Collocation at Gaussian points," *SIAM J. Numer. Anal.*, v. 10, 1973, pp. 582–606. MR 51 #9528.
10. J. DOUGLAS, JR. & T. DUPONT, "Some superconvergence results for Galerkin methods for the approximate solution of two point boundary problems." (Preprint.)
11. J. DOUGLAS, JR., T. DUPONT & M. F. WHEELER, "Some super-convergence results for an H^1 -Galerkin procedure for the heat equation," *Computing Methods in Applied Sciences and Engineering*. Part I (Proc. Internat. Sympos., Versailles, 1973), Springer-Verlag, Berlin and New York, 1974, pp. 288–311. MR 49 #4197.
12. A. FRIEDMAN, *Partial Differential Equations*, Holt, Rinehart and Winston, New York, 1969.
13. J. L. LIONS & E. MAGENES, *Non-Homogeneous Boundary Value Problems and Applications*. Vol. I, Springer-Verlag, Berlin and New York, 1972. MR 50 #2670.
14. J. A. NITSCHE, "Über ein Variationsprinzip zur Lösung von Dirichlet-Problemen bei Verwendung von Teilräumen, die keinen Randbedingungen unterworfen sind," *Abh. Math. Sem. Univ. Hamburg*, v. 36, 1971, pp. 9–15. MR 49 #6649.
15. J. A. NITSCHE & A. H. SCHATZ, "Interior estimates for Ritz-Galerkin methods," *Math. Comp.*, v. 28, 1974, pp. 937–958. MR 51 #9525.
16. I. J. SCHOENBERG, "Contributions to the problem of approximation of equidistant data by analytic functions," Parts A, B, *Quart. Appl. Math.*, v. 4, 1946, pp. 45–99, 112–141. MR 7, 487; 8, 55.
17. V. THOMÉE, "Spline approximation and difference schemes for the heat equation," *The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations* (Proc. Sympos., Univ. of Maryland, 1972), edited by A. K. Aziz, Academic Press, New York, 1972, pp. 711–746. MR 49 #11824.
18. V. THOMÉE & B. WENDROFF, "Convergence estimates for Galerkin methods for variable coefficient initial value problems," *SIAM J. Numer. Anal.*, v. 11, 1974, pp. 1059–1068. MR 51 #7309.

**2.6 Single step Galerkin approximations for parabolic problems
(1977)**

Single step Galerkin approximations for parabolic problems[2]

**2.7 Some convergence estimates for semidiscrete Galerkin type
approximations for parabolic equations (1977)**

Some convergence estimates for semidiscrete Galerkin type approximations for parabolic equations[37]

**2.8 Semidiscrete and single step fully discrete approximations for
second order hyperbolic equations (1979)**

[1]

2.9 Some estimates for a weighted L^2 projection (1991)

Some estimates for a weighted L^2 projection[?]

SOME ESTIMATES FOR A WEIGHTED L^2 PROJECTION

JAMES H. BRAMBLE AND JINCHAO XU

ABSTRACT. This paper is devoted to the error estimates for some weighted L^2 projections. Nearly optimal estimates are obtained. These estimates can be applied to the analysis of the usual multigrid method, multilevel preconditioner and domain decomposition method for solving elliptic boundary problems whose coefficients have large jump discontinuities.

1. INTRODUCTION

This work was motivated by the study of the numerical solution of elliptic boundary value problems that have large discontinuity jumps in coefficients. If these jumps become larger, the corresponding discretized (by finite elements, for example) equation may be harder to solve. In some special cases, however, multigrid or domain decomposition methods can be properly designed so that the numerically observed convergence rate is actually independent of these jumps. We find that the theoretical justification of this phenomenon lies in certain approximation and stability properties of some weighted L^2 projections with weights provided by the discontinuous coefficients (cf. [10, 11, 4]). The point is that we want to get estimates which are uniform with respect to the weights.

A careful study of this type of weighted L^2 projection will be made in this paper. We shall establish estimates that are nearly optimal under some special circumstances. In a sequel of this paper, we shall present some negative results to demonstrate that the expected estimates are not always possible, in general, and the results in the paper are sharp in a certain sense.

Related to the topic of this paper is the usual L^2 projection. Some error and stability estimates for such a projection are also presented with complete proofs.

As is done in [10], we will use the following notation:

$$x \lesssim y, \quad f \gtrsim g, \quad \text{and} \quad u \asymp v$$

Received May 18, 1990.

1980 *Mathematics Subject Classification* (1985 Revision). Primary 65M60, 65N15, 65N30.

Key words and phrases. Finite element space, weighted L^2 projections, multigrid, domain decomposition.

This work was supported in part under the National Science Foundation Grant Nos. DMS84-05352 and 8801346-02, also by the U.S. Army Research Office through the Mathematical Sciences Institute, Cornell University.

©1991 American Mathematical Society
0025-5718/91 \$1.00 + \$.25 per page

which means that

$$x \leq Cy, \quad f \geq cg, \quad \text{and} \quad cv \leq u \leq Cv,$$

where C and c are positive constants independent of the variables appearing in the inequalities and any other parameters related to meshes, spaces, etc.

The remainder of the paper is organized as follows. In §2, some preliminary material, such as the Sobolev spaces, finite element spaces, etc., will be presented. Section 3 is devoted to the analysis of the usual L^2 projection. The main estimates for weighted L^2 projections will be presented in §4.

2. PRELIMINARIES

Let $\Omega \subset \mathbf{R}^d$ ($1 \leq d \leq 3$) be a bounded domain. For simplicity, we assume that Ω is an interval for $d = 1$, a polygon for $d = 2$, and a polyhedron for $d = 3$. On Ω , $L^p(\Omega)$ denotes the usual Banach space consisting of p th power integrable functions. The Sobolev space of index (m, p) is defined by

$$W^{m,p}(\Omega) \stackrel{\text{def}}{=} \{v \in L^p(\Omega); D^\alpha v \in L^p(\Omega) \text{ if } |\alpha| \leq m\},$$

with a norm

$$\|v\|_{W^{m,p}(\Omega)} \stackrel{\text{def}}{=} \left(\sum_{|\alpha| \leq m} \|D^\alpha v\|_{L^p(\Omega)}^p \right)^{1/p},$$

where $\alpha = (\alpha_1, \dots, \alpha_d)$ is a multi-integer and

$$D^\alpha = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \cdots x_d^{\alpha_d}}, \quad |\alpha| = \sum_{i=1}^d \alpha_i.$$

For $p = 2$, by convention, we denote

$$H^m(\Omega) \stackrel{\text{def}}{=} W^{m,2}(\Omega).$$

We will have occasion to use the following seminorms:

$$|v|_{W^{m,p}(\Omega)} \stackrel{\text{def}}{=} \left(\sum_{|\alpha|=m} \|D^\alpha v\|_{L^p(\Omega)}^p \right)^{1/p}.$$

For $m = 1$, $H_0^1(\Omega)$ denotes the subspace of $H^1(\Omega)$ consisting of functions that vanish on $\partial\Omega$ in an appropriate sense. Similarly, for a measurable $\Gamma_0 \subset \partial\Omega$, $H_{\Gamma_0}^1(\Omega)$ is the space consisting of functions in H^1 that vanish on Γ_0 .

We quote the following well-known Sobolev continuous imbeddings [1]:

$$(2.1) \quad H^1(\Omega) \hookrightarrow \begin{cases} L^\infty(\Omega), & \text{if } d = 1, \\ L^p(\Omega) \ (1 \leq p < \infty), & \text{if } d = 2, \\ L^6(\Omega), & \text{if } d = 3. \end{cases}$$

Lemma 2.1. *We have*

$$\|u\|_{L^2(\partial\Omega)} \lesssim \varepsilon^{-1} \|u\|_{L^2(\Omega)} + \varepsilon \|u\|_{H^1(\Omega)} \quad \forall u \in H^1(\Omega), \quad \varepsilon \in (0, 1).$$

For a proof, we refer to [10].

The following result is a special case of Theorem 2.1 in [10]. (Cf. also [9].)

Lemma 2.2. *Assume D is a bounded domain in \mathbf{R}^2 with ∂D Lipschitz continuous. Then*

$$\|w\|_{L^\infty(D)} \lesssim |\log \varepsilon|^{1/2} \|w\|_{H^1(D)} + \varepsilon \|w\|_{W^{1,\infty}(D)} \quad \forall w \in W^{1,\infty}(D), \quad \varepsilon \in (0, 1).$$

Next we introduce the finite element space. For $0 < h < 1$, let \mathcal{T}_h be a triangulation of $\bar{\Omega}$ with simplices K of diameter less than or equal to h . We assume the family $\{\mathcal{T}_h\}$ is quasiuniform, i.e., there are constants $c_0 > 0$ and $c_1 > 0$ such that

$$\max_{K \in \mathcal{T}_h} \frac{h_K}{\rho_K} \leq c_0, \quad \frac{\max_{K \in \mathcal{T}_h}}{\min_{K \in \mathcal{T}_h}} \leq c_1 \quad \forall h,$$

where h_K is the diameter of K and ρ_K is the diameter of the largest ball contained in K . Corresponding to each triangulation \mathcal{T}_h , we define a finite element subspace $S_h \subset H_0^1(\Omega)$ that consists of continuous piecewise (with respect to the elements in \mathcal{T}_h) linear polynomials vanishing on $\partial\Omega$.

For a given triangulation \mathcal{T}_h , we consider a finer quasiuniform mesh $\underline{\mathcal{T}}_h$ with $\underline{h} < h$ which is obtained by refining \mathcal{T}_h in such a way that

$$S_h \subset S_{\underline{h}},$$

where $S_{\underline{h}} \subset H_0^1(\Omega)$ is the corresponding finite element space defined on $\underline{\mathcal{T}}_h$.

It is well known that, for any function $v \in S_h$

$$(2.2) \quad \|v\|_{L^2(\Omega)}^2 \asymp h^d \sum_{x \in \mathcal{N}_h} v^2(x),$$

where \mathcal{N}_h is the set of vertices of the triangulation \mathcal{T}_h , and

$$(2.3) \quad \|v\|_{L^\infty(\Omega)} \lesssim h^{-d/p} \|v\|_{L^p(\Omega)} \quad (1 \leq p \leq \infty).$$

The right-hand side of (2.2) is often called the discrete L^2 norm. Inequality (2.3) is the well-known inverse property of the finite element spaces (cf. [5]).

Lemma 2.3. *For all $v \in S_h(\Omega)$,*

$$(2.4) \quad \|v\|_{L^\infty(\Omega)} \lesssim \begin{cases} \|v\|_{H^1(\Omega)}, & \text{if } d = 1, \\ |\log h|^{1/2} \|v\|_{H^1(\Omega)}, & \text{if } d = 2, \\ h^{-1/2} \|v\|_{H^1(\Omega)}, & \text{if } d = 3. \end{cases}$$

Proof. The first inequality (for $d = 1$) follows from the usual Sobolev imbedding (in (2.1)). The second is well known in the literature (cf. [3]) and can be

easily obtained by using Lemma 2.2 together with the inverse inequality (2.3) with $\varepsilon = h$.

The proof of the last case for $d = 3$ is also almost trivial and can be furnished by using the inverse inequality (2.3) and Sobolev imbedding (2.1):

$$\|v\|_{L^\infty(\Omega)} \lesssim h^{-1/2} \|v\|_{L^6(\Omega)} \lesssim h^{-1/2} \|v\|_{H^1(\Omega)}. \quad \square$$

The most interesting part in Lemma 2.3 is perhaps for $d = 2$. That is just the limiting case in which the Sobolev imbedding fails. The following is another such example.

Lemma 2.4. *Assume Ω is a polyhedral domain in \mathbf{R}^3 . Then*

$$\|v\|_{L^2(\Gamma)} \lesssim |\log h|^{1/2} \|v\|_{H^1(\Omega)} \quad \forall v \in S^h(\Omega),$$

where Γ is any edge of Ω .

Proof. By breaking the domain Ω into (possibly overlapping) subdomains that have parallel faces in one direction, we may assume here, without loss of generality, that $\Omega = (0, 1)^3$ is the unit cube. Applying the inequality in Lemma 2.2 with the domain $D = (0, 1)^2$ and $w = v(x_1, x_2, x_3)$, we get

$$\begin{aligned} |v(0, 0, x_3)|^2 &\lesssim |\log \varepsilon| \int_D \left(v^2 + \sum_{i=1}^2 \left| \frac{\partial v}{\partial x_i} \right|^2 \right) dx_1 dx_2 \\ &\quad + \varepsilon^2 \|v\|_{W^{1,\infty}(\Omega)}^2 \quad \forall \varepsilon \in (0, 1). \end{aligned}$$

Integrating with respect to x_3 , we get

$$\int_0^1 |v(0, 0, x_3)|^2 dx_3 \lesssim |\log \varepsilon| \|w\|_{H^1(\Omega)}^2 + \varepsilon^2 \|v\|_{W^{1,\infty}(\Omega)}^2 \quad \forall \varepsilon \in (0, 1).$$

Taking $\varepsilon = h^{3/2}$ and applying the inverse inequality yields

$$\|v\|_{L^2(\Gamma)} \lesssim |\log h|^{1/2} \|v\|_{H^1(\Omega)} \quad \forall v \in S^h(\Omega),$$

where $\Gamma = \{(0, 0, x_3) : 0 \leq x_3 \leq 1\}$.

Similar arguments obviously apply to the other edges of Ω , and the proof is complete. \square

3. ORDINARY L^2 PROJECTIONS

In this section, we shall consider the usual L^2 projection with respect to the ordinary L^2 inner product (namely without weights).

Associated with the finite element space S_h , the L^2 projection $Q_h : L^2(\Omega) \mapsto S_h$ is defined by

$$(Q_h u, v) = (u, v) \quad \forall u \in L^2(\Omega), \quad v \in S_h.$$

The aim of this section is to establish some estimates for Q_h on H^1 in both the L^2 and H^1 norms, namely for all $u \in H_0^1(\Omega)$

$$(3.1) \quad \|u - Q_h u\|_{L^2(\Omega)} \lesssim h|u|_{H^1(\Omega)}$$

and

$$(3.2) \quad |Q_h u|_{H^1(\Omega)} \lesssim |u|_{H^1(\Omega)}.$$

The above estimates are closely related to the so-called simultaneous approximation property:

$$(3.3) \quad \inf_{\chi \in S_h} (\|u - \chi\|_{L^2(\Omega)} + h\|u - \chi\|_{H^1(\Omega)}) \lesssim h\|u\|_{H^1(\Omega)} \quad \forall u \in H_0^1(\Omega).$$

More specifically, we have

Lemma 3.1. (3.1) and (3.2) both hold if and only if (3.3) is true.

The proof, which uses the triangle inequality and also the inverse property, is straightforward.

Inequality (3.3) has been assumed in some papers on finite elements, but it seems that little attention is paid to its proof. The stability of the L^2 projection in the H^1 norm was perhaps first established by Bank and Dupont in [2]. Their proof, however, requires the full elliptic regularity condition (which is unnecessary). A discussion of this problem in two dimensions may also be found in Crouzeix and Thomée [6]. Recently, Scott and Zhang [8] have constructed a kind of interpolation operator for nonsmooth functions that can also be used to give a proof of this result. As we pointed out earlier, it can be directly obtained by assuming the simultaneous approximation property (3.3), which is actually the approach that Mandel, McCormick, and Bank take in [7]. For avoiding a logical circle, the question remains as to how the simultaneous approximation property is justified. Our approach here is to establish the stability by a different argument and obtain the simultaneous approximation property as a consequence.

L^2 error estimates. As we have assumed that $d \leq 3$, the Sobolev imbedding $H^2(\Omega) \hookrightarrow C(\bar{\Omega})$ holds. Therefore, the usual nodal value interpolant $I_h: C(\bar{\Omega}) \mapsto S_h$ is well defined in H^2 . It is well known that (cf. [5])

$$(3.4) \quad \|u - Q_h u\|_{L^2(\Omega)} \leq \|u - I_h u\|_{L^2(\Omega)} \lesssim h^2 |u|_{H^2(\Omega)} \quad \forall u \in H^2(\Omega) \cap H_0^1(\Omega).$$

On the other hand,

$$(3.5) \quad \|u - Q_h u\|_{L^2(\Omega)} \leq \|u\|_{L^2(\Omega)} \quad \forall u \in L^2(\Omega).$$

An application of the standard interpolation technique to the above two estimates yields

Theorem 3.2. For $u \in H_0^1(\Omega)$,

$$(3.6) \quad \|u - Q_h u\|_{L^2(\Omega)} \lesssim h |u|_{H^1(\Omega)}.$$

H^1 stability. The main ingredient in our analysis is a local L^2 projection $Q_\tau: L^2(\tau) \mapsto \mathcal{P}_1(\tau)$, for any given $\tau \in \mathcal{T}_h$, defined by

$$(Q_\tau u, \phi)_{L^2(\tau)} = (u, \phi)_{L^2(\tau)} \quad \forall u \in L^2(\tau), \phi \in \mathcal{P}_1(\tau).$$

Let $\hat{\tau}$ be the standard reference element, so that for any $\tau \in \mathcal{T}_h$ we have an affine diffeomorphism $F_\tau: \hat{\tau} \mapsto \tau$. For any function $v \in L^2(\tau)$, we adopt the following standard notation:

$$\hat{v}(\hat{x}) = v(F(\hat{x})), \quad \hat{x} \in \hat{\tau}.$$

If $Q_{\hat{\tau}}$ is defined similarly, it is then straightforward to verify that

$$(3.7) \quad \widehat{Q}_\tau u = Q_{\hat{\tau}} \hat{u}.$$

These locally defined operators have the desired stability and approximation properties, as shown by

Lemma 3.3. *For any $\tau \in \mathcal{T}_h$,*

$$(3.8) \quad |Q_\tau u|_{H^1(\tau)} \lesssim |u|_{H^1(\tau)} \quad \forall u \in H^1(\tau),$$

and

$$(3.9) \quad \|u - Q_\tau u\|_{L^2(\tau)} \lesssim h|u|_{H^1(\tau)} \quad \forall u \in H^1(\tau).$$

Proof. It follows from (3.7) that (3.8) is equivalent to

$$(3.10) \quad |Q_{\hat{\tau}} \hat{u}|_{H^1(\hat{\tau})} \lesssim |\hat{u}|_{H^1(\hat{\tau})} \quad \forall \hat{u} \in H^1(\hat{\tau}).$$

As all the norms on $\mathcal{P}_1(\hat{\tau})$ are equivalent, we have

$$|Q_{\hat{\tau}} \hat{u}|_{H^1(\hat{\tau})} \lesssim \|Q_{\hat{\tau}} \hat{u}\|_{L^2(\hat{\tau})} \leq \|\hat{u}\|_{L^2(\hat{\tau})} \lesssim \|\hat{u}\|_{H^1(\hat{\tau})},$$

which, since $Q_{\hat{\tau}} \hat{c} = \hat{c}$ for any $\hat{c} \in \mathbf{R}^2$, implies that

$$|Q_{\hat{\tau}} \hat{u}|_{H^1(\hat{\tau})} \lesssim \inf_{\delta \in \mathbf{R}^1} \|\hat{u} + \delta\|_{H^1(\hat{\tau})} \lesssim |\hat{u}|_{H^1(\hat{\tau})}.$$

This proves (3.10) and hence (3.8).

Now we turn to the proof of (3.9). By changing variables and using (3.7), we get

$$\begin{aligned} \|u - Q_\tau u\|_{L^2(\tau)} &\lesssim h^{d/2} \|\hat{u} - \hat{Q}_{\hat{\tau}} \hat{u}\|_{L^2(\hat{\tau})} \\ &\lesssim h^{d/2} \inf_{\hat{c} \in \mathbf{R}^1} \|\hat{u} + \hat{c}\|_{H^1(\hat{\tau})} \lesssim h^{d/2} |\hat{u}|_{H^1(\hat{\tau})} \\ &\lesssim h^{d/2} h^{1-d/2} |u|_{H^1(\tau)} \lesssim h |u|_{H^1(\tau)}. \end{aligned}$$

This completes the proof. \square

We are now in a position to state and prove our stability theorem.

Theorem 3.4. *For all $u \in H_0^1(\Omega)$,*

$$(3.11) \quad |Q_h u|_{H^1(\Omega)} \lesssim |u|_{H^1(\Omega)}.$$

Proof. It follows from the inverse inequality, Lemma 3.2, and Lemma 3.3, that

$$\begin{aligned} |Q_h u|_{H^1(\Omega)}^2 &= \sum_{\tau \in \mathcal{T}_h} |Q_h u|_{H^1(\tau)}^2 \leq 2 \sum_{\tau \in \mathcal{T}_h} \{|Q_h u - Q_\tau u|_{H^1(\tau)}^2 + |Q_\tau u|_{H^1(\tau)}^2\} \\ &\lesssim \sum_{\tau \in \mathcal{T}_h} \{h^{-2} |Q_h u - Q_\tau u|_{L^2(\tau)}^2 + |u|_{H^1(\tau)}^2\} \\ &\lesssim \sum_{\tau \in \mathcal{T}_h} \{h^{-2} \|u - Q_\tau u\|_{L^2(\tau)}^2 + |u|_{H^1(\tau)}^2\} + h^{-2} \|u - Q_h u\|_{L^2(\Omega)}^2 \\ &\lesssim |u|_{H^1(\Omega)}^2. \end{aligned}$$

The desired result then follows. \square

Remark 3.1. Notice that our proof of (3.11), which uses Q_τ , is carried out element-by-element. Such a “local” argument is crucial for us to establish the corresponding stability for the weighted L^2 projection (with trivial modification).

Simultaneous approximation properties. From the estimates we derived for Q_h , property (3.3) then becomes clear by Lemma 3.1. In fact, this simultaneous approximation property holds for more general boundary conditions. For example, if $\Gamma_0 \subset \partial\Omega$ is measurable, then we have

Proposition 3.5. *For any $u \in H_{\Gamma_0}^1(\Omega)$, there exists $v_h \in S_h \cap H_{\Gamma_0}^1(\Omega)$ such that (3.3) holds.*

4. WEIGHTED L^2 PROJECTION

This section, which is the core of the paper, is devoted to the analysis of the weighted L^2 projections. Both the L^2 error estimates and H^1 stability will be investigated.

Assume the domain Ω admits the following decomposition:

$$(4.1) \quad \overline{\Omega} = \bigcup_{i=1}^J \overline{\Omega}_i,$$

where the Ω_i are mutually disjoint. Let Γ denote the set of interfaces, i.e., $\Gamma = \bigcup_{i=1}^J \partial\Omega_i \setminus \partial\Omega$. For simplicity, we assume that Γ consists only of segments ($d = 2$) or plane polygons ($d = 3$). In other words, no part of any $\partial\Omega_i$ is curved.

Given a set of positive constants $\{\omega_i\}_{i=1}^J$, we introduce the following weighted inner products:

$$(4.2) \quad (u, v)_{L_\omega^2(\Omega)} = \sum_{i=1}^J \omega_i (u, v)_{L^2(\Omega_i)},$$

and

$$(4.3) \quad (u, v)_{H_\omega^1(\Omega)} = \sum_{i=1}^J \omega_i \int_{\Omega_i} \nabla u \cdot \nabla v \, dx,$$

with the induced norms denoted by $\|\cdot\|_{L_\omega^2(\Omega)}$ and $|\cdot|_{H_\omega^1(\Omega)}$, respectively. Moreover, we define a full weighted H^1 norm by

$$\|\cdot\|_{H_\omega^1(\Omega)}^2 = \|\cdot\|_{L_\omega^2(\Omega)}^2 + |\cdot|_{H_\omega^1(\Omega)}^2.$$

We assume that Ω is triangulated by a family of quasiuniform meshes $\{\mathcal{T}_h, h < 1\}$, as described earlier. An additional assumption we make here is that these triangulations will be lined up with the subdomains Ω_i 's. Namely, the restriction of each \mathcal{T}_h on each Ω_i is also a triangulation of Ω_i itself.

The weighted L^2 projection $Q_h^\omega: L^2(\Omega) \mapsto S_h$ is defined by

$$(4.4) \quad (Q_h^\omega u, v)_{L_\omega^2(\Omega)} = (u, v)_{L_\omega^2(\Omega)} \quad \forall u \in L^2(\Omega), v \in S_h.$$

We will derive error estimates for Q_h^ω of the following type:

$$\|(I - Q_h^\omega)u\|_{L_\omega^2(\Omega)} \leq Ch|\log h|^\gamma |u|_{H_\omega^1(\Omega)} \quad \forall u \in H_0^1(\Omega)$$

for some positive constant γ . The point here is that we require that the constant C appearing in the above estimate does not depend on the weights $\{\omega_i\}$. Again, we will use the notation “ \lesssim ” in place of “ $\leq C$ ”, where C is in particular independent of the weights.

The derivation of such an estimate is not as simple as it might appear. For example, the argument used in the proof of Theorem 3.2 cannot be applied easily here, even though we can get the estimates analogous to (3.4) and (3.5) with proper weights. It is unclear if the interpolation between weighted H^2 and L^2 spaces would give rise to the right space. Nevertheless, the proof for $d = 1$ is almost trivial. To be more precise, we have the following

Proposition 4.1. *For $d = 1$, we have for all $u \in H_0^1(\Omega)$*

$$\|(I - Q_h^\omega)u\|_{L_\omega^2(\Omega)} \lesssim h|u|_{H_\omega^1(\Omega)}$$

and

$$|Q_h^\omega u|_{H_\omega^1(\Omega)} \lesssim |u|_{H_\omega^1(\Omega)}.$$

Proof. Since $d = 1$, we have $H^1(\Omega) \hookrightarrow C(\overline{\Omega})$. Hence the nodal value interpolant $I_h: C(\overline{\Omega}) \mapsto S_h$ is well defined in H^1 . It is well known that (cf. [5]), for any $\tau \in \mathcal{T}_h$,

$$\|u - I_h u\|_{L^2(\tau)}^2 \lesssim h^2 |u|_{H^1(\tau)}^2 \quad \forall u \in H^1(\tau).$$

Summing up over all $\tau \in \mathcal{T}_h$ with proper weights, we then get

$$\|(I - I_h)u\|_{L_\omega^2(\Omega)} \leq Ch|u|_{H_\omega^1(\Omega)} \quad \forall u \in H_0^1(\Omega).$$

Our first inequality then follows, since $\|(I - Q_h^\omega)u\|_{L_\omega^2(\Omega)} \leq \|(I - I_h)u\|_{L_\omega^2(\Omega)}$. The proof of the second inequality is similar to Theorem 3.4 by Lemma 3.3. \square

The above approach cannot, in general, be extended to higher dimensions because of the lack of the imbedding $H^1(\Omega) \hookrightarrow C(\overline{\Omega})$, although a similar technique can be applied, as is done in §4.2.1 below in some special circumstances

when $d = 2$. The analysis for more general cases, especially for $d = 3$, is more complicated, and special techniques are needed.

4.1. The case of no internal cross point. By internal cross points we mean those points on Γ that belong to more than two $\bar{\Omega}_i$'s. If there is no such point on the interface, the analysis becomes very simple, and optimal estimates can be derived.

We shall first present a lemma that shows that the estimate we need can be reduced to the estimates on interfaces. To do this, let us introduce a weighted inner product on $L^2(\Gamma)$:

$$(u, v)_{L_\omega^2(\Gamma)} = \sum_{i=1}^J \int_{\partial\Omega_i \setminus \partial\Omega} \omega_i uv \, dx.$$

Denoting $S_h(\Gamma) \stackrel{\text{def}}{=} \{v|_\Gamma : v \in S_h\}$, let $P_\Gamma : L^2(\Gamma) \mapsto S_h(\Gamma)$ be the orthogonal projection with respect to $(\cdot, \cdot)_{L_\omega^2(\Gamma)}$.

Lemma 4.2. *For all $u \in H_0^1(\Omega)$,*

$$\|(I - Q_h^\omega)u\|_{L_\omega^2(\Omega)} \lesssim h|u|_{H_\omega^1(\Omega)} + h^{1/2}\|u - P_\Gamma u\|_{L_\omega^2(\Gamma)}.$$

Proof. On each domain Ω_i , by Proposition 3.5, there exists a $w_i \in S_h(\Omega_i)$ such that

$$(4.5) \quad \|u - w_i\|_{L^2(\Omega_i)}^2 + h^2\|u - w_i\|_{H^1(\Omega_i)}^2 \lesssim h^2\|u\|_{H^1(\Omega_i)}^2.$$

Let $w \in S_h$ be such that

$$w = \begin{cases} w_i, & \text{at the nodes in } \Omega_i, \\ P_\Gamma u, & \text{on } \Gamma. \end{cases}$$

Therefore, using (2.2),

$$\begin{aligned} \|u - w\|_{L_\omega^2(\Omega)}^2 &\lesssim \sum_{i=1}^J \omega_i \|u - w_i\|_{L^2(\Omega_i)}^2 + h^2 \sum_{i=1}^J \omega_i \sum_{p \in \Gamma_i} |(w - P_\Gamma u)(p)|^2 \\ &\lesssim \sum_{i=1}^J \omega_i \|u - w_i\|_{L^2(\Omega_i)}^2 + h\|w - P_\Gamma u\|_{L_\omega^2(\Gamma)}^2 \\ &\lesssim \sum_{i=1}^J \omega_i h^2\|u\|_{H^1(\Omega_i)}^2 + h\|u - P_\Gamma u\|_{L_\omega^2(\Gamma)}^2 \\ &\lesssim h^2|u|_{H_\omega^1(\Omega)}^2 + h\|u - P_\Gamma u\|_{L_\omega^2(\Gamma)}^2. \end{aligned}$$

The desired result then follows, since

$$\|u - Q_h^\omega u\|_{L_\omega^2(\Omega)} \leq \|u - w\|_{L_\omega^2(\Omega)}. \quad \square$$

From the above proof, we see that the validity of Lemma 4.2 has nothing to do with cross points. Nevertheless, we only know its application to the case that the interface has no internal cross points.

Theorem 4.3. *Assume the decomposition (4.1) has no internal cross points. Then, for all $u \in H_0^1(\Omega)$,*

$$(4.6) \quad \|(I - Q_h^\omega)u\|_{L_\omega^2(\Omega)} \lesssim h|u|_{H_\omega^1(\Omega)}$$

and

$$(4.7) \quad |Q_h^\omega u|_{H_\omega^1(\Omega)} \lesssim |u|_{H_\omega^1(\Omega)}.$$

Proof. Define a function $\phi \in S_h(\Gamma)$ by $\phi = P_{\Gamma_i} u$, on each Γ_i , where P_{Γ_i} is the orthogonal L^2 projection from $L^2(\Gamma_i)$ to the restriction of S_h to Γ_i . By the hypothesis that Γ has no cross point, ϕ is well defined. Note that on each Γ_i we have

$$\|u - \phi\|_{L^2(\Gamma_i)} \leq \|u - w_i\|_{L^2(\Gamma_i)}.$$

By Lemma 2.1,

$$\|u - w_i\|_{L^2(\Gamma_i)}^2 \lesssim h^{-1} \|u - w_i\|_{L^2(\Omega_i)}^2 + h \|u - w_i\|_{H^1(\Omega_i)}^2.$$

Hence,

$$\begin{aligned} h \|u - w_i\|_{L^2(\Gamma_i)}^2 &\lesssim \|u - w_i\|_{L^2(\Omega_i)}^2 + h^2 \|u - w_i\|_{H^1(\Omega_i)}^2 \\ &\lesssim h^2 |u|_{H^1(\Omega_i)}^2. \end{aligned}$$

Consequently,

$$\begin{aligned} h \|u - P_\Gamma u\|_{L_\omega^2(\Omega)}^2 &\lesssim h \sum_{i=1}^J \omega_i \|u - \phi\|_{L^2(\Gamma_i)}^2 \\ &\lesssim h^2 \sum_{i=1}^J \omega_i |u|_{H^1(\Omega_i)}^2 = h^2 |u|_{H_\omega^1(\Omega)}^2. \end{aligned}$$

Applying Lemma 4.2 gives (4.6).

The proof of (4.7) is identical to that of (3.11). This completes the proof. \square

4.2. General case. When the interface has some internal cross points, the problem becomes somewhat more subtle. We will derive certain estimates under some special circumstances.

4.2.1. Estimates for “finer” finite element functions. For $d = 2$, the embedding $H^1(\Omega) \hookrightarrow C(\bar{\Omega})$ is not true in general, but it is “almost right” for the functions in finite element subspaces, as is indicated by the second inequality in Lemma 2.3. This observation is the main motivation for the result in this subsection, and the argument is similar to that used in the proof of Proposition 4.1.

Lemma 4.4. *For any $u \in S_{\underline{h}}$ and $\tau \in \mathcal{T}_h$,*

$$\|(I - I_h)u\|_{L^2(\tau)} \lesssim \begin{cases} h \left(\log \frac{h}{\underline{h}} \right)^{1/2} |u|_{H^1(\tau)} & \text{if } d = 2, \\ h \left(\frac{h}{\underline{h}} \right)^{1/2} |u|_{H^1(\tau)} & \text{if } d = 3, \end{cases}$$

where $I_h: S_{\underline{h}} \mapsto S_h$ is the interpolation operator.

Proof. We first consider the case that $d = 2$. Let $\hat{\tau}$ be the standard reference element; then

$$\|(I - I_h)u\|_{L^2(\tau)} \lesssim h\|(\hat{I} - \hat{I}_h)\hat{u}\|_{L^2(\hat{\tau})}.$$

It follows from the discrete Sobolev inequality in Lemma 2.3 for $d = 2$ that

$$\|(\hat{I} - \hat{I}_h)\hat{u}\|_{L^2(\hat{\tau})} \leq 2\|\hat{u}\|_{L^\infty(\hat{\tau})} \lesssim \left(\log \frac{h}{\underline{h}}\right)^{1/2} \|\hat{u}\|_{H^1(\hat{\tau})}.$$

Replacing \hat{u} by $\hat{u} + c$ for any constant c , we have

$$\begin{aligned} \|(\hat{I} - \hat{I}_h)\hat{u}\|_{L^2(\hat{\tau})} &\lesssim \left(\log \frac{h}{\underline{h}}\right)^{1/2} \inf_{c \in \mathbb{R}^1} \|\hat{u} + c\|_{H^1(\hat{\tau})} \\ &\lesssim \left(\log \frac{h}{\underline{h}}\right)^{1/2} |\hat{u}|_{H^1(\hat{\tau})} \lesssim \left(\log \frac{h}{\underline{h}}\right)^{1/2} |u|_{H^1(\tau)}. \end{aligned}$$

The desired result for $d = 2$ then follows. The proof for $d = 3$ only differs in the type of Sobolev inequality (in (2.1)) used, and the details obviously need not be repeated here. This completes the proof. \square

As a direct consequence of Lemma 4.4, we have

Theorem 4.5. *For any $u \in S_h$,*

$$\|(I - Q_h^\omega)u\|_{L_\omega^2(\Omega)} \lesssim \begin{cases} h\left(\log \frac{h}{\underline{h}}\right)^{1/2} |u|_{H_\omega^1(\Omega)} & \text{if } d = 2, \\ h\left(\frac{h}{\underline{h}}\right)^{1/2} |u|_{H_\omega^1(\Omega)} & \text{if } d = 3 \end{cases}$$

and

$$|Q_h^\omega u|_{H_\omega^1(\Omega)} \lesssim \begin{cases} \left(\log \frac{h}{\underline{h}}\right)^{1/2} |u|_{H_\omega^1(\Omega)} & \text{if } d = 2, \\ \left(\frac{h}{\underline{h}}\right)^{1/2} |u|_{H_\omega^1(\Omega)} & \text{if } d = 3. \end{cases}$$

4.2.2. Estimates for general H^1 functions. In this subsection, we shall derive some estimates for functions in H^1 .

The following lemma shows that nearly optimal estimates can be obtained in general if the full weighted H^1 norms are used.

Lemma 4.6. *For all $u \in H_0^1(\Omega)$,*

$$(4.8) \quad \|(I - Q_h^\omega)u\|_{L_\omega^2(\Omega)} \lesssim h|\log h|^{1/2} \|u\|_{H_\omega^1(\Omega)}.$$

Proof. The proof will be carried out separately for different dimensions, even though the ideas in both cases are quite similar.

Case 1 : $d = 2$. Let w_i be as in the proof of Lemma 4.2 and define $w \in S_h$ by

$$w = \begin{cases} w_i, & \text{at the nodes inside } \Omega_i, \\ P_e u, & \text{at the nodes inside } e \subset \partial \Omega_i, \\ 0, & \text{at nodes elsewhere,} \end{cases}$$

where $e \subset \partial\Omega_i$ is any edge of Ω_i and $P_e: L^2(e) \mapsto S_h(e)$ is the orthogonal $L^2(e)$ projection.

By (2.2),

$$\begin{aligned} \|w_i - w\|_{L^2(\Omega_i)}^2 &\lesssim h^2 \sum_{x \in \partial\Omega_i} (w_i - w)^2(x) \lesssim h^2 \sum_{e \subset \partial\Omega_i} \sum_{x \in e} (w_i - w)^2(x) \\ &\lesssim h^2 \sum_{e \subset \partial\Omega_i} \left(\sum_{x \in e} (w_i - P_e u)^2(x) + \sum_{x \in \partial e} w_i^2(x) \right). \end{aligned}$$

We need to bound the two terms appearing in the last expression above. The first term is easy:

$$\begin{aligned} \sum_{e \subset \partial\Omega_i} h \|w_i - P_e u\|_{L^2(e)}^2 &\lesssim h \|u - w_i\|_{L^2(\partial\Omega_i)}^2 \\ &\lesssim \|u - w_i\|_{L^2(\Omega_i)}^2 + h^2 \|u - w_i\|_{H^1(\partial\Omega_i)}^2 \\ &\lesssim h^2 \|u\|_{H^1(\Omega_i)}^2, \end{aligned}$$

where we have used Lemma 2.1 and (4.5).

The second term can be bounded by the second discrete Sobolev inequality in Lemma 2.3:

$$\sum_{e \subset \partial\Omega_i} h^2 \sum_{x \in \partial e} w_i^2(x) \lesssim h^2 |\log h| \|w_i\|_{H^1(\Omega_i)}^2 \lesssim h^2 |\log h| \|u\|_{H^1(\Omega_i)}^2.$$

Consequently,

$$\|w - w_i\|_{L^2(\Omega_i)} \lesssim h |\log h|^{1/2} \|u\|_{H^1(\Omega_i)}.$$

Applying the above estimate, with the triangle inequality and (4.5), we get

$$(4.9) \quad \|u - w\|_{L^2(\Omega_i)} \leq \|u - w_i\|_{L^2(\Omega_i)} + \|w_i - w\|_{L^2(\Omega_i)} \lesssim h |\log h|^{1/2} \|u\|_{H^1(\Omega_i)}.$$

The desired result for $d = 2$ then follows.

Case 2 : $d = 3$. Again, let ω_i be as in the proof of Lemma 4.2 and define $w \in S_h$ by

$$w = \begin{cases} \omega_i, & \text{at the nodes inside } \Omega_i, \\ P_F u, & \text{at the nodes inside } F \subset \partial\Omega_i, \\ 0, & \text{at nodes elsewhere,} \end{cases}$$

where $F \subset \partial\Omega_i$ is any face of Ω_i and $P_F: L^2(F) \mapsto S_h(F)$ is the orthogonal $L^2(F)$ projection. Then

$$\begin{aligned} \|w_i - w\|_{L^2(\Omega_i)}^2 &\lesssim h^3 \sum_{x \in \partial\Omega_i} (w_i - w)^2(x) \lesssim h^3 \sum_{F \subset \partial\Omega_i} \sum_{x \in F} (w_i - w)^2(x) \\ &\lesssim h^3 \sum_{F \subset \partial\Omega_i} \left(\sum_{x \in F} (w_i - P_F u)^2(x) + \sum_{x \in \partial F} w_i^2(x) \right) \\ &\lesssim \sum_{F \subset \partial\Omega_i} (h \|w_i - P_F u\|_{L^2(F)}^2 + h^2 \|w_i\|_{L^2(\partial F)}^2). \end{aligned}$$

Again, we need to bound two terms appearing in the last expression above. For the first term, we have

$$\begin{aligned} \sum_{F \subset \partial\Omega_i} h \|w_i - P_F u\|_{L^2(F)}^2 &\lesssim h \|u - w_i\|_{L^2(\partial\Omega_i)}^2 \\ &\lesssim \|u - w_i\|_{L^2(\Omega_i)}^2 + h^2 \|u - w_i\|_{H^1(\partial\Omega_i)}^2 \\ &\lesssim h^2 \|u\|_{H^1(\Omega_i)}^2, \end{aligned}$$

where we have used Lemma 2.1 and (4.5).

The second term can be bounded by the discrete Sobolev inequality in Lemma 2.4:

$$\sum_{F \subset \partial\Omega_i} h^2 \|w_i\|_{L^2(\partial F)}^2 \lesssim h^2 |\log h| \|u\|_{H^1(\Omega_i)}^2 \lesssim h^2 |\log h| \|u\|_{H^1(\Omega_i)}^2.$$

Consequently,

$$\|w - w_i\|_{L^2(\Omega_i)} \lesssim h |\log h|^{1/2} \|u\|_{H^1(\Omega_i)}.$$

As in (4.9), the estimate for $d = 3$ follows. This completes the proof. \square

As a direct consequence of the above lemma, we have

Theorem 4.7. *If for all i , the $(d-1)$ -dimensional Lebesgue measure of $(\partial\Omega_i \cap \partial\Omega)$ is positive, then for all $u \in H_0^1(\Omega)$*

$$(4.10) \quad \|(I - Q_h^\omega)u\|_{L_\omega^2(\Omega)} \lesssim h |\log h|^{1/2} |u|_{H_\omega^1(\Omega)}$$

and

$$(4.11) \quad |Q_h^\omega u|_{H_\omega^1(\Omega)} \lesssim |\log h|^{1/2} |u|_{H_\omega^1(\Omega)}.$$

Proof. By hypothesis, we have $\|u\|_{H_\omega^1(\Omega)} \lesssim |u|_{H_\omega^1(\Omega)}$ by the Poincaré inequality. The estimate (4.10) then follows from (4.8). The estimate (4.11) can be proved similarly as (3.11) by using (4.10). \square

Remark 4.1. The assumption concerning the measure of $\text{meas}_{d-1}(\partial\Omega_i \cap \partial\Omega)$ in Theorem 4.7 cannot be removed, in general, and the deterioration $h^{-1/2}$ in the estimate of Theorem 4.5 is also best possible. All these issues will be discussed in a separate paper.

BIBLIOGRAPHY

1. R. Adams, *Sobolev spaces*, Academic Press, New York, 1975.
2. R. E. Bank and T. F. Dupont, *An optimal order process for solving elliptic finite element equations*, Math. Comp. **36** (1981), 967–975.
3. J. Bramble, *A second order finite difference analog of the first biharmonic boundary value problem*, Numer. Math. **9** (1966), 236–249.
4. J. Bramble, J. Pasciak, J. Wang, and J. Xu, *Convergence estimates for multigrid algorithms without regularity assumptions*, Math. Comp. **57** (1991) (to appear).
5. P. G. Ciarlet, *The finite element method for elliptic problems*, North-Holland, New York, 1978.

6. M. Crouzeix and V. Thomée, *The stability in L_p and W_p^1 of the L_2 -projection onto finite element function spaces*, Math. Comp. **48** (1987), 521–532.
7. J. Mandel, S. F. McCormick, and R. Bank, *Variational multigrid theory*, Multigrid Methods, (S. McCormick, ed.), SIAM, Philadelphia, 1988, pp. 131–178.
8. R. Scott and S. Zhang, *Finite element interpolation of nonsmooth functions satisfying boundary conditions*, Math. Comp. **54** (1990), 483–493.
9. J. Xu, *Error estimates and improved algorithms for infinite elements*, Proc. DD5 Internat. Conf., Beijing, China, 1984.
10. ——, *Theory of multilevel methods*, Ph.D. thesis, Cornell, 1989; AM report 48, Dept. of Math., Penn State Univ., July 1989.
11. ——, *Iterative methods by space decomposition and subspace correction: a unifying approach* (in preparation).

DEPARTMENT OF MATHEMATICS, CORNELL UNIVERSITY, ITHACA, NEW YORK 14853

E-mail address: bramble@mssun7.msi.cornell.edu

DEPARTMENT OF MATHEMATICS, PENN STATE UNIVERSITY, UNIVERSITY PARK, PENNSYLVANIA
16802

E-mail address: xu@math.psu.edu

2.10 A finite element method for interface problems in domains with smooth boundaries and interfaces (1996)

A finite element method for interface problems in domains with smooth boundaries and interfaces[13]

2.11 On the stability of the L2 projection in $H^1(\Omega)$ (2002)

On the stability of the L2 projection in $H^1(\Omega)$ [?]

ON THE STABILITY OF THE L^2 PROJECTION IN $H^1(\Omega)$

JAMES H. BRAMBLE, JOSEPH E. PASCIAK, AND OLAF STEINBACH

ABSTRACT. We prove the stability in $H^1(\Omega)$ of the L^2 projection onto a family of finite element spaces of conforming piecewise linear functions satisfying certain local mesh conditions. We give explicit formulae to check these conditions for a given finite element mesh in any number of spatial dimensions. In particular, stability of the L^2 projection in $H^1(\Omega)$ holds for locally quasiuniform geometrically refined meshes as long as the volume of neighboring elements does not change too drastically.

1. INTRODUCTION

Let $\{\varphi_k\}_{k=1}^M$ denote the nodal basis for a piecewise linear (continuous) finite element approximation space V_h based on a conforming triangulation (of simplices) $\{\tau_l\}_{l=1}^N$ of a polyhedral domain Ω in \mathbf{R}^n , $n = 1, 2, \dots$. We shall always assume that the triangulation is locally (but not globally) quasiuniform. The L^2 projection Q_h of a given function u onto the finite element space V_h is defined by

$$(1.1) \quad \langle Q_h u, v^h \rangle_{L^2(\Omega)} = \langle u, v^h \rangle_{L^2(\Omega)} \quad \text{for all } v^h \in V_h.$$

Here $\langle \cdot, \cdot \rangle_{L^2(\Omega)}$ denotes the inner product on $L^2(\Omega)$. The L^2 projection is obviously bounded on $L^2(\Omega)$; indeed, $\|Q_h u\|_{L^2(\Omega)} \leq \|u\|_{L^2(\Omega)}$ holds for all $u \in L^2(\Omega)$. In this paper we are concerned with the stability of the L^2 projection as a map $Q_h : H^1(\Omega) \rightarrow V_h \subset H^1(\Omega)$. In particular, we will prove the stability estimate

$$(1.2) \quad \|Q_h u\|_{H^1(\Omega)} \leq c \cdot \|u\|_{H^1(\Omega)} \quad \text{for all } u \in H^1(\Omega)$$

under certain conditions on the finite element trial space V_h , specifically, on the underlying triangulation.

The stability of Q_h in $H^1(\Omega)$ is of general interest, in particular, for Galerkin finite or boundary element methods for elliptic and parabolic boundary value problems [4, 6, 13]. For example, the stability estimate (1.2) is needed to analyze the properties of a Neumann series corresponding to a second kind boundary integral equation, see [9].

Using interpolation arguments we get

$$(1.3) \quad \|Q_h u\|_{H^s(\Omega)} \leq c \cdot \|u\|_{H^s(\Omega)} \quad \text{for all } u \in H^s(\Omega), \quad s \in [0, 1].$$

Received by the editor February 11, 2000 and, in revised form, May 24, 2000.

2000 *Mathematics Subject Classification*. Primary 65D05, 65N30, 65N50.

Key words and phrases. L^2 projection, finite elements, stability, adaptivity.

This work was supported by the National Science Foundation under grants numbered DMS-9626567 and DMS-9973328 and by the State of Texas under ARP/ATP grant #010366-168. This work was done while the third author was a Postdoctoral Research Associate at the Institute for Scientific Computation (ISC), Texas A & M University. The financial support by the ISC is gratefully acknowledged.

From (1.3) we can conclude the stability estimate (see for example [7])

$$(1.4) \quad c \cdot \|u^h\|_{H^s(\Omega)} \leq \sup_{v^h \in V_h} \frac{|\langle u^h, v^h \rangle_{L^2(\Omega)}|}{\|v^h\|_{\tilde{H}^{-s}(\Omega)}} \quad \text{for all } u^h \in V_h,$$

where $\tilde{H}^{-s}(\Omega) = (H^s(\Omega))'$ is defined by duality with respect to the L^2 inner product. The estimate (1.4) is essentially needed in the derivation of hybrid coupled finite element domain decomposition methods [1] and for hybrid boundary element methods [11] as well as in the construction of efficient preconditioners in finite and boundary element methods [12].

For globally quasiuniform triangulations, the estimate (1.2) is a direct consequence of a global inverse inequality. In [6], (1.2) was shown for nonuniform triangulations in one and two dimensions satisfying certain mesh conditions. The analysis is based on decay properties of the $L^2(\Omega)$ projection and results in conditions which depend on the global behavior of the mesh.

In this paper we prove the stability estimate (1.2) for arbitrary $n = 1, 2, \dots$ provided that local stability conditions are satisfied. This approach is valid for more general trial spaces, in particular, for trial functions of arbitrary polynomial degree, but for simplicity, we only consider the case of piecewise linear basis functions. In this case, we formulate explicit local mesh conditions which imply (1.2). These conditions can be easily checked for a given finite element mesh, allowing the user to redefine the mesh if necessary.

The remainder of this paper is organized as follows. Some preliminary notation is given in Section 2. In Section 3 we recall from [5, 8] the definition as well as some error and stability estimates for a quasi interpolation operator needed in our analysis. Our main result is formulated in Theorem 4.1. The proof is based on a general stability condition and several technical results given in Section 5. In Section 6 we discuss the stability condition in case of piecewise linear basis functions. Based on the eigenvalue analysis of locally defined weighted Gram matrices, we derive computable criteria for guaranteeing that the stability condition is satisfied.

2. NOTATION

Let

$$(2.1) \quad \mathcal{T}_h = \{\tau_l\}_{l=1}^N, \quad \Delta_l := \int_{\tau_l} dx \quad \text{for } l = 1, \dots, N.$$

As usual, we consider a family of meshes depending on h , the maximum diameter of any simplex. We assume that the triangulations are locally quasiuniform. This means that the diameter of the simplex divided by the diameter of the largest ball contained in the simplex is bounded independently of h for all simplices in all triangulations. Define

$$(2.2) \quad h_l := \Delta_l^{1/n} \quad \text{for } l = 1, \dots, N.$$

Let $\{x_k\}$ the set of all nodes of the mesh \mathcal{T}_h , where x_k is associated to the basis function φ_k ($\varphi_k(x_k) = 1$). We define

$$(2.3) \quad \omega_k := \text{supp } \varphi_k, \quad k = 1, \dots, M.$$

Let $I(k)$ denote the index set of all elements τ_l satisfying $\tau_l \subset \omega_k$. Then we define a local mesh size associated to the basis function φ_k by

$$(2.4) \quad \hat{h}_k := \frac{1}{\#I(k)} \sum_{l \in I(k)} h_l \quad \text{for } k = 1, \dots, M.$$

Here and in the rest of the paper, $\#$ denotes cardinality. Since the mesh \mathcal{T}_h is assumed to be locally quasiuniform, there exists a positive constant $\gamma \geq 1$ not depending on h such that

$$(2.5) \quad \gamma^{-1} \leq \frac{\hat{h}_k}{h_l} \leq \gamma \quad \text{for all } l \in I(k), \quad k = 1, \dots, M.$$

We define $J(l)$ to be the set of indices of the vertices in τ_l . Note that an inverse inequality holds locally [4], i.e.,

$$(2.6) \quad \|v^h\|_{H^1(\tau_l)} \leq c \cdot h_l^{-1} \cdot \|v^h\|_{L^2(\tau_l)} \quad \text{for all } v^h \in V_h, \quad l = 1, \dots, N.$$

Here and in the remainder of the paper, we use c with or without subscript to denote a generic positive constant which is independent of h .

3. QUASI INTERPOLATION

To prove the stability estimate (1.2) we need to use a projection operator P_h which is stable in $H^1(\Omega)$ and which satisfies local error estimates in $L^2(\tau_l)$ valid on all finite elements τ_l for $l = 1, \dots, N$. For this we will use the concept of quasi interpolation operators first introduced by Clement in [5]; see also [8].

We define local trial spaces of piecewise linear continuous functions by

$$(3.1) \quad V_h^k := \{v|_{\omega_k} : v \in V_h\}.$$

Let Q_h^k denote the L^2 projection onto V_h^k . We clearly have

$$(3.2) \quad \begin{aligned} \|Q_h^k u\|_{L^2(\omega_k)} &\leq \|u\|_{L^2(\omega_k)}, \\ \|(I - Q_h^k)u\|_{L^2(\omega_k)} &\leq c \cdot \hat{h}_k \cdot |u|_{H^1(\omega_k)}. \end{aligned}$$

Moreover, since the mesh is assumed to be locally quasiuniform, we have (see, e.g., [3])

$$(3.3) \quad \|Q_h^k u\|_{H^1(\omega_k)} \leq c \cdot \|u\|_{H^1(\omega_k)} \quad \text{for all } u \in H^1(\omega_k),$$

for $k = 1, \dots, M$.

Now we define a quasi interpolation operator by

$$(3.4) \quad (P_h u)(x) = \sum_{k=1}^M (Q_h^k u)(x_k) \cdot \varphi_k(x).$$

It is easy to check that P_h is a projection. Moreover, P_h is stable in $H^1(\Omega)$ and satisfies some local error estimates as asserted in the following lemma.

Lemma 3.1. *Let u be in $H^1(\Omega)$. There exists a positive constant c independent of h such that*

$$(3.5) \quad \|(I - P_h)u\|_{L^2(\tau_l)} \leq c \cdot \sum_{k \in J(l)} \hat{h}_k \cdot |u|_{H^1(\omega_k)} \quad \text{for } l = 1, \dots, N.$$

Moreover,

$$(3.6) \quad \|P_h u\|_{H^1(\Omega)} \leq c \cdot \|u\|_{H^1(\Omega)} \quad \text{for all } u \in H^1(\Omega).$$

Proof. The proof follows the general ideas already given in [5]. Let τ_l be an arbitrary but fixed finite element and let $\tilde{k} \in J(l)$ be a fixed index. For $x \in \tau_l$ we have the representation

$$(P_h u)(x) = (Q_h^{\tilde{k}} u)(x) + \sum_{k \in J(l), k \neq \tilde{k}} \left[(Q_h^k u)(x_k) - (Q_h^{\tilde{k}} u)(x_k) \right] \varphi_k(x).$$

Let $s = 0, 1$. Note that

$$\|\varphi_k\|_{H^s(\tau_l)} \leq c \cdot h_l^{n/2-s}.$$

Then, using (3.2) and (3.3), it follows that

$$\begin{aligned} \|(I - P_h)u\|_{H^s(\tau_l)} &\leq c_1 \cdot \hat{h}_{\tilde{k}}^{1-s} \cdot |u|_{H^1(\omega_{\tilde{k}})} \\ &\quad + c_2 \cdot h_l^{n/2-s} \sum_{k \in J(l), k \neq \tilde{k}} |(Q_h^k u)(x_k) - (Q_h^{\tilde{k}} u)(x_k)| \end{aligned}$$

Now

$$\|v^h\|_{L_\infty(\tau_l)} \leq c \cdot h_l^{-n/2} \cdot \|v^h\|_{L^2(\tau_l)} \quad \text{for all } v^h \in V_h, \quad l = 1, \dots, N.$$

Thus, (3.2) gives, for $x_k \in \tau_l$,

$$\begin{aligned} |(Q_h^k u)(x_k) - (Q_h^{\tilde{k}} u)(x_k)| &\leq \|Q_h^k u - Q_h^{\tilde{k}} u\|_{L_\infty(\tau_l)} \\ &\leq c \cdot h_l^{-n/2} \cdot \|Q_h^k u - Q_h^{\tilde{k}} u\|_{L^2(\tau_l)} \\ &\leq c \cdot h_l^{-n/2+1} \cdot \{|u|_{H^1(\omega_k)} + |u|_{H^1(\omega_{\tilde{k}})}\}. \end{aligned}$$

Hence,

$$\|(I - P_h)u\|_{H^s(\tau_l)} \leq c \cdot \sum_{k \in J(l)} \hat{h}_k^{1-s} \cdot |u|_{H^1(\omega_k)}$$

for $s = 0, 1$ and $l = 1, \dots, N$. Using this estimate for $s = 0$ gives (3.5), while for $s = 1$ we get (3.6) by summing over all elements. \square

4. MAIN RESULTS

In this section, we will formulate and prove the main result of this paper, the stability estimate (1.2) assuming some appropriate mesh conditions. For this, we define local weights

$$(4.1) \quad \gamma_k := \sqrt{\sum_{l \in I(k)} h_l^{-2} \cdot \|\varphi_k\|_{L^2(\tau_l)}^2} \quad \text{for } k = 1, \dots, M.$$

In addition, for each element τ_l , we define local matrices \mathbf{G}_l , \mathbf{D}_l and \mathbf{H}_l by

$$\mathbf{G}_l[j, i] = \langle \varphi_i^l, \varphi_j^l \rangle_{L^2(\tau_l)}, \quad \mathbf{D}_l = \text{diag} \left(\|\varphi_i^l\|_{L^2(\tau_l)}^2 \right), \quad \mathbf{H}_l = \text{diag} \left(\hat{h}_i^l \right),$$

for $i, j = 1, 2, 3$. Here φ_i^l and φ_j^l are the basis functions corresponding to the i th and j th vertex of the l th element, respectively, while \hat{h}_i^l is the related value of \hat{h} .

Now we are able to formulate local conditions to be used in the remainder of this section, specifically,

$$(4.2) \quad (\mathbf{H}_l^{-1} \mathbf{G}_l \mathbf{H}_l \underline{x}^l, \underline{x}^l) \geq c_0 \cdot (\mathbf{D}_l \underline{x}^l, \underline{x}^l) \quad \text{for all } \underline{x}^l \in \mathbf{R}^{\#J(l)}.$$

Here (\cdot, \cdot) denotes the inner product on $\mathbf{R}^{\#J(l)}$.

Remark 4.1. Let H denote the diagonal matrix,

$$H = \text{diag}(\hat{h}_k).$$

For $\underline{u} \in \mathbf{R}^M$, define $\underline{v} = H\underline{u}$ and $\underline{w} = H^{-1}\underline{u}$. The corresponding finite element functions are given by

$$u^h = \sum_{k=1}^M u_k \varphi_k, \quad v^h = \sum_{k=1}^M v_k \varphi_k, \quad w^h = \sum_{k=1}^M w_k \varphi_k.$$

We note that $(\mathbf{H}_l^{-1} \mathbf{G}_l \mathbf{H}_l \underline{u}^l, \underline{u}^l) = \langle v^h, w^h \rangle_{L^2(\tau_l)}$, where \underline{u}^l denotes the components of \underline{u} associated with the element τ_l . By local quasiuniformity, (4.2) is thus equivalent to

$$(4.3) \quad \|u^h\|_{L^2(\tau_l)}^2 \leq c \langle v^h, w^h \rangle_{L^2(\tau_l)}.$$

If all of the simplices have the same measure then $\mathbf{H}_l^{-1} \mathbf{G}_l \mathbf{H}_l = \mathbf{G}_l$, so

$$\langle v^h, w^h \rangle_{L^2(\tau_l)} = \langle u^h, u^h \rangle_{L^2(\tau_l)}$$

and (4.3) is trivial. The inequality will still hold provided that the measures of neighboring simplices do not vary too much. Explicit local conditions on the mesh for the case of piecewise linear elements are given in Section 6.

The following result is the main theorem of this paper.

Theorem 4.1. *Let condition (4.2) be satisfied. Then the L^2 projection $Q_h : H^1(\Omega) \rightarrow V_h \subset H^1(\Omega)$ is stable. In particular, there exists a positive constant c independent of h such that*

$$(4.4) \quad \|Q_h v\|_{H^1(\Omega)} \leq c \cdot \|v\|_{H^1(\Omega)} \quad \text{for all } v \in H^1(\Omega).$$

The proof of this theorem depends on the following lemma. A similar estimate was used in [2] to construct spectrally equivalent multilevel preconditioners in finite element methods in the case of globally quasiuniform meshes. The proof of the lemma will be given in the next section.

Lemma 4.1. *Let condition (4.2) be satisfied. Then there exists a positive constant c such that*

$$(4.5) \quad \sum_{l=1}^N h_l^{-2} \cdot \|v^h\|_{L^2(\tau_l)}^2 \leq c \cdot \sum_{k=1}^M \left[\frac{\langle v^h, \varphi_k \rangle_{L^2(\Omega)}}{\hat{h}_k \cdot \|\varphi_k\|_{L^2(\Omega)}} \right]^2$$

for all $v^h \in V_h$.

Proof of Theorem 4.1. Using the triangle inequality, (3.6), and (2.6), we get

$$\begin{aligned} \|Q_h v\|_{H^1(\Omega)}^2 &\leq 2 \cdot \left\{ \|P_h v\|_{H^1(\Omega)}^2 + \sum_{l=1}^N \|(Q_h - P_h)v\|_{H^1(\tau_l)}^2 \right\} \\ &\leq c \cdot \left\{ \|v\|_{H^1(\Omega)}^2 + \sum_{l=1}^N h_l^{-2} \cdot \|(Q_h - P_h)v\|_{L^2(\tau_l)}^2 \right\}. \end{aligned}$$

From Lemma 4.1 and the Schwarz inequality it follows that

$$\begin{aligned} \sum_{l=1}^N h_l^{-2} \cdot \| (Q_h - P_h)v \|^2_{L^2(\tau_l)} &\leq c \cdot \sum_{k=1}^M \left[\frac{\langle (Q_h - P_h)v, \varphi_k \rangle_{L^2(\Omega)}}{\hat{h}_k \cdot \|\varphi_k\|_{L^2(\Omega)}} \right]^2 \\ &= c \cdot \sum_{k=1}^M \left[\frac{\langle (I - P_h)v, \varphi_k \rangle_{L^2(\omega_k)}}{\hat{h}_k \cdot \|\varphi_k\|_{L^2(\Omega)}} \right]^2 \\ &\leq c \cdot \sum_{k=1}^M \hat{h}_k^{-2} \cdot \| (I - P_h)v \|^2_{L^2(\omega_k)} \end{aligned}$$

Hence, the assertion follows from (3.5). \square

5. PROOF OF LEMMA 4.1

In this section, we prove Lemma 4.1. We start by introducing some notation. We define two additional $M \times M$ diagonal matrices given by

$$(5.6) \quad D_\gamma = \text{diag}(\gamma_k), \quad D_\varphi = \text{diag}(\hat{h}_k \cdot \|\varphi_k\|_{L^2(\Omega)}),$$

where \hat{h}_k and γ_k are defined as in (2.4) and (4.1), respectively. The first step in the proof of Lemma 4.1 involves estimating the inverse of a scaled Gram matrix.

Lemma 5.1. *Let assumption (4.2) be satisfied. Then there exists a positive constant c such that*

$$\|\underline{x}\|_2 \leq c \cdot \|Ax\|_2 \quad \text{for all } \underline{x} \in \mathbf{R}^M,$$

where A is the scaled Gram matrix defined by

$$(5.7) \quad A = D_\varphi^{-1} G D_\gamma^{-1}$$

and G is the Gram matrix $G_{ij} = \langle \varphi_i, \varphi_j \rangle_{L^2(\Omega)}$.

Proof. Let $\underline{u}, \underline{u}^l, \underline{v}, \underline{w}, u^h, v^h$ and w^h be as in Remark 4.1. Setting $\tilde{G} = H^{-1}GH$ and using (4.2) gives

$$\begin{aligned} (\tilde{G}\underline{u}, \underline{u}) &= (G\underline{v}, \underline{w}) = \langle v^h, w^h \rangle_{L^2(\Omega)} = \sum_{l=1}^N \langle v^h, w^h \rangle_{L^2(\tau_l)} \\ &= \sum_{l=1}^N (\mathbf{H}_l^{-1} \mathbf{G}_l \mathbf{H}_l \underline{u}^l, \underline{u}^l) \geq c_0 \cdot \sum_{l=1}^N (\mathbf{D}_l \underline{u}^l, \underline{u}^l) = c_0 \cdot (D\underline{u}, \underline{u}). \end{aligned}$$

Here D is the diagonal matrix with entries $\|\varphi_k\|_{L^2(\Omega)}^2$. Let $D^{1/2} = \text{diag}(\|\varphi_k\|_{L^2(\Omega)})$. From

$$\begin{aligned} c_0 \cdot \|D^{1/2}\underline{u}\|_2^2 &= c_0 \cdot (D\underline{u}, \underline{u}) \leq (\tilde{G}\underline{u}, \underline{u}) \\ &= (D^{-1/2}\tilde{G}\underline{u}, D^{1/2}\underline{u}) \leq \|D^{-1/2}\tilde{G}\underline{u}\|_2 \|D^{1/2}\underline{u}\|_2, \end{aligned}$$

we conclude that

$$c_0 \cdot \|D^{1/2}\underline{u}\|_2 \leq \|D^{-1/2}\tilde{G}\underline{u}\|_2 \quad \text{for all } \underline{u} \in \mathbf{R}^M.$$

Taking $\tilde{\underline{u}} = D_\gamma \underline{u}$ gives

$$c_0 \cdot \|D^{1/2}D_\gamma^{-1}\tilde{\underline{u}}\|_2 \leq \|D^{-1/2}D_\varphi D_\varphi^{-1}\tilde{G}D_\gamma^{-1}\tilde{\underline{u}}\|_2 = \|D^{-1/2}D_\varphi \tilde{A}\tilde{\underline{u}}\|_2,$$

where $\tilde{A} = D_\varphi^{-1} \tilde{G} D_\gamma^{-1}$. The ratio of the diagonal entries satisfies

$$\frac{D^{1/2}[k, k]}{D_\gamma[k, k]} = \frac{\|\varphi_k\|_{L^2(\Omega)}}{\sqrt{\sum_{l \in I(k)} h_l^{-2} \|\varphi_k\|_{L^2(\tau_l)}^2}} \geq c \cdot \hat{h}_k$$

and

$$\frac{D_\varphi[k, k]}{D^{1/2}[k, k]} = \frac{\hat{h}_k \cdot \|\varphi_k\|_{L^2(\Omega)}}{\|\varphi_k\|_{L^2(\Omega)}} = \hat{h}_k$$

for all $k = 1, \dots, M$. Thus,

$$c \cdot \|H\tilde{u}\|_2 \leq \|H\tilde{A}\tilde{u}\|_2 \quad \text{for all } \tilde{u} \in \mathbf{R}^M.$$

Taking $\underline{x} = H\tilde{u}$ above gives

$$c \cdot \|\underline{x}\|_2 \leq \|H\tilde{A}H^{-1}\underline{x}\|_2 = \|HD_\varphi^{-1}H^{-1}GHD_\gamma^{-1}H^{-1}\underline{x}\|_2 = \|A\underline{x}\|_2$$

for all $\underline{x} \in \mathbf{R}^M$. This completes the proof of the lemma. \square

We now give the proof of Lemma 4.1.

Proof of Lemma 4.1. Let $\underline{v} \in \mathbf{R}^M$ and set $v^h = \sum_{k=1}^M \underline{v}_k \varphi_k \in V_h$. Then, the left hand side of (4.5) is bounded by

$$\begin{aligned} \sum_{l=1}^N h_l^{-2} \cdot \|v^h\|_{L^2(\tau_l)}^2 &\leq c \cdot \sum_{l=1}^N h_l^{-2} \sum_{k \in J(l)} \underline{v}_k^2 \cdot \|\varphi_k\|_{L^2(\tau_l)}^2 \\ &= c \cdot \sum_{k=1}^M \underline{v}_k^2 \sum_{l \in I(k)} h_l^{-2} \cdot \|\varphi_k\|_{L^2(\tau_l)}^2 \\ &= c \cdot \sum_{k=1}^M \underline{v}_k^2 \gamma_k^2 = c \cdot \sum_{k=1}^M \underline{x}_k^2 = c \cdot \|\underline{x}\|_2^2, \end{aligned}$$

where $\underline{x}_k = \gamma_k \underline{v}_k$. The right hand side in (4.5) is

$$\begin{aligned} \sum_{k=1}^M \left[\frac{\langle v^h, \varphi_k \rangle_{L^2(\Omega)}}{\hat{h}_k \|\varphi_k\|_{L^2(\Omega)}} \right]^2 &= \sum_{k=1}^M \left[\sum_{j=1}^M \underline{v}_j \cdot \frac{\langle \varphi_j, \varphi_k \rangle_{L^2(\Omega)}}{\hat{h}_k \|\varphi_k\|_{L^2(\Omega)}} \right]^2 \\ &= \sum_{k=1}^M \left[\sum_{j=1}^M \underline{x}_j \cdot \frac{\langle \varphi_j, \varphi_k \rangle_{L^2(\Omega)}}{\gamma_j \hat{h}_k \|\varphi_k\|_{L^2(\Omega)}} \right]^2 \\ &= \sum_{k=1}^M [(Ax)_k]^2 = \|Ax\|_2^2, \end{aligned}$$

using the matrix definition (5.7). Hence, (4.5) follows from Lemma 5.1. \square

Although we only considered the case of piecewise linear basis functions, the same approach may be used for higher order piecewise polynomial finite element spaces. In addition, the case of $V_h \subset H_0^1(\Omega)$ with basis functions vanishing along the boundary $\partial\Omega$ can be treated with only slight modifications. In this case we consider the index set $I(k)$ only for nodes x_k associated with a basis function $\varphi_k \in V_h$. Then all proofs given above apply. Note that the dimension of the local

matrices for boundary simplices decrease since the rows and columns corresponding to boundary nodes do not appear.

6. FINITE ELEMENT SPACES

The stability estimate in Theorem 4.1 is based on the condition (4.2). If we define the symmetric matrix

$$(6.1) \quad \mathbf{G}_l^S := \frac{1}{2} [\mathbf{H}_l \mathbf{G}_l \mathbf{H}_l^{-1} + \mathbf{H}_l^{-1} \mathbf{G}_l \mathbf{H}_l],$$

then (4.2) is the same as

$$(6.2) \quad (\mathbf{G}_l^S \underline{x}_l, \underline{x}_l) \geq c_0 \cdot (\mathbf{D}_l \underline{x}_l, \underline{x}_l) \quad \text{for all } \underline{x}_l \in \mathbf{R}^{\#J(l)}.$$

Let τ_l in \mathcal{T}_h be an arbitrary element. Note that $\#J(l) = n + 1$. A simple computation shows that

$$(6.3) \quad \mathbf{D}_l = d_n \cdot \Delta_l \cdot I \quad \text{with } d_n = \frac{2}{(n+1)(n+2)},$$

where I is the identity matrix in $n + 1$ dimensions. Moreover,

$$(6.4) \quad \mathbf{G}_l^S = \frac{1}{4} \cdot d_n \cdot \Delta_l \cdot \mathbf{A}_l,$$

where the matrix \mathbf{A}_l is defined by

$$\mathbf{A}_l[i, j] = \begin{cases} 4 & \text{for } i = j, \\ \frac{\hat{h}_i^l}{\hat{h}_j^l} + \frac{\hat{h}_j^l}{\hat{h}_i^l} & \text{for } i \neq j, \end{cases} \quad i, j = 1, \dots, n+1.$$

Here \hat{h}_j^l is the value of \hat{h} corresponding to the j th vertex of the l th element. Hence, to show (6.2) it is sufficient to consider the eigenvalues $\{\lambda_i\}$ of the matrix \mathbf{A}_l . To this end, we give the following proposition.

Proposition 6.1. *Let $\alpha_1, \dots, \alpha_{n+1}$ be real numbers with $\alpha_i \neq 0$ for $i = 1, \dots, n+1$, and consider the matrix*

$$\mathbf{A}[i, j] = \begin{cases} 4 & \text{for } i = j, \\ \frac{\alpha_i}{\alpha_j} + \frac{\alpha_j}{\alpha_i} & \text{for } i \neq j, \end{cases} \quad i, j = 1, \dots, n+1.$$

Then, all eigenvalues of \mathbf{A} are in the set $\{\lambda_+, \lambda_-, 2\}$, where

$$(6.5) \quad \lambda_{\pm} = 3 + n \pm \sqrt{\sum_{i=1}^{n+1} \alpha_i^2 \cdot \sum_{i=1}^{n+1} \alpha_i^{-2}}.$$

Proof. We first consider the case when $\alpha_1, \alpha_2, \dots, \alpha_{n+1}$ are not all equal. Let $\mathbf{M}_+ = (\alpha_1, \alpha_2, \dots, \alpha_{n+1})^t$ and $\mathbf{M}_- = (\alpha_1^{-1}, \alpha_2^{-1}, \dots, \alpha_{n+1}^{-1})^t$. Then $\mathbf{A} = 2I + \mathbf{N}$, where $\mathbf{N} = \mathbf{M}_- \cdot \mathbf{M}_+^t + \mathbf{M}_+ \cdot \mathbf{M}_-^t$. Note that the matrix \mathbf{N} is symmetric with range equal to the two dimensional subspace spanned by \mathbf{M}_+ and \mathbf{M}_- . Thus, zero is an eigenvalue of \mathbf{N} with multiplicity $n - 1$. We need only compute the two remaining eigenvalues. By expanding the corresponding eigenvectors in the basis $\{\mathbf{M}_+, \mathbf{M}_-\}$,

it is elementary to see that the remaining two eigenvalues of \mathbf{N} are eigenvalues of the matrix

$$\begin{pmatrix} n+1 & \sum_{i=1}^{n+1} \alpha_i^2 \\ \sum_{i=1}^{n+1} \alpha_i^{-2} & n+1 \end{pmatrix}$$

The proposition immediately follows for this case.

If $\alpha_1 = \alpha_2 = \dots = \alpha_{n+1}$ then \mathbf{M}_+ and \mathbf{M}_- are linearly dependent, so 2 is an eigenvalue of \mathbf{A} with multiplicity n . The remaining eigenvalue is $\lambda_+ = 2n+4$. This completes the proof of the proposition. \square

It follows from (6.4) that the local condition (6.2) is satisfied if all eigenvalues of \mathbf{A}_l are strictly positive and bounded away from zero. By the proposition, this is equivalent to

$$(6.6) \quad 3 + n - \sqrt{\sum_{i=1}^{n+1} (\hat{h}_i^l)^{-2} \cdot \sum_{i=1}^{n+1} (\hat{h}_i^l)^2} \geq c$$

with c independent of τ_l . Inequality (6.6) provides a mesh constraint for adaptive triangulations.

Remark 6.1. We can satisfy (6.6) for any $0 < c < 2$ provided that we design a mesh which gives rise to a γ (in (2.5)) sufficiently close to one. It immediately follows from (2.5) that

$$3 + n - \sqrt{\sum_{i=1}^{n+1} (\hat{h}_i^l)^{-2} \cdot \sum_{i=1}^{n+1} (\hat{h}_i^l)^2} \geq 2 - (n+1)(\gamma^2 - 1).$$

Thus (6.6) holds if

$$\gamma \leq \sqrt{1 + \frac{2-c}{n+1}}.$$

Thus, a mesh gives rise to an $H^1(\Omega)$ stable L^2 projection provided that the change in measures of neighboring simplices are controlled.

Finally, if a finite element mesh \mathcal{T}_h is given, the mesh condition (6.6) and therefore the stability assumption (4.2) can be checked explicitly (by direct computation). To illustrate the applicability of the mesh condition (6.6), we consider an adaptive finite element mesh for $n = 2$ as shown in Figure 1, generated by an adaptive algorithm described in [10]. In Table 1, we give values for

$$c = \max_{\tau_l} \left(5 - \sqrt{\sum_{i=1}^3 (\hat{h}_i^l)^{-2} \cdot \sum_{i=1}^3 (\hat{h}_i^l)^2} \right)$$

as a function of the refinement level L and the number of finite element nodes M .

TABLE 1. Computational results for c

L	0	1	2	3	4	5	6	7	8	9
M	8	17	28	53	87	155	291	532	1034	2003
c	2.00	1.93	1.59	1.52	1.52	1.66	1.61	1.09	1.49	1.50

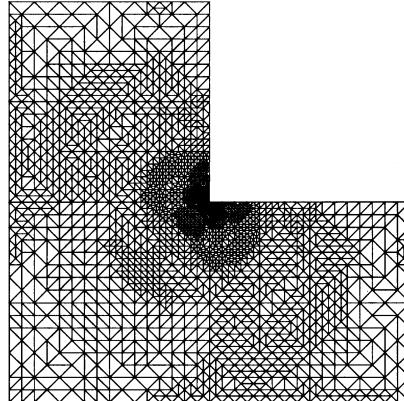


FIGURE 1. Adaptive finite element triangulation

REFERENCES

- [1] A. Agouzal, J.-M. Thomas, Une methode d'elements finis hybrides en decomposition de domaines. *Math. Modell. Numer. Anal.* 29 (1995) 749–764. MR 96g:65115
- [2] J. H. Bramble, J. E. Pasciak, P. S. Vassilevski, Computational scales of Sobolev norms with applications to preconditioning. *Math. Comp.* 69 (2000), 463–480. MR 2000k:65088
- [3] J. H. Bramble, J. Xu, Some estimates for a weighted L^2 projection. *Math. Comp.* 56 (1991) 463–476. MR 91i:65140
- [4] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*. North-Holland, 1978. MR 58:25001
- [5] P. Clement, Approximation by finite element functions using local regularization. *RAIRO Anal. Numer.* 9 R-2 (1975) 77–84. MR 53:4569
- [6] M. Crouzeix, V. Thomeé, The stability in L_p and W_p^1 of the L^2 -projection onto finite element function spaces. *Math. Comp.* 48 (1987) 521–532. MR 88f:41016
- [7] W. McLean, O. Steinbach, Boundary element preconditioners for a hypersingular integral equation on a curve. *Adv. Comput. Math.* 11 (1999) 271–286. MR 2000k:65236
- [8] L. R. Scott, S. Zhang, Finite element interpolation of nonsmooth functions satisfying boundary conditions. *Math. Comp.* 54 (1990) 483–493. MR 90j:65021
- [9] H. Schulz, O. Steinbach, A new a posteriori error estimator in direct boundary element methods. The Neumann problem. *Multifield Problems. State of the Art.* (A.-M. Sändig, W. Schiehlen, and W. L. Wendland, eds.) Springer-Verlag, Berlin, 201–208, 2000.
- [10] O. Steinbach, Adaptive finite element–boundary element solution of boundary value problems. *J. Comput. Appl. Math.* 106 (1999) 307–316. MR 2000b:65225
- [11] O. Steinbach, On a hybrid boundary element method. *Numer. Math.* 84 (2000), 679–695. MR 2001a:65154
- [12] O. Steinbach, W. L. Wendland, The construction of some efficient preconditioners in the boundary element method. *Adv. Comput. Math.* 9 (1998) 191–216. MR 99j:65219
- [13] L. B. Wahlbin, *Superconvergence in Galerkin Finite Element Methods*. Lecture Notes in Mathematics 1605, Springer, Berlin, 1995. MR 98j:65083

DEPARTMENT OF MATHEMATICS, TEXAS A & M UNIVERSITY, COLLEGE STATION, TEXAS 77843
E-mail address: bramble@math.tamu.edu

DEPARTMENT OF MATHEMATICS, TEXAS A & M UNIVERSITY, COLLEGE STATION, TEXAS 77843
E-mail address: pasciak@math.tamu.edu

MATHEMATISCHES INSTITUT A, UNIVERSITÄT STUTTGART, PFAFFENWALDRING 57, 70569
 STUTTGART, GERMANY
E-mail address: steinbach@mathematik.uni-stuttgart.de

2.12 A proof of the inf–sup condition for the Stokes equations on Lipschitz domains (2003)

A proof of the inf–sup condition for the Stokes equations on Lipschitz domains[8]

A PROOF OF THE INF-SUP CONDITION FOR THE STOKES EQUATIONS ON LIPSCHITZ DOMAINS

JAMES H. BRAMBLE

ABSTRACT. The purpose of this paper is to present a rather simple proof of an inequality of Nečas [9] which is equivalent to the inf-sup condition. This inequality is fundamental in the study of the Stokes equations. The boundary of the domain is only assumed to be Lipschitz.

1. INTRODUCTION

One of the most important inequalities in the theory of incompressible fluids is the so-called inf-sup condition, cf. [2] [3] [5] [10]. Let Ω be a bounded domain in R^N with a Lipschitz boundary. Let $L^2(\Omega)$ and $H_0^1(\Omega)$ be the standard Hilbert spaces of distributions on Ω , with norms $\|\cdot\|_{L^2(\Omega)}$ and $\|\cdot\|_{H_0^1(\Omega)}$. The L^2 -inner product, as well as the pairing between $H_0^1(\Omega)$ and its dual, $H^{-1}(\Omega)$ is denoted by (\cdot, \cdot) . Note that the space $\mathcal{D}(\Omega)$, of infinitely differentiable functions with compact support in Ω , is dense in both $L^2(\Omega)$ and $H_0^1(\Omega)$.

Let $L_0^2(\Omega) = \{u \in L^2(\Omega) | \int_{\Omega} u \, dx = 0\}$. Denote by \mathbf{v} vector functions with components $v_i, i = 1, \dots, N$. In the following, C denotes a constant which depends only on Ω unless otherwise stated. The important inf-sup condition is

$$(1.1) \quad \inf_{u \in L_0^2(\Omega)} \sup_{\mathbf{v} \in [H_0^1(\Omega)]^N} \frac{(u, \nabla \cdot \mathbf{v})}{\|u\|_{L^2(\Omega)} \|\mathbf{v}\|_{[H^1(\Omega)]^N}} \geq C > 0,$$

where $\nabla \cdot \mathbf{v} = \sum_{i=1}^N \frac{\partial v_i}{\partial x_i}$ is the standard divergence operator and $\|\mathbf{v}\|_{[H_0^1(\Omega)]^N}^2 = \sum_{i=1}^N \|v_i\|_{H^1(\Omega)}^2$. This is the same as

$$(1.2) \quad C\|u\|_{L^2(\Omega)} \leq \sup_{\mathbf{v} \in [H_0^1(\Omega)]^N} \frac{(u, \nabla \cdot \mathbf{v})}{\|\mathbf{v}\|_{[H^1(\Omega)]^N}} \quad \text{for all } u \in L_0^2(\Omega).$$

It is easy to see that both (1.1) and (1.2) are equivalent to

$$(1.3) \quad C\|u\|_{L^2(\Omega)} \leq \sum_{i=1}^N \sup_{v \in [H_0^1(\Omega)]} \frac{(u, \frac{\partial v}{\partial x_i})}{\|v\|_{H^1(\Omega)}} \quad \text{for all } u \in L_0^2(\Omega).$$

The norm on $H^{-1}(\Omega)$ is given by

$$(1.4) \quad \|u\|_{H^{-1}(\Omega)} = \sup_{v \in [H_0^1(\Omega)]} \frac{(u, v)}{\|v\|_{H^1(\Omega)}}.$$

Date: November 9, 2010.

This work was supported in part under the National Science Foundation Grant No. DMS-9973328 .

Noting that, for $u \in L^2(\Omega)$, the distributional derivative, $\frac{\partial u}{\partial x_i} \in H^{-1}(\Omega)$, (1.3) is the same as

$$(1.5) \quad C\|u\|_{L^2(\Omega)} \leq \sum_{i=1}^N \left\| \frac{\partial u}{\partial x_i} \right\|_{H^{-1}(\Omega)}, \quad \text{for all } u \in L_0^2(\Omega).$$

The following inequality

$$(1.6) \quad C\|u\|_{L^2(\Omega)} \leq \left(\|u\|_{H^{-1}(\Omega)}^2 + \sum_{i=1}^N \left\| \frac{\partial u}{\partial x_i} \right\|_{H^{-1}(\Omega)}^2 \right)^{1/2} \equiv \|u\|_{X(\Omega)}, \quad \text{for all } u \in L^2(\Omega),$$

was proved by Nečas [9].

It is not difficult to show that (1.6) is equivalent to (1.3) and therefore equivalent to (1.2). We will show in Appendix I that existence and uniqueness in the Stokes problem follows easily from (1.2). Thus (1.6) is of fundamental importance. To emphasize this point even more, it is easy to see that Korn's second inequality follows from (1.6), (cf. Duvault-Lions [3]) which is fundamental in the theory of elasticity. Direct proofs of Korn's second inequality, in the case of nonsmooth domains are given in, e.g., [4], [6] and [8]. The proofs are not elementary. As is done in [3], we will show in Appendix II that Korn's second inequality is a corollary of (1.6).

The purpose of this note is to provide a proof of (1.6). More precisely we shall give a proof of the following.

Theorem 1.1. *Let $\Omega \in R^N$ be a connected domain with a Lipschitz boundary. Then there exists a constant C such that, for all $u \in L^2(\Omega)$,*

$$(1.7) \quad C\|u\|_{L^2(\Omega)} \leq \|u\|_{X(\Omega)}.$$

Clearly if $\Omega = \cup_{j=1}^M \Omega_j$ and if (1.7) holds for each Ω_j , $j = 1, \dots, M < \infty$, then it holds for Ω .

We will use the fact that any Lipschitz domain can be written as the union of strongly star shaped Lipschitz domains. Hence we may assume, without loss, that Ω is strongly star shaped. A weaker version of this statement is proved by Bernardi [1]. The stronger version follows by a slight modification of her proof. A strongly star shaped Lipschitz domain is a domain which can be represented as follows.

$$\Omega = \{x \in R^N | r < g(x/r)\} \quad \text{and} \quad \partial\Omega = \{x \in R^N | r = g(x/r)\}$$

where $r = |x|$ and $g \in W_\infty^1$ with $\|g\|_{W_\infty^1} \leq \Lambda$. Evidently we have assumed that Ω is star shaped with respect to the origin. By x/r we mean $(x_1/r, \dots, x_N/r)$ and note that x_i/r is independent of r away from the origin.

2. A TRANSFORMATION OF B TO Ω

The plan is to map the unit ball B to Ω with ∂B mapped to the boundary $\partial\Omega$ and then to show that Theorem 1.1 holds if it holds for $\Omega = B$. We will use the variables y in B and set $|y| = \rho$. The boundary $\partial B = \{y \in R^N | \rho = 1\}$.

The simplest transformation would be defined by $x_i = y_i g(y/\rho)$. This transformation is not smooth enough to do the job unless $\partial\Omega$ is sufficiently smooth, so we follow the idea of Nečas and smooth g . We do this as follows.

Let $\Phi \in \mathcal{D}(R^N)$ with $\Phi(x) = 0$ if $|x| \geq 1$, $\Phi(x) \geq 0$ and $\int \Phi(x) dx = 1$. For $0 < h < 1/4$ and $|x| > 1/2$, define

$$g(h, x) = \frac{1}{h^N} \int_{R^N} g(y/\rho) \Phi((x - y)/h) dy.$$

The following are some important properties of $g(h, x)$.

- (1) $\lim_{h \rightarrow 0} g(h, y/\rho) = g(y/\rho)$
- (2) $|\frac{\partial}{\partial h} g(h, x)| \leq C$
- (3) $|\frac{\partial}{\partial x_i} g(h, x)| \leq C$
- (4) $|D^\alpha g(h, x)| \leq C/h, |\alpha| = 2$
- (5) For $|x| = 1, |g(h, x) - g(x)| \leq \bar{C}h$, where \bar{C} depends only on Λ , the Lipschitz bound, and bounds for $g(x/r)$.

We next set $h = \epsilon(1 - \rho)$. From the last inequality, 5, we obtain easily that, for ϵ small enough

$$g(\epsilon(1 - \rho), y/\rho) \geq g(y/\rho) - \bar{C}\epsilon \geq C > 0.$$

We are now in a position to define the transformation \mathcal{T} from the unit ball to Ω as follows. For $y \in B$ define \mathcal{T} by

$$(2.1) \quad x_i = y_i g(\epsilon(1 - \rho), y/\rho),$$

for $i = 1, \dots, N$ and ϵ small enough.

We next check that \mathcal{T} maps B to Ω . To this end notice that $r = \rho g(\epsilon(1 - \rho), y/\rho)$ and set $f(\rho) = \rho g(\epsilon(1 - \rho), y/|y|)$ for $y \in B$ fixed. Now $f(0) = 0$ and $f(1) = g(y/|y|)$. Now

$$f'(\rho) = g(\epsilon(1 - \rho), y/|y|) - \epsilon \frac{\partial}{\partial h} g(\epsilon(1 - \rho), y/|y|) > 0$$

if ϵ is small enough. Thus \mathcal{T} maps B to Ω .

We next compute the Jacobian of \mathcal{T} . For ease of notation, set $G = g(\epsilon(1 - \rho), y/\rho)$. From (2.1) the Jacobian matrix a_{ij} is

$$a_{ij} = \frac{\partial x_i}{\partial y_j} = \delta_{ij} G + y_i \frac{\partial G}{\partial y_j}.$$

Let \mathbf{v}^1 be the vector in R^N whose components are $v_j^1 = y_j/\rho$. Then

$$\sum_{j=1}^N a_{ij} v_j^1 = G v_i^1 + \rho v_i^1 \frac{\partial G}{\partial \rho} = (G - \epsilon \rho \frac{\partial G}{\partial h}) v_i^1,$$

so that \mathbf{v}^1 is an eigenvector of a_{ij} with eigenvalue $(G - \epsilon\rho\frac{\partial G}{\partial h})$. To find the rest of the eigenvalues, let $\{\mathbf{v}^k\}$ be an orthonormal set of vectors. Then $A_{kl} = \sum_{i,j} a_{ij} v_i^k v_j^l$ has the same eigenvalues as a_{ij} . Now

$$A_{kl} = \delta_{kl}G + \sum_i v_i^k v_i^l \rho \sum_j v_j^l \frac{\partial G}{\partial y_j}.$$

Hence A_{kl} is an upper diagonal matrix with $A_{kk} = G$ if $k > 1$. Thus we have found that the Jacobian determinant $J = (G - \epsilon\rho\frac{\partial G}{\partial h})G^{N-1}$.

Now from Property 4, $|D^\alpha g(1 - \rho, y/\rho)| \leq C/(1 - \rho)$, for any $|\alpha| = 2$. We have for two constants c_0 and c_1 that $c_0 \leq g(y/\rho) \leq c_1$. Hence

$$c_1(1 - \rho) \geq (1 - \rho)g(y/\rho) = g(x/r) - r + \rho(g(\epsilon(1 - \rho), y/\rho) - g(y/\rho)),$$

where we used the fact that $y/\rho = x/r$. Thus, using Property 5,

$$c_1(1 - \rho) \geq (1 - \rho)g(y/\rho) \geq g(x/r) - r - \rho\epsilon\bar{C}(1 - \rho).$$

Similarly

$$c_0(1 - \rho) \leq (1 - \rho)g(y/\rho) \leq g(x/r) - r + \rho\epsilon\bar{C}(1 - \rho).$$

Hence, for ϵ small enough we have

$$c_0(1 - \rho) \leq (g(x/r) - r) \leq c_1(1 - \rho).$$

It follows then that for $|\alpha| = 2$

$$|D^\alpha g(1 - \rho, y/\rho)| \leq C/(g(x/r) - r)$$

and therefore for $|\alpha| = 1$

$$|D^\alpha J| \leq C/(g(x/r) - r),$$

where D^α is any partial derivative of order one with respect to x_i or y_i .

3. SOME LEMMAS

We will need a few lemmas.

Lemma 3.1. *Let $u \in L^2(R^N)$. Then*

$$\|u\|_{L^2(R^N)} = \|u\|_{X(R^N)}.$$

Proof. Using the Fourier transform, \hat{u}

$$\|u\|_{L^2(R^N)}^2 = \|\hat{u}\|_{L^2(R^N)}^2 = \|(1+|\xi|^2)^{-1/2}\hat{u}\|_{L^2(R^N)}^2 + \sum_{i=1}^N \|(1+|\xi|^2)^{-1/2}\xi_i\hat{u}\|_{L^2(R^N)}^2 = \|u\|_{X(R^N)}^2.$$

□

Lemma 3.2. *Let $\Omega \subset R^N$ be a bounded domain. Let $\gamma \in \mathcal{D}(\Omega)$ be fixed. Then there exists a constant C_γ depending only on γ such that, for all $u \in L^2(\Omega)$,*

$$\|\gamma u\|_{L^2(\Omega)} \leq C_\gamma \|u\|_{X(\Omega)}.$$

Proof. Using Lemma 3.1 $\|\gamma u\|_{L^2(\Omega)} = \|\gamma u\|_{L^2(R^N)} = \|\gamma u\|_{X(R^N)}$. The lemma follows from the fact that multiplication by γ is a bounded operator from $X(\Omega)$ to $X(R^N)$. \square

We next give an elementary proof of a Hardy-type inequality.

Lemma 3.3. *Let $\phi \in \mathcal{D}(\Omega)$. Then*

$$(3.1) \quad \int_{\Omega} \frac{\phi^2}{(g(x/r) - r)^2} dx \leq 4 \|\phi\|_{H^1(\Omega)}^2.$$

Proof. Let $\Omega_\delta = \{x \in \Omega | |x| > \delta > 0\}$. Then

$$\int_{\Omega_\delta} \frac{\phi^2}{(g(x/r) - r)^2} dx = \int_{\Omega_\delta} \frac{\partial}{\partial r} \left(\frac{\phi^2}{g(x/r) - r} \right) dx - \int_{\Omega_\delta} \frac{1}{g(x/r) - r} \frac{\partial(\phi^2)}{\partial r} dx.$$

Now

$$\begin{aligned} \int_{\Omega_\delta} \frac{\partial}{\partial r} \left(\frac{\phi^2}{g(x/r) - r} \right) dx &= \sum_{i=1}^N \int_{\Omega_\delta} \frac{x_i}{r} \frac{\partial}{\partial x_i} \left[\frac{\phi^2}{g(x/r) - r} \right] dx \\ &= -(N-1) \int_{\Omega_\delta} \frac{\phi^2}{r(g(x/r) - r)} dx - \int_{|x|=\delta} \frac{\phi^2}{g(x/r) - \delta} ds, \end{aligned}$$

where the last integral is taken over the surface of the N-sphere of radius δ . Hence, for $\delta < \min g(x/r)$,

$$\int_{\Omega_\delta} \frac{\partial}{\partial r} \left(\frac{\phi^2}{g(x/r) - r} \right) dx \leq 0.$$

Thus

$$\int_{\Omega_\delta} \frac{\phi^2}{(g(x/r) - r)^2} dx \leq -2 \int_{\Omega_\delta} \frac{1}{g(x/r) - r} \phi \frac{\partial \phi}{\partial r} dx.$$

Using Schwarz's inequality and letting δ go to zero we obtain Lemma 3.3. \square

We now use this to prove the following.

Lemma 3.4. *Let $f \in C^1(\Omega) \cap C^0(\bar{\Omega})$ and satisfy $|D^\alpha f| \leq \tilde{C}/(g(x/r) - r)^{|\alpha|}$ for $x \in \Omega$ and $|\alpha| \leq 1$. Then there is constant C depending only on \tilde{C} such that*

$$(3.2) \quad \|f\phi\|_{H^1(\Omega)} \leq C \|\phi\|_{H^1(\Omega)}, \quad \text{for all } \phi \in \mathcal{D}(\Omega).$$

Proof. This follows immediately from the Lemma 3.3. \square

4. REDUCTION TO THE UNIT BALL

Let $B_\delta = \{y \in B | |y| > \delta\}$. In view of Lemma 3.2 it suffices to prove (1.6) for $\tilde{v} \in \mathcal{D}(\mathcal{T}(B_\delta))$. Thus we want to prove the following proposition.

Proposition 4.1. *Let $v \in \mathcal{D}(B_\delta)$ and $\tilde{v}(x) := v(\mathcal{T}^{-1}(x))$. Then*

$$\|v\|_{X(B)} \leq C \|\tilde{v}\|_{X(\Omega)}$$

and

$$\|\tilde{v}\|_{L^2(\Omega)} \leq C \|v\|_{L^2(B)}.$$

Proof. Let $\eta \in \mathcal{D}(B \setminus B_\delta)$ be such that $\eta(y) = 1$ for $|y| \leq \delta/2$. Since $v \in \mathcal{D}(B_\delta)$ we note that $\eta \frac{\partial v}{\partial y_i} = 0$. Hence

$$\left(\frac{\partial v}{\partial y_i}, \phi \right) = \left((1 - \eta) \frac{\partial v}{\partial y_i}, \phi \right) = \left(\frac{\partial v}{\partial y_i}, (1 - \eta)\phi \right).$$

Thus

$$\begin{aligned} \sup_{\phi \in \mathcal{D}(B)} \frac{\left(\frac{\partial v}{\partial y_i}, \phi \right)}{\|\phi\|_{H^1(B)}} &= \sup_{\phi \in \mathcal{D}(B)} \frac{\left(\frac{\partial v}{\partial y_i}, (1 - \eta)\phi \right)}{\|(1 - \eta)\phi\|_{H^1(B)}} \frac{\|(1 - \eta)\phi\|_{H^1(B)}}{\|\phi\|_{H^1(B)}} \\ &\leq C \sup_{\phi \in H_0^1(B_{\delta/2})} \frac{\left(\frac{\partial v}{\partial y_i}, \psi \right)}{\|\psi\|_{H^1(B)}} = C \sup_{\phi \in \mathcal{D}(B_{\delta/2})} \frac{\left(\frac{\partial v}{\partial y_i}, \psi \right)}{\|\psi\|_{H^1(B)}}. \end{aligned}$$

Next we make a change of variables.

$$\left(\frac{\partial v}{\partial y_i}, \psi \right) = \int_B \frac{\partial v}{\partial y_i} \psi \, dy = \sum_{j=1}^N \int_{\Omega} \frac{\partial \tilde{v}}{\partial x_j} \frac{\partial x_j}{\partial y_i} \tilde{\psi} J^{-1} \, dx,$$

where $J = [g(\epsilon(1 - \rho), y/\rho) - \epsilon \rho \frac{\partial g}{\partial h}(\epsilon(1 - \rho), y/\rho)] [g(\epsilon(1 - \rho), y/\rho)]^{N-1}$. Set $f_{ij} = \frac{\partial x_j}{\partial y_i} J^{-1}$. Then

$$\left(\frac{\partial v}{\partial y_i}, \psi \right) \leq \sum_{j=1}^N \left\| \frac{\partial \tilde{v}}{\partial x_j} \right\|_{H^{-1}(\Omega)} \|f_{ij} \tilde{\psi}\|_{H^1(\Omega)}.$$

Since

$$|D^\alpha f_{ij}| \leq C/(g(x/r) - r)$$

for $|\alpha| = 1$, it follows from Lemma 3.4 that

$$\|f_{ij} \tilde{\psi}\|_{H^1(\Omega)} \leq C \|\tilde{\psi}\|_{H^1(\Omega)} \leq C \|\psi\|_{H^1(B)}.$$

Hence

$$\left\| \frac{\partial v}{\partial y_i} \right\|_{H^{-1}(B)} \leq C \sum_{j=1}^N \left\| \frac{\partial \tilde{v}}{\partial x_j} \right\|_{H^{-1}(\Omega)}.$$

Clearly it follows in the same way that

$$\|v\|_{H^{-1}(B)} \leq C \|\tilde{v}\|_{H^{-1}(\Omega)}.$$

This proves the first inequality of the proposition. Since J is bounded, the second inequality of the proposition follows. \square

5. THE LEMMA FOR THE UNIT BALL

We want to prove that for the unit ball B in R^N

$$\|u\|_{L^2(B)} \leq C(\|u\|_{H^{-1}(B)} + \sum_{j=1}^N \left\| \frac{\partial u}{\partial x_j} \right\|_{H^{-1}(B)}).$$

Let $\eta \in \mathcal{D}(B)$ be such that $\eta(x) = 1$ for $|x| \leq 1/3$ and $\eta(x) = 0$ for $|x| \geq 1/2$. Then

$$\|\eta u\|_{L^2(B)} = \|\eta u\|_{L^2(R^N)} \leq \|\eta u\|_{X(R^N)}.$$

Now

$$\|\eta u\|_{X(R^N)} \leq \sup_{\phi \in \mathcal{D}(R^N)} \frac{(\eta u, \phi)}{\|\phi\|_1} + \sum_{j=1}^N \sup_{\phi \in \mathcal{D}(R^N)} \frac{(\frac{\partial}{\partial x_j}(\eta u), \phi)}{\|\phi\|_1} \leq C\|u\|_{X(B)}.$$

Hence we need to bound $v = (1 - \eta)u \in \mathcal{D}(B_{1/4})$ where $B_{1/4} = \{x | 1/4 < |x| < 1\}$. In order to prove that

$$\|v\|_{L^2(B)} \leq C\|u\|_{X(B)},$$

we define the following three operators, each of which maps $\mathcal{D}(R^N)$ to $H_D^1(B_{1/4})$, where $H_D^1(B_{1/4}) = \{v \in H^1(B_{1/4}) | v(x) = 0, |x| = 1\}$. For $\phi \in \mathcal{D}(R^N)$, define

$$\begin{aligned} P\phi &= \phi(x) + 3\phi\left(\frac{(2-r)}{r}x\right) - 4\phi\left(\frac{(3-2r)}{r}x\right), \\ \tilde{P}\phi &= \phi(x) - 3\phi\left(\frac{(2-r)}{r}x\right) + 2\phi\left(\frac{(3-2r)}{r}x\right) \end{aligned}$$

and

$$\hat{P}\phi = \phi(x) + 3[r/(2-r)]\phi\left(\frac{(2-r)}{r}x\right) - 4[r/(3-2r)]\phi\left(\frac{(3-2r)}{r}x\right).$$

The following relations hold:

$$\frac{\partial}{\partial r}(x_i/r) = 0, \left(x_j \frac{\partial}{\partial x_i} - x_i \frac{\partial}{\partial x_j} \right) r = 0.$$

Also

$$\frac{\partial}{\partial x_i} = \frac{x_i}{r} \frac{\partial}{\partial r} + \sum_{j=1}^N \frac{x_j}{r^2} \left(x_j \frac{\partial}{\partial x_i} - x_i \frac{\partial}{\partial x_j} \right).$$

By construction we have

$$P \frac{\partial \phi}{\partial r} = \frac{\partial}{\partial r}(\tilde{P}\phi)$$

and

$$P \left[\left(x_j \frac{\partial}{\partial x_i} - x_i \frac{\partial}{\partial x_j} \right) \phi \right] = \left(x_j \frac{\partial}{\partial x_i} - x_i \frac{\partial}{\partial x_j} \right) (\hat{P}\phi).$$

Define P^t by $(P^t v, \phi) := (v, P\phi)$ and note that for $v \in \mathcal{D}(B_{1/4})$ we have that $P^t v \in \mathcal{D}(R^N)$ and $P^t v = v$ in B . In fact the support of $P^t v$ is contained in the annulus $1/4 < r < 5/2$. Now we have

$$(\frac{\partial}{\partial x_i}(P^t v), \phi) = -(v, P(\frac{\partial \phi}{\partial x_i})) = - \left(v, P \left[\frac{\partial}{\partial r} \left(\frac{x_i}{r} \phi \right) \right] \right) - \sum_{j=1}^N \left(v, P \left[\frac{x_j}{r^2} \left(x_j \frac{\partial}{\partial x_i} - x_i \frac{\partial}{\partial x_j} \right) \right] \right).$$

Notice that if f is a function such that $f(x) = h(x/r)$ then for any $\alpha > 0$, $f(\alpha x) = f(x)$, i.e. $f(x)$ is homogeneous of degree 0. Hence for such a function $P[f\phi] = fP\phi$. Taking $f(x) = \frac{x_i}{r}$ we see that $P[\frac{x_i}{r}\phi] = \frac{x_i}{r}P\phi$. Thus

$$\left(v, P \left[\left(\frac{x_i}{r} \right) \frac{\partial \phi}{\partial r} \right] \right) = \left(v, \frac{x_i}{r} \left(\frac{\partial}{\partial r} \tilde{P}\phi \right) \right)$$

and

$$\left(v, P \left[\frac{x_j}{r} \left(\frac{x_j}{r} \frac{\partial}{\partial x_i} - \frac{x_i}{r} \frac{\partial}{\partial x_j} \right) \phi \right] \right) = \left(v, \frac{x_j}{r} \left(\frac{x_j}{r} \frac{\partial}{\partial x_i} - \frac{x_i}{r} \frac{\partial}{\partial x_j} \right) (\hat{P}\phi) \right).$$

We now see that

$$\left(\frac{\partial}{\partial x_i} (P^t v), \phi \right) = \left(\frac{x_i}{r} \frac{\partial v}{\partial r}, (\tilde{P}\phi - \hat{P}\phi) \right) + \left(\frac{\partial v}{\partial x_i}, \hat{P}\phi \right).$$

The mapping defined by $y_i = \frac{(2-r)}{r}x_i$, $r = |x|$, for $x \in B_{1/4}$ (and similarly the mapping $y_i = \frac{(3-2r)}{r}x_i$, $r = |x|$, for $x \in B_{1/4}$) is smooth. We verify this at the end of this section by computing its Jacobian. Therefore

$$\|P\phi\|_{H^1(B_{1/4})} + \|\tilde{P}\phi\|_{H^1(B_{1/4})} + \|\hat{P}\phi\|_{H^1(B_{1/4})} \leq C\|\phi\|_{H^1(R^N)}.$$

Hence

$$\begin{aligned} \|v\|_{L^2(B)} &= \|P^t v\|_{L^2(B)} \leq \|P^t v\|_{L^2(R^N)} \leq \|P^t v\|_{H^{-1}(R^N)} + \sum_{i=1}^N \left\| \frac{\partial}{\partial x_i} (P^t v) \right\|_{H^{-1}(R^N)} \\ &\leq \|v\|_{H^{-1}(B_{1/4})} + \sum_{i=1}^N \left\| \frac{\partial v}{\partial x_i} \right\|_{H^{-1}(B_{1/4})}. \end{aligned}$$

Thus we finally conclude that

$$\|v\|_{L^2(B)} \leq C\|v\|_{X(B)} \leq C\|u\|_{X(B)},$$

which completes the proof provided we check the smoothness of the the above-mentioned mappings.

To this end let

$$B_0 = \{x \in R^N | 0 < |x| < 1\}$$

and

$$B_1 = \{x \in R^N | 1 < |x| < 2\}.$$

Define

$$\mathcal{M} : B_0 \mapsto B_1$$

by

$$y_i = \frac{(2-r)}{r}x_i,$$

for $x \in B_0$ and $y \in B_1$. Now

$$b_{ij} := \frac{\partial y_i}{\partial x_j} = \frac{(2-r)}{r} \delta_{ij} - 2 \frac{x_i x_j}{r^3}$$

and the Jacobian of \mathcal{M} is

$$J = \det \left(\frac{\partial y_i}{\partial x_j} \right) = -((2-r)/r)^{N-1}.$$

To see this, note that the vector \mathbf{v}^1 , with components $\{v_i^1 = x_i/r, i = 1, \dots, N\}$, is an eigenvector with eigenvalue $\lambda_1 = -1$ of the matrix with entries b_{ij} ; i.e.

$$\sum_{j=1}^N b_{ij} v_j^1 = \frac{(2-r)}{r} v_i^1 - 2/r v_i^1 \sum_{j=1}^N v_j^1 v_j^1 = -v_i^1.$$

Now let $\{\mathbf{v}^k, k = 1, \dots, N\}$ be N orthonormal vectors, with entries $\{v_i^k, i = 1, \dots, N\}$ for each k ; i.e.

$$\sum_{i=1}^N v_i^k v_i^l = \delta_{kl}.$$

Then for $k > 1$

$$\sum_{j=1}^N b_{ij} v_j^k = \frac{(2-r)}{r} v_i^k - 2/r v_i^1 \sum_{j=1}^N v_j^1 v_j^k = \frac{(2-r)}{r} v_i^k.$$

Hence $\lambda_k = \frac{(2-r)}{r}$, for $k > 1$ and therefore

$$J = \det\left(\frac{\partial y_i}{\partial x_j}\right) = \lambda_1 \cdots \lambda_N = -((2-r)/r)^{N-1}.$$

6. APPENDIX I: EXISTENCE AND UNIQUENESS IN THE STOKES PROBLEM

We will show how (1.2) may be used to easily solve the following Stokes problem: For $\mathbf{f} \in [H^{-1}(\Omega)]^N$ and $g \in L_0^2(\Omega)$ find $\mathbf{u} \in [H_0^1(\Omega)]^N$ such that

$$(6.1) \quad D(\mathbf{u}, \mathbf{v}) + (p, \nabla \cdot \mathbf{v}) = (\mathbf{f}, \mathbf{v}) \quad \text{for all } \mathbf{v} \in [H_0^1(\Omega)]^N$$

and

$$(6.2) \quad (\nabla \cdot \mathbf{u}, q) = (g, q) \quad \text{for all } q \in L_0^2(\Omega).$$

Here $D(\cdot, \cdot)$ is the Dirichlet integral. Let T be the solution operator for the Dirichlet problem: $T : H^{-1}(\Omega) \rightarrow H_0^1(\Omega)$ defined by

$$D(Tf, \phi) = (f, \phi) \quad \text{for all } \phi \in H_0^1(\Omega).$$

The operator T is extended to vectors component-wise. Then it is easy to check that, taking $\mathbf{v} = T\nabla q$ in the first equation, a necessary condition is

$$(6.3) \quad (-\nabla \cdot T\nabla p, q) = (g - \nabla \cdot Tf, q) \quad \text{for all } q \in L_0^2(\Omega).$$

Noting that $D^{1/2}(\mathbf{v}, \mathbf{v})$ is equivalent to the norm on $[H_0^1(\Omega)]^N$ it follows easily that

$$\begin{aligned} (-\nabla \cdot T\nabla q, q)^{1/2} &= (T\nabla q, \nabla q)^{1/2} = D^{1/2}(T\nabla q, T\nabla q) \\ &= \sup_{\mathbf{v} \in [H_0^1(\Omega)]^N} \frac{(q, \nabla \cdot \mathbf{v})}{D^{1/2}(\mathbf{v}, \mathbf{v})} \geq C \|q\|_{L^2(\Omega)} \quad \text{for all } q \in L_0^2(\Omega). \end{aligned}$$

The bilinear form on the left hand side of (6.3) is continuous on $L_0^2(\Omega) \times L_0^2(\Omega)$ and by the last inequality it is coercive. Since $g - \nabla \cdot Tf \in L_0^2(\Omega)$ we may apply the Lax-Milgram Theorem and obtain the existence and uniqueness of $p \in L_0^2(\Omega)$

satisfying (6.3). With this p we can now solve (6.1) with $\mathbf{u} = Tf + T\nabla p$ and it follows that (6.2) is also satisfied. This solves the Stokes problem (6.1), (6.2).

7. APPENDIX II: KORN'S SECOND INEQUALITY

Let $\mathbf{u} \in [H^1(\Omega)]^N$ with components u_i and define the “strain tensor” $\epsilon_{ij} = 1/2(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i})$.

Theorem 7.1. *Let $\Omega \in R^N$ be a connected domain with a Lipschitz boundary. Then there exists a constant C such that,*

$$C\left(\sum_{i=1}^N \|u_i\|_{H^1(\Omega)}^2\right)^{1/2} \leq \left(\sum_{i=1}^N \|u_i\|_{L^2(\Omega)}^2\right)^{1/2} + \left(\sum_{i,j=1}^N \|\epsilon_{ij}\|_{L^2(\Omega)}^2\right)^{1/2},$$

for all $\mathbf{u} \in [H^1(\Omega)]^N$.

Proof. Since smooth functions are dense in $L^2(\Omega)$ and $H^1(\Omega)$ it suffices to consider only smooth vectors \mathbf{u} . The theorem follows by noting that, for i, j fixed and any k ,

$$\frac{\partial^2 u_i}{\partial x_j \partial x_k} = \frac{\partial \epsilon_{ik}}{\partial x_j} + \frac{\partial \epsilon_{ij}}{\partial x_k} - \frac{\partial \epsilon_{jk}}{\partial x_i}$$

and applying the Theorem 1.1 to $u = \frac{\partial u_i}{\partial x_j}$. \square

REFERENCES

- [1] C. Bernardi, Méthode d’éléments finis pour les équations de Navier-Stokes, Thèse. Univ. Paris. (1979).
- [2] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer, New York, 1991.
- [3] G. Duvaut and J.L. Lions, *Inequalities in Mechanics and Physics*, Series of Comprehensive Studies in Math., 219, Springer, 1976.
- [4] G. Fichera, Sull’esistenza e sul calcolo delle soluzioni dei problemi al contorno, relativi all’equilibrio di un corpo elastico, Ann. Scuola Normale Sup. Pisa s. III 4 (1950).
- [5] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier-Stokes equations*, Springer Series in Computational Mathematics, 5, Springer, 1986.
- [6] J. Gobert, Une inégalité fondamentale de la théorie de l’élasticité, Bull. Soc. Royale Science Liège, **31** année, No 3-4 (1962) 182-191.
- [7] O.A. Ladyzhenskaya, *The Mathematical Theory of Viscous Incompressible Flows*, Gordon and Breach, London, 1969.
- [8] J. Nitsche, On Korn’s second inequality, R.A.I.R.O. 15 (1981) 237-248.
- [9] J. Nečas, *Les Méthodes Directes en Théorie des Équations Elliptiques*, Masson, 1967.
- [10] R. Temam, *Navier-Stokes Equations*, North-Holland, New York, 1977.

JAMES H. BRAMBLE, DEPARTMENT OF MATHEMATICS, TEXAS A&M UNIVERSITY,
COLLEGE STATION, TX 77843
E-mail address: bramble@math.tamu.edu

2.13 Super-convergence

Multigrid Methods

3.1 New convergence estimates for multigrid algorithms

New convergence estimates for multigrid algorithms[[16](#)]

New Convergence Estimates for Multigrid Algorithms

By James H. Bramble and Joseph E. Pasciak*

Abstract. In this paper, new convergence estimates are proved for both symmetric and nonsymmetric multigrid algorithms applied to symmetric positive definite problems. Our theory relates the convergence of multigrid algorithms to a “regularity and approximation” parameter $\alpha \in (0, 1]$ and the number of relaxations m . We show that for the symmetric and nonsymmetric V cycles, the multigrid iteration converges for any positive m at a rate which deteriorates no worse than $1 - c j^{-(1-\alpha)/\alpha}$, where j is the number of grid levels. We then define a generalized V cycle algorithm which involves exponentially increasing (for example, doubling) the number of smoothings on successively coarser grids. We show that the resulting symmetric and nonsymmetric multigrid iterations converge for any α with rates that are independent of the mesh size. The theory is presented in an abstract setting which can be applied to finite element multigrid and finite difference multigrid methods.

1. Introduction. In recent years, multigrid methods have been used extensively as tools for obtaining approximations to solutions of partial differential equations (see the references in [5], [9]). In conjunction, there has been intensive research into the theoretical understanding of the convergence properties of these methods (cf. [2], [3], [4], [9], [11], [12]–[16], [18]). This paper will present a number of new results on the convergence of multigrid algorithms.

We shall be concerned with the analysis of many-level multigrid schemes. The first approach to this problem involved Fourier analysis and only applied to rather limited situations, i.e., rectangular domains [8]. More general results can be obtained by variational or finite element like formulations of multigrid. One approach is to obtain results for two-grid schemes and use those results to derive estimates for the many-level schemes [3], [9]. The problem with this technique is that it only leads to results for W cycles with sufficiently large m (the number of smoothing iterations). Another interesting approach was taken in [4] in which a direct analysis was made for the many-level scheme. This technique leads to results for the V cycle with any m but under the assumption of ‘full elliptic regularity’, i.e., $\alpha = 1$. Results for the W cycle algorithm for any m and α were given in [13], [14]. The theory presented in this paper also provides a direct analysis of many-level schemes.

Received January 5, 1987.

1980 *Mathematics Subject Classification* (1985 Revision). Primary 65N30; Secondary 65F10.

*This manuscript has been authored under contract number DE-AC02-76CH00016 with the U.S. Department of Energy. Accordingly, the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. This work was also supported in part under the National Science Foundation Grant No. DMS84-05352 and under the Air Force Office of Scientific Research, Contract No. ISSA86-0026 and by the U.S. Army Research Office through the Mathematical Science Institute, Cornell University.

©1987 American Mathematical Society
0025-5718/87 \$1.00 + \$.25 per page

We shall provide an analysis for symmetric and nonsymmetric multigrid cycling algorithms applied to symmetric positive definite problems. We relate our convergence estimates to a certain discrete regularity and approximation assumption which involves the parameter $\alpha \in (0, 1]$. We will show that both the symmetric and nonsymmetric \mathcal{V} cycle algorithms converge for any m and any α at a rate

$$(1.1) \quad \delta_j \leq \frac{M_\alpha(j_0 + j)^{(1-\alpha)/\alpha}}{m^\alpha + M_\alpha(j_0 + j)^{(1-\alpha)/\alpha}}$$

per iteration. Here M_α and j_0 are constants which are independent of j , the number of grid levels. In finite element and finite difference applications, j is proportional to $\log(1/h)$ where h is essentially the size of the finest grid. In such applications, our theorems guarantee that the convergence rate can only deteriorate like $1 - c/\log^{(1-\alpha)/\alpha}(1/h)$ as $h \rightarrow 0$. The estimates improve as m becomes large.

We next consider the symmetric and nonsymmetric \mathcal{W} cycle algorithms. Although the uniform convergence (independent of the number of levels) of this algorithm for $\alpha = 1$ and any m was shown in [13], [14], we derive convergence estimates in the context of our framework. We prove a convergence bound of the form

$$(1.2) \quad \delta \leq (1 + m/M_\alpha)^{-\alpha}.$$

Again, we have convergence for any m and any α , but for the \mathcal{W} cycle the rate is also bounded independently of the number of levels j .

Finally, we introduce a generalized \mathcal{V} cycle algorithm and derive the corresponding convergence estimates. In this algorithm, the number of smoothings varies on different grid levels. One example doubles the number of smoothings on each consecutive coarser grid level. Note that this doubling strategy gives rise to the same number of smoothings as the \mathcal{W} cycle but is simpler to code. We show that the generalized \mathcal{V} cycle also converges for any α with rates that are independent of j (see Theorems 5 and 6).

In our algorithms and analysis, the smoothing process is defined in terms of a general smoothing operator. Consequently, our analysis is applicable to many of the smoothing processes used in actual multigrid applications. We should however note that according to our analysis, little appears to be gained by using smoothing operators which are more complex than a weighted identity, although in practice some improvement may result.

The outline of the remainder of the paper is as follows. In Section 2 we define the multigrid algorithms. These algorithms are described by a simple induction process and lead to linear operators which ‘approximate’ the inverse of the problem to which they are applied. In Section 3 we give the multigrid convergence analysis. In Section 4 we show how the theorems of Section 3 can be used to guarantee rapid convergence of preconditioned iterative schemes using the multigrid operator as a preconditioner. Finite element and finite difference applications are considered in Section 5. In addition, we show that for finite differences, symmetric Gauss-Seidel iteration leads to a smoothing operator which satisfies all of the hypotheses of the theorem. Finally, in Section 6, we give the results of numerically computed convergence factors for multigrid algorithms. We also give numerical evidence which suggests that the ‘regularity and approximation’ assumption does not in general hold for all α . Hence, the new convergence estimates presented in this paper for $\alpha < 1$ are of theoretical importance.

2. The Multigrid Algorithms. In this section we describe both the symmetric and nonsymmetric multigrid cycling algorithms. Along the way, we derive

basic recursion relations which play major roles in the analysis of the methods. For convenience, the algorithms are developed in an abstract Hilbert space setting. The results most naturally apply to finite element multigrid algorithms but can also be applied to certain formulations of finite difference multigrid algorithms.

Let us assume that we are given a nested sequence of finite-dimensional vector spaces

$$\mathcal{M}_0 \subset \mathcal{M}_1 \subset \cdots \subset \mathcal{M}_j.$$

In addition, let $A(\cdot, \cdot)$ and $(\cdot, \cdot)_k$ be symmetric positive definite bilinear forms on \mathcal{M}_k for $k = 0, \dots, j$. We shall develop multigrid algorithms for the solution of the problem: Given $f \in \mathcal{M}_j$, find $v \in \mathcal{M}_j$ satisfying

$$(2.1) \quad A(v, \phi) = (f, \phi)_j \quad \text{for all } \phi \in \mathcal{M}_j.$$

To define the multigrid algorithms, we shall define auxiliary operators. For $k = 0, \dots, j$, define the operator $A_k: \mathcal{M}_k \mapsto \mathcal{M}_k$ by

$$(A_k w, \phi)_k = A(w, \phi) \quad \text{for all } \phi \in \mathcal{M}_k.$$

The operator A_k is clearly symmetric (in both the $A(\cdot, \cdot)$ and $(\cdot, \cdot)_k$ -inner products) and positive definite. Also define the projectors $P_k: \mathcal{M}_{k+1} \mapsto \mathcal{M}_k$ and $P_k^0: \mathcal{M}_{k+1} \mapsto \mathcal{M}_k$ by

$$A(P_k w, \phi) = A(w, \phi) \quad \text{for all } \phi \in \mathcal{M}_k,$$

and

$$(P_k^0 w, \phi)_k = (w, \phi)_{k+1} \quad \text{for all } \phi \in \mathcal{M}_k.$$

Note that P_k is symmetric in the A -inner product.

To define the smoothing process, we require a linear operator $R_k: \mathcal{M}_k \mapsto \mathcal{M}_k$ for $k = 1, \dots, j$. We assume that R_k is symmetric in the $(\cdot, \cdot)_k$ -inner product and set $K_k = (I - R_k A_k)$. We further assume that K_k is nonnegative in the sense that $A(K_k u, u) \geq 0$ for all $u \in \mathcal{M}_k$.

We first define the symmetric multigrid operator $B_k^s: \mathcal{M}_k \mapsto \mathcal{M}_k$ by induction.

Algorithm S.

Set $B_0^s = A_0^{-1}$. Assume that B_{k-1}^s has been defined and define $B_k^s g$ for $g \in \mathcal{M}_k$ as follows:

- (1) Set $x^0 = 0$ and $q^0 = 0$.
- (2) Define x^l for $l = 1, \dots, m(k)$ by

$$(2.2) \quad x^l = x^{l-1} + R_k(g - A_k x^{l-1}).$$

- (3) Define $x^{m(k)+1} = x^{m(k)} + q^p$ where q^i for $i = 1, \dots, p$ is defined by

$$q^i = q^{i-1} + B_{k-1}^s [P_{k-1}^0(g - A_k x^{m(k)}) - A_{k-1} q^{i-1}].$$

- (4) Set $B_k^s g = x^{2m(k)+1}$ where x^l is defined for $l = m(k) + 2, \dots, 2m(k) + 1$ by (2.2).

In this algorithm, $m(k)$ is a positive integer which may vary from level to level and determines the number of smoothing iterations on that level. Because of this variable smoothing, the above algorithm is more general than that usually described [2], [3], [5], [9]. If all of the $m(k)$ are the same, then this algorithm is the usual symmetric multigrid algorithm described in a notation which is convenient for our analysis. Note that B_k^s is clearly a linear operator for each k . In this algorithm, p is a positive integer. We shall study the cases $p = 1$ and $p = 2$ which correspond

respectively to the symmetric \mathcal{V} and \mathcal{W} cycles of multigrid. Generalizations to $p > 2$ are straightforward and will not be considered.

The definition of the nonsymmetric multigrid operator B_k^n is similar except that the smoothings of Step 4 are excluded. More precisely, we define $B_k^n: \mathcal{M}_k \mapsto \mathcal{M}_k$ by induction.

Algorithm N.

Set $B_0^n = A_0^{-1}$. Assume that B_{k-1}^n has been defined and define $B_k^n g$ for $g \in \mathcal{M}_k$ as follows:

- (1) Set $x^0 = 0$ and $q^0 = 0$.
- (2) Define x^l for $l = 1, \dots, m(k)$ by (2.2).
- (3) Define $B_k^n g = x^{m(k)} + q^p$ where q^i for $i = 1, \dots, p$ is defined by

$$(2.3) \quad q^i = q^{i-1} + B_{k-1}^n [P_{k-1}^0(g - A_k x^{m(k)}) - A_{k-1} q^{i-1}].$$

The above algorithm defines a linear operator B_k^n which is equivalent to the standard nonsymmetric multigrid algorithms described in [3], [9] when $m(k)$ is constant.

Remark 2.1. One computationally effective algorithm with variable m is the \mathcal{V} cycle algorithm ($p = 1$) with $m(k) = m_0 2^{j-k}$. Note that, for this algorithm, the total number of smoothing iterations on each level is the same as that for the \mathcal{W} cycle ($p = 2$) with $m(k) = m_0$ for all k . We shall prove in Section 3 (see Theorems 5 and 6) that, like the corresponding \mathcal{W} cycle, this \mathcal{V} cycle converges for any m_0 and α with rates that are independent of j . In addition, this generalized \mathcal{V} cycle is easier to implement.

Let $g = A_k x$. It is straightforward to check that q^p defined by (2.3) satisfies

$$(2.4) \quad q^p = (I - (I - B_{k-1}^n A_{k-1})^p) A_{k-1}^{-1} P_{k-1}^0 A_k (x - x^{m(k)}).$$

A trivial computation gives that

$$(2.5) \quad x - x^{m(k)} = K_k^{m(k)} x.$$

Noting that $P_{k-1}^0 A_k = A_{k-1} P_{k-1}$ and combining (2.4) and (2.5) gives

$$(2.6) \quad I - B_k^n A_k = [(I - P_{k-1}) + (I - B_{k-1}^n A_{k-1})^p P_{k-1}] K_k^{m(k)}.$$

Equation (2.6) gives a fundamental recurrence relation for the nonsymmetric multigrid algorithms. The analogous recurrence in the symmetric case is

$$(2.7) \quad I - B_k^s A_k = K_k^{m(k)} [(I - P_{k-1}) + (I - B_{k-1}^s A_{k-1})^p P_{k-1}] K_k^{m(k)},$$

which follows from similar reasonings.

Note that (2.7) implies that

$$(2.8) \quad \begin{aligned} A((I - B_k^s A_k)u, v) &= A((I - P_{k-1})K_k^{m(k)} u, K_k^{m(k)} v) \\ &\quad + A((I - B_{k-1}^s A_{k-1})^p P_{k-1} K_k^{m(k)} u, K_k^{m(k)} v). \end{aligned}$$

Remark 2.2. An obvious argument using (2.8) and induction gives that $I - B_k^s A_k$ is a symmetric operator in the A -inner product. Consequently, B_k^s is symmetric in $(\cdot, \cdot)_k$.

Remark 2.3. A similar argument gives that

$$(2.9) \quad A((I - B_k^s A_k)u, u) \geq 0 \quad \text{for all } u \in \mathcal{M}_k.$$

We finally note that by the definition of P_k ,

$$(2.10) \quad \begin{aligned} A((I - B_k^n A_k)u, (I - B_k^n A_k)u) \\ = A((I - P_{k-1})K_k^{m(k)} u, K_k^{m(k)} u) \\ + A((I - B_{k-1}^n A_k)^p P_{k-1} K_k^{m(k)} u, (I - B_{k-1}^n A_k)^p P_{k-1} K_k^{m(k)} u). \end{aligned}$$

3. Multigrid Analysis. In this section we give an analysis of the multigrid algorithms described in the previous section. The goal of this section is to prove norm inequalities of the form

$$(3.1) \quad A((I - B_k^s A_k)u, u) \leq \delta_k A(u, u) \quad \text{for all } u \in \mathcal{M}_k,$$

and

$$(3.2) \quad A((I - B_k^n A_k)u, (I - B_k^n A_k)u) \leq \delta_k A(u, u) \quad \text{for all } u \in \mathcal{M}_k.$$

We shall relate δ_k to two *a priori* assumptions. Let $0 < \alpha \leq 1$. The first assumption is a “regularity and approximation” assumption of the form

$$(3.3) \quad A((I - P_{k-1})u, u) \leq C_\alpha^2 \left(\frac{\|A_k u\|_k^2}{\lambda_k} \right)^\alpha A(u, u)^{1-\alpha} \quad \text{for all } u \in \mathcal{M}_k,$$

where λ_k is the largest eigenvalue of A_k . More precisely, we assume that (3.3) holds with C_α independent of k for $k = 1, \dots, j$. As will be demonstrated in Section 5, in finite element applications, the derivation of inequalities of the form (3.3) uses regularity estimates for the elliptic operator being approximated and the approximation properties of the subspace. The second assumption is that the smoothing operator R_k satisfies the inequality

$$(3.4) \quad \frac{\|u\|_k^2}{\lambda_k} \leq C_R (R_k u, u)_k \quad \text{for all } u \in \mathcal{M}_k.$$

Again, we assume that (3.4) holds with a constant C_R independent of k . By an obvious change of variable, we see that (3.4) is equivalent to

$$(3.5) \quad \frac{\|A_k u\|_k^2}{\lambda_k} \leq C_R A((I - K_k)u, u) \quad \text{for all } u \in \mathcal{M}_k.$$

Remark 3.1. Note that a simple choice of the smoothing operator is $R_k = \bar{\lambda}_k^{-1} I_k$ where $\bar{\lambda}_k$ is any upper bound for λ_k (we need K_k nonnegative) and I_k is the identity on \mathcal{M}_k . Furthermore, if one takes $\bar{\lambda}_k = \lambda_k$ then (3.4) holds with $C_R = 1$.

We can now state and prove the theorem for estimating δ_k in (3.1) for the symmetric \mathcal{V} cycle.

THEOREM 1. *Assume that (3.3) and (3.4) hold and define B_j^s by Algorithm S with $p = 1$ and $m(k) = m$ for all k . Then (3.1) holds with*

$$(3.6) \quad \delta_j = 1 - \varepsilon_j,$$

where

$$(3.7) \quad \varepsilon_j = \frac{m^\alpha}{m^\alpha + M_\alpha(j_0 + j)^{(1-\alpha)/\alpha}}.$$

In (3.7), j_0 is any positive constant and M_α is defined by first defining

$$(3.8) \quad \tilde{M}_\alpha = \left(\frac{1 + j_0}{j_0} \right)^s \frac{C_R (\alpha C_\alpha^2)^{1/\alpha}}{2},$$

where

$$(3.9) \quad s = \begin{cases} \frac{1-\alpha}{\alpha} & \text{for } \alpha \geq 1/2, \\ \left(\frac{1-\alpha}{\alpha}\right)^2 & \text{for } \alpha < 1/2, \end{cases}$$

and then setting

$$(3.10) \quad M_\alpha = \left(\frac{1 + \tilde{M}_\alpha}{\tilde{M}_\alpha} \right)^{(1-\alpha)/\alpha} \tilde{M}_\alpha.$$

Proof. We shall prove that

$$(3.11) \quad A((I - B_i^s A_i)u, u) \leq \delta_i A(u, u) \quad \text{for all } u \in \mathcal{M}_i,$$

by induction on i . For $i = 0$, there is nothing to prove. Assume that (3.11) holds for $i = k - 1$. By (2.8) and the induction hypothesis,

$$(3.12) \quad \begin{aligned} A((I - B_k^s A_k)u, u) &\leq A((I - P_{k-1})K_k^m u, K_k^m u) \\ &\quad + \delta_{k-1} A(P_{k-1} K_k^m u, K_k^m u) \end{aligned}$$

$$(3.13) \quad = (1 - \delta_{k-1}) A((I - P_{k-1})K_k^m u, K_k^m u) + \delta_{k-1} A(K_k^m u, K_k^m u).$$

By (3.3) and a generalized arithmetic-geometric mean inequality,

$$(3.14) \quad \begin{aligned} &A((I - P_{k-1})K_k^m u, K_k^m u) \\ &\leq C_\alpha^2 \left\{ \alpha \gamma_k \frac{\|A_k K_k^m u\|_k^2}{\lambda_k} + (1 - \alpha) \gamma_k^{-\alpha/(1-\alpha)} A(K_k^m u, K_k^m u) \right\} \end{aligned}$$

holds for any positive γ_k . By (3.5) and the symmetry of K_k ,

$$(3.15) \quad \frac{\|A_k K_k^m u\|_k^2}{\lambda_k} \leq C_R A((I - K_k) K_k^{2m} u, u).$$

Since the spectrum of K_k is contained in the interval $[0,1]$,

$$(3.16) \quad \begin{aligned} A((I - K_k) K_k^{2m} u, u) &\leq \frac{1}{2m} \sum_{i=0}^{2m-1} A((I - K_k) K_k^i u, u) \\ &= \frac{1}{2m} A((I - K_k^{2m}) u, u). \end{aligned}$$

Combining (3.13)–(3.16) gives

$$(3.17) \quad \begin{aligned} A((I - B_k^s A_k)u, u) &\leq [(1 - \delta_{k-1})C_\alpha^2(1 - \alpha)\gamma_k^{-\alpha/(1-\alpha)} + \delta_{k-1}] A(K_k^{2m} u, u) \\ &\quad + (1 - \delta_{k-1})C_\alpha^2 C_R \frac{\alpha}{2m} \gamma_k A((I - K_k^{2m}) u, u). \end{aligned}$$

To prove that (3.11) holds for k , it suffices to choose γ_k so that

$$(3.18) \quad (1 - \delta_{k-1})C_\alpha^2(1 - \alpha)\gamma_k^{-\alpha/(1-\alpha)} + \delta_{k-1} \leq \delta_k$$

and

$$(3.19) \quad (1 - \delta_{k-1})C_\alpha^2 C_R \frac{\alpha}{2m} \gamma_k \leq \delta_k.$$

We set γ_k by

$$(3.20) \quad (1 - \delta_{k-1})C_\alpha^2 C_R \frac{\alpha}{2m} \gamma_k = \delta_{k-1}.$$

Note that by definition, δ_k is an increasing function of k and hence (3.20) immediately implies (3.19). We need only verify that (3.18) holds for this choice of γ_k . This is equivalent to showing

$$(3.21) \quad (1 - \delta_{k-1}) C_\alpha^2 (1 - \alpha) \gamma_k^{-\alpha/(1-\alpha)} \leq \delta_k - \delta_{k-1}.$$

Let $\mathcal{D}(k) = m^\alpha + M_\alpha(j_0 + k)^{(1-\alpha)/\alpha}$. A direct computation gives that

$$(3.22) \quad \delta_k - \delta_{k-1} = \frac{M_\alpha m^\alpha [(j_0 + k)^{(1-\alpha)/\alpha} - (j_0 + k - 1)^{(1-\alpha)/\alpha}]}{\mathcal{D}(k)\mathcal{D}(k-1)}.$$

The left-hand side of (3.21) can be written

$$(3.23) \quad (C_\alpha^2)^{1/(1-\alpha)} \frac{(1 - \alpha)}{(j_0 + k - 1)\mathcal{D}(k-1)} \left(\frac{C_R \alpha}{2M_\alpha} \right)^{\alpha/(1-\alpha)}.$$

A straightforward exercise in calculus, noting that $k \geq 1$, gives

$$(3.24) \quad \begin{aligned} & (j_0 + k - 1)[(j_0 + k)^{(1-\alpha)/\alpha} - (j_0 + k - 1)^{(1-\alpha)/\alpha}] \\ & \geq \left(\frac{1 - \alpha}{\alpha} \right) (j_0 + k)^{(1-\alpha)/\alpha} \left(\frac{j_0}{1 + j_0} \right)^l, \end{aligned}$$

where

$$l = \begin{cases} 1 & \text{if } \alpha \geq 1/2, \\ \frac{1 - \alpha}{\alpha} & \text{if } \alpha < 1/2. \end{cases}$$

Combining (3.22)–(3.24) shows that (3.21) holds if

$$(3.25) \quad \begin{aligned} & (\alpha C_\alpha^2)^{1/(1-\alpha)} \left(\frac{C_R}{2} \right)^{\alpha/(1-\alpha)} \\ & \leq M_\alpha^{\alpha/(1-\alpha)} \left(\frac{j_0}{1 + j_0} \right)^l \frac{m^\alpha M_\alpha (j_0 + k)^{(1-\alpha)/\alpha}}{\mathcal{D}(k)}. \end{aligned}$$

We note that \tilde{M}_α was chosen so that

$$(\alpha C_\alpha^2)^{1/(1-\alpha)} \left(\frac{C_R}{2} \right)^{\alpha/(1-\alpha)} = \left(\frac{j_0}{1 + j_0} \right)^l \tilde{M}_\alpha^{\alpha/(1-\alpha)}.$$

The definition of M_α implies

$$(\alpha C_\alpha^2)^{1/\alpha} \left(\frac{C_R}{2} \right)^{\alpha/(1-\alpha)} \leq \left(\frac{j_0}{1 + j_0} \right)^l M_\alpha^{\alpha/(1-\alpha)} \frac{M_\alpha}{1 + M_\alpha}.$$

Inequality (3.25) then follows from

$$\frac{M_\alpha}{1 + M_\alpha} \leq \frac{m^\alpha M_\alpha (j_0 + k)^{(1-\alpha)/\alpha}}{\mathcal{D}(k)}.$$

This completes the proof of Theorem 1.

Remark 3.2. As is clear from the proof of the theorem, any positive j_0 leads to a bound for δ_j with ε_j given by (3.7). The best choice is evidently a value which minimizes δ_j .

Our second theorem applies to the nonsymmetric \mathcal{V} cycle.

THEOREM 2. *Assume that (3.3) and (3.4) hold and define B_j^n by Algorithm N with $p = 1$ and $m(k) = m$ for all k . Then (3.2) holds with*

$$\delta_j = 1 - \varepsilon_j,$$

where ε_j is defined as in Theorem 1 (see (3.7)–(3.10)).

Proof. We shall prove that

$$(3.26) \quad A((I - B_i^n A_i)u, (I - B_i^n A_i)u) \leq \delta_i A(u, u) \quad \text{for all } u \in M_i,$$

by induction on i . For $i = 0$, there is nothing to prove. Assume that (3.26) holds for $i = k - 1$. By (2.10) and the induction hypothesis,

$$(3.27) \quad \begin{aligned} A((I - B_k^n A_k)u, (I - B_k^n A_k)u) &\leq A((I - P_{k-1})K_k^m u, K_k^m u) \\ &\quad + \delta_{k-1} A(P_{k-1} K_k^m u, K_k^m u). \end{aligned}$$

Notice that the terms on the right-hand sides of Eqs. (3.12) and (3.27) are identical. The arguments following (3.12) show that the right-hand side of (3.27) can be bounded by δ_k . This completes the proof of Theorem 2.

Remark 3.3. Theorems 1 and 2 give bounds for the convergence factor δ_k in (3.1) and (3.2) in terms of α , m and j . In finite element applications, j (the number of levels) is proportional to a logarithm of the mesh size. Theorems 1 and 2 show that the multigrid \mathcal{V} cycles still converge for any m but the convergence rate may deteriorate with a power of j depending upon the α for which (3.3) holds. A discussion of the relationship between α and domain/operator regularity is given in Section 5 (see Proposition 5.1).

Remark 3.4. The theorems guarantee that the convergence factor δ_j goes to zero faster than c/m^α when j is held fixed and m tends to infinity. These rates are in agreement with the results of numerical computations presented in Section 6.

Remark 3.5. The results given in Theorems 1 and 2 for $\alpha < 1$ are new. The proof gives rise to a somewhat simpler analysis than that already in the literature [2], [4], [9] for $\alpha = 1$. Indeed, by (3.17), if $\alpha = 1$, it suffices to take δ_k to be the solution δ of

$$(1 - \delta)C_\alpha^2 \frac{C_R}{2m} = \delta,$$

i.e.,

$$(3.28) \quad \delta = \frac{C_\alpha^2 C_R}{2m + C_\alpha^2 C_R}.$$

This value for δ agrees with the results derived by the earlier analysis [2], [4]. Note also that the expression for δ_j given by (3.6) tends to (3.28) as $\alpha \rightarrow 1$.

We next give results for the \mathcal{W} cycle multigrid algorithms. As previously mentioned, the uniform convergence of the \mathcal{W} cycle algorithm independent of the number of levels was shown in [13], [14]. Our results give convergence bounds which exhibit the explicit dependence on α and m .

THEOREM 3. *Assume that (3.3) and (3.4) hold and define B_j^s by Algorithm S with $p = 2$ and $m(k) = m$ for all k . Then (3.1) holds with $\delta_k = \delta$ (independent of k) given by*

$$(3.29) \quad \delta = (1 + m/M_\alpha)^{-\alpha},$$

where

$$(3.30) \quad M_\alpha = 2^{1/\alpha} (C_\alpha^2)^{1/\alpha} \frac{\alpha C_R}{2} (1 - \alpha)^{(1-\alpha)/\alpha}.$$

Proof. We proceed as in the proof of Theorem 1. However, since $p = 2$, (3.12) and (3.13) get replaced by

$$(3.31) \quad A((I - B_k^s A_k)u, u) \leq A((I - P_{k-1})K_k^m u, K_k^m u) \\ + \delta^2 A(P_{k-1} K_k^m u, K_k^m u)$$

$$(3.32) \quad \leq (1 - \delta^2)A((I - P_{k-1})K_k^m u, K_k^m u) + \delta^2 A(K_k^m u, K_k^m u).$$

The same reasoning which leads to (3.17) gives that

$$(3.33) \quad A((I - B_k^s A_k)u, u) \leq [(1 - \delta^2)C_\alpha^2(1 - \alpha)\gamma^{-\alpha/(1-\alpha)} + \delta^2]A(K_k^{2m} u, u) \\ + (1 - \delta^2)C_\alpha^2 C_R \frac{\alpha}{2m} \gamma A((I - K_k^{2m})u, u)$$

holds for any positive γ . We define γ by the equation

$$(3.34) \quad (1 - \delta^2)C_\alpha^2 C_R \frac{\alpha}{2m} \gamma = \delta.$$

It then suffices to show that for γ defined by (3.34),

$$(1 - \delta^2)C_\alpha^2(1 - \alpha)\gamma^{-\alpha/(1-\alpha)} + \delta^2 \leq \delta,$$

or

$$(3.35) \quad (1 - \delta)^{\alpha/(1-\alpha)} (C_\alpha^2)^{1/(1-\alpha)} (1 - \alpha) \left(\frac{C_R \alpha}{2m} \right)^{\alpha/(1-\alpha)} \leq \left(\frac{\delta}{1 + \delta} \right)^{1/(1-\alpha)}.$$

A straightforward manipulation shows that (3.35) will be satisfied if

$$(3.36) \quad 2^{-1/\alpha} M_\alpha = (C_\alpha^2)^{1/\alpha} \frac{\alpha C_R}{2} (1 - \alpha)^{(1-\alpha)/\alpha} \leq \left(\frac{\delta}{1 + \delta} \right)^{1/\alpha} m(1 - \delta)^{-1}.$$

However, for $y = 1/\delta$,

$$(3.37) \quad \left(\frac{\delta}{1 + \delta} \right)^{1/\alpha} \frac{m}{M_\alpha} (1 - \delta)^{-1} = \frac{y^{1+1/\alpha} - y}{(y - 1)(1 + y)^{1/\alpha}}$$

$$(3.38) \quad \geq \frac{y^{1/\alpha}}{(1 + y)^{1/\alpha}} \geq 2^{-1/\alpha},$$

where we used $y^{1/\alpha} > y$ for $y > 1$. The theorem follows from (3.36) and (3.38).

THEOREM 4. *Assume that (3.3) and (3.4) hold and define B_j^n by Algorithm N with $p = 2$ and $m(k) = m$ for all k . Then (3.2) holds with $\delta_k = \delta$ given by (3.29).*

Proof. The proof of Theorem 4 is a slight modification of the proof of Theorem 3. We use (2.10) to get

$$(3.39) \quad A((I - B_k^s A_k)u, (I - B_k^s A_k)u) \leq A((I - P_{k-1})K_k^m u, K_k^m u) \\ + \delta^2 A(P_{k-1} K_k^m u, K_k^m u).$$

Notice that the terms on the right-hand sides of Eqs. (3.31) and (3.39) are identical. The arguments following (3.31) show that the right-hand side of (3.39) can be bounded by δ . This completes the proof of Theorem 4.

Remark 3.6. Theorems 3 and 4 show that the multigrid \mathcal{W} cycles converge for any m and any α . The convergence estimates tend to zero as m gets larger and deteriorate as α tends to zero.

We next consider a generalized symmetric \mathcal{V} cycle.

THEOREM 5. *Assume that (3.3) and (3.4) hold and define B_j^s by Algorithm S with $p = 1$. Assume that $m(k)$ satisfies*

$$(3.40) \quad \beta_0 m(k) \leq m(k-1) \leq \beta_1 m(k).$$

Here we assume that β_0 and β_1 are constants which are greater than one and independent of k . Then (3.1) holds with

$$(3.41) \quad \delta_j = 1 - \varepsilon_j,$$

where

$$(3.42) \quad \varepsilon_j = \frac{m(j)^\alpha}{m(j)^\alpha + M_\alpha}.$$

In (3.42), M_α is defined by

$$M_\alpha = \left(\frac{1 + \tilde{M}_\alpha}{\tilde{M}_\alpha} \right)^{(1-\alpha)/\alpha} \tilde{M}_\alpha,$$

where

$$\tilde{M}_\alpha = (C_\alpha^2)^{1/\alpha} (1 - \alpha)^{(1-\alpha)/\alpha} \left(\frac{\alpha C_R \beta_1}{2} \right) (\beta_0^\alpha - 1)^{-(1-\alpha)/\alpha}.$$

Proof. The proof of this theorem follows along the lines of the proof of Theorem 1. In fact, following the arguments in the proof of Theorem 1 (see (3.21)), we see that it suffices to show that

$$(3.43) \quad (1 - \delta_{k-1}) C_\alpha^2 (1 - \alpha) \gamma_k^{-\alpha/(1-\alpha)} \leq \delta_k - \delta_{k-1}$$

holds with δ_k defined by (3.41)–(3.42), where γ_k is defined by

$$(3.44) \quad (1 - \delta_{k-1}) C_\alpha^2 C_R \frac{\alpha}{2m(k)} \gamma_k = \delta_{k-1}.$$

Let $D(k) = m(k)^\alpha + M_\alpha$. A direct computation gives that

$$(3.45) \quad \delta_k - \delta_{k-1} \geq \frac{M_\alpha m(k)^\alpha (\beta_0^\alpha - 1)}{D(k) D(k-1)}.$$

The left-hand side of (3.43) can be written

$$(3.46) \quad (C_\alpha^2)^{1/(1-\alpha)} \frac{(1 - \alpha)}{D(k-1)} \left(\frac{\alpha C_R m(k-1)}{2M_\alpha m(k)} \right)^{\alpha/(1-\alpha)}.$$

Combining (3.45) and (3.46) shows that (3.43) holds if

$$(3.47) \quad (C_\alpha^2)^{1/(1-\alpha)} (1 - \alpha) \left(\frac{\alpha C_R \beta_1}{2} \right)^{\alpha/(1-\alpha)} \leq (\beta_0^\alpha - 1) M_\alpha^{\alpha/(1-\alpha)} \frac{m(k)^\alpha M_\alpha}{D(k)}.$$

However, the definitions of \tilde{M}_α and M_α imply

$$\begin{aligned} & (C_\alpha^2)^{1/(1-\alpha)} (1 - \alpha) \left(\frac{\alpha C_R \beta_1}{2} \right)^{\alpha/(1-\alpha)} \\ &= (\beta_0^\alpha - 1) \tilde{M}_\alpha^{\alpha/(1-\alpha)} \leq (\beta_0^\alpha - 1) M_\alpha^{\alpha/(1-\alpha)} \frac{M_\alpha}{1 + M_\alpha} \\ &\leq (\beta_0^\alpha - 1) M_\alpha^{\alpha/(1-\alpha)} \frac{m(k)^\alpha M_\alpha}{D(k)}. \end{aligned}$$

This completes the proof of the theorem.

Remark 2.1 gives one example of a \mathcal{V} cycle multigrid algorithm with variable m satisfying the hypothesis of the theorem ($\beta_0 = \beta_1 = 2$). For a fixed $\beta_0 > 1$, one can define a sequence of $m(k)$'s as follows. Let $m(j)$ be an arbitrary positive integer and define, by induction, $m(k)$ to be the least integer greater than or equal to $\beta_0 m(k+1)$. For example, taking $\beta_0 = 3/2$, $\beta_1 = 2$ and $m(j) = 1$ gives rise to the sequence $m(j), m(j-1), \dots = 1, 2, 3, 5, 8, 12, 18, \dots$.

The following result gives the analogous theorem in the nonsymmetric case.

THEOREM 6. *Assume that (3.3) and (3.4) hold and define B_j^n by Algorithm N with $p = 1$ and $m(k)$ satisfying (3.40). Then (3.2) holds with δ_j given by (3.41)–(3.42).*

Proof. The theorem easily follows from (3.27) and the estimates derived in the proof of Theorem 5.

4. Using Multigrid to Solve (2.1) and Related Systems. In this section we shall consider a number of iterative techniques which use multigrid to solve (2.1) and related systems. We shall see that Theorems 1–6 give rise to estimates on the rate of convergence for the resulting iterative schemes. These observations are not new, however are included to indicate some of the ways in which multigrid can be used to solve problems.

We first consider the symmetric multigrid operators. Combining (2.9), (3.1) and an obvious change of variable gives that

$$(4.1) \quad (1 - \delta)(A_j^{-1}u, u)_j \leq (B_j^s u, u)_j \leq (A_j^{-1}u, u)_j \quad \text{for all } u \in \mathcal{M}_j$$

holds for δ given by (3.6), (3.29) or (3.41). Inequalities of the form of (4.1) imply that B_j^s is a good preconditioner for A_j , and hence many preconditioned iterative techniques can be applied to solve (2.1) or similar problems corresponding to other comparable operators. Specifically, let $\tilde{A}_j: \mathcal{M}_j \mapsto \mathcal{M}_j$ be another symmetric (with respect to the $(\cdot, \cdot)_j$ -inner product) operator which satisfies comparability estimates of the form

$$(4.2) \quad c_0(\tilde{A}_j u, u)_j \leq (A_j u, u)_j \leq c_1(\tilde{A}_j u, u)_j \quad \text{for all } u \in \mathcal{M}_j.$$

Then B_j^s can be used as a preconditioner for the solution of the problem: Given $f \in \mathcal{M}_j$, find $v \in \mathcal{M}_j$ satisfying

$$(4.3) \quad \tilde{A}_j v = f.$$

The most straightforward preconditioned iterative method is the linear method given by

$$(4.4) \quad v^{n+1} = v^n + \tau B_j^s(f - \tilde{A}_j v^n).$$

Let

$$|||\cdot||| \equiv (\tilde{A}_j \cdot, \cdot)^{1/2}$$

and set $e^n = v - v^n$. Then iteration (4.4) is convergent for an appropriate choice of τ . Furthermore, if $\tilde{A}_j = A_j$, then (4.4) is convergent for $\tau < 2$ and

$$(4.5) \quad |||e^{n+1}||| \leq \rho |||e^n|||,$$

where $\rho = \max(|1 - \tau|, |1 - \tau + \tau\delta|)$. Taking $\tau = 1$ gives a rate of δ per step, while the optimal convergence is achieved by setting $\tau = 2/(2 - \delta)$.

It is not difficult to see that $e^n = (I + P_\tau(B_j^s \tilde{A}_j))e^0$ for an appropriate polynomial P_τ . More generally, the iterative scheme defined by

$$v^n = v^0 - P(B_j^s A_j)(v - v^0)$$

gives rise to the error equation

$$e^n = (I + P(B_j^s \tilde{A}_j))e^0.$$

An accelerated linear scheme is defined by making an appropriate choice of P . An optimal choice of the P can be made by use of Chebyshev polynomials (see, for example, [7]). For such a choice, the error $e^n \equiv v - v^n$ can be bounded by

$$(4.6) \quad |||e^n||| \leq 2 \left(\frac{\sqrt{K} - 1}{\sqrt{K} + 1} \right)^n |||e^0|||,$$

where K (the condition number of $B_j^s \tilde{A}_j$) is bounded by $c_1/(c_0(1 - \delta))$. The coefficients of the polynomial P depend upon *a priori* estimates for the largest and smallest eigenvalues of $B_j^s \tilde{A}_j$.

It is possible to set up nonlinear iterations which use B_j^s as a preconditioner for the solution of (4.3). One well-known candidate results from the conjugate gradient method. The solution v of (4.3) satisfies

$$(4.7) \quad B_j^s \tilde{A}_j v = B_j^s f.$$

The operator $B_j^s \tilde{A}_j$ is symmetric in the $(\tilde{A}_j \cdot, \cdot)_j$ -inner product (see Remark 2.2) and is positive definite by (4.1) and (4.2). The conjugate gradient method can be directly applied to (4.7) in this inner product to produce a sequence $\{v_n\}$ which converges to v . In fact, the error for this iteration is also bounded by (4.6). The conjugate gradient scheme has the additional property that estimates for eigenvalues of $B_j^s \tilde{A}_j$ need not be known *a priori* for its application.

We next consider the use of nonsymmetric multigrid algorithms for the solution of (2.1). The linear iteration (4.4) with $\tau = 1$ converges and satisfies the inequality

$$(4.8) \quad |||e^{n+1}||| \leq \delta^{1/2} |||e^n|||,$$

where δ satisfies (3.2).

Remark 4.1. The estimates for δ_j satisfying (3.1) given by Theorem 1 are identical to those for δ_j satisfying (3.2) given by Theorem 2. Using the estimates of the theorems, we find that $\tau = 1$ in (4.5) should converge at a rate which is twice as fast as that of (4.8). Thus, the extra smoothings used in the symmetric algorithm lead to an iterative method which converges in half the number of iterations. A similar observation holds for the symmetric and nonsymmetric \mathcal{W} and generalized \mathcal{V} cycle algorithms.

Remark 4.2. We have considered applying the multigrid algorithms to the solution of the ‘algebraic’ problem (2.1) or the related problem (4.3). In the case of finite element or finite difference applications, the so-called full multigrid process can be used to get an ‘accurate’ approximation to the solution of (2.1) with computational work proportional to that required for a reduction on the finest grid level (cf. [3]).

5. Estimate (3.3) and Applications. We give a proof of the regularity and approximation estimate (3.3) in this section in the case of a typical finite element application of multigrid. Although the proof is given in [3], we include

it for completeness. We next discuss application to finite differences. We finally consider an example where symmetric Gauss-Seidel iteration is used to define R_k .

We shall consider the problem of approximating the solution U of

$$(5.1) \quad \begin{aligned} LU &= F && \text{in } \Omega, \\ U &= 0 && \text{on } \partial\Omega. \end{aligned}$$

Here Ω is a domain in n -dimensional Euclidean space and L is given by

$$Lv = - \sum_{i,j=1}^n \frac{\partial}{\partial x_i} \left(a_{ij} \frac{\partial v}{\partial x_j} \right),$$

with $\{a_{ij}\}$ uniformly positive definite and bounded on Ω . The form A used in the multigrid algorithm is the bilinear form corresponding to the operator L and is defined by

$$(5.2) \quad A(v, w) = \sum_{i,j=1}^n \int_{\Omega} a_{ij} \frac{\partial v}{\partial x_i} \frac{\partial w}{\partial x_j} dx.$$

This form is defined for all v and w in the Sobolev space $H^1(\Omega)$ (the space of distributions in $L^2(\Omega)$ with square integrable first derivatives). Clearly, U is the solution of

$$A(U, \theta) = (F, \theta) \quad \text{for all } \theta \in H_0^1(\Omega),$$

where $H_0^1(\Omega)$ is the subspace of $H^1(\Omega)$ of functions which vanish in the appropriate sense on $\partial\Omega$ and (\cdot, \cdot) denotes the L^2 -inner product on Ω .

We assume that Ω has been triangulated with a sequence of quasi-uniform triangulations $\Omega = \bigcup_i \tau_k^i$ of size h_k for $k = 0, \dots, j$, where the quasi-uniformity constants are independent of k (cf. [3]). We further assume that there is a constant c , independent of k , such that $h_{k-1} \leq ch_k$. These triangulations should be nested in the sense that any triangle τ_{k-1}^l can be written as a union of triangles of $\{\tau_k^i\}$. We define \mathcal{M}_k to be the set of piecewise linear functions (with respect to the triangulation $\bigcup_i \tau_k^i$) which vanish on $\partial\Omega$.

To avoid the inversion of L^2 Gram matrices, we define $(\cdot, \cdot)_k$ as a discrete L^2 -inner product. Let $\{y_k^i\}$ be the collection of nodes corresponding to the triangulation for \mathcal{M}_k . We set

$$(5.3) \quad (u, v)_k = h_k^n \sum_i u(y_k^i) v(y_k^i).$$

Note that the quasi-uniformity of the triangulations implies that the discrete form $(\cdot, \cdot)_k$ is equivalent to the form (\cdot, \cdot) on the subspace \mathcal{M}_k . We seek the Galerkin approximation $v \in \mathcal{M}_j$ (to the solution U of (5.1)) defined by

$$(5.4) \quad A(v, \phi) = (F, \phi) \quad \text{for all } \phi \in \mathcal{M}_j.$$

Equation (5.4) can be rewritten as

$$(5.5) \quad A(v, \phi) = (\tilde{F}, \phi)_j \quad \text{for all } \phi \in \mathcal{M}_j,$$

with an obvious choice of $\tilde{F} \in \mathcal{M}_j$. We derive (3.3) under the following elliptic regularity assumption. There exists a constant C such that

$$(5.6) \quad \|U\|_{H^{1+\alpha}(\Omega)} \leq C \|F\|_{H^{\alpha-1}(\Omega)}$$

holds for solutions U of (5.1). The norms $H^s(\Omega)$ are Sobolev norms of order s and are defined in, for example, [10], [15].

The following lemma is given in [3] and depends on the quasi-uniformity assumptions on the mesh defining \mathcal{M}_k .

LEMMA 5.1. *Let $0 \leq s \leq 1$. There are constants c_0 and c_1 which are independent of $\mathcal{M}_0 \subset \cdots \subset \mathcal{M}_j$ and satisfy*

$$c_0 \|A_k^{s/2} u\|_k \leq \|u\|_{H^s(\Omega)} \leq c_1 \|A_k^{s/2} u\|_k \quad \text{for all } u \in \mathcal{M}_k.$$

Using the above lemma, we shall prove the following proposition.

PROPOSITION 5.1. *Assume that (5.6) holds for some α in $(0, 1]$. Then (3.3) also holds for that α .*

Remark 5.1. It is well known that estimates of the form of (5.6) do not in general hold for all α , and the range of α 's for which they hold depends upon the regularity of the coefficients defining L and the smoothness of $\partial\Omega$. We have computational evidence (see Example 6.1) which suggests that in such instances (3.3) will not hold for all α with constants C_α^2 independent of h . Thus, the discrete result (3.3) is tied strongly to the elliptic regularity result (5.6).

Proof of Proposition 5.1. Let $u \in \mathcal{M}_k$. Applying Schwarz's inequality and Lemma 5.1 gives

$$\begin{aligned} A((I - P_{k-1})u, u) &\leq |||A_k^{\alpha/2} u||| |||A_k^{-\alpha/2}(I - P_{k-1})u||| \\ (5.7) \quad &\leq |||A_k^{\alpha/2} u||| \left\| A_k^{(1-\alpha)/2}(I - P_{k-1})u \right\|_k \\ &\leq |||A_k^{\alpha/2} u||| \|(I - P_{k-1})u\|_{H^{1-\alpha}(\Omega)}. \end{aligned}$$

By Hölder's inequality,

$$(5.8) \quad |||A_k^{\alpha/2} u||| \leq \left(A(u, u)^{1-\alpha} \|A_k u\|_k^{2\alpha} \right)^{1/2}.$$

By standard error analysis techniques for finite element methods, employing duality and (5.6) (cf. [1], [6]), it follows that

$$(5.9) \quad \|(I - P_{k-1})u\|_{H^{1-\alpha}(\Omega)} \leq ch_{k-1}^\alpha A((I - P_{k-1})u, u)^{1/2}.$$

By the quasi-uniformity assumption of the mesh, $\lambda_k \leq ch_k^{-2}$. Hence, the proposition results from combining (5.7)–(5.9).

We next consider applying the theorems in the finite difference case. Here we consider a uniform rectangular grid with nodes (i, l) . We are to invert the five-point operator which is defined for grid points $(i, l) \in \Omega$ by

$$(5.10) \quad (\tilde{A}_j u)_{i,l} = h_j^{-2} (4u_{i,l} - u_{i-1,l} - u_{i+1,l} - u_{i,l-1} - u_{i,l+1}),$$

where we set $u_{i,l} = (\tilde{A}_j u)_{i,l} = 0$ for $(i, l) \notin \Omega$. We define a triangulation by breaking each rectangle into two triangles and set \mathcal{M}_j to be the space of piecewise linear functions with respect to this triangulation which vanish on the nodes not in Ω . It is well known that

$$(\tilde{A}_j u, u)_j = (A_j u, u)_j,$$

where A_j is the operator defined by (5.2) and (5.3) with L equal to minus the Laplacian $(-\Delta)$. Consequently, if we define a nested sequence of subspaces \mathcal{M}_k as described above, then Proposition 5.1 applies and shows that the multigrid algorithms (defined using these spaces, (5.2), (5.3) and A_j) satisfy the convergence estimates of the theorems. These are clearly multigrid algorithms for \tilde{A}_j . Note that for these algorithms, both the ‘interpolation operators’ and ‘residual computation operators’ which appear in the so-called ‘finite difference’ multigrid are defined by

the way nodal basis elements in \mathcal{M}_{k+1} are combined to get nodal basis elements in \mathcal{M}_k .

We finally consider an example where symmetric Gauss-Seidel iteration is used to define R_k . We consider the case where $\tilde{A}_j = A_j$ is given by (5.10). Note that we do not require that the domain be a square. Then the matrix M_k corresponding to A_j in the usual nodal basis with any ordering can be written

$$M_k = \frac{4}{h_k^2} (I - L_k - U_k),$$

where U_k and L_k are strictly upper and lower triangular matrices and I is the identity matrix. We define R_k by

$$R_k = \frac{h_k^2}{4} (I - L_k)^{-1} (I - U_k)^{-1}.$$

The matrix corresponding to K_k is

$$I - (I - L_k)^{-1} (I - U_k)^{-1} (I - L_k - U_k)$$

which is similar to

$$(5.11) \quad (I - U_k)^{-1} U_k L_k (I - L_k)^{-1}.$$

The matrix (5.11) is clearly nonnegative and hence K_k is also nonnegative. We need only verify (3.4). The largest eigenvalue of A_j is greater than or equal to $4/h^2$, and hence (3.4) is equivalent to

$$(5.12) \quad \|(I - U_k)u\|_k^2 \leq C_R \|u\|_k^2.$$

But the sum of the absolute value of entries in any column of $L_k + U_k$ is less than or equal to one, from which it follows that

$$(5.13) \quad \|U_k u\|_k \leq \|u\|_k.$$

Hence, (3.4) holds with $C_R = 4$.

6. Numerical Examples. Numerical experiments are presented in this section which illustrate some of the theoretical properties derived earlier in this paper. In some cases, actual values for C_α are computed and the results of numerical experiments are compared with the theoretical bounds given by the theorems of Section 3 with these values of C_α . These results suggest that, for finite element applications, the theorems provide reasonable convergence bounds. We also give an example which suggests that (3.3) does not in general hold for $\alpha = 1$ with constant C_α independent of h .

All of the examples presented will be for the problem

$$\begin{aligned} -\Delta U &= F && \text{in } \Omega, \\ U &= 0 && \text{on } \partial\Omega, \end{aligned}$$

where Ω is either the unit square, the unit square with the upper right-hand sub-square removed (i.e., the ‘L-shaped’ domain), or the slit domain whose boundary consists of the boundary of the unit square together with the set $\{(1/2, y) \mid y \in [1/2, 1]\}$. We will consider a regular rectangular mesh where each rectangle is split into two triangles with piecewise linear elements and use the finite element multigrid framework discussed in Section 5. As discussed in Section 5, the resulting

multigrid algorithm can be used to solve either the corresponding finite element or finite difference equations.

Example 1. In this example, we compute the best constant $C_1(h_k)$ satisfying the inequality

$$A((I - P_{k-1})u, u) \leq C_1^2(h_k) \left(\frac{\|A_k u\|_k^2}{\lambda_k} \right) \quad \text{for all } u \in M_k,$$

for both the square and slit domains. For the square, Proposition 5.1 gives that $C_1(h_k)$ can be bounded independently of h_k . In fact, by using an odd extension and an analysis (cf. [2]) on the periodic domain, it can be shown that $C_1^2(h_k) \leq 4 + 2\sqrt{2} = 6.828\dots$. The computed values for $C_1^2(h_k)$ in Table 6.1 satisfy the above estimate. In contrast, for the slit domain, Proposition 5.1 can only be applied for $0 < \alpha < 1/2$. Note that the results in Table 6.1 clearly suggest that it is not possible to satisfy (3.3) with $\alpha = 1$ for the slit domain. In fact, $C_1^2(h_k)$ seems to be growing like ch_k^{-1} . This example clearly indicates that the multigrid analysis for $\alpha < 1$ is important since the assumption (3.3) cannot be expected to hold in general for $\alpha = 1$.

TABLE 6.1
Values for $C_1^2(h_k)$ for the Square and Slit Domains.

h_k	Square	Slit
1/16	5.8	7.8
1/32	6.3	12
1/64	6.6	20
1/128	6.6	36

Example 2. In this example, we consider the symmetric \mathcal{V} cycle of multigrid applied to the slit domain. We shall consider two algorithms. The first algorithm uses $m(k) = 1$, while the second uses variable smoothing defined by $m(k) = 2^{j-k}$. Both algorithms use $R_k = \tilde{\lambda}_k^{-1} I_k$ where $\tilde{\lambda}_k$ is the largest eigenvalue for the A_k with the square domain. We compute the best values of δ_j satisfying (3.1) for the corresponding algorithms and set $\varepsilon_j = 1 - \delta_j$. According to Proposition 5.1, (3.3) holds for α in the interval $(0, 1/2)$ and from Theorem 1 we expect that ε_j ($m = 1$) should go to zero nearly like

$$(6.1) \quad \tilde{\varepsilon}_j \ (m = 1) = (1 + M_\alpha(j_0 + j))^{-1}.$$

Unfortunately, we do not know the corresponding constant $C_{1/2}$ defining $M_{1/2}$ in Theorem 1. Instead, we have chosen $M_{1/2} = .268$ and $j_0 = 1.93$ in (6.1) to fit the computed results. The fact that this function fits the computed values of ε_j ($m = 1$) so closely indicates that the $\log(h)$ is really reflected in the computational behavior. However, the logarithmic growth is quite slow and the corresponding reductions (i.e., $\delta_j = .722$ at $h_j = 1/256$) are rather remarkable. We also include the values of ε_j (variable m) corresponding to the variable smoothing algorithm. Note that, as predicted by Theorem 5, these values for ε_j remain bounded away from zero independently of h .

TABLE 6.2
Values of ε_j for \mathcal{V} Cycle Algorithms on the Slit Domain.

h_j	ε_j ($m = 1$)	$\tilde{\varepsilon}_j$ ($m = 1$)	ε_j (variable m)
1/8	.45	.43	.414
1/16	.386	.390	.428
1/32	.347	.350	.424
1/64	.318	.320	.422
1/128	.296	.295	.420
1/256	.278	.273	.420

Example 3. In this example, we compare the values of ε_j for the L-shaped and square domains. We use the \mathcal{V} cycle with $m = 1$ and $R_k = \tilde{\lambda}_k^{-1}I_k$. In the case of the L-shaped domain, Proposition 5.1 implies that (3.3) holds for $0 < \alpha < 2/3$. As can be seen from Table 6.3, the computed values of ε_j for the L-shaped domain are somewhat smaller than those for the square domain and somewhat larger than those for the slit domain. The ε_j 's corresponding to the L-shaped domain also decrease faster than those for the square while not as fast as those for the slit. This is in qualitative agreement with the theorems.

TABLE 6.3
Computed Values of ε_j for the Square and L-shaped Domains.

h_j	ε_j (square)	ε_j (L-shaped)
1/8	.48	.46
1/16	.43	.42
1/32	.42	.40
1/64	.41	.38
1/128	.41	.37
1/256	.41	.36

Example 4. We include this example to indicate that the behavior for large m suggested by the theorems is consistent with the results of actual computations. We shall use the symmetric \mathcal{V} cycle on a grid of size $h_j = 1/64$ with $R_k = \tilde{\lambda}_k^{-1}I_k$ and $m(k) = m$ for all k .

We first consider the case of $\alpha = 1$, i.e., the square domain. We shall compare the computed best constant δ_j (computed) satisfying (3.1) with the theoretical estimate given by Theorem 1 as a function of m . In this case,

$$C_1^2 \leq 6.828\dots$$

as noted earlier. In this application, $R_k = \tilde{\lambda}_k^{-1}I_k = \lambda_k^{-1}I_k$ and hence $C_R = 1$. Theorem 1 guarantees that

$$\delta_j \text{ (computed)} \leq \delta_j \text{ (theoretical)} = \frac{C_1^2}{2m + C_1^2}.$$

Table 6.4 gives the values of δ_j (computed) and δ_j (theoretical) as a function of m . We also fit the computed results (at $m = 13$) with a function of the form

$$(6.2) \quad \delta_j \text{ (fit-1)} = \frac{C}{2m + C}.$$

TABLE 6.4
Computed, Theoretical, and Fit Values of δ_j on the Square.

m	δ_j (computed)	δ_j (theoretical)	δ_j (fit-1)
1	.59	.77	.55
5	.20	.40	.19
13	.085	.20	.085
25	.045	.12	.046
41	.027	.08	.029

Note that the fitted function provides a very good approximation to the computed values of δ_j for large m .

We next consider the slit domain. We again report the computed values of δ_j and compare with fitted functions. Since, for this example, (3.3) holds for $0 < \alpha < 1/2$, we should expect to be able to fit δ_j with

$$(6.3) \quad \delta_j \text{ (fit-1/2)} = \frac{C}{\sqrt{2m} + C}.$$

Table 6.5 compares this fit (at $m = 13$) with the computed results for δ_j . Note that the computed results are going to zero slightly faster than the fit. For comparison, we also have fit the computed results (δ_j (fit-1)) to a function of the form (6.2). The computed results go to zero more slowly than the fit to (6.2). Thus the actual computed results seem to show an asymptotic behavior which is somewhere between m^{-1} and $m^{-1/2}$. This is consistent with the results of the theorems since the theorems provide only pessimistic convergence bounds.

TABLE 6.5
Computed and Fit Values of δ_j on the Slit.

m	δ_j (computed)	δ_j (fit-1/2)	δ_j (fit-1)
1	.682	.38	.69
5	.287	.22	.30
13	.146	.146	.146
25	.089	.11	.08
41	.060	.09	.05

Department of Mathematics
Cornell University
Ithaca, New York 14853

Brookhaven National Laboratory
Upton, New York 11973

1. A. K. AZIZ & I. BABUŠKA, "Survey lectures on the mathematical foundations of the finite element method," in *The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations. Part I* (A. K. Aziz, ed.), Academic Press, New York, 1972, pp. 1-362.

2. R. E. BANK & C. C. DOUGLAS, "Sharp estimates for multigrid rates of convergence with general smoothing and acceleration," *SIAM J. Numer. Anal.*, v. 22, 1985, pp. 617-633.

3. R. E. BANK & T. F. DUPONT, "An optimal order process for solving elliptic finite element equations," *Math. Comp.*, v. 36, 1981, pp. 35–51.
4. D. BRAESS & W. HACKBUSCH, "A new convergence proof for the multigrid method including the V -cycle," *SIAM J. Numer. Anal.*, v. 20, 1983, pp. 967–975.
5. A. BRANDT, "Multi-level adaptive solutions to boundary-value problems," *Math. Comp.*, v. 31, 1977, pp. 333–390.
6. P. G. CIARLET, *The Finite Element Method for Elliptic Problems*, North-Holland, New York, 1978.
7. T. DUPONT, R. P. KENDALL & H. H. RACHFORD, "An approximate factorization procedure for solving self-adjoint elliptic difference equations," *SIAM J. Numer. Anal.*, v. 5, 1968, pp. 559–573.
8. R. P. FEDORENKO, "The speed of convergence of one iterative process," *USSR Comput. Math. and Math. Phys.*, v. 1, 1961, pp. 1092–1096.
9. W. HACKBUSCH, *Multi-Grid Methods and Applications*, Springer-Verlag, New York, 1985.
10. J. L. LIONS & E. MAGENES, *Problèmes aux Limites non Homogènes et Applications*, Dunod, Paris, 1968.
11. J. F. MAITRE & F. MUSY, "Algebraic formalization of the multigrid method in the symmetric and positive definite case—a convergence estimation for the V -cycle," *Multigrid Methods for Integral and Differential Equations* (D. J. Paddon and H. Holstein, eds.), Clarendon Press, Oxford, 1985.
12. J. MANDEL, "Multigrid convergence for nonsymmetric, indefinite variational problems and one smoothing step," in *Proc. Copper Mtn. Conf. Multigrid Methods*, Applied Math. Comput., 1986, pp. 201–216.
13. J. MANDEL, "Algebraic study of multigrid methods for symmetric, definite problems." (Preprint)
14. J. MANDEL, S. F. MCCORMICK & J. RUGE, "An algebraic theory for multigrid methods for variational problems." (Preprint)
15. S. F. MCCORMICK, "Multigrid methods for variational problems: Further results," *SIAM J. Numer. Anal.*, v. 21, 1984, pp. 255–263.
16. S. F. MCCORMICK, "Multigrid methods for variational problems: General theory for the V -cycle," *SIAM J. Numer. Anal.*, v. 22, 1985, pp. 634–643.
17. J. NEČAS, *Les Méthodes Directes en Théorie des Équations Elliptiques*, Academia, Prague, 1967.
18. R. A. NICOLAIDES, "On the l^2 convergence of an algorithm for solving finite element equations," *Math. Comp.*, v. 31, 1977, pp. 892–906.

3.2 The analysis of multigrid algorithms for nonsymmetric and indefinite elliptic problems

The analysis of multigrid algorithms for nonsymmetric and indefinite elliptic problems[25]

The Analysis of Multigrid Algorithms for Nonsymmetric and Indefinite Elliptic Problems*

By James H. Bramble, Joseph E. Pasciak, and Jinchao Xu

Abstract. We prove some new estimates for the convergence of multigrid algorithms applied to nonsymmetric and indefinite elliptic boundary value problems. We provide results for the so-called ‘symmetric’ multigrid schemes. We show that for the variable \mathcal{V} -cycle and the \mathcal{W} -cycle schemes, multigrid algorithms with any amount of smoothing on the finest grid converge at a rate that is independent of the number of levels or unknowns, provided that the initial grid is sufficiently fine. We show that the \mathcal{V} -cycle algorithm also converges (under appropriate assumptions on the coarsest grid) but at a rate which may deteriorate as the number of levels increases. This deterioration for the \mathcal{V} -cycle may occur even in the case of full elliptic regularity. Finally, the results of numerical experiments are given which illustrate the convergence behavior suggested by the theory.

1. Introduction. In recent years, multigrid methods have been used extensively as tools for obtaining the solution of the discrete systems which arise in the numerical approximation of partial differential equations (cf. [6], [8]). In conjunction, there has been intensive research aimed at attaining a more thorough theoretical understanding of the multigrid technique [1]–[5], [8], [13]–[18], [21]. In this paper, we shall provide some new iterative convergence estimates for multigrid algorithms applied to nonsymmetric and indefinite problems.

The theory for the analysis of multigrid methods applied to symmetric positive definite problems is most completely developed [2], [4], [5], [13], [15], [21]. Generally, these results assume a ‘regularity and approximation’ hypothesis which involves a parameter $0 < \alpha \leq 1$. The results in these papers guarantee convergence rates for multigrid \mathcal{V} -cycle, the variable \mathcal{V} -cycle (cf. [5]) and the \mathcal{W} -cycle algorithms for various α . In particular, [5], [15] give iterative convergence results for the symmetric problem which are valid for any amount of smoothing and any α .

The theory for multigrid methods applied to nonsymmetric and indefinite problems is not so completely developed. Two types of algorithms are the so-called ‘symmetric’ and ‘nonsymmetric’ multigrid schemes. The nonsymmetric scheme uses a relaxation procedure based on the original equations whereas the symmetric

Received November 9, 1987.

1980 *Mathematics Subject Classification* (1985 Revision). Primary 65N30; Secondary 65F10.

*This manuscript has been authored under contract number DE-AC02-76CH00016 with the U.S. Department of Energy. Accordingly, the U.S. Government retains a non-exclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. This work was also supported in part under the National Science Foundation Grant No. DMS84-05352 and under the Air Force Office of Scientific Research, Contract No. ISSA86-0026 and by the U.S. Army Research Office through the Mathematical Science Institute, Cornell University.

©1988 American Mathematical Society
0025-5718/88 \$1.00 + \$.25 per page

scheme uses a relaxation based on the symmetric positive definite system associated with the normal equations. Some results only hold under rather restrictive assumptions involving the relation between the number of smoothings m and the size of the coarsest grid h_j . For example, Bank [1] gives \mathcal{W} -cycle results for both schemes and for arbitrary α which, however, require first that m be sufficiently large, and secondly that h_j be sufficiently small (depending on m). Mandel [14] gives results for the nonsymmetric \mathcal{W} -cycle scheme and the \mathcal{V} -cycle scheme (assuming full regularity $\alpha = 1$) which are valid for any m if h_j is chosen sufficiently small (depending on m).

In this paper, we shall prove some new iterative convergence estimates for the symmetric multigrid scheme applied to nonsymmetric and indefinite problems. We give results for the \mathcal{V} -cycle, variable \mathcal{V} -cycle and \mathcal{W} -cycle algorithms for any amount of smoothing under the assumption of $\alpha > 3/4$. Our theorems for the variable \mathcal{V} -cycle and \mathcal{W} -cycle algorithms require that h_j be sufficiently small (independent of the amount of smoothing) and guarantee an iterative convergence rate which is uniformly independent of the number of levels and the mesh size on the finest grid. The assumption that h_j is sufficiently small is not very restrictive since such an assumption must be made for solvability on the coarsest grid. The results for the \mathcal{V} -cycle algorithm are somewhat weaker. We show that the \mathcal{V} -cycle converges if h_j is small enough (depending on the number of levels and α), at a rate which deteriorates as more and more levels are used. Even in the case $\alpha = 1$, the \mathcal{V} -cycle convergence estimates deteriorate like $1 - c/\ln(h^{-1})$.

We derive our iterative convergence estimates for multigrid algorithms in an abstract setting. The use of this abstract approach more clearly identifies the relevant hypotheses.

The outline of the remainder of the paper is as follows. In Section 2 we describe the abstract framework to be used in the paper. The assumptions used in our analysis and some preliminary definitions are also given there. Section 3 shows how this framework can be applied in the case of nonsymmetric and indefinite uniformly elliptic second-order boundary value problems. Section 4 defines the multigrid operator and provides a basic recurrence relation used in our subsequent analysis. The convergence estimates given in this paper are based on three technical lemmas. In Section 5 we prove our multigrid theorems, assuming the technical lemmas. Section 6 provides the proof of the lemmas and represents the core of our analysis. Finally, the results of numerical experiments illustrating the earlier derived theory are given in Section 7.

Throughout this paper, c and C , with or without subscript will denote a generic positive constant which may take on different values in different places. These constants will always be independent of the mesh parameters.

2. Abstract Framework and Assumptions. In this section, we first give an abstract framework for our nonsymmetric multigrid application. This abstract presentation more clearly identifies the relevant hypotheses used in the iterative convergence analysis to be developed. We then list the assumptions required for the multigrid analysis presented in later sections. To keep the paper from becoming too abstract, we show how a model application to a second-order problem fits into this framework in the next section.

We start with a Hilbert scale (cf. [11]) of spaces $\{H^\gamma\}$ for $\gamma \in [0, 2]$. The norm on H^γ will be denoted by $\|\cdot\|_{H^\gamma}$. We assume that $H^s \subset H^t$ whenever $t < s$. The largest space (i.e., $\gamma = 0$) will be denoted H with norm $\|\cdot\|_H$ and inner product (\cdot, \cdot) . The space H^γ is assumed to be compactly contained in H^δ whenever $\gamma > \delta$. Let \mathcal{M} be a closed subspace of H^1 . The spaces H^s for $-1 \leq s < 0$ are defined by duality and with norm given by

$$\|v\|_{H^s} \equiv \sup_{\phi \in \mathcal{M}} \frac{(v, \phi)}{\|\phi\|_{H^{-s}}}.$$

Assume that we are given a nested sequence of ‘approximation’ subspaces

$$\mathcal{M}_1 \subset \mathcal{M}_2 \subset \cdots \subset \mathcal{M}_J \subset \mathcal{M}.$$

In addition, let $\hat{A}(\cdot, \cdot)$ be a positive definite symmetric quadratic form on $\mathcal{M} \times \mathcal{M}$ satisfying

$$(2.1) \quad c\|v\|_{H^1}^2 \leq \hat{A}(v, v) \leq C\|v\|_{H^1}^2 \quad \text{for all } v \in \mathcal{M}$$

and $D(\cdot, \cdot)$ be a quadratic form on $\mathcal{M} \times \mathcal{M}$. We shall be interested in approximating the solution of

$$(2.2) \quad A(u, \phi) \equiv \hat{A}(u, \phi) + D(u, \phi) = (f, \phi) \quad \text{for all } \phi \in \mathcal{M},$$

for a given function $f \in H$. We shall assume that (2.2) is uniquely solvable for any $f \in H$.

We will be interested in applying multigrid procedures to develop a rapidly converging iterative algorithm for the solution of the Galerkin approximation of (2.2) in the subspace \mathcal{M}_J . Specifically, we seek the function $U \in \mathcal{M}_J$ which satisfies

$$(2.3) \quad A(U, \chi) = (f, \chi) \quad \text{for all } \chi \in \mathcal{M}_J.$$

Our multigrid algorithms will require the use of discrete inner products $(\cdot, \cdot)_k$ on $\mathcal{M}_k \times \mathcal{M}_k$ for $k = 1, \dots, J$. The corresponding norm will be denoted $\|\cdot\|_k$. In the algorithms, these inner products are used instead of (\cdot, \cdot) to avoid the inversion of Gram matrices. This means that the problem of computing $W \in \mathcal{M}_k$ satisfying

$$(2.4) \quad (W, \theta)_k = F(\theta) \quad \text{for all } \theta \in \mathcal{M}_k$$

for a given linear functional F should be simple.

We next list the assumptions required for our multigrid analysis.

(A.1): The first assumption involves elliptic regularity for the forms $A(\cdot, \cdot)$ and $\hat{A}(\cdot, \cdot)$. We assume that solutions u of (2.2) and the corresponding equation

$$\hat{A}(u, \theta) = (f, \theta) \quad \text{for all } \theta \in \mathcal{M}$$

satisfy

$$(2.5) \quad \|u\|_{H^{1+\alpha}} \leq c\|f\|_{H^{\alpha-1}}$$

for some $\alpha \in (3/4, 1]$ independent of f .

(A.2): We assume first that D satisfies

$$(2.6) \quad |D(v, w)| \leq C\|v\|_{H^1}\|w\|_H \quad \text{for all } v, w \in \mathcal{M}.$$

It is an immediate consequence of (2.6) that the operator $D : \mathcal{M} \mapsto H$ defined by

$$(Dv, \theta) = D(v, \theta) \quad \text{for all } \theta \in H$$

is well defined and satisfies

$$(2.7) \quad \|Dv\|_H \leq C\|v\|_{H^1}.$$

We further assume that D maps $H^{1+\alpha}$ into H^α , i.e.,

$$(2.8) \quad \|Dv\|_{H^\alpha} \leq C\|v\|_{H^{1+\alpha}}.$$

Let $D^* : H \mapsto H^{-1}$ be defined by

$$(D^*w, \phi) = (w, D\phi).$$

We assume that D^* is a bounded operator from H^1 into $H^{-1/2-\varepsilon}$ for any positive ε .

(A.3): We require approximation properties for the subspaces $\{\mathcal{M}_k\}$. These are given in terms of a parameter h_k which satisfies

$$c\kappa^k \leq h_k \leq C\kappa^k$$

for constants c, C and $\kappa < 1$ independent of k . We assume that for v in H^s and $s \in [1, 1 + \alpha]$, there exists $\chi \in \mathcal{M}_k$ such that

$$\|v - \chi\|_H + h_k\|v - \chi\|_{H^1} \leq Ch_k^s\|v\|_{H^s}.$$

(A.4): We require that the inverse inequality,

$$\|W\|_{H^\beta} \leq Ch_k^{\gamma-\beta}\|W\|_{H^\gamma} \quad \text{for all } W \in \mathcal{M}_k$$

holds for all $\beta > \gamma$ with $\beta, \gamma \in [0, 1 + \alpha]$.

(A.5): We require first that the discrete inner product $(\cdot, \cdot)_k$ be equivalent to (\cdot, \cdot) on \mathcal{M}_k , i.e.,

$$(2.9) \quad c\|\chi\|_H \leq \|\chi\|_k \leq C\|\chi\|_H.$$

In addition, we assume that the discrete inner products accurately approximate the inner product on H in the sense that

$$(2.10) \quad |(\psi, \chi) - (\psi, \chi)_k| \leq Ch_k\|\psi\|_{H^1}\|\chi\|_k \quad \text{for all } \psi, \chi \in \mathcal{M}_k.$$

We next introduce some discrete operators which play a fundamental role both in the analysis and the algorithms to be considered in this paper:

(O.1): The operator $A_k : \mathcal{M}_k \mapsto \mathcal{M}_k$ is defined by

$$(A_k W, \theta)_k = A(W, \theta) \quad \text{for all } \theta \in \mathcal{M}_k.$$

(O.2): The operator $P_k : \mathcal{M} \mapsto \mathcal{M}_k$ is defined by

$$(2.11) \quad A(P_k w, \theta) = A(w, \theta) \quad \text{for all } \theta \in \mathcal{M}_k.$$

(O.3): The operator $\hat{A}_k : \mathcal{M}_k \mapsto \mathcal{M}_k$ is defined by

$$(\hat{A}_k W, \theta)_k = \hat{A}(W, \theta) \quad \text{for all } \theta \in \mathcal{M}_k.$$

(O.4): The operator $\hat{P}_k : \mathcal{M} \mapsto \mathcal{M}_k$ is defined by

$$\hat{A}(\hat{P}_k w, \theta) = \hat{A}(w, \theta) \quad \text{for all } \theta \in \mathcal{M}_k.$$

(O.5): The operator $D_k : \mathcal{M}_k \mapsto \mathcal{M}_k$ is defined by

$$(D_k W, \theta)_k = D(W, \theta) \quad \text{for all } \theta \in \mathcal{M}_k.$$

(O.6): The operator $I_k : \mathcal{M}_{k+1} \mapsto \mathcal{M}_k$ is defined by

$$(I_k W, \theta)_k = (W, \theta)_{k+1} \quad \text{for all } \theta \in \mathcal{M}_k.$$

(O.7): The operator $P_k^0 : H \mapsto \mathcal{M}_k$ is defined by

$$(P_k^0 v, \theta)_k = (v, \theta) \quad \text{for all } \theta \in \mathcal{M}_k.$$

All of the above operators except possibly P_k are clearly well defined. We shall assume, however, that h_k is less than some positive constant ν with ν chosen small enough so that the above assumptions imply a unique solution to (2.11) (cf. [20]). This also implies that A_k is invertible.

We note that (2.3) is equivalent to

$$A_k U = P_k^0 f.$$

We define two scales of norms on \mathcal{M}_k which we shall use in our analysis. The operator \hat{A}_k is symmetric and positive definite on \mathcal{M}_k in the $(\cdot, \cdot)_k$ inner product. We define the scale of norms $\{\|\cdot\|_{k,s}\}$ for any real s by

$$\|W\|_{k,s} = \|\hat{A}_k^{s/2} W\|_k \quad \text{for all } W \in \mathcal{M}_k.$$

Similarly, the operator $A_k^* A_k$ is also symmetric and positive definite on \mathcal{M}_k (here, $*$ denotes the adjoint with respect to $(\cdot, \cdot)_k$). We define the scale of norms $\{\|\cdot\|_{k,s}\}$ for any real s by

$$\|W\|_{k,s} = ((A_k^* A_k)^{s/2} W, W)_k^{1/2} \quad \text{for all } W \in \mathcal{M}_k.$$

Let $L_k = (A_k^* A_k)^{1/2}$; then clearly

$$\|W\|_{k,s} = \|L_k^{s/2} W\|_k \quad \text{for all } W \in \mathcal{M}_k.$$

We will often consider the norms of operators from a space into itself. If $T : S \mapsto S$ is an operator on a generic space S with norm $\|\cdot\|$, then the norm of T will be denoted by $\|T\|$ and is given by

$$\|T\| = \sup_{\phi \in S} \frac{\|T\phi\|}{\|\phi\|}.$$

3. An Application to the Second-Order Problem. We consider a model second-order problem in this section and show that the hypotheses of Section 2 are satisfied. This application involves a finite element approximation of a nonsymmetric and indefinite elliptic problem in N -dimensional Euclidean space.

Let Ω be a domain in R^N . The spaces $H^s = H^s(\Omega)$ will be the Sobolev spaces of order s on Ω [12], [19]. We shall be interested in approximating the solution of the problem

$$(3.1) \quad \mathcal{L}u = f \quad \text{in } \Omega,$$

$$(3.2) \quad \frac{\partial u}{\partial v} = 0 \quad \text{on } \partial\Omega,$$

or

$$(3.3) \quad u = 0 \quad \text{on } \partial\Omega,$$

where

$$\mathcal{L}u = - \sum_{i,j=1}^N \frac{\partial}{\partial x_i} a_{ij}(x) \frac{\partial u}{\partial x_j} + \sum_{i=1}^N b_i(x) \frac{\partial u}{\partial x_i} + c(x)u$$

and $\frac{\partial}{\partial v}$ denotes the outward co-normal derivative on $\partial\Omega$.

We assume that the matrix $\{a_{ij}(x)\}$ is symmetric and uniformly positive definite.

Under appropriate smoothness assumptions for the domain Ω and coefficients defining \mathcal{L} , it is possible to prove that the solutions of (3.1)–(3.3) satisfy estimates of the form (2.5) [7], [10]. For two-dimensional polygonal domains, with coefficients in $C^1(\Omega)$, (2.5) holds for $\alpha > 3/4$, if all interior angles of the polygon are bounded by $4\pi/3$. For more general applications, we implicitly assume the appropriate hypotheses so that (2.5) holds for $\alpha > 3/4$.

The space \mathcal{M} is a subset of $H^1(\Omega)$ satisfying appropriate boundary conditions. In the case of boundary condition (3.3), \mathcal{M} is the completion of $C_0^\infty(\Omega)$ in the $H^1(\Omega)$ -norm. For boundary condition (3.2), $\mathcal{M} = H^1(\Omega)$ unless $c(x) = 0$ for all x , in which case \mathcal{M} consists of those functions in $H^1(\Omega)$ which are orthogonal to constants.

A weak formulation of (3.1)–(3.3) is: Find $u \in \mathcal{M}$ such that

$$(3.4) \quad A(u, v) = (f, v) \quad \text{for all } v \in \mathcal{M},$$

where (\cdot, \cdot) is the usual $L^2(\Omega)$ inner product and

$$A(u, v) = \sum_{i,j=1}^N \int_{\Omega} a_{ij} \frac{\partial u}{\partial x_j} \frac{\partial v}{\partial x_i} dx + \sum_{i=1}^N \int_{\Omega} b_i \frac{\partial u}{\partial x_i} v dx + \int_{\Omega} cuv dx.$$

Note that, in general, $A(\cdot, \cdot)$ is nonsymmetric and indefinite. We assume that (3.4) has a unique solution.

We define $\hat{A}(\cdot, \cdot)$ by

$$\hat{A}(u, v) = \sum_{i,j=1}^N \int_{\Omega} a_{ij} \frac{\partial u}{\partial x_j} \frac{\partial v}{\partial x_i} dx + \int_{\Omega} uv dx.$$

Then, obviously

$$D(u, v) = \sum_{i=1}^N \int_{\Omega} b_i \frac{\partial u}{\partial x_i} v dx + \int_{\Omega} (c-1)uv dx.$$

We next check assumption (A.2). Inequality (2.6) follows immediately from the Schwarz inequality. The operator D is given by

$$Du = \sum_{i=1}^N b_i \frac{\partial u}{\partial x_i} + (c-1)u,$$

and hence (2.8) clearly holds. Finally, we note that for $w \in H^1(\Omega)$ and $\phi \in \mathcal{M}$,

$$(3.5) \quad (D^*w, \phi) = (w, D\phi) = -(Dw, \phi) + \sum_{i=1}^N \left\{ \int_{\partial\Omega} b_i n_i w \phi ds - \left(\frac{\partial b_i}{\partial x_i} w, \phi \right) \right\},$$

where n_i is the component of the outward normal in the i th direction. We assume that b_i is in $C^1(\Omega)$ and that c is in $L^\infty(\Omega)$. The boundary term in (3.5) vanishes in the case of boundary conditions (3.3) and hence $D^* : H^1(\Omega) \mapsto L^2(\partial\Omega)$ in this case. In the case of boundary conditions (3.2), by a well-known trace inequality,

$$\left| \sum_{i=1}^N \int_{\partial\Omega} b_i n_i w \phi \, ds \right| \leq C \|w\|_{H^{1/2+\epsilon}(\Omega)} \|\phi\|_{H^{1/2+\epsilon}(\Omega)},$$

from which it follows that $D^* : H^1(\Omega) \mapsto H^{-1/2-\epsilon}(\Omega)$. Thus (A.2) holds for either application.

We next consider the finite element approximation subspaces. For simplicity, we shall only describe a piecewise linear application in two dimensions. The application to higher-dimensional problems and more general approximation subspaces is straightforward. We write $\Omega = \bigcup \tau_i^1$, where $\tau_1 \equiv \{\tau_i^1\}$ is a collection of triangles with mutually disjoint interiors. We assume that these triangles are of quasi-uniform size h_1 . This means that there are positive constants c and C such that the diameter of every triangle is bounded by Ch_1 and each triangle contains a circle of radius ch_1 . We define a sequence of triangulations by induction. Assume that the triangulation $\tau_{k-1} = \{\tau_i^{k-1}\}$ has been defined. The triangles of τ_k are formed by connecting the midpoints of the edges of the triangles in τ_{k-1} . Thus, each triangle in τ_{k-1} gives rise to four triangles in τ_k .

The approximation subspace \mathcal{M}_k consists of functions which are continuous and piecewise linear with respect to the triangulation τ_k . In the case of Dirichlet boundary conditions, we additionally require that the functions in τ_k vanish on $\partial\Omega$. In the case of boundary conditions (3.2) and $c(x) = 0$, we also require that the functions in τ_k have zero mean value. For these spaces, $h_k = 2^{-k+1}h_1$ and classical techniques in the theory of finite elements imply that (A.3) and (A.4) hold.

We finally define the discrete inner products. Let x_{ij}^k , $j = 1, 2, 3$, denote the vertices of the i th triangle of the k th grid. Define

$$(3.6) \quad (\phi, \chi)_k = 1/3 \sum_i |\tau_i^k| \sum_{j=1}^3 \phi(x_{ij}^k) \chi(x_{ij}^k).$$

Here $|\tau_i^k|$ denotes the area of the triangle τ_i^k . It is not difficult to show that (A.5) holds for this inner product. Note that (3.6) can be rewritten

$$(3.7) \quad (\phi, \chi)_k = \sum_i \omega_i^k \phi(y_i^k) \chi(y_i^k),$$

where $\{y_i^k\}$ are the nodes of the k th grid and ω_i^k is an appropriate weight function. Note that (3.7) implies that the solution of problems of the form (2.4) reduces to division by the weights $\{\omega_i^k\}$.

4. Multigrid Algorithms. We will define the multigrid algorithms in this section and develop certain recurrence relations which will be used in the iterative convergence analysis given later in the paper. The multigrid algorithm defines a linear operator B_k on \mathcal{M}_k which is an approximate inverse for A_k . We will consider the so-called ‘symmetric multigrid scheme’. Here ‘symmetric’ refers to the fact that the relaxation process used results from an iterative scheme for the symmetric operator $A_k^* A_k$.

We define the operator $B_k : \mathcal{M}_k \mapsto \mathcal{M}_k$ by induction on k . As we shall see in later sections, for stability, the coarsest grid in the multigrid process must not be too coarse. To this end, we shall define our algorithms starting from the intermediate grid level j , $1 \leq j < J$. In this algorithm, we assume that the operator B_j equals A_j^{-1} , although some results still hold when B_j is defined differently (see Remark 5.2).

The Multigrid Algorithm.

Set $B_j = A_j^{-1}$. Assume that $B_{k-1} : \mathcal{M}_{k-1} \mapsto \mathcal{M}_{k-1}$ has been defined and define $B_k g$ for $g \in \mathcal{M}_k$ and $k = j + 1, \dots, J$ as follows:

- (1) Set $x^0 = 0$ and $q^0 = 0$.
- (2) For $l = 1, \dots, m(k)$, define

$$(4.1) \quad x^l = x^{l-1} + \mu_k^{-2} A_k^*(g - A_k x^{l-1}),$$

where μ_k is the largest eigenvalue of $L_k = (A_k^* A_k)^{1/2}$.

- (3) Define $B_k g = x^{m(k)} + q^p$, where q^i , for $i = 1, 2, \dots, p$, is defined by

$$(4.2) \quad q^i = q^{i-1} + B_{k-1}[I_{k-1}(g - A_k x^{m(k)}) - A_{k-1} q^{i-1}].$$

The heuristic motivation for the above algorithm is as follows. Step (2) is a smoothing process and is designed to reduce the high-frequency components of the error. The low-frequency components of the error are then reduced by the coarser grid correction (3).

Remark 4.1. We have used μ_k^2 in (4.1) for convenience. In actual algorithms, any reasonable bound for the largest eigenvalue of the system $A_k^* A_k$ can be used.

Let $g = A_k x$ and $K_k = I - \mu_k^{-2} A_k^* A_k$. Clearly

$$x - x^{m(k)} = K_k^{m(k)} x.$$

It is straightforward to check that q^p satisfies

$$q^p = (I - (I - B_{k-1} A_{k-1})^p) P_{k-1}(x - x^{m(k)}).$$

Combining the above equalities gives

$$(4.3) \quad I - B_k A_k = [(I - P_{k-1}) + (I - B_{k-1} A_{k-1})^p P_{k-1}] K_k^{m(k)}.$$

The relation (4.3) provides a fundamental identity for the analysis of the multigrid algorithm.

The goal of this paper is to prove inequalities of the form

$$(4.4) \quad \|I - B_k A_k\|_{k,1}^2 \leq \delta_k.$$

Such inequalities immediately imply that the linear iteration

$$U^{n+1} = U^n + B_k(F - A_k U^n)$$

converges to the solution U of

$$A_k U = F$$

with a rate of $\sqrt{\delta_k}$ per step in the norm $\|\cdot\|_{k,1}$. Equality (4.3) gives a way of relating the reduction δ_k to that of the $(k-1)$ -grid and hence provides a key ingredient for a mathematical induction argument.

5. The Convergence Theorems and their Proofs. We give our convergence results for multigrid algorithms in this section. We first give results for the variable \mathcal{V} -cycle. Next, we consider the \mathcal{V} -cycle with constant $m(k) = m$. Finally, we consider the \mathcal{W} -cycle algorithms. The proofs of these theorems depend on three lemmas. These lemmas are central to the analysis of the paper and will be proved in the next section. In this section, we prove our multigrid theorems, assuming the lemmas.

We start by stating the lemmas. The first lemma gives a so-called ‘regularity and approximation’ estimate for the projection operator P_k .

LEMMA 5.1. *If h_j is sufficiently small, there exists a positive constant C not depending on k such that*

$$\|(I - P_{k-1})v\|_1^2 \leq C(\mu_k^{-1} \|L_k v\|_k^2)^\alpha (L_k v, v)_k^{1-\alpha} \quad \text{for all } v \in \mathcal{M}_k.$$

The next two lemmas represent an essential part of the analysis of this paper. Their proof uses the Dunford-Taylor integral formula for operators and is given in the next section.

LEMMA 5.2. *If h_j is sufficiently small, there exists a positive constant C not depending on k such that for all $v \in \mathcal{M}_k$, $\chi \in \mathcal{M}_{k-1}$,*

$$(5.1) \quad (L_k(I - P_{k-1})v, \chi)_k \leq Ch_k^{\alpha-1/2-\varepsilon} \|(I - P_{k-1})v\|_{k,1} \|\chi\|_{k,1}$$

holds for any positive ε .

LEMMA 5.3. *If h_j is sufficiently small, there exists a positive constant C not depending on k such that for all $\chi \in \mathcal{M}_{k-1}$,*

$$|\|\chi\|_{k,1}^2 - \|\chi\|_{k-1,1}^2| \leq Ch_k^{4\alpha-3}(1 + |\ln h_k|) \|\chi\|_{k,1}^2.$$

We can now state and prove the convergence theorem for the variable \mathcal{V} -cycle algorithm.

THEOREM 1. *Let $p = 1$ and assume that $m(k)$ satisfies*

$$(5.2) \quad \beta_0 m(k) \leq m(k-1) \leq \beta_1 m(k)$$

where β_0 and β_1 are constants greater than one and independent of k for $k = j+2, \dots, J$. Let γ be positive and less than $\min(\alpha - 1/2, 4\alpha - 3)$. Then there exist positive constants M and ν not depending on k such that when $h_j \leq \nu$, (4.4) holds with

$$(5.3) \quad \delta_k = \frac{M}{M + m(k)^{\alpha/2}}$$

for $k = j+1, \dots, J$.

Proof. We will prove the theorem by induction. For the purpose of this proof, let $m(j) = \beta_0 m(j+1)$ (note that $m(j)$ does not appear in the definition of the multigrid process). Clearly, (4.4) holds for $k = j$ with δ_k given by (5.3). Let $k \in \{j+1, \dots, J\}$ and assume that (4.4) holds for $k-1$ with δ_{k-1} given by (5.3). It follows from the recursive relation (4.3) that

$$\begin{aligned} \|(I - B_k A_k)v\|_{k,1}^2 &= \|(I - P_{k-1})\tilde{v}\|_{k,1}^2 + \|(I - B_{k-1}A_{k-1})P_{k-1}\tilde{v}\|_{k,1}^2 \\ &\quad + 2(L_k(I - P_{k-1})\tilde{v}, (I - B_{k-1}A_{k-1})P_{k-1}\tilde{v})_k, \end{aligned}$$

where $\tilde{v} = K_k^{m(k)} v$. Applying Lemma 5.2 gives

$$\begin{aligned} & \|(I - B_k A_k)v\|_{k,1}^2 \\ & \leq (1 + Ch_k^{\alpha-1/2-\varepsilon}) (\|(I - P_{k-1})\tilde{v}\|_{k,1}^2 + \|(I - B_{k-1}A_{k-1})P_{k-1}\tilde{v}\|_{k,1}^2). \end{aligned}$$

Using Lemma 5.3 and the induction hypothesis, we deduce that

$$\begin{aligned} \|(I - B_{k-1}A_{k-1})P_{k-1}\tilde{v}\|_{k,1}^2 & \leq (1 + Ch_k^\gamma) \|(I - B_{k-1}A_{k-1})P_{k-1}\tilde{v}\|_{k-1,1}^2 \\ & \leq \delta_{k-1}(1 + Ch_k^\gamma) \|P_{k-1}\tilde{v}\|_{k-1,1}^2 \\ & \leq \delta_{k-1}(1 + Ch_k^\gamma) \|P_{k-1}\tilde{v}\|_{k,1}^2 \end{aligned}$$

holds for any fixed γ less than $4\alpha - 3$. We remind the reader that here and throughout the paper, C denotes a generic positive constant which may take on different values from line to line. It follows from Lemma 5.2 that

$$\begin{aligned} \|\tilde{v}\|_{k,1}^2 & = \|P_{k-1}\tilde{v}\|_{k,1}^2 + \|(I - P_{k-1})\tilde{v}\|_{k,1}^2 + 2(L_k(I - P_{k-1})\tilde{v}, P_{k-1}\tilde{v})_k \\ & \geq (1 - Ch_k^{\alpha-1/2-\varepsilon})(\|P_{k-1}\tilde{v}\|_{k,1}^2 + \|(I - P_{k-1})\tilde{v}\|_{k,1}^2) \end{aligned}$$

and thus for ν sufficiently small

$$\|P_{k-1}\tilde{v}\|_{k,1}^2 \leq (1 + Ch_k^{\alpha-1/2-\varepsilon}) \|\tilde{v}\|_{k,1}^2 - \|(I - P_{k-1})\tilde{v}\|_{k,1}^2.$$

Requiring, in addition, that $\gamma < \alpha - 1/2$ and combining the above inequalities gives

$$\|(I - B_k A_k)v\|_{k,1}^2 \leq (1 + Ch_j^\gamma) \{(1 - \delta_{k-1}) \|(I - P_{k-1})\tilde{v}\|_{k,1}^2 + \delta_{k-1} \|\tilde{v}\|_{k,1}^2\}.$$

By Lemma 5.1, the Schwarz inequality, and a generalized arithmetic geometric mean inequality,

$$\begin{aligned} \|(I - P_{k-1})\tilde{v}\|_{k,1}^2 & \leq C (\mu_k^{-1} (L_k^2 \tilde{v}, \tilde{v})_k)^\alpha (L_k \tilde{v}, \tilde{v})_k^{1-\alpha} \\ & \leq C (\mu_k^{-2} (L_k^3 \tilde{v}, \tilde{v})_k)^{\alpha/2} (L_k \tilde{v}, \tilde{v})_k^{1-\alpha/2} \\ & \leq C \left\{ \eta_k \mu_k^{-2} (L_k^3 \tilde{v}, \tilde{v})_k + \eta_k^{-\alpha/(2-\alpha)} (L_k \tilde{v}, \tilde{v})_k \right\} \end{aligned}$$

holds for any positive constant η_k . Using the definition of K_k and the fact that its eigenvalues are in the interval $[0, 1)$ gives

$$\begin{aligned} \mu_k^{-2} (L_k^3 \tilde{v}, \tilde{v})_k & = (L_k(I - K_k)K_k^{m(k)} v, K_k^{m(k)} v)_k \\ & \leq (2m(k))^{-1} \sum_{l=0}^{2m(k)-1} (L_k(I - K_k)K_k^l v, v)_k \\ & = (2m(k))^{-1} (L_k(I - K_k^{2m(k)}) v, v)_k. \end{aligned}$$

Combining the above inequalities gives

$$\begin{aligned} & \|(I - B_k A_k)v\|_{k,1}^2 \\ (5.4) \quad & \leq (1 + C_1 h_k^\gamma) \left\{ C_0(1 - \delta_{k-1}) \eta_k m(k)^{-1} (L_k(I - K_k^{2m(k)}) v, v)_k \right. \\ & \quad \left. + [C_0(1 - \delta_{k-1}) \eta_k^{-\alpha/(2-\alpha)} + \delta_{k-1}] (L_k K_k^{2m(k)} v, v)_k \right\}. \end{aligned}$$

Setting $C_2 = C_0(1 + C_1)$, we see that the theorem will follow if we can choose η_k , h_j and M such that

$$(5.5) \quad C_2(1 - \delta_{k-1}) \eta_k m(k)^{-1} \leq \delta_k$$

and

$$(5.6) \quad C_2(1 - \delta_{k-1})\eta_k^{-\alpha/(2-\alpha)} + C_1 h_j^\gamma \delta_{k-1} \leq \delta_k - \delta_{k-1}.$$

We choose η_k by

$$(5.7) \quad C_2(1 - \delta_{k-1})\eta_k m(k)^{-1} = \delta_{k-1},$$

from which (5.5) immediately follows. Solving for η_k in (5.7) and using this result in (5.6) implies that it is sufficient to choose M and h_j so that

$$(5.8) \quad C_3(1 - \delta_{k-1})^{2/(2-\alpha)} m(k)^{-\alpha/(2-\alpha)} + C_1 h_j^\gamma \delta_{k-1}^{2/(2-\alpha)} \leq (\delta_k - \delta_{k-1}) \delta_{k-1}^{\alpha/(2-\alpha)}.$$

Let $\mathcal{D}(k) \equiv M + m(k)^{\alpha/2}$ and $\beta_k = m(k-1)/m(k) \in [\beta_0, \beta_1]$; then

$$(5.9) \quad 1 - \delta_{k-1} = \frac{(\beta_k m(k))^{\alpha/2}}{\mathcal{D}(k-1)}.$$

A direct computation using (5.9) and the identity $\delta_k = M/\mathcal{D}(k)$ shows that (5.8) is equivalent to

$$(5.10) \quad C_3 \beta_k^{\alpha/(2-\alpha)} M^{-\alpha/(2-\alpha)} + C_1 h_j^\gamma M \leq \frac{M m(k)^{\alpha/2}}{\mathcal{D}(k)} (\beta_k^{\alpha/2} - 1).$$

Note that if $M \geq 1$ then

$$C_4 \equiv (\beta_0^{\alpha/2} - 1)/2 \leq \frac{M m(k)^{\alpha/2}}{M + m(k)^{\alpha/2}} (\beta_k^{\alpha/2} - 1),$$

hence it suffices to have

$$C_5 M^{-\alpha/(2-\alpha)} + C_1 h_j^\gamma M \leq C_4,$$

where $C_5 = C_3 \beta_1^{\alpha/(2-\alpha)}$. Thus, taking $M \geq 1$ large enough so that

$$C_5 M^{-\alpha/(2-\alpha)} \leq C_4/2$$

and

$$h_j \leq \nu \leq C_4^{1/\gamma} (2C_1 M)^{-1/\gamma}$$

completes the proof of the theorem.

We next prove a theorem for the standard \mathcal{V} -cycle algorithm.

THEOREM 2. *Consider the \mathcal{V} -cycle algorithm ($p = 1$) with $m(k) = m$ for all k . Let γ be positive and less than $\min(\alpha - 1/2, 4\alpha - 3)$. Then there exist positive constants M , c , and ν not depending on k such that when $h_j \leq \min(\nu, c(j-1)^{-2/(\alpha\gamma)})$, (4.4) holds with*

$$(5.11) \quad \delta_k = \frac{M k^{(2-\alpha)/\alpha}}{(M k^{(2-\alpha)/\alpha} + m^{\alpha/2})}$$

for $k = j+1, \dots, J$.

Remark 5.1. The theorem suggests that the \mathcal{V} -cycle may be less robust than the variable \mathcal{V} -cycle. Note that the convergence estimate for the \mathcal{V} -cycle algorithm deteriorates as k becomes larger, even in the case $\alpha = 1$. Furthermore, the theorem suggests that for stability, the coarsest grid must become finer as the number of grid levels increases.

Proof. The proof of this theorem is essentially contained in the proof of Theorem 1 and the proof of Theorem 1 of [5]. Indeed, (5.4) is valid with $m(k) = m$ and hence it suffices to choose η_k , M , h_j , and c so that (5.5) and (5.6) are satisfied. We choose η_k by (5.7) and reduce (5.5)–(5.6) to (5.8). Making similar algebraic manipulations (compare with (5.10)), we see it suffices to choose the parameters so that

$$\begin{aligned} C_3 M^{-\alpha/(2-\alpha)} + C_1 h_j^\gamma (k-1)^{2/\alpha} \\ \leq \frac{Mm^{\alpha/2}}{\mathcal{D}(k)} [k^{(2-\alpha)/\alpha} - (k-1)^{(2-\alpha)/\alpha}] (k-1), \end{aligned}$$

where $\mathcal{D}(k) \equiv Mk^{(2-\alpha)/\alpha} + m^{\alpha/2}$. Noting that $k \geq 2$ and $(2-\alpha)/\alpha > 0$, elementary arguments imply

$$k^{(2-\alpha)/\alpha} \leq C_6 [k^{(2-\alpha)/\alpha} - (k-1)^{(2-\alpha)/\alpha}] (k-1).$$

Thus, it suffices to prove

$$C_3 M^{-\alpha/(2-\alpha)} + C_1 h_j^\gamma M (k-1)^{2/\alpha} \leq C_6 \frac{Mm^{\alpha/2} k^{(2-\alpha)/\alpha}}{\mathcal{D}(k)}.$$

We set

$$M = \left(\frac{1 + \tilde{M}}{\tilde{M}} \right)^{(2-\alpha)/\alpha} \tilde{M}$$

and define \tilde{M} by

$$C_3 \tilde{M}^{-\alpha/(2-\alpha)} = C_6/2.$$

Then

$$C_3 M^{-\alpha/(2-\alpha)} \leq \frac{C_6}{2} \frac{M}{1+M} \leq \frac{C_6}{2} \frac{Mm^{\alpha/2} k^{(2-\alpha)/\alpha}}{\mathcal{D}(k)}.$$

We then set

$$c = \left(\frac{C_6 M}{2C_1(1+M)} \right)^{1/\gamma},$$

from which it follows that $h_j \leq c(j-1)^{-2/(\gamma\alpha)}$ implies

$$C_1 h_j^\gamma M (k-1)^{2/\alpha} \leq \frac{C_6}{2} \frac{M}{1+M} \leq \frac{C_6}{2} \frac{Mm^{\alpha/2} k^{(2-\alpha)/\alpha}}{\mathcal{D}(k)}.$$

Combining the above inequalities proves the theorem.

The last theorem which we shall prove is for the \mathcal{W} -cycle algorithm.

THEOREM 3. *Consider the \mathcal{W} -cycle algorithm ($p = 2$) with $m(k) \equiv m$ for all k . Let γ be positive and less than $\min(\alpha - 1/2, 4\alpha - 3)$. Then there exist positive constants M and ν such that when $h_j \leq \nu$, (4.4) holds with*

$$(5.12) \quad \delta_k \equiv \delta = (1 + m/M)^{-\alpha/2}$$

for $k = j + 1, \dots, J$.

Proof. The proof of this theorem is essentially contained in the proof of Theorem 1 and the proof of Theorem 3 of [5]. Since the term involving $(I - B_{k-1}A_{k-1})$ appears squared in (4.3), following the proof of Theorem 1, we see that (5.4) holds

with δ_{k-1} replaced by δ^2 . We see that the theorem will follow if we can choose $\eta_k = \eta$, ν and M such that

$$(5.13) \quad C_2(1 - \delta^2)\eta m^{-1} \leq \delta$$

and

$$(5.14) \quad C_2(1 - \delta^2)\eta^{-\alpha/(2-\alpha)} + C_1 h_j^\gamma \delta^2 \leq \delta - \delta^2.$$

We choose η so that (5.13) holds with equality. Solving for η and using this result in (5.14) implies that it is sufficient to choose M and ν so that

$$(5.15) \quad C_3(1 - \delta^2)^{2/(2-\alpha)} m^{-\alpha/(2-\alpha)} + C_1 h_j^\gamma \delta^{(4-\alpha)/(2-\alpha)} \leq (1 - \delta) \delta^{2/(2-\alpha)}.$$

It is elementary to see that

$$(5.16) \quad 2^{-2/(2-\alpha)} \leq (1 - \delta)^{-\alpha/(2-\alpha)} \left(\frac{m}{M}\right)^{\alpha/(2-\alpha)} \left(\frac{\delta}{1 + \delta}\right)^{2/(2-\alpha)}$$

for δ given by (5.12). Define M by

$$M^{\alpha/(2-\alpha)} 2^{-2/(2-\alpha)} = 2C_3.$$

Then

$$C_3(1 - \delta^2)^{2/(2-\alpha)} m^{-\alpha/(2-\alpha)} \leq \frac{1}{2}(1 - \delta) \delta^{2/(2-\alpha)}.$$

Choosing

$$(5.17) \quad h_j \leq \nu \leq \left(\frac{(1 - \delta)}{2C_1 \delta}\right)^{1/\gamma}$$

implies

$$C_1 h_j^\gamma \delta^{(4-\alpha)/(2-\alpha)} \leq \frac{1}{2}(1 - \delta) \delta^{2/(2-\alpha)}.$$

This completes the proof of the theorem.

Remark 5.2. The multigrid process described in Section 4 requires that the problem on the coarsest grid be solved exactly, i.e., $B_j = A_j^{-1}$. It is possible to relax this restriction and still apply the results of this paper. We consider, for example, the variable \mathcal{V} -cycle multigrid algorithm. From the proof of Theorem 1 it is immediate that the theorem will still hold as long as B_j satisfies

$$(5.18) \quad \|(I - B_j A_j)\|_{k,1}^2 \leq \delta_j,$$

where

$$(5.19) \quad \delta_j = \frac{M}{M + \beta_0^{\alpha/2} m(j+1)^{\alpha/2}}.$$

One obvious choice for an iterative definition of B_j is $B_j g = x^{m_j}$ where x^l for $l = 1, \dots, m_j$ is given by (4.1) with $k = j$. Here m_j is some integer to be specified. An iterative definition of B_j has the advantage that no additional coding is necessary (in contrast to the use of $B_j = A_j^{-1}$, where direct solvers for nonsymmetric and indefinite problems must be introduced into the code). There are two additional factors involved in the use of an iterative process for B_j . First, one would like to avoid the coarsest grids so that $h_j \leq \nu$ is satisfied. Secondly, the computational work on the coarsest grid should not increase the asymptotic work of the algorithm.

We consider the application described in Section 3. We should like the multigrid algorithm to achieve a reduction δ_J which is independent of h_J , with computational effort bounded by a constant times the number of grid points in the finest grid. Let $N(k)$ denote the number of degrees of freedom in the k th grid level. We assume $N(k)/N(k-1) \geq c_0 > \beta_1$, and hence the amount of work on the grids $1, \dots, J$ will be bounded by $O(N(J))$ [3], [6]. It is not difficult to see that for B_j defined as above,

$$\|I - B_j A_j\|_{k,1}^2 \leq (1 - ch_j^4)^{m_j}.$$

On the other hand, if we take $\beta_0 = \beta_1 = 2$,

$$\frac{M}{M + 2^{\alpha/2} m(j+1)^{\alpha/2}} \geq C(h_J/h_j)^{\alpha/2}.$$

Consequently, to satisfy (5.18)–(5.19), we need only take

$$(5.20) \quad m_j = O(h_J^{-1} h_j^{-3}).$$

The work constraint is then $h_J^{-1} h_j^{-5} \leq ch_j^{-2}$. Thus setting $h_j = h_j^{1/5}$ and defining m_j by (5.20) gives rise to a multigrid algorithm which yields a uniform reduction independent of h_J , with an operation count bounded by a constant times the number of degrees of freedom on the finest grid.

6. The Proof of Lemmas 5.1, 5.2 and 5.3. This section will provide the proofs of Lemmas 5.1–5.3. Before proceeding, let us state two propositions and two preliminary lemmas.

PROPOSITION 6.1. *There are positive constants c and C not depending on $v \in \mathcal{M}$ such that*

$$\|v\|_{H^1}^2 \leq C\{A(v,v) + c\|v\|_H^2\}.$$

PROPOSITION 6.2. *For $v \in H^{1+\alpha}$ and $0 \leq \delta \leq \alpha$,*

$$(6.1) \quad \|(I - \hat{P}_k)v\|_{H^{1-\delta}} \leq Ch_k^\delta \|(I - \hat{P}_k)v\|_{H^1}.$$

If h_j is sufficiently small, then P_k is well defined and

$$(6.2) \quad \|(I - P_k)v\|_{H^1} \leq C \inf_{\chi \in \mathcal{M}_k} \|v - \chi\|_{H^1},$$

for all $v \in \mathcal{M}$.

Proposition 6.1 follows immediately from (2.6). (6.1) follows from a standard duality argument and (6.2) can be proved by using the techniques given in [20].

We next introduce the preliminary lemmas. The first lemma was essentially proved in [1].

LEMMA 6.1. *Let $0 \leq s \leq 1$. There exist positive constants c_1, c_2 and c_3 such that*

$$\|\chi\|_{H^s} \leq c_1 \|\chi\|_{k,s} \leq c_2 \|\chi\|_{k,s} \leq c_3 \|\chi\|_{H^s} \quad \text{for all } \chi \in \mathcal{M}_k.$$

In addition, there are constants c and C satisfying

$$c \|\chi\|_{k,2} \leq \|\chi\|_{k,2} \leq C \|\chi\|_{k,2} \quad \text{for all } \chi \in \mathcal{M}_k.$$

LEMMA 6.2. *There exists a positive constant C which does not depend upon $v \in H^1$ such that*

$$(6.3) \quad \|(I - P_k^0)v\|_H \leq Ch_k \|v\|_{H^1},$$

$$(6.4) \quad \|P_k^0 v\|_{H^s} \leq C\|v\|_{H^s} \quad \text{for all } 0 \leq s \leq 1.$$

Proof. Let πv denote the H projection of v into \mathcal{M}_k . Using (A.3), (A.4) and standard techniques of finite element analysis gives

$$\begin{aligned} \|\pi v\|_{H^1} &\leq C\|v\|_{H^1}, \\ \|(I - \pi)v\|_H &\leq Ch_k \|v\|_{H^1}. \end{aligned}$$

For $\chi \in \mathcal{M}_k$, by (2.10),

$$((P_k^0 - \pi)v, \chi)_k = (\pi v, \chi) - (\pi v, \chi)_k \leq Ch_k \|\pi v\|_{H^1} \|\chi\|_k,$$

hence

$$\|(P_k^0 - \pi)v\|_k \leq Ch_k \|\pi v\|_{H^1}.$$

Estimate (6.3) follows from the triangle inequality.

For (6.4), by interpolation, it suffices to verify the cases $s = 0$ and $s = 1$. The case for $s = 0$ follows immediately from the definition of P_k^0 and (2.9). For $s = 1$, the argument is standard and proceeds as follows:

$$\begin{aligned} \|P_k^0 v\|_{H^1} &\leq \|(P_k^0 - \pi)v\|_{H^1} + \|\pi v\|_{H^1} \\ &\leq Ch_k^{-1} \|(P_k^0 - \pi)v\|_k + C\|v\|_{H^1} \leq C\|v\|_{H^1}. \end{aligned}$$

This completes the proof of the lemma.

We can now prove Lemma 5.1.

Proof of Lemma 5.1. Following the argument in [5], we can easily show (using our assumptions and definitions) that

$$\|(I - \hat{P}_{k-1})v\|_1^2 \leq C(h_k^2 \|\hat{A}_k v\|_k^2)^\alpha \hat{A}(v, v)^{1-\alpha} \quad \text{for all } v \in \mathcal{M}_k.$$

We note that (A.4) and Lemma 6.1 imply that $h_k^2 \leq C\mu_k^{-1}$. The lemma now follows from (6.2) and Lemma 6.1.

The proofs of Lemmas 5.2 and 5.3 require some technical perturbation estimates. We consider the term on the left-hand side of (5.1). Let $G_k = L_k - A_k$; then since

$$(A_k(I - P_{k-1})v, \chi)_k = 0,$$

we have

$$\begin{aligned} (6.5) \quad (L_k(I - P_{k-1})v, \chi)_k &= (G_k(I - P_{k-1})v, \chi)_k = ((I - P_{k-1})v, G_k^* \chi)_k \\ &\leq \|(I - P_{k-1})v\|_k \|G_k^* \chi\|_k. \end{aligned}$$

Thus, we must estimate $G_k^* = L_k - \hat{A}_k - D_k^*$.

In light of (6.5), we see that it would be useful to estimate the difference $L_k - \hat{A}_k$. Note that L_k is defined as the positive square root of the discrete operator $L_k^2 \equiv A_k^* A_k$. An alternative expression for L_k is given by the Dunford-Taylor integral representation (cf. [9]):

$$(6.6) \quad L_k = (2\pi i)^{-1} \int_{\Gamma} z^{1/2} \mathcal{R}_z(L_k^2) dz,$$

where $\mathcal{R}_z(L_k^2) \equiv (z - L_k^2)^{-1}$ and Γ is a simple closed curve in the right half (complex) plane which encloses the spectrum of L_k^2 . Let $\kappa_1, \kappa_2 > 0$ be such that the eigenvalues of L_k^2 and \hat{A}_k^2 are in the interval $[2\kappa_1, 2\kappa_2]$. In this paper, we will take Γ as illustrated in Figure 6.1, i.e.,

$$\begin{aligned}\Gamma = & \{(\kappa_1, y) \mid y \in [-\kappa_1, \kappa_1]\} \cup \{(t, t) \mid t \in [\kappa_1, 2\kappa_2]\} \\ & \cup \{(t, -t) \mid t \in [\kappa_1, 2\kappa_2]\} \cup \{(2\kappa_2, y) \mid y \in [-2\kappa_2, 2\kappa_2]\}.\end{aligned}$$

Using an expression similar to (6.6) for \hat{A}_k gives

$$(6.7) \quad L_k - \hat{A}_k = (2\pi i)^{-1} \int_{\Gamma} z^{1/2} \mathcal{R}_z(L_k^2)(L_k^2 - \hat{A}_k^2) \mathcal{R}_z(\hat{A}_k^2) dz.$$

To estimate (6.7) we shall use the bounds given in the following lemma.

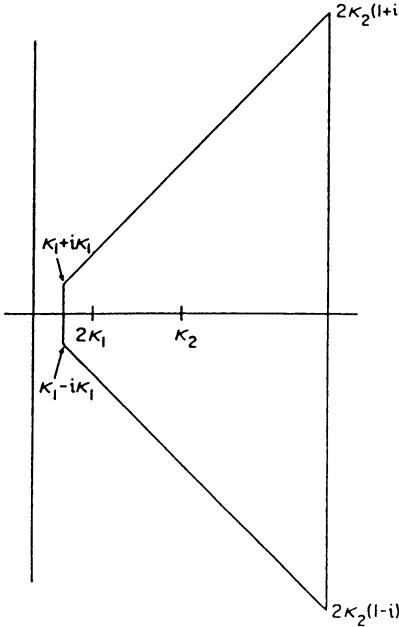


FIGURE 6.1
The curve Γ used in (6.6).

LEMMA 6.3. *Let S and T be symmetric positive definite operators on \mathcal{M}_k satisfying*

$$\begin{aligned}2\kappa_1 \|\chi\|_k^2 &\leq (S^2 \chi, \chi)_k \leq \kappa_2 \|\chi\|_k^2, \\ 2\kappa_1 \|\chi\|_k^2 &\leq (T^2 \chi, \chi)_k \leq \kappa_2 \|\chi\|_k^2,\end{aligned}$$

for all $\chi \in \mathcal{M}_k$. Assume that $\kappa_1 \geq c$ independently of k . We allow S , T and κ_2 to depend on k . Then

$$(6.8) \quad \int_{\Gamma} |z|^{1/2} \|S \cdot \mathcal{R}_z(S^2)\|_k \|\mathcal{R}_z(T^2)\|_k d|z| \leq C(1 + \ln(\kappa_2/\kappa_1)),$$

and for any $\chi \in \mathcal{M}_k$,

$$(6.9) \quad \int_{\Gamma} |z|^{1/2} \|S^{1/2} \cdot \mathcal{R}_z(S^2) \chi\|_k^2 d|z| \leq C \|\chi\|_k^2.$$

Proof. By symmetry, it suffices to derive the above bounds for the curve $\Gamma_+ \equiv \Gamma_1 \cup \Gamma_2 \cup \Gamma_3$, where $\Gamma_1 \equiv \{(\kappa_1, y) \mid y \in [0, \kappa_1]\}$, $\Gamma_2 \equiv \{(t, t) \mid t \in [\kappa_1, 2\kappa_2]\}$ and $\Gamma_3 \equiv \{(2\kappa_2, y) \mid y \in [0, 2\kappa_2]\}$. By expansion in terms of eigenvectors, it is easy to see that

$$(6.10) \quad \|S^\beta \mathcal{R}_z(S^2)\|_k \leq \max_{\lambda \in [\sqrt{2\kappa_1}, \sqrt{\kappa_2}]} \lambda^\beta |\lambda^2 - z|^{-1}, \quad \beta = 0, 1/2, 1.$$

A similar inequality obviously holds for T .

Let

$$\mathcal{F}_1(\gamma) \equiv \int_{\gamma} |z|^{1/2} \|S \cdot \mathcal{R}_z(S^2)\|_k \|\mathcal{R}_z(T^2)\|_k d|z|$$

and

$$\mathcal{F}_2(\gamma) \equiv \int_{\gamma} |z|^{1/2} \|S^{1/2} \cdot \mathcal{R}_z(S^2)\chi\|_k^2 d|z|.$$

Then by (6.10) and elementary estimates,

$$\begin{aligned} \mathcal{F}_1(\Gamma_1) &\leq C \int_{\Gamma_1} |z|^{1/2} \kappa_1^{-3/2} d|z| \leq C, \\ \mathcal{F}_1(\Gamma_2) &\leq C \int_{\Gamma_2} |z|^{-1} d|z| \leq C \ln(2\kappa_2/\kappa_1), \\ \mathcal{F}_1(\Gamma_3) &\leq C \int_0^\infty \frac{\kappa_2}{\kappa_2^2 + y^2} dy \leq C. \end{aligned}$$

This verifies (6.8). Similar arguments give

$$\begin{aligned} \mathcal{F}_2(\Gamma_1) &\leq C \left(\int_{\Gamma_1} |z|^{1/2} \kappa_1^{-3/2} d|z| \right) \|\chi\|_k^2 \leq C \|\chi\|_k^2, \\ \mathcal{F}_2(\Gamma_3) &\leq C \left(\int_0^\infty \frac{\kappa_2}{\kappa_2^2 + y^2} dy \right) \|\chi\|_k^2 \leq C \|\chi\|_k^2. \end{aligned}$$

To bound $\mathcal{F}_2(\Gamma_2)$, we expand in terms of the eigenvectors of S . Let $\{\lambda_i, \theta_i\}$ denote the eigenvalue-eigenvector pairs for the operator S . Without loss of generality, we may assume that $\{\theta_i\}$ form an orthonormal basis for \mathcal{M}_j . Clearly, $2\kappa_1 \leq \lambda_i^2 \leq \kappa_2$ holds for each i . Decomposing

$$\chi = \sum_i c_i \theta_i$$

gives

$$\|S^{1/2} \cdot \mathcal{R}_z(S^2)\chi\|_k^2 = \sum_i \frac{\lambda_i c_i^2}{|\lambda_i^2 - z|^2}.$$

Integrating term by term yields

$$(6.11) \quad \mathcal{F}_2(\Gamma_2) = \sum_i 2^{3/4} c_i^2 \int_{\kappa_1}^{2\kappa_2} \frac{\lambda_i t^{1/2}}{(\lambda_i^2 - t)^2 + t^2} dt.$$

Elementary manipulations show that the integrals in (6.11) are bounded uniformly in κ_1, κ_2 , and λ_i . Hence $\mathcal{F}_2(\Gamma_2) \leq C \|\chi\|_k^2$. This completes the proof of the lemma.

We now state and prove a lemma for estimating $L_k - \hat{A}_k$.

LEMMA 6.4. *Let h_j be sufficiently small. Then there exists a constant C such that for all $\chi \in \mathcal{M}_k$,*

$$\|L_k \chi - \hat{A}_k \chi\|_{H^1} \leq C h_k^{\alpha-1} (1 + |\ln h_k|) \|\hat{A}_k \chi\|_k$$

and

$$\|L_k \chi - \hat{A}_k \chi\|_k \leq C h_k^{\alpha-1} \|\chi\|_{H^1}.$$

Proof. By Lemma 6.1,

$$\|L_k \chi - \hat{A}_k \chi\|_{H^1} \leq C \|L_k^{1/2} (L_k - \hat{A}_k) \chi\|_k.$$

By (6.7), for any $\chi, \theta \in \mathcal{M}_k$,

$$(L_k^{1/2} (L_k - \hat{A}_k) \chi, \theta)_k = (2\pi i)^{-1} \int_{\Gamma} z^{1/2} (E_k \mathcal{R}_z(\hat{A}_k^2) \hat{A}_k \chi, L_k \mathcal{R}_z(L_k^2) \theta)_k dz,$$

where

$$E_k = L_k^{-1/2} (L_k^2 - \hat{A}_k^2) \hat{A}_k^{-1}.$$

By the Schwarz inequality and (6.8), with $S = L_k$ and $T = \hat{A}_k$,

$$(L_k^{1/2} (L_k - \hat{A}_k) \chi, \theta)_k \leq C (1 + |\ln h_k|) \|E_k\|_k \|\hat{A}_k \chi\|_k \|\theta\|_k.$$

Note that we have used the fact that κ_1 is bounded uniformly from below and by (A.3), we can take $\kappa_2 \leq C h_k^{-4}$. Similarly, by (6.7),

$$(L_k \chi - \hat{A}_k \chi, \theta)_k = (2\pi i)^{-1} \int_{\Gamma} z^{1/2} (E_k \hat{A}_k^{1/2} \mathcal{R}_z(\hat{A}_k^2) \hat{A}_k^{1/2} \chi, L_k^{1/2} \mathcal{R}_z(L_k^2) \theta)_k dz.$$

By the Schwarz inequality, Lemma 6.1 and (6.9),

$$|(L_k \chi - \hat{A}_k \chi, \theta)_k| \leq C \|E_k\|_k \|\chi\|_{H^1} \|\theta\|_k.$$

Thus, the proof of the lemma will be complete if we can show that

$$(6.12) \quad \|E_k\|_k \leq C h_k^{\alpha-1}.$$

Obviously,

$$(6.13) \quad L_k^2 - \hat{A}_k^2 = \hat{A}_k D_k + D_k^* \hat{A}_k + D_k^* D_k$$

and hence

$$(6.14) \quad \|E_k\|_k \leq \|L_k^{-1/2} \hat{A}_k D_k \hat{A}_k^{-1}\|_k + \|L_k^{-1/2} D_k^* \hat{A}_k \hat{A}_k^{-1}\|_k + \|L_k^{-1/2} D_k^* D_k \hat{A}_k^{-1}\|_k.$$

Using Lemmas 6.1 and 6.2 and (2.7) gives

$$(6.15) \quad \|L_k^{-1/2} D_k^* \hat{A}_k \hat{A}_k^{-1}\|_k = \|D_k L_k^{-1/2}\|_k = \|P_k^0 D L_k^{-1/2}\|_k \leq C.$$

Similarly,

$$\|L_k^{-1/2} D_k^* D_k \hat{A}_k^{-1}\|_k \leq C \|D_k L_k^{-1/2}\|_k \|D_k \hat{A}_k^{-1}\|_k \leq C.$$

For the first term of (6.14), using Lemma 6.1 gives

$$\|L_k^{-1/2} \hat{A}_k D_k \hat{A}_k^{-1}\|_k \leq \|L_k^{-1/2} \hat{A}_k^{1/2}\|_k \|\hat{A}_k^{1/2} D_k \hat{A}_k^{-1}\|_k \leq C \|\hat{A}_k^{1/2} D_k \hat{A}_k^{-1}\|_k.$$

Combining the above estimates, making an obvious change of variable, and applying Lemma 6.2 implies that the proof of the lemma will be complete if we show

$$(6.16) \quad \|D_k \chi\|_{H^1} \leq C h_k^{\alpha-1} \|\hat{A}_k \chi\|_k \quad \text{for all } \chi \in \mathcal{M}_k.$$

Fix $\chi \in \mathcal{M}_k$ and let $w \in \mathcal{M}$ be the solution to

$$\hat{A}(w, \phi) = (\hat{A}_k \chi, \phi) \quad \text{for all } \phi \in \mathcal{M}.$$

Clearly $\chi = \hat{P}_k w$. Now

$$\|D_k \chi\|_{H^1} \leq \|P_k^0 D(\chi - w)\|_{H^1} + \|P_k^0 Dw\|_{H^1}.$$

Applying (2.7), (A.3), (A.4), and Lemma 6.2 gives

$$\|P_k^0 D(\chi - w)\|_{H^1} \leq Ch_k^{-1} \|\chi - w\|_{H^1} \leq h_k^{\alpha-1} \|w\|_{H^{1+\alpha}}.$$

Finally, by (A.4), Lemma 6.2 and (2.8),

$$\|P_k^0 Dw\|_{H^1} \leq Ch_k^{\alpha-1} \|P_k^0 Dw\|_{H^\alpha} \leq Ch_k^{\alpha-1} \|Dw\|_{H^\alpha} \leq Ch_k^{\alpha-1} \|w\|_{1+\alpha}.$$

Inequality (6.16) now follows combining the above estimates with (A.1). This completes the proof of Lemma 6.4.

We can now prove Lemma 5.2.

Proof of Lemma 5.2. By (6.5), Lemma 6.1 and Proposition 6.2, it suffices to show that

$$\|G_k^* \chi\|_k \leq Ch^{-1/2-\varepsilon} \|\chi\|_{H^1}.$$

In turn, by Lemma 6.4 and the triangle inequality, noting that $\alpha > 1/2$, it suffices to show

$$(6.17) \quad \|D_k^* \chi\|_k \leq Ch^{-1/2-\varepsilon} \|\chi\|_{H^1}.$$

Let $\theta \in \mathcal{M}_k$; then by (A.2), (A.4) and Lemma 6.1,

$$(D_k^* \chi, \theta)_k = (D^* \chi, \theta) \leq C \|\chi\|_{H^1} \|\theta\|_{H^{1/2+\varepsilon}} \leq Ch_k^{-1/2-\varepsilon} \|\chi\|_{H^1} \|\theta\|_k.$$

Inequality (6.17) immediately follows. This completes the proof of the lemma.

We shall need two additional lemmas for the proof of Lemma 5.3. The first involves stability and approximation for the operator I_k .

LEMMA 6.5. *There exists a positive constant C such that for all $\chi \in \mathcal{M}_k$*

$$(6.18) \quad \|(I - I_{k-1})\chi\|_H \leq Ch_k \|\chi\|_{H^1}$$

and

$$(6.19) \quad \|I_{k-1}\chi\|_{H^1} \leq C \|\chi\|_{H^1}.$$

Proof. Note that by (2.9) and (2.10), for $\varphi \in \mathcal{M}_{k-1}$,

$$((I_{k-1} - P_{k-1}^0)\chi, \varphi)_{k-1} = (\chi, \varphi)_k - (\chi, \varphi) \leq Ch_k \|\chi\|_{H^1} \|\varphi\|_{k-1}.$$

This implies that

$$\|(I_{k-1} - P_{k-1}^0)\chi\|_{k-1} \leq Ch_k \|\chi\|_{H^1}.$$

The lemma then follows from Lemmas 6.1 and 6.2 and (A.4).

LEMMA 6.6. *There exists a positive constant C such that for all $\chi \in \mathcal{M}_{k-1}$*

$$(6.20) \quad \|\hat{A}_k \chi\|_k \leq Ch_k^{\alpha-1} \|\hat{A}_{k-1} \chi\|_{k-1},$$

$$(6.21) \quad \|L_k \chi\|_k \leq Ch_k^{\alpha-1} \|L_{k-1} \chi\|_{k-1},$$

$$(6.22) \quad \|\hat{A}_{k-1} \chi\|_{k-1} \leq Ch_k^{\alpha-1} \|I_{k-1} L_k \chi\|_{k-1}$$

and

$$(6.23) \quad \|L_k \chi\|_k \leq Ch_k^{2\alpha-2} \|I_{k-1} L_k \chi\|_{k-1}.$$

Proof. By Proposition 2, Lemma 6.1, (A.3) and (A.4), for all $\varphi \in \mathcal{M}_k$,

$$\|\varphi - \hat{P}_{k-1} \varphi\|_k \leq Ch_k^\alpha \|\hat{\varphi}\|_{H^1} \leq Ch_k^{\alpha-1} \|\varphi\|_k,$$

hence

$$\|\hat{P}_{k-1} \varphi\|_k \leq Ch_k^{\alpha-1} \|\varphi\|_k.$$

Therefore, for $\chi \in \mathcal{M}_{k-1}$,

$$\begin{aligned} (\hat{A}_k \chi, \varphi)_k &= \hat{A}(\chi, \varphi) = \hat{A}(\chi, \hat{P}_{k-1} \varphi) \\ &= (\hat{A}_{k-1} \chi, \hat{P}_{k-1} \varphi)_{k-1} \leq Ch_k^{\alpha-1} \|\hat{A}_{k-1} \chi\|_{k-1} \|\varphi\|_k. \end{aligned}$$

This proves (6.20). Inequality (6.21) then follows from (6.20) and Lemma 6.1.

We next prove (6.22). Noting that $\hat{A}_{k-1} = I_{k-1} \hat{A}_k$, the triangle inequality and Lemma 6.4 give

$$\begin{aligned} \|\hat{A}_{k-1} \chi\|_{k-1} &= \|I_{k-1} \hat{A}_k \chi\|_{k-1} \leq (\|I_{k-1} L_k \chi\|_{k-1} + \|(\hat{A}_k - L_k) \chi\|_k) \\ &\leq Ch_k^{\alpha-1} (\|I_{k-1} L_k \chi\|_{k-1} + \|\chi\|_{H^1}). \end{aligned}$$

Finally, we note that by Lemma 6.1 and (2.9),

$$\|\chi\|_{H^1}^2 \leq C(L_k \chi, \chi)_k \leq C\|I_{k-1} L_k \chi\|_{k-1} \|\chi\|_{H^1},$$

and hence

$$\|\chi\|_{H^1} \leq C\|I_{k-1} L_k \chi\|_{k-1}.$$

Combining the above inequalities completes the proof of (6.22). Inequality (6.23) follows immediately from (6.22), (6.20) and Lemma 6.1.

We are now ready to prove Lemma 5.3. However, before doing so, we note a few properties of our operators which are immediate consequences of the defining relations. As noted earlier, $\hat{A}_{k-1} = I_{k-1} \hat{A}_k$. Similarly, $D_{k-1} = I_{k-1} D_k$. In addition, the operator I_{k-1} is symmetric on both \mathcal{M}_k with the $(\cdot, \cdot)_k$ inner product as well as \mathcal{M}_{k-1} with the $(\cdot, \cdot)_{k-1}$ inner product.

Proof of Lemma 5.3. For $\chi \in \mathcal{M}_{k-1}$,

$$|\|\chi\|_{k,1}^2 - \|\chi\|_{k-1,1}^2| = ((\tilde{L}_{k-1} - L_{k-1})\chi, \chi)_{k-1},$$

where $\tilde{L}_{k-1} \equiv I_{k-1} L_k$. Note that the operator $\tilde{L}_{k-1} : \mathcal{M}_{k-1} \mapsto \mathcal{M}_{k-1}$ is symmetric and the eigenvalues of \tilde{L}_{k-1}^2 are in the interval $[c, Ch_k^{-4}]$ for appropriate constants c and C (independent of k). Applying an expression analogous to (6.7) gives

$$\begin{aligned} &\left((L_{k-1} - \tilde{L}_{k-1})\chi, \chi \right)_{k-1} \\ &= (2\pi i)^{-1} \int_{\Gamma} z^{1/2} (F_k L_{k-1} \mathcal{R}_z(L_{k-1}^2)\chi, \tilde{L}_{k-1} \mathcal{R}_z(\tilde{L}_{k-1}^2)\chi)_{k-1} dz, \end{aligned}$$

where

$$F_k = \tilde{L}_{k-1}^{-1} (L_{k-1}^2 - \tilde{L}_{k-1}^2) L_{k-1}^{-1}.$$

By the Schwarz inequality and (6.9),

$$\begin{aligned} \left| \left((L_{k-1} - \tilde{L}_{k-1})\chi, \chi \right)_{k-1} \right| &\leq C\|F_k\|_{k-1} \|L_{k-1}^{1/2} \chi\|_{k-1} \|\tilde{L}_{k-1}^{1/2} \chi\|_{k-1} \\ &\leq C\|F_k\|_{k-1} \|\chi\|_{k,1}^2, \end{aligned}$$

where the second inequality follows from Lemma 6.1 and the identity $(\tilde{L}_{k-1}\chi, \chi)_{k-1} = (L_k\chi, \chi)_k$. To complete the proof of the lemma, we need only bound $\|F_k\|_{k-1}$.

We start first with the identity

$$F_k = Q_1 + Q_2 + Q_3,$$

where

$$\begin{aligned} Q_1 &= (I - I_{k-1})(L_k - \hat{A}_k)L_{k-1}^{-1}, \\ Q_2 &= \tilde{L}_{k-1}^{-1}I_{k-1}(L_k - \hat{A}_k)(I - I_{k-1})\hat{A}_kL_{k-1}^{-1}, \\ Q_3 &= \tilde{L}_{k-1}^{-1}(L_{k-1}^2 - I_{k-1}L_k^2 - \hat{A}_{k-1}^2 + I_{k-1}\hat{A}_k^2)L_{k-1}^{-1}. \end{aligned}$$

Obviously, it suffices to bound the norms $\|Q_i\|_{k-1}$ for $i = 1, 2, 3$.

Let $\chi, \theta \in \mathcal{M}_{k-1}$. For Q_1 , by Lemmas 6.1, 6.4, 6.5 and 6.6, we have

$$\begin{aligned} \|Q_1\chi\|_{k-1} &\leq Ch_k\|(L_k - \hat{A}_k)L_{k-1}^{-1}\chi\|_{H^1} \leq Ch_k^\alpha(1 + |\ln h_k|)\|L_kL_{k-1}^{-1}\chi\|_k \\ &\leq Ch_k^{2\alpha-1}(1 + |\ln h_k|)\|\chi\|_{k-1}. \end{aligned}$$

For Q_2 , we have

$$\begin{aligned} |(Q_2\chi, \theta)_{k-1}| &= |(\hat{A}_kL_{k-1}^{-1}\chi, (I - I_{k-1})(L_k - \hat{A}_k)\tilde{L}_{k-1}^{-1}\theta)_k| \\ &\leq \|\hat{A}_kL_{k-1}^{-1}\chi\|_k\|(I - I_{k-1})(L_k - \hat{A}_k)\hat{A}_k^{-1}\|_k\|\hat{A}_k\tilde{L}_{k-1}^{-1}\theta\|_k. \end{aligned}$$

Thus, applying Lemmas 6.1, 6.4, 6.5, and 6.6 gives

$$\|Q_2\|_{k-1} \leq Ch_k^{4\alpha-3}(1 + |\ln h_k|).$$

For Q_3 , we obviously have

$$\|Q_3\|_{k-1} \leq \|Q_{3,1}\|_{k-1} + \|Q_{3,2}\|_{k-1} + \|Q_{3,3}\|_{k-1},$$

where

$$\begin{aligned} Q_{3,1} &= \tilde{L}_{k-1}^{-1}\hat{A}_{k-1}(I_{k-1} - I)D_kL_{k-1}^{-1}, \\ Q_{3,2} &= \tilde{L}_{k-1}^{-1}(D_{k-1}^*\hat{A}_{k-1} - I_{k-1}D_k^*\hat{A}_k)L_{k-1}^{-1}, \\ Q_{3,3} &= \tilde{L}_{k-1}^{-1}(D_{k-1}^*D_{k-1} - I_{k-1}D_k^*D_k)L_{k-1}^{-1}. \end{aligned}$$

For $Q_{3,1}$, by (6.16), (6.22), and Lemma 6.5,

$$\begin{aligned} \|Q_{3,1}\chi\|_{k-1} &\leq Ch_k\|\tilde{L}_{k-1}^{-1}\hat{A}_{k-1}\|_{k-1}\|D_kL_{k-1}^{-1}\chi\|_{H^1} \\ &\leq Ch^{3\alpha-2}\|\chi\|_{k-1}. \end{aligned}$$

For $Q_{3,2}$,

$$\begin{aligned} |(Q_{3,2}\chi, \theta)_{k-1}| &= |(\hat{A}_kL_{k-1}^{-1}\chi, (I - I_{k-1})D_k\tilde{L}_{k-1}^{-1}\theta)_k| \\ &\leq \|\hat{A}_kL_{k-1}^{-1}\chi\|_k\|(I - I_{k-1})D_k\hat{A}_k^{-1}\|_k\|\hat{A}_k\tilde{L}_{k-1}^{-1}\theta\|_k. \end{aligned}$$

Applying (6.16) and Lemmas 6.5 and 6.6 gives

$$\|Q_{3,2}\|_{k-1} \leq Ch_k^{4\alpha-3}.$$

Finally, for $Q_{3,3}$,

$$\begin{aligned} |(Q_{3,3}\chi, \theta)_{k-1}| &= |(D_kL_{k-1}^{-1}\chi, (I - I_{k-1})D_k\tilde{L}_{k-1}^{-1}\theta)_k| \\ &\leq \|D_kL_{k-1}^{-1}\|_k\|L_kL_{k-1}^{-1}\chi\|_k\|(I - I_{k-1})D_k\hat{A}_k^{-1}\|_k\|\hat{A}_k\tilde{L}_{k-1}^{-1}\theta\|_k. \end{aligned}$$

Applying (6.15), (6.16), and Lemmas 6.1, 6.5 and 6.6 gives

$$\|Q_{3,3}\|_{k-1} \leq Ch_k^{4\alpha-3}.$$

Combining the above inequalities proves Lemma 5.3.

7. Numerical Results. In this section, we give the results of numerical experiments involving the multigrid algorithms. These model computations show that the assumption $h_j \leq \nu$ is necessary for convergence in practice. In contrast, the degradation of the convergence rate for $\alpha = 1$ suggested by Theorem 2 was not observed in the reported computations.

In our numerical examples, we consider the symmetric and indefinite problem

$$(7.1) \quad \begin{aligned} -\mu u - \Delta u &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega, \end{aligned}$$

where Ω is the unit square in R^2 . No examples for the nonsymmetric problem are given.

The eigenvalues for the operator of (7.1) are $(j^2 + k^2)\pi^2 - \mu$ where j, k are positive integers. We will consider the cases $\mu = 30$ and $\mu = 65$. The case $\mu = 30$ has only one negative eigenvalue. When $\mu = 65$, there are two negative eigenvalues, one of which is of multiplicity two.

To triangulate Ω , we first partition it into a regular rectangular mesh and then split each rectangle into two triangles (see Figure 7.1). We use the continuous piecewise linear finite element subspace on the resulting triangulations described in Section 3 and use the discrete inner products given by (3.6). For the purpose of this computation, we deviate from the finite element approximation in that the lower-order term in (7.1) is approximated by an appropriately weighted diagonal term. This is the so-called ‘lumped mass’ finite difference operator. With this discretization it is possible to actually compute the action of L_k and its inverse.

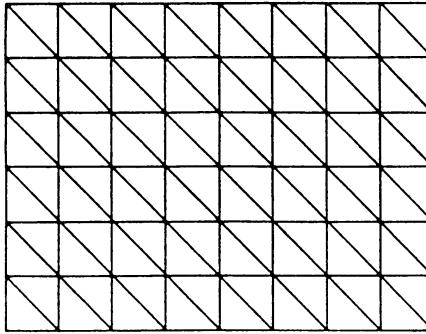


FIGURE 7.1
The regular triangular mesh defining \mathcal{M}_1 .

For these examples, it is computationally feasible to actually compute the best possible δ_k satisfying (4.4). Note that (4.4) is equivalent to the inequality,

$$((I - A_k B_k^*) L_k (I - B_k A_k) v, v)_k \leq \delta_k (L_k v, v)_k \quad \text{for all } v \in \mathcal{M}_k.$$

Thus, the best value of δ_k equals the largest eigenvalue of the operator

$$(7.2) \quad \mathcal{C} = L_k^{-1} (I - A_k B_k^*) L_k (I - B_k A_k).$$

The largest eigenvalue of \mathcal{C} can then be computed by, for example, the power method, if routines for computing the action of \mathcal{C} are available. We obviously know how to compute A_k and B_k . For the constant coefficient problem on a rectangular domain with a regular mesh, the operator L_k and its inverse can be efficiently computed by use of the Fast Fourier Transform. We are left to compute B_k^* .

For a symmetric problem, the operator B_k^* is also a multigrid operator and is given by the following algorithm [18]:

Algorithm for computing B_k^* . Let $B_j^* : \mathcal{M}_j \mapsto \mathcal{M}_j$ denote the adjoint of A_j^{-1} . Assume that $B_{k-1}^* : \mathcal{M}_{k-1} \mapsto \mathcal{M}_{k-1}$ has been defined and define B_k^*g for $g \in \mathcal{M}_k$ as follows:

(I) Set $q^0 = 0$.

(II) Define $x^0 = q^p$ where q^i , for $i = 1, 2, \dots, p$, is defined by

$$q^i = q^{i-1} + B_{k-1}^*[I_{k-1}g - A_{k-1}q^{i-1}].$$

(III) For $l = 1, \dots, m(k)$, define

$$x^l = x^{l-1} + \mu_k^{-2} A_k^*(g - A_k x^{l-1}).$$

Arguments similar to those leading to (4.3) imply that the operator B_k^* defined by (I)–(III) satisfies the equation

$$(7.3) \quad I - B_k^* A_k = K_k^{m(k)} [(I - P_{k-1}) + (I - B_{k-1}^* A_{k-1})^p P_{k-1}].$$

A straightforward mathematical induction argument using (7.3) and the symmetry of A_k implies that the operator defined by (I)–(III) is the adjoint of B_k .

Table 7.1 gives the computed largest eigenvalue for the discrete system (7.2). We vary the mesh size on the finest grid and use $h_j = 1/8$ for the coarsest grid. We give the convergence parameter δ_j as a function of $h_J = 2^{-2-J}$ for $J = 2, 3, 4, 5$ for the variable \mathcal{V} -cycle algorithm ($\beta_0 = \beta_1 = 2$), the standard \mathcal{V} -cycle algorithm ($m(k) = 1$) and the \mathcal{W} -cycle algorithm ($m(k) = 1$). In the variable \mathcal{V} -cycle case, we use $m(J) = 1$.

The results of Table 7.1 illustrate that the multigrid process can be used to develop convergent iterative algorithms for the solution of the equations on the finest grid level. The rate of iterative convergence for these algorithms appears to be bounded independently of the number of grid levels as suggested by the theory.

TABLE 7.1
 δ_k for ‘symmetric’ multigrid schemes with $h_j = 1/8$
applied to (7.1) with $\mu = 30$.

h_J	Var \mathcal{V} -cycle	\mathcal{V} -cycle	\mathcal{W} -cycle
1/16	.88	.88	.88
1/32	.88	.90	.88
1/64	.88	.90	.88
1/128	.88	.90	.88

The next table illustrates the importance of satisfying the assumption $h_j \leq \nu$. For this example, we again consider (7.1) with $\mu = 30$ but use $h_j = 1/4$. Values of δ_J greater than one indicate instability of the multigrid scheme. Note that only the \mathcal{W} -cycle examples with $h_J < 1/16$ and the variable \mathcal{V} -cycle example with $h_J = 1/16$ were stable. It should not be inferred from these results that the \mathcal{W} -cycle is generally more stable than the \mathcal{V} -cycle algorithms. Later examples will show that it shares the same type of stability problems.

TABLE 7.2
 δ_k for ‘symmetric’ multigrid schemes with $h_j = 1/4$
 applied to (7.1) with $\mu = 30$.

h_J	Var \mathcal{V} -cycle	\mathcal{V} -cycle	\mathcal{W} -cycle
1/16	.93	1.06	1.02
1/32	1.09	1.07	.88
1/64	1.09	1.07	.88
1/128	1.08	1.07	.88

The next table illustrates how the convergence rate of the multigrid schemes depends on the number of smoothings used. Table 7.3 gives δ_J as a function of $m(J)$, the number of smoothings on the finest grid level. In this example, $h_J = 1/128$ and $h_j = 1/8$ and we again use $\beta_0 = \beta_1 = 2$ in the variable \mathcal{V} -cycle scheme. The theory developed earlier indicates that, for stability, ν can be chosen independently of $m(J)$. This is consistent with the numerical results which remain stable without the use of smaller h_j as $m(J)$ increases. In contrast, the ‘nonsymmetric’ scheme requires the use of smaller h_j as the number of smoothings increases [1], [14].

TABLE 7.3
 δ_k for ‘symmetric’ multigrid schemes applied to (7.1)
 with $\mu = 30$, $h_j = 1/8$ and $h_J = 1/128$.

$m(J)$	Var \mathcal{V} -cycle	\mathcal{V} -cycle	\mathcal{W} -cycle
1	.88	.90	.88
3	.68	.74	.68
5	.52	.64	.52
7	.43	.56	.43
9	.39	.54	.39

Table 7.3 also shows that the rate of convergence δ_J decreases with larger $m(J)$ as theoretically predicted.

For the final example, we consider $\mu = 65$. In this case, we had to use $h_j = 1/16$ to get a stable algorithm. The computed values of δ_J for $1/32 \leq h_J \leq 1/128$ for the variable \mathcal{V} -cycle, the \mathcal{V} -cycle, and \mathcal{W} -cycle algorithms were approximately .88, .9 and .88, respectively. These results, as well as those given in Table 7.1, do not exhibit the convergence degradation for the \mathcal{V} -cycle algorithm suggested by Theorem 2.

Table 7.4 gives computed values of δ_J when $h_j = 1/4$ was used. These results again illustrate the importance of the theoretical assumption $h_j \leq \nu$. Note that a value of δ_j of a thousand implies that two steps of multigrid will amplify certain frequencies of the error by a factor of a thousand. Such an amplification leads to a rapidly divergent numerical scheme. This example also illustrates that the \mathcal{W} -cycle algorithm displays the same type of stability problems as the \mathcal{V} -cycle algorithms. In fact, the \mathcal{W} -cycle schemes were so unstable at smaller h_J , that it was impossible to compute the corresponding values of δ_J due to computer exponential overflow.

TABLE 7.4
 δ_k for ‘symmetric’ multigrid schemes applied to (7.1)
with $\mu = 65$ and $h_j = 1/4$.

h_J	Var \mathcal{V} -cycle	\mathcal{V} -cycle	\mathcal{W} -cycle
1/16	956	1060	8.0×10^5
1/32	826	1115	6.5×10^{11}
1/64	634	1121	*
1/128	484	1120	*

When the above examples converge, we observe almost identical results for the variable \mathcal{V} -cycle and the \mathcal{W} -cycle algorithms. Note that both algorithms have the same number of smoothing iterations on the various grid levels. For these examples, the extra grid transfer involved in the \mathcal{W} -cycle algorithm does not seem to yield a faster convergence rate.

Department of Mathematics
Cornell University
Ithaca, New York 14853
E-mail: bramble@mathvax.msi.cornell.edu

Brookhaven National Laboratory
Upton, New York 11973
E-mail: pasciak@bnlux0.bnl.gov

Department of Mathematics
Cornell University
Ithaca, New York 14853

1. R. E. BANK, “A comparison of two multilevel iterative methods for nonsymmetric and indefinite elliptic finite element equations,” *SIAM J. Numer. Anal.*, v. 18, 1981, pp. 724–743.
2. R. E. BANK & C. C. DOUGLAS, “Sharp estimates for multigrid rates of convergence with general smoothing and acceleration,” *SIAM J. Numer. Anal.*, v. 22, 1985, pp. 617–633.
3. R. E. BANK & T. DUPONT, “An optimal order process for solving elliptic finite element equations,” *Math. Comp.*, v. 36, 1981, pp. 35–51.
4. D. BRAESS & W. HACKBUSCH, “A new convergence proof for the multigrid method including the V -cycle,” *SIAM J. Numer. Anal.*, v. 20, 1983, pp. 967–975.
5. J. H. BRAMBLE & J. E. PASCIAK, “New convergence estimates for multigrid algorithms,” *Math. Comp.*, v. 49, 1987, pp. 311–329.
6. A. BRANDT, “Multi-level adaptive solutions to boundary-value problems,” *Math. Comp.*, v. 31, 1977, pp. 333–390.
7. P. GRISVARD, “Behavior of the solutions of an elliptic boundary value problem in a polygonal or polyhedral domain,” in *Numerical Solution of Partial Differential Equations*, III (B. Hubbard, ed.), Academic Press, New York, 1976, pp. 207–274.
8. W. HACKBUSCH, *Multi-Grid Methods and Applications*, Springer-Verlag, New York, 1985.
9. T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1976.
10. V. A. KONDRAT'EV, “Boundary problems for elliptic equations with conical or angular points,” *Trans. Moscow Math. Soc.*, v. 16, 1967, pp. 227–313.
11. S. G. KREIN & Y. I. PETUNIN, *Scales of Banach spaces*, Russian Math. Surveys, vol. 21, 1966, pp. 85–160.
12. J. L. LIONS & E. MAGENES, *Problèmes aux Limites non Homogènes et Applications*, Dunod, Paris, 1968.

13. J. F. MAITRE & F. MUSY, "Algebraic formalization of the multigrid method in the symmetric and positive definite case—A convergence estimation for the V -cycle," in *Multigrid Methods for Integral and Differential Equations* (D. J. Paddon and H. Holstein, eds.), Clarendon Press, Oxford, 1985.
14. J. MANDEL, "Multigrid convergence for nonsymmetric, indefinite variational problems and one smoothing step," in *Proc. Copper Mtn. Conf. Multigrid Methods*, Appl. Math. Comput., 1986, pp. 201–216.
15. J. MANDEL, *Algebraic Study of Multigrid Methods for Symmetric, Definite Problems*. (Preprint.)
16. J. MANDEL, S. F. MCCORMICK & J. RUGE, *An Algebraic Theory for Multigrid Methods for Variational Problems*. (Preprint.)
17. S. F. MCCORMICK, "Multigrid methods for variational problems: Further results," *SIAM J. Numer. Anal.*, v. 21, 1984, pp. 255–263.
18. S. F. MCCORMICK, "Multigrid methods for variational problems: General theory for the V -cycle," *SIAM J. Numer. Anal.*, v. 22, 1985, pp. 634–643.
19. J. NEČAS, *Les Méthodes Directes en Théorie des Équations Elliptiques*, Academia, Prague, 1967.
20. A. H. SCHATZ, "An observation concerning Ritz-Galerkin methods with indefinite bilinear forms," *Math. Comp.*, v. 28, 1974, pp. 959–962.
21. H. YSERENTANT, "The convergence of multi-level methods for solving finite-element equations in the presence of singularities," *Math. Comp.*, v. 47, 1986, pp. 399–409.

3.3 Parallel multilevel preconditioners

New convergence estimates for multigrid algorithms[26]

PARALLEL MULTILEVEL PRECONDITIONERS

JAMES H. BRAMBLE, JOSEPH E. PASCIAK, AND JINCHAO XU

ABSTRACT. In this paper, we provide techniques for the development and analysis of parallel multilevel preconditioners for the discrete systems which arise in numerical approximation of symmetric elliptic boundary value problems. These preconditioners are defined as a sum of independent operators on a sequence of nested subspaces of the full approximation space. On a parallel computer, the evaluation of these operators and hence of the preconditioner on a given function can be computed concurrently.

We shall study this new technique for developing preconditioners first in an abstract setting, next by considering applications to second-order elliptic problems, and finally by providing numerically computed condition numbers for the resulting preconditioned systems. The abstract theory gives estimates on the condition number in terms of three assumptions. These assumptions can be verified for quasi-uniform as well as refined meshes in any number of dimensions. Numerical results for the condition number of the preconditioned systems are provided for the new algorithms and compared with other well-known multilevel approaches.

1. INTRODUCTION

We shall provide some new techniques for the development and analysis of preconditioners for the discrete systems which arise in approximation to the solutions of elliptic boundary value problems. It has been demonstrated that preconditioned iteration techniques often lead to the most computationally effective algorithms for the solution of the large algebraic systems corresponding to boundary value problems in two- and three-dimensional Euclidean space (cf. [3] and the included references). The use of preconditioned iteration will become even more important on computers with parallel architecture.

This paper provides an approach for developing completely parallel multilevel preconditioners. In order to illustrate the resulting algorithms, we shall describe the simplest application of the technique to a model elliptic problem. Let Ω be a polygonal domain in R^2 and consider the problem of approximating the

Received February 13, 1989.

1980 *Mathematics Subject Classification* (1985 Revision). Primary 65N30; Secondary 65F10.

This manuscript has been authored under contract number DE-AC02-76CH00016 with the U.S. Department of Energy. Accordingly, the U.S. Government retains a non-exclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. This work was also supported in part under the National Science Foundation Grant No. DMS84-05352 and by the U.S. Army Research Office through the Mathematical Science Institute, Cornell University.

©1990 American Mathematical Society
0025-5718/90 \$1.00 + \$.25 per page

solution u of

$$(1.1) \quad \begin{aligned} Lu &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega, \end{aligned}$$

where

$$Lu = - \sum_{i,j=1}^2 \frac{\partial}{\partial x_i} a_{ij} \frac{\partial u}{\partial x_j} + au.$$

We assume that the matrix $\{a_{ij}(x)\}$ is symmetric and uniformly positive definite and $a(x) \geq 0$ in Ω .

We first define a sequence of multilevel finite element spaces in the usual way. Since Ω is polygonal, we can define a ‘coarse’ triangulation $\tau_1 = \bigcup_l \tau_1^l$, where τ_1^l represents an individual triangle and τ_1 denotes the triangulation. Successively finer triangulations $\{\tau_k, k = 2, \dots, J\}$ are defined by breaking each triangle of a coarser triangulation into four triangles by connecting the midpoints of the edges. The subspace \mathcal{M}_k is defined to be the continuous functions defined on Ω which are piecewise linear with respect to τ_k and vanish on $\partial\Omega$. We shall be interested in developing a preconditioner for the solution of the Galerkin equations on the J th subspace, i.e., $U \in \mathcal{M}_J$ satisfying

$$(1.2) \quad A(U, \phi) = (f, \phi) \quad \text{for all } \phi \in \mathcal{M}_J.$$

Here $A(\cdot, \cdot)$ denotes the generalized Dirichlet integral defined by

$$(1.3) \quad A(u, v) = \sum_{i,j=1}^2 \int_{\Omega} a_{ij} \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_j} dx + \int_{\Omega} a u v dx$$

and (\cdot, \cdot) denotes the L^2 inner product on Ω .

Let $\{\phi_k^l\}$ denote the usual nodal basis for the subspace \mathcal{M}_k , i.e., the l th basis function is one on the l th node of the k th triangulation and vanishes on all others. The preconditioner \mathcal{B} is defined by

$$(1.4) \quad \mathcal{B}v = \sum_{k=1}^J \sum_l (v, \phi_k^l) \phi_k^l.$$

The above preconditioner is simply a double sum, the terms of which can be computed concurrently. This results in an inherently parallel algorithm.

As is well-known, the rate of convergence of an iterative method can be estimated in terms of the condition number of the preconditioned system. We provide a theory for the estimation of the condition number for this type of multilevel preconditioner in terms of a number of a priori assumptions. In the above example, this theory can be used to show that the relevant condition number is at worst $O(J^2)$. Moreover, these results hold for problems in two, three, and higher dimensions as well as problems with only locally quasi-uniform mesh approximation.

We note that many alternative preconditioning techniques have been proposed for such discrete systems. For example, domain decomposition preconditioners have been developed ([5], [6], [7], [8], [13], and the included references). These domain decomposition preconditioners are inherently parallel, however become somewhat complex in three-dimensional applications. Alternatively, multigrid [4], [9], [14], [17] and hierarchical multigrid [2], [20] techniques give rise to different multilevel preconditioners. The standard multigrid algorithms do not allow for completely parallel computations, since the computations on a given level use results from the previous levels. Theoretical results for the usual multigrid algorithms are available, in general, for problems in any number of spatial dimensions but only for quasi-uniform mesh approximation. Good results hold for the hierarchical basis method in two dimensions with refined meshes but degenerate when applied to three-dimensional problems. Finally, preconditioners based on approximate LU factorization are often proposed; however, a comprehensive theory is yet to be developed [11], [12], [18].

The outline of the remainder of the paper is as follows. A general abstract theory for the development and analysis of parallel multilevel preconditioners is given in §2. In §3, this theory is applied to second-order elliptic boundary value problems, and the serial and parallel complexity of the resulting algorithms is discussed. We apply the abstract theory to a second-order problem with a locally refined mesh in §4. Finally, the results of numerical experiments illustrating the theory of the earlier sections are given in §5.

2. GENERAL THEORY

In this section, we develop a general theory for the construction of parallel multilevel preconditioners. This theory is presented in an abstract setting to most clearly illustrate the relevant analytic techniques and assumptions. The development of this class of preconditioners is based on a certain orthogonal decomposition of the approximation space. The parallel multilevel preconditioners are then abstractly defined in terms of this decomposition by the replacement of orthogonal projections by more computationally efficient operators. Applications to second-order elliptic boundary value problems are given in §§3 and 4.

We start with the basic abstract framework. We assume that we are given a nested sequence of finite-dimensional spaces,

$$(2.1) \quad \mathcal{M}_1 \subset \mathcal{M}_2 \subset \cdots \subset \mathcal{M}_J \equiv \mathcal{M}, \quad J \geq 2.$$

The space \mathcal{M} and hence all of its subspaces are equipped with two inner products (\cdot, \cdot) and $A(\cdot, \cdot)$. The first part of this section will consider properties of a certain orthogonal decomposition of \mathcal{M} with respect to the inner product (\cdot, \cdot) and the sequence of spaces (2.1). We shall investigate the spectral properties of these spaces with respect to the form $A(\cdot, \cdot)$ since, ultimately, we are interested in computing the solution to the Galerkin equations: Given $f \in \mathcal{M}$,

find $u \in \mathcal{M}$ satisfying

$$(2.2) \quad A(u, v) = (f, v) \quad \text{for all } v \in \mathcal{M}.$$

We shall use the following notation in the development and analysis. For each $k = 1, \dots, J$, we introduce the following operators:

(1) The projection $P_k : \mathcal{M} \rightarrow \mathcal{M}_k$ is defined for $u \in \mathcal{M}$ by

$$A(P_k u, v) = A(u, v) \quad \text{for all } v \in \mathcal{M}_k.$$

(2) The projection $Q_k : \mathcal{M} \rightarrow \mathcal{M}_k$ is defined for $u \in \mathcal{M}$ by

$$(Q_k u, v) = (u, v) \quad \text{for all } v \in \mathcal{M}_k.$$

(3) The operator $A_k : \mathcal{M}_k \rightarrow \mathcal{M}_k$ is defined for $u \in \mathcal{M}_k$ by

$$(A_k u, v) = A(u, v) \quad \text{for all } v \in \mathcal{M}_k.$$

We shall also denote $A = A_J$ and define

$$\mathcal{O}_k = \{\phi \mid \phi = (Q_k - Q_{k-1})\psi, \psi \in \mathcal{M}\},$$

where $Q_0 = 0$. We shall study the spectral properties of A with respect to the decomposition

$$(2.3) \quad \mathcal{M} = \mathcal{O}_1 + \cdots + \mathcal{O}_J.$$

It follows from the above definitions that

$$(2.4) \quad \begin{aligned} Q_k A &= A_k P_k, \\ Q_k Q_l &= Q_l Q_k = Q_l \quad \text{for } l \leq k. \end{aligned}$$

From the second equation of (2.4), it follows that

$$(Q_k - Q_{k-1})(Q_l - Q_{l-1}) = 0$$

if $k \neq l$, and hence the decomposition (2.3) is orthogonal, i.e., $(u, v) = 0$ whenever $u \in \mathcal{O}_l$, $v \in \mathcal{O}_k$ with $l \neq k$.

We consider first the operator

$$(2.5) \quad B = \sum_{k=1}^J \lambda_k^{-1} (Q_k - Q_{k-1}),$$

where λ_k denotes the spectral radius of A_k . Clearly, B is symmetric and positive definite, and

$$(2.6) \quad A(BAv, v) = \sum_{k=1}^J \lambda_k^{-1} \left\| (Q_k - Q_{k-1})Av \right\|^2,$$

where $\|\cdot\|^2 = (\cdot, \cdot)$. Note that B is block diagonal with respect to the decomposition (2.3) and each diagonal block is a multiple of the identity matrix.

The operator B may be thought of as an “approximate inverse” for A . Thus, we shall be interested in estimating the condition number $K(BA)$ of BA . We note that $K(BA) \leq c_1/c_0$ for any positive constants c_0, c_1 satisfying

$$(2.7) \quad c_0 A(v, v) \leq A(BAv, v) \leq c_1 A(v, v) \quad \text{for all } v \in \mathcal{M}.$$

Remark 2.1. The form of the operator B can be motivated by the spectral decomposition of the operator A . Indeed, for a special example, namely, \mathcal{M}_k the space spanned by the eigenvectors corresponding to the smallest k distinct eigenvalues of A , the operator B defined by (2.5) is in fact equal to A^{-1} .

In general, we can see that

$$(2.8) \quad A(BAv, v) \leq JA(v, v) \quad \text{for all } v \in \mathcal{M}.$$

Indeed, by (2.4), (2.6) and the definition of λ_k ,

$$A(BAv, v) \leq \sum_{k=1}^J \lambda_k^{-1} \|Q_k Av\|^2 \leq \sum_{k=1}^J A(P_k v, P_k v).$$

Inequality (2.8) follows from the fact that P_k is a bounded operator with operator norm one in the norm induced by the $A(\cdot, \cdot)$ inner product.

The lower estimate of (2.7) will require some hypotheses concerning the spaces \mathcal{M}_k . We first consider the following assumptions on the operators Q_k : For $k = 1, \dots, J$, there exists a constant $C_1 > 0$ such that

$$(A.1) \quad \|(I - Q_{k-1})v\|^2 \leq C_1 \lambda_k^{-1} A(v, v) \quad \text{for all } v \in \mathcal{M}.$$

We can now prove the following theorem.

Theorem 1. *Assume that (A.1) holds. Then*

$$(2.9) \quad C_1^{-1} J^{-1} A(v, v) \leq A(BAv, v) \leq JA(v, v) \quad \text{for all } v \in \mathcal{M}.$$

Proof. By (2.8), we need only prove the first inequality of (2.9). We note that

$$\begin{aligned} A(v, v) &= \sum_{k=1}^J A((Q_k - Q_{k-1})v, v) \\ &= \sum_{k=1}^J ((I - Q_{k-1})v, (Q_k - Q_{k-1})Av). \end{aligned}$$

By the Schwarz inequality and (A.1), it follows that

$$A(v, v) \leq C_1^{1/2} \sum_{k=1}^J A(v, v)^{1/2} \lambda_k^{-1/2} \|(Q_k - Q_{k-1})Av\|.$$

Applying the Schwarz inequality to the sum gives

$$A(v, v)^{1/2} \leq C_1^{1/2} J^{1/2} A(BAv, v)^{1/2},$$

which is the lower inequality of (2.9).

Corollary 1. *For any real s ,*

$$(2.10) \quad B^s = \sum_{k=1}^J \lambda_k^{-s} (Q_k - Q_{k-1}).$$

Moreover, for any $s \in [0, 1]$,

$$(2.11) \quad J^{-s}(A^s v, v) \leq (B^{-s} v, v) \leq (C_1 J)^s(A^s v, v) \quad \text{for all } v \in \mathcal{M}.$$

Proof. The orthogonality of the decomposition (2.3) immediately implies (2.10). Clearly, B^s and A^s are Hilbert scales. By interpolation, it suffices to verify (2.11) for $s = 0$ and $s = 1$. The case $s = 0$ is trivial and $s = 1$ is given by Theorem 1.

We have included Corollary 1 for the purpose of future applications which will not be described in this paper. In particular, it will be used for the development of preconditioners for certain boundary operators which arise in domain decomposition techniques for second-order boundary value problems [10].

In the next corollary, we consider the case of the sum of two operators. Let $\widehat{A}(\cdot, \cdot)$ be another symmetric positive definite form and let \widehat{A} , $\{\widehat{A}_k\}$ and $\{\widehat{\lambda}_k\}$ be defined analogously in terms of $\widehat{A}(\cdot, \cdot)$. Consider the operator $\overline{B} : \mathcal{M} \mapsto \mathcal{M}$ defined by

$$\overline{B} = \sum_{k=1}^J (\lambda_k + \widehat{\lambda}_k)^{-1} (Q_k - Q_{k-1}).$$

Theorem 1 immediately implies the following corollary.

Corollary 2. *Assume that (A.1) holds for both A and \widehat{A} . Then,*

$$J^{-1}((A + \widehat{A})v, v) \leq (\overline{B}^{-1}v, v) \leq C_1 J((A + \widehat{A})v, v) \quad \text{for all } v \in \mathcal{M}.$$

Proof. A change of variable shows that (2.9) is equivalent to

$$J^{-1}(Av, v) \leq (B^{-1}v, v) \leq C_1 J(Av, v) \quad \text{for all } v \in \mathcal{M}.$$

Corollary 2 follows adding this and the analogous inequality involving \widehat{A} .

The most natural application of the above corollary is to the discrete systems which arise in parabolic time-stepping algorithms. At each time level, a function $U^n \in \mathcal{M}$ satisfying

$$(I + \tau A)U^n = F^n,$$

with known $F^n \in \mathcal{M}$ must be computed. Here τ is a positive number which is related to the time step size. We shall not consider further application of Corollary 2 in this paper.

We next apply the above results to analyze parallel multilevel preconditioners for A . An operator $\mathcal{B} : \mathcal{M} \mapsto \mathcal{M}$ is a good preconditioner for A if it satisfies:

- (1) The action of \mathcal{B} on vectors of \mathcal{M} is economical to compute.
- (2) The condition number $K(\mathcal{B}A)$ of the preconditioned system is not too large.

Item (1) above guarantees that the cost per iteration in a preconditioned scheme using \mathcal{B} for solving (2.2) will not be unreasonable. Item (2) guarantees that the number of iterations in a preconditioned scheme will not be too large. Note that by Theorem 1, B satisfies (2). \mathcal{B} may in fact satisfy (1) in many applications

but generally it is desirable to avoid evaluating the action of Q_k . Hence we shall develop more computationally effective algorithms by modifying (2.5).

To get a computationally effective preconditioner, we write (2.5) in the form

$$B = \sum_{k=1}^{J-1} (\lambda_k^{-1} - \lambda_{k+1}^{-1}) Q_k + \lambda_J^{-1} I.$$

Notice that if $\{\lambda_k\}_{k=1}^J$ satisfies the growth condition $\lambda_{k+1} \geq \sigma \lambda_k$ for $\sigma > 1$, then the operator

$$(2.12) \quad \hat{B} = \sum_{k=1}^J \lambda_k^{-1} Q_k$$

satisfies

$$(1 - \sigma^{-1})(\hat{B}u, u) \leq (Bu, u) \leq (\hat{B}u, u) \quad \text{for all } u \in \mathcal{M}.$$

We consider a slightly more general operator defined by replacing $\lambda_k^{-1} I$ in (2.12) with a symmetric positive definite operator $R_k : \mathcal{M}_k \mapsto \mathcal{M}_k$, i.e.,

$$(2.13) \quad \mathcal{B} = \sum_{k=1}^J R_k Q_k.$$

Clearly, \mathcal{B} is symmetric and positive definite on \mathcal{M} . The cost of evaluating the action of the preconditioner \mathcal{B} on a vector in \mathcal{M} will be discussed in later sections but will obviously depend on an appropriate choice of R_k .

For our subsequent analysis, we shall need to make the following assumption concerning the operator R_k . We assume that

$$(A.2) \quad C_2 \frac{\|u\|^2}{\lambda_k} \leq (R_k u, u) \leq C_3 (A_k^{-1} u, u) \quad \text{for all } u \in \mathcal{M}_k,$$

where C_2 and C_3 are positive constants not depending on J . Clearly, the choice $R_k = \lambda_k^{-1} I$ corresponding to (2.12) satisfies (A.2).

The preconditioner (2.13) can be thought of as a parallel version of a V-cycle multigrid algorithm. The operator R_k plays the role of a smoothing procedure. The major difference between (2.13) and the V-cycle multigrid scheme is that the smoothing on every level of (2.13) is applied to the original fine grid residual. In contrast, the multigrid V-cycle applies the smoothing to the residual computed using the corrections from the previously visited grid. Obviously, the different terms in (2.13) can be computed in parallel while, in contrast, computations on a given grid level in a standard multigrid algorithm must wait for the results from previous levels. The connection between (2.13) and the multigrid V-cycle will be more fully discussed in §3. However, it is not surprising that assumptions which are equivalent to (A.2) have been made in the analysis of the usual serial multigrid algorithms [4], [9], [15], [16].

Remark 2.2. A particularly interesting choice of R_k can be motivated as follows. As noted above, $R_k = \lambda_k^{-1} I$ satisfies (A.2). Let $\{\psi_k^l\}$ be an orthonormal

basis for \mathcal{M}_k . Then

$$(2.14) \quad \lambda_k^{-1} u = \sum_l (u, \psi_k^l) \psi_k^l \quad \text{for all } u \in \mathcal{M}_k.$$

In practice, an orthonormal basis for \mathcal{M}_k is seldom available. However, for finite element applications with quasi-uniform grids, the right-hand side of (2.14) with normalized nodal basis functions $\{\bar{\psi}_k^l\}$ defines an R_k satisfying (A.2) (see §3). Moreover, we note that for $u \in \mathcal{M}$,

$$R_k Q_k u = \lambda_k^{-1} \sum_l (u, \bar{\psi}_k^l) \bar{\psi}_k^l,$$

and hence $R_k Q_k$ is computable without the solution of Gram matrix systems. This will be discussed in more detail in §3.

With \mathcal{B} defined in (2.13), we have the following corollary.

Corollary 3. *Under assumptions (A.1) and (A.2),*

$$(2.15) \quad C_1^{-1} C_2 J^{-1} A(v, v) \leq A(\mathcal{B} Av, v) \leq C_3 J A(v, v) \quad \text{for all } v \in \mathcal{M}.$$

Proof. By (A.2), for $v \in \mathcal{M}$,

$$(2.16) \quad A(\mathcal{B} Av, v) = \sum_{k=1}^J (R_k A_k P_k v, A_k P_k v) \leq C_3 \sum_{k=1}^J A(P_k v, P_k v),$$

from which the second inequality of (2.15) follows. For the first inequality, by Theorem 1 and (A.2),

$$\begin{aligned} C_1^{-1} J^{-1} A(v, v) &\leq A(BAv, v) \\ &\leq \sum_{k=1}^J \lambda_k^{-1} \|Q_k Av\|^2 \leq C_2^{-1} A(\mathcal{B} Av, v). \end{aligned}$$

This completes the proof of the corollary.

We next provide an alternative hypothesis for a lower estimate in (2.15). This is the so-called “regularity and approximation” assumption often used in multigrid analysis (cf. [4], [14], [17]). We assume that for a fixed $\alpha \in (0, 1]$, there exists a positive constant C_4 not depending on $k = 1, \dots, J$ satisfying

$$(A.3) \quad A((I - P_{k-1})v, v) \leq (C_4 \lambda_k^{-1} \|A_k v\|^2)^\alpha A(v, v)^{1-\alpha} \quad \text{for all } v \in \mathcal{M}_k,$$

where $P_0 = 0$. In finite element applications, the above assumption is usually proved by using elliptic regularity for the continuous problem and the approximation properties of the space \mathcal{M}_{k-1} [1], [4]. In such applications, assumption (A.3) may be stronger than (A.1), e.g., when $\alpha = 1$, (A.3) implies (A.1). We can now prove the following theorem.

Theorem 2. Assume that (A.2) and (A.3) hold. Then

$$(2.17) \quad \begin{aligned} C_2 C_4^{-1} J^{1-1/\alpha} A(v, v) &\leq A(\mathcal{B} A v, v) \\ &\leq C_3 J A(v, v) \quad \text{for all } v \in \mathcal{M}. \end{aligned}$$

Proof. We need only prove the first inequality in (2.17). Writing

$$v = \sum_{k=1}^J (P_k - P_{k-1})v,$$

and using the properties of P_k and (A.3), gives

$$\begin{aligned} A(v, v) &= \sum_{k=1}^J A((I - P_{k-1})P_k v, P_k v) \\ &\leq C_4^\alpha \sum_{k=1}^J (\lambda_k^{-1} \|A_k P_k v\|^2)^\alpha A(v, v)^{1-\alpha}. \end{aligned}$$

By (A.2),

$$A(v, v) \leq (C_2^{-1} C_4)^\alpha \sum_{k=1}^J (R_k A_k P_k v, A_k P_k v)^\alpha (Av, v)^{1-\alpha}.$$

By Hölder's inequality, for a sequence of nonnegative numbers $\{b_k\}$, we clearly have

$$\sum_{k=1}^J b_k^\alpha \leq J^{1-\alpha} \left(\sum_{k=1}^J b_k \right)^\alpha,$$

from which it follows that

$$(2.18) \quad \begin{aligned} A(v, v)^\alpha &\leq (C_2^{-1} C_4)^\alpha J^{1-\alpha} \left(\sum_{k=1}^J (R_k A_k P_k v, A_k P_k v) \right)^\alpha \\ &\leq (C_2^{-1} C_4)^\alpha J^{1-\alpha} A(\mathcal{B} A v, v)^\alpha. \end{aligned}$$

The first inequality of (2.17) follows from (2.18) in an obvious manner. This completes the proof of Theorem 2.

Remark 2.3. Included in (A.1) and (A.3) is the implicit assumption that C_1 and C_4 are greater than or equal to $K(A_1)$. In finite element applications, $K(A_1)$ will not be large if the grid size of the coarsest grid is of unit size. However, if a good preconditioner R_j is available for any finer grid, i.e., R_j satisfies in addition

$$(2.19) \quad (R_j^{-1} u, u) \leq C_5 (A_j u, u),$$

then it suffices to use

$$\mathcal{B} = \sum_{k=j}^J R_k Q_k.$$

In such applications, (A.1) or (A.3) need only be satisfied for $k > j$. Note that $R_j = A_j^{-1}$ will be effective provided that the j th grid size is relatively small. Many alternative choices are possible.

3. THE QUASI-UNIFORM APPLICATION

In this section, we shall illustrate the application of the abstract theory and algorithms discussed in the previous section to a second-order elliptic boundary value problem approximated using finite element functions on a quasi-uniform mesh. We first show that the hypotheses of the previous section are satisfied. We also consider the computational complexity of the resulting algorithm in both serial and parallel computing applications. For brevity, we consider only the most basic finite element applications. Many other applications are possible, including examples of elliptic problems in higher dimensions.

Let $\mathcal{M}_1 \subset \cdots \subset \mathcal{M}_J \equiv \mathcal{M}$ be the finite element spaces defined in the introduction subsequent to (1.1), $A(\cdot, \cdot)$ be the generalized Dirichlet form defined in (1.3) and (\cdot, \cdot) be the L^2 inner product on Ω .

We will apply the results of §2 to Problem (1.2) with the above sequence of spaces. Let h_k denote the size of the k th triangulation. It easily follows that there are constants c_0 and c_1 , not depending on k and satisfying

$$(3.1) \quad c_0 h_k^{-2} \leq \lambda_k \leq c_1 h_k^{-2}.$$

Inequality (A.1) with $k > 2$ is well-known. For $k = 1$, we have that

$$\|v\|^2 \leq \Lambda^{-1} A(v, v) \quad \text{for all } v \in \mathcal{M},$$

where Λ is the smallest eigenvalue of A and is obviously bounded away from zero (independently of J). We shall suppose in this application that \mathcal{M}_1 is such that h_1 is proportional to the diameter of Ω , so that $C_1 \geq \lambda_1/\Lambda$, which is not large.

We next consider the operator R_k motivated by Remark 2.2, i.e.,

$$(3.2) \quad R_k v = \sum_l (v, \phi_k^l) \phi_k^l \quad \text{for } v \in \mathcal{M}_k,$$

where the sum is taken over all nodes of τ_k . As observed in Remark 2.2, the action of $R_k Q_k$ can be computed without explicitly computing Q_k . Moreover, using R_k defined by (3.2) in (2.13) leads to the preconditioner of (1.4).

We now show that (A.2) holds for this R_k . Any $u \in \mathcal{M}_k$ may be represented by

$$(3.3) \quad u = \sum_l \alpha_l \phi_k^l,$$

where α_l is the value of u at the l th node of τ_k . Let $\vec{\alpha}$ denote the corresponding vector with entries $\{\alpha_l\}$, G_k denote the matrix with entries $(G_k)_{lm} = (\phi_k^l, \phi_k^m)$ and $\langle \cdot, \cdot \rangle$ denote the Euclidean inner product. Note that

$$(3.4) \quad (R_k u, u) = \sum_l (u, \phi_k^l)^2 = \langle G_k \vec{\alpha}, G_k \vec{\alpha} \rangle.$$

By the quasi-uniformity of τ_k , $h_k^2 \sum_l \alpha_l^2$ is a norm which is equivalent to $\|u\|^2 = \langle G_k \vec{\alpha}, \vec{\alpha} \rangle$. This equivalence is uniform with respect to k . It immediately follows that $\langle G_k \vec{\alpha}, G_k \vec{\alpha} \rangle$ is uniformly equivalent to $h_k^4 \langle \vec{\alpha}, \vec{\alpha} \rangle$. Thus, by (3.1),

$$(3.5) \quad c_0 \frac{\|u\|^2}{\lambda_k} \leq (R_k u, u) \leq c_1 \frac{\|u\|^2}{\lambda_k} \quad \text{for all } u \in \mathcal{M}_k,$$

with c_0 and c_1 independent of k . Assumption (A.2) follows immediately from the definition of λ_k .

For this problem, (A.3) will always be satisfied for some $\alpha \in (0, 1]$ (cf. for example, [1], [4]). The size of α depends on the elliptic regularity of Problem (1.1). Thus, in the case when Ω is a convex polygonal domain and the coefficients defining L are smooth, $\alpha = 1$ and we conclude from Theorem 2 that

$$K(\mathcal{B}A) \leq cJ.$$

In the case of a so-called crack problem (with smooth coefficients), the largest interior angle is 2π and the regularity of (1.1) is such that (A.3) does not hold for $\alpha \geq 1/2$. Hence Corollary 3 yields the better estimate and shows that

$$K(\mathcal{B}A) \leq CJ^2.$$

Remark 3.1. It is possible to apply the theory of §2 to elliptic problems in three or more dimensions. Many examples are possible, and we consider the simplest. In three dimensions, we let the coarse mesh be a union of equally sized cubes. Finer meshes are obtained by breaking each cube of a coarser mesh into eight smaller cubes in the obvious way. The subspaces $\{\mathcal{M}_k\}$ are defined to be the functions on Ω which are continuous and piecewise trilinear with respect to the k th mesh and vanish on $\partial\Omega$. The nodes of these spaces are the vertices of the cubes defining the mesh. We may take

$$(3.6) \quad \mathcal{B}u = \sum_{k=1}^J h_k^{-1} \sum_l (u, \phi_k^l) \phi_k^l,$$

where $\{\phi_k^l\}$ denotes the set of nodal basis functions. We emphasize here again that all the terms in (3.6) are independent and hence may be computed concurrently.

Remark 3.2. Assumption (A.1) is often easier to verify than (A.3). For example, we consider the two-dimensional problem (1.1) when the coefficients of the operator L are discontinuous. If the jumps in the coefficients are only along the lines of the coarse mesh, then it is possible to prove that (A.1) holds with $C_1 \leq CJ$, where the constant C depends on the local variation of the coefficients of L on the coarse grid triangles but not on the magnitude of the jumps across triangles [19]. This leads to a conditioning result of the form

$$K(\mathcal{B}A) \leq CJ^3.$$

The dependence of constant C_4 (in (A.3)) on the size of the jumps is a much more difficult question, since it requires the knowledge of the dependence of the elliptic regularity constants on such jumps.

In the remainder of this section, we consider computational issues involved in implementing the above algorithm in serial and parallel computing architectures. However, before proceeding, we make the following observation. Even though we have defined \mathcal{B} as an operator on \mathcal{M} , in a preconditioned iterative scheme we are only required to compute $\mathcal{B}v$ given the data $W_J^l = (v, \phi_J^l)$. This is because when $v = A_J\theta$, we always compute $\{(A_J\theta, \phi_J^l) = A(\theta, \phi_J^l)\}$ and hence avoid the solution of the Gram matrix problem required for the computation of $A_J\theta$.

We first consider the serial version of the algorithm. Let $v \in \mathcal{M}$ be given and define $W_k^l = (v, \phi_k^l)$. Let W_k denote the vector with entries $(W_k)_l = W_k^l$. We need to compute the action of $\mathcal{B}v$ given W_J . We define W_{k-1} from W_k in a recursive manner. Note that each basis function in \mathcal{M}_{k-1} can be written as a local linear combination of basis functions for \mathcal{M}_k . Thus, each value of W_{k-1}^l can be written as a local linear combination of values of W_k . Moreover, the work involved in computing W_{k-1} from W_k is proportional to the number of unknowns in \mathcal{M}_{k-1} . Consequently, the work involved in computing the vectors $\{W_k\}$, $k = 1, \dots, J$, is bounded by a constant times the number of unknowns in \mathcal{M} . Once the vectors $\{W_k\}$ are known, we are left to compute the representation of $\mathcal{B}v$ in the basis for \mathcal{M} . To do this, we compute the representation of

$$\mathcal{B}_m v \equiv \sum_{k=1}^m \sum_l (v, \phi_k^l) \phi_k^l,$$

in the basis for \mathcal{M}_m , for $m = 1, \dots, J$. The result at $m = J$ is of course the basis representation for $\mathcal{B}v$. For $m = 1$, the representation is already given by W_1 . The representation of $\mathcal{B}_m v$ for $m > 1$ is calculated from that of $\mathcal{B}_{m-1} v$ by interpolating the $\mathcal{B}_{m-1} v$ results (i.e., expanding them in terms of the m th basis) and adding the m th level contribution from W_m . The work of calculating the representation of $\mathcal{B}_m v$, given that for $\mathcal{B}_{m-1} v$, is on the order of the number of unknowns in \mathcal{M}_m , and thus the total work for this algorithm is bounded by a constant times the number of unknowns on the finest grid.

Remark 3.3. The serial implementation of the operator \mathcal{B} is closely related to the multigrid V-cycle algorithm. The step of computing W_{k-1} from W_k in \mathcal{B} is nothing more than the step which “transfers the residuals” from grid level k to $k - 1$ in a multigrid V-cycle algorithm. However, the multigrid algorithm requires extra computation since it must smooth and then compute new residuals on the k th level before transferring. The second step in the serial algorithm for \mathcal{B} is also duplicated in the “coarser to finer interpolation” step in the multigrid V-cycle algorithm. The symmetric multigrid V-cycle requires extra computation since it requires additional smoothing on each grid level.

Thus the serial \mathcal{B} algorithm, in terms of complexity, is similar to a multigrid V-cycle algorithm without smoothing.

We next consider parallel implementation of the preconditioner \mathcal{B} . The execution of (1.4) can obviously be made parallel in many ways by breaking up the terms into various numbers of parallel tasks. The optimal splitting of the sum is clearly dependent on characteristics of the individual parallel computer, for example, memory management considerations, task initialization overhead, the number of parallel processors, etc. We note, however, the simplicity of the form of (1.4) allows for almost complete freedom for parallel splitting.

It is of theoretical interest to consider the algorithm on a shared memory machine with an unlimited number of processors. As above, the implementation $\mathcal{B}v$ involves two steps, the calculation of the coefficients W_k^l and the computation of the representation of $\mathcal{B}v$ in the basis for \mathcal{M} . Each coefficient can be computed independently and involves a linear combination (not necessarily local) of the values of W_j . With enough processors, a linear combination of m numbers can be computed in $\log_2 m$ time. Hence the coefficient vectors $\{W_k\}$ can be computed in $\log_2 N$ time where N is the dimension of \mathcal{M} . Each coefficient of $\mathcal{B}v$ involves a linear combination of $M_n J$ contributions from the J grid levels (here, M_n is the maximum number of neighbors for any given level). Thus, computation of $\mathcal{B}v$ can be done in time bounded by CJ .

4. A LOCAL REFINEMENT APPLICATION

In this section, we shall consider the application of the parallel multilevel algorithm to the finite element equations corresponding to a problem with mesh refinement. Such mesh refinements are necessary for accurate modeling of problems with various types of singular behavior. For simplicity, we shall make no attempt at generality. Instead, we shall illustrate the technique by considering an example from which many obvious generalizations are possible. For this example, the domain Ω will be the unit square and we shall approximate the solution to (1.1). The form $A(\cdot, \cdot)$ and the inner product (\cdot, \cdot) will be as in §3. The sequence of grids which we shall consider will be progressively more refined as we approach the corner (1,1). Such a mesh would be effective if, for example, the function f in (1.1) behaved like a δ -function distribution at the point (1,1).

To define the mesh, we first start with a sequence of subspaces $\mathcal{M}_1, \dots, \mathcal{M}_j$ defined using uniform grids of size $h_k = 2^{-k}$, $k = 1, \dots, j$, as described in the quasi-uniform case (see §1). The $(j+1)$ st triangulation is then defined by refining only those triangles in the upper quarter, $[1/2, 1] \times [1/2, 1]$. Similarly, the $(j+2)$ nd triangulation is defined by refining only those triangles in the $(j+1)$ st grid which are in the region $[3/4, 1] \times [3/4, 1]$, etc. (see Figure 4.1). The spaces \mathcal{M}_k for $k = j+1, \dots, J$ are defined to be the continuous functions on Ω which are piecewise linear with respect to the k th grid. Note that this introduces slave nodes into the computation, i.e., the vertices of the triangles

on the boundary of the k th refinement region which are not nodes for the $(k - 1)$ st subspace (see Figure 4.1). These nodes are slaves, since the values of functions on these nodes are determined by the values of neighboring nodes and the continuity condition on the subspace. Thus, they do not represent degrees of freedom in the subspace.

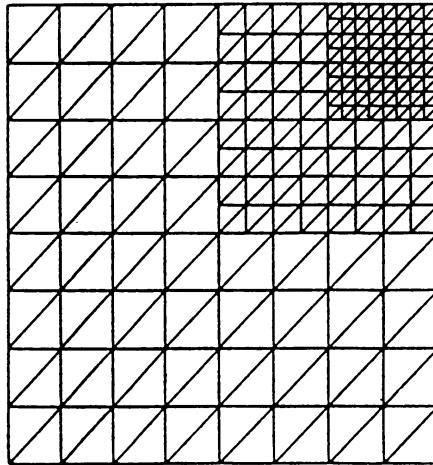


FIGURE 4.1
A mesh with two refinement levels

We shall first show that (A.1) is satisfied. The argument in §3 gives that (A.1) holds for $k = 1, \dots, j + 1$ since this is just the quasi-uniform case. In order to complete the proof of (A.1), we shall introduce some additional notation.

Let us first define $h_k = 2^{-k}$ even in the case when $k = j + 1, \dots, J$, so that h_k corresponds to the size of the finest triangle in the mesh defining \mathcal{M}_k . In addition, let $\Omega_k = (1 - 2^{j-k}, 1) \times (1 - 2^{j-k}, 1)$. Notice that the mesh size of the triangulation defining \mathcal{M}_k , restricted to Ω_k is h_k , and the functions in \mathcal{M}_l with support in Ω/Ω_k for $l \geq k$ are in \mathcal{M}_k . Let $\widetilde{\mathcal{M}}_k$ be the space of piecewise linear functions (which vanish on $\partial\Omega$) defined from the regular uniform triangulation of Ω of size h_k . Note that both \mathcal{M}_k and $\widetilde{\mathcal{M}}_k$ have the same mesh restricted to Ω_k and $\mathcal{M}_k \subset \widetilde{\mathcal{M}}_k$. Finally, \tilde{Q}_k will denote the L^2 projection onto $\widetilde{\mathcal{M}}_k$.

We prove (A.1) for $k > j + 1$. Let $v \in \mathcal{M}$, $l = k - 1$, and consider the function $w \in \mathcal{M}_l$ defined by

$$w = \begin{cases} \tilde{Q}_l v & \text{at the nodes of } \mathcal{M}_l \text{ in the interior of } \Omega_l, \\ v & \text{at the remaining nodes of } \mathcal{M}_l. \end{cases}$$

By the definitions of \tilde{Q}_l and w , and the triangle inequality,

$$(4.1) \quad \begin{aligned} \|(I - Q_l)v\| &\leq \|v - w\| = \|v - w\|_{\Omega_l} \\ &\leq \|(I - \tilde{Q}_l)v\| + \|\tilde{Q}_l v - w\|_{\Omega_l}, \end{aligned}$$

where $\|\cdot\|_{\Omega_l}$ denotes the L^2 norm on Ω_l . Clearly,

$$\|(I - \tilde{Q}_l)v\| \leq Ch_l A^{1/2}(v, v) \leq C\lambda_{l+1}^{-1/2} A^{1/2}(v, v),$$

and hence it suffices to estimate the second term on the right-hand side of (4.1) by the first. But by the definition of w ,

$$\|\tilde{Q}_l v - w\|_{\Omega_l}^2 \leq Ch_l^2 \sum_i (\tilde{Q}_l v(x_l^i) - v(x_l^i))^2,$$

where the sum is taken over the nodes x_l^i on $\partial\Omega_l$. Clearly,

$$h_l^2 \sum_i (\tilde{Q}_l v(x_l^i) - v(x_l^i))^2 \leq C \|(I - \tilde{Q}_l)v\|^2.$$

This proves (A.1).

We next define a sequence of operators $\{R_k\}$ satisfying (A.2). For $k \leq j$, R_k is given by (3.2). Let $\{x_k^l\}$ denote the nodes of the k th grid, and let $\{\phi_k^l\}$ denote the corresponding nodal basis functions. For each node x_k^l with $k > j$ we define

$$h_{kl} = \begin{cases} h_k & \text{if } x_k^l \in \overline{\Omega}_k, \\ h_m & \text{if } x_k^l \in \overline{\Omega}_m / \overline{\Omega}_{m+1}, j \leq m < k. \end{cases}$$

Note that if $x_k^l \in \overline{\Omega}_k / \overline{\Omega}_{k+1}$, then x_k^l is a node for each finer subspace and gets assigned the same value h_k . We then define

$$(4.2) \quad R_k u = h_k^2 \sum_l h_{kl}^{-2} (u, \phi_k^l) \phi_k^l.$$

We will show that

$$(4.3) \quad c_0 \frac{\|u\|^2}{\lambda_k} \leq (R_k u, u) \leq c_1 \frac{\|u\|^2}{\lambda_k} \quad \text{for all } u \in \mathcal{M}_k$$

holds with c_0 and c_1 not depending on k . We proceed as in §3. For $u \in \mathcal{M}_k$, we define $\vec{\alpha}$ by (3.3) and $(G_k)_{lm} = (\phi_k^l, \phi_k^m)$. Let D_k denote the diagonal matrix with diagonal entries $\{h_{kl}\}$. As in §3, $\sum_l h_{kl}^2 \alpha_l^2 = \langle D_k \vec{\alpha}, \vec{\alpha} \rangle$ is a norm which is uniformly equivalent to $\|u\|^2 = \langle G_k \vec{\alpha}, \vec{\alpha} \rangle$. It immediately follows that there are constants c_0 and c_1 , not depending on k , satisfying

$$c_0 \langle D_k \vec{\alpha}, \vec{\alpha} \rangle \leq \langle D_k^{-1} G_k \vec{\alpha}, G_k \vec{\alpha} \rangle \leq c_1 \langle D_k \vec{\alpha}, \vec{\alpha} \rangle.$$

Inequality (4.3) then follows from

$$\langle D_k^{-1} G_k \vec{\alpha}, G_k \vec{\alpha} \rangle = \sum_l h_{kl}^{-2} (u, \phi_k^l)^2 = h_k^{-2} (R_k u, u)$$

and (3.1). Hence (A.2) holds for R_k given by (4.2).

We can apply Corollary 3 to show that $K(\mathcal{B}A) \leq CJ^2$, where \mathcal{B} is defined by (2.13) with R_k and \mathcal{M}_k as above. For this application, we have not been able to prove the regularity and approximation assumption (A.3).

For the purpose of implementation, it is more efficient to reorder the terms defining \mathcal{B} . For $k = j, \dots, J$ let \mathcal{N}_k be the nodes of \mathcal{M}_k in $\overline{\Omega}_k$, and for $k < J$ let \mathcal{N}_k^1 be the nodes of \mathcal{M}_k in $\overline{\Omega}_k/\overline{\Omega}_{k+1}$. For a function $u \in \mathcal{M}$, it is not difficult to see by induction on J that

$$(4.4) \quad \begin{aligned} \mathcal{B}u &= \sum_{k=1}^{j-1} R_k Q_k u + \sum_{x_J^l \in \mathcal{I}_j} (u, \phi_J^l) \phi_J^l \\ &\quad + \sum_{k=j}^{J-1} \left[\sum_{x_k^l \in \mathcal{I}_k^1} \gamma_k^J (u, \phi_k^l) \phi_k^l + \sum_{x_k^l \in \mathcal{I}_k \setminus \mathcal{I}_k^1} (u, \phi_k^l) \phi_k^l \right], \end{aligned}$$

where $\gamma_k^J = h_k^{-2} \sum_{m=k}^J h_m^2$. Note that the R_k terms in the first sum of (4.4) involves the same sums which appear in the uniform case of §3. In addition, the calculation corresponding to the k th mesh in (4.4) for $k = j, \dots, J$ only involves nodal basis functions on $\overline{\Omega}_k$.

Finally, we define a simpler preconditioner $\widehat{\mathcal{B}}$ by replacing γ_k^J by one in (4.4), i.e.,

$$(4.5) \quad \widehat{\mathcal{B}}u = \sum_{k=1}^j \sum_l (u, \phi_k^l) \phi_k^l + \sum_{k=j+1}^J \sum_{x_k^l \in \mathcal{I}_k} (u, \phi_k^l) \phi_k^l.$$

Note that in (4.5), the k th refinement grid only adds a sum over the nodes in $\overline{\Omega}_k$. We note that for $u \in \mathcal{M}$, by (4.4),

$$\begin{aligned} (\mathcal{B}u, u) &= \sum_{k=1}^{j-1} \sum_l (u, \phi_k^l)^2 + \sum_{x_J^l \in \mathcal{I}_j} (u, \phi_J^l)^2 \\ &\quad + \sum_{k=j}^{J-1} \left[\sum_{x_k^l \in \mathcal{I}_k^1} \gamma_k^J (u, \phi_k^l)^2 + \sum_{x_k^l \in \mathcal{I}_k \setminus \mathcal{I}_k^1} (u, \phi_k^l)^2 \right], \end{aligned}$$

with an analogous expression for $\widehat{\mathcal{B}}$. Clearly, $1 \leq \gamma_k^J \leq 4/3$, from which it follows that

$$(\widehat{\mathcal{B}}u, u) \leq (\mathcal{B}u, u) \leq \frac{4}{3}(\widehat{\mathcal{B}}u, u) \quad \text{for all } u \in \mathcal{M}.$$

From the discussion in §3, it is clear that the first sum in (4.5) is a preconditioner for the problem on \mathcal{M}_j , i.e., the finest uniform grid. As we shall see, this sum can also be replaced by any uniform preconditioner for A_j without adversely affecting the asymptotic behavior of the overall condition number. Indeed, let the operator R_j be a preconditioner for A_j (satisfying (2.19) and

the second inequality of (A.2)), and define for $u \in \mathcal{M}$,

$$(4.6) \quad \widehat{B}u = R_j Q_j u + \sum_{k=j+1}^J \sum_{x_k^l \in \mathcal{V}_k^l} (u, \phi_k^l) \phi_k^l.$$

Note that by Remark 2.3, the operator

$$\widetilde{B}u = \sum_{k=j}^J R_k Q_k u$$

satisfies $K(\widetilde{B}A) \leq C(J-j)^2$. We will show that \widetilde{B} is uniformly equivalent to \widehat{B} . Reordering the terms as in (4.4), we have

$$(4.7) \quad \begin{aligned} \widetilde{B}u &= R_j Q_j u + \sum_{x_j^l \in \mathcal{V}_j^l} (\gamma_j^l - 1)(u, \phi_j^l) \phi_j^l \\ &\quad + \sum_{k=j+1}^{J-1} \left[\sum_{x_k^l \in \mathcal{V}_k^l} \gamma_k^l (u, \phi_k^l) \phi_k^l + \sum_{x_k^l \in \mathcal{V}_k^l / \mathcal{V}_k^{l-1}} (u, \phi_k^l) \phi_k^l \right] \\ &\quad + \sum_{x_J^l \in \mathcal{V}_J^l} (u, \phi_J^l) \phi_J^l. \end{aligned}$$

It clearly follows from (4.7) and $1 \leq \gamma_k^l \leq 4/3$ that the operator \widetilde{B} is uniformly equivalent to the operator

$$\widehat{B}u + \sum_{x_j^l \in \mathcal{V}_j^l} (u, \phi_j^l) \phi_j^l.$$

But, by (3.5) and (2.19),

$$\begin{aligned} \sum_{x_j^l \in \mathcal{V}_j^l} (u, \phi_j^l)^2 &\leq C \lambda_j^{-1} \|Q_j u\| \leq C(A_j^{-1} Q_j u, Q_j u) \\ &\leq C(R_j Q_j u, Q_j u) \leq C(\widehat{B}u, u), \end{aligned}$$

from which the equivalence of \widehat{B} and \widetilde{B} follows. Thus, $K(\widehat{B}A) \leq C(J-j)^2$.

Remark 4.1. Clearly, we could generalize this example to include much more general refinements for problems in R^2 as well as higher-dimensional space. Note that the refinement only changes the preconditioner $\widehat{\mathcal{B}}$ (resp. \widehat{B}) by adding additional terms in (4.5) (resp. (4.6)) involving nodes from the refinement region. Thus, this approach is well suited to dynamic adaptive refinement techniques. New refinement regions add terms to the sum, whereas the “de-refinement” of existing regions only takes away terms from the sum. The operator \widehat{B} is even more useful in this context, since it allows the easy inclusion of this refinement preconditioner into existing large-scale uniform grid codes. Preconditioners for the uniform grid already available in the existing code can

be used, supplemented with additional routines implementing the terms due to the refinement.

5. NUMERICAL RESULTS

In this section, we provide the results of numerical examples illustrating the theory developed in the earlier sections. To demonstrate the performance of the proposed algorithms, we shall provide numerical results for a two-dimensional problem with full elliptic regularity and one with less than full elliptic regularity, a two-dimensional example with a geometric mesh refinement and a three-dimensional example. In all of the reported results, the experimentally observed behavior of the condition number of the preconditioned system was in agreement with the theory presented earlier. In the first example, we also compare the results of the new method with those obtained using the hierarchical preconditioning method [20] and a classical V-cycle multigrid preconditioner [4].

For our first example, we consider Problem (1.1) when $L = -\Delta \equiv -\partial^2/\partial x_1^2 - \partial^2/\partial x_2^2$ and Ω is the unit square. This example satisfies the regularity and approximation assumption (A.3) for $\alpha = 1$ as well as (A.1).

We will use a finite element discretization of (1.1) and develop a sequence of grids in a standard way. To define the coarsest grid, we start by breaking the square into four smaller squares of side length $1/2$ and then dividing each smaller square into two triangles by connecting the lower left-hand corner with the upper right-hand corner. Subsequently, finer grids are developed as in the introduction, i.e., by dividing each triangle into the four triangles formed by the edges of the original triangle and the lines connecting the centers of these edges. The space \mathcal{M}_i is defined to be the set of continuous functions on Ω which are piecewise linear on the i th triangulation and vanish on $\partial\Omega$.

We shall compare three preconditioners for (1.2). The first preconditioner \mathcal{B} is defined by the multilevel algorithm (2.13) with R_k given by (3.2) and fits into the framework considered in §3. For comparison, we also provide results for the hierarchical preconditioner B_H [20] and a preconditioner B_M defined by a standard symmetric V-cycle of multigrid [4]. The multigrid algorithm uses one sweep of Jacobi smoothing whenever a grid level is visited, and hence results in two smoothing steps on each grid for each evaluation of the preconditioner. The multigrid algorithm uses $h_0 = 1/4$ for the coarsest grid, while both the hierarchical and the parallel multilevel algorithms use $h_0 = 1/2$.

Table 5.1 gives the condition numbers K of the preconditioned systems $B_H A$, $\mathcal{B} A$, and $B_M A$ corresponding, respectively, to the hierarchical preconditioner, the preconditioner defined by (2.13), and the V-cycle multigrid preconditioner. We note that for these examples, a preconditioned conjugate gradient algorithm using the new preconditioner would be expected to take twice as many iterations as the corresponding algorithm using the V-cycle of multigrid. However, even in a serial implementation, the multigrid algorithm involves

TABLE 5.1
Condition numbers when Ω is the square

h_J	$K(B_H A)$	$K(\mathcal{B} A)$	$K(B_M A)$
1/16	19	7.0	2.3
1/32	31	8.1	2.4
1/64	43	9.0	2.4
1/128	58	9.8	2.4

substantially more computational effort per step. The new method outperforms the hierarchical preconditioner.

This test problem illustrates an example where all three methods work reasonably well. However, we note that \mathcal{B} is preferred over standard multigrid when the parallel aspects of the algorithm are important. In addition, \mathcal{B} generalizes to higher-dimensional problems without convergence rate deterioration (see Table 5.5) and hence would be preferred to the hierarchical method in three-dimensional computations.

We next consider the above preconditioners on a problem with less than full elliptic regularity. We again consider (1.1) with L given by the Laplacian and Ω equal to the “slit domain”, i.e., Ω is the set of points in the interior of the unit square excluding the line $\{(1/2, y) \mid y \in [1/2, 1]\}$. This example does not satisfy the a priori estimates used in the proof of the regularity and approximation assumption (A.3) for $\alpha \geq 1/2$. However, assumption (A.1) is satisfied.

TABLE 5.2
Condition numbers when Ω is the slit domain

h_J	$K(B_H A)$	$K(\mathcal{B} A)$	$K(B_M A)$
1/16	14.6	7.9	2.6
1/32	25.17	10.0	2.9
1/64	38.2	12.6	3.1
1/128	53.8	14.9	3.4

Table 5.2 gives the condition numbers K of the preconditioned systems $B_H A$, $\mathcal{B} A$, and $B_M A$ corresponding, respectively, to the hierarchical preconditioner, the preconditioner defined by (2.13), and the V-cycle multigrid preconditioner. The results are in general agreement with the theoretical estimates

$$\begin{aligned} K(B_H A) &\leq C \ln^2(1/h_J), \\ K(\mathcal{B} A) &\leq C \ln^2(1/h_J), \end{aligned}$$

for the respective methods.

We next provide numerical results for the refinement example of §4. We once again consider the solution of (1.1) with L the Laplacian and Ω the unit square. The sequence of spaces $\mathcal{M}_1 \subset \cdots \subset \mathcal{M}_j$ are as developed in §4 and provide results for the preconditioner $\widehat{\mathcal{B}}$ defined by (4.5). As noted in §4, some such refinement would be necessary if, for example, the function f had a δ -function behavior at the point (1,1). Table 5.3 gives the condition number of the preconditioned system $\widehat{\mathcal{B}}A$ as a function of the mesh size of the uniform grid h_j and the number of refinement levels l . The size of the finest triangle can be computed by dividing the uniform mesh size by 2^l . In all of the runs, the coarsest grid level corresponded to $h_0 = 1/2$. The numerical results seem to indicate that an increase in the number of uniform levels has a greater effect on the condition number than an increase in the number of refinement levels.

TABLE 5.3
Condition numbers for the refinement example

h_j	$l = 1$	$l = 2$	$l = 3$	$l = 4$
1/8	6.3	6.5	6.7	6.9
1/16	7.7	7.9	8.05	8.1
1/32	8.8	9.0	9.1	9.2
1/64	9.6	9.7	9.8	9.9

We next present results for the refinement operator defined by (4.6). The problem and sequence of subspaces are as just described but only the subspaces \mathcal{M}_k , $k \geq j$, are used. In (4.6), we use a multigrid preconditioner (cf. [4]) scaled by 4 to define R_j , the operator on the finest uniform grid. The scaling was introduced to balance the size of the two terms in (4.6). Table 5.4 gives the condition number of the preconditioned system $\widehat{\mathcal{B}}A$ as a function of the mesh size of the uniform grid h_j and the number of refinement levels l .

TABLE 5.4
Condition numbers for $\widehat{\mathcal{B}}A$ using multigrid preconditioning on level j

h_j	$l = 1$	$l = 2$	$l = 3$	$l = 4$
1/8	4.3	6.0	6.4	6.6
1/16	4.7	6.7	7.6	8.1
1/32	4.9	7.0	8.4	9.2
1/64	5.0	7.1	8.5	9.6

As a final example, we illustrate the preconditioning technique on a three-dimensional problem. We consider a Galerkin approximation to the Laplace equation

$$(5.1) \quad \begin{aligned} -\Delta u &= f && \text{in } \Omega, \\ u &= 0 && \text{on } \partial\Omega, \end{aligned}$$

where $\Delta = \partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2$ and Ω is the unit cube. We define the coarse mesh by dividing Ω into eight smaller cubes of size $h_0 = 1/2$. Successively finer meshes are formed by dividing each cube of a coarser mesh into eight smaller cubes. The finite element space \mathcal{M}_k is defined to be the set of continuous functions on Ω which are trilinear with respect to the k th mesh and vanish on $\partial\Omega$.

Table 5.5 gives the condition number K of the preconditioned system $\mathcal{B}A$ where \mathcal{B} is defined by (3.6). This example satisfies full elliptic regularity, and the regularity and approximation assumption (A.3) holds with $\alpha = 1$. Thus, the theory predicts only a logarithmic growth in the condition number, which is in agreement with the reported results. Note the finite element spaces are of rather large dimension, in fact, the $h_j = 1/64$ example has over a quarter of a million unknowns.

TABLE 5.5
Condition numbers for the three-dimensional example

h_j	$K(\mathcal{B}A)$
1/8	4.1
1/16	5.2
1/32	6.0
1/64	6.6

BIBLIOGRAPHY

1. R. E. Bank and T. Dupont, *An optimal order process for solving finite element equations*, Math. Comp. **36** (1981), 35–51.
2. R. E. Bank, T. F. Dupont, and H. Yserentant, *The hierarchical basis multigrid method*, Numer. Math. **52** (1988), 427–458.
3. G. Birkhoff and A. Schoenstadt, eds., *Elliptic problem solvers II*, Academic Press, New York, 1984.
4. J. H. Bramble and J. E. Pasciak, *New convergence estimates for multigrid algorithms*, Math. Comp. **49** (1987), 311–329.
5. J. H. Bramble, J. E. Pasciak, and A. H. Schatz, *The construction of preconditioners for elliptic problems by substructuring. I*, Math. Comp. **47** (1986), 103–134.
6. J. H. Bramble, J. E. Pasciak, and A. H. Schatz, *The construction of preconditioners for elliptic problems by substructuring. II*, Math. Comp. **49** (1987), 1–16.
7. J. H. Bramble, J. E. Pasciak, and A. H. Schatz, *The construction of preconditioners for elliptic problems by substructuring. III*, Math. Comp. **51** (1988), 415–430.
8. J. H. Bramble, J. E. Pasciak, and A. H. Schatz, *The construction of preconditioners for elliptic problems by substructuring. IV*, Math. Comp. **53** (1989), 1–24.
9. J. H. Bramble, J. E. Pasciak, and J. Xu, *The analysis of multigrid algorithms with nonnested spaces or noninherited quadratic forms*, Math. Comp. **56** (1991), (to appear).
10. J. H. Bramble, J. E. Pasciak, and J. Xu, *A multilevel preconditioner for domain decomposition boundary systems*, (in preparation).
11. P. Concus, G. H. Golub, and G. Meurant, *Block preconditioning for the conjugate gradient method*, SIAM J. Sci. Statist. Comput. **6** (1985), 220–252.

12. T. Dupont, R. P. Kendall, and H. H. Rachford, *An approximate factorization procedure for solving self-adjoint elliptic difference equations*, SIAM J. Numer. Anal. **5** (1968), 559–573.
13. R. Glowinski, G. H. Golub, G. A. Meurant, and J. Périaux, eds., Proc. 1st Internat. Conf. on Domain Decomposition Methods for Partial Differential Equations, SIAM, Philadelphia, PA, 1988.
14. W. Hackbusch, *Multi-grid methods and applications*, Springer-Verlag, New York, 1985.
15. S. F. McCormick, *Multigrid methods for variational problems: Further results*, SIAM J. Numer. Anal. **21** (1984), 255–263.
16. S. F. McCormick, *Multigrid methods for variational problems: General theory for the V-cycle*, SIAM J. Numer. Anal. **22** (1985), 634–643.
17. J. Mandel, S. F. McCormick and R. Bank, *Variational multigrid theory*, Multigrid Methods (S. F. McCormick, ed.), SIAM, Philadelphia, PA, 1987, pp. 131–177.
18. J. A. Meyerink and H. A. van der Vorst, *Iterative methods for the solution of linear systems of which the coefficient matrix is a symmetric M-matrix*, Math. Comp. **31** (1977), 148–162.
19. J. Xu, *Theory of multilevel methods*, thesis, Cornell Univ., 1988.
20. H. Yserentant, *On the multi-level splitting of finite element spaces*, Numer. Math. **49** (1986), 379–412.

DEPARTMENT OF MATHEMATICS, CORNELL UNIVERSITY, ITHACA, NEW YORK 14853. *E-mail:* bramble@mssun7.msi.cornell.edu

BROOKHAVEN NATIONAL LABORATORY, UPTON, NEW YORK 11973. *E-mail:* pasciak@bnl.gov

DEPARTMENT OF MATHEMATICS, PENNSYLVANIA STATE UNIVERSITY, UNIVERSITY PARK, PENNSYLVANIA 16802. *E-mail:* xu@rayleigh.psu.edu

3.4 The analysis of multigrid algorithms with nonnested spaces or noninherited quadratic forms

The analysis of multigrid algorithms with nonnested spaces or noninherited quadratic forms[27]

THE ANALYSIS OF MULTIGRID ALGORITHMS WITH NONNESTED SPACES OR NONINHERITED QUADRATIC FORMS

JAMES H. BRAMBLE, JOSEPH E. PASCIAK, AND JINCHAO XU

ABSTRACT. We provide a theory for the analysis of multigrid algorithms for symmetric positive definite problems with nonnested spaces and noninherited quadratic forms. By this we mean that the form on the coarser grids need not be related to that on the finest, i.e., we do not stay within the standard variational setting. In this more general setting, we give new estimates corresponding to the \mathcal{V} cycle, \mathcal{W} cycle and a \mathcal{V} cycle algorithm with a variable number of smoothings on each level. In addition, our algorithms involve the use of nonsymmetric smoothers in a novel way.

We apply this theory to various numerical approximations of second-order elliptic boundary value problems. In our first example, we consider certain finite difference multigrid algorithms. In the second example, we consider a finite element multigrid algorithm with nested spaces, which however uses a prolongation operator that does not coincide with the natural subspace imbedding. The third example gives a multigrid algorithm derived from a loosely coupled sequence of approximation grids. Such a loosely coupled grid structure results from the most natural standard finite element application on a domain with curved boundary. The fourth example develops and analyzes a multigrid algorithm for a mixed finite element method using the so-called Raviart-Thomas elements.

1. INTRODUCTION

In recent years, multigrid methods have been used extensively as tools for obtaining approximations to the solutions of partial differential equations (see the references in [8, 16, 21]). In conjunction, there has been intensive research into the theoretical understanding of these methods (cf. [2, 3, 4, 5, 6, 16, 20, 21, 22, 30] and others). In this paper, we shall extend the theory for symmetric problems so that it applies in a more general framework.

The analysis of this paper can be broken down into two distinct parts. In the first part (§§2, 3, and 4), we provide a general theoretical framework for

Received August 29, 1988; revised June 29, 1989 and October 9, 1989.

1980 *Mathematics Subject Classification* (1985 Revision). Primary 65N30; Secondary 65F10.

This manuscript has been authored under contract number DE-AC02-76CH00016 with the U.S. Department of Energy. Accordingly, the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. This work was also supported in part under the National Science Foundation Grant No. DMS84-05352 and under the Air Force Office of Scientific Research, Contract No. ISSA86-0026 and by the U.S. Army Research Office through the Mathematical Science Institute, Cornell University.

©1991 American Mathematical Society
0025-5718/91 \$1.00 + \$.25 per page

the analysis of multigrid algorithms for symmetric problems with nonnested subspaces. Our algorithms allow the use of nonsymmetric smoothers in a novel way. In the second part (§§5–8), we use the general theory to provide iterative convergence estimates for a number of applications. The results obtained for these examples are new.

One of the most powerful tools for the development of an iterative convergence analysis for multigrid algorithms involves the use of the variational multigrid framework (see, e.g., [3, 4, 5, 21]). This is motivated from the study of the finite element multigrid technique, where the coarser multigrid spaces are nested and the discrete operators on the subspaces are given in terms of a form defined on a larger space (which contains all of the subspaces). As is well known, this analysis generalizes in a straightforward manner to the case of nonnested spaces under the constraint that the form on the coarser grid is equal to the form on the finer applied to the interpolated image (see (2.5), [13, 21, 22]).

This paper provides an analysis which allows the above-mentioned constraint to be violated. In §3, we consider the case when the equality constraint is replaced by a corresponding inequality (see (A.2)). With this weakened assumption, a “regularity and approximation” assumption (see (A.3)), and an appropriate smoother, we show that all of the results in [5] hold. This means that for the \mathcal{V} cycle, the variable \mathcal{V} cycle and the \mathcal{W} cycle algorithm, with any amount of smoothing, $I - B_j A_j$ is a reducer. Here B_j is the corresponding multigrid operator (symmetric and positive definite) and A_j is the operator which we are trying to invert.

In §4, we consider the case when the inequality constraint (A.2) no longer holds. In this case, $I - B_j A_j$ may no longer be a reducer. However, for the \mathcal{V} and variable \mathcal{V} cycle algorithms, the operator B_j is still symmetric and positive definite and hence can be used as a preconditioner. Section 4 provides bounds on the spectrum of $B_j A_j$. We prove that for the variable \mathcal{V} cycle algorithm with the additional regularity and approximation assumption, the system $B_j A_j$ is uniformly well-conditioned (independent of the number of multigrid levels). Thus, we can construct rapidly converging iterative schemes for computing the action of A_j^{-1} using B_j (corresponding to the variable \mathcal{V} multigrid cycle) as a preconditioner. We next provide a result for the \mathcal{W} cycle algorithm without assuming (A.2). In this case, $I - B_j A_j$ is still a reducer (uniformly, independent of j) for the \mathcal{W} cycle provided that m (the number of smoothing steps) is chosen sufficiently large. We finally provide a result for the \mathcal{V} cycle algorithm which is valid if (A.2) holds, up to a perturbation.

Earlier papers provided a technique for proving \mathcal{W} cycle results without the variational framework under the assumption that the number of smoothings was sufficiently large [3, 16]. This approach was used extensively in [16] and, for example, [9, 32].

It is interesting to note that if (A.2) is not satisfied, the operator B_j corresponding to the \mathcal{W} cycle algorithm with a fixed number of smoothings may

be indefinite and hence of little use in an iterative algorithm for computing the action of A_j^{-1} . This is illustrated computationally in an example in §9. This indicates that the variable \mathcal{V} cycle algorithm is more robust than the \mathcal{W} cycle algorithm.

We note that many of the results given in this paper (and also [2, 4, 5, 6, 20] in the variational case) provide multigrid analysis for any number of smoothings on the finest grid. Such results are important to the code developer in that they guarantee that algorithms will work with just one smoothing. Accordingly, it is not necessary to experiment with various amounts of smoothing and one need not be concerned that the number of smoothing iterations may become so large as to make the algorithm no longer practical.

In the second part of the paper, we apply the earlier developed theory to a number of examples. The major part of the analysis necessary for the application of our theory involves the proof of the so-called “regularity and approximation” property. Its proof generally uses the elliptic regularity of the underlying problem as well as the approximation properties of the numerical method.

Sections 5 through 8 consider four different applications of the general theory. In §5, we consider a finite difference example with a lower-order term discretized by the “lumped mass” approximation. Section 6 considers a finite element example with alternative prolongation operators. Section 7 studies a finite element example where the multigrid algorithm was derived with loosely coupled grids. This example can be used to develop and analyze multigrid algorithms for problems with curved boundaries. Section 8 considers a multigrid algorithm for a mixed finite element approximation using the “Raviart-Thomas” elements. In all of these applications, the equality constraint mentioned above does not hold and hence the usual “variational” theory does not apply.

Unless otherwise stated, c , C , and M , with or without subscript will denote generic positive constants which may take on different values in different places. These constants will always be independent of mesh parameters and the number of levels in the multigrid algorithms.

2. MULTIGRID ALGORITHMS

In this section, we describe the symmetric multigrid algorithm in the general setting. We also derive a basic recursion relation which plays a major role in the analysis given in §3.

Let us assume that we are given a sequence of finite-dimensional vector spaces

$$\mathcal{M}_0, \mathcal{M}_1, \dots, \mathcal{M}_j$$

along with linear operators $I_k: \mathcal{M}_{k-1} \mapsto \mathcal{M}_k$ for $k = 1, \dots, j$. The operators $\{I_k\}$ will sometimes be called “prolongation” operators. In addition, we assume that we are given symmetric positive definite quadratic forms $A_k(\cdot, \cdot)$ and $(\cdot, \cdot)_k$ defined on $\mathcal{M}_k \times \mathcal{M}_k$ for $k = 0, \dots, j$. The norm corresponding to $(\cdot, \cdot)_k$ will be denoted by $\|\cdot\|_k$. Examples of families of spaces, operators, and forms will be given in later sections.

We shall develop multigrid algorithms for the solution of the problem: Given $f \in \mathcal{M}_j$, find $v \in \mathcal{M}_j$ satisfying

$$(2.1) \quad A_j(v, \phi) = (f, \phi)_j \quad \text{for all } \phi \in \mathcal{M}_j.$$

To define these algorithms, we first define auxiliary operators. For $k = 0, \dots, j$, define the operator $A_k : \mathcal{M}_k \mapsto \mathcal{M}_k$ by

$$(A_k w, \phi)_k = A_k(w, \phi) \quad \text{for all } \phi \in \mathcal{M}_k.$$

The operator A_k is clearly symmetric (in both the $A_k(\cdot, \cdot)$ and $(\cdot, \cdot)_k$ inner products) and positive definite. Also define the operators $P_{k-1} : \mathcal{M}_k \mapsto \mathcal{M}_{k-1}$ and $P_{k-1}^0 : \mathcal{M}_k \mapsto \mathcal{M}_{k-1}$ by

$$A_{k-1}(P_{k-1}w, \phi) = A_k(w, I_k \phi) \quad \text{for all } \phi \in \mathcal{M}_{k-1},$$

and

$$(P_{k-1}^0 w, \phi)_{k-1} = (w, I_k \phi)_k \quad \text{for all } \phi \in \mathcal{M}_{k-1}.$$

It is easy to see that $I_k P_{k-1}$ is a symmetric operator with respect to the A_k form. Note that, in general, neither P_k^0 nor P_k is a projection.

To define the smoothing process, we require linear operators $R_k : \mathcal{M}_k \mapsto \mathcal{M}_k$ for $k = 1, \dots, j$. This operator may be symmetric or nonsymmetric with respect to the inner product $(\cdot, \cdot)_k$. If R_k is nonsymmetric, then we define R_k^ℓ to be its adjoint and set

$$R_k^{(l)} = \begin{cases} R_k & \text{if } l \text{ is odd,} \\ R_k^\ell & \text{if } l \text{ is even.} \end{cases}$$

The multigrid operator $B_k : \mathcal{M}_k \mapsto \mathcal{M}_k$ is defined by induction and is given as follows.

Multigrid Algorithm

Set $B_0 = A_0^{-1}$. Assume that B_{k-1} has been defined and define $B_k g$ for $g \in \mathcal{M}_k$ as follows:

(1) Set $x^0 = 0$ and $q^0 = 0$.

(2) Define x^l for $l = 1, \dots, m(k)$ by

$$(2.2) \quad x^l = x^{l-1} + R_k^{(l+m(k))}(g - A_k x^{l-1}).$$

(3) Define $y^{m(k)} = x^{m(k)} + I_k q^p$, where q^i for $i = 1, \dots, p$ is defined by

$$(2.3) \quad q^i = q^{i-1} + B_{k-1}[P_{k-1}^0(g - A_k x^{m(k)}) - A_{k-1} q^{i-1}].$$

(4) Define y^l for $l = m(k) + 1, \dots, 2m(k)$ by

$$y^l = y^{l-1} + R_k^{(l+m(k))}(g - A_k y^{l-1}).$$

(5) Set $B_k g = y^{2m(k)}$.

In this algorithm, $m(k)$ is a positive integer which may vary from level to level and determines the number of smoothing iterations on that level. Because of this variable smoothing, the above algorithm is more general than that usually described [2, 3, 8, 16]. If R_k is symmetric and all of the $m(k)$ are the same, then the algorithm is the usual symmetric multigrid cycling algorithm described in a notation which is convenient for our analysis. Note that B_k is clearly a linear operator for each k . In this algorithm, p is a positive integer. We shall study the cases $p = 1$ and $p = 2$, which correspond respectively to the symmetric \mathcal{V} and \mathcal{W} cycles of multigrid.

In the above algorithm, we alternate between R_k and R_k^t in Step 2. In Step 4, we use the adjoints of the Step 2 smoothings applied in the reverse order. This results in a symmetric operator B_j . As far as we know, this form of the multigrid algorithm has not previously been suggested. The exact form of the above algorithm is motivated by the theory presented in later sections. Nonsymmetric smoothers were previously considered in [22, 23], but the theory there assumed a “variational” multigrid setup and full elliptic regularity.

Set $K_k = I - R_k A_k$; then $K_k^* = I - R_k^t A_k$ is the adjoint with respect to $A_k(\cdot, \cdot)$. We now make the following basic assumption:

(A.1) The spectrum of $K_k^* K_k$ is in the interval $[0, 1]$.

Remark 2.1. We note that the Richardson iteration is an example of a symmetric R_k satisfying (A.1). In addition, one sweep of the Gauss-Seidel iteration with any ordering is an example of a nonsymmetric iteration satisfying (A.1).

Set

$$\tilde{K}_k^{(m)} = \begin{cases} (K_k^* K_k)^{m/2} & \text{if } m \text{ is even,} \\ (K_k^* K_k)^{(m-1)/2} K_k^* & \text{if } m \text{ is odd.} \end{cases}$$

Let I denote the identity operator. It is straightforward to check that

$$(2.4) \quad I - B_k A_k = (\tilde{K}_k^{(m(k))})^* [(I - I_k P_{k-1}) + I_k (I - B_{k-1} A_{k-1})^p P_{k-1}] \tilde{K}_k^{(m(k))},$$

cf. (2.7) of [5]. In (2.4), $*$ denotes the adjoint with respect to the inner product $A_k(\cdot, \cdot)$.

Equation (2.4) gives a fundamental recurrence relation for the multigrid operator B_k . A straightforward argument using (2.4) and mathematical induction implies that $I - B_k A_k$ is a symmetric operator on \mathcal{M}_k (even when R_k is non-symmetric) with respect to the A_k form. This immediately implies that B_k is symmetric with respect to the $(\cdot, \cdot)_k$ inner product.

In the above framework, the multigrid spaces need not be related to each other. Note that in the so-called “variational” case studied in [2, 4, 5, 20], it is assumed that

$$(2.5) \quad A_k(I_k u, I_k v) = A_{k-1}(u, v) \quad \text{for all } u, v \in \mathcal{M}_{k-1}.$$

Hence, the forms on all of the coarser grids are defined in terms of, or inherited

from, the form on the finest. The purpose of this paper is to analyze more general multigrid algorithms not satisfying assumption (2.5).

3. GENERAL MULTIGRID THEORY ASSUMING (A.2)

We provide a general multigrid theory in this and the following section. In this section, we consider the case when (2.5) is replaced by the assumption that for $k = 1, \dots, j$,

$$(A.2) \quad A_k(I_k u, I_k u) \leq A_{k-1}(u, u) \quad \text{for all } u \in \mathcal{M}_{k-1}.$$

The reason for such an assumption will become clear as the analysis develops. As illustrated in Theorem 1, this assumption along with (A.1) is sufficient to guarantee that $I - B_j A_j$ is a reducer and that the linear multigrid algorithm converges. In §4, we consider the case when (A.2) fails to hold.

Remark 3.1. Note that by definition, P_{k-1} is the adjoint of I_k and hence (A.2) holds if and only if

$$(3.1) \quad A_{k-1}(P_{k-1} u, P_{k-1} u) \leq A_k(u, u) \quad \text{for all } u \in \mathcal{M}_k.$$

Inequality (3.1) is also equivalent to the nonnegativity of the operator $I - I_k P_{k-1}$ on \mathcal{M}_k . A straightforward argument using (2.4) and mathematical induction implies that $I - B_k A_k$ is also nonnegative. If (A.2) does not hold, it is unlikely that $I - B_k A_k$ is nonnegative.

The goal of this section is to prove that $I - B_k A_k$ is a reducer and to estimate its rate of reduction under the Assumption (A.2). It suffices to show that the inequality

$$(3.2) \quad |A_k((I - B_k A_k)u, u)| \leq \delta_k A_k(u, u) \quad \text{for all } u \in \mathcal{M}_k$$

holds for a constant $\delta_k < 1$ and estimate the dependence of δ_k on k and additional assumptions. The above inequality implies that $I - B_k A_k$ is a contraction with contraction number δ_k . Moreover, if (A.2) holds, then $I - B_k A_k$ is nonnegative and (3.2) is the same as

$$(3.3) \quad A_k((I - B_k A_k)u, u) \leq \delta_k A_k(u, u) \quad \text{for all } u \in \mathcal{M}_k.$$

The first theorem guarantees convergence of the multigrid algorithm under minimal assumptions.

Theorem 1. *Assume that (A.1) and (A.2) hold. Then (3.2) holds for some $\delta_k < 1$.*

The proof of the above theorem will be given later in this section. We note that the hypotheses for the theorem are rather weak. The spaces \mathcal{M}_k need not be related except for the existence of the linear maps I_k . Moreover, the maps I_k need not be injective and the assumptions on R_k are minimal. Thus, the above theorem can be thought of as a result for “algebraic multigrid” since it requires none of the stronger “regularity and approximation” assumptions used in the convergence analysis for the partial differential equation approximation applications.

Moreover, the theorem can still be used to develop multigrid algorithms, even when the forms A_k on the spaces do not a priori satisfy (A.2). Note that (A.2) can be satisfied by simply scaling the forms, i.e., $A_k(\cdot, \cdot) \leftarrow \alpha_k A_k(\cdot, \cdot)$. Clearly, there is no difficulty in applying the multigrid algorithm with the scaled forms. Theorem 1 then implies stability and convergence; however, this convergence may be unacceptably slow without further conditions being satisfied. Multigrid results without this scaling of forms (i.e., when (A.2) fails to hold) are given in §4.

For stronger convergence estimates, we shall make additional a priori assumptions. Let $0 < \alpha \leq 1$. The first assumption is a “regularity and approximation” assumption of the form

$$(A.3) \quad |A_k((I - I_k P_{k-1})u, u)| \leq C_\alpha^2 \left(\frac{\|A_k u\|_k^2}{\lambda_k} \right)^\alpha A_k(u, u)^{1-\alpha} \quad \text{for all } u \in \mathcal{M}_k,$$

where λ_k is the largest eigenvalue of A_k . More precisely, we assume that (A.3) holds with C_α independent of k for $k = 1, \dots, j$.

Let $R_{k,\omega}$ correspond to the Richardson smoothing iteration defined by $R_{k,\omega} = \omega \lambda_k^{-1} I$ and $K_{k,\omega} = (I - R_{k,\omega} A_k)$ be the corresponding reducer. We assume that there exists an ω in $(0, 2)$ not depending on k such that

$$(A.4) \quad A_k(K_k u, K_k u) \leq A_k(K_{k,\omega} u, K_{k,\omega} u) \quad \text{for all } u \in \mathcal{M}_k,$$

i.e., the smoothing process converges as fast as Richardson’s method for some ω .

The above assumption was used by McCormick in [22, 23]. It is not difficult to show that (A.4) is equivalent to the existence of a positive constant C_R not depending on k and satisfying

$$(3.4) \quad \frac{\|u\|_k^2}{\lambda_k} \leq C_R (\tilde{R}_k u, u)_k \quad \text{for all } u \in \mathcal{M}_k,$$

where $\tilde{R}_k = (I - K_k^* K_k) A_k^{-1}$. The inequality (3.4) is convenient for our analysis. We also note that (A.1) immediately follows from (3.4) and hence (A.4) implies (A.1).

Remark 3.2. The Richardson method is a symmetric iteration satisfying (A.4) for $\omega \in (0, 2)$. In finite element and finite difference applications, under reasonable assumptions, one sweep of Gauss-Seidel iteration with any ordering gives rise to a nonsymmetric R_k which also satisfies (A.4).

Remark 3.3. For the theory presented in [5], it was assumed that K_k was non-negative and that

$$(3.5) \quad \frac{\|u\|_k^2}{\lambda_k} \leq C_R (R_k u, u)_k \quad \text{for all } u \in \mathcal{M}_k$$

was satisfied with the (assumed symmetric) operator R_k . It is easy to see that this implies (3.4) with the same constant C_R .

We can now state and prove the theorem for estimating δ_k in (3.2) for the symmetric \mathcal{V} cycle.

Theorem 2. *Assume that (A.2–A.4) hold and define B_j with $p = 1$ and $m(k) = m$ for all k . Then (3.2) holds with*

$$(3.6) \quad \delta_k = \frac{Mk^{\frac{1-\alpha}{\alpha}}}{Mk^{\frac{1-\alpha}{\alpha}} + m^\alpha}.$$

Remark 3.4. The convergence estimates in Theorem 2 and those to be stated later in this section have exactly the same form as the theorems in [5]. The interested reader is referred to [5] for various explicit expressions for M in terms of the constants α , C_α and C_R .

Proof of Theorem 2. We shall prove (3.3) by induction on k . For $k = 0$, there is nothing to prove. Assume that (3.3) holds for $k - 1$. By (2.4),

$$(3.7) \quad A_k((I - B_k A_k)u, u) = A_k((I - I_k P_{k-1})\tilde{u}, \tilde{u}) + A_{k-1}((I - B_{k-1} A_{k-1})P_{k-1}\tilde{u}, P_{k-1}\tilde{u}),$$

where $\tilde{u} = \tilde{K}_k^{(m)}u$. Applying the induction hypothesis and the definition of P_{k-1} gives

$$(3.8) \quad \begin{aligned} A_k((I - B_k A_k)u, u) &\leq A_k((I - I_k P_{k-1})\tilde{u}, \tilde{u}) + \delta_{k-1} A_k(I_k P_{k-1}\tilde{u}, \tilde{u}) \\ &= (1 - \delta_{k-1})A_k((I - I_k P_{k-1})\tilde{u}, \tilde{u}) + \delta_{k-1} A_k(\tilde{u}, \tilde{u}). \end{aligned}$$

Applying (A.3) and a generalized arithmetic-geometric mean inequality gives

$$(3.9) \quad A_k((I - I_k P_{k-1})\tilde{u}, \tilde{u}) \leq C_\alpha^2 \left\{ \alpha \gamma_k \frac{\|A_k \tilde{u}\|_k^2}{\lambda_k} + (1 - \alpha) \gamma_k^{-\alpha/(1-\alpha)} A_k(\tilde{u}, \tilde{u}) \right\}.$$

Applying (3.4) gives

$$(3.10) \quad \frac{\|A_k \tilde{u}\|_k^2}{\lambda_k} \leq C_R A_k((I - \bar{K}_k) \bar{K}_k^m u, u),$$

where

$$(3.11) \quad \bar{K}_k = \begin{cases} K_k^* K_k & \text{if } m \text{ is even,} \\ K_k K_k^* & \text{if } m \text{ is odd.} \end{cases}$$

The remainder of the proof of the theorem is exactly the same as the proof of Theorem 1 in [5]. \square

From the proof of Theorem 2, it is apparent that the framework for nonnested spaces and noninherited forms developed in §2 fits into the machinery of [5]. The next two theorems follow in a similar manner. The first gives a result for the \mathcal{W} cycle algorithm, while the second gives a result for the variable \mathcal{V} cycle algorithm.

Theorem 3. Assume that (A.2–A.4) hold and define B_j with $p = 2$ and $m(k) = m$ for all k . Then (3.2) holds with $\delta_k = \delta$ (independent of k) given by

$$(3.12) \quad \delta = \frac{M}{M + m^\alpha}.$$

Theorem 4. Assume that (A.2–A.4) hold and define B_j with $p = 1$. Assume that $m(k)$ satisfies

$$(3.13) \quad \beta_0 m(k) \leq m(k - 1) \leq \beta_1 m(k).$$

Here, we assume that β_0 and β_1 are constants which are greater than one and independent of k . Then (3.2) holds with

$$(3.14) \quad \delta_k = \frac{M}{M + m(k)^\alpha}.$$

Remark 3.5. We have only provided results for the “symmetric” multigrid cycling schemes, i.e., those in which one smooths both before and after coarse grid correction. The above analysis seems to fail for the nonsymmetric multigrid schemes (described in, for example, [5, 13, 22]) due to the fact that $I_k P_{k-1}$ is no longer a projection and the product of the so-called slash cycles [22] is no longer the symmetric \mathcal{V} cycle.

Proof of Theorem 1. We now prove Theorem 1. Note that since R_k is positive definite and all spaces are finite-dimensional, (3.4) holds for some constant $C_R(j)$ which may depend on $\{R_k\}$, $k = 1, \dots, j$. Similarly, the definiteness of A_k implies that (A.3) holds for some constant $C_\alpha(j)$ which may depend on $\{A_k\}$, $\{I_k\}$, and $\{\langle \cdot, \cdot \rangle_k\}$. Theorems 2–4 still hold with some convergence parameter $\delta_k < 1$ depending on $C_R(j)$ and $C_\alpha(j)$ since, in this case, the constant M is not independent of k . This proves Theorem 1. \square

4. GENERAL MULTIGRID THEORY WITHOUT (A.2)

In this section, we provide an analysis for the multigrid algorithm which allows (A.2) to be violated. In this case, $I - B_k A_k$ may no longer be a reducer. Nevertheless, the operator B_j corresponding to the variable \mathcal{V} and the \mathcal{V} cycle multigrid algorithms is positive definite and hence can be used as a preconditioner in an iterative method for solving (2.1). The \mathcal{W} cycle may, however, be indefinite without increasing the number of smoothings. We first give a theorem with minimal hypotheses, which guarantees that the operator B_k corresponding to the \mathcal{V} or variable \mathcal{V} cycle algorithm is symmetric and positive definite. We next consider additional hypotheses which are sufficient to guarantee iterative convergence rates for variable \mathcal{V} , \mathcal{V} , and \mathcal{W} cycle multigrid algorithms.

Theorem 5. Assume that (A.1) holds and $p = 1$. Then B_j is a symmetric positive definite operator on \mathcal{M}_j .

Proof. As already observed in §2, B_j is a symmetric operator with respect to the $(\cdot, \cdot)_k$ inner product. By (2.4),

$$(4.1) \quad \begin{aligned} (B_k A_k u, A_k u)_k &= A_k((I - \bar{K}_k^{m(k)})u, u) \\ &\quad + (B_{k-1} A_{k-1} P_{k-1} \tilde{K}_k^{(m(k))} u, A_{k-1} P_{k-1} \tilde{K}_k^{(m(k))} u)_{k-1} \end{aligned}$$

for all $u \in \mathcal{M}_k$. Since the eigenvalues of \bar{K}_k are in $[0, 1]$, it follows that $I - \bar{K}_k^{m(k)}$ is a positive definite operator. Thus, by (4.1) and induction, B_k is positive definite. This completes the proof of Theorem 5. \square

Remark 4.1. In general, the theorem does not hold for the \mathcal{W} cycle multigrid algorithm. We give a computational example in §9 where (A.2) is violated and the B_j corresponding to the \mathcal{W} cycle multigrid algorithm with $m = 1$ has negative eigenvalues. Thus, this \mathcal{W} cycle algorithm cannot be used in a preconditioning strategy or to develop a reducer. Computational results given in §9 for the same problem indicate that the corresponding variable \mathcal{V} and \mathcal{W} cycle algorithms give rise to effective preconditioners and hence lead to rapidly converging iterative schemes.

The above theorem can be thought of as a result for algebraic multigrid since the hypotheses are minimal. The theorem guarantees that corresponding preconditioned iterative algorithms (for example, preconditioned conjugate gradient) for the solution of (2.1) will be stable and convergent. The convergence rates of these algorithms may be unacceptably slow without further conditions being satisfied.

In the remainder of this section, we shall make additional assumptions which will lead to theorems which guarantee that the iterative algorithms converge at more reasonable rates. For the \mathcal{V} cycle algorithms, this involves deriving bounds on the largest and smallest eigenvalues of the operator $B_j A_j$. Equivalently, we shall provide positive constants η_0 and η_1 which may depend on k and satisfy the inequalities

$$(4.2) \quad \eta_0 A_k(u, u) \leq A_k(B_k A_k u, u) \leq \eta_1 A_k(u, u) \quad \text{for all } u \in \mathcal{M}_k.$$

Note that if (4.2) holds, then the preconditioned conjugate gradient method converges with an asymptotic reduction rate of

$$\frac{1 - \sqrt{\eta_0/\eta_1}}{1 + \sqrt{\eta_0/\eta_1}}$$

per iterative step.

The next theorem provides estimates for η_0 and η_1 for the variable \mathcal{V} cycle algorithm.

Theorem 6. *Assume that (A.3) and (A.4) hold and define B_j with $p = 1$. Assume that $\{m(k)\}$ satisfy (3.13). Then the constants η_0 and η_1 in (4.2) satisfy*

$$\eta_0 \geq \frac{m(k)^\alpha}{M + m(k)^\alpha} \quad \text{and} \quad \eta_1 \leq \frac{M + m(k)^\alpha}{m(k)^\alpha},$$

i.e., the system $B_j A_j$ is well-conditioned independently of j .

We shall use the following lemma in the proof of Theorem 6.

Lemma 4.1. *Assume that $p = 1$ and that $\bar{\delta}_i$ for $i = 1, 2, \dots, k$ satisfies the inequality*

$$(4.3) \quad -A_i((I - I_i P_{i-1})\tilde{u}, \tilde{u}) \leq \bar{\delta}_i A_i(u, u) \quad \text{for all } u \in \mathcal{M}_i$$

where $\tilde{u} = \tilde{K}_i^{(m(i))}u$. Then

$$\eta_1 \leq \prod_{i=1}^k (1 + \bar{\delta}_i).$$

Proof. It suffices to show that

$$(4.4) \quad -A_k((I - B_k A_k)u, u) \leq (\tau_k - 1)A_k(u, u) \quad \text{for all } u \in \mathcal{M}_k$$

where $\tau_0 = 1$ and $\tau_k = \prod_{i=1}^k (1 + \bar{\delta}_i)$. We prove (4.4) by induction. For $k = 0$, there is nothing to prove. Assume (4.4) holds for $k - 1$. Then by (2.4), the induction assumption and (4.3),

$$\begin{aligned} & -A_k((I - B_k A_k)u, u) \\ &= -A_k((I - I_k P_{k-1})\tilde{u}, \tilde{u}) - A_{k-1}((I - B_{k-1} A_{k-1})P_{k-1}\tilde{u}, P_{k-1}\tilde{u}) \\ &\leq -A_k((I - I_k P_{k-1})\tilde{u}, \tilde{u}) + (\tau_{k-1} - 1)A_{k-1}(P_{k-1}\tilde{u}, P_{k-1}\tilde{u}) \\ &\leq [\bar{\delta}_k + (\tau_{k-1} - 1)(1 + \bar{\delta}_k)]A_k(u, u) = (\tau_k - 1)A_k(u, u). \end{aligned}$$

This completes the proof of the lemma. \square

Proof of Theorem 6. We note that Assumption (A.2) was used in the proof of Theorem 4 only to reduce to the proof of inequality (3.3). The subsequent arguments showing (3.3) remain valid without (A.2) and lead to the inequality

$$A_k((I - B_k A_k)u, u) \leq \delta_k A_k(u, u) \quad \text{for all } u \in \mathcal{M}_k,$$

where δ_k is given by (3.14). It immediately follows that (4.2) holds with $\eta_0 = 1 - \delta_k$.

To estimate η_1 , we note that by (A.3),

$$(4.5) \quad -A_k((I - I_k P_{k-1})\tilde{u}, \tilde{u}) \leq C_\alpha^2 (\lambda_k^{-1} \|A_k \tilde{u}\|_k^2)^\alpha A_k(\tilde{u}, \tilde{u})^{1-\alpha}.$$

By (3.10) and [5, (3.16)],

$$(4.6) \quad \frac{\|A_k \tilde{u}\|_k^2}{\lambda_k} \leq \frac{C}{m(k)} A_k((I - \bar{K}_k^{m(k)})u, u).$$

Since the eigenvalues of \bar{K}_k are in the interval $[0, 1)$,

$$(4.7) \quad -A_k((I - I_k P_{k-1})\tilde{u}, \tilde{u}) \leq C m(k)^{-\alpha} A_k(u, u).$$

Elementary arguments imply that

$$\prod_{k=1}^j \left(1 + \frac{C}{m(k)^\alpha}\right) \leq 1 + \frac{C}{m(j)^\alpha},$$

and hence the bound for η_1 follows from Lemma 4.1. This completes the proof of Theorem 6. \square

We next give a theorem for the \mathcal{W} cycle algorithm.

Theorem 7. *Assume that all the hypotheses except (A.2) hold for Theorem 3. Then, for the \mathcal{W} cycle algorithm with m sufficiently large, (3.2) holds with $\delta_k = \delta$ where δ is given by (3.12). Furthermore, the same conclusion holds if "m sufficiently large" is replaced by the assumption*

$$(A.5) \quad A_k(I_k u, I_k u) \leq 2A_{k-1}(u, u) \quad \text{for all } u \in \mathcal{M}_{k-1}.$$

Proof. We first consider the case without Assumption (A.5). We first show that

$$(4.8) \quad -A_k((I - B_k A_k)u, u) \leq \delta A_k(u, u) \quad \text{for all } u \in \mathcal{M}_k.$$

By (2.4), it clearly suffices to show that

$$(4.9) \quad -A_k((I - I_k P_{k-1})\tilde{u}, \tilde{u}) \leq \delta A_k(u, u)$$

where $\tilde{u} = \tilde{K}_k^{(m)} u$.

Inequality (4.9) immediately follows from (4.7) if m and M are chosen sufficiently large. With (4.8) verified, the proof of the opposite inequality,

$$A_k((I - B_k A_k)u, u) \leq \delta A_k(u, u)$$

follows in the same way as Theorem 3. This completes the proof of Theorem 7 without the assumption of (A.5).

If (A.5) holds, then

$$-A_k((I - I_k P_{k-1})\tilde{u}, \tilde{u}) \leq A_k(\tilde{u}, \tilde{u}).$$

Hence,

$$(4.10) \quad -A_k((I - I_k P_{k-1})\tilde{u}, \tilde{u}) \leq (1 - \delta^2)|A_k((I - I_k P_{k-1})\tilde{u}, \tilde{u})| + \delta^2 A_k(\tilde{u}, \tilde{u}).$$

We note that the first step in the analysis of the \mathcal{W} cycle algorithm in Theorem 3 (see [5, (3.32)]) is to show that the right-hand side of (4.10) bounds $A_k((I - B_k A_k)u, u)$. The remainder of the proof of Theorem 3 bounds the right-hand side of (4.10) by $\delta A_k(u, u)$ (see the proof of Theorem 3 of [5]). This completes the proof of Theorem 7. \square

Remark 4.2. It is elementary but tedious to see from the proof of Theorem 7 (and in particular, the proof of Theorem 3 of [5]) that the constant 2 in (A.5) can always be replaced by $2 + \varepsilon$, where the size of ε depends upon the size of C_R and C_α . Larger C_R and C_α require smaller ε .

The last theorem of this section provides a result for the \mathcal{V} cycle algorithm. For this result, we assume that

$$(A.6) \quad A_k(I_k u, I_k u) \leq (1 + c\lambda_k^{-\gamma})A_{k-1}(u, u) \quad \text{for all } u \in \mathcal{M}_{k-1}$$

which holds for some γ in the interval $(0, 1]$. In many applications, λ_k grows like h_k^{-2} , where h_k is the mesh size. Thus, (A.6) is a perturbation of (A.2) up to some power of h_k . We have the following theorem for the \mathcal{V} cycle algorithm.

Theorem 8. Assume that (A.3), (A.4) and (A.6) hold, and define B_j with $p = 1$ and $m(k) = m$. Assume further that the maximum eigenvalue $\lambda_k \geq \kappa^k$ for some $\kappa > 1$. Then η_1 in (4.2) can be chosen independently of k and $\eta_0 \leq 1 - \delta_k$, where δ_k is given by (3.6).

Proof. Since I_k and P_{k-1} are adjoint operators, (A.6) implies that

$$-A_k((I - I_k P_{k-1})\tilde{u}, \tilde{u}) \leq c\lambda_k^{-\gamma} A_k(\tilde{u}, \tilde{u}) \leq c\lambda_k^{-\gamma} A_k(u, u).$$

Elementary manipulations imply

$$\prod_{k=1}^{\infty} (1 + c\lambda_k^{-\gamma})^{-1} < \infty.$$

Lemma 4.1 shows that η_1 can be bounded independently of k . The bound for η_0 follows from the proof of Theorem 2 (see also the proof of Theorem 6). \square

Remark 4.3. We note that if (A.6) is satisfied, the results of Theorem 7 still hold if the assumption “ m is sufficiently large” is replaced by the assumption “the coarse grid is sufficiently fine”.

5. A FINITE DIFFERENCE APPLICATION

In this section, we consider a finite difference application approximating the solution of the problem

$$(5.1) \quad \begin{aligned} -\Delta u + u &= f && \text{in } \Omega, \\ u &= 0 && \text{on } \partial\Omega. \end{aligned}$$

Here, we do not assume that Ω is a rectangle, and hence standard multigrid analysis for square or periodic domains does not apply. We define a multigrid algorithm in terms of the general approach of §2. In this as well as the remaining sections, we shall consider only simple model problems, even though the techniques obviously extend to more general applications.

We set up a sequence of nodes in the usual way. Without loss of generality, we assume that the domain $\Omega \subset [0, 1]^2$ and $h_k = 2^{-k}/M$ for some integer $M > 1$. Let $N_k = 2^k M - 1$. The nodes of the finite difference approximation on the k th level are given by

$$x_{ij}^k = (ih_k, jh_k) \quad \text{for } i, j = 1, \dots, N_k.$$

We assume that the boundary of the domain Ω aligns with the mesh lines on the coarsest grid, i.e., Ω is the union of coarse grid rectangles. Let Ω_k denote the nodes of the k th grid which are in the interior of Ω . The space \mathcal{M}_k is defined to be the vector space of nodal values defined on Ω_k . The prolongation operator I_k is defined as follows:

- (1) If x_{ij}^k is a node on the $(k-1)$ st grid, then $(I_k V)(x_{ij}^k) = V(x_{ij}^k)$.

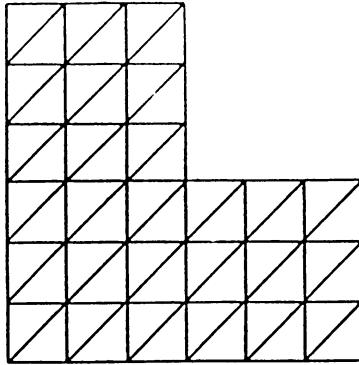


FIGURE 5.1
The triangular mesh

- (2) If x_{ij}^k is a node on an edge in the x or y direction between two nodes in the $(k-1)$ st grid, then $(I_k V)(x_{ij}^k)$ is the average of the nodal values of V at the two nodes.
- (3) Otherwise, $(I_k V)(x_{ij}^k)$ is the average of the nodal value of the node immediately above and to the right with that of the node immediately below and to the left.

The operator I_k corresponds to piecewise linear interpolation on the triangulation of size h_{k-1} in Figure 5.1. For a function $V \in \mathcal{M}_k$, V_{ij} will denote the value of V at x_{ij}^k . The quadratic forms $A_k(\cdot, \cdot)$ and $(\cdot, \cdot)_k$ are defined by

$$(5.2) \quad A_k(U, V) = \sum_{x_{ij}^k \in \Omega_k} [h_k^2 U_{ij} + 4U_{i,j} - U_{i+1,j} - U_{i-1,j} - U_{i,j+1} - U_{i,j-1}] V_{ij}$$

and

$$(5.3) \quad (U, V)_k = h_k^2 \sum_{x_{ij}^k \in \Omega_k} U_{ij} V_{ij}.$$

In (5.2) and the remainder of this section, nodal values are set to zero when the node is not inside of Ω . It is obvious that, with the above definitions, the solution of (2.1) corresponds to the standard finite difference approximation to the solution of (5.1). In fact, the problems on the coarser grids are also standard finite difference approximations. Because of the way the lower-order term of (5.1) is approximated, (2.5) does not hold and the “variational” theory does not apply.

To prove regularity and approximation for this and the remaining applications in this paper, we shall need to use various Sobolev spaces. For nonnegative integers m , the Sobolev space $H^m(\Omega)$ is defined to be the set of functions in $L^2(\Omega)$ whose distributional derivatives up to order m are in $L^2(\Omega)$ (see, e.g., [19, 25]). For $\sigma \in (0, 1)$, the space $H^\sigma(\Omega)$ is defined to be the set of functions

in $L^2(\Omega)$ for which the norm

$$\|v\|_{H^\sigma(\Omega)}^2 \equiv \int_{\Omega} \int_{\Omega} \frac{(v(x) - v(y))^2}{|x - y|^{2+2\sigma}} dx dy$$

is finite. For $s \in (m, m+1)$ and m an integer greater than zero, we define $H^s(\Omega)$ to be the set of functions in $H^m(\Omega)$ for which the norm

$$\|v\|_{H^s(\Omega)}^2 \equiv \|v\|_{H^m(\Omega)}^2 + \sum_{|k|=m} \|D^k u\|_{H^{s-m}(\Omega)}$$

is finite. It can be shown that these spaces coincide with those defined by Hilbert scale interpolation between $L^2(\Omega)$ and $H^{m+1}(\Omega)$ (with equivalent norms) provided that there exists an extension operator which is simultaneously bounded from $L^2(\Omega) \hookrightarrow L^2(R^2)$ and $H^{m+1}(\Omega) \hookrightarrow H^{m+1}(R^2)$ (cf. [18, Theorem 9.3]). The existence of such operators is known in the case of domains with minimally smooth boundary [28] which includes the case of plane polygons.

For a nodal function $V \in \mathcal{M}_k$, let \tilde{V} denote the piecewise linear function on the triangulation depicted by Figure 5.1 which interpolates V at the nodes of \mathcal{M}_k . It can be shown that

$$(5.4) \quad |(\tilde{W}, \tilde{V}) - (W, V)_k| \leq ch_k \|\tilde{W}\|_{H^1(\Omega)} \|V\|_k \quad \text{for all } V, W \in \mathcal{M}_k.$$

Here, (\cdot, \cdot) denotes the L^2 inner product on Ω and $\|\cdot\|_k = (\cdot, \cdot)_k^{1/2}$.

Regularity results for problem (5.1) have been proven in [17]. The space $H^{-1}(\Omega)$ is defined as the set of distributions on Ω for which the norm

$$\|f\|_{H^{-1}(\Omega)} = \sup_{\phi \in H_0^1(\Omega)} \frac{(f, \phi)}{\|\phi\|_{H^1(\Omega)}}$$

is finite. For $s \in (0, 1)$, the spaces $H^{-s}(\Omega)$ are defined by interpolation between $L^2(\Omega)$ and $H^{-1}(\Omega)$ with norms denoted by $\|\cdot\|_{H^{-s}(\Omega)}$. In [17], it is shown that the solution u of (5.1) satisfies inequalities of the form

$$(5.5) \quad \|u\|_{H^{1+\beta}(\Omega)} \leq c \|f\|_{H^{-1+\beta}(\Omega)},$$

where $0 < \beta \leq 1$ is a constant which depends upon $\partial\Omega$. In particular, a result of [17] shows that (5.5) holds for some $\beta > 1/2$ for any polygonal domain in R^2 with interior angles less than 2π .

For our purposes, we shall need an alternative representation of the negative Sobolev norms defined above. Consider the problem: Given $g \in L^2(\Omega)$, find $w \in H_0^1(\Omega)$ satisfying

$$D(w, \chi) = (g, \chi) \quad \text{for all } \chi \in H_0^1(\Omega),$$

where $D(\cdot, \cdot)$ denotes the Dirichlet form on Ω . By Rellich's Lemma [25], $H_0^1(\Omega)$ is compactly contained in $L^2(\Omega)$. It follows that the solution operator $T: L^2(\Omega) \mapsto H_0^1(\Omega)$ defined by $T(g) = w$ has a complete orthonormal basis

of eigenfunctions $\{\psi_i\}$ [27]. The eigenvalues of T corresponding to these eigenfunctions will be denoted $\{\sigma_i\}$. We define the spaces

$$\dot{H}^s = \left\{ \phi = \sum c_i \psi_i \text{ such that } \sum \sigma_i^{-s} c_i^2 < \infty \right\}$$

with norm

$$(5.6) \quad |||\phi|||_s = \left(\sum \sigma_i^{-s} c_i^2 \right)^{1/2}.$$

It is not difficult to show that $\dot{H}^1 = H_0^1(\Omega)$ with equivalent norms. By duality and interpolation, it immediately follows that $\dot{H}^{-s} = H^{-s}(\Omega)$ for $s \in [0, 1]$. Thus inequality (5.5) can be rewritten

$$(5.7) \quad \|u\|_{H^{1+\beta}(\Omega)} \leq c |||f|||_{-1+\beta}.$$

We can now prove the following proposition.

Proposition 5.1. *Let \mathcal{M}_k , $A_k(\cdot, \cdot)$, $(\cdot, \cdot)_k$, and I_k be defined as above. Then (A.2) holds and (A.3) holds for $\alpha = \beta/2$.*

Combining the proposition with Theorems 2–4 gives results for the corresponding multigrid algorithms (with appropriately chosen R_k). Note that we get uniform (independent of h_j) convergence for the \mathcal{W} cycle and variable \mathcal{V} cycle algorithms. With the \mathcal{V} cycle, we may see some logarithmic ($j \sim \ln h_j^{-1}$) deterioration even in the case of full elliptic regularity ($\beta = 1$).

Proof of Proposition 5.1. We first prove (A.2). We write $A_k(\cdot, \cdot) = (\cdot, \cdot)_k + D_k(\cdot, \cdot)$. It is not difficult to see that

$$(5.8) \quad D_k(V, W) = D(\tilde{V}, \tilde{W}) \quad \text{for all } V, W \in \mathcal{M}_k.$$

This uses the assumption that $\partial\Omega$ aligns with the mesh lines of the coarsest grid. Consequently, to prove (A.2), it suffices to show that

$$(5.9) \quad (I_k V, I_k V)_k \leq (V, V)_{k-1} \quad \text{for all } V \in \mathcal{M}_{k-1}.$$

Note that the form $(\cdot, \cdot)_k$ can also be written

$$(W, W)_k = h_k^2/6 \sum_{\tau_i} \sum_{j=1}^3 W(y_{ij})^2,$$

where the first sum is taken over the triangles of the k th mesh and $\{y_{ij}\}_{j=1}^3$ denotes the three vertices of the i th triangle. It clearly suffices to prove the inequality analogous to (5.9) on a typical triangle of the $(k-1)$ st grid. Note that $I_k V = \tilde{V}$ on the nodes of the k th mesh and \tilde{V} is piecewise linear on each triangle of the $(k-1)$ st mesh. Consider a triangle of the $(k-1)$ st grid and a linear function which takes on the values a , b , and c at the nodes. Computing the local contributions to $(I_k V, I_k V)_k$ and $(V, V)_{k-1}$ corresponding to this triangle, we see that (5.9) follows from

$$a^2 + b^2 + c^2 + 3(a+b)^2/4 + 3(b+c)^2/4 + 3(a+c)^2/4 \leq 4(a^2 + b^2 + c^2).$$

This completes the proof of (A.2).

We first introduce some additional notation for the proof of (A.3). Let \tilde{A} be the form on $H_0^1(\Omega) \times H_0^1(\Omega)$ defined by

$$\tilde{A}(w, v) = (w, v) + D(w, v).$$

In addition, let $S_k \subset H_0^1(\Omega)$ denote the collection of piecewise linear functions on the triangulation of size h_k (see Figure 5.1). Let \bar{P}_k denote the elliptic projection onto S_k with respect to $\tilde{A}(\cdot, \cdot)$, i.e., $\bar{P}_k w$ for $w \in H_0^1(\Omega)$ is the unique function in S_k satisfying

$$\tilde{A}(\bar{P}_k w, v) = \tilde{A}(w, v) \quad \text{for all } v \in S_k.$$

From the definitions and (5.8), we have for $W \in \mathcal{M}_k$ and $\Phi \in \mathcal{M}_{k-1}$,

$$\tilde{A}(\widetilde{P}_{k-1} W - \bar{P}_{k-1} \widetilde{W}, \widetilde{\Phi}) = (P_{k-1} W, \widetilde{\Phi}) - (P_{k-1} W, \Phi)_{k-1} + (W, I_k \Phi)_k - (\widetilde{W}, \widetilde{\Phi}).$$

Here we used the identity $I_k \widetilde{\Phi} = \widetilde{\Phi}$. Thus, by (3.1) and (5.4),

$$(5.10) \quad \tilde{A}(P_{k-1} W - \bar{P}_{k-1} \widetilde{W}, P_{k-1} W - \bar{P}_{k-1} \widetilde{W}) \leq ch_k^2 A_k(W, W).$$

To complete the proof of (A.3), we shall need the following lemma. Its proof is the same as that given for the finite element case in [3].

Lemma 5.1. *Let $0 \leq s \leq 1$. There are constants c_0 and c_1 which are independent of j such that*

$$(5.11) \quad \begin{aligned} c_0 \|A_k^{s/2} V\|_k &\leq \|\widetilde{V}\|_{H^s(\Omega)} \\ &\leq c_1 \|A_k^{s/2} V\|_k \quad \text{for all } V \in \mathcal{M}_k, \quad k = 0, \dots, j. \end{aligned}$$

Continuing with the proof of Proposition 5.1, we have by Lemma 5.1,

$$(5.12) \quad A_k((I - I_k P_{k-1})W, W) \leq C \|A_k^{(1+\beta)/2} W\|_k \|\widetilde{W} - P_{k-1} W\|_{H^{1-\beta}(\Omega)}.$$

Now

$$\|\widetilde{W} - P_{k-1} W\|_{H^{1-\beta}(\Omega)} \leq \|\widetilde{W} - \bar{P}_{k-1} \widetilde{W}\|_{H^{1-\beta}(\Omega)} + \|P_{k-1} W - \bar{P}_{k-1} \widetilde{W}\|_{H^1(\Omega)}.$$

By finite element duality [1] (see also (6.12) and the following estimates; this makes use of the assumption (5.7)) we have that

$$\|\widetilde{W} - \bar{P}_{k-1} \widetilde{W}\|_{H^{1-\beta}(\Omega)} \leq ch_k^\beta \|\widetilde{W}\|_{H^1(\Omega)},$$

and combining with (5.10) gives

$$(5.13) \quad \|\widetilde{W} - P_{k-1} W\|_{H^{1-\beta}(\Omega)} \leq ch_k^\beta A_k(W, W)^{1/2}.$$

Clearly,

$$(5.14) \quad \|A_k^{(1+\beta)/2} W\|_k \leq \|A_k W\|_k^\beta A_k(W, W)^{(1-\beta)/2}.$$

Combining (5.10), (5.12), (5.13), and (5.14) proves (A.3) for $\alpha = \beta/2$ and hence completes the proof of the proposition. \square

Remark 5.1. One obvious application of the results of this section is to the case when Ω is an L-shaped domain. As far as we know, this is the first proof which guarantees convergence for a multigrid algorithm for this five-point operator on this domain with linear interpolation as a prolongation operator.

Remark 5.2. It is possible to analyze the analogous multigrid algorithm in the case when the lower-order term in (5.1) has variable coefficients. In that case, it will be unlikely that (A.2) holds, and hence one should use the theory of §4.

6. FINITE ELEMENT EXAMPLES WITH ALTERNATIVE PROLONGATIONS

In this section, we consider two finite element examples with nested spaces where the prolongation operator does not correspond to the natural imbedding of the coarser space into the finer. An immediate consequence of the use of these prolongation operators is that (2.5) no longer holds, and hence the variational theory does not apply. The first example leads to an algorithm which is equivalent to a rather reasonable finite difference multigrid application [8], and our theory provides new estimates for its convergence. The second example provides an instance when (A.2) is violated, in fact, (A.5) is sharp as $k \rightarrow \infty$ (see Remark 6.1).

We consider the simplest of all finite element applications. We start with a domain Ω which is a union of rectangles and consider the problem

$$(6.1) \quad \begin{aligned} -\Delta u &= f && \text{in } \Omega, \\ u &= 0 && \text{on } \partial\Omega. \end{aligned}$$

We shall provide two different finite element subspaces for this problem.

In the first case, we define a coarse grid triangulation by dividing each rectangle into two triangles, using one of the diagonals of the rectangle. Finer grids are defined by successively dividing each triangle into four by connecting the midpoints of the edges of the triangle (see Figure 6.1). The finite element subspace \mathcal{M}_k is defined to be the space of continuous piecewise linear functions on the k th grid which vanish on $\partial\Omega$.

In the second case, we consider the corresponding sequence of rectangular grids, i.e., the original collection of rectangles is successively refined by dividing each rectangle into four subrectangles in the obvious way. The finite element subspace $\widetilde{\mathcal{M}}_k$ is defined to be the space of continuous piecewise bilinear functions on the k th rectangular grid which vanish on $\partial\Omega$.

The Galerkin approximation to the solution u of (6.1) is, of course, defined as the function $u_j \in \mathcal{M}_j$ (resp. $\widetilde{\mathcal{M}}_j$) satisfying

$$D(u_j, \chi) = (f, \chi) \quad \text{for all } \chi \in \mathcal{M}_j \text{ (resp. } \widetilde{\mathcal{M}}_j\text{).}$$

We define $A_k(\cdot, \cdot) = D(\cdot, \cdot)$ and

$$(6.2) \quad (u, v)_k = h_k^2 \sum_{ij} u(x_{ij}^k) v(x_{ij}^k).$$

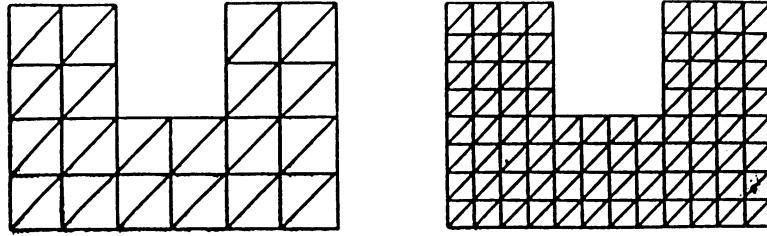


FIGURE 6.1
The grids for \mathcal{M}_0 and \mathcal{M}_1

The sum in (6.2) is taken over the nodal points x_{ij}^k of the k th mesh, and $h_k = 2^{-k} h_0$, where h_0 corresponds to the size of the rectangles of the coarsest mesh.

Note that we get the standard variational finite element multigrid algorithms if we define I_k to be the imbedding of \mathcal{M}_{k-1} into \mathcal{M}_k (resp. $\widetilde{\mathcal{M}}_{k-1}$ into $\widetilde{\mathcal{M}}_k$). Instead, in the first case, for $u \in \mathcal{M}_{k-1}$, we define the values of $I_k u$ at the nodes of \mathcal{M}_k by first interpolating u into \mathcal{M}_{k-1} and subsequently interpolating the result into \mathcal{M}_k . Note that the natural imbedding uses linear interpolation on the $(k-1)$ st triangulation and differs from I_k only in that it assigns $(b+c)/2$ to the center node in Figure 6.2 instead of $(a+b+c+d)/4$. Analogously, in the second case, we define $\widetilde{I}_k u$ at the nodes of $\widetilde{\mathcal{M}}_k$ by interpolation into the subspace \mathcal{M}_{k-1} followed by interpolation into $\widetilde{\mathcal{M}}_k$. Thus at the fine grid nodes, the interpolation operator for the first problem corresponds to the natural imbedding for the second, and vice versa.

The multigrid algorithm in the first case can be thought of as a finite difference multigrid application. Indeed, the stiffness matrix is the standard five-point difference stencil. Moreover, from the finite difference point of view, the prolongation I_k is as reasonable as any other [8].

We can now give the following proposition.

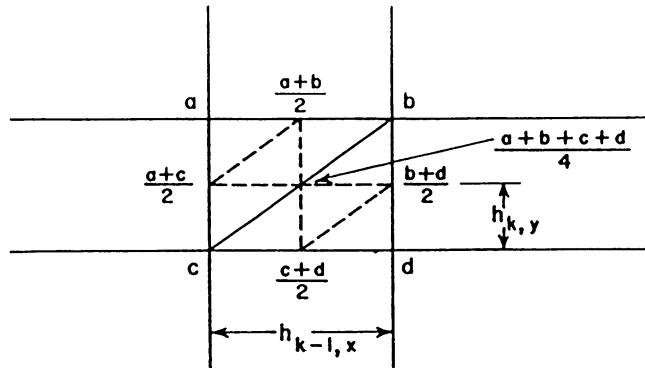


FIGURE 6.2
Nodal values for the I_k interpolation

Proposition 6.1. *Let \mathcal{M}_k , $A_k(\cdot, \cdot)$, $(\cdot, \cdot)_k$, and I_k be defined as above and assume that (5.7) holds for the solution u of (6.1). Then (A.2) holds, and (A.3) holds for $\alpha \leq \beta/2$. If \mathcal{M}_k and I_k are replaced by $\widetilde{\mathcal{M}}_k$ and \widetilde{I}_k , then (A.2) may no longer hold but (A.3) still holds with $\alpha \leq \beta/2$.*

Combining the proposition with Theorems 2–4 gives results for the corresponding multigrid algorithms using \mathcal{M}_k and I_k and an appropriate smoothing process. Note that we get uniform (independent of h_j) convergence for the \mathcal{W} cycle and variable \mathcal{V} cycle algorithms. With the \mathcal{V} cycle, we may see deterioration in the convergence rate like $1 - c/\ln(h_j^{-1})$ even in the case of full elliptic regularity ($\beta = 1$).

In the case of $\widetilde{\mathcal{M}}_k$ and \widetilde{I}_k , (A.2) will not hold in general. Hence, the multigrid operator $I - B_k A_k$ need not be a reducer. In contrast, the \mathcal{V} cycle multigrid strategies employing the multigrid operator as a preconditioner will always be stable and convergent.

Proof of Proposition 6.1. We first prove (A.2) in the case of \mathcal{M}_k , I_k , i.e.,

$$(6.3) \quad D(I_k W, I_k W) \leq D(W, W) \quad \text{for all } W \in \mathcal{M}_{k-1}.$$

It clearly suffices to prove the corresponding local inequality

$$(6.4) \quad D_R(I_k W, I_k W) \leq D_R(W, W) \quad \text{for all } W \in \mathcal{M}_{k-1},$$

where the Dirichlet form is over the domain R and R is a typical rectangle of the $(k-1)$ st grid. Let R be the rectangle pictured in Figure 6.2 and assume that W takes on the values indicated on the corners of the rectangle. Then (6.4) is equivalent to ($r = h_{ky}/h_{kx}$)

$$(6.5) \quad \begin{aligned} & \frac{1}{8}[(a+b-c-d)^2/r + r(b+d-c-a)^2] \\ & + \frac{1}{4}[r(a-b)^2 + r(c-d)^2 + (d-b)^2/r + (c-a)^2/r] \\ & \leq \frac{1}{2}[r(a-b)^2 + r(c-d)^2 + (d-b)^2/r + (c-a)^2/r]. \end{aligned}$$

Inequality (6.5) clearly holds and hence (A.2) follows.

We next prove (A.3) in the case of \mathcal{M}_k , I_k . Lemma 5.1 with v replacing V and \tilde{V} in (5.11) and $A_k(\cdot, \cdot)$ as defined above was proved in [3]. Consequently,

$$(6.6) \quad A_k((I - I_k P_{k-1})w, w) \leq \|A_k^{(1+\beta)/2} w\|_k \|(I - I_k P_{k-1})w\|_{H^{1-\beta}(\Omega)}$$

holds for all $w \in \mathcal{M}_k$. Let \bar{P}_k denote the elliptic projection onto \mathcal{M}_k , i.e.,

$$D(\bar{P}_k v, \phi) = D(v, \phi) \quad \text{for all } \phi \in \mathcal{M}_k.$$

By standard finite element techniques (the duality argument) [1, 12],

$$(6.7) \quad \|(I - \bar{P}_{k-1})w\|_{H^{1-\beta}(\Omega)} \leq ch_k^\beta \|w\|_{H^1(\Omega)}.$$

Here we have used hypothesis (5.7).

Let \bar{I}_k denote the standard interpolation operator onto the subspace \mathcal{M}_k . Applying the Bramble-Hilbert Lemma directly in the fractional-order spaces and [14], noting that on each rectangle of the $(k-1)$ st grid, $I_k \bar{I}_{k-1} - I$ annihilates linear functions, we conclude that

$$(6.8) \quad \|(I_k \bar{I}_{k-1} - I)v\|_{H^{1-\beta}(\Omega)} \leq ch_k^{\beta+\delta} \|v\|_{H^{1+\delta}(\Omega)},$$

and

$$(6.9) \quad \|(\bar{I}_k - I)v\|_{H^{1-\beta}(\Omega)} \leq ch_k^{\beta+\delta} \|v\|_{H^{1+\delta}(\Omega)}$$

holds for $0 \leq \beta \leq 1$ and $0 < \delta \leq 1$. We note that by the inverse property

$$(6.10) \quad \|v\|_{H^{1+\delta}(\Omega)} \leq ch_{k-1}^{-\delta} \|v\|_{H^1(\Omega)};$$

(6.8) and (6.9) hold for $\delta = 0$ when $v \in \mathcal{M}_{k-1}$. The inequality (6.10) will be proved in the appendix.

By the triangle inequality,

$$\begin{aligned} \|(I - I_k P_{k-1})w\|_{H^{1-\beta}(\Omega)} &\leq \|(I - \bar{P}_{k-1})w\|_{H^{1-\beta}(\Omega)} + \|(I - I_k \bar{I}_{k-1})P_{k-1}w\|_{H^{1-\beta}(\Omega)} \\ &\quad + \|(P_{k-1} - \bar{P}_{k-1})w\|_{H^{1-\beta}(\Omega)}. \end{aligned}$$

By (6.6), (6.7), (6.8) and (3.1), (A.3) will follow if we can show

$$(6.11) \quad \|(P_{k-1} - \bar{P}_{k-1})w\|_{H^{1-\beta}(\Omega)} \leq ch_k^\beta \|w\|_{H^1(\Omega)}.$$

We use a duality argument to derive (6.11). Since $H^1 = H_0^1(\Omega)$ is contained in $H^1(\Omega)$, by interpolation

$$\begin{aligned} (6.12) \quad \|(P_{k-1} - \bar{P}_{k-1})w\|_{H^{1-\beta}(\Omega)} &\leq C \|(P_{k-1} - \bar{P}_{k-1})w\|_{1-\beta} \\ &= C \sup_{\phi \in C_0^\infty(\Omega)} \frac{((P_{k-1} - \bar{P}_{k-1})w, T^{(\beta-1)/2}\phi)}{\|\phi\|_{L^2(\Omega)}}. \end{aligned}$$

The power of T is, of course, defined in terms of its eigenfunction expansion and the equality above is an immediate consequence of the definition of the norm in (5.6). Let ζ solve

$$\begin{aligned} -\Delta\zeta &= T^{(\beta-1)/2}\phi && \text{in } \Omega, \\ \zeta &= 0 && \text{on } \partial\Omega. \end{aligned}$$

Then

$$(6.13) \quad \begin{aligned} ((P_{k-1} - \bar{P}_{k-1})w, T^{(\beta-1)/2}\phi) &= D((P_{k-1} - \bar{P}_{k-1})w, \zeta - \bar{I}_{k-1}\zeta) \\ &\quad + D(w, (I_k - I)\bar{I}_{k-1}\zeta). \end{aligned}$$

By (6.9) and (3.1), the first term in (6.13) can be bounded in absolute value by

$$D((P_{k-1} - \bar{P}_{k-1})w, \zeta - \bar{I}_{k-1}\zeta) \leq Ch_k^\beta \|w\|_{H^1(\Omega)} \|\zeta\|_{H^{1+\beta}(\Omega)}.$$

For the second term,

$$\|(I_k - I)\bar{I}_{k-1}\zeta\|_{H^1(\Omega)} \leq \|(I_k \bar{I}_{k-1} - I)\zeta\|_{H^1(\Omega)} + \|(I - I_{k-1})\zeta\|_{H^1(\Omega)}.$$

Applying (6.8) and (6.9) shows that the second term in (6.13) can be bounded similarly. Thus, by (5.7) we have

$$\begin{aligned} ((P_{k-1} - \bar{P}_{k-1})w, T^{(\beta-1)/2}\phi) &\leq ch_k^\beta \|w\|_{H^1(\Omega)} \|\zeta\|_{H^{1+\beta}(\Omega)} \\ &\leq ch_k^\beta \|\phi\|_{L^2(\Omega)} \|w\|_{H^1(\Omega)}. \end{aligned}$$

Combining the above inequalities proves (6.11) and hence completes the proof of the proposition in the case of \mathcal{M}_k, I_k .

The proof in the case $\widetilde{\mathcal{M}}_k$ and \widetilde{I}_k is similar except that we use the inequality corresponding to (6.8) to deduce the boundedness of \cdot'_{k-1} in $H^1(\Omega)$. This completes the proof of Proposition 6.1. \square

Remark 6.1. In general, (A.2) does not hold in the second case. In fact, when $h_x = h_y$, there is a local function defined on the four nodes of size h_{k-1} in Figure 6.2 such that

$$D_R(\widetilde{I}_k W, \widetilde{I}_k W) = 2D_R(W, W).$$

In addition, we have computed the minimal constant μ_k satisfying

$$(6.14) \quad A_k(I_k u, I_k u) \leq \mu_k A_{k-1}(u, u) \quad \text{for all } u \in \mathcal{M}_{k-1},$$

for the slit domain (see Example 9.1) and found that, for this example, $\mu_k \rightarrow 2$ as $k \rightarrow \infty$.

7. A FINITE ELEMENT EXAMPLE WITH LOOSELY COUPLED GRIDS

In this section, we consider a finite element example using a sequence of loosely coupled grids. By loosely coupled, we mean that the triangulation on the k th grid is quasi-uniform of size h_k . In general, the grids and their corresponding finite element subspaces are nonnested. Our results apply to the natural finite element method applied to a problem with curved boundaries where a sequence of grids are generated which successively more closely approximate the boundary of the original domain. In general, (A.2) will not hold. We will show that (A.3) holds with appropriate α and C_α independent of the number of levels. Thus, the preconditioning results of Theorem 6 hold.

Let Ω be a domain in R^2 with piecewise smooth boundary $\partial\Omega$. We consider the numerical approximation to the solution u of the problem

$$(7.1) \quad \begin{aligned} Lu &= f && \text{in } \Omega, \\ u &= 0 && \text{on } \partial\Omega, \end{aligned}$$

where

$$Lv = - \sum_{i,j=1}^2 \frac{\partial}{\partial x_i} a_{ij} \frac{\partial v}{\partial x_j}$$

with $\{a_{ij}(x)\}$ smooth, symmetric, and uniformly positive definite.

We assume that we have defined a sequence of grids $\{T_k\}$ for $k = 0, \dots, j$ approximating Ω such that the k th grid consists of triangles of quasi-uniform

size h_k . In most applications, h_k is roughly twice the size of h_{k+1} , although for our theory we need only assume that there are positive constants c_0 and c_1 not depending on k satisfying

$$(7.2) \quad c_0 h_k \leq h_{k+1} \leq c_1 h_k.$$

We define \mathcal{M}_k to be the set of functions which are piecewise linear on T_k and vanish on the nodes of T_k on $\partial\Omega$. For good approximation, the boundary nodes of the triangulation T_k should lie on $\partial\Omega$. Note that we have not assumed that the nodes of the triangles of the k th grid are related in any way to the nodes of the triangles of the $(k-1)$ st.

For convenience of exposition, we shall only consider the case where every triangle of each T_k lies in $\bar{\Omega}$. We consider the functions in \mathcal{M}_k to be extended by zero to Ω and thus can think of \mathcal{M}_k as being contained in $H^{1+\beta}(\Omega)$ for $\beta < 1/2$. The forms on \mathcal{M}_k are defined by

$$A_k(u, v) = A(u, v) \quad \text{for all } u, v \in \mathcal{M}_k$$

where

$$A(v, w) = \sum_{i,j=1}^2 \int_{\Omega} a_{ij} \frac{\partial v}{\partial x_j} \frac{\partial w}{\partial x_i} dx.$$

The prolongation operators I_k are defined by the natural interpolation operator associated with the subspace \mathcal{M}_k , i.e.,

$$I_k w(x_i) = w(x_i)$$

for nodes x_i of T_k and functions $w \in \mathcal{M}_{k-1}$. The discrete inner products are defined by

$$(u, v)_k = h_k^2 \sum u(x_i) v(x_i)$$

where the sum is taken over all the nodes of T_k .

The following proposition shows that (A.3) holds with appropriate α . Combining this result with Theorems 6 and 7 implies conditioning results for the variable \mathcal{V} cycle and \mathcal{W} cycle multigrid algorithms. The variable \mathcal{V} cycle results hold with $m(j) = 1$ while the \mathcal{W} cycle results hold only assuming that m is sufficiently large.

Proposition 7.1. *Let \mathcal{M}_k , $A_k(\cdot, \cdot)$, $(\cdot, \cdot)_k$, and I_k be defined as above. Furthermore, assume that (5.7) holds for the solution u of (7.1). Then (A.3) holds for $\alpha < \min(\beta/2, 1/4)$.*

Proof. Without loss of generality, we assume $\beta < 1/2$. The argument given in [3] can be used to show that

$$c_0 \|A_k^{(1-\beta)/2} W\|_k \leq \|W\|_{H^{1-\beta}(\Omega)} \leq C_1 \|A_k^{(1-\beta)/2} W\|_k \quad \text{for all } W \in \mathcal{M}_k.$$

The proof proceeds as the proof of Proposition 6.1. By (6.6), it suffices to estimate the norm in $H^{1-\beta}(\Omega)$ of $(I - I_k P_{k-1})w$. Let \bar{P}_{k-1} denote the elliptic projection into \mathcal{M}_{k-1} defined by

$$A(\bar{P}_{k-1} v, \phi) = A(v, \phi) \quad \text{for all } \phi \in \mathcal{M}_{k-1}.$$

The triangle inequality gives

$$(7.3) \quad \begin{aligned} & \| (I - I_k P_{k-1}) w \|_{H^{1-\beta}(\Omega)} \\ & \leq \| (I - \bar{P}_{k-1}) w \|_{H^{1-\beta}(\Omega)} + \| (I - I_k) P_{k-1} w \|_{H^{1-\beta}(\Omega)} \\ & \quad + \| (P_{k-1} - \bar{P}_{k-1}) w \|_{H^{1-\beta}(\Omega)}. \end{aligned}$$

As in (6.9), the estimate

$$(7.4) \quad \| (I - I_k) v \|_{H^{1-\beta}(\Omega)} \leq c h_k^{\beta+\delta} \| v \|_{H^{1+\delta}(\Omega)}$$

holds for $0 \leq \beta \leq 1$ and $0 < \delta \leq 1$. Again, by (6.10), (7.4) holds for $\delta = 0$ when v is in \mathcal{M}_{k-1} . By (7.4) with $\delta = \beta = 0$, P_{k-1} is a bounded operator with respect to the norm induced by the $A(\cdot, \cdot)$ inner product. Hence, we have that

$$\| (I - I_k) P_{k-1} w \|_{H^{1-\beta}(\Omega)} \leq C h_k^\beta \| P_{k-1} w \|_{H^1(\Omega)} \leq C h_k^\beta \| w \|_{H^1(\Omega)}.$$

The estimate corresponding to (6.7) is well known under the assumption (5.7). Thus, we are left to bound the third term of (7.3).

As in the proof of Proposition 6.1, we use a duality argument. Replacing T^{-1} by L in (6.12)–(6.13) gives

$$\| (P_{k-1} - \bar{P}_{k-1}) w \|_{H^{1-\beta}(\Omega)} \leq C \sup_{\phi \in C_0^\infty(\Omega)} \frac{A((P_{k-1} - \bar{P}_{k-1}) w, \zeta)}{\|\phi\|_{L^2(\Omega)}},$$

where ζ is defined by

$$\begin{aligned} L\zeta &= L^{(1-\beta)/2} \phi \quad \text{in } \Omega, \\ \zeta &= 0 \quad \text{on } \partial\Omega. \end{aligned}$$

But

$$A((P_{k-1} - \bar{P}_{k-1}) w, \zeta) = A(w, (I_k - I) \bar{P}_{k-1} \zeta).$$

In the appendix, we show that for $\beta < 1/2$,

$$(7.5) \quad \| \bar{P}_{k-1} v \|_{H^{1+\beta}(\Omega)} \leq C \| v \|_{H^{1+\beta}(\Omega)}$$

and hence (7.4) implies

$$A(w, (I_k - I) \bar{P}_{k-1} \zeta) \leq C h_k^\beta \| \zeta \|_{H^{1+\beta}(\Omega)} \| w \|_{H^1(\Omega)}.$$

Combining the above estimates with (5.7) proves

$$\| (I - I_k P_{k-1}) w \|_{H^{1-\beta}(\Omega)} \leq C h_k^\beta \| w \|_{H^1(\Omega)}.$$

This completes the proof of the proposition. \square

Remark 7.1. Slightly stronger results can be obtained with further assumptions on the relationships between grids. For example, one natural way of developing a sequence of grids is as follows. Each coarse grid triangle gives rise to four triangles in the finer grid connecting the midpoints of interior edges and the midpoint (with respect to arc-length) along the boundary curve for boundary edges (see Figure 7.1). In this case, under reasonable smoothness assumptions

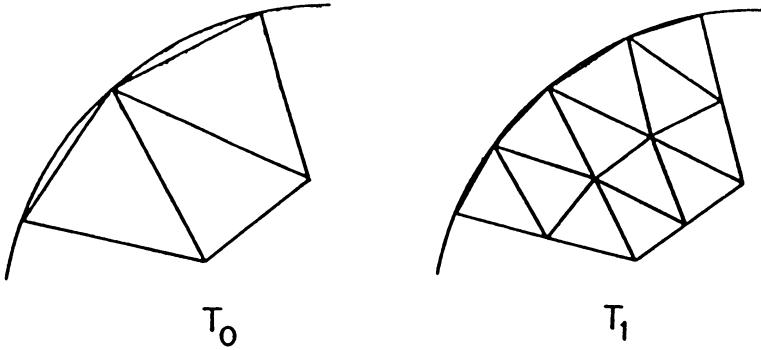


FIGURE 7.1
The mesh of Remark 7.1

on the boundary, it is possible to prove by a perturbation argument that (A.6) holds with $\gamma = 1/4$ [31]. Hence, Theorems 7 and 8, and Remark 4.3, provide results for the corresponding \mathcal{V} and \mathcal{W} cycle algorithms.

8. A MIXED FINITE ELEMENT EXAMPLE

In this section, we develop a multigrid technique for a mixed method finite element approximation of a second-order elliptic problem. We consider the so-called “Raviart-Thomas” elements on triangles and the analogous elements on rectangles [26]. In this example, assumption (A.2) will be satisfied. A similar treatment of the “Brezzi-Douglas-Marini” elements [10] may also be carried out.

Let Ω be a bounded domain in R^N for $N = 2$ or $N = 3$. We consider the problem

$$(8.1) \quad \begin{aligned} -\nabla \cdot (\kappa \nabla w) &= f \quad \text{in } \Omega, \\ w &= 0 \quad \text{on } \partial\Omega. \end{aligned}$$

We assume that $\kappa = \kappa(x)$ is smooth and bounded from below by some constant $\kappa_0 > 0$.

The mixed approximation to (8.1) can be developed as follows. We define $\mathbf{P} = \kappa \nabla w$ and note that the pair (\mathbf{P}, w) satisfies

$$(8.2) \quad \begin{aligned} (\kappa^{-1} \mathbf{P}, \mathbf{Q}) + (w, \nabla \cdot \mathbf{Q}) &= 0 \quad \text{for all } \mathbf{Q} \in H(\text{div}; \Omega), \\ (\nabla \cdot \mathbf{P}, v) &= -(f, v) \quad \text{for all } v \in L^2(\Omega), \end{aligned}$$

where

$$H(\text{div}; \Omega) = \{\Phi \in [L^2(\Omega)]^N \text{ such that } \nabla \cdot \Phi \in L^2(\Omega)\}.$$

$H(\text{div}; \Omega)$ is a Hilbert space [29] with norm given by

$$(\|\Phi\|_{L^2(\Omega)}^2 + \|\nabla \cdot \Phi\|_{L^2(\Omega)}^2)^{1/2}.$$

The pair (\mathbf{P}, w) is approximated in mixed finite element subspace pairs \mathcal{Q}_h, V_h contained respectively in $H(\text{div}; \Omega)$ and $L^2(\Omega)$. Associated with these pairs is an integer r which is related to the approximation order. We assume that the

reader is familiar with the construction of these spaces as described in [26]. The mixed approximation is defined to be the pair $(\mathbf{P}_h, w_h) \in \mathcal{Q}_h \times V_h$ satisfying

$$(8.3) \quad \begin{aligned} (\kappa^{-1} \mathbf{P}_h, Q_h) + (w_h, \nabla \cdot Q_h) &= 0 \quad \text{for all } Q_h \in \mathcal{Q}_h, \\ (\nabla \cdot \mathbf{P}_h, v) &= -(f, v) \quad \text{for all } v \in V_h. \end{aligned}$$

Techniques for solving systems of the form (8.3) have been considered (e.g., [7, 15]). We believe the multigrid technique to be described is new.

To describe our multigrid algorithm, we shall need some additional operator notation. Define the operators $A: \mathcal{Q}_h \mapsto \mathcal{Q}_h$, $B: V_h \mapsto \mathcal{Q}_h$, and $B^*: \mathcal{Q}_h \mapsto V_h$ by

$$\begin{aligned} (Ap, q) &= (\kappa^{-1} p, q) \quad \text{for all } q \in \mathcal{Q}_h, \\ (Bv, q) &= (v, \nabla \cdot q) \quad \text{for all } q \in \mathcal{Q}_h, \\ (B^* p, v) &= (\nabla \cdot p, v) \quad \text{for all } v \in V_h. \end{aligned}$$

Clearly, Ap is the $L^2(\Omega)$ orthogonal projection of $\kappa^{-1} p$ onto \mathcal{Q}_h , $B^* p$ is the $L^2(\Omega)$ orthogonal projection of $\nabla \cdot p$ onto V_h , and B is the adjoint of B^* .

With the above notation, (8.3) can be rewritten

$$\begin{pmatrix} A & B \\ B^* & 0 \end{pmatrix} \begin{pmatrix} \mathbf{P}_h \\ w_h \end{pmatrix} = \begin{pmatrix} 0 \\ -P^0 f \end{pmatrix},$$

where P^0 denotes the $L^2(\Omega)$ orthogonal projection onto V_h . Thus, the solution w_h of (8.3) satisfies

$$(8.4) \quad B^* A^{-1} B w_h = P^0 f.$$

We shall develop a multigrid algorithm for (8.4). In the remainder of this section, we restrict ourselves to the case of R^2 .

For the multigrid algorithm to be developed, we assume that the cost of evaluating A^{-1} applied to a vector in \mathcal{Q}_h is not too expensive. This is true in the case of tensor product elements on a regular rectangular grid, where the evaluation of the action of A^{-1} involves banded solves (of bandwidth proportional to r) along lines of constant y for the v_x -component and along lines of constant x for the v_y -component. The overall cost is thus proportional to the total number of unknowns. In the case of triangles, one might consider iterative evaluation of A^{-1} , a well-conditioned system. To highlight the ideas, from here on, we limit our discussion to the case of tensor product elements on rectangles and assume that Ω is the union of such rectangles.

We develop a sequence of rectangular subgrids $\{R_k\}$, $k = 0, \dots, j-1$, in the usual way. We start by a coarse partitioning of Ω into rectangular elements. Each successively finer grid is defined by partitioning coarser grid rectangles into four equal size subrectangles. The mixed element approximation subspaces are defined with respect to the finest grid R_{j-1} only, and we define $\mathcal{M}_j = V_h$. The space \mathcal{M}_k for $k < j$ is defined to be the set of continuous piecewise bilinear functions with respect to the k th grid which vanish on $\partial\Omega$.

Note that \mathcal{M}_j and \mathcal{M}_{j-1} are defined with respect to the same grid. The spaces $\mathcal{M}_0 \subset \mathcal{M}_1 \subset \cdots \subset \mathcal{M}_{j-1}$ form a nested sequence of spaces, but $\mathcal{M}_{j-1} \subset \mathcal{M}_j$ only if $r > 0$.

We next define the multigrid forms. The form A_j is defined directly from the mixed method by

$$A_j(\theta, \chi) = (B^* A^{-1} B \theta, \chi) \quad \text{for all } \theta, \chi \in \mathcal{M}_j.$$

The forms on the spaces \mathcal{M}_k for $k < j$ are defined by

$$A_k(\theta, \chi) = \int_{\Omega} \kappa \nabla \theta \cdot \nabla \chi \, dx \quad \text{for all } \theta, \chi \in \mathcal{M}_k.$$

We assume that “discrete” inner products $(\cdot, \cdot)_k$ are defined satisfying

$$(8.5) \quad c \|\theta\|_{L^2(\Omega)}^2 \leq (\theta, \theta)_k \leq C \|\theta\|_{L^2(\Omega)}^2 \quad \text{for all } \theta \in \mathcal{M}_k,$$

for $k = 0, \dots, j$, with c, C independent of k . Note that, for $k < j$ we can use (6.2) to define $(\cdot, \cdot)_k$. In addition, $(\cdot, \cdot)_j$ can be defined to be the $L^2(\Omega)$ inner product.

To complete the definition of the multigrid algorithm, we need only define the operators I_k , $k = 1, \dots, j$. Except in the case $r = 0$, all spaces are nested, and hence I_k can be defined by the natural injection. For $r = 0$, only \mathcal{M}_{j-1} is not contained in \mathcal{M}_j . In this case, we define $I_j \theta$ to be the function in \mathcal{M}_j whose value on a grid rectangle is the mean value of θ on that rectangle.

For sufficiently smooth κ , regularity results of the form (5.7) hold for the solution of (8.1) as discussed earlier.

We now give the proposition which shows that the hypotheses (A.2) and (A.3) hold with the above operator definitions. Combining these results with the theorems of §§3 and 4 implies iterative convergence estimates for the corresponding multigrid algorithms (with appropriate smoothers).

Proposition 8.1. *Let \mathcal{M}_k , $A_k(\cdot, \cdot)$, $(\cdot, \cdot)_k$, and I_k be defined as above. Then (A.2) holds. Furthermore, if the solution u of (8.1) satisfies (5.7) for some $\beta \in (1/2, 1]$, then (A.3) holds for $\alpha = \beta/2$.*

Proof. We first show that (A.2) holds. Since the above setup corresponds to the usual finite element multigrid for $k < j$, the stronger result (2.5) holds for $k \neq j$. Thus, we need only verify (A.2) for $k = j$. It is easy to see that for $\theta \in \mathcal{M}_j$,

$$(8.6) \quad A_j(\theta, \theta) = \sup_{\chi \in \mathcal{C}_h} \frac{(\theta, \nabla \cdot \chi)^2}{(\kappa^{-1} \chi, \chi)}.$$

If $r > 0$, then for $\theta \in \mathcal{M}_{j-1}$,

$$A_j(\theta, \theta) = \sup_{\chi \in \mathcal{C}_h} \frac{(\theta, \nabla \cdot \chi)^2}{(\kappa^{-1} \chi, \chi)} = \sup_{\chi \in \mathcal{C}_h} \frac{(\kappa^{1/2} \nabla \theta, \kappa^{-1/2} \chi)^2}{(\kappa^{-1} \chi, \chi)} \leq A_{j-1}(\theta, \theta).$$

For $r = 0$, since $\nabla \cdot \chi$ is constant on the rectangles of size h_j ,

$$\begin{aligned} A_j(I_j\theta, I_j\theta) &= \sup_{\chi \in \mathcal{Q}_h} \frac{(I_j\theta, \nabla \cdot \chi)^2}{(\kappa^{-1}\chi, \chi)} = \sup_{\chi \in \mathcal{Q}_h} \frac{(\theta, \nabla \cdot \chi)^2}{(\kappa^{-1}\chi, \chi)} \\ &= \sup_{\chi \in \mathcal{Q}_h} \frac{(\kappa^{1/2}\nabla\theta, \kappa^{-1/2}\chi)^2}{(\kappa^{-1}\chi, \chi)} \leq A_{j-1}(\theta, \theta). \end{aligned}$$

Combining the above inequalities verifies (A.2).

We next prove (A.3). Again, since for $k < j$, the above setup corresponds to the usual finite element multigrid, we need only consider $k = j$. Fix $u \in \mathcal{M}_j$ and define $f \in \mathcal{M}_j$ to be the solution of

$$(8.7) \quad (f, \theta) = A_j(u, \theta) \quad \text{for all } \theta \in \mathcal{M}_j.$$

Clearly, f is well defined. Moreover, u is the mixed approximation to the function $W \in H_0^1(\Omega)$ satisfying

$$\int_{\Omega} \kappa \nabla W \cdot \nabla \phi \, dx = (f, \phi) \quad \text{for all } \phi \in H_0^1(\Omega).$$

Note that $P_{j-1}u$ satisfies

$$A_{j-1}(P_{j-1}u, \chi) = (f, \chi) \quad \text{for all } \chi \in \mathcal{M}_{j-1}$$

and hence $P_{j-1}u$ is the standard (conforming) finite element approximation to W in \mathcal{M}_{j-1} .

By the Schwarz inequality,

$$A_j((I - I_j P_{j-1})u, u) \leq A_j((I - I_j P_{j-1})u, (I - I_j P_{j-1})u)^{1/2} A_j(u, u)^{1/2}.$$

Consequently, (A.3) will follow if we can show

$$\begin{aligned} (8.8) \quad &A_j((I - I_j P_{j-1})u, (I - I_j P_{j-1})u) \\ &\leq Ch_j^{2\beta} \|A_j u\|_j^{2\beta} A_j(u, u)^{1-\beta} \quad \text{for all } u \in \mathcal{M}_j. \end{aligned}$$

Fix $u \in \mathcal{M}_j$ and let $q \in \mathcal{Q}_h$ satisfy $Aq + Bu = 0$. Applying known error estimates for the mixed finite element method with $\beta > 1/2$ [24] and the standard finite element method [1, 12] gives

$$\begin{aligned} (8.9) \quad &A_j((I - I_j P_{j-1})u, (I - I_j P_{j-1})u)^{1/2} = \sup_{\chi \in \mathcal{Q}_h} \frac{(u - P_{j-1}u, \nabla \cdot \chi)}{(\kappa^{-1}\chi, \chi)^{1/2}} \\ &= \sup_{\chi \in \mathcal{Q}_h} \frac{(\nabla W - \kappa^{-1}q, \chi)}{(\kappa^{-1}\chi, \chi)^{1/2}} + \sup_{\chi \in \mathcal{Q}_h} \frac{(\nabla(P_{j-1}u - W), \chi)}{(\kappa^{-1}\chi, \chi)^{1/2}} \\ &\leq Ch_j^\beta \|W\|_{H^{1+\beta}(\Omega)} \leq Ch_j^\beta \|f\|_{H^{-1+\beta}(\Omega)}. \end{aligned}$$

The last inequality of (8.9) used (5.7).

Clearly, by (8.5) and (8.7),

$$\|f\|_{H^{-1+\beta}(\Omega)} \leq C \|f\|_{H^{-1}(\Omega)}^{1-\beta} \|f\|_{L^2(\Omega)}^\beta \leq C \|f\|_{H^{-1}(\Omega)}^{1-\beta} \|A_j u\|_j^\beta.$$

In addition,

$$\begin{aligned}\|f\|_{H^{-1}(\Omega)} &= \sup_{\phi \in H_0^1(\Omega)} \frac{(f, P^0 \phi)}{\|\phi\|_{H^1(\Omega)}} = \sup_{\phi \in H_0^1(\Omega)} \frac{(A_j u, P^0 \phi)_j}{\|\phi\|_{H^1(\Omega)}} \\ &\leq A_j(u, u)^{1/2} \sup_{\phi \in H_0^1(\Omega)} \frac{A_j(P^0 \phi, P^0 \phi)^{1/2}}{\|\phi\|_{H^1(\Omega)}}.\end{aligned}$$

But the mixed element spaces satisfy $\nabla \cdot \mathcal{Q}_h \subset V_h$, and hence

$$\begin{aligned}A_j(P^0 \phi, P^0 \phi) &= \sup_{\chi \in \mathcal{Q}_h} \frac{(P^0 \phi, \nabla \cdot \chi)^2}{(\kappa^{-1} \chi, \chi)} = \sup_{\chi \in \mathcal{Q}_h} \frac{(\phi, \nabla \cdot \chi)^2}{(\kappa^{-1} \chi, \chi)} \\ &= \sup_{\chi \in \mathcal{Q}_h} \frac{(\nabla \phi, \chi)^2}{(\kappa^{-1} \chi, \chi)} \leq C \|\phi\|_{H^1(\Omega)}^2.\end{aligned}$$

Combining the above results gives

$$(8.10) \quad \|f\|_{H^{-1+\beta}(\Omega)} \leq C A_j(u, u)^{(1-\beta)/2} \|A_j u\|_j^\beta.$$

Combining (8.9)–(8.10) verifies (8.8). This completes the proof of the proposition. \square

9. NUMERICAL RESULTS

We provide the results of a few numerical experiments to illustrate the theory developed in the earlier sections. We have made no attempt to provide numerical results for all of the examples. Instead, we provide examples only to illustrate the theorems in §§3 and 4.

Example 9.1. We consider the Laplace equation on a slit domain. Specifically, we define Ω to be the points interior to the unit square which are not on the line $(1/2, y)$ for $y \in [1/2, 1]$, and we approximate the solution to (6.1). We define \mathcal{M}_k to be the space $\widetilde{\mathcal{M}}_k$ of piecewise bilinear functions on the k th rectangular grid as developed in §6. The prolongation operator I_k corresponds to the linear interpolant \tilde{I}_k with respect to the triangular mesh defined in §6. For this example, the form A_k corresponds to the Dirichlet form on the subspace, i.e.,

$$A_k(u, u) = D(u, u) \quad \text{for all } u \in \mathcal{M}_k.$$

In this application, (A.5) is satisfied, but (A.2) is not.

Table 9.1 gives the condition number K for the system $B_j A_j$ and the reduction factor δ ($\delta = \delta_j$ in (3.2)) for the system $I - B_j A_j$ as a function of the mesh size on the finest grid. We compare the \mathcal{V} cycle (K_v, δ_v) , the variable \mathcal{V} cycle (K_{vv}, δ_{vv}) and the \mathcal{W} cycle (K_w, δ_w) multigrid algorithm. We use Richardson smoothing, and hence (A.4) is satisfied. The variable \mathcal{V} cycle used twice the number of smoothings on each consecutively coarser grid (i.e., $\beta_0 = \beta_1 = 2$) and one smoothing on the finest grid. The \mathcal{V} and \mathcal{W} cycle

TABLE 9.1
Convergence results for Example 9.1

h_j	$K_{vv}(\delta_{vv})$	$K_w(\delta_w)$	$K_v(\delta_v)$
1/8	2.1 (.45)	2.1 (.45)	2.1 (.45)
1/16	2.2 (.45)	2.2 (.45)	2.4 (.50)
1/32	2.2 (.46)	2.2 (.46)	2.7 (.54)
1/64	2.2 (.46)	2.2 (.46)	3.0 (.57)
1/128	2.2 (.46)	2.2 (.46)	3.2 (.59)

algorithms used $m(k) = m = 1$. For all of the runs, the coarse grid corresponded to a mesh of size $1/4$ and the coarse grid problems were essentially solved by applying 40 smoothing steps. Note that for this example, the computational results for the variable \mathcal{V} and the \mathcal{W} cycle multigrid algorithms are essentially identical. This is reasonable since both algorithms have exactly the same number of total smoothings on each grid in the multi-level iteration. This example satisfies the hypotheses of Theorems 6 and 7, and the observed behavior of the variable \mathcal{V} and \mathcal{W} cycle algorithms agree with the theory. However, the behavior of the \mathcal{V} cycle algorithm is perhaps better than one would expect from the theory of §§3 and 4.

Example 9.2. This example illustrates what can happen to the multigrid algorithms when the minimal constant μ_k satisfying (6.14) is greater than 2. We consider the same setup as in Example 9.1 except that we define A_k by

$$(9.1) \quad A_k(u, u) = \tau_k D(u, u) \quad \text{for all } u \in \mathcal{M}_k,$$

where $\tau_j = 1$ and for $k < j$,

$$\tau_k = \prod_{i=k}^{j-1} (1 + 6h_i).$$

A result of this scaling is that (A.5) no longer holds. Clearly, $\tau_k = 1 + O(h_k)$ and it is not difficult to show that (A.3) still holds.

Even though the scaling introduced in (9.1) is purely artificial, it is not unreasonable to expect similar differences in forms in actual applications. Such differences might be observed if the operator involved had variable coefficients and the forms on the individual grids were computed by numerical integration.

Table 9.2 gives computational results for this example. The condition number for the variable \mathcal{V} cycle algorithm (K_{vv}) and the \mathcal{V} cycle (K_v) algorithm as a function of h_j is reported. In addition, the largest (η_1^w) and smallest (η_0^w) eigenvalue of the operator $B_j A_j$ is given in the case of the \mathcal{W} cycle algorithm with $m = 1$. Finally, the minimum value of μ_k satisfying (6.14) is also given. Note that for these computations, the \mathcal{W} cycle algorithm leads to an indefinite operator B_j for more than two levels. Thus, the extra smoothing requirement

TABLE 9.2
Convergence results for Example 9.2

h_j	K_{vv}	K_v	$\eta_0^w (\eta_1^w)$	μ_k
1/8	3.8	3.8	.57 (2.2)	3.7
1/16	5.2	6.1	-.4 (1.6)	3.2
1/32	5.7	8.5	-1.2 (1.4)	2.7
1/64	5.2	10.6	-4.4 (1.3)	2.4
1/128	4.6	12.5	-30 (1.2)	2.2

in Theorem 7 is needed to produce a stable \mathcal{W} cycle multigrid algorithm. In contrast, the hypotheses for Theorem 6 are satisfied and the computational results for the variable \mathcal{V} cycle algorithm illustrate the uniform conditioning of the $B_j A_j$ guaranteed by the theory. As in Example 9.1, the behavior of the \mathcal{V} cycle seems better than that predicted by the theory. The \mathcal{V} cycle does show more deterioration in the condition number compared to Example 9.1, but nevertheless would lead to a reasonable preconditioned strategy for solving (2.1).

Remark 9.1. Although not reported, the largest eigenvalues of $B_j A_j$ for variable \mathcal{V} and \mathcal{W} cycle computations in Table 9.2 were always greater than 2. Accordingly, $I - B_j A_j$ is not a reducer. Obviously, there exists a constant $\gamma < 1$ so that $I - \gamma B_j A_j$ is a reducer with a good reduction rate. An iterative algorithm with reduction matrix $I - \gamma B_j A_j$ can be trivially constructed and is, equivalently, a linear preconditioned iteration for the computation of the action of A_j^{-1} applied to a function in \mathcal{M}_j . Note that for the \mathcal{W} cycle algorithm with one smoothing and more than two levels, there does not exist a constant γ so that $I - \gamma B_j A_j$ is a reducer. For a stable iterative technique utilizing the \mathcal{W} cycle algorithm, one would have to increase m .

10. APPENDIX

We give a proof of (6.10) and (7.5) in this section. We will prove the results for piecewise linear functions on triangles. The proof for bilinear functions is similar. Let m be a nonnegative integer. Assume that we are given a quasi-uniform triangulation $\{\cup \tau_i\}$ of size h on Ω and consider the space S_h of discontinuous piecewise polynomials up to degree m on this triangulation. We will prove that the inverse inequality

$$(10.1) \quad \|v\|_{H^\alpha(\Omega)} \leq Ch^{-\alpha} \|v\|_{L^2(\Omega)} \quad \text{for all } v \in S_h$$

holds for $\alpha < 1/2$.

Assuming that (10.1) holds, we can prove (6.10) and (7.5) as follows. Clearly,

$$(10.2) \quad \|v\|_{H^{1+\alpha}(\Omega)} \leq C \left(\|v\|_{H^1(\Omega)} + \sum_{i=1}^2 \left\| \frac{\partial v}{\partial x_i} \right\|_{H^\alpha(\Omega)} \right).$$

Inequality (6.10) follows applying (10.1) to $\frac{\partial v}{\partial x_i}$. To prove (7.5), we let the above triangulation correspond to the triangulation defining \mathcal{M}_{k-1} and let π denote the $L^2(\Omega)$ orthogonal projection operator onto \mathcal{M}_{k-1} . Obviously, it suffices to prove that

$$\|\bar{P}_{k-1}u - u\|_{H^{1+\beta}(\Omega)} \leq C\|u\|_{H^{1+\beta}(\Omega)}.$$

Clearly,

$$\begin{aligned} & \|\bar{P}_{k-1}u - u\|_{H^{1+\beta}(\Omega)} \\ & \leq \|u\|_{H^1(\Omega)} + \sum_{i=1}^2 \left(\left\| (I - \pi) \frac{\partial u}{\partial x_i} \right\|_{H^\beta(\Omega)} + \left\| \pi \frac{\partial u}{\partial x_i} - \frac{\partial}{\partial x_i} (\bar{P}_{k-1}u) \right\|_{H^\beta(\Omega)} \right). \end{aligned}$$

By (10.1) and approximation properties,

$$\begin{aligned} \left\| \pi \frac{\partial u}{\partial x_i} - \frac{\partial}{\partial x_i} (\bar{P}_{k-1}u) \right\|_{H^\beta(\Omega)} & \leq Ch_k^{-\beta} \left\| \pi \frac{\partial u}{\partial x_i} - \frac{\partial}{\partial x_i} (\bar{P}_{k-1}u) \right\|_{L^2(\Omega)} \\ & \leq C\|u\|_{H^{1+\beta}(\Omega)}. \end{aligned}$$

Inequality (7.5) then follows from the fact that π is a bounded operator on $H^\beta(\Omega)$.

We provide a proof of (10.1) in the remainder of this section. To do this, we use the real method of interpolation of Lions and Peetre (see [11]) which asserts that we may take

$$\|u\|_{H^\beta(\Omega)}^2 = \int_0^\infty K(u, t)^2 t^{-2\beta-1} dt,$$

where

$$K(u, t)^2 = \inf_{v \in H^1(\Omega)} (\|u - v\|_{L^2(\Omega)}^2 + t^2 \|v\|_{H^1(\Omega)}^2).$$

In fact, a direct computation shows that the norm above is equal to a constant (depending on β) multiple of the Hilbert scale norm. Taking $v = 0$ in the definition of $K(u, t)^2$ gives

$$(10.3) \quad \int_h^\infty K(u, t)^2 t^{-2\beta-1} dt \leq (2\beta)^{-1} h^{-2\beta} \|u\|_{L^2(\Omega)}^2.$$

Thus, we are left to estimate the integral from 0 to h .

For $t \leq h$, we define v (depending on t) as follows. Let ϕ_i be a smooth function defined on the i th triangle τ_i of the triangulation defining S_h satisfying

$$\phi_i(x) = \begin{cases} 0 & \text{if } x \text{ is not in } \tau_i, \\ 1 & \text{if } x \text{ is in } \tau_i \text{ and the distance from } x \text{ to } \partial\tau_i \text{ is } \geq t. \end{cases}$$

In addition, assume that ϕ_i satisfies

$$\phi'_i(x) \leq Ct^{-1} \quad \text{for all } x \in \tau_i.$$

We then define $v \in H^1(\Omega)$ by

$$v = \sum_i \phi_i u.$$

By the quasi-uniformity of the triangulation and the smoothness of ϕ_i ,

- (1) $\|u - v\|_{L^2(\tau_i)}^2 \leq Cht \|u\|_{L^\infty(\tau_i)}^2$ and
- (2) $\|v\|_{H^1(\tau_i)}^2 \leq C(\|u\|_{H^1(\tau_i)}^2 + ht^{-1} \|u\|_{L^\infty(\tau_i)}^2).$

We clearly have

$$\|u\|_{L^\infty(\tau_i)}^2 \leq ch^{-2} \|u\|_{L^2(\tau_i)}^2$$

and

$$\|u\|_{H^1(\tau_i)}^2 \leq ch^{-2} \|u\|_{L^2(\tau_i)}^2.$$

Consequently,

$$K(u, t)^2 \leq cth^{-1} \|u\|_{L^2(\Omega)}^2.$$

It follows that

$$(10.4) \quad \int_0^h K(u, t)^2 t^{-2\beta-1} dt \leq Ch^{-2\beta} \|u\|_{L^2(\Omega)}^2.$$

Combining (10.3) and (10.4) completes the proof of (10.1).

BIBLIOGRAPHY

1. A. K. Aziz and I. Babuška, *Survey lectures on the mathematical foundations of the finite element method*, Part I, The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations (A. K. Aziz, ed.), Academic Press, New York, 1972, pp. 1–362.
2. R. E. Bank and C. C. Douglas, *Sharp estimates for multigrid rates of convergence with general smoothing and acceleration*, SIAM J. Numer. Anal. **22** (1985), 617–633.
3. R. E. Bank and T. Dupont, *An optimal order process for solving finite element equations*, Math. Comp. **36** (1981), 35–51.
4. D. Braess and W. Hackbusch, *A new convergence proof for the multigrid method including the V-cycle*, SIAM J. Numer. Anal. **20** (1983), 967–975.
5. J. H. Bramble and J. E. Pasciak, *New convergence estimates for multigrid algorithms*, Math. Comp. **49** (1987), 311–329.
6. J. H. Bramble, J. E. Pasciak, and J. Xu, *The analysis of multigrid algorithms for nonsymmetric and indefinite elliptic problems*, Math. Comp. **51** (1988), 389–414.
7. J. H. Bramble and J. E. Pasciak, *A preconditioning technique for indefinite systems resulting from mixed approximations of elliptic problems*, Math. Comp. **50** (1988), 1–18.
8. A. Brandt, *Multi-level adaptive solutions to boundary-value problems*, Math. Comp. **31** (1977), 333–390.
9. S. C. Brenner, *An optimal-order multigrid method for P1 nonconforming finite elements*, Math. Comp. **52** (1989), 1–16.
10. F. Brezzi, J. Douglas, Jr., and L. D. Marini, *Two families of mixed finite elements for second order elliptic problems*, Numer. Math. **47** (1985), 217–235.
11. P. L. Butzer and H. Berens, *Semi-groups of operators and approximation*, Springer-Verlag, Berlin and New York, 1967.
12. P. G. Ciarlet, *The finite element method for elliptic problems*, North-Holland, New York, 1978.

13. C. C. Douglas, *Multi-grid algorithms with applications to elliptic boundary-value problems*, SIAM J. Numer. Anal. **21** (1984), 236–254.
14. T. Dupont and R. Scott, *Polynomial approximation of functions in Sobolev spaces*, Math. Comp. **34** (1980), 441–463.
15. R. Glowinski and M. F. Wheeler, *Domain decomposition and mixed methods for elliptic problems*, Proc. 1st Internat. Conf. on Domain Decomposition Methods, SIAM, Philadelphia, PA, 1988, pp. 144–172.
16. W. Hackbusch, *Multi-grid methods and applications*, Springer-Verlag, New York, 1985.
17. R. B. Kellogg, Interpolation between subspaces of a Hilbert space, Technical Note BN-719, Univ. of Maryland, Inst. of Fluid Dynamics and Appl. Math., 1971.
18. S. G. Krein and Y. I. Petunin, *Scales of Banach spaces*, Russian Math. Surveys **21** (1966), 85–160.
19. J. L. Lions and E. Magenes, *Problèmes aux limites non homogènes et applications*, vol. 1, Dunod, Paris, 1968.
20. J. Mandel, S. F. McCormick, and J. Ruge, *An algebraic theory for multigrid methods for variational problems*, preprint.
21. J. Mandel, S. McCormick, and R. Bank, *Variational multigrid theory*, Multigrid Methods, (S. McCormick, ed.), SIAM, Philadelphia, PA, pp. 131–178.
22. S. F. McCormick, *Multigrid methods for variational problems: Further results*, SIAM J. Numer. Anal. **21** (1984), 255–263.
23. S. F. McCormick, *Multigrid methods for variational problems: General theory for the V-cycle*, SIAM J. Numer. Anal. **22** (1985), 634–643.
24. F. A. Milner, *Mixed finite element methods for quasilinear second-order elliptic problems*, Math. Comp. **44** (1985), 303–320.
25. J. Nečas, *Les méthodes directes en théorie des équations elliptiques*, Academia, Prague, 1967.
26. P. A. Raviart and J. M. Thomas, *A mixed finite element method for 2-nd order elliptic problems*, Mathematical Aspects of Finite Element Methods, Lecture Notes in Math., vol. 606 (I. Galligani and E. Magenes, eds.), Springer-Verlag, New York, 1977, pp. 292–315.
27. F. Riesz and B. Sz.-Nagy, *Functional analysis*, Ungar, New York, 1955.
28. E. M. Stein, *Singular integrals and differentiability properties of functions*, Princeton Univ. Press, Princeton, NJ, 1970.
29. R. Temam, *Navier-Stokes equations*, North-Holland, New York, 1977.
30. R. Verfürth, *A multilevel algorithm for mixed problems*, SIAM J. Numer. Anal. **21** (1984), 264–271.
31. J. Xu, *Theory of multilevel methods*, Thesis, Cornell University, Ithaca, NY, 1989.
32. S. Zhang, *Multi-level iterative techniques*, Thesis, Math. Res. Rep. 88020, Penn. State Univ., 1988.

DEPARTMENT OF MATHEMATICS, CORNELL UNIVERSITY, ITHACA, NEW YORK 14853
E-mail address: bramble@mathvax.msi.cornell.edu

BROOKHAVEN NATIONAL LABORATORY, UPTON, NEW YORK 11973
E-mail address: pasciak@bnl.gov

DEPARTMENT OF MATHEMATICS, PENNSYLVANIA STATE UNIVERSITY, UNIVERSITY PARK, PENNSYLVANIA 16802
E-mail address: xu@rayleigh.psu.edu

3.6. THE ANALYSIS OF SMOOTHERS FOR MULTIGRID ALGORITHMS

3.5 Convergence estimates for multigrid algorithms without regularity assumptions

Convergence estimates for multigrid algorithms without regularity assumptions[23]

3.6 The analysis of smoothers for multigrid algorithms

The analysis of smoothers for multigrid algorithms[18]

THE ANALYSIS OF SMOOTHERS FOR MULTIGRID ALGORITHMS

JAMES H. BRAMBLE AND JOSEPH E. PASCIAK

ABSTRACT. The purpose of this paper is to provide a general technique for defining and analyzing smoothing operators for use in multigrid algorithms. The smoothing operators considered are based on subspace decomposition and include point, line, and block versions of Jacobi and Gauss-Seidel iteration as well as generalizations. We shall show that these smoothers will be effective in multigrid algorithms provided that the subspace decomposition satisfies two simple conditions. In many applications, these conditions are trivial to verify.

1. INTRODUCTION

In recent years, multigrid methods have been used extensively as tools for obtaining approximations to the solutions of partial differential equations (see the references in [10, 11, 12]). In conjunction, there has been intensive research into the theoretical understanding of these methods (cf. [1, 2, 3, 4, 6, 9, 11, 12, 13, 14, 18] and others). Many of the above papers present various analyses of multigrid methods which are often based on certain assumptions concerning the smoothing process. These assumptions are sometimes verified for specific examples. It is the purpose of this paper to present a general approach for developing smoothing operators and show that they work in multigrid methods provided that a few simple hypotheses are satisfied in the construction. For other estimates concerning smoothing operators in multigrid procedures, we refer the reader to [16] and the extended bibliography included there.

The smoothers for a given space are defined to be either the additive or multiplicative iterative scheme associated with a decomposition of the space (see (3.2) and Algorithm 3.1). Different smoothers result from distinct decompositions. Depending on the choice of subspaces in this decomposition, the technique can be used to generate many of the popular smoothing schemes used in multigrid iteration. For example, it can be used to generate point, line, and block Jacobi smoothing as well as point, line, and block Gauss-Seidel smoothing.

The construction of iterative schemes based on subspace decomposition is not a new idea. In fact, this technique has been used extensively for the construc-

Received October 8, 1990; revised July 1, 1991.

1991 *Mathematics Subject Classification*. Primary 65N30; Secondary 65F10.

This manuscript has been authored under contract number DE-AC02-76CH00016 with the U.S. Department of Energy. Accordingly, the U.S. Government retains a non-exclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. This work was also supported in part under the National Science Foundation Grant No. DMS84-05352 and by the U.S. Army Research Office through the Mathematical Sciences Institute, Cornell University.

©1992 American Mathematical Society
0025-5718/92 \$1.00 + \$.25 per page

tion of preconditioners using overlapping domain decomposition (also known as Schwarz domain decomposition methods) [8, 20]. The hypotheses on the subspace decomposition required in this paper for multigrid smoothers are easier to satisfy than those required for the construction of effective preconditioners. Thus, the subspace decomposition used in most smoothing procedures will not give rise to a good preconditioner. We illustrate this in the following discussion.

Let \mathcal{M}_k be a finite-dimensional space with inner product $(\cdot, \cdot)_k$ and consider the problem of computing the solution $u \in \mathcal{M}_k$ of

$$A_k u = f,$$

for given $f \in \mathcal{M}_k$. Here, A_k is a symmetric and positive definite operator on \mathcal{M}_k . The additive and multiplicative versions of the smoothers are defined in terms of a decomposition

$$\mathcal{M}_k = \sum_{i=1}^l \mathcal{M}_k^i.$$

The basic hypothesis which is used to show that the resulting smoother is effective in a multigrid iteration is that there exists a constant c_0 such that every $v \in \mathcal{M}_k$ can be decomposed into $v = \sum v_i$ with $v_i \in \mathcal{M}_k^i$ satisfying

$$(1.1) \quad \sum_{i=1}^l (v_i, v_i)_k \leq c_0 (v, v)_k.$$

The corresponding hypothesis which is used to show that the smoother is an effective preconditioner for A_k is that there exists a constant C_0 such that every $v \in \mathcal{M}_k$ can be decomposed into $v = \sum v_i$ with $v_i \in \mathcal{M}_k^i$ satisfying

$$(1.2) \quad \sum_{i=1}^l (A_k v_i, v_i)_k \leq C_0 (A_k v, v)_k.$$

In finite element discretization of second-order elliptic partial differential equations, for functions in \mathcal{M}_k , $(\cdot, \cdot)_k$ is often equivalent to the L^2 inner product on the domain of consideration. In contrast, $(A_k \cdot, \cdot)_k$ is usually equivalent to the norm on the Sobolev space of order one. Condition (1.1) is often trivial to verify for many subspace decompositions. Most subspace decompositions which are used as multigrid smoothers (and satisfy (1.1) with bounded c_0) satisfy (1.2) only with a constant C_0 which grows large as the mesh parameter becomes small. Thus, the multigrid smoother would not be effective as a stand-alone iterative method.

The outline of the remainder of this paper is as follows. Section 2 describes the basic multigrid algorithm in an abstract setting and gives some of the conditions on the smoothers which are commonly assumed in various multigrid analyses. The general smoothing procedures based on subspace decomposition are described and analyzed in §3. In §4, we give theorems providing estimates for multigrid algorithms using these smoothers. Computer implementation of the smoothers is discussed in §5. In particular, it is shown that the commonly used multigrid smoothers can be generated by this technique with appropriate selection of the subspace decompositions. Finally, in §6, we discuss the finite element multigrid application.

2. THE MULTIGRID ALGORITHMS

In this section, we describe a symmetric multigrid cycling algorithm. For convenience, this algorithm is developed in an abstract Hilbert space setting and uses general smoothing operators. We then state two conditions involving these smoothing operators which are assumed for various analyses of multigrid.

We start by describing the general multigrid algorithm in an abstract setting. We assume that we are given a sequence of finite-dimensional inner product spaces

$$\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_j.$$

The inner product on \mathcal{M}_k will be denoted by $(\cdot, \cdot)_k$. In addition, we assume that symmetric positive definite operators $A_k : \mathcal{M}_k \mapsto \mathcal{M}_k$ for $k = 1, \dots, j$ and “interpolation” operators $I_k : \mathcal{M}_{k-1} \mapsto \mathcal{M}_k$ are given. The multigrid algorithm gives rise to iterative procedures for the solution of the problem on \mathcal{M}_j , i.e., given $f \in \mathcal{M}_j$ find $u \in \mathcal{M}_j$ satisfying

$$(2.1) \quad A_j u = f.$$

The final ingredient needed to define the general multigrid algorithm is a sequence of linear (smoothing) operators $R_k : \mathcal{M}_k \mapsto \mathcal{M}_k$, for $k = 2, \dots, j$. We shall always take $R_1 = A_1^{-1}$. The point of this paper is to present a general approach for the definition of these operators as well as a unified analysis for showing that these operators are effective in multigrid procedures. We set

$$R_k^{(l)} = \begin{cases} R_k & \text{if } l \text{ is odd,} \\ R_k' & \text{if } l \text{ is even.} \end{cases}$$

Here, and throughout this manuscript, t will denote adjoint with respect to the inner product $(\cdot, \cdot)_k$.

We next define a general multigrid process for iteratively computing the solution of (2.1). This process is defined in the following algorithm in terms of a mathematical induction involving the subspace level. On each subspace \mathcal{M}_k , the multigrid iterative procedure can be viewed as a process which acts on both a function $F_k \in \mathcal{M}_k$ and an “approximation” W_k to the solution of

$$(2.2) \quad A_k U_k = F_k$$

and produces an improved approximation in \mathcal{M}_k to U_k (denoted by $\text{Mg}_k(W_k, F_k)$).

Algorithm 2.1. For $k = 1$, define $\text{Mg}_1(W_1, F_1) = A_1^{-1}F_1$, i.e., solve (2.2) exactly. For $k > 1$, $\text{Mg}_k(W_k, F_k)$ is defined in terms of $\text{Mg}_{k-1}(\cdot, \cdot)$ as follows:

- (1) Set $X^0 = W_k$ and $Q^0 = 0 \in \mathcal{M}_{k-1}$.
- (2) Define X^i , for $i = 1, \dots, m(k)$ by

$$(2.3) \quad X^i = X^{i-1} + R_k^{(i+m(k))}(F_k - A_k X^{i-1}).$$

- (3) Set $Y^{m(k)} = X^{m(k)} + I_k Q^p$, where Q^i for $i = 1, \dots, p$ is given by

$$(2.4) \quad Q^i = \text{Mg}_{k-1}(Q^{i-1}, P_{k-1}^0(F_k - A_k X^m)).$$

Here, P_{k-1}^0 is defined by

$$(2.5) \quad (P_{k-1}^0 v, \phi)_{k-1} = (v, I_k \phi)_k \quad \text{for all } \phi \in \mathcal{M}_{k-1}.$$

(4) For $i = m(k) + 1, \dots, 2m(k)$ define Y^i by

$$Y^i = Y^{i-1} + R_k^{(i+m(k))}(F_k - A_k Y^{i-1}).$$

(5) Set $\mathbf{Mg}_k(W_k, F_k) = Y^{2m(k)}$.

The above algorithm is more general than those often described [2, 4, 12], in that it allows the use of general symmetric and nonsymmetric smoothers. Note that we have placed very few restrictions on the definition of the linear operators A_k , I_k , and R_k at this time.

The above multigrid procedure can be used to solve (2.1) by the following iteration,

$$(2.6) \quad u^l = \mathbf{Mg}_j(u^{l-1}, f),$$

for initial iterate u^0 and $l = 1, 2, \dots$. Let $e^l = u - u^l$. Then it is possible to show that

$$e^l = E_j e^{l-1}$$

holds for a linear error reduction operator $E_j : \mathcal{M}_j \mapsto \mathcal{M}_j$. Accordingly, we can define a preconditioner B_j associated with the multigrid process by

$$B_j = (I - E_j)A_j^{-1} \quad \text{or} \quad E_j = I - B_j A_j.$$

Consequently, (2.6) is nothing more than the preconditioned linear iterative scheme

$$(2.7) \quad u^l = u^{l-1} + B_j(f - A_j u^{l-1}).$$

Alternatively, the linear operator B_j can be directly defined by the following algorithm.

Algorithm 2.2. Set $B_1 = A_1^{-1}$. Assume that B_{k-1} has been defined and define $B_k g$ for $g \in \mathcal{M}_k$ as follows:

(1) Set $x^0 = 0$ and $q^0 = 0$.

(2) Define x^i for $i = 1, \dots, m(k)$ by

$$(2.8) \quad x^i = x^{i-1} + R_k^{(i+m(k))}(g - A_k x^{i-1}).$$

(3) Define $y^{m(k)} = x^{m(k)} + I_k q^p$, where q^i for $i = 1, \dots, p$ is defined by

$$(2.9) \quad q^i = q^{i-1} + B_{k-1}[P_{k-1}^0(g - A_k x^{m(k)}) - A_{k-1} q^{i-1}].$$

(4) Define y^i for $i = m(k) + 1, \dots, 2m(k)$ by

$$(2.10) \quad y^i = y^{i-1} + R_k^{(i+m(k))}(g - A_k y^{i-1}).$$

(5) Set $B_k g = y^{2m(k)}$.

In the above algorithm, we alternate between R_k and R_k^t in Step 2. In Step 4, we use the adjoints of the Step 2 smoothings applied in the reverse order. This results in a symmetric operator B_j . This form of the multigrid algorithm has been suggested in [5]. Nonsymmetric multigrid procedures which, for example,

do not include the smoothing of Step 4 have also been analyzed [4, 11, 12], etc. The hypotheses required on the smoothing operators are exactly the same as those used in the symmetric case.

There are two standard conditions concerning the smoothing operators which are often assumed as hypotheses in the analysis of multigrid algorithms. To describe these, we first define $K_k = I - R_k A_k$ and note that $K_k^* = I - R_k^t A_k$. Here, and throughout this paper, $*$ will denote adjoint with respect to the inner product $(A_k \cdot, \cdot)_k$.

- (C.1) There is a constant C_R which does not depend on k such that the smoothing procedure satisfies

$$(2.11) \quad \frac{\|u\|_k^2}{\lambda_k} \leq C_R (\bar{R}_k u, u)_k \quad \text{for all } u \in \mathcal{M}_k.$$

Here, \bar{R}_k is either $(I - K_k^* K_k) A_k^{-1}$ or $(I - K_k K_k^*) A_k^{-1}$, λ_k is the largest eigenvalue of A_k , and $\|\cdot\|_k^2$ denotes the norm corresponding to the inner product $(\cdot, \cdot)_k$.

- (C.2) Let $T_k = R_k A_k$. There is a constant $\theta < 2$ not depending on k satisfying

$$(2.12) \quad (A_k T_k v, T_k v)_k \leq \theta (A_k T_k v, v)_k.$$

The point of the present paper is to define general smoothing procedures and prove estimates of the form of (2.11) and (2.12) under simple hypotheses.

We shall state some convergence estimates from [5, 7], and [9] for Algorithm 2.2 in a later section. The following remarks show that the above two conditions are used in other multigrid theories as well.

Remark 2.1. Let $R_{k,\omega}$ correspond to the Richardson smoothing iteration defined by $R_{k,\omega} = \omega \lambda_k^{-1} I$ and $K_{k,\omega} = (I - R_{k,\omega} A_k)$ be the corresponding reducer. Inequality (2.11) in the case of $\bar{R}_k = (I - K_k^* K_k) A_k^{-1}$ can be rewritten as

$$(2.13) \quad (A_k K_k u, K_k u)_k \leq (A_k K_{k,\omega} u, u)_k \quad \text{for all } u \in \mathcal{M}_k,$$

with $\omega = 1/C_R$. This means that the smoothing process applied to any $u \in \mathcal{M}_k$ converges as fast as Richardson's method for some $\omega \in (0, 1)$. A hypothesis of the form of the above inequality was essentially used in [5, 14] and [15]. The Richardson method is perhaps the most natural smoothing procedure.

Remark 2.2. The following condition on the smoothing operator is used by Bank et al. (see [12, (4.6)]): There is a positive constant c satisfying

$$(2.14) \quad \frac{\|A_k K_k u\|_k^2}{\lambda_k} \leq c (A_k (I - K_k^* K_k) u, u)_k \quad \text{for all } u \in \mathcal{M}_k.$$

Note that for the appropriate definition of \bar{R}_k , (2.11) can be rewritten as

$$(2.15) \quad \frac{\|A_k v\|_k^2}{\lambda_k} \leq C_R (A_k (I - K_k K_k^*) v, v)_k \quad \text{for all } v \in \mathcal{M}_k.$$

We shall show that (2.15) implies (2.14). Indeed, taking $v = K_k u$ in (2.15) gives

$$\frac{\|A_k K_k u\|_k^2}{\lambda_k} \leq C_R (A_k (I - \bar{K}_k) \bar{K}_k u, u)_k$$

where $\bar{K}_k = K_k^* K_k$. Note that the spectrum of \bar{K}_k is in $[0,1)$, and hence

$$(A_k(I - \bar{K}_k)\bar{K}_k u, u)_k \leq (A_k(I - \bar{K}_k)u, u)_k = (A_k(I - K_k^* K_k)u, u)_k.$$

Combining the above two estimates proves (2.14).

3. GENERAL SMOOTHING PROCEDURES IN MULTIGRID ALGORITHMS

In this section, we shall define smoothing operators in terms of subspace decompositions. These procedures are related to overlapping domain decomposition and the classical Schwarz method and are generalizations of Jacobi and Gauss-Seidel iteration procedures. In this section, we shall show that the hypotheses (2.11) and (2.12) will follow from scaling, when appropriate, and an easily verified function decomposition inequality. Explicit examples providing such decompositions in the case of finite element multigrid applications are given in a later section.

The technique which we shall study for developing smoothers involves the use of a variant of overlapping domain decomposition. These methods are also referred to as "Schwarz overlapping" methods. We shall develop a smoother for the problem on \mathcal{M}_k . One starts with a decomposition of the space,

$$(3.1) \quad \mathcal{M}_k = \sum_{i=1}^l \mathcal{M}_k^i.$$

This sum may or may not be a direct sum.

Given the decomposition (3.1), there are two types of smoothers which can be defined. The first will be called the additive smoother and is defined by

$$(3.2) \quad R_k = \gamma \sum_{i=1}^l A_{k,i}^{-1} Q_k^i.$$

Here, $A_{k,i} : \mathcal{M}_k^i \mapsto \mathcal{M}_k^i$ is defined by

$$(A_{k,i}v, \chi)_k = (A_k v, \chi)_k \quad \text{for all } \chi \in \mathcal{M}_k^i$$

and $Q_k^i : \mathcal{M}_k \mapsto \mathcal{M}_k^i$ is the projection onto \mathcal{M}_k^i with respect to the inner product $(\cdot, \cdot)_k$. In addition, γ is a positive scaling factor which will be chosen later. We note that R_k is a symmetric operator with respect to the inner product $(\cdot, \cdot)_k$. Implementation issues involving the above smoother will be discussed in a later section.

To analyze the additive algorithm, we shall use a limited interaction hypothesis. To describe this property, we first introduce the projection $P_k^i : \mathcal{M}_k \mapsto \mathcal{M}_k^i$ defined by

$$(A_k P_k^i v, \chi)_k = (A_k v, \chi)_k \quad \text{for all } \chi \in \mathcal{M}_k^i.$$

We then define

$$\kappa_{im} = \begin{cases} 0 & \text{if } P_k^i P_k^m = 0, \\ 1 & \text{otherwise,} \end{cases}$$

and set

$$n_0 = \max_i \sum_{m=1}^l \kappa_{im}.$$

In our applications, n_0 remains small, even when l becomes large. Note that the matrix $\{\kappa_{im}\}$ is symmetric.

We shall use the following two conditions:

- (1) The subspaces satisfy a limited interaction property, i.e.,

$$(3.3) \quad n_0 \leq c_1,$$

with c_1 independent of k .

- (2) There exists a positive constant c_0 not depending on k such that for each $u \in \mathcal{M}_k$, there is a decomposition $u = \sum_{i=1}^l u_i$ with $u_i \in \mathcal{M}_k^i$ satisfying

$$(3.4) \quad \sum_{i=1}^l \|u_i\|_k^2 \leq c_0 \|u\|_k^2.$$

In applications, the above conditions are often trivial to verify. Moreover, under these hypotheses, we can prove the following theorem.

Theorem 3.1. *Let R_k be defined by (3.2) and assume that (3.3) and (3.4) are satisfied. Let $\theta \in (0, 2)$ and set $\gamma = \theta/c_1$. Then (2.12) holds, and (2.11) holds with $C_R = c_0 c_1 / [\theta(2 - \theta)]$.*

Before proving the theorem, we prove the following lemma [8].

Lemma 3.1. *Let n_0 be defined as above and $u_i, v_i \in \mathcal{M}_k^i$ for $i = 1, \dots, l$. Then*

$$(3.5) \quad \left(\sum_{i,m=1}^l |(A_k u_i, v_m)_k| \right)^2 \leq n_0^2 \sum_{i=1}^l (A_k u_i, u_i)_k \sum_{m=1}^l (A_k v_m, v_m)_k.$$

Proof. We note that

$$\begin{aligned} \left(\sum_{i,m=1}^l |(A_k u_i, v_m)_k| \right)^2 &= \left(\sum_{i,m=1}^l \kappa_{im} |(A_k u_i, v_m)_k| \right)^2 \\ &\leq \sum_{i,m=1}^l \kappa_{im} (A_k u_i, u_i)_k \sum_{i,m=1}^l \kappa_{im} (A_k v_m, v_m)_k \\ &\leq n_0^2 \sum_{i=1}^l (A_k u_i, u_i)_k \sum_{m=1}^l (A_k v_m, v_m)_k. \end{aligned}$$

This completes the proof of the lemma.

Proof of Theorem 3.1. We first show that

$$(3.6) \quad \frac{\|u\|_k^2}{\lambda_k} \leq c_0 \gamma^{-1} (R_k u, u)_k \quad \text{for all } u \in \mathcal{M}_k.$$

Let $u = \sum_{i=1}^l u_i$ be the decomposition of (3.4). Then

$$\begin{aligned} \|u\|_k^2 &= \sum_{i=1}^l (u_i, Q_k^i u)_k \leq \left(\sum_{i=1}^l \|u_i\|_k^2 \right)^{1/2} \left(\sum_{i=1}^l \|Q_k^i u\|_k^2 \right)^{1/2} \\ &\leq c_0^{1/2} \|u\|_k \left(\sum_{i=1}^l \|Q_k^i u\|_k^2 \right)^{1/2}. \end{aligned}$$

Hence,

$$(3.7) \quad \|u\|_k^2 \leq c_0 \sum_{i=1}^l \|Q_k^i u\|_k^2.$$

Now, it is immediate from the definitions that the largest eigenvalue of the operator $A_{k,i}$ is bounded by λ_k . Consequently,

$$(3.8) \quad \sum_{i=1}^l \|Q_k^i u\|_k^2 \leq \lambda_k \sum_{i=1}^l (A_{k,i}^{-1} Q_k^i u, u)_k = \lambda_k \gamma^{-1} (R_k u, u)_k.$$

Combining (3.7) and (3.8) proves (3.6).

We will show that the spectral radius of $T_k = R_k A_k$ is less than or equal to θ provided that we take $\gamma = \theta/c_1$. Let us temporarily assume this. Note, that R_k is a symmetric operator in the $(\cdot, \cdot)_k$ inner product. By (3.6), it is also positive definite, and hence its square root is well defined. We then have for $u \in \mathcal{M}_k$,

$$(3.9) \quad \begin{aligned} (\bar{R}_k u, u)_k &= ((2R_k - R_k A_k R_k)u, u)_k \\ &= ((2I - R_k^{1/2} A_k R_k^{1/2})R_k^{1/2} u, R_k^{1/2} u)_k \\ &\geq (2 - \theta)(R_k u, u)_k. \end{aligned}$$

The theorem follows combining (3.6) and (3.9), once we provide the desired estimate for the spectral radius of $R_k A_k$.

Let $v_i = P_k^i u$. By the definition of R_k and the identity $Q_k^i A_k = A_{k,i} P_k^i$,

$$(3.10) \quad \begin{aligned} (A_k R_k A_k u, u)_k &= \gamma \sum_{i=1}^l (A_k v_i, v_i)_k = \gamma \left(A_k \sum_{i=1}^l v_i, u \right)_k \\ &\leq \gamma \left(\left(A_k \sum_{i=1}^l v_i, \sum_{i=1}^l v_i \right)_k \right)^{1/2} ((A_k u, u)_k)^{1/2}. \end{aligned}$$

Applying Lemma 3.1 proves the desired bound, i.e.,

$$(3.11) \quad (A_k R_k A_k u, u)_k \leq \theta (A_k u, u)_k.$$

This completes the proof of the theorem.

Remark 3.1. The overlapping domain decomposition techniques (e.g. (3.2)) can be used directly to develop preconditioners for the operator A_k (see [8]). However, in this case the subspaces must be chosen in a much more restricted way. To prove that the additive preconditioner (3.2) provides a good preconditioner, one replaces (3.4) by the existence of a decomposition satisfying

$$(3.12) \quad \sum_{i=1}^l (A_k u_i, u_i)_k \leq C_0 (A_k u, u)_k.$$

As we shall see from later examples, it is much easier to construct subspaces satisfying (3.4). In general, the subspaces used for developing smoothers will not satisfy (3.12).

Remark 3.2. An obvious alternative to hypothesis (3.3) in the case of Theorem 3.1 is the assumption that

$$(3.13) \quad \sum_{i=1}^l (A_k P_k^i u, P_k^i u)_k \leq c_1 (A_k u, u)_k \quad \text{for all } u \in \mathcal{M}_k.$$

With such an assumption, (3.11) follows immediately from the first equality of (3.10) and provides a simpler proof. However, the limited interaction condition was introduced because it is also used for the analysis of the multiplicative algorithms to be subsequently described.

Remark 3.3. When developing preconditioners (instead of smoothers), it is often important to include a “coarse” subspace \mathcal{M}_k^0 which interacts with all of the other subspaces, i.e., $\kappa_{0k} \neq 0$ for all k . This is not necessary in the case of smoothers. However, it would still be possible to analyze the resulting algorithm using the above arguments and those presented in [8].

We define the multiplicative smoother based on the above subspace decomposition of \mathcal{M}_k in the following algorithm.

Algorithm 3.1. Let $f \in \mathcal{M}_k$. We define $R_k f \in \mathcal{M}_k$ as follows:

- (1) Set $v_0 = 0$.
- (2) Define v_i for $i = 1, \dots, l$ by

$$(3.14) \quad v_i = v_{i-1} + A_{k,i}^{-1} Q_k^i (f - A_k v_{i-1}).$$

- (3) Set $R_k f = v_l$.

It immediately follows from the identity $A_{k,i} P_k^i = Q_k^i A_k$ that

$$(3.15) \quad K_k = (I - P_k^l) \cdots (I - P_k^1).$$

That is, the error propagator associated with the smoother defined by Algorithm 3.1 is a product of orthogonal projections onto the complements of the subspaces. The next theorem provides an estimate for (2.11) and (2.12) with this definition of R_k .

Theorem 3.2. Let R_k be defined by Algorithm 3.1 and assume that (3.3) and (3.4) hold. Then (2.11) holds with

$$(3.16) \quad C_R = (2c_0(1 + c_1^2)).$$

In addition, (2.12) holds with $\theta = 2c_1/(c_1 + 1)$.

Proof. The proof of this theorem uses techniques of [8]. First, we define the operator

$$(3.17) \quad E_i = (I - P_k^i)(I - P_k^{i-1}) \cdots (I - P_k^1)$$

for $i = 1, \dots, l$. For convenience, we let $E_0 = I$ and note that $E_l = K_k$.

We will prove inequality (2.11) for $\bar{R}_k = (I - K_k^* K_k) A_k^{-1}$ by proving the equivalent inequality (2.13). Note that E_j^* is obtained by reversing the order of the factors in (3.17). With this observation, it is possible to use the same proof for the case of $\bar{R}_k = (I - K_k K_k^*) A_k^{-1}$.

We shall first derive some identities involving the above operators. We clearly have for $i = 1, \dots, l$,

$$(3.18) \quad E_{i-1} - E_i = P_k^i E_{i-1},$$

from which it follows that

$$(3.19) \quad I - E_i = \sum_{m=1}^i P_k^m E_{m-1}.$$

It is obvious from (3.18) that for $v \in \mathcal{M}_k$,

$$(3.20) \quad (A_k E_{i-1} v, E_{i-1} v)_k - (A_k E_i v, E_i v)_k = (A_k P_k^i E_{i-1} v, P_k^i E_{i-1} v)_k.$$

Summing (3.20) gives that

$$(3.21) \quad (A_k v, v)_k - (A_k E_l v, E_l v)_k = \sum_{i=1}^l (A_k P_k^i E_{i-1} v, E_{i-1} v)_k.$$

We note that (2.13) can be rewritten as

$$(3.22) \quad \omega \lambda_k^{-1} \|A_k v\|_k^2 \leq (A_k v, v)_k - (A_k E_l v, E_l v)_k.$$

But, by (3.7),

$$\begin{aligned} \|A_k v\|_k^2 &\leq c_0 \sum_{i=1}^l \|Q_k^i A_k v\|_k^2 \\ &= c_0 \sum_{i=1}^l (A_k, i P_k^i v, A_k, i P_k^i v)_k \leq c_0 \lambda_k \sum_{i=1}^l (A_k P_k^i v, P_k^i v)_k. \end{aligned}$$

Hence, (3.22) will follow if we can show that

$$(3.23) \quad \sum_{i=1}^l (A_k P_k^i v, P_k^i v)_k \leq 2(1 + c_1^2) \sum_{i=1}^l (A_k P_k^i E_{i-1} v, E_{i-1} v)_k.$$

It is shown in [8] that (3.23) holds under assumption (3.3) (cf. inequality (2.23) of [8]).

We include the proof of (3.23) for completeness. By (3.19),

$$(3.24) \quad (A_k P_k^i v, v)_k = (A_k P_k^i v, E_{i-1} v)_k + \sum_{m=1}^{i-1} (A_k P_k^i v, P_k^m E_{m-1} v)_k.$$

Summing gives

$$\sum_{i=1}^l (A_k P_k^i v, P_k^i v)_k = \sum_{i=1}^l (A_k P_k^i v, P_k^i E_{i-1} v)_k + \sum_{i=1}^l \sum_{m=1}^{i-1} (A_k P_k^i v, P_k^m E_{m-1} v)_k.$$

Thus, by the arithmetic-geometric mean inequality and Lemma 3.1,

$$\begin{aligned} \left(\sum_{i=1}^l (A_k P_k^i v, P_k^i v)_k \right)^2 &\leq 2 \left\{ \sum_{i=1}^l (A_k P_k^i v, P_k^i v)_k \sum_{i=1}^l (A_k P_k^i E_{i-1} v, E_{i-1} v)_k \right. \\ &\quad \left. + \left[\sum_{i, m=1}^l |(A_k P_k^i v, P_k^m E_{m-1} v)_k| \right]^2 \right\} \\ &\leq 2(1 + n_0^2) \sum_{i=1}^l (A_k P_k^i v, P_k^i v)_k \sum_{i=1}^l (A_k P_k^i E_{i-1} v, E_{i-1} v)_k. \end{aligned}$$

This completes the proof of (3.23).

Finally, we provide the estimate for θ . Note that for $u \in \mathcal{M}_k$, by Lemma 3.1,

$$\begin{aligned} (A_k T_k u, T_k u)_k &= (A_k(I - E_l)u, (I - E_l)u)_k = \sum_{i,m=1}^l (A_k P_k^i E_{i-1} u, P_k^m E_{m-1} u)_k \\ &\leq n_0 \sum_{i=1}^l (A_k P_k^i E_{i-1} u, E_{i-1} u)_k. \end{aligned}$$

Applying (3.21) gives

$$\begin{aligned} (3.25) \quad (A_k T_k u, T_k u)_k &\leq n_0[(A_k u, u)_k - (A_k E_l u, E_l u)_k] \\ &= n_0[2(A_k T_k u, u)_k - (A_k T_k u, T_k u)_k]. \end{aligned}$$

This shows that (2.12) holds for $\theta \leq 2n_0/(n_0 + 1)$ and hence completes the proof of the theorem.

Remark 3.4. We note that Theorem 3.1 provides the estimate $C_R = c_0 c_1$ for (2.11) when $\theta = 1$ and the smoother is defined by (3.2). In contrast, Theorem 3.2 provides the estimate $C_R = 2c_0(1 + c_1^2)$ when Algorithm 3.1 is used. This suggests that the additive version may work better in practice. As far as we know, this is not the case. In all of the examples which we have considered, numerical evidence suggests that the multiplicative smoother always works slightly better than the additive smoother using the same subspaces.

4. CONVERGENCE ESTIMATES FOR MULTIGRID ALGORITHMS

In this section, we apply the results of the theorems of the previous section to get convergence for multigrid Algorithm 2.2. We make no attempt to survey all possible applications but, instead, provide the theorems to illustrate the type of convergence results available utilizing the estimates on the smoothing operators provided by Theorems 3.1 and 3.2. Modifying a proof given in [9], we also provide a “no-regularity” convergence estimate in the case of the product smoothing operator defined by Algorithm 3.1.

As observed earlier, the multigrid process gives rise to the iterative reduction matrix $I - B_j A_j$, where B_j is given by Algorithm 2.2. Thus, bounds for the iterative convergence rate of either (2.6) or (2.7) follow from norm estimates for the operator $I - B_j A_j$. Alternatively, one can use the operator B_j directly in a preconditioned iteration for the solution of (2.1). Since B_j is symmetric in the inner product $(\cdot, \cdot)_j$ (cf. [5]), bounds for preconditioned iterative schemes follow from estimates for the condition number $K(B_j A_j)$, which is defined to be the ratio of the largest eigenvalue of $B_j A_j$ to the smallest.

We start by illustrating the convergence and preconditioning results for Algorithm 2.2 under the following regularity and approximation hypothesis: There exists a fixed number $\alpha \in (0, 1]$ and a positive constant C_α which does not depend on A_k such that for $k = 2, \dots, j$

$$(4.1) \quad |(A_k(I - I_k P_{k-1})u, u)_k| \leq C_\alpha^2 \left(\frac{\|A_k u\|_k^2}{\lambda_k} \right)^\alpha (A_k u, u)_k^{1-\alpha} \quad \text{for all } u \in \mathcal{M}_k.$$

Here $P_{k-1}: \mathcal{M}_k \mapsto \mathcal{M}_{k-1}$ is defined by $P_{k-1}v = w$, where w is the unique function in \mathcal{M}_{k-1} satisfying

$$(A_{k-1}w, \phi)_{k-1} = (A_k v, I_k \phi)_k \quad \text{for all } \phi \in \mathcal{M}_{k-1}.$$

The following two theorems are a consequence of Theorems 3.1 and 3.2 and the results in [5]. The first gives estimates for the reduction operator $I - B_j A_j$ in the norm $\| \cdot \|_j = ((A_j \cdot, \cdot)_j)^{1/2}$. The second gives estimates for the condition number $K(B_j A_j)$.

Theorem 4.1. *Let R_k be defined by either (3.2) or Algorithm 3.1 and assume that (4.1), (3.3), and (3.4) hold. Furthermore, assume that*

$$(A_k I_k v, I_k v)_k \leq (A_{k-1} v, v)_{k-1} \quad \text{for all } v \in \mathcal{M}_{k-1}.$$

Let B_j be defined by Algorithm 2.2 with $p = 1$ and $m(k) = m$ for all k . Then

$$(4.2) \quad \| (I - B_j A_j) v \|_j \leq \delta \| v \|_j \quad \text{for all } v \in \mathcal{M}_j,$$

where

$$\delta = \frac{M_\alpha j^{(1-\alpha)/\alpha}}{M_\alpha j^{(1-\alpha)/\alpha} + m^\alpha}.$$

If, instead, $m(k)$ satisfies

$$(4.3) \quad \beta_0 m(k) \leq m(k-1) \leq \beta_1 m(k)$$

(β_0 and β_1 are constants which are greater than one and independent of k), then (4.2) holds with

$$\delta = \frac{M_\alpha}{M_\alpha + m(j)^\alpha}.$$

The constant M_α above is independent of j .

Theorem 4.2. *Let R_k be defined by either (3.2) or Algorithm 3.1 and assume that (4.1), (3.3), and (3.4) hold. Let B_j be defined by Algorithm 2.2 with $p = 1$ and $\{m(k)\}$ satisfying (4.3). Then $K(B_j A_j) \leq \eta_1 / \eta_0$, where η_0 and η_1 are given by*

$$\eta_0 = \frac{m(j)^\alpha}{M_\alpha + m(j)^\alpha}$$

and

$$\eta_1 = \frac{M_\alpha + m(j)^\alpha}{m(j)^\alpha},$$

i.e., the system $B_j A_j$ is well conditioned independently of j .

Multigrid is often applied to sequences of operators approximating the solution of an elliptic partial differential equation. In this case, the validity of (4.1) is inherently related to the regularity properties of solutions of this partial differential equation. Alternative hypotheses which avoid these regularity assumptions have been used to provide multigrid results (see [7, 8, 9]). These are as follows:

- (1) The subspaces $\mathcal{M}_1, \dots, \mathcal{M}_j$ are nested and the operators are inherited, i.e., $\mathcal{M}_{k-1} \subseteq \mathcal{M}_k$ and

$$(4.4) \quad (A_k v, v)_k = (A_j v, v)_j \quad \text{for all } v \in \mathcal{M}_k.$$

- (2) There exists a sequence of linear operators $Q_k: \mathcal{M}_j \mapsto \mathcal{M}_k$ for $k = 1, \dots, j$, with $Q_j = I$ satisfying the following properties. There are constants C_1 and C_2 not depending on k for which

$$(4.5) \quad \begin{aligned} \|(Q_k - Q_{k-1})u\|_k^2 &\leq C_1 \lambda_k^{-1} A(u, u) \quad \text{for } k = 2, \dots, j, \\ A(Q_k u, Q_k u) &\leq C_2 A(u, u) \quad \text{for } k = 1, \dots, j-1. \end{aligned}$$

The inequalities in (4.5) hold for all $u \in \mathcal{M}_j$. Inequalities of the form of (4.5) can be verified without the use of elliptic regularity estimates (see [7]).

The hypotheses required for the smoother in the case of the “regularity-free” estimates are less stringent. Loosely, the smoother R_k need only “smooth” on a subspace of \mathcal{M}_k containing the image of $Q_k - Q_{k-1}$. To this end, let $\widetilde{\mathcal{M}}_k$ be a subspace of \mathcal{M}_k which contains the range of the operator $Q_k - Q_{k-1}$. Assume that we are given a decomposition

$$(4.6) \quad \widetilde{\mathcal{M}}_k = \sum_{i=1}^l \mathcal{M}_k^i$$

satisfying assumptions (3.3) and (3.4) (with $\widetilde{\mathcal{M}}_k$ replacing \mathcal{M}_k). Let $\widetilde{R}_k: \widetilde{\mathcal{M}}_k \mapsto \widetilde{\mathcal{M}}_k$ be defined by either (3.2) (with $\gamma = 1/c_1$) or Algorithm 3.1 using these spaces. In addition, set $R_k = \widetilde{R}_k \tilde{P}_k^0: \mathcal{M}_k \mapsto \widetilde{\mathcal{M}}_k$, where \tilde{P}_k^0 denotes the $(\cdot, \cdot)_k$ orthogonal projection onto $\widetilde{\mathcal{M}}_k$. Note that Theorems 3.1 and 3.2 provide estimates for a constant C_R satisfying

$$(4.7) \quad \frac{\|u\|_k^2}{\lambda_k} \leq C_R (\overline{R}_k u, u)_k \quad \text{for all } u \in \widetilde{\mathcal{M}}_k.$$

Here,

$$(4.8) \quad \overline{R}_k = \widetilde{R}_k + \widetilde{R}_k^t - \widetilde{R}_k^t \widetilde{A}_k \widetilde{R}_k,$$

where $\widetilde{A}_k: \widetilde{\mathcal{M}}_k \mapsto \widetilde{\mathcal{M}}_k$ is defined by

$$(\widetilde{A}_k v, \phi)_k = (A_k v, \phi)_k \quad \text{for all } \phi \in \widetilde{\mathcal{M}}_k.$$

The following theorem provides estimates for the rate of convergence of the multigrid algorithm with this R_k under the above assumptions. The proof in the case of (3.2) follows directly from results in [9] and Theorem 3.1. The proof of the theorem in the case of Algorithm 3.1 is a modification to that given in [9]. A somewhat more restricted result in the case of nonsymmetric R_k (also based on [9]) was given in [19].

Theorem 4.3. *Assume that (4.4) and (4.5) hold. Let R_k be defined as above and assume that (3.3) and (3.4) hold with $\widetilde{\mathcal{M}}_k$ replacing \mathcal{M}_k . Let B_j be defined by either Algorithm 2.2 or the nonsymmetric smoothing (with corresponding operator denoted by B_j^n) version (see [9]). Then (4.2) holds with $\delta = \delta_j^2$, where*

$$(4.9) \quad \delta_j = 1 - \frac{1}{C(j-1)}$$

and $C = [(1 + C_2^{1/2})(2c_1)^{1/2} + (C_R C_1)^{1/2}]^2$. The constant C_R satisfies (4.7) and is provided by either Theorem 3.1 or Theorem 3.2. In the case of the nonsymmetric smoothing version,

$$(4.10) \quad |||(I - B_j^n A_j)v|||_j \leq \delta_j |||v|||_j \quad \text{for all } v \in \mathcal{M}_j.$$

Proof. For the purpose of this proof, we shall let $A(\cdot, \cdot) = (A_j \cdot, \cdot)_j$. We need only prove the result in the case of nonsymmetric R_k . Moreover, we shall prove the result for $p = 1$. The results for higher p follow from arguments given in [9].

First of all, it was observed in [9] that under the above assumptions,

$$(I - B_j A_j) = (I - B_j^n A_j)^* (I - B_j^n A_j),$$

where B_j^n denotes the multigrid operator which involves smoothing only before correction. Set

$$\bar{K}_k^{(m(k))} = \begin{cases} (K_k^* K_k)^{m(k)/2} & \text{if } m(k) \text{ is even,} \\ (K_k^* K_k)^{(m(k)-1)/2} K_k^* & \text{if } m(k) \text{ is odd.} \end{cases}$$

It was also observed in [9] that for $T_k = (I - (\bar{K}_k^{(m(k))})^*) P_k$,

$$(4.11) \quad (I - B_j^n A_j)^* = (I - T_j)(I - T_{j-1}) \cdots (I - T_1).$$

We use a product analysis similar to that used in Theorem 3.2 and also Theorem 1 of [9]. To this end, we set $E_0 = I$ and

$$E_k = (I - T_k)(I - T_{k-1}) \cdots (I - T_1) = (I - T_k)E_{k-1}.$$

As in the proof of Theorem 3.2 (compare with (3.21)),

$$(4.12) \quad \begin{aligned} A(u, u) - A(E_j u, E_j u) &= \sum_{k=1}^j [A(E_{k-1} u, E_{k-1} u) - A(E_k u, E_k u)] \\ &= \sum_{k=1}^j A((2I - T_k)E_{k-1} u, T_k E_{k-1} u). \end{aligned}$$

Note that $I - B_j^n A_j = E_j^*$, and hence inequalities (4.2) and (4.10) will follow if we can show that

$$(4.13) \quad \begin{aligned} A(u, u) &\leq C(j-1)[A(u, u) - A(E_j u, E_j u)] \\ &= C(j-1) \sum_{k=1}^j A((2I - T_k)E_{k-1} u, T_k E_{k-1} u). \end{aligned}$$

Proceeding as in [9], we use the fact that $Q_j = I$ and write

$$u = \sum_{k=2}^j (Q_k - Q_{k-1})u + Q_1 u.$$

Thus,

$$\begin{aligned}
 A(u, u) &= \sum_{k=2}^j A(u, (Q_k - Q_{k-1})u) + A(u, Q_1 u) \\
 (4.14) \quad &= \sum_{k=2}^j A(E_{k-1}u, (Q_k - Q_{k-1})u) + A(u, Q_1 u) \\
 &\quad + \sum_{k=2}^j A((I - E_{k-1})u, (Q_k - Q_{k-1})u).
 \end{aligned}$$

For the first sum on the right-hand side of (4.14), we see that

$$\begin{aligned}
 \sum_{k=2}^j A(E_{k-1}u, (Q_k - Q_{k-1})u) &= \sum_{k=2}^j A(P_k E_{k-1}u, (Q_k - Q_{k-1})u) \\
 &= \sum_{k=2}^j (\tilde{P}_k^0 A_k P_k E_{k-1}u, (Q_k - Q_{k-1})u)_k \\
 &\leq \sum_{k=2}^j \|\tilde{P}_k^0 A_k P_k E_{k-1}u\|_k \|(Q_k - Q_{k-1})u\|_k.
 \end{aligned}$$

Applying (4.5) gives

$$\sum_{k=2}^j A(E_{k-1}u, (Q_k - Q_{k-1})u) \leq (C_1)^{1/2} A^{1/2}(u, u) \sum_{k=2}^j \lambda_k^{-1/2} \|\tilde{P}_k^0 A_k P_k E_{k-1}u\|_k.$$

For \bar{R}_k defined by (4.8), Theorem 3.2 gives

$$\begin{aligned}
 \sum_{k=2}^j A(E_{k-1}u, (Q_k - Q_{k-1})u) \\
 \leq (C_R C_1(j-1))^{1/2} A^{1/2}(u, u) \left(\sum_{k=2}^j (\bar{R}_k \tilde{P}_k^0 A_k P_k E_{k-1}u, \tilde{P}_k^0 A_k P_k E_{k-1}u) \right)^{1/2}.
 \end{aligned}$$

It is easy to check that for $v \in \widetilde{\mathcal{M}}_k$,

$$\begin{aligned}
 (4.15) \quad \bar{R}_k v &= \tilde{R}_k v + \tilde{R}_k^t v - \tilde{R}_k^t \tilde{A}_k \tilde{R}_k v \\
 &= R_k v + R_k^t v - R_k^t A_k R_k v = (I - K_k^* K_k) A_k^{-1} v.
 \end{aligned}$$

The last equality in (4.15) defines an extension of \bar{R}_k to $\widetilde{\mathcal{M}}_k$. This extension, which we shall still denote by \bar{R}_k , is symmetric with respect to $(\cdot, \cdot)_k$. Moreover, R_k^t is defined by cycling through Algorithm 3.1 in reverse order, and hence its image is contained in $\widetilde{\mathcal{M}}_k$. Thus, it follows that $\bar{R}_k \tilde{P}_k^0 = \bar{R}_k$, and hence

$$\begin{aligned}
 \sum_{k=2}^j A(E_{k-1}u, (Q_k - Q_{k-1})u) &\leq (C_R C_1(j-1))^{1/2} A^{1/2}(u, u) \\
 &\quad \cdot \left(\sum_{k=2}^j A((I - K_k^* K_k) P_k E_{k-1}u, P_k E_{k-1}u) \right)^{1/2}.
 \end{aligned}$$

Let $w = P_k E_{k-1} u$. The spectrum of $K_k^* K_k$ is in $[0, 1]$, and hence

$$\begin{aligned} A((I - K_k^* K_k)w, w) &\leq A((I - \bar{K}_k^{(m(k))} (\bar{K}_k^{(m(k))})^*)w, w) \\ &= A((2I - T_k)w, T_k w) = A((2I - T_k)E_{k-1} u, T_k E_{k-1} u). \end{aligned}$$

Thus,

$$\begin{aligned} (4.16) \quad \sum_{k=2}^j A(E_{k-1} u, (Q_k - Q_{k-1})u) &\leq (C_R C_1(j-1))^{1/2} A^{1/2}(u, u) \\ &\cdot \left(\sum_{k=2}^j A((2I - T_k)E_{k-1} u, T_k E_{k-1} u) \right)^{1/2}. \end{aligned}$$

For the remaining terms in (4.14), we have

$$\begin{aligned} (4.17) \quad \sum_{k=2}^j A((I - E_{k-1})u, (Q_k - Q_{k-1})u) + A(u, Q_1 u) \\ = \sum_{k=2}^{j-1} A((E_k - E_{k-1})u, Q_k u) + A((I - E_{j-1})u, u). \end{aligned}$$

But, $E_{k-1} - E_k = T_k E_{k-1}$, and it follows that

$$I - E_{j-1} = \sum_{k=1}^{j-1} T_k E_{k-1}.$$

From this and (4.17),

$$\begin{aligned} (4.18) \quad \sum_{k=2}^j A((I - E_{k-1})u, (Q_k - Q_{k-1})u) + A(u, Q_1 u) \\ = \sum_{k=2}^{j-1} A(T_k E_{k-1} u, (I - Q_k)u) + A(T_1 u, u) \\ \leq (j-1)^{1/2} (1 + C_2^{1/2}) \left(\sum_{k=1}^{j-1} A(T_k E_{k-1} u, T_k E_{k-1} u) \right)^{1/2} A^{1/2}(u, u). \end{aligned}$$

We shall show that

$$(4.19) \quad A(T_k w, T_k w) \leq 2c_1 A((2I - T_k)w, T_k w).$$

If $m(k)$ is even, then this is evidently true, since $c_1 \geq 1$ and T_k is symmetric in $A(\cdot, \cdot)$ with spectrum in $[0, 1]$. For $m(k)$ odd, set $\tilde{K}_k = (K_k K_k^*)^{(m(k)-1)/2}$. Then \tilde{K}_k is symmetric in $A(\cdot, \cdot)$ with spectrum in $[0, 1]$ and $T_k = (I - \tilde{K}_k K_k)P_k$. Clearly, it suffices to prove (4.19) for $w \in \mathcal{M}_k$. Now

$$\begin{aligned} A(T_k w, T_k w) &= A((I - \tilde{K}_k K_k)w, (I - \tilde{K}_k K_k)w) \\ &\leq 2 \left[A((I - K_k)w, (I - K_k)w) + A((I - \tilde{K}_k)K_k w, (I - \tilde{K}_k)K_k w) \right]. \end{aligned}$$

By Theorem 3.2,

$$A((I - K_k)w, (I - K_k)w) \leq \frac{2c_1}{c_1 + 1} A((I - K_k)w, w).$$

This can be rewritten (see (3.25)) as

$$A((I - K_k)w, (I - K_k)w) \leq c_1 [A(w, w) - A(K_k w, K_k w)].$$

Using the symmetry of \tilde{K}_k and the fact that its spectrum is in $[0, 1]$ gives

$$A((I - \tilde{K}_k)K_k w, (I - \tilde{K}_k)K_k w) \leq A(K_k w, K_k w) - A(\tilde{K}_k K_k w, \tilde{K}_k K_k w).$$

Combining the above and noting that $c_1 \geq 1$ gives

$$\begin{aligned} A(T_k w, T_k w) &\leq 2c_1 [A(w, w) - A(\tilde{K}_k K_k w, \tilde{K}_k K_k w)] \\ &= 2c_1 A((2I - T_k)w, T_k w), \end{aligned}$$

i.e., (4.19) holds.

Combining (4.14), (4.16), (4.18), and (4.19) gives

$$A(u, u) \leq C(j-1) \left(\sum_{k=1}^j A((2I - T_k)E_{k-1} u, T_k E_{k-1} u) \right),$$

for $C = [(1 + C_2^{1/2})(2c_1)^{1/2} + (C_R C_1)^{1/2}]^2$. This completes the proof of the theorem.

Remark 4.1. The requirement that inequality (4.7) need only hold on $\tilde{\mathcal{M}}_k$ is important in local refinement applications. These are discussed in more detail in [9]. However, to apply Theorems 3.1 and 3.2, one need only provide a decomposition of the subspace $\tilde{\mathcal{M}}_k$. The resulting smoothing operator R_k only involves computation in the subdomain where the new nodes are being added.

5. IMPLEMENTATION OF THE SMOOTHING PROCEDURE

In this section, we consider implementation of the smoothing procedures described in §3. We shall see that the additive schemes correspond to generalizations of block Jacobi iteration. The product schemes correspond to generalizations of block Gauss-Seidel iteration. The observations that, e.g., Gauss-Seidel iteration is a product scheme of the form of Algorithm 3.1 are not new (cf., for example, [17]). We include this section only to stress the point that the results provided earlier apply to the smoothers commonly used in multigrid algorithms.

We first consider computer implementation of the parts of Algorithm 2.2 which are relevant to the smoothing procedures. Assume that a decomposition of \mathcal{M}_k of the form of (3.1) is given which satisfies (3.3) and (3.4). Moreover, assume that there is a basis $\{\phi_k^i\}$ for \mathcal{M}_k such that each \mathcal{M}_k^i has a basis consisting of a subset of $\{\phi_k^i\}$. Let M denote the stiffness matrix associated with this basis, i.e., $M_{im} = (A_k \phi_k^i, \phi_k^m)_k$. In implementation, one seldom is required to solve (2.1) but rather the equivalent matrix equation

$$MU = F,$$

where U is related to the solution of (2.1) by $u = \sum_i U_i \phi_j^i$ and F is a known vector of coefficients ($F_i = (f, \phi_j^i)_j$). Consequently, in the multigrid implementation, we are required to compute the action of R_k on a function $g \in \mathcal{M}_k$ which is represented by the inner product vector $G_i = (g, \phi_k^i)_k$.

We first consider the case of the additive smoothing operator defined by (3.2). Let S_i denote the indices of $\{\phi_k^m\}$ which correspond to the basis functions of \mathcal{M}_k^i . We note that the vector W^i representing the function $w^i = A_{k,i}^{-1}Q_k^i g$ satisfies the equation

$$M^i W^i = G^i,$$

where

$$M_{lm}^i = \begin{cases} M_{lm} & \text{if } l, m \in S_i, \\ 0 & \text{if } l \neq m \text{ and either } l \notin S_i \text{ or } m \notin S_i, \\ 1 & \text{if } l = m \text{ and } l \notin S_i \end{cases}$$

and

$$G_m^i = \begin{cases} G_m & \text{if } m \in S_i, \\ 0 & \text{otherwise.} \end{cases}$$

We now consider the case when all of the subspaces are disjoint. Then, we may partition the basis elements into groups corresponding to the subspaces. Under this ordering, the vector $W = \{W_i\}$ representing the function $w = \sum A_{k,i}^{-1}Q_k^i g$ satisfies the equation

$$\bar{M}_a W = G,$$

where M_a is the block diagonal part of the matrix M . In the case when each subspace has one degree of freedom associated with a given basis function, then W is given by

$$W_i = (A_k \phi_k^i, \phi_k^i)_k^{-1} G_i,$$

i.e., the smoother corresponds to the Jacobi method applied to the diagonally scaled stiffness matrix.

We next consider the case of the multiplicative smoother. Again, we look at the case when all of the subspaces are disjoint and the basis elements are ordered into groups accordingly. The matrix M has a block structure corresponding to this ordering with blocks denoted by $M^{i,m}$, $i, m = 1, \dots, l$. As usual, we write

$$M = L + D + U,$$

where L , D , and U are respectively, block lower diagonal, block diagonal, and block upper diagonal. Let F^i be the vector of data corresponding to these blocks, i.e., $F^i = ((f, \phi_{m_1}^k)_k, \dots, (f, \phi_{m_2}^k)_k)^t$, where m_1 and m_2 denote the first and last basis element corresponding to the subspace \mathcal{M}_k^i . Let the vectors V_i be the vectors of coefficients representing the functions v_i appearing in Algorithm 3.1. These are partitioned in a similar manner. We first note that the i th step in (3.14) only changes the i th component. Thus, the i th component of V_l is defined by the equation

$$M_{i,i} V_l^i = F_i - \sum_{m < i} M_{i,m} V_{i-1}^m = F_i - \sum_{m < i} M_{i,m} V_l^m - \sum_{m > i} M_{i,m} V_0^m.$$

This can be rewritten as

$$(L + D) V_l = -U V_0 + F,$$

and corresponds to block Gauss-Seidel iteration applied to the stiffness matrix M .

6. TYPICAL FINITE ELEMENT APPLICATIONS

In this section, we discuss developing smoothers using the techniques of §3 for finite element multigrid applied to a second-order elliptic boundary value problem. First, we consider the case when the subspace is defined in terms of a quasi-uniform triangulation which approximates the original domain. This often leads to spaces $\{\mathcal{M}_k\}$ which are not nested. The case of mesh refinement is discussed next. We will see that it is easy to apply the techniques presented earlier to develop smoothers which only require computation where new nodes are being added and give rise to effective multigrid algorithms.

We shall consider the problem of approximating the solution U of

$$(6.1) \quad \begin{aligned} LU &= F \quad \text{in } \Omega, \\ U &= 0 \quad \text{on } \partial\Omega. \end{aligned}$$

Here Ω is a domain (not necessarily polygonal) in n -dimensional Euclidean space and L is given by

$$Lv = - \sum_{i,j=1}^n \frac{\partial}{\partial x_i} \left(a_{ij} \frac{\partial v}{\partial x_j} \right),$$

with $\{a_{ij}\}$ uniformly positive definite and bounded on $\overline{\Omega}$. The form A corresponding to the above operator is given by

$$(6.2) \quad A(v, w) = \sum_{i,j=1}^n \int_{\Omega} a_{ij} \frac{\partial v}{\partial x_i} \frac{\partial w}{\partial x_j} dx.$$

This form is defined for all v and w in the Sobolev space $H^1(\Omega)$ (the space of distributions in $L^2(\Omega)$ with square-integrable first derivatives). Clearly, $U \in H_0^1(\Omega)$ is the solution of

$$A(U, \theta) = (F, \theta) \quad \text{for all } \theta \in H_0^1(\Omega),$$

where $H_0^1(\Omega)$ is the subspace of $H^1(\Omega)$ of functions which vanish in the appropriate sense on $\partial\Omega$ and (\cdot, \cdot) denotes the L^2 inner product on Ω .

We will first discuss the case of quasi-uniform triangulation. We assume that Ω has been approximately triangulated with a sequence of quasi-uniform triangulations $\Omega_k = \bigcup_i \tau_k^i$ of size h_k for $k = 1, \dots, j$, where the quasi-uniformity constants are independent of k . We define \mathcal{M}_k to be the set of piecewise linear functions (with respect to the triangulation $\bigcup_i \tau_k^i$) which vanish on $\partial\Omega_k$. If Ω is polygonal, then it is possible to take $\Omega_k = \Omega$ and construct triangulations which are nested.

We next define the inner product $(\cdot, \cdot)_k$. We do not have complete freedom here, since we must choose an inner product so that either (4.1) or (4.5) are satisfied, depending on the application. Let $\{y_k^i\}$ be the collection of nodes corresponding to the triangulation for \mathcal{M}_k . It suffices to take

$$(6.3) \quad (u, v)_k = h_k^n \sum_i u(y_k^i) v(y_k^i).$$

Note that the quasi-uniformity of the triangulations implies that the norm $\|\cdot\|_k$ is equivalent to the $L^2(\Omega)$ norm on the subspace \mathcal{M}_k . The operators A_k , $k = 1, \dots, l$, are then defined by

$$(6.4) \quad (A_k v, \phi)_k = A(v, \phi) \quad \text{for all } \phi \in \mathcal{M}_k.$$

Finally, the operators I_k are defined by nodal interpolation, i.e., $I_k w$ is defined to be the unique function in \mathcal{M}_k which equals w at the nodes of \mathcal{M}_k .

Let $\{\phi_k^i\}_{i=1}^{N_k}$ denote the usual nodal basis associated with the subspace \mathcal{M}_k . Partition the integers $\{1, 2, \dots, N_k\}$ into sets S_1, S_2, \dots, S_l and define \mathcal{M}_k^i to be the span of the basis functions with indices in S_i . The discussion in the previous section shows that implementation of (3.2) and Algorithm 3.1 reduce to block Jacobi and block Gauss-Seidel iteration on the stiffness matrices. These subspaces provide a direct sum decomposition of the space \mathcal{M}_k and hence, the decomposition $u = \sum u_i$ with $u_i \in \mathcal{M}_k^i$ is uniquely defined. In addition, since the matrix with entries $N_{im} = (\phi_k^i, \phi_k^m)_k$ is diagonal,

$$\sum_{i=1}^l \|u_i\|_k^2 = \|u\|_k^2,$$

i.e., (3.4) holds with $c_0 = 1$.

The constant c_1 appearing in (3.3) is related to the geometry of the subspaces. For $i = 1, 2, \dots, l$, let Ω_k^i denote the union of the supports of the basis functions defining \mathcal{M}_k^i . Note that κ_{im} is nonzero only if $\Omega_k^i \cap \Omega_k^m \neq \emptyset$. Let χ_k^i be the number of subdomains $\{\Omega_k^m\}$ which intersect Ω_k^i . Then we can take c_1 in (3.3) to be the maximum of $\{\chi_k^i\}$ for $i = 1, 2, \dots, l$. In the case of point relaxation (i.e., $\mathcal{M}_k^i = \{c\phi_k^i\}$), c_1 can be taken to be one plus the maximum number of triangles which meet at a given vertex. Alternatively, for line relaxation, the grid consists of a regular rectangular mesh and the i th subspace is defined to be, for example, the span of the basis functions on the i th horizontal mesh line. In this case, $c_1 = 3$. Obviously, many other examples are possible.

We next consider the case when the mesh results from a local refinement. To illustrate this situation, we consider the case of two spatial dimensions. We note that for refinement applications, it is only possible to prove (4.1) with a C_α which grows with powers of the ratio of the diameters of largest to smallest triangle in the refined mesh. In contrast, estimates of the form (4.5) hold with constants independent of the mesh parameters. Consequently, we shall only consider the case of nested spaces and inherited operators.

We start with the definition of the nested refined grids. These grids are defined in terms of a given sequence of nested subdomains

$$\Omega_j \subseteq \Omega_{j-1} \subseteq \dots \subseteq \Omega_0 = \Omega.$$

We assume that we are given a coarse triangulation of $\Omega = \bigcup_m \tau_0^m$. This coarse triangulation provides the first grid $\{\tau_0^m\}$. Given that a grid $\{\tau_{k-1}^m\}$ has been defined, the grid $\{\tau_k^m\}$ is defined by refining those triangles of $\{\tau_{k-1}^m\}$ which are in Ω_k . This refinement is done, for example, by breaking each triangle of the mesh $\{\tau_{k-1}^m\}$ in Ω_k into four triangles by connecting the midpoints of the edges. We assume $\partial\Omega_k$ aligns with the mesh $\{\tau_{k-1}^m\}$.

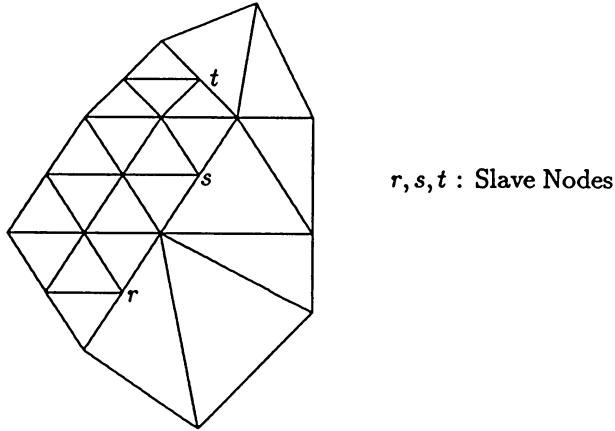


FIGURE 6.1
A mesh transition region

The space \mathcal{M}_k is defined to be the set of continuous functions on Ω which are piecewise linear with respect to the grid $\{\tau_k^m\}$ and vanish on $\partial\Omega$. We note that the continuity constraint implies that there are no new degrees of freedom corresponding to nodes on $\partial\Omega_k$ (see Figure 6.1). These new nodes on $\partial\Omega_k$ will be called slave nodes since, by continuity, their values are determined by the values of their neighboring nodes (which were already in the previous grid). It is easy to see that the space \mathcal{M}_k has a nodal basis consisting of the vertices of $\{\tau_k^m\}$ excluding the slave nodes.

For this application, we shall take $(\cdot, \cdot)_k$ to be the $L^2(\Omega)$ inner product. The operators $\{A_k\}$ are defined by (6.4) and the operators I_k are defined to be the natural injection of \mathcal{M}_{k-1} into \mathcal{M}_k .

A sequence of operators Q_k , $k = 1, \dots, l$, are constructed in [9] satisfying (4.5). These operators, in addition, satisfy $(Q_k - Q_{k-1})v \in \widetilde{\mathcal{M}}_k$ for all $v \in \mathcal{M}_j$, where

$$(6.5) \quad \widetilde{\mathcal{M}}_k = \{\phi \in \mathcal{M}_k \mid \text{supp } \phi \subseteq \Omega_k\}.$$

Now, to apply Theorem 4.3, we need only provide a decomposition of the subspace $\widetilde{\mathcal{M}}_k$. Note that $\widetilde{\mathcal{M}}_k$ is a finite element space corresponding to a quasi-uniform triangulation of Ω_k . Accordingly, the constructions given above can be used. Note that this leads to smoothing algorithms which only require computation involving the nodes of Ω_k and not on all of the nodes of the space \mathcal{M}_k .

BIBLIOGRAPHY

1. R. E. Bank and C. C. Douglas, *Sharp estimates for multigrid rates of convergence with general smoothing and acceleration*, SIAM J. Numer. Anal. **22** (1985), 617–633.
2. R. E. Bank and T. Dupont, *An optimal order process for solving finite element equations*, Math. Comp. **36** (1981), 35–51.
3. D. Braess and W. Hackbusch, *A new convergence proof for the multigrid method including the V-cycle*, SIAM J. Numer. Anal. **20** (1983), 967–975.
4. J. H. Bramble and J. E. Pasciak, *New convergence estimates for multigrid algorithms*, Math. Comp. **49** (1987), 311–329.
5. J. H. Bramble, J. E. Pasciak, and J. Xu, *The analysis of multigrid algorithms with nonnested spaces or noninherited quadratic forms*, Math. Comp. **56** (1991), 1–34.

6. ———, *The analysis of multigrid algorithms for nonsymmetric and indefinite elliptic problems*, Math. Comp. **51** (1988), 389–414.
7. ———, *Parallel multilevel preconditioners*, Math. Comp. **55** (1990), 1–22.
8. J. H. Bramble, J. E. Pasciak, J. Wang, and J. Xu, *Convergence estimates for product iterative methods with applications to domain decomposition*, Math. Comp. **56** (1991), 1–21.
9. ———, *Convergence estimates for multigrid algorithms without regularity assumptions*, Math. Comp. **56** (1991), 23–45.
10. A. Brandt, *Multi-level adaptive solutions to boundary-value problems*, Math. Comp. **31** (1977), 333–390.
11. W. Hackbusch, *Multi-grid methods and applications*, Springer-Verlag, New York, 1985.
12. J. Mandel, S. McCormick, and R. Bank, *Variational multigrid theory*, Multigrid Methods (S. McCormick, ed.), SIAM, Philadelphia, PA, 1987, pp. 131–178.
13. J. Mandel, S. F. McCormick, and J. Ruge, *An algebraic theory for multigrid methods for variational problems*, (Preprint).
14. S. F. McCormick, *Multigrid methods for variational problems: General theory for the V-cycle*, SIAM J. Numer. Anal. **22** (1985), 634–643.
15. ———, *Multigrid methods for variational problems: Further results*, SIAM J. Numer. Anal. **21** (1984), 255–263.
16. S. McCormick (Ed.), *Multigrid methods*, SIAM, Philadelphia, PA, 1987.
17. S. F. McCormick and J. Ruge, *Unigrid for multigrid simulation*, Math. Comp. **41** (1983), 43–62.
18. R. Verfüth, *A multilevel algorithm for mixed problems*, SIAM J. Numer. Anal. **21** (1984), 264–284.
19. J. Wang, *Convergence analysis without regularity assumptions for multigrid algorithms based on SOR smoothing*, (preprint).
20. O. Widlund, *Optimal iterative refinement methods*, Technical Report No. 391, Courant Institute of Mathematical Sciences, 1988.

DEPARTMENT OF MATHEMATICS, CORNELL UNIVERSITY, ITHACA, NEW YORK 14853
E-mail address: bramble@mssun7.msi.cornell.edu

DEPARTMENT OF APPLIED SCIENCE, BROOKHAVEN NATIONAL LABORATORY, UPTON, NEW YORK 11973
E-mail address: pasciak@bnl.gov

3.7 New estimates for multilevel algorithms including the V-cycle

New estimates for multilevel algorithms including the V-cycle[?]

NEW ESTIMATES FOR MULTILEVEL ALGORITHMS INCLUDING THE V-CYCLE

JAMES H. BRAMBLE AND JOSEPH E. PASCIAK

ABSTRACT. The purpose of this paper is to provide new estimates for certain multilevel algorithms. In particular, we are concerned with the simple additive multilevel algorithm discussed recently together with J. Xu and the standard V-cycle algorithm with one smoothing step per grid. We shall prove that these algorithms have a uniform reduction per iteration independent of the mesh sizes and number of levels, even on nonconvex domains which do not provide full elliptic regularity. For example, the theory applies to the standard multigrid V-cycle on the L-shaped domain, or a domain with a crack, and yields a uniform convergence rate. We also prove uniform convergence rates for the multigrid V-cycle for problems with nonuniformly refined meshes. Finally, we give a new multigrid approach for problems on domains with curved boundaries and prove a uniform rate of convergence for the corresponding multigrid V-cycle algorithms.

1. INTRODUCTION

In recent years, multigrid methods have been used extensively to efficiently solve the discrete equations which arise in the numerical approximation of partial differential equations (see the references in [13, 18, 21]). In conjunction, there has been intensive research into the theoretical understanding of the convergence properties of these methods (cf. [2, 3, 6, 7, 9–12, 18, 20, 21]). In this paper, we present a new general theory based on two assumptions which are different from those made in earlier works. By using the new theory, we are able to derive some surprising uniform convergence bounds for a number of problems. The earlier theories suggested that the rates of convergence for these applications deteriorated as the number of multigrid levels increased.

Previously, there were two general approaches for proving convergence of multigrid algorithms. The first was based on the so-called regularity and approximation assumption [7]. The verification of this hypothesis used both the approximation properties of the discrete method as well as the regularity properties of the approximated partial differential equation. The theory of [7, 16] only provides a uniform convergence rate for the V-cycle algorithm in the case

Received by the editor August 26, 1991 and, in revised form, May 12, 1992.

1991 *Mathematics Subject Classification*. Primary 65N30; Secondary 65F10.

This manuscript has been authored under contract number DE-AC02-76CH00016 with the U.S. Department of Energy. Accordingly, the U.S. Government retains a non-exclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. This work was also supported in part under the National Science Foundation Grant No. DMS-9007185 and by the U.S. Army Research Office through the Mathematical Sciences Institute, Cornell University.

©1993 American Mathematical Society
0025-5718/93 \$1.00 + \$.25 per page

of full elliptic regularity. It gives a deteriorating estimate, for example, in the case of an L-shaped domain or a domain with a crack boundary.

The second general approach is based solely on approximation and is given in [9, 12]. The “no regularity” theory gives rise to estimates which deteriorate at least linearly with the number of levels in the multigrid scheme.

In contrast, the theory developed in this paper uses two assumptions. The first assumption replaces the regularity and approximation assumption by a much weaker inequality on the whole space (see (3.1)). As we will demonstrate, this inequality often can be verified in applications where full elliptic regularity fails to hold. In [22], this assumption was shown to hold for the standard application using Besov space arguments. In this paper, we show that the estimate can be verified in this case using regularity properties of elliptic problems which depend only on the domain.

The second assumption is that the underlying discrete operator should be appropriately small when restricted to coarser grid spaces (see (3.5)). This assumption was motivated by an inequality proved by Zhang [23] for the standard application. Zhang’s proof used local arguments. In this paper, we provide a more general approach, which has been successfully applied even in cases with nonlocal operators (cf. [4, 10]).

We provide a general theory which shows how the rate of convergence of multigrid algorithms can be bounded in terms of the constants appearing in these two new assumptions. We will give three applications showing that the general theory can be used to prove stronger convergence estimates for V-cycle algorithms. The first applies the general theory to second-order uniformly elliptic problems in d -dimensional Euclidean space. We will show that the V-cycle algorithm, with only one smoothing per grid per iteration, leads to a uniformly convergent algorithm independent of the number of levels. This is true on the L-shaped domain and domains with a crack boundary. In addition, these results are valid for both the additive (multilevel) and multiplicative (standard multigrid) form of the algorithm.

The second application involves an example with a general mesh refinement. The best earlier results known for this problem were given in [9] and [12], where it was shown that the convergence rate for the additive and multiplicative algorithms could deteriorate at worst at a rate of $1 - c/J^2$ and $1 - c/J$, respectively. Here, J is the number of levels and c is a positive constant independent of the number of levels. Applying the general theory developed in this paper, we prove, for instance, that the convergence rate for the V-cycle for this example is bounded uniformly below one, independently of J .

The final example applies to domains with curved boundaries. We construct a simple set of nested multilevel spaces and show that our general theory may be used to prove uniform estimates for the additive multilevel schemes and that the V-cycle multigrid scheme (with one smoothing step per grid) has an associated contraction number bounded uniformly below one.

The outline of the remainder of the paper is as follows. Following [7], we provide a general framework for the development of multilevel algorithms in §2. Section 3 provides a general theory for the analysis of these algorithms based on the new assumptions mentioned above. Section 4 verifies the new assumptions in the case of the quasi-uniform finite element approximation. The theorems of §3 then give new convergence estimates for this application. The case of

mesh refinement is considered in §5. Application of the new theory also leads to uniform convergence estimates. Finally, we consider a multigrid algorithm for domains with curved boundaries in §6. Again, uniform rates of convergence are proven.

2. GENERAL ADDITIVE AND MULTIPLICATIVE MULTIGRID ALGORITHMS

Following [7], we first give a general framework for the development of multigrid algorithms in this section. We next define the additive and multiplicative versions of the multigrid algorithms. The additive version provides a preconditioner for the resulting operator. The multiplicative algorithm also gives rise to a preconditioner which can be used in a simple linear iteration (the standard multigrid approach) or as a convergence accelerator in a conjugate gradient iteration. For convenience, the algorithms are presented in an abstract Hilbert space setting. The results most naturally apply to finite element multigrid algorithms but can also be applied to certain formulations of finite difference multigrid algorithms.

Let us assume that we are given a nested sequence of finite-dimensional vector spaces

$$M_1 \subset M_2 \subset \cdots \subset M_J.$$

Associated with this sequence, we assume that we are given additional subspaces, $\widetilde{M}_i \subseteq M_i$ for $i = 2, 3, \dots, J$. The multilevel algorithms will involve smoothing only on the subspaces $\{\widetilde{M}_i\}$. In addition, let $A(\cdot, \cdot)$ and (\cdot, \cdot) be symmetric positive definite bilinear forms on M_J . Let $\|\cdot\|$ denote the norm corresponding to (\cdot, \cdot) . We shall study multigrid algorithms for the solution of the problem: Given $f \in M_J$, find $v \in M_J$ satisfying

$$(2.1) \quad A(v, \phi) = (f, \phi) \quad \text{for all } \phi \in M_J.$$

The multigrid algorithms are described in terms of auxiliary operators. For $k = 1, \dots, J$, define the operator $A_k : M_k \mapsto M_k$ by

$$(2.2) \quad (A_k w, \phi) = A(w, \phi) \quad \text{for all } \phi \in M_k.$$

The operator A_k is clearly symmetric (in both the $A(\cdot, \cdot)$ and (\cdot, \cdot) inner products) and positive definite. Also define the orthogonal projectors $P_k, Q_k : M_J \mapsto M_k$ by

$$A(P_k w, \phi) = A(w, \phi) \quad \text{for all } \phi \in M_k,$$

and

$$(Q_k w, \phi) = (w, \phi) \quad \text{for all } \phi \in M_k.$$

With M_k replaced by \widetilde{M}_k , the operators \widetilde{A}_k , \widetilde{P}_k , and \widetilde{Q}_k are defined analogously. It is easy to check the fundamental identity $Q_l A_k = A_l P_l$ whenever $l < k$. This and the analogous identity $\widetilde{Q}_k A_k = \widetilde{A}_k \widetilde{P}_k$ will be used in various places throughout this paper.

Equation (2.1) can then be rewritten

$$(2.3) \quad A_J u_J = f.$$

Both the additive and multiplicative (standard) versions of the multigrid algorithms can be thought of as defining an operator $B_J : M_J \mapsto M_J$ which approximately inverts A_J . The goal of the analysis is to provide estimates for either the spectrum of $B_J A_J$ or an appropriate norm of $I - B_J A_J$.

To introduce smoothing into the multigrid algorithms, we shall use “generic” smoothing operators $R_k: M_k \mapsto \tilde{M}_k$, for $k = 2, \dots, J$. Examples of these operators are given in [6]. The properties which they satisfy will be discussed in the subsequent analysis. We set $R_1 = A_1^{-1}$, i.e., we solve on the coarsest space. The additive multigrid preconditioner is then defined by

$$(2.4) \quad B_J^a = \sum_{k=1}^J R_k Q_k.$$

To analyze the above preconditioner, we must provide estimates for the spectrum of the operator

$$(2.5) \quad B_J^a A_J = \sum_{k=1}^J T_k,$$

where $T_k = R_k Q_k A_J = R_k A_k P_k$. Note that $T_1 = P_1$.

We shall always take R_k to be symmetric with respect to the (\cdot, \cdot) inner product when used in the additive algorithm. This implies that $R_k = R_k \bar{Q}_k$ and hence $T_k = R_k \tilde{A}_k \tilde{P}_k$. This also results in a symmetric operator B_J^a . In general, preconditioned iterative techniques for symmetric problems are much more effective when applied with symmetric preconditioners. The use of a nonsymmetric preconditioner is inappropriate in this case.

The standard multigrid algorithm is often defined as a process which produces a function $\text{MG}_k(w_k, g_k)$. Here, k is the grid level and $w_k, g_k \in M_k$. The function w_k can be thought of as a given approximation to the solution u_k of

$$(2.6) \quad A_k u_k = g_k.$$

The result of the multigrid process is to produce $\text{MG}_k(w_k, g_k) \in M_k$, an improved approximation to the solution $u_k = A_k^{-1} g_k$. A standard presentation of this algorithm is given below.

Algorithm 2.1. For $k = 1$, define $\text{MG}_1(w_1, g_1) = A_1^{-1} g_1$. For $k > 1$, $\text{MG}_k(w_k, g_k)$ is defined in terms of $\text{MG}_{k-1}(\cdot, \cdot)$ as follows:

(1) Set

$$(2.7) \quad x_k = w_k + R_k^t(g_k - A_k w_k).$$

(2) Set $y_k = x_k + q$ where

$$(2.8) \quad q = \text{MG}_{k-1}(0, Q_{k-1}(g_k - A_k x_k)).$$

(3) Set $\text{MG}_k(w_k, g_k) = y_k + R_k(g_k - A_k y_k)$.

The first and third steps above correspond to smoothing. The second step is a correction step. The operator R_k^t is the (\cdot, \cdot) adjoint of the operator R_k . Many generalizations of the above algorithm exist involving more smoothing and correction iterations [7, 11, 18, 21]. We only consider the above algorithm since the results of this paper are most interesting in this case. We note that the results immediately extend to more general algorithms (see the remark after the second theorem of §3) with more than one correction step (e.g., the W-cycle algorithm) as well as algorithms with more than one smoothing step per level.

The above algorithm results in a very simple error reduction process. For any k , let u_k solve (2.6), $v_k = \text{MG}_k(w_k, g_k)$, $e_k^0 = u_k - w_k$, and $e_k^1 = u_k - v_k$. Note that the error e_k^1 is the resulting error after one application of the multigrid process on the k th subspace with an initial error e_k^0 . We shall demonstrate that these errors are related by a linear operator F_k , i.e., $e_k^1 = F_k e_k^0$. This is obviously true for $k = 1$, where $F_1 = 0$. Assume that $e_{k-1}^1 = F_{k-1} e_{k-1}^0$ holds for all $w_{k-1}, g_{k-1} \in M_{k-1}$. From (2.7) we have that $u_k - x_k = K_k^* e_k^0$, where $K_k^* = I - R_k^t A_k$. The function q produced in Step 2 approximates the function

$$\begin{aligned}\tilde{q} &= A_{k-1}^{-1} Q_{k-1}(g_k - A_k x_k) \\ &= A_{k-1}^{-1} Q_{k-1} A_k (u_k - x_k) = P_{k-1}(u_k - x_k).\end{aligned}$$

By assumption,

$$\tilde{q} - q = F_{k-1} \tilde{q},$$

or

$$u_k - x_k - q = u_k - x_k - (I - F_{k-1}) P_{k-1}(u_k - x_k).$$

Thus,

$$u_k - y_k = [(I - P_{k-1}) + F_{k-1} P_{k-1}](u_k - x_k).$$

Finally, $e_k^1 = K_k(u_k - y_k)$, where $K_k = (I - R_k A_k)$. Hence,

$$e_k^1 = K_k[(I - P_{k-1}) + F_{k-1} P_{k-1}]K_k^* e_k^0.$$

Thus, we see that the errors on the k th level are related by the linear operator F_k defined by the recurrence

$$(2.9) \quad F_k = K_k[(I - P_{k-1}) + F_{k-1} P_{k-1}]K_k^*.$$

The multigrid process is often applied repeatedly to develop an iterative method for solving problem (2.3). Given an initial approximation u^0 , subsequent approximations are defined by

$$(2.10) \quad u^{l+1} = \text{MG}_J(u^l, g) \quad \text{for } l = 1, \dots.$$

From the above discussion, the error $e^l = u - u^l$ is given by $e^l = (F_J)^l e^0$. Consequently, the multigrid iterative process corresponds to a linear iterative procedure. This can be written equivalently as

$$(2.11) \quad u^{l+1} = u^l + B_J^m(g - A_J u^l)$$

for the operator $B_J^m = (I - F_J)A_J^{-1}$. Alternatively, this operator B_J^m can be directly defined by the following algorithm.

Algorithm 2.2. Define $B_1^m = A_1^{-1}$. For $k > 1$, $B_k^m g$ for $g \in M_k$ is defined as follows:

(1) Set

$$(2.12) \quad x = R_k^t(g).$$

(2) Set $y = x + q$, where q is given by

$$(2.13) \quad q = B_{k-1}^m Q_{k-1}(g - A_k x).$$

(3) Set $B_k^m g = y + R_k(g - A_k y)$.

It is straightforward to show that B_k^m satisfies (cf. [7])

$$I - B_k^m A_k = K_k[(I - P_{k-1}) + (I - B_{k-1} A_{k-1})P_{k-1}]K_k^*$$

with B_k^m , B_{k-1}^m defined by Algorithm 2.2. This shows that $F_k = I - B_k^m A_k$, for $k = 1, \dots, J$, i.e., the linear iteration (2.11) with B_j^m defined by Algorithm 2.2 is equivalent to the multigrid iteration (2.10). This is an important observation in that it allows the use of the multigrid process to define preconditioning operators B_j^m . For example, the operator B_J^m can be used as a preconditioner with the conjugate gradient method to develop more effective iteration procedures in many applications. It also allows us to use the operator presentation of Algorithm 2.2 for the analysis of the multigrid iteration.

It was shown in [9] that the error reduction operator associated with Algorithm 2.2 (the standard multigrid algorithm) can be written

$$(2.14) \quad (I - B_J^m A_J) = (I - T_J)(I - T_{J-1}) \cdots (I - T_2)(I - T_1) \\ \cdot (I - T_1^*)(I - T_2^*) \cdots (I - T_{J-1}^*)(I - T_J^*).$$

This identity depends upon the assumption that the subspaces are imbedded and that one form is used to define the operators on all levels (see (2.2)).

Note that T_k^* is the adjoint of T_k with respect to the $A(\cdot, \cdot)$ inner product. Comparing (2.5) and (2.14) clearly shows the relation between additive and multiplicative multilevel algorithms.

Remark 2.1. The multigrid algorithms are often defined in terms of inner products $(\cdot, \cdot)_k$ which may vary as a function of k . In this case, $(\cdot, \cdot)_k$ replaces the (\cdot, \cdot) inner product in (2.1) and the operator Q_{k-1} is replaced by $Q'_{k-1}: M_k \mapsto M_{k-1}$ defined by

$$(Q'_{k-1}v, \psi)_{k-1} = (v, \psi)_k \quad \text{for all } \psi \in M_{k-1}.$$

The reason for introducing (possibly discrete) inner products on each level is that it may appear that the projection Q_k requires the inversion of Gram matrices. In fact, for appropriately defined smoothers [6], this inversion is avoided and Q_k never explicitly appears in the computational algorithm [9].

3. A GENERAL FRAMEWORK FOR THE ANALYSIS OF MULTIGRID ALGORITHMS

We provide a general theory for multigrid algorithms in this section which is based on a number of abstract assumptions. Two of these assumptions are different from those used in earlier analyses of multigrid algorithms. In later sections, we will apply this theory to prove stronger results concerning the convergence rate of multigrid algorithms in certain applications.

We first describe the new assumptions. The first is much weaker than the full regularity and approximation assumption (cf. [7]). Let λ_k denote the largest eigenvalue of A_k . The new assumption is that there exists a constant $C_0 \geq 1$ satisfying

$$(3.1) \quad A(v, v) \leq C_0 \left[A(P_1 v, v) + \sum_{k=2}^J \frac{\|\tilde{A}_k \tilde{P}_k v\|^2}{\lambda_k} \right] \quad \text{for all } v \in M_J.$$

Remark 3.1. The full regularity and approximation assumption is that there is a constant C_K not depending on k such that

$$(3.2) \quad A((I - P_{k-1})u, u) \leq C_K \frac{\|A_k u\|^2}{\lambda_k} \quad \text{for all } u \in M_k.$$

Let v be in M_J . Taking $u = P_k v$ in (3.2) and summing over k gives

$$A(v, v) \leq A(P_1 v, v) + C_K \sum_{k=2}^J \frac{\|A_k P_k v\|^2}{\lambda_k} \quad \text{for all } v \in M_J.$$

Thus, (3.2) implies (3.1) in the case when $\tilde{M}_k = M_k$. In general, the converse is not true.

The following lemma, which will be crucial in applying the general theory, illustrates that the above assumption is much weaker than the standard full regularity and approximation assumption.

Lemma 3.1. *Let A and \mathcal{A} be equivalent quadratic forms on M_J . By this we mean that there are positive constants c_0 and c_1 (not depending on J) satisfying*

$$(3.3) \quad c_0 A(v, v) \leq \mathcal{A}(v, v) \leq c_1 A(v, v) \quad \text{for all } v \in M_J.$$

Let $\tilde{\mathcal{A}}_k$, $\tilde{\mathcal{P}}_k$, $\tilde{\mathcal{P}}_k$, and Λ_k be the quantities defined with respect to \mathcal{A} corresponding to \tilde{A}_k , P_k , \tilde{P}_k , and λ_k . Assume that (3.1) holds. Then

$$(3.4) \quad \mathcal{A}(v, v) \leq C_0 c_1 / c_0 \left[\mathcal{A}(\tilde{\mathcal{P}}_1 v, v) + \sum_{k=2}^J \frac{\|\tilde{\mathcal{A}}_k \tilde{\mathcal{P}}_k v\|^2}{\Lambda_k} \right] \quad \text{for all } v \in M_J.$$

Proof. Let v be in M_J . We note that (3.1) can be rewritten as

$$A(v, v) \leq C_0 \left[(A_1^{-1} Q_1 A_J v, A_J v) + \sum_{k=2}^J \frac{\|\bar{Q}_k A_j v\|^2}{\lambda_k} \right].$$

Setting $w = A_J v$ gives the equivalent inequality

$$(A_J^{-1} w, w) \leq C_0 \left[(A_1^{-1} Q_1 w, w) + \sum_{k=2}^J \frac{\|\bar{Q}_k w\|^2}{\lambda_k} \right].$$

It then follows from (3.3) that

$$c_0 (\mathcal{A}_J^{-1} w, w) \leq C_0 c_1 \left[(\mathcal{A}_1^{-1} Q_1 w, w) + \sum_{k=2}^J \frac{\|\bar{Q}_k w\|^2}{\Lambda_k} \right] \quad \text{for all } w \in M_J.$$

This is just a restatement of (3.4) and hence the proof is complete. \square

Remark 3.2. We allow for the constant C_0 appearing in (3.1) to depend on J . The results of the general theorems will always depend in a simple way on this constant. We will provide applications where (3.1) can be proved with C_0 independent of J , even though it is known that, for these applications, the corresponding C_K in (3.2) must tend to infinity. It is also shown in [12] that (3.1) holds with $C_0 = CJ$ for many applications. \square

For $k = 2, \dots, J$, set $\tilde{T}_k = \lambda_k^{-1} \tilde{A}_k \tilde{P}_k$ and $\tilde{T}_1 = P_1$. The second assumption is that the operator \tilde{T}_k is “small” when applied to functions in M_l with $l \leq k$. More precisely, we assume that there is a positive number $\varepsilon < 1$ and a positive constant \tilde{C} satisfying

$$(3.5) \quad A(\tilde{T}_k w, w) \leq (\tilde{C}\varepsilon^{k-l})^2 A(w, w) \quad \text{for all } w \in M_l.$$

Additional assumptions required for the theory are standard and will be stated when needed. However, we note that (3.1) can be rewritten as

$$(3.6) \quad A(v, v) \leq C_0 \sum_{k=1}^J A(\tilde{T}_k v, v) \quad \text{for all } v \in M_J.$$

The first theorem of this section provides an estimate for the condition number associated with the additive multilevel method. For this result, we use the following hypothesis on the smoothing operator: For $k = 2, \dots, J$, we assume that R_k is a symmetric operator with respect to (\cdot, \cdot) and satisfies

$$(3.7) \quad C_1 \frac{\|w\|^2}{\lambda_k} \leq (R_k w, w) \leq C_2 \frac{\|w\|^2}{\lambda_k} \quad \text{for all } w \in \tilde{M}_k.$$

Without loss of generality, we assume that $C_1 \leq 1 \leq C_2$. Note that for $k > 1$,

$$A(\tilde{T}_k v, v) = \frac{\|\tilde{A}_k \tilde{P}_k v\|^2}{\lambda_k},$$

and hence (3.7) implies

$$(3.8) \quad C_1 A(\tilde{T}_k v, v) \leq A(T_k v, v) \leq C_2 A(\tilde{T}_k v, v) \quad \text{for all } v \in M_J.$$

Theorem 3.1. *Assume that R_k satisfies (3.7) and that (3.1) and (3.5) hold. Then the condition number $K(B_J^a A_J)$ satisfies*

$$K(B_J^a A_J) \leq \left(\frac{\tilde{C}}{1-\varepsilon} \right)^2 \frac{C_2 C_0}{C_1}.$$

Proof. It suffices to estimate the constants c_1 and c_2 satisfying the inequalities

$$(3.9) \quad c_1 A(\psi, \psi) \leq A(B_J^a A_J \psi, \psi) \leq c_2 A(\psi, \psi) \quad \text{for all } \psi \in M_J.$$

We first bound the sum on the right-hand side of (3.6). Clearly, for $\psi \in M_J$,

$$\sum_{k=1}^J A(\tilde{T}_k \psi, \psi) = \sum_{k=1}^J \sum_{l=1}^k A(\tilde{T}_k \psi, (P_l - P_{l-1})\psi),$$

with $P_0 = 0$. Applying the Schwarz inequality with respect to $A(\tilde{T}_k \cdot, \cdot)$ and (3.5) gives that

$$(3.10) \quad \sum_{k=1}^J A(\tilde{T}_k \psi, \psi) \leq \tilde{C} \sum_{k=1}^J \sum_{l=1}^k \varepsilon^{k-l} A(\tilde{T}_k \psi, \psi)^{1/2} A((P_l - P_{l-1})\psi, \psi)^{1/2}.$$

Let $\mathcal{E} = 1 + \varepsilon + \varepsilon^2 + \dots = (1 - \varepsilon)^{-1}$. Then for $\{\alpha_k\}$, $\{\beta_k\}$ arbitrary real vectors, the following inequality is elementary:

$$(3.11) \quad \begin{aligned} \sum_{k=1}^J \sum_{l=1}^k \varepsilon^{k-l} \alpha_k \beta_l &\leq \frac{\eta}{\mathcal{E}} \sum_{k=1}^J \sum_{l=1}^k \varepsilon^{k-l} \alpha_k^2 + \frac{\mathcal{E}}{4\eta} \sum_{k=1}^J \sum_{l=1}^k \varepsilon^{k-l} \beta_l^2 \\ &\leq \eta \sum_{k=1}^J \alpha_k^2 + \frac{\mathcal{E}^2}{4\eta} \sum_{k=1}^J \beta_k^2, \end{aligned}$$

for $\eta > 0$. Combining (3.10) and (3.11) gives

$$\sum_{k=1}^J A(\tilde{T}_k \psi, \psi) \leq \tilde{C} \eta \sum_{k=1}^J A(\tilde{T}_k \psi, \psi) + \frac{\tilde{C} \mathcal{E}^2}{4\eta} A(\psi, \psi).$$

Taking $\eta = (2\tilde{C})^{-1}$ above and (3.6) show that

$$(3.12) \quad C_0^{-1} A(\psi, \psi) \leq \sum_{k=1}^J A(\tilde{T}_k \psi, \psi) \leq \tilde{C}^2 \mathcal{E}^2 A(\psi, \psi).$$

Combining (3.12) with (3.8) shows that (3.9) holds with $c_1 = C_1/C_0$ and $c_2 = C_2 \tilde{C}^2 \mathcal{E}^2$. This completes the proof of the theorem. \square

The following corollary is an obvious consequence of the theorem. It shows that we may reduce the analysis of the additive algorithm to that for any equivalent quadratic form.

Corollary 3.1. *Assume that R_k satisfies (3.7) and that (3.1) and (3.5) hold. Let \mathcal{A} be an equivalent quadratic form on M_J (satisfying (3.3)) and \mathcal{B}_J^a be the additive preconditioner corresponding to \mathcal{A} , i.e.,*

$$\mathcal{B}_J^a = \mathcal{A}_1^{-1} Q_1 + \sum_{k=2}^J R_k Q_k.$$

Then the condition number $K(\mathcal{B}_J^a \mathcal{A}_J)$ is bounded independently of J .

Remark 3.3. The upper inequality of (3.7) does not hold for many smoothing operators but it does hold for the additive point smoother (cf. [12, inequality (4.3)]). The additive point smoother is defined on spaces \tilde{M}_k (of dimension \tilde{N}_k) with a nodal finite element basis $\{\phi_k^i\}$ by

$$(3.13) \quad R_k v = \sum_{i=1}^{\tilde{N}_k} A(\phi_k^i, \phi_k^i)^{-1} (v, \phi_k^i) \phi_k^i.$$

Note that the upper inequality in (3.7) fails to hold with general R_k if R_k is too much like A_k^{-1} . In fact, the largest eigenvalue of B_J^a is J when $R_k = A_k^{-1}$, for $k = 1, \dots, J$. This shows that the convergence of the additive multilevel algorithm may deteriorate if the smoothers do not behave like point methods. For example, the upper inequality in (3.7) holds for point and line relaxation schemes but fails to hold for some block schemes. In contrast, the analysis for the multiplicative version (standard multigrid) will not require the upper inequality of (3.7) and no convergence deterioration can be seen in the resulting algorithms. \square

We next provide an analysis of the multiplicative form of the multigrid algorithm (Algorithm 2.1). For the multigrid algorithms, we shall also allow the use of nonsymmetric smoothers. In this case, the lower inequality of (3.7) is replaced by the following conditions on the smoother.

(C.1) There is a constant $C_R \geq 1$ which does not depend on k such that the smoothing procedure satisfies

$$(3.14) \quad \frac{\|u\|^2}{\lambda_k} \leq C_R (\bar{R}_k u, u) \quad \text{for all } u \in \tilde{M}_k.$$

Here, $\bar{R}_k = (I - K_k^* K_k) A_k^{-1}$ (recalling that $K_k = I - R_k A_k$). Note that (3.14) holds with $C_R = 1$ for $k = 1$ since $\bar{R}_1 = A_1^{-1}$. In addition, smoothers in multigrid algorithms must be properly scaled as stated in the following condition.

(C.2) There is a constant $\theta < 2$ not depending on k such that

$$(3.15) \quad A(T_k v, T_k v) \leq \theta A(T_k v, v) \quad \text{for all } v \in M_k.$$

Finally, \tilde{M}_k should be an invariant subspace under R_k' . Explicitly, we require that

$$(C.3) \quad R_k = R_k \bar{Q}_k.$$

The above conditions are shown to hold in [6] for the smoothing operators corresponding to many variations (including line- and point-based schemes) of Jacobi and Gauss-Seidel iterative procedures. It is easy to see that the lower inequality of (3.7) implies (3.14) (with a slightly different constant) in the case of symmetric R_k . In addition, (C.3) is automatically satisfied for symmetric smoothers since, by definition, the range of R_k is contained in \tilde{M}_k .

The following theorem provides the general convergence result for the multigrid algorithm.

Theorem 3.2. *Let R_k satisfy (C.1)–(C.3) and assume that (3.1) and (3.5) hold. Then*

$$(3.16) \quad 0 \leq A((I - B_J^m A_J)v, v) \leq (1 - 1/C_M)A(v, v) \quad \text{for all } v \in M_J$$

holds for

$$(3.17) \quad C_M = 2C_0 \left(C_R + \frac{\tilde{C}^2 \varepsilon^2}{(1 - \varepsilon)^2} \frac{\theta}{2 - \theta} \right).$$

Proof. Set $E_0 = I$, and for $k = 1, \dots, J$, define

$$E_k = (I - T_k)(I - T_{k-1}) \cdots (I - T_1).$$

Note that by (2.14),

$$A((I - B_J^m A_J)v, v) = A(E_J^* v, E_J^* v),$$

and the lower inequality of (3.16) follows. The proof of the upper inequality requires bounding the norm of E_J^* or, equivalently, the norm of its adjoint E_J .

We first derive some identities involving the above operators. Clearly, for $k = 1, \dots, J$,

$$(3.18) \quad E_{k-1} - E_k = T_k E_{k-1},$$

from which it follows that

$$(3.19) \quad I - E_k = \sum_{m=1}^k T_m E_{m-1}.$$

It is obvious from (3.18) that

$$(3.20) \quad A(E_{k-1} v, E_{k-1} v) - A(E_k v, E_k v) = A((2I - T_k)E_{k-1} v, T_k E_{k-1} v).$$

Let $\bar{T}_k = \bar{R}_k A_k P_k = (I - K_k^* K_k)P_k$. Clearly,

$$(3.21) \quad A(\bar{T}_k E_{k-1} v, E_{k-1} v) = A((2I - T_k)E_{k-1} v, T_k E_{k-1} v).$$

Summing (3.20) gives that

$$(3.22) \quad A(v, v) - A(E_J v, E_J v) = \sum_{k=1}^J A(\bar{T}_k E_{k-1} v, E_{k-1} v).$$

Note that the upper inequality of (3.16) immediately follows if we prove that

$$(3.23) \quad A(v, v) \leq C_M [A(v, v) - A(E_J v, E_J v)].$$

Thus, the proof reduces to showing that $A(v, v)$ can be bounded by C_M times the sum on the right-hand side of (3.22).

By (3.6),

$$(3.24) \quad \begin{aligned} A(v, v) &\leq C_0 \sum_{k=1}^J A(\tilde{T}_k v, v) \\ &\leq 2C_0 \left(\sum_{k=1}^J A(\tilde{T}_k E_{k-1} v, E_{k-1} v) \right. \\ &\quad \left. + \sum_{k=2}^J A(\tilde{T}_k(I - E_{k-1})v, (I - E_{k-1})v) \right). \end{aligned}$$

For $k = 2, \dots, J$, condition (C.1) implies

$$A(\tilde{T}_k E_{k-1} v, E_{k-1} v) = \frac{\|\tilde{A}_k \tilde{P}_k E_{k-1} v\|^2}{\lambda_k} \leq C_R (\bar{R}_k \tilde{A}_k \tilde{P}_k E_{k-1} v, \tilde{A}_k \tilde{P}_k E_{k-1} v).$$

By (C.3), the image of \bar{R}_k is in \bar{M}_k and $\bar{R}_k \bar{Q}_k = \bar{R}_k$. Consequently,

$$(\bar{R}_k \tilde{A}_k \tilde{P}_k E_{k-1} v, \tilde{A}_k \tilde{P}_k E_{k-1} v) = A(\bar{T}_k E_{k-1} v, E_{k-1} v).$$

Moreover, since $\tilde{T}_1 = \bar{T}_1$,

$$(3.25) \quad \sum_{k=1}^J A(\tilde{T}_k E_{k-1} v, E_{k-1} v) \leq C_R \sum_{k=1}^J A(\bar{T}_k E_{k-1} v, E_{k-1} v).$$

For the second sum on the right-hand side of (3.24), we use (3.19) and (3.5) to get

$$\begin{aligned} \sum_{k=2}^J A(\tilde{T}_k(v - E_{k-1} v), v - E_{k-1} v) &= \sum_{k=2}^J \sum_{l=1}^{k-1} \sum_{m=1}^{k-1} A(\tilde{T}_k T_l E_{l-1} v, T_m E_{m-1} v) \\ &\leq \tilde{C}^2 \sum_{k=2}^J \sum_{l=1}^{k-1} \sum_{m=1}^{k-1} \epsilon^{2k-l-m} A(T_l E_{l-1} v, T_l E_{l-1} v)^{1/2} A(T_m E_{m-1} v, T_m E_{m-1} v)^{1/2} \\ &\leq \frac{\tilde{C}^2}{2} \sum_{k=2}^J \sum_{l=1}^{k-1} \sum_{m=1}^{k-1} \epsilon^{2k-l-m} [A(T_l E_{l-1} v, T_l E_{l-1} v) + A(T_m E_{m-1} v, T_m E_{m-1} v)]. \end{aligned}$$

It suffices to bound either of the two terms on the right-hand side above. By

(C.2),

$$\begin{aligned}
& \sum_{k=2}^J \sum_{l=1}^{k-1} \sum_{m=1}^{k-1} \varepsilon^{2k-l-m} A(T_l E_{l-1} v, T_l E_{l-1} v) \\
& \leq \theta \varepsilon \mathcal{E} \sum_{k=2}^J \sum_{l=1}^{k-1} \varepsilon^{k-l} A(T_l E_{l-1} v, E_{l-1} v) \\
& = \theta \varepsilon \mathcal{E} \sum_{l=1}^{J-1} \sum_{k=l+1}^J \varepsilon^{k-l} A(T_l E_{l-1} v, E_{l-1} v) \\
& \leq \theta \varepsilon^2 \mathcal{E}^2 \sum_{l=1}^{J-1} A(T_l E_{l-1} v, E_{l-1} v).
\end{aligned}$$

Using (C.2) and (3.21) gives that

$$A(T_l E_{l-1} v, E_{l-1} v) \leq (2 - \theta)^{-1} A(\bar{T}_l E_{l-1} v, E_{l-1} v).$$

Thus,

$$\begin{aligned}
& \sum_{k=2}^J A(\tilde{T}_k(v - E_{k-1} v), v - E_{k-1} v) \\
(3.26) \quad & \leq \frac{\tilde{C}^2 \varepsilon^2}{(1 - \varepsilon)^2} \frac{\theta}{2 - \theta} \sum_{k=1}^{J-1} A(\bar{T}_k E_{k-1} v, E_{k-1} v).
\end{aligned}$$

Combining (3.24), (3.25), and (3.26) shows that (3.23) holds with C_M given by (3.17). This completes the proof of the theorem. \square

Remark 3.4. The above theorem holds for many generalizations of the multigrid algorithm given in Algorithm 2.1. For example, it holds for W-cycle and other algorithms which use more than one iteration for the coarse-grid correction (Step 2). The proof follows from an argument given in §2 of [9]. The result also holds for algorithms which use more smoothings per grid level as long as one alternates between R_k and R'_k (see, [6, 11]). The modification to the proof is minor and illustrated in the proof of Theorem 4.3 of [6]. Finally, an analogous contraction result holds for nonsymmetric cycling algorithms where smoothing is only done either before or after the correction step, i.e., Step 1 or Step 3 is skipped.

4. THE QUASI-UNIFORM FINITE ELEMENT APPROXIMATION

In this section, we verify the hypotheses for the general multigrid theory in the case of a model second-order elliptic problem. We first describe the model problem and its finite element approximation. In particular, a nested sequence of quasi-uniform approximation spaces are defined in a standard fashion. Next, some notation concerning Sobolev spaces and the corresponding norms is provided. Finally, the conditions (3.1) and (3.5) are shown to hold with constants that are independent of the mesh parameters. Application of the general theory of the previous section then implies that the multilevel algorithms converge with rates that are independent of the number of levels, even in many examples which do not satisfy full elliptic regularity.

Let Ω be a bounded domain in R^d with polygonal boundary $\partial\Omega$. We will include the case when $\Omega \subset R^2$ is a domain with a crack. We consider the Dirichlet problem

$$(4.1) \quad \begin{aligned} Lu &= f && \text{in } \Omega, \\ u &= 0 && \text{on } \partial\Omega, \end{aligned}$$

where

$$Lv = - \sum_{i,j=1}^d \frac{\partial}{\partial x_i} \left(a_{ij} \frac{\partial v}{\partial x_j} \right) + av.$$

Let $\bar{\Omega} = \bigcup \bar{\Omega}_l$, for a fixed number of subdomains $\{\Omega_l\}$, where each Ω_l has a Lipschitz continuous boundary. For each i, j, l assume that a_{ij} is in $W_p^\gamma(\Omega_l)$ for some $\gamma \in (0, 1/2)$ and $p > d/\gamma$. Here, $W_p^\gamma(\Omega)$ is the Sobolev space of order γ defined in terms of the norm $L^p(\Omega)$ (cf. [17]). This condition implies that the coefficients are continuous on Ω_l but may jump across the boundaries. In the case of the additive multilevel algorithm, this assumption can be replaced by the assumption that the functions a_{ij} are in $L^\infty(\Omega)$. We further assume that the matrix $\{a_{ij}(x)\}$ is uniformly positive definite almost everywhere. In addition, we assume, for simplicity, that $a(x) \in L^\infty(\Omega)$ is nonnegative.

Remark 4.1. We impose Dirichlet boundary conditions in the above problem for simplicity. The techniques provided in this section can also be applied to problems with mixed boundary conditions. An example of such an application is given at the end of this section. \square

The form $A(\cdot, \cdot)$ is defined for this example by the generalized Dirichlet form corresponding to (4.1), i.e.,

$$(4.2) \quad A(v, w) = \sum_{i,j=1}^d \int_{\Omega} a_{ij} \frac{\partial v}{\partial x_i} \frac{\partial w}{\partial x_j} dx + \int_{\Omega} avw dx.$$

This is defined for all v and w in the Sobolev space $H^1(\Omega)$ (the space of distributions with square-integrable first derivatives).

Note that by the assumptions on the coefficients in (4.1), the quadratic form $A(\cdot, \cdot)$ is uniformly equivalent to the form corresponding to the constant-coefficient operator $-\Delta$. Thus, for the purpose of proving (3.1), it suffices to consider the case of the Laplacian. We will assume that there is an α in $(0, 1]$ such that solutions u of (4.1) with $L = -\Delta$ satisfy the following regularity estimate :

$$(4.3) \quad \|u\|_{1+\alpha} \leq C \|f\|_{-1+\alpha}.$$

Here, $\|\cdot\|_{-1+\alpha}$ is the interpolated norm between $L^2(\Omega)$ and $H^{-1}(\Omega)$ (the dual of $H_0^1(\Omega)$). Thus, we assume that the domain results in some elliptic regularity for smooth coefficient problems (but not necessarily full elliptic regularity). This assumption is weak, since (4.3) may not hold for any $\alpha > 0$ for equations with bad coefficients. Such an inequality holds for plane domains with polygonal boundaries, including domains with cracks (cf. [15]).

We shall consider the case of quasi-uniform finite element approximation of the solution of (4.1). To define the approximation spaces, we will first define the

underlying mesh partitioning. We assume that a unit-size coarse finite element partitioning of the original domain Ω is given ($\overline{\Omega} = \bigcup \tau_1^i$). For example, we take this partitioning in terms of triangles in the case of two spatial dimensions. For examples of such constructions see [14].

Associated with the mesh partitioning, we are given a rule for refinement. For example, in two dimensions, a triangle can be refined into four by connecting the midpoints of the edges.

The mesh triangulations can now be defined by mathematical induction. The coarse triangulation defined above provides the first grid $\{\tau_1^m\}$. Given that a grid $\{\tau_{k-1}^m\}$ has been defined, the grid $\{\tau_k^m\}$ is defined by refining $\{\tau_{k-1}^m\}$ using the refinement rule. We assume that the mesh size of the k th triangulation is on the order of δ^k for fixed $\delta < 1$.

The finite element space M_k is defined to be a space of piecewise polynomial functions with respect to the mesh $\{\tau_k^i\}$ which are continuous on Ω and vanish on $\partial\Omega$ (cf. [1, 14]). Note that a nodal finite element basis is not required for the application of this section. In the present example, $\widetilde{M}_k = M_k$, i.e., we smooth on all functions of M_k . The simplest two-dimensional case is when M_k consists of functions which are piecewise linear with respect to the triangulation and the refinements are defined by breaking triangles into four by connecting the midpoints of the edges.

We next define the Galerkin approximation to the solution of (4.1). Multiplying (4.1) by a smooth function ϕ which vanishes on $\partial\Omega$, and integrating by parts, gives that u satisfies

$$A(u, \phi) = (f, \phi).$$

Here, (\cdot, \cdot) denotes the $L^2(\Omega)$ inner product. The Galerkin approximation to u is the unique function $u_J \in M_J$ satisfying

$$(4.4) \quad A(u_J, \psi) = (f, \psi) \quad \text{for all } \psi \in M_J.$$

We shall analyze the multilevel iterative methods of §2 for solving (4.4).

To verify the hypotheses of §3, we require some notation for norms in Sobolev spaces. Let s be a nonnegative real number and $\|\cdot\|_{s, \widetilde{\Omega}}$ denote the Sobolev norm of order s on a domain $\widetilde{\Omega}$ (cf. [17, 19]). The $\widetilde{\Omega}$ will be left out of this notation when $\widetilde{\Omega} = \Omega$, and the s will be left out in the case of $L^2(\widetilde{\Omega})$ ($s = 0$).

Remark 4.2. The above assumptions on the mesh sizes imply that λ_k (where A_k is defined by (2.2)) is on the order of δ^{-2k} .

Remark 4.3. For this example, it was shown in [12] that (3.1) holds with $C_0 = CJ$. The next lemma improves this bound and shows that (3.1) holds with $C_0 = C$ (independent of J). This inequality was proved earlier in [22] using Besov space techniques. We base our proof on the well-known fact that a wide class of domains give rise to elliptic regularity.

Lemma 4.1. *Assume that (4.3) holds for $L = -\Delta$ and some $\alpha \in (0, 1]$. Then, for general L , there exists a positive constant C_0 not depending on the number of mesh levels J such that (3.1) holds.*

Proof. Here and in the remainder of this paper, we shall use c with or without subscript to denote a generic positive constant. Such constants will always be

independent of the number of levels in the multilevel algorithm. As already observed, by Lemma 3.1, it suffices to prove (3.1) with $L = -\Delta$. Clearly, for $w \in M_J$,

$$\begin{aligned} A(w, w) &= A(w, Q_1 w) + \sum_{k=2}^J (A_k P_k w, (Q_k - Q_{k-1})w) \\ &\leq \left(A(P_1 w, w) + \sum_{k=2}^J \frac{\|A_k P_k w\|^2}{\lambda_k} \right)^{1/2} \\ &\quad \left(A(Q_1 w, Q_1 w) + \sum_{k=2}^J \lambda_k \|(Q_k - Q_{k-1})w\|^2 \right)^{1/2}. \end{aligned}$$

Consequently, it suffices to show that

$$(4.5) \quad A(Q_1 w, Q_1 w) + \sum_{k=2}^J \lambda_k \|(Q_k - Q_{k-1})w\|^2 \leq c A(w, w).$$

The following approximation and boundedness properties for the operators $\{Q_k\}$ are well known: For $v \in H_0^1(\Omega)$,

$$(4.6) \quad \begin{aligned} \|(Q_k - Q_{k-1})v\|^2 &\leq c \lambda_k^{-1} A(v, v) \quad \text{for } k = 2, \dots, J, \\ A(Q_k v, Q_k v) &\leq c A(v, v) \quad \text{for } k = 1, \dots, J. \end{aligned}$$

In addition, it is well known that (4.3) and the approximation properties of the finite element spaces imply that

$$(4.7) \quad \|(I - P_{k-1})v\|_{1-\alpha}^2 \leq c \lambda_k^{-\alpha} A((I - P_{k-1})v, v).$$

Here we have used the fact that $\lambda_k \leq ch_k^{-2}$ for this application. From (4.6), it suffices to bound the terms appearing in the sum on the left-hand side of (4.5). Let β satisfy $0 < \beta < \alpha$. Then

$$(4.8) \quad \begin{aligned} \sum_{k=2}^J \lambda_k \|(Q_k - Q_{k-1})w\|^2 &= \sum_{k=2}^J \lambda_k \left\| \sum_{l=k}^J (Q_k - Q_{k-1})(P_l - P_{l-1})w \right\|^2 \\ &\leq \sum_{k=2}^J \lambda_k \left(\sum_{l=k}^J \lambda_l^{-\beta} \right) \left(\sum_{l=k}^J \lambda_l^\beta \|(Q_k - Q_{k-1})(P_l - P_{l-1})w\|^2 \right). \end{aligned}$$

By Remark 4.2, the first sum over l above is bounded by $c \lambda_k^{-\beta}$. Thus, the boundedness of Q_k on $L^2(\Omega)$, the first inequality of (4.6), and interpolation gives

$$\sum_{k=2}^J \lambda_k \|(Q_k - Q_{k-1})w\|^2 \leq c \sum_{k=2}^J \lambda_k^{\alpha-\beta} \sum_{l=k}^J \lambda_l^\beta \|(P_l - P_{l-1})w\|_{1-\alpha}^2.$$

Applying (4.7) and changing the order of summation gives

$$\begin{aligned} \sum_{k=2}^J \lambda_k \|(Q_k - Q_{k-1})w\|^2 &\leq c \sum_{k=2}^J \lambda_k^{\alpha-\beta} \sum_{l=k}^J \lambda_l^{\beta-\alpha} A((P_l - P_{l-1})w, w) \\ &= c \sum_{l=2}^J \lambda_l^{\beta-\alpha} A((P_l - P_{l-1})w, w) \sum_{k=2}^l \lambda_k^{\alpha-\beta}. \end{aligned}$$

Since α is greater than β , Remark 4.2 implies that the sum over k above is bounded by a constant times $\lambda_l^{\alpha-\beta}$. Thus,

$$\sum_{k=2}^J \lambda_k \| (Q_k - Q_{k-1})w \|^2 \leq c \sum_{l=2}^J A((P_l - P_{l-1})w, w) \leq cA(w, w).$$

This completes the proof of the lemma. \square

Remark 4.4. We note that replacing $(Q_k - Q_{k-1})$ by $(I - Q_{k-1})$ in (4.8) and following the proof of the lemma gives that

$$\sum_{k=2}^J \lambda_k \| (I - Q_{k-1})w \|^2 \leq cA(w, w).$$

This inequality will be used in §5. \square

The next lemma provides a proof of (3.5) for the application described above.

Lemma 4.2. *Let l be less than or equal to k . Then there is a constant \tilde{C} not depending on the mesh parameters satisfying*

$$(4.9) \quad A(\tilde{T}_k v, v) \leq \tilde{C}(h_k/h_l)^{2\gamma} A(v, v) \quad \text{for all } v \in M_l.$$

Proof. The proof of Lemma 4.2 is based on the following lemma. Its proof will be given after the proof of Lemma 4.2.

Lemma 4.3. *There exists a constant c such that for all $\eta > 0$, $\phi \in H^1(\Omega)$ and $\psi \in H^{1+\gamma}(\Omega)$,*

$$(4.10) \quad |A(\phi, \psi)| \leq c(\eta^{-1} \|\phi\|^2 + \eta^{\gamma/(1-\gamma)} \|\phi\|_1^2)^{1/2} \|\psi\|_{1+\gamma}.$$

Assuming Lemma 4.3, we now complete the proof of Lemma 4.2. Let $w \in M_l$. We clearly have that for $k > 2$,

$$(4.11) \quad A(\tilde{T}_k w, w) = \frac{\|A_k w\|^2}{\lambda_k} = \lambda_k^{-1} \left(\sup_{\phi \in M_k} \frac{A(w, \phi)}{\|\phi\|} \right)^2.$$

By Lemma 4.3,

$$\begin{aligned} |A(w, \phi)| &\leq c(\eta^{-1} \|\phi\|^2 + \eta^{\gamma/(1-\gamma)} \|\phi\|_1^2)^{1/2} \|w\|_{1+\gamma} \\ &\leq c(\eta^{-1} + \eta^{\gamma/(1-\gamma)} h_k^{-2})^{1/2} \|\phi\| \|w\|_{1+\gamma}. \end{aligned}$$

Taking $\eta = h_k^{2(1-\gamma)}$ and using the inverse property (see the Appendix of [11] for a proof)

$$\|w\|_{1+\gamma} \leq c h_l^{-\gamma} \|w\|_1$$

gives

$$(4.12) \quad |A(w, \phi)| \leq c h_k^{-1+\gamma} h_l^{-\gamma} \|w\|_1 \|\phi\|.$$

Inequality (4.9) follows combining (4.11), (4.12) and the fact that $\lambda_k \geq c h_k^{-2}$. This completes the proof of Lemma 4.2. \square

Proof of Lemma 4.3. This lemma was essentially given in [5] and we will follow its proof. However, the version stated here gives a somewhat more explicit form

of the bounds and requires less regularity on the coefficients than that of Lemma 4.1 of [5].

Let ϕ be in $H^1(\Omega)$ and ψ be in $H^{1+\gamma}(\Omega)$. There is no problem bounding the lowest-order term of (4.2). We need only consider the derivative terms in (4.2). Fix l , and let \mathbf{E}_l denote the extension operator defined on $L^2(\Omega_l)$ given by Theorem 1.4.3.1 of [17]. For a function v defined on Ω_l , let \tilde{v} denote the extension of v by zero to R^d . Since $\gamma < 1/2$, Corollary 1.4.4.5 of [17] gives that the norm $\|\tilde{v}\|_{W_2^\gamma(R^d)}$ is equivalent to the norm $\|v\|_{\gamma, \Omega_l}$ for all $v \in H^\gamma(\Omega_l)$. Thus,

$$\int_{\Omega_l} a_{ij} \frac{\partial \phi}{\partial x_i} \frac{\partial \psi}{\partial x_j} dx = \left(\mathcal{F} \left(\frac{\partial (\mathbf{E}_l \phi)}{\partial x_i} \right), \mathcal{F} \left(a_{ij} \frac{\partial \psi}{\partial x_j} \right) \right),$$

where \mathcal{F} denotes the Fourier transform. By the Schwarz inequality,

$$\begin{aligned} (4.13) \quad & \int_{\Omega_l} a_{ij} \frac{\partial \phi}{\partial x_i} \frac{\partial \psi}{\partial x_j} dx \\ & \leq c \left(\int_{R^d} \frac{|\zeta|^2}{(1+|\zeta|^2)^\gamma} |\mathcal{F}(\mathbf{E}_l \phi)(\zeta)|^2 d\zeta \right)^{1/2} \left\| a_{ij} \frac{\partial \psi}{\partial x_j} \right\|_{\gamma, \Omega_l} \\ & \leq c(\eta^{-1} \|\phi\|^2 + \eta^{\gamma/(1-\gamma)} \|\phi\|_1^2)^{1/2} \left\| \frac{\partial \psi}{\partial x_j} \right\|_\gamma, \end{aligned}$$

where $\eta > 0$ is arbitrary. For the last inequality, we used Theorem 1.4.4.2 of [17], which states that multiplication by $a_{ij} \in W_p^\gamma(\Omega_l)$ for $p > d/\gamma$ is a bounded operator on $H^\gamma(\Omega_l)$. The lemma immediately follows by summing over l . \square

Remark 4.5. A more constructive proof of Lemma 4.2 is possible in the case of smooth coefficients and nodal finite element approximation spaces. In this case, one proves directly that

$$A(T_k w, w) \leq c(h_k/h_l) A(w, w)$$

holds for $T_k = R_k A_k P_k$, with R_k defined by (3.13). The lemma then follows with $\gamma = 1/2$ from

$$\frac{\|v\|^2}{\lambda_k} \leq c(R_k v, v) \quad \text{for all } v \in M_k,$$

which is a general smoothing property proved in [6].

Combining Lemmas 4.1 and 4.2 with Theorems 3.1, 3.2, and Corollary 3.1 gives the following theorems.

Theorem 4.1. Let B_J^a be defined by (2.4) with $A(\cdot, \cdot)$, (\cdot, \cdot) , and $\{M_k, \widetilde{M}_k\}$ as described in this section. Assume that R_k (which is symmetric) satisfies (3.7). Then the condition number $K(B_J^a A_J)$ is bounded by a constant which is independent of J .

Theorem 4.2. Let B_J^m be defined by Algorithm 2.1 with $A(\cdot, \cdot)$, (\cdot, \cdot) , and $\{M_k, \widetilde{M}_k\}$ as described in this section. Assume that R_k satisfies (C.1)–(C.3). Then

$$(4.14) \quad 0 \leq A((I - B_J^m A_J)v, v) \leq (1 - 1/C_M) A(v, v) \quad \text{for all } v \in M_J.$$

The constant C_M in (4.14) is independent of J .

Remark 4.6. A uniform result similar to the upper estimate in Theorem 4.1 in the case of R_k defined by (3.13) was announced by Zhang at the Fifth International Symposium on Domain Decomposition Methods [23]. In addition, a version of Theorem 4.1 is proved in [22] using Besov-space equivalences.

Remark 4.7. The results of this section hold for many applications with mixed boundary conditions. We illustrate this by considering a simple example. Specifically, we consider (4.1) with $\Omega = (0, 1) \times (0, 2)$ but with the boundary conditions

$$\begin{aligned} 2 \frac{\partial u}{\partial v} &= 0 && \text{on } \Gamma_N, \\ u &= 0 && \text{on } \partial\Omega/\Gamma_N. \end{aligned}$$

Here, Γ_N is the line segment $\{(1, y) | y \in (1, 2)\}$ and $\frac{\partial}{\partial v}$ denotes the outward conormal derivative on Γ_N . There is no problem with the proof of Lemma 4.2 in this case since the basic ingredient in the proof, Lemma 4.3 does not depend on boundary conditions. Furthermore, the resulting form $A(\cdot, \cdot)$ is equivalent to that corresponding to the Laplacian (with boundary condition $\partial u/\partial n = 0$ on Γ_N). The Laplacian with this boundary condition satisfies (4.3) for any α in $(0, 1/2)$. Thus, Lemma 4.1 holds.

5. GENERAL MESH REFINEMENT

In this section, we apply the general theory to an approximation which utilizes a locally refined mesh. Such mesh refinements are convenient for accurate modeling of problems with various types of singular behavior. For simplicity, we will consider the piecewise linear finite element approximation, although we will allow a very general form of refinement.

As in the previous section, we consider problem (4.1) and start with a coarse triangulation. The refinement triangulation $\{\tau_k^i\}$ is defined in terms of a sequence of (open) mesh domains

$$\Omega_J \subseteq \Omega_{J-1} \subseteq \cdots \subseteq \Omega_1 = \Omega.$$

The only restrictions on the mesh domains $\{\Omega_k\}$ are that the boundary of Ω_k , for $k > 1$, consists of edges of mesh triangles in the mesh $\{\tau_{k-1}^i\}$, and that there is at least one edge of $\{\tau_{k-1}^i\}$ contained in Ω_k . These mesh domains control the region of refinement. If τ_{k-1}^i is a triangle contained in Ω_k , then it is broken into four smaller triangles (in the triangulation $\{\tau_k^i\}$) by the lines connecting the midpoints of the edges. Alternatively, if τ_{k-1}^i is in the complement of Ω_k , then it is not subdivided but is directly included into the k th triangulation. A simple example of this construction is the case of a unit square with local refinement near the corner $(1, 1)$. In this case, we take $\Omega_k = \Omega$ for $k = 1, \dots, j$ and $\Omega_k = [1 - 2^{k-j}, 1] \times [1 - 2^{k-j}, 1]$ for $k = j + 1, \dots, J$.

We consider the piecewise linear finite element approximation, although many extensions are possible. The space M_k is defined to be the set of piecewise linear functions with respect to the mesh $\{\tau_k^i\}$ which are continuous on Ω and vanish on $\partial\Omega$. The continuity condition implies that the finer grid nodes on a coarse-fine boundary are slave nodes in the sense that the values of the

function there are completely determined by the values of the function on the nearby coarse grid points.

For this application, if $\Omega_{k-1} \neq \Omega_k$, then the subspace \tilde{M}_k on which we smooth is a proper nonzero subspace of M_k . In fact, we define \tilde{M}_k to be the functions in M_k which are zero outside of Ω_k . Thus, we smooth on a given level just in the region where new nodes are being added in the refinement scheme.

For this application, there is no difficulty in proving (3.5). The largest eigenvalue of M_k is on the order of h_k^{-2} , where h_k is the size of the smallest triangle defining the mesh of M_k . The argument given in the proof of Lemma 4.2 applies with little change and gives that (3.5) holds with a uniform constant \tilde{C} .

To apply Theorems 3.1 and 3.2, we need to prove (3.1). By Lemma 3.1, we need only consider the case when the coefficients defining L are smooth. Let \bar{M}_k denote the quasi-uniform finite element space obtained from refining over the entire domain at each level starting with $\{\tau_1^i\}$, that is, the quasi-uniform space resulting from the above construction with $\Omega = \Omega_1 = \Omega_2 = \dots = \Omega_J$. Let \bar{Q}_k denote the (\cdot, \cdot) orthogonal projection onto \bar{M}_k . A sequence of operators $\mathcal{Q}_k : M_J \mapsto M_k$ was constructed in [9] (see the sequence $\{Q_k\}$ defined in §5 of [9]) which satisfy

- (1) \mathcal{Q}_J is the identity.
- (2) The range of $\mathcal{Q}_k - \mathcal{Q}_{k-1}$ is contained in \tilde{M}_k .
- (3) The inequalities

$$\begin{aligned} \| (I - \mathcal{Q}_k) v \| &\leq c \| (I - \bar{Q}_k) v \|, \\ A(\mathcal{Q}_1 v, \mathcal{Q}_1 v) &\leq c A(v, v) \end{aligned}$$

hold for all $v \in M_J$ and with constant c independent of k and J .

Following the proof of Lemma 4.1, for $w \in M_J$, we have

$$\begin{aligned} A(w, w) &= A(w, \mathcal{Q}_1 w) + \sum_{k=2}^J (\tilde{A}_k \tilde{P}_k w, (\mathcal{Q}_k - \mathcal{Q}_{k-1}) w) \\ &\leq \left(A(P_1 w, w) + \sum_{k=2}^J \frac{\|\tilde{A}_k \tilde{P}_k w\|^2}{\lambda_k} \right)^{1/2} \\ &\quad \left(A(\mathcal{Q}_1 w, \mathcal{Q}_1 w) + \sum_{k=2}^J \lambda_k \|(\mathcal{Q}_k - \mathcal{Q}_{k-1}) w\|^2 \right)^{1/2}. \end{aligned}$$

Consequently, it suffices to show that

$$(5.1) \quad A(\mathcal{Q}_1 w, \mathcal{Q}_1 w) + \sum_{k=2}^J \lambda_k \|(\mathcal{Q}_k - \mathcal{Q}_{k-1}) w\|^2 \leq c A(w, w).$$

Clearly, by (3),

$$\begin{aligned} \|(\mathcal{Q}_k - \mathcal{Q}_{k-1}) w\|^2 &\leq 2 [\|(I - \mathcal{Q}_k) w\|^2 + \|(I - \mathcal{Q}_{k-1}) w\|^2] \\ &\leq c \| (I - \bar{Q}_{k-1}) w \|^2. \end{aligned}$$

Thus,

$$\begin{aligned} A(\mathcal{Q}_1 w, \mathcal{Q}_1 w) + \sum_{k=2}^J \lambda_k \|(\mathcal{Q}_k - \mathcal{Q}_{k-1})w\|^2 \\ \leq c \left(A(w, w) + \sum_{k=2}^J \lambda_k \|(I - \bar{Q}_{k-1})w\|^2 \right). \end{aligned}$$

Inequality (5.1) follows from Remark 4.4. This shows that (3.1) holds for the refinement application of this section.

We can combine the above observations with Theorems 3.1, 3.2 and Corollary 3.1 to get:

Theorem 5.1. *Let B_J^q be defined by (2.4) with $A(\cdot, \cdot)$, (\cdot, \cdot) , and $\{M_k, \widetilde{M}_k\}$ as described in this section. Assume that R_k (which is symmetric) satisfies (3.7). Then the condition number $K(B_J^q A_J)$ is bounded by a constant which is independent of J .*

Theorem 5.2. *Let B_J^m be defined by Algorithm 2.1 with $A(\cdot, \cdot)$, (\cdot, \cdot) , and $\{M_k, \widetilde{M}_k\}$ as described in this section. Assume that R_k satisfies (C.1)–(C.3). Then*

$$(5.2) \quad 0 \leq A((I - B_J^m A_J)v, v) \leq (1 - 1/C_M)A(v, v) \quad \text{for all } v \in M_J.$$

The constant C_M in (5.2) is independent of J .

6. A CURVED BOUNDARY APPLICATION

In this section, we consider applying the theory developed earlier to a finite element approximation of a boundary value problem with a curved boundary. To remain in the framework of nested spaces, we consider coarser-grid multigrid spaces M_k which vanish in a neighborhood of order h_k of the domain boundary. Even though these spaces provide a poor approximation, we will show that they lead to multigrid algorithms which converge with uniform rates of reduction.

For convenience, we shall consider a convex domain in R^2 with smooth boundary. Many extensions are possible. We will consider problem (4.1) with the same assumptions on the coefficients as made in §4. The form $A(\cdot, \cdot)$ is defined by (4.2). We also assume that (4.3) holds for $L = -\Delta$ and some α in $(0, 1]$.

We start with a coarse approximate triangulation $\{\tau_1^i\}$ of Ω . By construction, a node will be either in the interior of Ω or on $\partial\Omega$. Without loss of generality, we assume that no triangle of $\{\tau_1^i\}$ has all three vertices on $\partial\Omega$. The triangulation $\{\tau_k^i\}$ will be defined from $\{\tau_{k-1}^i\}$ as follows:

- (1) If τ_{k-1}^i is a triangle with two vertices in Ω then τ_{k-1}^i is broken into four finer-grid triangles by the lines connecting the centers of the edges.
- (2) A triangle τ_{k-1}^i with two vertices on $\partial\Omega$ results in four finer-grid triangles as illustrated in Figure 6.1. The new boundary point is the midpoint along the boundary arc between the two boundary vertices of τ_{k-1}^i .

Note that not all triangles in $\{\tau_{k-1}^i\}$ can be written as the union of the triangles in $\{\tau_k^i\}$. As a consequence, we will define the coarser multilevel spaces $k < J$ in a different manner than that used for the finest space.

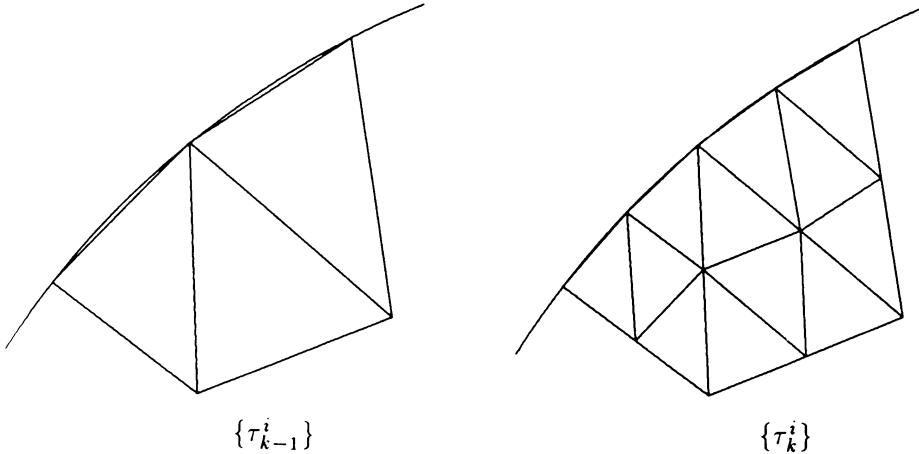


FIGURE 6.1. The mesh refinement near the boundary

We use the above grid strategy to define the sequence of approximation meshes up to J and define Ω_k to be the interior of $\bigcup_i \bar{\tau}_k^i$. The fine-grid approximation space M_J is a space of piecewise linear functions with respect to $\{\tau_J^i\}$ which are continuous on Ω_J and vanish on $\partial\Omega_J$.

For $k < J$, let \bar{M}_k denote the analogous space of piecewise linear functions which are defined in terms of the triangulation $\{\tau_k^i\}$. Let Ω_k^0 , for $k < J$, be the interior of the union of the closures of triangles of $\{\tau_k^i\}$ which do not have vertices on $\partial\Omega$. The spaces M_k for $k < J$, to be used in the multigrid algorithm, are defined by

$$(6.1) \quad M_k = \{\phi \in \bar{M}_k \mid \text{supp}(\phi) \in \bar{\Omega}_k^0\}.$$

Because of the convexity of Ω , it is easy to see that $\Omega_k \subseteq \Omega$ for $k = 1, \dots, J$. Consequently, functions in \bar{M}_k extended by zero are in $H_0^1(\Omega)$. Thus, we shall consider the spaces \bar{M}_k and M_k as being contained in $H_0^1(\Omega)$. Moreover, the triangles of $\{\tau_{k-1}^i\}$ in Ω_k^0 can be written as the union of triangles in $\{\tau_k^i\}$. This implies that the multigrid spaces defined by (6.1) are nested, i.e.,

$$M_1 \subset M_2 \subset \dots \subset M_J \subset H_0^1(\Omega).$$

The argument of Lemma 4.2 goes through without change for this application. To apply the theory of §4, we need only prove (3.1). From the proof of Lemma 4.1, it clearly suffices to prove the inequalities (4.6) and (4.7) for this application. To this end, we introduce the following lemma.

Lemma 6.1. *Let Ω^η denote the strip $\{x \in \Omega \mid \text{dist}(x, \partial\Omega) < \eta\}$ and $0 < s < 1/2$. Then for all $v \in H^{1+s}(\Omega)$,*

$$(6.2) \quad \|v\|_{1, \Omega^\eta} \leq c\eta^s \|v\|_{1+s}.$$

In addition, for $v \in H_0^1(\Omega)$,

$$(6.3) \quad \|v\|_{\Omega^\eta} \leq c\eta \|v\|_1.$$

Proof. Note that if ω^η is a reference square of side length η , then

$$\|w\|_{\omega^\eta}^2 \leq \eta^2/2 \int_{\omega^\eta} |\nabla w|^2 dx$$

holds for all functions w vanishing on one edge of ω^η . Inequality (6.3) then easily follows by the use of local maps. A relatively straightforward interpolation argument (see Chapter 1, Section 15 of [19]) using the real method [19] between (6.3) and the trivial inequality $\|w\|_{\Omega^\eta} \leq \|w\|$ gives that (for $s \neq 1/2$)

$$(6.4) \quad \|w\|_{\Omega^\eta} \leq c(s)\eta^s \|w\|_s \quad \text{for all } w \in H_0^s(\Omega).$$

By Corollary 1.4.4.5 of [17], $H_0^s(\Omega) = H^s(\Omega)$ for $s < 1/2$. Let $v \in H^{1+s}(\Omega)$. Applying (6.4) to v and the first derivatives of v implies (6.2). This completes the proof of the lemma. \square

We next prove (4.7). Note that by Lemma 3.1, we need only prove (4.7) for $A(\cdot, \cdot)$ corresponding to the Laplacian. Let $\alpha \in (0, 1/2)$ be such that (4.3) holds. Based on the standard finite element duality argument (cf., [1, 14]), (4.7) will follow if we show that for all $w \in H^{1+\alpha}(\Omega) \cap H_0^1(\Omega)$,

$$(6.5) \quad \|(I - P_{k-1})w\|_1 \leq c\lambda_k^{-\alpha/2} \|w\|_{1+\alpha}.$$

Fix w in $H^{1+\alpha}(\Omega) \cap H_0^1(\Omega)$. Let χ denote the function in M_{k-1} which interpolates w on the nodes of Ω_{k-1}^0 and $\bar{\chi}$ be the function in \bar{M}_{k-1} which interpolates w at the nodes of Ω_{k-1} . From the definition of P_{k-1} , we immediately have that

$$(6.6) \quad \begin{aligned} A((I - P_{k-1})w, (I - P_{k-1})w) &\leq A(w - \chi, w - \chi) \\ &\leq c(\|w - \bar{\chi}\|_{1, \Omega_{k-1}^0}^2 + \|\bar{\chi} - \chi\|_{1, \Omega_{k-1}^0}^2 + \|w\|_{1, \Omega \setminus \Omega_{k-1}^0}^2). \end{aligned}$$

Applying the Bramble-Hilbert Lemma and well-known techniques, we conclude that

$$(6.7) \quad \|w - \bar{\chi}\|_{1, \Omega_{k-1}}^2 \leq ch_k^{2\alpha} \|w\|_{1+\alpha}^2.$$

Note that $\bar{\chi} - \chi$ is a mesh function (in \bar{M}_{k-1}) which vanishes on all nodes except those on $\partial\Omega_{k-1}^0$. Consequently,

$$(6.8) \quad \|\bar{\chi} - \chi\|_{1, \Omega_{k-1}^0}^2 \leq c \sum_{x_i} \bar{\chi}(x_i)^2,$$

where the sum is taken over the nodes x_i of \bar{M}_{k-1} on $\partial\Omega_{k-1}^0$. Let $\tilde{\Omega}_{k-1} = \Omega_{k-1} \setminus \Omega_{k-1}^0$. Then (6.7) and (6.8) imply that

$$(6.9) \quad \begin{aligned} \|\bar{\chi} - \chi\|_{1, \Omega_{k-1}^0}^2 &\leq c \|\bar{\chi}\|_{1, \tilde{\Omega}_{k-1}}^2 \leq c(\|w - \bar{\chi}\|_{1, \tilde{\Omega}_{k-1}}^2 + \|w\|_{1, \tilde{\Omega}_{k-1}}^2) \\ &\leq c(h_k^{2\alpha} \|w\|_{1+\alpha}^2 + \|w\|_{1, \Omega \setminus \Omega_{k-1}^0}^2). \end{aligned}$$

Applying Lemma 6.1 gives

$$(6.10) \quad \|w\|_{1, \Omega \setminus \Omega_{k-1}^0}^2 \leq ch_k^{2\alpha} \|w\|_{1+\alpha}^2.$$

Combining (6.6), (6.7), (6.9) and (6.10) proves (6.5), i.e., (4.7) holds for this application.

We now prove (4.6). By the triangle inequality, the first inequality will follow if we can show that for all $v \in M_J$,

$$(6.11) \quad \|(I - Q_k)v\|^2 \leq c\lambda_k^{-1} A(v, v) \quad \text{for } k = 1, \dots, J-1.$$

Let \bar{Q}_k denote the $L^2(\Omega_k)$ orthogonal projector onto the space \bar{M}_k . By using the interpolant, it is easy to prove that

$$\|(I - \bar{Q}_k)v\|_{\Omega_k}^2 \leq ch_k^{1+\alpha} \|v\|_{1+\alpha}^2$$

for all $v \in H_0^1(\Omega) \cap H^{1+\alpha}(\Omega)$. Interpolating (using the real method [19]) between this inequality and the trivial inequality

$$\|(I - \bar{Q}_k)v\|_{\Omega_k}^2 \leq \|v\|^2$$

gives

$$\|(I - \bar{Q}_k)v\|_{\Omega_k}^2 \leq c\lambda_k^{-1} \|v\|_1^2 \quad \text{for all } v \in H_0^1(\Omega).$$

Fix $v \in H_0^1(\Omega)$ and let θ_k denote the function in M_k which interpolates $\bar{Q}_k v$ at the nodes in Ω_k^0 . Then, Lemma 6.1 and the triangle inequality give

$$\begin{aligned} \|(I - Q_k)v\|^2 &\leq \|v - \theta_k\|^2 \\ &\leq c(\|(I - \bar{Q}_k)v\|_{\Omega_k^0}^2 + \|\bar{Q}_k v - \theta_k\|_{\Omega_k^0}^2 + \|v\|_{\Omega \setminus \Omega_k^0}^2) \\ &\leq c(\lambda_k^{-1} \|v\|_1^2 + \|\bar{Q}_k v - \theta_k\|_{\Omega_k^0}^2 + \lambda_k^{-1} \|v\|_1^2). \end{aligned}$$

The difference $\bar{Q}_k v - \theta_k$ is a mesh function which vanishes on the nodes of Ω_k^0 and hence

$$\begin{aligned} \|\bar{Q}_k v - \theta_k\|_{\Omega_k^0}^2 &\leq ch_k^2 \sum_{x_i} (\bar{Q}_k v(x_i))^2 \leq c\|\bar{Q}_k v\|_{\Omega_k \setminus \Omega_k^0}^2 \\ &\leq c(\|\bar{Q}_k v - v\|_{\Omega_k \setminus \Omega_k^0}^2 + \|v\|_{\Omega_k \setminus \Omega_k^0}^2) \leq c\lambda_k^{-1} \|v\|_1^2. \end{aligned}$$

The sum on x_i above is over the nodes on $\partial\Omega_k^0$. Combining the above inequalities proves (6.11).

We finally prove the second inequality of (4.6). Fix $v \in H_0^1(\Omega)$ and let $\bar{\theta}_k$ denote the $L^2(\Omega_k)$ orthogonal projection of v onto the space of discontinuous piecewise linear functions with respect to the triangulation $\{\tau_k^i\}$. Note that

$$(6.12) \quad \begin{aligned} \|v - \bar{\theta}_k\|_{\Omega_k}^2 &\leq c\lambda_k^{-1} \|v\|_{1,\Omega_k}^2, \\ \sum_{\tau_k^i \subset \Omega_k} \|\bar{\theta}_k\|_{1,\tau_k^i}^2 &\leq c\|v\|_{1,\Omega_k}^2. \end{aligned}$$

Consequently, by (6.11) and (6.12),

$$\sum_{\tau_k^i \subset \Omega_k} \|\bar{\theta}_k - Q_k v\|_{1,\tau_k^i}^2 \leq c\lambda_k \|\bar{\theta}_k - Q_k v\|_{\Omega_k}^2 \leq cA(v, v).$$

The second inequality of (4.6) follows by the triangle inequality and the second inequality of (6.12).

Since (4.6) and (4.7) are valid, the argument proving Lemma 4.1 implies that (3.1) holds for this application. Applying the theory of §3 gives the following theorems.

Theorem 6.1. *Let B_J^q be defined by (2.4) with $A(\cdot, \cdot)$, (\cdot, \cdot) , and $\{M_k\}$ as described in this section. Set $\tilde{M}_k = M_k$ for $k = 2, \dots, J$. Assume that R_k*

(which is symmetric) satisfies (3.7). Then the condition number $K(B_J^a A_J)$ is bounded by a constant which is independent of J .

Theorem 6.2. Let B_J^m be defined by Algorithm 2.1 with $A(\cdot, \cdot)$, (\cdot, \cdot) , and $\{M_k\}$ as described in this section. Set $\tilde{M}_k = M_k$ for $k = 2, \dots, J$. Assume that R_k satisfies (C.1)–(C.3). Then

$$(6.13) \quad 0 \leq A((I - B_J^m A_J)v, v) \leq (1 - 1/C_M)A(v, v) \quad \text{for all } v \in M_J.$$

The constant C_M in (6.13) is independent of J .

Remark 6.1. Clearly, the techniques of §§5 and 6 could be combined to yield similar results for problems with curved boundaries and nonuniform mesh refinements.

BIBLIOGRAPHY

1. A. K. Aziz and I. Babuška, *Survey lectures on the mathematical foundations of the finite element method*, Part I, The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations (A. K. Aziz, ed.), Academic Press, New York, 1972, pp. 1–362.
2. R. E. Bank and T. Dupont, *An optimal order process for solving finite element equations*, Math. Comp. **36** (1981), 35–51.
3. D. Braess and W. Hackbusch, *A new convergence proof for the multigrid method including the V-cycle*, SIAM J. Numer. Anal. **20** (1983), 967–975.
4. J. H. Bramble, Z. Leyk, and J. E. Pasciak, *The analysis of multigrid algorithms for pseudo-differential operators of order minus one*, preprint.
5. ———, *Iterative schemes for non-symmetric and indefinite elliptic boundary value problems*, BNL Rep. 45870.
6. J. H. Bramble and J. E. Pasciak, *The analysis of smoothers for multigrid algorithms*, Math. Comp. **58** (1992), 467–488.
7. ———, *New convergence estimates for multigrid algorithms*, Math. Comp. **49** (1987), 311–329.
8. J. H. Bramble, J. E. Pasciak, J. Wang, and J. Xu, *Convergence estimates for product iterative methods with applications to domain decomposition*, Math. Comp. **57** (1991), 1–21.
9. ———, *Convergence estimates for multigrid algorithms without regularity assumptions*, Math. Comp. **57** (1991), 23–45.
10. J. H. Bramble, J. E. Pasciak and J. Xu, *A multilevel preconditioner for domain decomposition boundary systems*, Proceedings of the 10th Internat. Conf. on Comput. Methods. in Appl. Sci. and Engr., Nova Sciences, New York, 1992.
11. ———, *The analysis of multigrid algorithms with nonnested spaces or noninherited quadratic forms*, Math. Comp. **56** (1991), 1–34.
12. ———, *Parallel multilevel preconditioners*, Math. Comp. **55** (1990), 1–22.
13. A. Brandt, *Multi-level adaptive solutions to boundary-value problems*, Math. Comp. **31** (1977), 333–390.
14. P. G. Ciarlet, *The finite element method for elliptic problems*, North-Holland, New York, 1978.
15. M. Dauge, *Elliptic boundary value problems on corner domains*, Lecture Notes in Math., vol. 1341, Springer-Verlag, Berlin and New York, 1988.
16. N. H. Decker, S. V. Parter, and J. Mandel, *On the role of regularity in multigrid methods*, Multigrid Methods, Proceedings of the Third Copper Mountain Conference (S. McCormick, ed.), Marcel Dekker, New York, 1988, pp. 143–156.
17. P. Grisvard, *Elliptic problems in non smooth domains*, Pitman, Boston, 1985.
18. W. Hackbusch, *Multi-grid methods and applications*, Springer-Verlag, New York, 1985.

19. J. L. Lions and E. Magenes, *Problèmes aux limites non homogènes et applications*, vol. 1, Dunod, Paris, 1968.
20. J.-F. Maitre and F. Musy, *Algebraic formalisation of the multigrid method in the symmetric and positive definite case—a convergence estimation for the V-cycle*, Multigrid Methods for Integral and Differential Equations (D. J. Paddon and H. Holstien, eds), Clarendon Press, Oxford, 1985, pp. 213–223.
21. J. Mandel, S. McCormick, and R. Bank, *Variational multigrid theory*, Multigrid Methods, (S. McCormick, ed.), SIAM, Philadelphia, PA, 1987, pp. 131–178.
22. P. Oswald, *On discrete norm estimates related to multilevel preconditioners in the finite element method*, preprint.
23. X. Zhang, *Multi-level additive Schwarz methods*, Courant Inst. Math. Sci., Dept. Comp. Sci. Rep. (August, 1991).

DEPARTMENT OF MATHEMATICS, CORNELL UNIVERSITY, ITHACA, NEW YORK 14853
E-mail address: bramble@mssun7.msi.cornell.edu

DEPARTMENT OF APPLIED SCIENCE, BROOKHAVEN NATIONAL LABORATORY, UPTON, NEW YORK 11973
E-mail address: pasciak@bnl.gov

3.8 The analysis of multigrid algorithms for pseudodifferential operators of order minus one

The analysis of multigrid algorithms for pseudodifferential operators of order minus one[15]

3.9 Uniform convergence of multigrid V-cycle iterations for indefinite and nonsymmetric problems

Uniform convergence of multigrid V-cycle iterations for indefinite and nonsymmetric problems[14]

UNIFORM CONVERGENCE OF MULTIGRID V-CYCLE ITERATIONS FOR INDEFINITE AND NONSYMMETRIC PROBLEMS

JAMES H. BRAMBLE
DO Y. KWAK
AND
JOSEPH E. PASCIAK

To appear: SIAM Journal of Numerical Analysis
Dedicated to Professor Seymour Parter on the occasion
of the sixty fifth anniversary of his birthday.

1992

ABSTRACT. In this paper, we present an analysis of a multigrid method for nonsymmetric and/or indefinite elliptic problems. In this multigrid method various types of smoothers may be used. One type of smoother which we consider is defined in terms of an associated symmetric problem and includes point and line, Jacobi and Gauss-Seidel iterations. We also study smoothers based entirely on the original operator. One is based on the normal form, that is, the product of the operator and its transpose. Other smoothers studied include point and line, Jacobi and Gauss-Seidel (with certain orderings). We show that the uniform estimates of [6] for symmetric positive definite problems carry over to these algorithms. More precisely, the multigrid iteration for the nonsymmetric and/or indefinite problem is shown to converge at a uniform rate provided that the coarsest grid in the multilevel iteration is sufficiently fine (but not depending on the number of multigrid levels).

1. INTRODUCTION.

The purpose of this paper is to study certain multigrid methods for second order elliptic boundary value problems including problems which may be nonsymmetric and/or indefinite. Multigrid methods are among the most efficient methods available for solving the discrete equations associated with approximate solutions of elliptic partial differential equations. Since their introduction by Fedorenko [15],

1991 *Mathematics Subject Classification.* Primary 65N30; Secondary 65F10.

This manuscript has been authored under contract number DE-AC02-76CH00016 with the U.S. Department of Energy. Accordingly, the U.S. Government retains a non-exclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. This work was also supported in part under the National Science Foundation Grant No. DMS-9007185 and by the U.S. Army Research Office through the Mathematical Sciences Institute, Cornell University. The second author was also partially supported by the Korea Science and Engineering Foundation.

there has been intensive research toward the mathematical understanding of such methods. The reader is referred to [19], [17] and [3] and the bibliographies therein. Most of these works concern symmetric, positive definite elliptic problems although a few consider nonsymmetric and/or indefinite problems. In particular, [1], [18], [10] and [24] deal with such multigrid algorithms and are most closely related to the subject of this paper. All of these papers share the requirement that the coarse grid be sufficiently fine. We shall briefly describe their contents.

The paper of Bank [1] derives uniform convergence estimates for the W-cycle multigrid iteration with both a standard Jacobi smoother and a smoother which uses the operator times its adjoint. In each case, sufficiently many smoothings are required and a sufficiently fine coarse grid depending on the number of smoothings is needed. Some regularity for the elliptic partial differential equation was also required.

Mandel studied the V-cycle iteration and showed that it was effective with only one smoothing and a sufficiently fine coarse grid. His result requires that the underlying partial differential equation satisfy the “full elliptic regularity” hypothesis and generalizes the results of Braess and Hackbusch [2] for the symmetric positive definite problem.

Bramble, Pasciak and Xu [10] studied the symmetric smoother introduced by Bank and showed that the W-cycle and variable V-cycle worked without making the undesirable requirement of “sufficiently many smoothings”. Somewhat more than minimal regularity was needed.

In [24], Wang showed that, for the standard V-cycle with one smoothing, the “reduction factor” for the iteration error was bounded by $1 - C/J + C_1 h_1$ where J is the number of levels, h_1 is the size of the coarsest grid and C and C_1 are constants. This estimate deteriorates with the number of levels and will be less than one only if the coarse grid is subsequently finer as the number of levels increase. Minimal elliptic regularity was assumed.

In this paper uniform iterative convergence estimates for V-cycle multigrid methods applied to nonsymmetric and/or indefinite problems are proved under rather weak assumptions (e.g., the domain need not be convex). Uniform estimates were shown to hold in [6] and [8] for the V-cycle with one smoothing step in the symmetric positive definite case under such hypotheses. We show that these results carry over to the nonsymmetric and/or indefinite case for a variety of smoothers. The coarse grid must be fine enough but need not depend on the number of levels J . Such a condition seems unavoidable since, in many cases, it is needed even for the approximate problem to make sense.

In recent years, some other techniques have been proposed to handle the non-symmetric indefinite case. One approach in [14], [4] and [7] is to precondition with a symmetric operator and then solve certain normal equations by the conjugate gradient method. One possible advantage of such a method is that some nonsymmetric problems which are not “compact perturbations” of symmetric ones may be treated. Of course, the usual normal equations may be formed and then preconditioned (cf. [7] and [20]); this approach seems to be rather restrictive in that good preconditioners may be difficult to construct. Other recent approaches have included Schwarz type methods [12] and two-level methods in which a “coarse space” is introduced to reduce the problem to one with a positive definite symmetric part (cf. [4], [13]).

and [25]).

The remainder of the paper is organized as follows: In Section 2, we describe a model problem and introduce the multigrid method. In Section 3, smoothers based on the symmetric problem (and used in our nonsymmetric and/or indefinite applications) are defined and the relevant properties which they satisfy are stated. Section 4 develops smoothers based on the original problem. The main results of the paper, which provide iterative convergence rates for the multigrid algorithms with the smoothers of Sections 3 and 4, are given in Section 5.

2. THE PROBLEM AND MULTIGRID ALGORITHM.

We set up the model nonsymmetric problem and the simplest multigrid algorithm in this section. We consider, for simplicity, the Dirichlet problem in two spatial dimensions approximated by piecewise linear finite elements on a quasi-uniform mesh. The multigrid convergence results hold for many extensions and generalizations as discussed at the end of Section 5.

We consider as our model problem the following second order elliptic equation with homogeneous boundary conditions.

$$(2.1) \quad \begin{aligned} -\sum_{i,j=1}^2 \frac{\partial}{\partial x_j} (a_{ij} \frac{\partial u}{\partial x_i}) + \sum_{i=1}^2 b_i \frac{\partial u}{\partial x_i} + au &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega, \end{aligned}$$

where Ω is a polygonal domain (possibly nonconvex) in R^2 and $\{a_{ij}(x)\}$ is bounded symmetric, and uniformly positive definite for $x \in \Omega$. We assume that a_{ij} is in the Sobolev space $W_p^\gamma(\Omega)$ for $p > 2/\gamma$ (see, [16] for the definition of $W_p^\gamma(\Omega)$). Further, we assume that b_i is continuously differentiable on $\bar{\Omega}$ and that $|a|$ is bounded. Finally, we assume that the solution of (2.1) exists.

Let $H^1(\Omega)$ denote the Sobolev space of order one on Ω (cf., [16]) and let $H_0^1(\Omega)$ denote those functions in $H^1(\Omega)$ whose trace vanish on $\partial\Omega$. For $v, w \in H_0^1(\Omega)$, define

$$(2.2) \quad A(v, w) = \sum_{i,j=1}^2 \int_\Omega a_{ij} \frac{\partial v}{\partial x_i} \frac{\partial w}{\partial x_j} dx + \sum_{i=1}^2 \int_\Omega b_i \frac{\partial v}{\partial x_i} w dx + \int_\Omega avw dx.$$

The solution u of (2.1) satisfies

$$(2.3) \quad A(u, v) = (f, v) \quad \text{for all } v \in H_0^1(\Omega),$$

where (\cdot, \cdot) denotes the inner product in $L^2(\Omega)$.

For the analysis, we introduce a symmetric positive definite form $\hat{A}(\cdot, \cdot)$ which has same second order part as $A(\cdot, \cdot)$. We define $\hat{A}(\cdot, \cdot)$ by

$$\hat{A}(u, v) = \sum_{i,j=1}^2 \int_\Omega a_{ij} \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_j} dx + \int_\Omega uv dx.$$

The difference is denoted by

$$D(u, v) = A(u, v) - \hat{A}(u, v).$$

The form $D(\cdot, \cdot)$ satisfies the inequalities

$$(2.4) \quad |D(u, v)| \leq C \|u\|_1 \|v\| \quad \text{and} \quad |D(u, v)| \leq C \|u\| \|v\|_1.$$

Here $\|\cdot\|_1$ and $\|\cdot\|$ denote the norms in $H^1(\Omega)$ and $L^2(\Omega)$ respectively. The second inequality above follows from integration by parts. Here and throughout the paper, c or C , with or without subscript, will denote a generic positive constant. These constants can take on different values in different occurrences but will always be independent of the mesh size and the number of levels in multigrid algorithms.

By the assumptions on the coefficients appearing in the definition of $\hat{A}(\cdot, \cdot)$, it follows that the norm $\hat{A}(v, v)^{1/2}$ for $v \in H^1(\Omega)$ is equivalent to the norm on $H^1(\Omega)$. Thus, we take

$$\|v\|_1 = \hat{A}(v, v)^{1/2}.$$

We develop a sequence of nested triangulations of Ω in the usual way. We assume that a coarse triangulation $\{\tau_1^i\}$ of Ω is given. Successively finer triangulations $\{\tau_m^i\}$ for $m > 1$ are defined by subdividing each triangle (in a coarser triangulation) into four by connecting the midpoints of the edges. The mesh size of $\{\tau_1^i\}$ will be denoted to be d_1 and can be taken to be the diameter of the largest triangle. By similarity, the mesh size of $\{\tau_m^i\}$ is $2^{1-m}d_1$.

For theoretical and practical purposes, the coarsest grid in the multilevel algorithms must be sufficiently fine. In practice, however, the coarse grid is still considerably coarser than the solution grid. Let L and J be greater than or equal to one and set M_k , for $k = 1, \dots, J$, to be the functions which are piecewise linear with respect to the triangulation $\{\tau_{k+L}^i\}$, continuous on Ω and vanish on $\partial\Omega$. Since the triangulations are nested, it follows that

$$M_1 \subset M_2 \subset \dots \subset M_J.$$

The space M_k has a mesh size of $h_k = 2^{1-L-k}d_1 = 2^{1-k}h_1$.

Fix k in $\{1, 2, \dots\}$. Let us temporarily assume that for every $u \in M_k$,

$$(2.5) \quad A(u, v) = 0 \quad \text{for all } v \in M_k \quad \text{implies} \quad u = 0.$$

This assumption immediately implies the existence and uniqueness of solutions to problems of the form: Given a linear functional $F(\cdot)$ defined on M_k , find $u \in M_k$ satisfying

$$A(u, \phi) = F(\phi) \quad \text{for all } \phi \in M_k.$$

In particular, the projection operator $P_k : H^1(\Omega) \mapsto M_k$ satisfying

$$A(P_k u, v) = A(u, v) \quad \text{for all } v \in M_k,$$

is well defined.

Clearly, if (2.2) has a positive definite symmetric part then (2.5) holds. More generally, if solutions of (2.1) satisfy regularity estimates of the form

$$(2.6) \quad \|u\|_{1+\alpha} \leq C\|f\|_{-1+\alpha},$$

then, it is well known (cf., [22]) that there exists a constant h_0 such that for $h_k \leq h_0$, (2.5) holds and furthermore

$$(2.7) \quad \|(I - P_k)u\| \leq ch_k^\alpha \|(I - P_k)u\|_1.$$

and finally,

$$(2.8) \quad \|P_k u\|_1 \leq C \|u\|_1.$$

Even if regularity estimates of the form of (2.6) are not known to hold, then (2.5) is known from a recent result by Schatz and Wang [23].

Lemma 2.1 [23]. *There exists an h_0 such that (2.5) holds for $h_k \leq h_0$. Moreover, given $\epsilon > 0$, there exists an $h_0(\epsilon) > 0$ such that for all $h_k \in (0, h_0]$, (2.8) holds and*

$$(2.9) \quad \|(I - P_k)u\| \leq \epsilon \|(I - P_k)u\|_1.$$

Remark 2.1. The above ϵ will appear in our subsequent analysis. We note that ϵ can be taken arbitrarily small. However, L will be taken large enough so that (2.5), (2.8) and (2.9) hold. Thus, the coarse grid size (i.e., L) for any estimate in which ϵ appears will depend on ϵ .

In our analysis, we shall use the orthogonal projectors $\hat{P}_k : H_0^1(\Omega) \mapsto M_k$ and $Q_k : L^2(\Omega) \mapsto M_k$ which, respectively, denote the elliptic projection corresponding to $\hat{A}(\cdot, \cdot)$ and the $L^2(\Omega)$ projection. These are defined by

$$\hat{A}(\hat{P}_k u, v) = \hat{A}(u, v) \quad \text{for all } v \in M_k,$$

and

$$(Q_k u, v) = (u, v) \quad \text{for all } v \in M_k.$$

The multigrid algorithms will be defined in terms of an additional inner product $(\cdot, \cdot)_k$ on $M_k \times M_k$. Examples of this inner product in our applications will be given in the next section. Additional operators are defined in terms of this inner product as follows: For each k , define $A_k : M_k \rightarrow M_k$ and $\hat{A}_k : M_k \rightarrow M_k$ by

$$(A_k u, v)_k = A(u, v) \quad \text{for all } v \in M_k,$$

and

$$(\hat{A}_k u, v)_k = \hat{A}(u, v) \quad \text{for all } v \in M_k.$$

Finally, the restriction operator $P_{k-1}^0 : M_k \mapsto M_{k-1}$ is defined by

$$(P_{k-1}^0 u, v)_{k-1} = (u, v)_k \quad \text{for all } v \in M_{k-1}.$$

We seek the solution of

$$(2.10) \quad A(u, v) = (f, v), \quad \text{for all } v \in M_J.$$

This can be rewritten in the above notation as

$$(2.11) \quad A_J u = Q_J f.$$

We describe the simplest V-cycle multigrid algorithm for iteratively computing the solution u of (2.3). Given an initial iterate $u_0 \in M_J$, we define a sequence approximating u by

$$(2.12) \quad u_{i+1} = \text{Mg}_J(u_i, Q_J f).$$

Here $\text{Mg}_J(\cdot, \cdot)$ is a map of $M_J \times M_J$ into M_J and is defined as follows.

Definition MG. Set $\text{Mg}_1(v, w) = A_1^{-1}w$. Let $k > 1$ and v, w be in M_k . Assuming that $\text{Mg}_{k-1}(\cdot, \cdot)$ has been defined, we define $\text{Mg}_k(v, w)$ by:

- (1) $x_k = v + R_k(w - A_k v)$.
- (2) $\text{Mg}_k(v, w) = x_k + q$, where q is defined by

$$q = \text{Mg}_{k-1}(0, P_{k-1}^0(w - A_k x_k)).$$

Here $R_k : M_k \mapsto M_k$ is a linear smoothing operator. Note that in this V-cycle, we smooth only as we proceed to coarser grids.

In Section 3, we define R_k in terms of smoothing operators defined for the form $\hat{A}(\cdot, \cdot)$. Specifically, the smoothing procedure for the symmetric problem will be denoted $\hat{R}_k : M_k \mapsto M_k$ and we set $R_k = \hat{R}_k$. In Section 4, we consider smoothers which are directly defined in terms of the original operator A_k .

A straightforward mathematical induction argument shows that $\text{Mg}_J(\cdot, \cdot)$ is a linear map from $M_J \times M_J$ into M_J . Moreover, the scheme is consistent in the sense that $v = \text{Mg}_J(v, A_J v)$ for all $v \in M_J$. It easily follows that the linear operator $E = \text{Mg}_J(\cdot, 0)$ is the error reduction operator for (2.12), that is

$$u - u_{i+1} = E(u - u_i).$$

Let $T_k = R_k A_k P_k$ for $k > 1$ and set $T_1 = P_1$. Using the facts that $P_{k-1}^0 A_k = A_{k-1} P_{k-1}$ and $P_{k-1} P_k = P_{k-1}$ and Definition MG, a straightforward manipulation gives that for $k > 1$ and any $u \in M_J$,

$$u - \text{Mg}_k(0, A_k P_k u) = (I - T_k)u - \text{Mg}_{k-1}(0, A_{k-1} P_{k-1}(I - T_k)u).$$

Let $E_k u = u - \text{Mg}_k(0, A_k P_k u)$. In terms of E_k , the above identity is the same as

$$E_k = E_{k-1}(I - T_k).$$

Moreover, by consistency, $E = E_J$ and hence

$$(2.13) \quad E = (I - T_1)(I - T_2) \cdots (I - T_J).$$

The product representation of the error operator given above will be a fundamental ingredient in the convergence analysis presented in Section 4. Similar representations in the case of multigrid algorithms for symmetric problems were given in [9].

The above algorithm is a special case of more general multigrid algorithms in that we only use pre-smoothing. Alternatively, we could define an algorithm with just post-smoothing or both pre- and post-smoothing. The analysis of these algorithms is similar to that above and will not be presented.

Often algorithms with more than one smoothing are considered [3], [17], [19]. This is not advised in the above algorithm since the smoothing iteration is generally unstable.

3. SMOOTHERS BASED ON THE SYMMETRIC PROBLEM.

In this section, we consider smoothers which are based on the symmetric problem. The symmetric smoother will be denoted by \hat{R}_k . We state a number of abstract conditions concerning these smoothing operators. We then give three examples of smoothing procedures which satisfy these assumptions. In Section 5, we provide convergence estimates for multigrid algorithms with $R_k = \hat{R}_k$ in Definition MG.

The first two conditions are standard assumptions used in earlier multigrid analyses. For $k > 1$, let $\hat{K}_k = I - \hat{R}_k \hat{A}_k$ (defined on M_k) and $\hat{T}_k = \hat{R}_k \hat{A}_k \hat{P}_k$ (defined on M_J). We assume that:

- (1) There is a constant C_R such that

$$(C.1) \quad \frac{(u, u)_k}{\lambda_k} \leq C_R (\bar{R}_k u, u)_k, \quad \text{for all } u \in M_k,$$

where $\bar{R}_k = (I - \hat{K}_k^* \hat{K}_k) \hat{A}_k^{-1}$ and λ_k is the largest eigenvalue of \hat{A}_k . Here and in the remainder of this paper, $*$ denotes the adjoint with respect to the inner product $\hat{A}(\cdot, \cdot)$.

- (2) There is a constant $\theta < 2$ not depending on k satisfying

$$(C.2) \quad \hat{A}(\hat{T}_k v, \hat{T}_k v) \leq \theta \hat{A}(\hat{T}_k v, v) \quad \text{for all } v \in M_k.$$

Provided that (C.2) holds, (C.1) is equivalent to

$$(3.1) \quad \frac{(u, u)_k}{\lambda_k} \leq C (\hat{R}_k u, u)_k, \quad \text{for all } u \in M_k.$$

When \hat{R}_k is symmetric with respect to $(\cdot, \cdot)_k$, (C.2) states that the norm of \hat{T}_k is less than or equal to θ . Even in the case of non-symmetric \hat{R}_k , (C.2) implies stability of $(I - \hat{T}_k)$. In fact, for any $w \in M_J$, (C.2) implies that

$$(3.2) \quad \begin{aligned} \hat{A}((I - \hat{T}_k)w, (I - \hat{T}_k)w) &= \hat{A}(w, w) - 2\hat{A}(\hat{T}_k w, w) + \hat{A}(\hat{T}_k w, \hat{T}_k w) \\ &\leq \hat{A}(w, w) - (2 - \theta)\hat{A}(\hat{T}_k w, w) \leq \hat{A}(w, w). \end{aligned}$$

The final condition is that for $k > 1$, there exists a constant C satisfying

$$(C.3) \quad (\hat{T}_k u, \hat{T}_k u)_k \leq C \lambda_k^{-1} \hat{A}(\hat{T}_k u, u) \quad \text{for all } u \in M_k.$$

A simple change of variable shows that (C.3) is the same as

$$(\hat{R}_k v, \hat{R}_k v)_k \leq C \lambda_k^{-1} (\hat{R}_k v, v)_k \quad \text{for all } v \in M_k.$$

In the case when \hat{R}_k is symmetric, this is equivalent to

$$(3.3) \quad (\hat{R}_k v, v)_k \leq C \lambda_k^{-1} (v, v)_k \quad \text{for all } v \in M_k$$

and is the opposite inequality of (3.1). Note that both (C.2) and (C.3) hold on M_J .

Remark 3.1. If Conditions (C.1)–(C.3) hold for a smoother R_k then they hold for its adjoint R_k^t with respect to the inner product $(\cdot, \cdot)_k$. This means that (C.1) holds for $\bar{R}_k = (I - \hat{K}_k \hat{K}_k^*) \hat{A}_k^{-1}$ and that (C.2) and (C.3) hold with \hat{T}_k^* replacing \hat{T}_k . In the case of (C.2) and (C.3), the corresponding inequalities hold with the same constants as those appearing in the original inequalities.

Example 1. The first example of a smoother is the operator

$$\hat{R}_k = \bar{\lambda}_k^{-1} I$$

where I denotes the identity operator on M_k and $\lambda_k \leq \bar{\lambda}_k \leq C \lambda_k$. In this case, (3.1) holds with $C = \bar{\lambda}_k / \lambda_k$, (C.2) holds with $\theta = 1$ and (3.3) holds with $C = \lambda_k / \bar{\lambda}_k$. To avoid the inversion of L^2 Gram matrices in the multigrid algorithm, we use the inner product

$$(3.4) \quad (u, v)_k = h_k^2 \sum_i u(x_i) v(x_i).$$

Here the sum is taken over all nodes x_i of the subspace M_k . Note that $(\cdot, \cdot)_k$ is uniformly (independent of k) equivalent to (\cdot, \cdot) on M_k .

The remaining smoothers correspond to Jacobi and Gauss-Seidel, point and line iteration methods. We shall present these smoothers in terms of subspace decompositions. Specifically, we write

$$(3.5) \quad M_k = \sum_{i=1}^l M_k^i$$

where M_k^i is the one dimensional subspace spanned by the nodal basis function ϕ_k^i or the subspace spanned by the nodal basis functions along a line. The number of such spaces $l = l(k)$ will often depend on k . These spaces satisfy the following inequality.

$$(3.6) \quad \|v\| \leq C h_k \|v\|_1 \quad \text{for all } v \in M_k^i.$$

Example 2. For the second example, we consider the additive smoother defined by

$$(3.7) \quad \hat{R}_k = \gamma \sum_{i=1}^l \hat{A}_{k,i}^{-1} Q_{k,i}.$$

Here $\hat{A}_{k,i} : M_k^i \rightarrow M_k^i$ is defined by

$$(\hat{A}_{k,i}v, \chi)_k = \hat{A}(v, \chi) \quad \text{for all } \chi \in M_k^i$$

and $Q_{k,i} : M_k \rightarrow M_k^i$ is the projection onto M_k^i with respect to the inner product $(\cdot, \cdot)_k$. The constant γ is a scaling factor which is chosen to ensure that (C.2) is satisfied (see, e.g., [11], [5]). Note that \hat{R}_k is symmetric with respect to the inner product $(\cdot, \cdot)_k$. In addition, (3.1) and (3.3) are shown to hold in [11] with point Jacobi. When the subspaces M_k^i are defined in terms of lines, (3.1) was proved in [5]. The estimate (3.3) easily follows in the line case using the support properties of the basis functions and (3.6). For this example, we take $(\cdot, \cdot)_k = (\cdot, \cdot)$ for all k .

Example 3. We next consider the multiplicative smoother. Given $f \in M_k$, we define \hat{R}_k by

- (1) Set $v_0 = 0 \in M_k$.
- (2) Define v_i , for $i = 1, \dots, l$, by

$$v_i = v_{i-1} + \hat{A}_{k,i}^{-1} Q_{k,i}(f - \hat{A}_k v_{i-1}).$$

- (3) Set $\hat{R}_k f = v_l$.

Conditions (C.1) and (C.2) are known for this operator (see, e.g., [5]). The next lemma shows that (C.3) holds for this choice of \hat{R}_k . For this case, we also take $(\cdot, \cdot)_k = (\cdot, \cdot)$ for all k .

Lemma 3.1. *(C.3) holds when \hat{R}_k is defined to be the multiplicative smoother of Example 3.*

Proof. The proof uses the techniques for analyzing smoothers presented in [5]. Fix $k > 1$ and let

$$(3.8) \quad \hat{\mathcal{E}}_i = (I - \hat{P}_k^i)(I - \hat{P}_k^{i-1}) \cdots (I - \hat{P}_k^1)$$

where \hat{P}_k^i denotes the $\hat{A}(\cdot, \cdot)$ projection onto M_k^i . Note that $(I - \hat{T}_k) = \hat{\mathcal{E}}_l$ and $\hat{\mathcal{E}}_{i-1} = \hat{\mathcal{E}}_i + \hat{P}_k^i \hat{\mathcal{E}}_{i-1}$. Hence

$$\hat{T}_k = I - \hat{\mathcal{E}}_l = \sum_{i=1}^l \hat{P}_k^i \hat{\mathcal{E}}_{i-1}$$

and for every $u \in M_k$, (cf., [5])

$$\begin{aligned} \hat{A}((2I - \hat{T}_k)u, \hat{T}_k u) &= \hat{A}(u, u) - A(\hat{\mathcal{E}}_l u, \hat{\mathcal{E}}_l u) \\ &= \sum_{i=1}^l \hat{A}(\hat{P}_k^i \hat{\mathcal{E}}_{i-1} u, \hat{\mathcal{E}}_{i-1} u). \end{aligned}$$

Since $h_k^2 \leq c\lambda_k^{-1}$, the proof of the lemma will be complete if we can show that

$$(3.9) \quad (\hat{T}_k u, \hat{T}_k u) \leq ch_k^2 \sum_{i=1}^l \hat{A}(\hat{P}_k^i \hat{\mathcal{E}}_{i-1} u, \hat{\mathcal{E}}_{i-1} u).$$

Expanding the left hand side of (3.9) gives

$$(3.10) \quad (\hat{T}_k u, \hat{T}_k u) = \sum_{i=1}^l \sum_{j=1}^l (\hat{P}_k^i \hat{\mathcal{E}}_{i-1} u, \hat{P}_k^j \hat{\mathcal{E}}_{j-1} u).$$

Because of the support properties of $\{\phi_k^i\}$, the subspaces $\{M_k^i\}$ satisfy a limited interaction property in that for every i , the number of subspaces j for which $(v^i, v^j) \neq 0$, with $v^i \in M_k^i$ and $v^j \in M_k^j$ is bounded by a fixed constant n_0 not depending on k or l . Lemma 3.1 of [5] implies that the double sum of (3.10) can be bounded by n_0 times its diagonal, i.e.

$$(3.11) \quad (\hat{T}_k u, \hat{T}_k u) \leq n_0 \sum_{i=1}^l (\hat{P}_k^i \hat{\mathcal{E}}_{i-1} u, \hat{P}_k^i \hat{\mathcal{E}}_{i-1} u).$$

Applying (3.6) gives

$$(3.12) \quad (\hat{P}_k^i \hat{\mathcal{E}}_{i-1} u, \hat{P}_k^i \hat{\mathcal{E}}_{i-1} u) \leq C h_k^2 \hat{A}(\hat{P}_k^i \hat{\mathcal{E}}_{i-1} u, \hat{\mathcal{E}}_{i-1} u).$$

Combining (3.11) and (3.12) proves (3.9). This completes the proof of the lemma.

Remark 3.2. The same analysis could be used for successive overrelaxation type iteration. In that case,

$$\hat{\mathcal{E}}_l = (I - \beta \hat{P}_k^l)(I - \beta \hat{P}_k^{l-1}) \cdots (I - \beta \hat{P}_k^1)$$

where $\beta \in (0, 2)$ is the relaxation parameter.

4. SMOOTHERS BASED ON A_k .

In this section, we consider smoothing operators R_k which are defined directly in terms of the nonsymmetric and/or indefinite operator A_k . The first smoother is one that was originally analyzed in [1] and subsequently studied in [10].

Example 4. For our first example of a smoother based on A_k , we consider R_k defined by

$$R_k = \bar{\lambda}_k^{-2} A_k^t.$$

Here, A_k^t is the adjoint of A_k with respect to the inner product $(\cdot, \cdot)_k$ and $\bar{\lambda}_k$ is as in Example 1. A possible motivation for such a choice is that, on M_k , the iteration

$$v^i = v^{i-1} + \bar{\lambda}_k^{-2} A_k^t (f - A_k v^{i-1})$$

is stable in the norm $(\cdot, \cdot)_k^{1/2}$ provided that $\bar{\lambda}_k^2$ is greater than or equal to half the largest eigenvalue of $A_k^t A_k$.

Example 5. This example is closely related to the second example of the previous section. As in that example, we define the line or point subspaces $\{M_k^i\}$ for $i = 1, \dots, l$. Note that the form $A(\cdot, \cdot)$ satisfies a Gårding inequality

$$c_1 \hat{A}(u, u) - c \|u\|^2 \leq A(u, u) \quad \text{for all } u \in H_0^1(\Omega).$$

Consequently, by (3.6),

$$(c_1 - Ch_k^2) \hat{A}(u, u) \leq A(u, u) \quad \text{for all } u \in M_k^i.$$

We will assume that h_2 is sufficiently small so that

$$(4.1) \quad Ch_2^2 \leq c_1/2.$$

This means that $A(\cdot, \cdot)$ restricted to M_k^i has a positive definite symmetric part. Hence, the projector $P_k^i : M_k \mapsto M_k^i$ satisfying

$$A(P_k^i v, w) = A(v, w) \quad \text{for all } w \in M_k^i$$

is well defined and satisfies

$$(4.2) \quad \|P_k^i u\|_1 \leq C \|u\|_{1, \Omega_k^i}.$$

The second norm is taken only over the subdomain Ω_k^i which is the set of points of Ω where the functions in M_k^i are nonzero. In addition, the operator $A_{k,i} : M_k^i \mapsto M_k^i$ defined by

$$(A_{k,i}v, w)_k = A(v, w) \quad \text{for all } v, w \in M_k^i,$$

is invertible. We set R_k by

$$R_k = \gamma \sum_{i=1}^l A_{k,i}^{-1} Q_{k,i}.$$

We choose γ as in Example 2 so that the symmetric smoother defined by (3.7) satisfies (C.2).

Example 6. Our final example is that of Gauss-Seidel directly applied to the non-symmetric/indefinite equations. We assume that the subspaces $\{M_k^i\}$ satisfy the conditions of the previous example and in addition, that l is bounded independently of k . This is possible by doing what is commonly referred to as a coloring scheme. Starting with subspaces satisfying (3.6), we group together those whose supports overlap at most on sets of measure zero and thus reducing the size of l . For a regular mesh on the square, we could group together all of the subspaces associated with the odd lines (similarly, those associated with the even lines). Since we are grouping subspaces with essentially disjoint supports, (3.6) holds on the larger subspaces. The block Gauss-Seidel algorithm (based on A_k) is given as follows:

- (1) Set $v_0 = 0 \in M_k$.
- (2) Define v_i , for $i = 1, \dots, l$, by

$$v_i = v_{i-1} + A_{k,i}^{-1} Q_{k,i} (f - A_k v_{i-1}).$$

- (3) Set $R_k f = v_l$.

5. ANALYSIS OF THE MULTIGRID ITERATION (2.12).

We provide an analysis of the multigrid iteration (2.12) in this section. This analysis is based on the product representation of the error operator (2.13). All of the analysis of this section is based on perturbation from the uniform convergence estimates for multigrid applied to symmetric problems.

We start by stating a result from [6] estimating the rate of convergence for the multigrid algorithm applied to the symmetric problem. Specifically, we replace A_k by \hat{A}_k and R_k by \hat{R}_k in Definition MG. Set $\hat{T}_1 = \hat{P}_1$. From the earlier discussion, the error operator associated with this iteration applied to finding solution of the symmetric problem

$$\hat{A}_J u = Q_J f$$

is given by $\hat{E} = \hat{E}_J$ where

$$(5.1) \quad \hat{E}_k = (I - \hat{T}_1)(I - \hat{T}_2) \cdots (I - \hat{T}_k).$$

We then have the following theorem.

Theorem 5.1 [6]. *For $k > 1$, let \hat{R}_k satisfy (C.1) and (C.2). Under the assumptions on the domain Ω and the coefficients of (2.1) given in Section 2, there exists a positive constant $\hat{\delta} < 1$ not depending on J such that*

$$\hat{A}(\hat{E}_J u, \hat{E}_J u) \leq \hat{\delta}^2 A(u, u) \quad \text{for all } u \in M_J.$$

To analyze the multigrid algorithms using the smoothers of Section 3, we use the perturbation operator

$$Z_k = T_k - \hat{T}_k.$$

We note that for any $u, v \in M_J$, for $k > 1$,

$$(5.2) \quad \hat{A}(Z_k u, v) = D(u, \hat{T}_k^* v).$$

Indeed, by definition,

$$\begin{aligned} \hat{A}(T_k u, v) &= (T_k u, \hat{A}_k \hat{P}_k v)_k = (A_k P_k u, \hat{R}_k^t \hat{A}_k \hat{P}_k v)_k \\ &= (A_k P_k u, \hat{T}_k^* v)_k = A(P_k u, \hat{T}_k^* v) \\ &= A(u, \hat{T}_k^* v) = \hat{A}(u, \hat{T}_k^* v) + D(u, \hat{T}_k^* v). \end{aligned}$$

The equality (5.2) immediately follows.

To handle the case of $k = 1$, we have

$$(5.3) \quad \hat{A}(Z_1 u, v) = D((I - P_1)u, \hat{P}_1 v).$$

In fact, by definition,

$$\begin{aligned} \hat{A}(P_1 u, v) &= \hat{A}(P_1 u, \hat{P}_1 v) \\ &= A(u, \hat{P}_1 v) - D(P_1 u, \hat{P}_1 v) \\ &= \hat{A}(\hat{P}_1 u, v) + D((I - P_1)u, \hat{P}_1 v). \end{aligned}$$

The following theorem provides an estimate for the multigrid algorithm when the smoothers of Section 3 are used.

Theorem 5.2. Let $R_k = \hat{R}_k$ and assume that (C.1)–(C.3) hold. Given $\epsilon > 0$, there exists an $h_0 > 0$ such that for $h_1 \leq h_0$,

$$\hat{A}(Eu, Eu) \leq \delta^2 \hat{A}(u, u) \quad \text{for all } u \in M_J,$$

for $\delta = \hat{\delta} + c(h_1 + \epsilon)$. Here $\hat{\delta}$ is less than one (independently of J) and is given by Theorem 5.1.

Proof. For an arbitrary operator $\mathcal{O} : M_J \mapsto M_J$, let $\|\mathcal{O}\|_{\hat{A}}$ denote its operator norm, i.e.,

$$\|\mathcal{O}\|_{\hat{A}} = \sup_{u, v \in M_J} \frac{\hat{A}(\mathcal{O}u, v)}{\hat{A}(u, u)^{1/2} \hat{A}(v, v)^{1/2}}.$$

Applying (2.4), (2.9) and (2.8) to (5.3) gives

$$|\hat{A}(Z_1 u, v)| \leq C\epsilon \|(I - P_1)u\|_1 \|v\|_1 \leq C\epsilon \|u\|_1 \|v\|_1.$$

This means that the operator norm of Z_1 is bounded by $C\epsilon$. Since the operator norm of $(I - \hat{P}_1)$ is less than or equal to one, the triangle inequality implies that the operator norm of $(I - P_1) = (I - \hat{P}_1 - Z_1)$ is bounded by $1 + C\epsilon$.

For $k > 1$, applying (2.4), (C.3), Remark 3.1, and (3.2) to (5.2) gives

$$\begin{aligned} |\hat{A}(Z_k u, v)| &\leq ch_k \|u\|_1 \hat{A}(\hat{T}_k v, v)^{1/2} \\ &\leq ch_k \|u\|_1 \|v\|_1, \end{aligned}$$

i.e., the operator norm of Z_k is bounded by ch_k . Since, by (3.2), the operator norm of $(I - \hat{T}_k)$ is less than or equal to one, the triangle inequality implies that the operator norm of $(I - T_k) = (I - \hat{T}_k - Z_k)$ is less than or equal to $1 + ch_k$. Hence, it follows that

$$\|E_k\|_{\hat{A}} \leq (1 + C\epsilon) \prod_{i=2}^k (1 + ch_i) \leq C.$$

It is immediate from the definitions that

$$(5.4) \quad E_k - \hat{E}_k = (E_{k-1} - \hat{E}_{k-1})(I - \hat{T}_k) - E_{k-1}Z_k.$$

By (3.2) and the above estimates, for $k > 1$,

$$\begin{aligned} (5.5) \quad \|E_k - \hat{E}_k\|_{\hat{A}} &\leq \|E_{k-1} - \hat{E}_{k-1}\|_{\hat{A}} \|I - \hat{T}_k\|_{\hat{A}} + \|E_{k-1}\|_{\hat{A}} \|Z_k\|_{\hat{A}} \\ &\leq \|E_{k-1} - \hat{E}_{k-1}\|_{\hat{A}} + Ch_k. \end{aligned}$$

Repetitively applying (5.5) and using

$$\|E_1 - \hat{E}_1\|_{\hat{A}} = \|Z_1\|_{\hat{A}} \leq C\epsilon$$

gives that

$$\|E_J - \hat{E}_J\|_{\hat{A}} \leq C\epsilon + C \sum_{k=2}^J h_k \leq c(h_1 + \epsilon).$$

The theorem follows from the triangle inequality and Theorem 5.1.

Remark 5.1. Note that ϵ can be made arbitrarily small by taking h_1 small enough. Consequently, Theorem 5.2 shows that the multigrid iteration converges with a rate which is independent of J provided that the coarse grid is fine enough. The coarse grid mesh size can also be taken to be independent of J .

We next consider the case of Example 4. For this example, we consider first the multigrid algorithm for the symmetric problem which uses

$$(5.6) \quad \hat{R}_k = \bar{\lambda}_k^{-2} \hat{A}_k$$

as a smoother. From the discussion in Section 2, the iteration (2.12) with \hat{R}_k (given by (5.6)) and \hat{A}_k replacing, respectively, R_k and A_k in Definition MG, gives rise to the error operator given by (5.1) where, as above, for $k > 1$, $\hat{T}_k = \hat{R}_k \hat{A}_k \hat{P}_k$. The smoother (5.6) does not satisfy (C.1) and so the first step in the analysis of the nonsymmetric and/or indefinite example is to provide a uniform estimate for \hat{E}_J given by (5.1). Such an estimate is provided in the following theorem. Its proof is given in the appendix.

Theorem 5.3. *Let \hat{E}_J be given by (5.1) where $\hat{T}_k = \hat{R}_k \hat{A}_k \hat{P}_k$ and \hat{R}_k is defined by (5.6). Then,*

$$\hat{A}(\hat{E}_J u, \hat{E}_J u) \leq \hat{\delta}^2 A(u, u) \quad \text{for all } u \in M_J.$$

Here $\hat{\delta}$ is less than one and independent of J .

We can now prove the convergence estimate for multigrid applied to (2.1) using the smoother of Example 4.

Theorem 5.4. *Let R_k be defined by Example 4. Given $\epsilon > 0$, there exists an $h_0 > 0$ such that for $h_1 \leq h_0$,*

$$\hat{A}(E u, E u) \leq \delta^2 \hat{A}(u, u) \quad \text{for all } u \in M_J,$$

for $\delta = \hat{\delta} + c(h_1 + \epsilon)$. Here $\hat{\delta}$ is less than one (independently of J) and is given by Theorem 5.3.

Proof. For $k > 1$, we consider the perturbation operator

$$Z_k = T_k - \hat{T}_k = \bar{\lambda}_k^{-2} (A_k^t A_k P_k - \hat{A}_k^2 \hat{P}_k).$$

Clearly,

$$(5.7) \quad Z_k = \bar{\lambda}_k^{-2} [A_k^t (A_k P_k - \hat{A}_k \hat{P}_k) + (A_k^t - \hat{A}_k) \hat{A}_k \hat{P}_k].$$

As in (5.2),

$$\bar{\lambda}_k^{-1} \hat{A}((A_k P_k - \hat{A}_k \hat{P}_k) u, v) = \bar{\lambda}_k^{-1} D(u, \hat{A}_k \hat{P}_k v)$$

from which it follows using (2.4) that

$$\|\bar{\lambda}_k^{-1} (A_k P_k - \hat{A}_k \hat{P}_k)\|_{\hat{A}} \leq c h_k.$$

A similar argument shows that

$$\|\bar{\lambda}_k^{-1}(A_k^t - \hat{A}_k)\hat{P}_k\|_{\hat{A}} \leq ch_k.$$

It is not difficult to show that

$$\|A_k^t\|_{\hat{A}} \leq C\bar{\lambda}_k.$$

Combining the above estimates with (5.7) gives

$$\begin{aligned} \|Z_k\|_{\hat{A}} &\leq \|\bar{\lambda}_k^{-1}A_k^t\|_{\hat{A}}\|\bar{\lambda}_k^{-1}(A_kP_k - \hat{A}_k\hat{P}_k)\|_{\hat{A}} \\ &+ \|\bar{\lambda}_k^{-1}(A_k^t - \hat{A}_k)\hat{P}_k\|_{\hat{A}}\|\bar{\lambda}_k^{-1}\hat{A}_k\hat{P}_k\|_{\hat{A}} \leq ch_k. \end{aligned}$$

The remainder of the proof is exactly the same as that of Theorem 5.2. This completes the proof of the theorem.

We next consider the case of Example 5. We use perturbation from the multigrid algorithm for \hat{A} which uses the smoother \hat{R}_k defined by Example 2. Theorem 5.1 provides a uniform estimate for the operator norm of \hat{E}_J .

Theorem 5.5. *Let R_k be defined by Example 5. Given $\epsilon > 0$, there exists an $h_0 > 0$ such that for $h_1 \leq h_0$,*

$$\hat{A}(Eu, Eu) \leq \delta^2 \hat{A}(u, u) \quad \text{for all } u \in M_J,$$

for $\delta = \hat{\delta} + c(h_1 + \epsilon)$. Here $\hat{\delta}$ is less than one (independently of J) and is given by Theorem 5.1 applied to \hat{R}_k defined in Example 2.

Proof. For this case, the perturbation operator Z_k is given by

$$Z_k = \gamma \sum_{i=1}^l (P_k^i - \hat{P}_k^i).$$

As in (5.3),

$$\hat{A}((P_k^i - \hat{P}_k^i)u, v) = D((I - P_k^i)u, \hat{P}_k^i v).$$

Applying (2.4), (3.6) and (4.2) gives

$$(5.8) \quad \hat{A}((P_k^i - \hat{P}_k^i)u, v) \leq ch_k \|u\|_{1,\Omega_k^i} \|v\|_{1,\Omega_k^i}$$

and hence

$$\hat{A}(Z_k u, v) \leq ch_k \sum_{i=1}^l \|u\|_{1,\Omega_k^i} \|v\|_{1,\Omega_k^i}.$$

Using the limited overlap properties of the domains Ω_k^i gives

$$\|Z_k\|_{\hat{A}} \leq ch_k.$$

The remainder of the proof of the theorem is exactly the same as that given in the proof of Theorem 5.2.

We finally consider the case of Example 6. We use perturbation from the multigrid algorithm for \hat{A} which uses the smoother \hat{R}_k defined by Example 3. Theorem 5.1 provides a uniform estimate for the operator norm of \hat{E}_J .

Theorem 5.6. Let R_k be defined by Example 6. Given $\epsilon > 0$, there exists an $h_0 > 0$ such that for $h_1 \leq h_0$,

$$\hat{A}(Eu, Eu) \leq \delta^2 \hat{A}(u, u) \quad \text{for all } u \in M_J,$$

for $\delta = \hat{\delta} + c(h_1 + \epsilon)$. Here $\hat{\delta}$ is less than one (independently of J) and is given by Theorem 5.1 applied with \hat{R}_k defined as in Example 3.

Proof. The perturbation operator for this example is

$$Z_k = T_k - \hat{T}_k = \hat{\mathcal{E}}_l - \mathcal{E}_l$$

where $\hat{\mathcal{E}}_l$ is given by (3.8) and

$$\mathcal{E}_i = (I - P_k^i)(I - P_k^{i-1}) \cdots (I - P_k^1).$$

As in (5.4),

$$\hat{\mathcal{E}}_i - \mathcal{E}_i = (I - \hat{P}_k^i)(\hat{\mathcal{E}}_{i-1} - \mathcal{E}_{i-1}) - (\hat{P}_k^i - P_k^i)\mathcal{E}_{i-1}.$$

Clearly, the operator norm of $(I - \hat{P}_k^i)$ is bounded by one. Moreover, by (5.8) the operator norm of $(\hat{P}_k^i - P_k^i)$ is bounded by ch_k . It follows that

$$\|I - P_k^i\|_{\hat{A}} \leq 1 + ch_k$$

and, for $j = 1, \dots, l$,

$$\|\mathcal{E}_j\|_{\hat{A}} \leq C.$$

Thus,

$$(5.9) \quad \begin{aligned} \|\hat{\mathcal{E}}_i - \mathcal{E}_i\|_{\hat{A}} &\leq \|\hat{\mathcal{E}}_{i-1} - \mathcal{E}_{i-1}\|_{\hat{A}} + \|\hat{P}_k^i - P_k^i\|_{\hat{A}} \|\mathcal{E}_{i-1}\|_{\hat{A}} \\ &\leq \|\hat{\mathcal{E}}_{i-1} - \mathcal{E}_{i-1}\|_{\hat{A}} + Ch_k. \end{aligned}$$

Repetitively applying (5.9) and using the fact that for this example, l is bounded independently of k gives that for $k > 1$,

$$\|Z_k\|_{\hat{A}} \leq Ch_k.$$

The remainder of the proof of this theorem is the same as that of Theorem 5.2.

Remark 5.1. Many extensions and generalizations of the techniques given above are possible. These techniques lead to uniform estimates for multigrid iteration methods for solving nonsymmetric and/or indefinite problems for the following applications.

- (1) Approximations using higher order nodal finite element spaces.
- (2) Three dimensional problems.
- (3) Problems with discontinuous coefficients as discussed in [6].
- (4) More general boundary conditions.
- (5) Problems with local mesh refinement as described in [11].
- (6) Finite element approximation of problems on domains with nonpolygonal boundaries as discussed in [6].

In addition, the perturbation analysis given above can be combined with results for additive multilevel algorithms, for example, Theorem 3.1 of [6]. This leads to new estimates for additive multilevel preconditioning iterations applied to indefinite and nonsymmetric problems. Provided that the coarse grid is sufficiently fine, the operator

$$P = \sum_{k=1}^J T_k$$

has a uniformly (independent of J) positive definite symmetric part with respect to the inner product $\hat{A}(\cdot, \cdot)$ and has a uniformly bounded operator norm. These results extend to all of the applications discussed in Remark 5.1.

6. APPENDIX

We provide a proof of Theorem 5.3 in this appendix. We will apply the analysis given in the proof of Theorem 3.2 of [6]. Note that we cannot directly apply Theorem 3.2 of [6] since the smoother $\hat{R}_k = \bar{\lambda}_k^{-2} \hat{A}_k$ does not satisfy (C.1). We note, however, that Theorem 5.3 will follow from the proof of Theorem 3.2 of [6] if we show that (C.2) holds as well as (3.5) and (3.6) of [6] with \tilde{T}_k replaced by \hat{T}_k defined above. Clearly, (C.2) holds with $\theta = 1$. The remaining two inequalities corresponding to (3.5) and (3.6) of [6] are

$$(6.1) \quad \hat{A}(\hat{T}_k v, v) \leq (\tilde{C}\eta^{k-l})^2 \hat{A}(v, v) \quad \text{for all } v \in M_l, \quad l < k$$

and

$$(6.2) \quad \hat{A}(v, v) \leq C \sum_{k=1}^J \hat{A}(\hat{T}_k v, v) \quad \text{for all } v \in M_J.$$

Here η is less than one and independent of k and l .

From the definition of $\bar{\lambda}_k$, we obviously have

$$\hat{A}(\hat{T}_k v, v) \leq \bar{\lambda}_k^{-1} \hat{A}(\hat{A}_k v, v) = \hat{A}(\tilde{T}_k v, v).$$

As in [6], we have set $\tilde{T}_k = \bar{\lambda}_k^{-1} \hat{A}_k$. Inequality (6.1) follows from Lemma 4.2 of [6].

Inequality (6.2) can be rewritten,

$$(6.3) \quad \hat{A}(u, u) \leq C \left(\hat{A}(\hat{P}_1 u, u) + \sum_{k=2}^J \bar{\lambda}_k^{-2} \left\| \hat{A}_k \hat{P}_k u \right\|_1^2 \right).$$

To prove this we proceed as follows. Let $u \in M_J$ and $Q_0 = 0$. Then

$$(6.4) \quad \begin{aligned} \hat{A}(u, u) &= \sum_{k=1}^J \hat{A}(u, (Q_k - Q_{k-1})u) \\ &\leq \left(\hat{A}(\hat{P}_1 u, u) + \sum_{k=2}^J \bar{\lambda}_k^{-2} \left\| \hat{A}_k \hat{P}_k u \right\|_1^2 \right)^{1/2} \left(\hat{A}(Q_1 u, Q_1 u) \right. \\ &\quad \left. + \sum_{k=2}^J \bar{\lambda}_k^2 (\hat{A}_k^{-1} (Q_k - Q_{k-1})u, (Q_k - Q_{k-1})u)_k \right)^{1/2}. \end{aligned}$$

Now, for $k > 1$,

$$\begin{aligned}
& (\hat{A}_k^{-1}(Q_k - Q_{k-1})u, (Q_k - Q_{k-1})u)_k \\
&= \sup_{\phi \in M_k} \frac{(\hat{A}_k^{-1/2}(Q_k - Q_{k-1})u, \phi)_k^2}{(\phi, \phi)_k} \\
&= \sup_{\psi \in M_k} \frac{((Q_k - Q_{k-1})u, (Q_k - Q_{k-1})\psi)_k^2}{\|\psi\|_1^2}.
\end{aligned}$$

By well known approximation properties,

$$((Q_k - Q_{k-1})\psi, (Q_k - Q_{k-1})\psi)_k^{1/2} \leq C \|(Q_k - Q_{k-1})\psi\| \leq Ch_k \|\psi\|_1.$$

Combining the above estimates gives

$$\begin{aligned}
& \hat{A}(Q_1 u, Q_1 u) + \sum_{k=2}^J \bar{\lambda}_k^2 (\hat{A}_k^{-1}(Q_k - Q_{k-1})u, (Q_k - Q_{k-1})u)_k \\
(6.5) \quad & \leq C \left(\hat{A}(Q_1 u, Q_1 u) + \sum_{k=2}^J \bar{\lambda}_k \|(Q_k - Q_{k-1})u\|^2 \right) \\
& \leq C \hat{A}(u, u).
\end{aligned}$$

The last inequality of (6.5) is (4.5) of [6] and also can be found in [21]. Combining (6.4) and (6.5) proves (6.3) and hence completes the proof of the theorem.

REFERENCES

1. R. Bank, *A comparison of two multilevel iterative methods for nonsymmetric and indefinite elliptic finite element equations*, SIAM J. Numer. Anal. **18** (1981), 724–743.
2. Braess, D. and Hackbusch, W., *A new convergence proof for the multigrid method including the V-cycle*, SIAM J. Numer. Anal. **20** (1983), 967–975.
3. J.H. Bramble, *Multigrid Methods*, Cornell Mathematics Department Lecture Notes, 1992.
4. J.H. Bramble, Z. Leyk, and J.E. Pasciak, *Iterative schemes for non-symmetric and indefinite elliptic boundary value problems*, Math. Comp. (to appear).
5. J.H. Bramble and J.E. Pasciak, *The analysis of smoothers for multigrid algorithms*, Math. Comp. **58** (1992), 467–488.
6. J.H. Bramble and J.E. Pasciak, *New estimates for multigrid algorithms including the V-cycle*, Math. Comp. (to appear).
7. J.H. Bramble and J.E. Pasciak, *Preconditioned iterative methods for nonselfadjoint or indefinite elliptic boundary value problems*, Unification of finite element methods, (Ed. H. Kardestuncer), Elsevier Science Publ. (North-Holland), New York, 1984, pp. 167 – 184.
8. J.H. Bramble and J.E. Pasciak, *Uniform convergence estimates for multigrid V-cycle algorithms with less than full elliptic regularity* (1992), Brookhaven Nat. Lab. #BNL-47892.
9. J.H. Bramble, J.E. Pasciak, J. Wang, and J. Xu, *Convergence estimates for multigrid algorithms without regularity assumptions*, Math. Comp. **57** (1991), 23–45.
10. J.H. Bramble, J.E. Pasciak and J. Xu, *The analysis of multigrid algorithms for nonsymmetric and indefinite elliptic problems*, Math. Comp. **51** (1988), 389–414.
11. J.H. Bramble, J.E. Pasciak and J. Xu, *Parallel multilevel preconditioners*, Math. Comp. **55** (1990), 1–22.
12. X.-C. Cai and O. Widlund, *Domain decomposition algorithms for indefinite elliptic problems*, SIAM J. Sci. Stat. Comp. (to appear).

13. X.-C. Cai and J. Xu, *A preconditioned GMRES method for nonsymmetric and indefinite problems*, (Preprint).
14. H.C. Elman, *Iterative methods for large, sparse, nonsymmetric systems of linear equations*, Yale Univ. Dept. of Comp. Sci. Rep. 229, (1982).
15. Fedorenko, R.P., *The speed of convergence of one iterative process*, USSR Comput. Math. and Math. Phys. (1961), 1092–1096.
16. P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.
17. Hackbusch, W., *Multi-Grid Methods and Applications*, Springer-Verlag, New York, 1985.
18. Mandel, J., *Multigrid convergence for nonsymmetric, indefinite variational problems and one smoothing step*, Proc. Copper Mtn. Conf. Multigrid Methods, vol. 19, Applied Math. Comput., 1986, pp. 201–216.
19. J. Mandel, S. McCormick and R. Bank, *Variational multigrid theory*, Multigrid Methods, Ed. S. McCormick, SIAM, Philadelphia, Penn., 1987, pp. 131–178.
20. T.A. Manteuffel and S.V. Parter, *Preconditioning and boundary conditions*, SIAM J. Numer. Anal. **27** (1990), 656–694.
21. P. Oswald, *On discrete norm estimates related to multilevel preconditioners in the finite element method*, (Preprint).
22. A.H. Schatz, *An observation concerning Ritz-Galerkin methods with indefinite bilinear forms*, Math. Comp. **28** (1974), 959–962.
23. A.H. Schatz and J. Wang.
24. J. Wang, *Convergence analysis of multigrid algorithms for non-selfadjoint and indefinite elliptic problems* (1991), Proceedings of the 5'th Copper Mountain Conference on Multigrid Methods.
25. J. Xu, *A new class of iterative methods for nonsymmetric boundary value problems*, (preprint).

DEPARTMENT OF MATHEMATICS
 WHITE HALL, CORNELL UNIVERSITY
 ITHACA, NY 14853-7901
 E-MAIL: BRAMBLE@MATH.MSI.CORNELL.EDU

DEPARTMENT OF MATHEMATICS
 KOREA ADVANCED INSTITUTE OF SCIENCE AND TECHNOLOGY
 TAEJON, KOREA 305-701
 E-MAIL: DYKWAK%MATH1.KAIST.AC.KR

DEPARTMENT OF APPLIED SCIENCE
 BROOKHAVEN NATIONAL LABORATORY
 UPTON, NY 11973
 E-MAIL: PASCIAK@BNL.GOV

3.10 The analysis of multigrid methods

The analysis of multigrid methods[30]

|||||| HEAD

3.11 Non-overlapping domain decomposition methods

1. An iterative method for elliptic problems on regions partitioned into substructures
2. The construction of preconditioners for elliptic problems by substructuring. I
3. The construction of preconditioners for elliptic problems by substructuring. II
4. The construction of preconditioners for elliptic problems by substructuring. III
5. The construction of preconditioners for elliptic problems by substructuring. IV

3.12 Overlapping domain decomposition methods

1. Convergence estimates for product iterative methods with applications to domain decomposition

=====

4

Domain Decomposition Methods

4.1 The construction of preconditioners for elliptic problems by substructuring. I

The construction of preconditioners for elliptic problems by substructuring. I [19]

The Construction of Preconditioners for Elliptic Problems by Substructuring. I

By J. H. Bramble,* J. E. Pasciak* and A. H. Schatz*

Dedicated to Professor Joachim Nitsche on the occasion
of the sixtieth anniversary of his birthday.

Abstract. We consider the problem of solving the algebraic system of equations which arise from the discretization of symmetric elliptic boundary value problems via finite element methods. A new class of preconditioners for these discrete systems is developed based on substructuring (also known as domain decomposition). The resulting preconditioned algorithms are well suited to emerging parallel computing architectures. The proposed methods are applicable to problems on general domains involving differential operators with rather general coefficients. A basic theory for the analysis of the condition number of the preconditioned system (which determines the iterative convergence rate of the algorithm) is given. Techniques for applying the theory and algorithms to problems with irregular geometry are discussed and the results of extensive numerical experiments are reported.

1. Introduction. The aim of this series of papers is to propose and analyze methods for efficiently solving the equations resulting from finite element discretizations of second-order elliptic boundary value problems on general domains in R^2 and R^3 . In particular, we shall be concerned with constructing easily invertible and “effective” preconditioners for the resulting system of discrete equations which can be used in a preconditioned iterative algorithm to achieve a rapid solution method. The methods to be presented are well suited to parallel computing architectures.

In this paper we shall restrict ourselves to boundary value problems in R^2 . Let Ω be a bounded domain in R^2 with a piecewise smooth boundary $\partial\Omega$. As a model problem for a second-order uniformly elliptic equation we shall consider the Dirichlet problem

$$(1.1) \quad Lu = f \quad \text{in } \Omega, \quad u = 0 \quad \text{on } \partial\Omega,$$

where

$$Lv = - \sum_{i,j=1}^2 \frac{\partial}{\partial x_i} \left(a_{ij} \frac{\partial v}{\partial x_j} \right),$$

Received April 9, 1985; revised November 27, 1985.

1980 *Mathematics Subject Classification*. Primary 65N30; Secondary 65F10.

*This manuscript has been authored under contract number DE-AC02-76CH00016 with the U.S. Department of Energy. Accordingly, the U.S. Government retains a non-exclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. This work was also supported in part under the National Science Foundation Grant No. DMS84-05352 and under the Air Force Office of Scientific Research, Contract No. ISSA86-0026..

©1986 American Mathematical Society
0025-5718/86 \$1.00 + \$.25 per page

with a_{ij} uniformly positive definite, bounded and piecewise smooth on Ω . The generalized Dirichlet form is given by

$$(1.2) \quad A(v, \phi) = \sum_{i,j=1}^2 \int_{\Omega} a_{ij} \frac{\partial v}{\partial x_i} \frac{\partial \phi}{\partial x_j} dx,$$

which is defined for all v and ϕ in the Sobolev space $H^1(\Omega)$ (the space of distributions with square-integrable first derivatives). The $L^2(\Omega)$ -inner product is denoted

$$(v, \phi) = \int_{\Omega} v\phi dx.$$

The subspace $H_0^1(\Omega)$ is the completion of the smooth functions with support in Ω with respect to the norm in $H^1(\Omega)$. The weak formulation of the problem defined by (1.1) is: Find $u \in H_0^1(\Omega)$ such that

$$(1.3) \quad A(u, \phi) = (f, \phi)$$

for all $\phi \in H_0^1(\Omega)$. This leads immediately to the standard Galerkin approximation. Let $S_h^0(\Omega)$ be a finite-dimensional subspace of $H_0^1(\Omega)$. The Galerkin approximation is defined as the solution of the following problem: Find $U \in S_h^0(\Omega)$ such that

$$(1.4) \quad A(U, \Phi) = (f, \Phi)$$

for all $\Phi \in S_h^0(\Omega)$. Once a basis $\{\chi_i\}_{i=1}^N$ for $S_h^0(\Omega)$ is chosen, (1.4) leads to a system of linear algebraic equations. Write $U = \sum_{i=1}^N \alpha_i \chi_i$. Then (1.4) becomes

$$(1.5) \quad \sum_{i=1}^N \alpha_i A(\chi_i, \chi_j) = (f, \chi_j),$$

$j = 1, \dots, N$. We shall choose $S_h^0(\Omega)$ so that firstly, the function U will be a good approximation to u and secondly, efficient algorithms for the solution of the underlying linear system (1.5) can be developed. In particular, we will consider subspaces $S_h^0(\Omega)$ of $H_0^1(\Omega)$ which are defined so that certain related subproblems can be efficiently solved. We will see that this leads to algorithms for the solution of the global linear system which is well suited to parallel processing.

The strategy of choosing $S_h^0(\Omega)$ so that efficient algorithms exist for the solution of the resulting linear system is not unusual. For example, for the Laplace operator on a rectangular region, a subspace $S_h^0(\Omega)$ of piecewise linear functions on a uniform triangulation leads to the usual 5-point approximation to the Laplacian. The resulting equations may be solved "fast" using, for example, fast Fourier transform techniques. In this case, other choices of $S_h^0(\Omega)$ may lead to good approximate solutions, but these solutions may be more difficult to obtain computationally. Another example of a special choice of $S_h^0(\Omega)$ which leads often to a fast algorithm is one which may be thought of as connected with a nested set of grids. For such spaces, a "multigrid" algorithm may be applied.

The underlying method which we will consider is a preconditioned iterative method. The choice of a particular iterative method within a certain class is not essential, but for the purpose of this exposition we may think of the well-known conjugate gradient method [12], [15] which is often used in practice. Roughly, the application of a preconditioned method may be described as follows. Let A be the $N \times N$ matrix with entries $A(\chi_i, \chi_j)$, α the column vector whose components are

as in (1.5), and F the vector with components (f, χ_j) . Then (1.5) may be written as

$$(1.6) \quad A\alpha = F.$$

Generally, the matrix A is not well-conditioned so that a direct application of the conjugate gradient method to the symmetric positive-definite system (1.6) will not be a very efficient algorithm. The preconditioned conjugate gradient method (PCG) consists of choosing a positive-definite symmetric matrix B and writing the equivalent system

$$(1.7) \quad B^{-1}A\alpha = B^{-1}F.$$

In the present context the matrix B will be associated with another bilinear form $B(\cdot, \cdot)$ defined on $S_h^0(\Omega) \times S_h^0(\Omega)$. The system (1.7) is symmetric with respect to the inner product defined by

$$(1.8) \quad [\alpha, \beta] \equiv \sum_{i,j=1}^N B_{ij} \alpha_i \beta_j.$$

Thus, the conjugate gradient method may be applied to (1.7) with respect to (1.8). The importance of making a “good” choice for B is well known. The matrix B should have two properties. First, the solution of the problem

$$(1.9) \quad B\beta = b$$

should be easy to obtain. This is tantamount to applying the operator B^{-1} to the vector b . Secondly, B should be spectrally close to A in the sense that the condition number K of $B^{-1}A$ should not be large. Clearly, $K \leq \lambda_1/\lambda_0$, where λ_0 and λ_1 are constants such that

$$\lambda_0[\beta, \beta] \leq [B^{-1}A\beta, \beta] \leq \lambda_1[\beta, \beta] \quad \text{for all } \beta \in R^N.$$

In terms of the form $B(\cdot, \cdot)$, the first property means that the solution W of

$$(1.10) \quad B(W, \Phi) = (g, \Phi), \quad \text{for all } \Phi \in S_h^0(\Omega)$$

for a given function g should be easier to obtain than the solution of (1.4). The spectral condition, in terms of the forms, is

$$(1.11) \quad \lambda_0 B(V, V) \leq A(V, V) \leq \lambda_1 B(V, V) \quad \text{for all } V \in S_h^0(\Omega).$$

These two properties will guarantee, firstly, that the work per iterative step in applying the preconditioned method will be small, and, secondly, that the number of steps to reduce the error to a given size will also be small so that an efficient algorithm will result.

In this paper we will describe and analyze a technique for constructing the bilinear form $B(\cdot, \cdot)$ so that the action of the corresponding matrix problem B^{-1} is easy to compute. As a preliminary step, the domain is subdivided into subdomains. Our preconditioner will be defined so that the computation of its inverse applied to a vector only involves solving in parallel related Galerkin (or matrix) equations on subregions of Ω and some interconnecting equations, which may also be solved in parallel. The preconditioner B will be the first of our domain decomposition preconditioners to be developed in this series of papers and will sometimes be denoted DD1.

In Section 2, the preconditioner B^{-1} will be defined and the essential step in the iterative algorithm of computing the action of B^{-1} will be described in detail.

The main result concerning the condition number K is also stated in this section as Theorem 1. Section 3 is devoted to a complete proof of Theorem 1. In Section 4 we show how $S_h^0(\Omega)$ can be constructed and how various coefficients introduced in the definition of the preconditioner (see Section 2) can be chosen so that the related subproblems can be efficiently solved, even in rather complex domain geometry. Section 5 contains a reexamination of the process of applying the action of B^{-1} in the context of “block Gauss elimination”. Finally, in Section 6 we describe the results of numerical calculations which show that the theoretical estimates are fully realized in practice.

For other works dealing with the numerical solution of boundary value problems via substructuring we refer to [1], [2], [4]–[7], [10]. We emphasize that a novel feature of our approach is that more than two subdomains can meet at an interior point of the original domain. In addition, our results remain valid independently of the number of such points. As a simple example, our approach applies to a checkerboard subdivision of a square.

2. The Construction of $B(\cdot, \cdot)$ and the Preconditioning Algorithm. As mentioned in the introduction, the preconditioner which we will construct involves the solution of smaller related problems on subdomains and subdomain boundaries. For the sake of simplicity of exposition we shall proceed with the discussion only for the special case of polygonal domains and piecewise linear approximations.

More precisely, we shall begin with the following assumptions with regard to Ω .

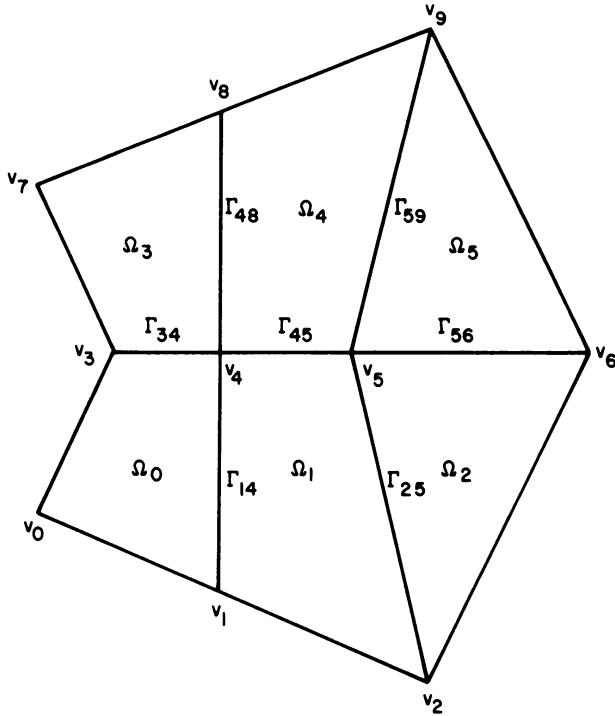
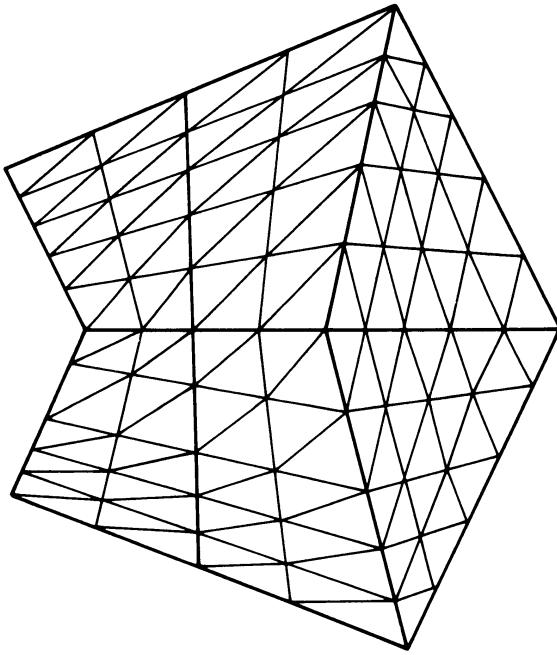
A.1: Ω is a polygonal domain.

A.2: For each h , $0 < h < 1$ a parameter, Ω has been given a quasi-uniform triangulation Ω^h . By this we mean that there exists a positive constant c_1 independent of h such that each triangle $\tau^h \in \Omega^h$ contains a ball of radius $c_1 h$ and is contained in a ball of radius h .

A.3: For each triangulation Ω^h , Ω may be written as the union of disjoint regions Ω_k , which are either quadrilaterals or triangles whose sides coincide with the mesh lines of the original triangulation and which are quasi-uniform of size $d \geq h$ with constants, as above, which are independent of d and h . If Ω_k is a quadrilateral, we require additionally that the lengths of each side be bounded from below by $c_1 d$ and that any interior angle α satisfy $0 < C_0 \leq \alpha \leq C_1 < \pi$. The collection of regions Ω_k will frequently be referred to as the subdomains.

The vertices of the $\{\Omega_k\}$ will be labeled v_j (ordered in some way) and Γ_{ij} will denote the straight line segment with endpoints v_i and v_j . Throughout this paper we shall only consider Γ_{ij} when Γ_{ij} is an edge of some Ω_k . Furthermore, we associate with each Ω_k the triangulation inherited from the original triangulation Ω^h . The examples given in Figures 2.1 and 2.2 should help clarify the situation.

For each h , let $S_h(\Omega)$ be the space of continuous piecewise linear functions defined relative to the triangulation Ω^h and $S_h^0(\Omega)$ be the subspace of $S_h(\Omega)$ consisting of those functions which vanish on $\partial\Omega$. $S_h^0(\Omega_j)$ will denote the subspace of $S_h^0(\Omega)$ of functions whose supports are contained in $\bar{\Omega}_j$ (in particular, they vanish on $\partial\Omega_j$ and outside $\bar{\Omega}_j$). In addition, $S_h(\Omega_j)$ will be the set of functions which are restrictions of those in $S_h^0(\Omega)$ to $\bar{\Omega}_j$. Subspaces on the boundaries of the subdomains will be denoted as follows. $S_h(\partial\Omega_j)$ will denote the restrictions of $S_h(\Omega_j)$ to $\partial\Omega_j$ and $S_h^0(\Gamma_{ij})$, the subspace of $S_h(\partial\Omega_j)$ consisting of functions whose support is contained on the edge Γ_{ij} . In what follows, c and C (with or without subscript) will denote generic positive constants which are independent of h , d and the Ω_k .

FIGURE 2.1. *The domain Ω and subdomains.*FIGURE 2.2. *The domain with mesh.*

We construct our preconditioner B by constructing its corresponding bilinear form $B(\cdot, \cdot)$ defined on $S_h^0(\Omega) \times S_h^0(\Omega)$. We first introduce another form $\tilde{A}(\cdot, \cdot)$ which is defined by first setting

$$\tilde{A}_k(U, V) = \sum_{i,j=1}^2 \int_{\Omega_k} a_{ij}^k \frac{\partial U}{\partial x_i} \frac{\partial V}{\partial x_j} dx,$$

and then defining

$$\tilde{A}(U, V) = \sum_k \tilde{A}_k(U, V).$$

Here for each k , a_{ij}^k is a piecewise smooth (possibly discontinuous) uniformly positive-definite matrix. The reason for the form of \tilde{A} will become clear as we proceed with the development. We note, however, that

$$C_0 \tilde{A}(U, U) \leq A(U, U) \leq C_1 \tilde{A}(U, U)$$

for positive constants C_0 and C_1 . Thus, the problem of finding a preconditioner for A is the same as finding one for \tilde{A} .

We next decompose functions in $S_h^0(\Omega)$ as follows: Write $W = W_P + W_H$ where $W_P \in S_h^0(\Omega_1) \oplus \cdots \oplus S_h^0(\Omega_{n_r})$ and satisfies

$$\tilde{A}_k(W_P, \Phi) = \tilde{A}_k(W, \Phi) \quad \text{for all } \Phi \in S_h^0(\Omega_k)$$

for each k . Notice that W_P is determined on Ω_k by the values of W on Ω_k and that

$$\tilde{A}_k(W_H, \Phi) = 0 \quad \text{for all } \Phi \in S_h^0(\Omega_k).$$

Thus on each Ω_k , W is decomposed into a function W_P which vanishes on $\partial\Omega_k$ and a function $W_H \in S_h(\Omega_k)$ which satisfies the above homogeneous equations and has the same boundary values as W . We shall refer to such a function W_H as “discrete \tilde{A}_k -harmonic.”

Remark 2.1. The matrices with entries a_{ij}^k are in principle arbitrary but, as will be seen in Section 4, they may be chosen in such a way that the subproblems determining W_P and W_H may be easily solved once the values of W_H on the subdomain boundaries are known.

We note that the above decomposition is orthogonal in the \tilde{A} -inner product and hence,

$$\tilde{A}(W, W) = \tilde{A}(W_P, W_P) + \tilde{A}(W_H, W_H).$$

We shall define $B(\cdot, \cdot)$ by replacing the $\tilde{A}(W_H, W_H)$ term above. To do this, we decompose $W_H \in S_h(\Omega_k)$ into $W_H = W_E + W_V$, where $W_V \in S_h(\Omega_k)$ is the discrete \tilde{A}_k -harmonic function whose values on $\partial\Omega_k$ are the linear function along each Γ_{ij} with the same values as W at the vertices. Thus W_E is a discrete \tilde{A}_k -harmonic function in Ω_k for each k which vanishes at all of the vertices.

Before defining the form $B(\cdot, \cdot)$, we note that for any discrete \tilde{A}_k -harmonic function W with zero mean value on Ω_k ,

$$(2.1) \quad \gamma_0 \tilde{A}_k(W, W) \leq |W|_{1/2, \partial\Omega_k}^2 \leq \gamma_1 \tilde{A}_k(W, W),$$

where γ_0 and γ_1 are positive constants and $|\cdot|_{1/2, \partial\Omega_k}$ is the norm on the Sobolev space $H^{1/2}(\partial\Omega_k)$. This will be proved in Section 3 but is noted here to motivate our construction. Now it will also be shown in the next section that if $W = 0$ at the vertices, then the norm $|W|_{1/2, \partial\Omega_k}^2$ may be replaced in (2.1) by $\sum_{\Gamma_{ij}} \alpha_{ij} \langle a^{-1} \tilde{l}_0^{1/2} W, W \rangle_{\Gamma_{ij}}$ with new values of γ_0 and γ_1 such that $\gamma_1/\gamma_0 \leq C(1 + \ln(d/h)^2)$. Here \tilde{l}_0 is the operator defined for each Γ_{ij} on $S_h^0(\Gamma_{ij})$ by

$$(2.2) \quad \langle a^{-1} \tilde{l}_0 W, \Phi \rangle_{\Gamma_{ij}} = \langle a W', \Phi' \rangle_{\Gamma_{ij}} \quad \text{for all } \Phi \in S_h^0(\Gamma_{ij}).$$

The prime denotes differentiation with respect to arc length s along Γ_{ij} . In (2.2) a is, for simplicity, a positive piecewise constant function on Γ_{ij} , $0 < a_0 \leq a \leq a_1$, with a_0 and a_1 independent of Γ_{ij} and h , and

$$\langle \phi, \psi \rangle_{\Gamma_{ij}} = \int_{\Gamma_{ij}} \phi \psi \, ds.$$

Note that \tilde{l}_0 is symmetric and positive definite in the inner product $\langle a^{-1} \cdot, \cdot \rangle_{\Gamma_{ij}}$ and hence its square root is well-defined.

Here again we may, in principle, choose the function a quite arbitrarily, but, as will be seen in Section 4, computational considerations dictate a natural choice. α_{ij} is a positive constant which will also be chosen explicitly later. We however require that $0 < C_0 \leq \alpha_{ij} \leq C_1$ for constants C_0 and C_1 which are independent of h , d , and the Ω_k 's.

Finally, as is shown in the next section, for W_V as defined above,

$$C_0 A_k(W_V, W_V) \leq \sum_{\Gamma_{ij}} \alpha_{ij} (W_V(v_i) - W_V(v_j))^2 \leq C_1 A_k(W_V, W_V)$$

holds for some positive constants C_0 and C_1 .

With the above statements in mind, we now define the form $B(\cdot, \cdot)$ by

$$(2.3) \quad \begin{aligned} B(W, \Phi) = & \tilde{A}(W_P, \Phi_P) + \sum_{\Gamma_{ij}} \alpha_{ij} \langle a^{-1} \tilde{l}_0^{1/2} W_E, \Phi_E \rangle_{\Gamma_{ij}} \\ & + \sum_{\Gamma_{ij}} \alpha_{ij} (W_V(v_i) - W_V(v_j)) (\Phi_V(v_i) - \Phi_V(v_j)). \end{aligned}$$

The following theorem is proved in Section 3:

THEOREM 1. *There are positive constants λ_0 , λ_1 and C such that*

$$\lambda_0 B(W, W) \leq A(W, W) \leq \lambda_1 B(W, W) \quad \text{for all } W \in S_h^0(\Omega),$$

where $\lambda_1/\lambda_0 \leq C(1 + \ln(d/h)^2)$. If all of the vertices of the Ω_k lie on $\partial\Omega$, then $\lambda_1/\lambda_0 \leq C$.

Thus the condition number grows at most like $(1 + \ln(d/h)^2)$ as h tends to zero. This means that the preconditioned iteration will converge rapidly and corresponds to the second of the two desirable properties mentioned earlier.

The first property previously discussed states that problem (1.10) should be much more easily solved than the original (1.4). This means that the solution of the corresponding matrix equation (1.9) is relatively easy to obtain.

We shall demonstrate how (1.10) can be solved efficiently. In fact, we shall see that the defining equations have been chosen to conveniently lend themselves to a “block Gauss elimination” procedure. Here, we shall describe the process used to solve (1.10). The matrix interpretation is given in Section 5.

Given g , the problem of solving (1.10) reduces to finding the functions W_P and W_H . The function W_P restricted to Ω_k satisfies

$$(2.4) \quad \tilde{A}_k(W_P, \Phi) = (g, \Phi) \quad \text{for all } \Phi \in S_h^0(\Omega_k).$$

Thus the function W_P on Ω_k can be obtained by solving the corresponding Dirichlet problem (2.4). Note that the problems on different subdomains are independent of each other so that they may be solved in parallel.

With W_P now known, we are left with the equation

$$(2.5) \quad \begin{aligned} & \sum_{\Gamma_{ij}} \alpha_{ij} \langle a^{-1} \tilde{l}_0^{1/2} W_E, \Phi_E \rangle_{\Gamma_{ij}} + \sum_{\Gamma_{ij}} \alpha_{ij} (W_V(v_i) - W_V(v_j)) (\Phi_V(v_i) - \Phi_V(v_j)) \\ & = (g, \Phi) - \tilde{A}(W_P, \Phi_P) = (g, \Phi) - \tilde{A}(W_P, \Phi), \end{aligned}$$

the last equality holding since $\tilde{A}(W_P, \Phi_H) = 0$. Notice that the value of $(g, \Phi) - \tilde{A}(W_P, \Phi)$, for each Φ , depends only on the value of Φ on the Γ_{ij} 's. Thus (2.5) gives rise to a set of equations on the restriction of $S_h^0(\Omega)$ to $\bigcup \Gamma_{ij}$. To solve these equations, we proceed as follows: For each Γ_{ij} choose Φ in the subspace of $S_h^0(\Omega)$ whose elements vanish in the interior mesh points of every Ω_k and on all other Γ 's and, in particular, at the endpoints of Γ_{ij} . Thus, on this subspace, (2.5) decouples into the independent problems of finding $W_E \in S_h^0(\Gamma_{ij})$ given by

$$(2.6) \quad \alpha_{ij} \langle a^{-1} \tilde{l}_0^{1/2} W_E, \Phi \rangle_{\Gamma_{ij}} = (g, \Phi) - \tilde{A}(W_P, \Phi)$$

for each Γ_{ij} . The computational aspects of solving for W_E on each Γ_{ij} are fully discussed in Section 4; however, note that these are local problems with unknowns corresponding to the nodes on Γ_{ij} and may be solved in parallel.

Next we must solve for W_V on the edges. We consider the subspace of $S_h^0(\Omega)$ consisting of functions which are linear between the endpoints of each Γ_{ij} and vanish at mesh points in Ω^h which are interior to any Ω_j . Clearly, such a subspace has dimension equal to the number of interior vertices, i.e., vertices of the Ω_k which do not lie on $\partial\Omega$. For each Φ in this subspace, $\Phi_E = 0$ and (2.5) reduces to

$$(2.7) \quad \sum_{\Gamma_{ij}} \alpha_{ij} (W_V(v_i) - W_V(v_j)) (\Phi_V(v_i) - \Phi_V(v_j)) = (g, \Phi) - \tilde{A}(W_P, \Phi).$$

A basis for this subspace may be chosen as follows: Choose Φ^1, \dots, Φ^M , where M is the number of vertices not on $\partial\Omega$ and $\Phi^i(v_j) = \delta_{ij}$ where $\delta_{ij} = 1$ if $i = j$ and 0 otherwise. This choice gives rise to a difference equation for the function W_V on the interior vertices which is independent of (2.6) and may be solved concurrently. The values W_V at the vertices determine W_V on the edges and hence $W_H = W_E + W_V$ is known on all of the edges Γ_{ij} .

The last step consists of determining W_H in each Ω_k so that

$$(2.8) \quad \tilde{A}_k(W_H, \Phi) = 0 \quad \text{for all } \Phi \in S_h^0(\Omega_k).$$

The problem of finding the solution of (2.8), given the values of W_H on the boundary of the subdomains, reduces to independent Dirichlet solves on the subdomains. Hence the solution of (1.10) is determined by $W = W_P + W_H$.

We summarize the process by outlining the steps for obtaining the solution of

$$B(W, \psi) = (g, \psi) \quad \text{for all } \psi \in S_h^0(\Omega),$$

and hence for computing the action of B^{-1} .

Algorithm DD1.

1. Find W_P by solving Dirichlet problems on subregions. The solution of the individual Dirichlet problems on the subdomains may be done in parallel.
2. Find W_E on Γ_{ij} by solving one-dimensional equations on each Γ_{ij} .
3. Find W_V on $\bigcup \Gamma_{ij}$ by first finding W_V on the interior vertices of Ω_k by solving a coarse mesh difference equation and then extending piecewise linearly to the edges

Γ_{ij} . The solution on the different segments Γ_{ij} of Step 2 and Step 3 may be done in parallel.

4. Find W_H by extending the values of $W_E + W_V$ on the Γ_{ij} discrete \tilde{A}_k -harmonically to the subregions; i.e., solve Dirichlet problems on the subregions. As in Step 1, the solutions of the individual Dirichlet problems on the various subdomains may be done in parallel.

5. Set $W = W_P + W_H$.

We shall now discuss several features of this preconditioning algorithm.

Remark 2.2. The process described above is just that which is required for applying the “action” of the matrix B^{-1} to an arbitrary vector. We again emphasize that it involves solving some local problems on subdomains which are independent of each other so that they can be solved concurrently on computers with parallel architecture.

Remark 2.3. As remarked previously in this section, the matrices of coefficients a_{ij}^k defining the forms \tilde{A}_k need only, in principle, be chosen so that they are uniformly positive definite (which implies the spectral equivalence of \tilde{A} and A). However, as will be seen in later sections, a judicious choice of a_{ij}^k can often be made which results in subdomain problems which can be “fast” solved. In Section 4, we shall explicitly show one method of choosing the coefficients a_{ij}^k in a simple way so that known efficient direct methods may be used to solve the problems on the subdomains.

Remark 2.4. The theoretical results for this algorithm remain valid independent of the number of subdomains and interior vertices used in the decomposition of Ω . This is important when the coefficients $a_{ij}(x)$ are rapidly varying, in which case preconditioners with smaller subdomains more closely reflect the behavior of the coefficients and give rise to more rapidly convergent algorithms. The freedom to use many subdomains may also prove to be important in developing the most efficient preconditioner for a computer with a large number of parallel processors.

Remark 2.5. For simplicity of presentation, we have assumed that Ω is a polygonal domain and the subdomains Ω_j are either quadrilaterals or triangles and that the subspaces consist of piecewise linear functions. The algorithm and theorem can be extended, under reasonable assumptions, to the case where Ω is a bounded domain with piecewise smooth boundary and the subdomains Ω_j have either piecewise smooth boundaries or are mesh domains which approximate piecewise smooth boundaries. We can also extend the above algorithms to a class of higher-order piecewise polynomial subspaces. We intend to deal with these extensions in a later paper.

Remark 2.6. We could also define another preconditioner by replacing the form on the left-hand side of (2.7) by a form, corresponding to a weighted identity operator, leading to the equations for W_V given by

$$(2.9) \quad \sum_{i=1}^M \alpha_i W_V(v_i) \Phi_V(v_i) = (g, \Phi) - \tilde{A}(W_p, \Phi),$$

where the sum is taken over all interior vertices of the Ω_j 's. Note that in the case $M = 1$, this formulation coincides with (2.7). The choice of basis functions Φ^i as indicated after (2.7) leads to

$$W_V(v_i) = \frac{(g, \Phi^i) - \tilde{A}(W_p, \Phi^i)}{\alpha_i}, \quad i = 1, \dots, M.$$

In this case one can prove, using the techniques given in Section 3, that Theorem 1 holds with

$$(2.10) \quad \lambda_1/\lambda_0 \leq Cd^{-2}(1 + \ln(d/h)^2),$$

where C is independent of d and h . The estimate indicates that this procedure may be reasonable if d is large, i.e., there are very few subdomains, but will become inefficient as d becomes small, i.e., as the number of subdomains is increased. This is illustrated by the results of Example 6 of Section 6.

Remark 2.7. In the case that the forms $\tilde{A}_k(\cdot, \cdot)$ and $A(\cdot, \cdot)$ coincide on functions in $S_h(\Omega_k)$ then the variables which are interior to Ω_k can be eliminated from the iterative process. Consequently, if the above forms coincide on every subdomain, then the iterative process can be reduced to a boundary iteration. The resulting algorithm is more efficient than the general algorithm in that each iteration does not require the solution of (2.4). However, much of the generality and flexibility of the general algorithm is lost.

3. A Proof of Theorem 1. In this section, we prove the main theorem of the paper which provides bounds on the condition number for the preconditioned system corresponding to (1.7). This, as previously noted, reduces to the estimation of the quantities λ_0 and λ_1 appearing in quadratic form inequalities (1.11). This will be done here in the special case of assumptions A.1, A.2, A.3, and where the finite element subspaces are as in Section 2.

We shall first need some preliminaries. We remind the reader that c or C , with or without subscript, will denote a generic positive constant which is independent of h , d , the subdivision Ω_k and the triangulation Ω^h .

The derivation of the estimates in this section requires the use of various norms defined on the subdomain boundaries. Let Ω_i be a subdomain of Ω^h (as defined in Section 2) and β_i be the set of indices jk with $\Gamma_{jk} \in \partial\Omega_i$, hence $\partial\Omega_i = \bigcup \Gamma_{jk}$ for $jk \in \beta_i$. The Sobolev space of order one half on $\partial\Omega_i$ will be denoted $H^{1/2}(\partial\Omega_i)$ and is defined in [11], [14], [16]. With d as in A.3 (roughly the diameter of Ω_i), we define the weighted norm on $H^{1/2}(\partial\Omega_i)$ by

$$(3.1) \quad |w|_{1/2, \partial\Omega_i} = \left(\int_{\partial\Omega_i} \int_{\partial\Omega_i} \frac{(w(x) - w(y))^2}{|x - y|^2} ds(x) ds(y) + d^{-1} |w|_{L^2(\partial\Omega_i)}^2 \right)^{1/2},$$

where s is arc length along $\partial\Omega_i$. If v is a smooth function on $\partial\Omega_i$ with support contained in one of the edges $\Gamma_{jk} \subset \partial\Omega_i$, then the integral term in (3.1) reduces to

$$\int_{\Gamma_{jk}} \int_{\Gamma_{jk}} \frac{(v(x) - v(y))^2}{|x - y|^2} ds(x) ds(y) + 2 \int_{\Gamma_{jk}} \int_{\partial\Omega_i / \Gamma_{jk}} \frac{v(x)^2}{|x - y|^2} ds(y) ds(x).$$

A straightforward computation gives that

$$c \int_{\partial\Omega_i / \Gamma_{jk}} |x - y|^{-2} ds(y) \leq |x - v_k|^{-1} + |x - v_j|^{-1} \leq C \int_{\partial\Omega_i / \Gamma_{jk}} |x - y|^{-2} ds(y).$$

Thus, for smooth v with support contained on Γ_{jk} , the norm in (3.1) is equivalent to

$$(3.2) \quad \left(\int_{\Gamma_{jk}} \int_{\Gamma_{jk}} \frac{(v(x) - v(y))^2}{|x - y|^2} ds(x) ds(y) + \int_{\Gamma_{jk}} \frac{v(x)^2}{|x - v_k|} + \frac{v(x)^2}{|x - v_j|} ds(x) \right)^{1/2}.$$

The space $\mathring{H}^{1/2}(\Gamma_{jk})$ is defined to be the completion of the smooth functions with compact support in Γ_{jk} with respect to the norm (3.2). We shall denote by $|\cdot|_{1/2,\Gamma_{jk}}$ the norm on $\mathring{H}^{1/2}(\Gamma_{jk})$ given by (3.2). It is well known that the space $\mathring{H}^{1/2}(\Gamma_{jk})$ is the interpolation space which is halfway between $H_0^1(\Gamma_{jk})$ and $L^2(\Gamma_{jk})$ [14], [16]. We note that the operator $(-\partial^2/\partial s^2)$ with domain of definition $H_0^1(\Gamma_{jk})$ is positive definite and selfadjoint on $L^2(\Gamma_{jk})$ and domain of $(-\partial^2/\partial s^2)^{1/2} = \mathring{H}^{1/2}(\Gamma_{jk})$. Consequently, the corresponding norm given by

$$(3.3) \quad \left(\left\langle (-\partial^2/\partial s^2)^{1/2} w, w \right\rangle_{\Gamma_{jk}} \right)^{1/2}$$

is equivalent to the norm of (3.2) on $\mathring{H}^{1/2}(\Gamma_{jk})$. We note that the discrete operator l_0 defined on $S_h^0(\Gamma_{jk})$ by

$$\langle l_0 W, \phi \rangle_{\Gamma_{jk}} = \langle W', \phi' \rangle_{\Gamma_{jk}} \quad \text{for all } \phi \in S_h^0(\Gamma_{jk}),$$

is a finite-dimensional approximation to $(-\partial^2/\partial s^2)$. Using A.2, it can be shown by interpolation [13, Theorem 9.1] that

$$(3.4) \quad c|W|_{1/2,\Gamma_{jk}}^2 \leq \left\langle l_0^{1/2} W, W \right\rangle_{\Gamma_{jk}} \leq C|W|_{1/2,\Gamma_{jk}}^2 \quad \text{for all } W \in S_h^0(\Gamma_{jk}).$$

We also note that by the assumptions on the coefficients defining \tilde{l}_0 in (2.2),

$$c \langle l_0 W, W \rangle_{\Gamma_{jk}} \leq \left\langle a^{-1} \tilde{l}_0 W, W \right\rangle_{\Gamma_{jk}} \leq C \langle l_0 W, W \rangle_{\Gamma_{jk}} \quad \text{for all } W \in S_h^0(\Gamma_{jk}).$$

By A.3 and a similar interpolation argument,

$$(3.5) \quad \begin{aligned} c \left\langle l_0^{1/2} W, W \right\rangle_{\Gamma_{jk}} &\leq \left\langle a^{-1} \tilde{l}_0^{1/2} W, W \right\rangle_{\Gamma_{jk}} \\ &\leq C \left\langle l_0^{1/2} W, W \right\rangle_{\Gamma_{jk}} \quad \text{for all } W \in S_h^0(\Gamma_{jk}). \end{aligned}$$

We shall need several lemmas which will be used in the proof of the main theorem.

LEMMA 3.1. *For $V \in S_h^0(\Gamma_{jk})$, let \bar{V} be the function which is equal to V on Γ_{jk} and is equal to zero on the remaining edges of Ω_i . Let \tilde{v} denote the \tilde{A}_i -harmonic extension of \bar{V} satisfying $\tilde{v} = \bar{V}$ on $\partial\Omega_i$ and*

$$(3.6) \quad \tilde{A}_i(\tilde{v}, \phi) = 0 \quad \text{for all } \phi \in H_0^1(\Omega_i).$$

Then

$$c\tilde{A}_i(\tilde{v}, \tilde{v}) \leq \left\langle a^{-1} \tilde{l}_0^{1/2} V, V \right\rangle_{\Gamma_{jk}} \leq C\tilde{A}_i(\tilde{v}, \tilde{v}).$$

Proof. Let $D_i(u, \phi) = \int_{\Omega_i} \nabla u \cdot \nabla \phi \, dx$, and let $v^* \in H^1(\Omega)$ satisfy $v^* = \bar{V}$ on $\partial\Omega_i$, and $D_i(v^*, \phi) = 0$ for all $\phi \in H_0^1(\Omega_i)$. Then, using a trace inequality, A.3, and the uniform positive definiteness of the $\{a_{jk}^i\}$,

$$(3.7) \quad \begin{aligned} |\bar{V}|_{1/2,\partial\Omega_i}^2 &\leq cD_i(\tilde{v}, \tilde{v}) \leq c\tilde{A}_i(\tilde{v}, \tilde{v}) \\ &\leq c\tilde{A}_i(v^*, v^*) \leq CD_i(v^*, v^*). \end{aligned}$$

Using A.3 and a well-known *a priori* inequality, we have $D_i(v^*, v^*) \leq C|\bar{V}|_{1/2,\partial\Omega_i}^2$, and hence

$$(3.8) \quad c|\bar{V}|_{1/2,\partial\Omega_i}^2 \leq \tilde{A}_i(\tilde{v}, \tilde{v}) \leq C|\bar{V}|_{1/2,\partial\Omega_i}^2.$$

The lemma easily follows from the equivalence of norms (3.1) and (3.2) for functions \bar{V} , (3.4), (3.5) and (3.8). \square

We shall need some *a priori* estimates for discrete \tilde{A}_i -harmonic functions.

LEMMA 3.2. Let $W \in S_h(\Omega)$ be discrete \tilde{A}_i -harmonic, then

$$(3.9) \quad \tilde{A}_i(W, W) \leq C |W|_{1/2, \partial\Omega_i}^2.$$

Furthermore, the following hold:

(i) If W has mean value zero on Ω_i , then

$$(3.10) \quad c |W|_{1/2, \partial\Omega_i}^2 \leq \tilde{A}_i(W, W).$$

(ii) If W vanishes at the vertices of Ω_i , then

$$(3.11) \quad \tilde{A}_i(W, W) \leq C \sum_{jk \in \beta_i} \alpha_{jk} \left\langle a^{-1} \tilde{l}_0^{1/2} W, W \right\rangle_{\Gamma_{jk}}.$$

(iii) If W is a linear function on each edge $\Gamma_{jk} \subset \partial\Omega_i$, then

$$(3.12) \quad \tilde{A}_i(W, W) \leq C \sum_{jk \in \beta_i} \alpha_{jk} (W(v_j) - W(v_k))^2.$$

Proof. The inequalities (3.9), (3.10), and (3.11) were essentially proved in [4]. For completeness we shall include a proof of (3.9) and (3.10) at the end of this section. Here we shall show how (3.11) and (3.12) follow from (3.9), (3.10) and earlier inequalities. We prove (3.11) as follows. Let Γ_{jk} be any edge of Ω_i and let W_{jk} be the discrete \tilde{A}_i -harmonic function which is equal to W on Γ_{jk} and vanishes on all the other edges of $\partial\Omega_i$. Clearly, $W = \sum_{jk \in \beta_i} W_{jk}$ and the triangle inequality yields

$$(3.13) \quad \tilde{A}_i(W, W) \leq C \sum_{jk \in \beta_i} \tilde{A}_i(W_{jk}, W_{jk}).$$

Applying (3.9), the equivalence of norms (3.1) and (3.2) on the functions W_{jk} , (3.4) and (3.5) gives

$$(3.14) \quad \begin{aligned} \tilde{A}_i(W_{jk}, W_{jk}) &\leq C |W_{jk}|_{1/2, \partial\Omega_i}^2 \\ &\leq C |W_{jk}|_{1/2, \Gamma_{jk}}^2 \leq C \left\langle a^{-1} \tilde{l}_0^{1/2} W, W \right\rangle_{\Gamma_{jk}}. \end{aligned}$$

Combining (3.13) and (3.14) proves (3.11).

We next prove (3.12). Applying (3.9) to the function $W - \beta$, where β is a constant to be determined later, gives

$$(3.15) \quad \tilde{A}_i(W, W) = \tilde{A}_i(W - \beta, W - \beta) \leq C |W - \beta|_{1/2, \partial\Omega_i}^2.$$

If Ω_i is a triangle (resp. quadrilateral), let W^* be the linear (resp. bilinear) function on Ω_i which has the same boundary values as $W - \beta$. Choosing β so that the average of W^* on Ω_i is zero, applying a trace and Poincaré inequality gives

$$(3.16) \quad |W - \beta|_{1/2, \partial\Omega_i}^2 \leq CD_i(W^*, W^*).$$

An elementary calculation yields

$$(3.17) \quad \begin{aligned} D_i(W^*, W^*) &\leq C \sum_{jk \in \beta_i} (W(v_j) - W(v_k))^2 \\ &\leq C \sum_{jk \in \beta_i} \alpha_{jk} (W(v_j) - W(v_k))^2. \end{aligned}$$

Thus (3.12) follows from (3.15), (3.16), and (3.17). \square

An important ingredient for our analysis is a certain type of discrete Sobolev inequality. Let $\hat{\Omega}$ be a polygonal domain which satisfies a cone condition with radius d and angle γ . Let \tilde{S}_h for $0 < h < 1$ be any family of subspaces of $W_\infty^1(\hat{\Omega})$ satisfying the inverse inequality

$$(3.18) \quad \|\nabla W\|_{L^\infty(\hat{\Omega})} \leq C_1 h^{-1} \|W\|_{L^\infty(\hat{\Omega})} \quad \text{for all } W \in \tilde{S}_h,$$

where $d \geq h$. We then have the following lemma.

LEMMA 3.3. *There exists a positive constant C independent of h and d and depending only on γ and C_1 in (3.18) such that*

$$(3.19) \quad \|W\|_{L^\infty(\hat{\Omega})}^2 \leq C \left(d^{-2} \|W\|_{L^2(\hat{\Omega})}^2 + \ln(d/h) D_{\hat{\Omega}}(W, W) \right) \quad \text{for all } W \in \tilde{S}_h,$$

where $D_{\hat{\Omega}}(\cdot, \cdot)$ denotes the Dirichlet form on $\hat{\Omega}$.

Various discrete Sobolev inequalities have appeared in the literature [3], [19]. Since the results in the literature do not correspond exactly to the given lemma, we shall include an elementary proof of the lemma after the proof of Theorem 1.

Some consequences of Lemma 3.3 which are important in our present considerations are the following discrete type Sobolev inequalities.

LEMMA 3.4. *Let W be in $S_h(\Omega_i)$.*

(i) *If $W(p) = 0$ for some point $p \in \bar{\Omega}_i$, then*

$$(3.20) \quad \|W\|_{L^\infty(\Omega_i)}^2 \leq C(1 + \ln(d/h)) \tilde{A}_i(W, W).$$

(ii) *For any function $W \in S_h(\Omega_i)$,*

$$(3.21) \quad \sum_{jk \in \beta_i} \alpha_{jk} (W(v_j) - W(v_k))^2 \leq C(1 + \ln(d/h)) \tilde{A}_i(W, W).$$

Proof. In order to prove (3.20) we first observe that by A.2, (3.19) is satisfied for $\tilde{S}_h = S_h(\Omega_i)$. Let α be the average value of W on Ω_i . Applying the Poincaré inequality yields

$$d^{-2} \|W - \alpha\|_{L^2(\Omega_i)}^2 \leq CD_i(W, W) \leq C\tilde{A}_i(W, W).$$

Thus, applying Lemma 3.3 to the function $W - \alpha$ gives

$$\|W - \alpha\|_{L^\infty(\Omega_i)}^2 \leq C(1 + \ln(d/h)) \tilde{A}_i(W, W).$$

Note that since $W(p) = 0$,

$$(3.22) \quad |\alpha| \leq \|W - \alpha\|_{L^\infty(\Omega_i)},$$

and (3.20) follows by the triangle inequality.

Since

$$\sum_{jk \in \beta_i} \alpha_{jk} (W(v_j) - W(v_k))^2 \leq C \sum_{jk \in \beta_i} (W(v_j) - W(v_k))^2,$$

the inequality (3.21) follows by applying (3.20) to the function $W(x) - W(v_m)$, where v_m is a vertex of Ω_i . This completes the proof of the lemma. \square

LEMMA 3.5. Let $W \in S_h(\Omega_i)$ satisfy $W = 0$ on the vertices of Ω_i and let $W_L \in S_h(\Omega_i)$ be a discrete \tilde{A}_i -harmonic function which is linear on each edge $\Gamma_{jk} \subset \partial\Omega_i$. Then

$$(3.23) \quad \sum_{jk \in \beta_i} \alpha_{jk} \left\langle a^{-1} \tilde{l}_0^{1/2} W, W \right\rangle_{\Gamma_{jk}} \leq C(1 + \ln(d/h)^2) \tilde{A}_i(W + W_L, W + W_L).$$

Proof. We shall first prove (3.23) in the case that $W_L = 0$. Let Γ_{jk} be any edge of Ω_i . It follows from (3.5), (3.4), (3.2), and (3.1) that

$$(3.24) \quad \alpha_{jk} \left\langle a^{-1} \tilde{l}_0^{1/2} W, W \right\rangle_{\Gamma_{jk}} \leq c \left\{ |W|_{1/2, \partial\Omega_i}^2 + \int_{\Gamma_{jk}} \frac{W(x)^2}{|x - v_k|} + \frac{W(x)^2}{|x - v_j|} ds(x) \right\}.$$

Let α be the average value of W on Ω_i . Applying (3.22), (3.10), and (3.20) leads to

$$\begin{aligned} |W|_{1/2, \partial\Omega_i}^2 &\leq C(|\alpha|^2 + |W - \alpha|_{1/2, \partial\Omega_i}^2) \\ &\leq C(\|W - \alpha\|_{L^\infty(\Gamma_{jk})}^2 + \tilde{A}_i(W, W)) \\ &\leq C(1 + \ln(d/h)) \tilde{A}_i(W, W). \end{aligned}$$

Hence, it suffices to show that

$$(3.25) \quad \begin{aligned} I(W) &\equiv I_1(W) + I_2(W) \\ &\equiv \int_{\Gamma_{jk}} \frac{W(x)^2}{|x - v_k|} ds(x) + \int_{\Gamma_{jk}} \frac{W(x)^2}{|x - v_j|} ds(x) \\ &\leq C(1 + \ln(d/h)^2) \tilde{A}_i(W, W). \end{aligned}$$

Without loss of generality, we assume that v_k is the origin and that Γ_{jk} is the line segment with $x_1 = 0$ and $x_2 \in [0, Y]$. Then,

$$I_1(W) = \int_0^Y \frac{W(0, y)^2}{y} dy.$$

Let y_1 be the y -value of the node on Γ_{jk} closest to zero. We bound the preceding integral by considering

$$(3.26) \quad \int_0^Y \frac{W(0, y)^2}{y} dy = \int_0^{y_1} \frac{W(0, y)^2}{y} dy + \int_{y_1}^Y \frac{W(0, y)^2}{y} dy.$$

Note that by A.2, $ch \leq y_1 \leq Ch$. Therefore, by the mean-value theorem (using the hypothesis that $W(0, 0) = 0$),

$$(3.27) \quad \int_0^{y_1} \frac{W(0, y)^2}{y} dy \leq Ch^2 \left\| \frac{\partial W(0, \cdot)}{\partial y} \right\|_{L^\infty([0, y_1])}^2 \leq C \|W\|_{L^\infty(\Omega_i)}^2,$$

where the second inequality used the inverse property for the subspace $S_h^0(\Omega_i)$. Hence, by (3.20) and (3.27),

$$\int_0^{y_1} \frac{W(0, y)^2}{y} dy \leq C(1 + \ln(d/h)) \tilde{A}_i(W, W).$$

For the second term in (3.26) we have

$$\int_{y_1}^Y \frac{W(0, y)^2}{y} dy \leq \|W\|_{L^\infty(\Omega_i)}^2 \int_{y_1}^Y \frac{dy}{y} \leq C(1 + \ln(d/h)^2) \tilde{A}_i(W, W).$$

Combining the above estimates gives a bound for the first term in (3.25); the second term is estimated similarly. Hence (3.23) follows in the case that $W_L = 0$.

To prove (3.23) in the general case, let W_\perp be the function in $S_h(\Omega_i)$ which satisfies $W_\perp(v_m) = W_L(v_m)$ for all vertices v_m of $\partial\Omega_i$ and $\tilde{A}_i(W_\perp, \phi) = 0$ for all $\phi \in S_h(\Omega_i)$ with $\phi(v_m) = 0$ on all vertices of $\partial\Omega_i$. Notice that $W + W_L - W_\perp$ vanishes at the vertices of Ω_i . Applying the arithmetic-geometric mean inequality and the special case of (3.23) proved above gives

$$\begin{aligned} & \alpha_{jk} \left\langle a^{-1} \tilde{l}_0^{1/2} W, W \right\rangle_{\Gamma_{jk}} \\ & \leq 2\alpha_{jk} \left\langle a^{-1} \tilde{l}_0^{1/2} (W + W_L - W_\perp), (W + W_L - W_\perp) \right\rangle_{\Gamma_{jk}} \\ & \quad + 2\alpha_{jk} \left\langle a^{-1} \tilde{l}_0^{1/2} (W_L - W_\perp), (W_L - W_\perp) \right\rangle_{\Gamma_{jk}} \\ & \leq C (1 + \ln(d/h)^2) \tilde{A}_i((W + W_L - W_\perp), (W + W_L - W_\perp)) \\ & \quad + 2\alpha_{jk} \left\langle a^{-1} \tilde{l}_0^{1/2} (W_L - W_\perp), (W_L - W_\perp) \right\rangle_{\Gamma_{jk}}. \end{aligned}$$

Since the functions $(W + W_L - W_\perp)$ and W_\perp are orthogonal in the $\tilde{A}_i(\cdot, \cdot)$ -inner product,

$$\tilde{A}_i((W + W_L - W_\perp), (W + W_L - W_\perp)) \leq \tilde{A}_i(W + W_L, W + W_L).$$

Thus to complete the proof of the lemma we need only show that

$$\begin{aligned} (3.28) \quad & \left\langle a^{-1} \tilde{l}_0^{1/2} (W_L - W_\perp), (W_L - W_\perp) \right\rangle_{\Gamma_{jk}} \\ & \leq C(1 + \ln(d/h)^2) \tilde{A}_i(W + W_L, W + W_L). \end{aligned}$$

Since $W_L - W_\perp$ vanishes at the vertices of Ω_i , applying inequality (3.24) and the subsequent arguments give

$$\begin{aligned} & \left\langle a^{-1} \tilde{l}_0^{1/2} (W_L - W_\perp), (W_L - W_\perp) \right\rangle_{\Gamma_{jk}} \\ & \leq C(1 + \ln(d/h)) \tilde{A}_i(W_L - W_\perp, W_L - W_\perp) + I(W_L - W_\perp), \end{aligned}$$

where I is defined in (3.25). Since W_\perp is orthogonal to $W_L - W_\perp$ in the $\tilde{A}_i(\cdot, \cdot)$ -inner product, we have in view of (3.12) and (3.21) that

$$\begin{aligned} & \tilde{A}_i(W_L - W_\perp, W_L - W_\perp) \leq \tilde{A}_i(W_L, W_L) \\ & \leq c \sum_{jk \in \beta_i} \alpha_{jk} [(W(v_j) + W_L(v_j)) - (W(v_k) + W_L(v_k))]^2 \\ & \leq C(1 + \ln(d/h)) \tilde{A}_i(W + W_L, W + W_L). \end{aligned}$$

Hence, in order to complete the proof of (3.28), it suffices to show that

$$(3.29) \quad I(W_L - W_\perp) \leq C(1 + \ln(d/h)^2) \tilde{A}_i(W + W_L, W + W_L).$$

Now with I_1 and I_2 as in (3.25) we have by the arithmetic-geometric mean inequality,

$$(3.30) \quad I_1(W_L - W_\perp) \leq 2I_1(W_\perp - W_\perp(v_k)) + 2I_1(W_L - W_L(v_k)).$$

For the first term on the right we apply (3.25) and then use the fact that $\tilde{A}_i(W_\perp, W_\perp) \leq \tilde{A}_i(W + W_L, W + W_L)$ to obtain

$$I_1(W_\perp - W_\perp(v_k)) \leq C(1 + \ln(d/h)^2)\tilde{A}_i(W + W_L, W + W_L).$$

A simple calculation using the linearity of W_L on Γ_{jk} and (3.21) yields

$$\begin{aligned} I_1(W_L - W_L(v_k)) &\leq C\alpha_{jk}(W_L(v_j) - W_L(v_k))^2 \\ &\leq C(1 + \ln(d/h))\tilde{A}_i(W + W_L, W + W_L). \end{aligned}$$

Thus

$$I_1(W_L - W_\perp) \leq C(1 + \ln(d/h)^2)\tilde{A}_i(W + W_L, W + W_L).$$

Obviously, the same bound holds for $I_2(W_L - W_\perp)$, which completes the proof of (3.29) and hence the lemma. \square

We are now in a position to prove Theorem 1.

Proof of Theorem 1. By the uniform positive definiteness of the matrices $\{a_{jk}\}$ and $\{a_{jk}^i\}$,

$$c\tilde{A}(W, W) \leq A(W, W) \leq C\tilde{A}(W, W) \quad \text{for all } W \in S_h^0(\Omega).$$

Hence, it suffices to compare the quadratic forms $\tilde{A}(\cdot, \cdot)$ with $B(\cdot, \cdot)$. As in Section 2, we decompose $W \in S_h^0(\Omega)$ into $W = W_P + W_E + W_V$. With $W_H = W_E + W_V$, we have (as noted in Section 2)

$$\tilde{A}(W, W) = \tilde{A}(W_P, W_P) + \tilde{A}(W_H, W_H)$$

and

$$B(W, W) = \tilde{A}(W_P, W_P) + B(W_H, W_H).$$

Hence, it suffices to compare $\tilde{A}(W_H, W_H)$ with $B(W_H, W_H)$. More specifically, the proof will be complete when we have shown that

$$(3.31) \quad \tilde{A}(W_H, W_H) \leq CB(W_H, W_H),$$

and

$$(3.32) \quad B(W_H, W_H) \leq C(1 + \ln(d/h)^2)\tilde{A}(W_H, W_H).$$

Consider a subdomain Ω_i . Using the arithmetic-geometric mean inequality, (3.11) and (3.12) yield

$$\begin{aligned} \tilde{A}_i(W_H, W_H) &\leq 2(\tilde{A}_i(W_E, W_E) + \tilde{A}_i(W_V, W_V)) \\ &\leq C \sum_{jk \in \beta_i} \alpha_{jk} \left(\left\langle a^{-1} \tilde{l}_0^{1/2} W_E, W_E \right\rangle_{\Gamma_{jk}} + (W_V(v_j) - W_V(v_k))^2 \right). \end{aligned}$$

Summing with respect to i gives (3.31). In view of (3.21) applied to W_V , and (3.23) applied to W_E and W_V (replacing W and W_L , respectively, in (3.23)), we have on each Ω_i ,

$$\begin{aligned} \sum_{jk \in \beta_i} \alpha_{jk} \left(\left\langle a^{-1} \tilde{l}_0^{1/2} W_E, W_E \right\rangle_{\Gamma_{jk}} + (W_V(v_j) - W_V(v_k))^2 \right) \\ \leq C(1 + \ln(d/h)^2)\tilde{A}_i(W_H, W_H), \end{aligned}$$

and summing with respect to i gives (3.32) which completes the proof of the theorem in the case where interior vertices are present.

We now turn to the case where all the vertices of the Ω_i lie on $\partial\Omega$. Since the function W_V vanishes at all the vertices, it follows that $W_V \equiv 0$ on Ω so that $W_H = W_E$ and hence the preconditioning form B simplifies to

$$B(W, W) = \tilde{A}(W_P, W_P) + \sum_{\Gamma_{jk}} \alpha_{jk} \left\langle a^{-1} \tilde{l}_0^{1/2} W_E, W_E \right\rangle_{\Gamma_{jk}}.$$

Reasoning as in the proof of the first part of this theorem, we have $\tilde{A}(W_E, W_E) \leq CB(W_E, W_E)$ which is a special case of (3.31). Hence, we need only show the sharper version of (3.32),

$$(3.33) \quad B(W_E, W_E) \leq C\tilde{A}(W_E, W_E).$$

Let us first note that $W_E \equiv 0$ on $\partial\Omega$ and by our assumption that the vertices of the Ω_i are in this case fixed independent of h , there are a fixed number of interior edges Γ_{jk} on which possibly $W_E \neq 0$. Consider any interior edge Γ_{jk} . Identifying Γ_{jk} as two segments, say Γ_{jk}^1 and Γ_{jk}^2 with opposite orientations, it is not difficult to prove that either

(i) Γ_{jk} separates Ω into two polygonal domains, one of which (say $\tilde{\Omega}_1$) has a boundary $\partial\tilde{\Omega}_1$ consisting of Γ_{jk}^1 and nonempty parts of $\partial\Omega$ meeting each vertex of Γ_{jk}^1 , or

(ii) Γ_{jk}^1 and Γ_{jk}^2 may be considered to be part of the boundary $\partial\tilde{\Omega}_1$ of a polygonal domain bounding a subdomain, say $\tilde{\Omega}_1$ of Ω , where Γ_{jk}^1 and Γ_{jk}^2 are separated (there is a positive distance with respect to arc length along $\partial\tilde{\Omega}_1$ between them) by components of $\partial\Omega$.

In either case, by equivalence of norms,

$$\alpha_{jk} \left\langle a^{-1} \tilde{l}_0^{1/2} W_E, W_E \right\rangle_{\Gamma_{jk}} \equiv \alpha_{jk} \left\langle a^{-1} \tilde{l}_0^{1/2} W_E, W_E \right\rangle_{\Gamma_{jk}^1} \leq c |W_E|_{1/2, \partial\tilde{\Omega}_1}^2.$$

Using a trace theorem for $\tilde{\Omega}_1$ we have

$$\alpha_{jk} \left\langle a^{-1} \tilde{l}_0^{1/2} W_E, W_E \right\rangle_{\Gamma_{jk}} \leq C\tilde{A}(W_E, W_E).$$

The inequality (3.33) follows after summing over all interior edges Γ_{jk} . This completes the proof of Theorem 1.

In the remainder of this section, we shall give the proof of Lemma 3.3 and the inequalities (3.9) and (3.10) of Lemma 3.2.

Proof of (3.9) and (3.10). It is not difficult to see (by scaling Ω_i to unit size and using A.3) that it suffices to prove (3.9) and (3.10) under the assumption that $d = 1$.

We first prove (3.9). Let $W \in S_h(\Omega_i)$ be \tilde{A}_i -discrete harmonic and w be the \tilde{A}_i -harmonic function defined by

$$\tilde{A}_i(w, \phi) = 0 \quad \text{for all } \phi \in H_0^1(\Omega_i), \quad w = W \quad \text{on } \partial\Omega_i.$$

Using the well-known *a priori* inequality for harmonic functions, $\tilde{A}_i(w, w) \leq c |W|_{1/2, \partial\Omega_i}^2$, and the triangle inequality, it suffices to prove

$$(3.34) \quad \tilde{A}_i(w - W, w - W) \leq c |W|_{1/2, \partial\Omega_i}^2.$$

Now from the definition of w and W it follows easily that

$$\tilde{A}_i(w - W, w - W) \leq \inf \tilde{A}_i(\Phi - w, \Phi - w),$$

with the infimum taken over functions $\Phi \in S_h(\Omega_i)$ with $\Phi = W$ on $\partial\Omega_i$. By well-known properties of $S_h(\Omega_i)$, we see that for $0 < \varepsilon < 1/2$,

$$\inf \tilde{A}_i(\Phi - w, \Phi - w) \leq Ch^{2\varepsilon} \|w\|_{H^{1+\varepsilon}(\Omega_i)}^2.$$

Now using a well-known *a priori* inequality (cf. [14], [16]) and an “inverse property” implied by A.2, we see that

$$h^{2\varepsilon} \|w\|_{H^{1+\varepsilon}(\Omega_i)}^2 \leq Ch^{2\varepsilon} |W|_{H^{1/2+\varepsilon}(\partial\Omega_i)}^2 \leq C |W|_{1/2, \partial\Omega_i}^2,$$

which proves (3.34) and hence also (3.9).

We next prove (3.10). Let $W \in S_h(\Omega)$ have mean value zero on Ω_i . Applying a trace inequality gives $|W|_{1/2, \partial\Omega_i}^2 \leq c\{D_i(W, W) + \|W\|_{L^2(\Omega)}^2\}$. Using a Poincaré inequality and the assumptions on the coefficients a_{jk}^i defining \tilde{A}_i yields

$$|W|_{1/2, \partial\Omega_i}^2 \leq CD_i(W, W) \leq C\tilde{A}_i(W, W),$$

which proves (3.10). \square

Proof of Lemma 3.3. Let W be in \tilde{S}_h and x be a point of $\hat{\Omega}$ where $|W(x)| = \|W\|_{L^\infty(\hat{\Omega})}$. Let $\Lambda \subset \hat{\Omega}$ be a cone of radius d , angle γ , and vertex x . Without loss of generality assume that $x = 0$. For any point y in Λ , the Fundamental Theorem of Calculus gives that

$$W(0) = W(y) - \int_0^{|y|} \nabla W \left(\frac{ty}{|y|} \right) \cdot \frac{y}{|y|} dt.$$

Breaking up the integral into two regions and integrating over θ gives

$$\begin{aligned} \gamma |W(0)| &\leq \left| \int_0^\gamma W(y(\theta)) d\theta \right| + \left| \int_0^\gamma \int_0^{\delta h} \nabla W \left(\frac{ty(\theta)}{|y(\theta)|} \right) \cdot \frac{y(\theta)}{|y(\theta)|} dt d\theta \right| \\ &\quad + \left| \int_{\Lambda/\Lambda_{\delta h}} \nabla W(\zeta) \cdot \frac{\zeta}{|\zeta|^2} d\zeta \right|, \end{aligned}$$

where $\delta > 0$ is to be chosen and $\Lambda_{\delta h}$ is the cone contained in Λ of radius δh , angle γ and vertex $(0, 0)$. The second term above is estimated by

$$\left| \int_0^\gamma \int_0^{\delta h} \nabla W \left(\frac{ty(\theta)}{|y(\theta)|} \right) \cdot \frac{y(\theta)}{|y(\theta)|} dt d\theta \right| \leq \gamma \delta h \|\nabla W\|_{L^\infty(\hat{\Omega})} \leq c\delta |W(0)|$$

and then kicked back. Applying the Schwarz inequality to the first and the third terms gives

$$\begin{aligned} |W(0)| &\leq C \left\{ \left(\int_0^\gamma |W(y(\theta))|^2 d\theta \right)^{1/2} + \|\nabla W\|_{L^2(\hat{\Omega})} \left(\int_{\Lambda/\Lambda_{\delta h}} \frac{d\zeta}{|\zeta|^2} \right)^{1/2} \right\} \\ (3.35) \quad &\leq C \left\{ \left(\int_0^\gamma |W(y(\theta))|^2 d\theta \right)^{1/2} + \ln \left(\frac{d}{h} \right)^{1/2} \|\nabla W\|_{L^2(\hat{\Omega})} \right\}. \end{aligned}$$

Squaring Eq. (3.35), applying the arithmetic-geometric mean inequality, multiplying by $|y|$ and integrating from 0 to d with respect to $|y|$ gives

$$\frac{d^2}{2} W(0)^2 \leq C \left(\|W\|_{L^2(\hat{\Omega})}^2 + \frac{d^2}{2} \ln \left(\frac{d}{h} \right) D(W, W) \right)$$

which completes the proof of the lemma. \square

4. The Choice of the Coefficients Defining \tilde{A} ; the “Fast Solvability” of the Subproblems. As stated in Section 2, the application of our preconditioning algorithm involves solving the subproblems defined in (2.4), (2.6), (2.7) and (2.8). We have shown that Theorem 1 holds when the coefficients defining these problems are chosen among a rather large class of functions. Once a specific choice of coefficients has been made, one is then faced with the problem of choosing a method for solving the resulting equations. There are many ways in which this can be done. In general, a good choice may depend on the particular problem and the architecture of the particular computing machine. It is for this reason that we have stated Theorem 1 in a somewhat general form. In a forthcoming paper we shall give a detailed discussion of some possible approaches which are applicable to a large variety of boundary value problems and which have the additional feature that the subproblems (2.4), (2.6) and (2.8) may be solved by known “fast” methods. Our aim in this section is to give a very brief discussion of a very special case of one of these methods, which will be used in the calculation of the examples given in Section 6.

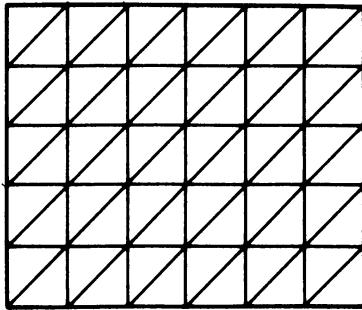


FIGURE 4.1. *A regular mesh on a rectangle.*

A globally efficient algorithm results when the original domain is split into sub-regions whose subproblems can be efficiently solved. The possibility of “fast solvability” for a subproblem is inherently linked to the coefficients defining \tilde{A}_k and the geometry of the mesh on the subdomain. We shall first consider some simple sub-domain problems where “fast solvers” are available. We will then show how these solvers can be used with Algorithm DD1 to solve much more complex problems on the original domain.

We shall begin by discussing “fast solvability” in the special case where a sub-domain is a rectangle which has been triangulated with a regular mesh and then indicate how this may be extended to more general subregions. For m and n positive integers and $0 < h < 1$, let $\Omega_k = R$ denote the rectangle $R = \{(x_1, x_2) \mid 0 < x_1 < mh, 0 < x_2 < nh\}$. Assume that R has been triangulated with a uniform mesh of size h as indicated in Figure 4.1. Here, for simplicity of notation, the dependence of R on m , n , h and k has been omitted. The choice of m , n , and h will be clear in the context in which they are used. As before, let $S_h(R)$ be the space of piecewise linear functions defined relative to the given triangulation and $S_h^0(R)$ the subspace of $S_h(R)$ whose functions vanish on ∂R . We shall restrict ourselves to a particularly simple choice of the form \tilde{A}_k , namely a constant times the Dirichlet form for the Laplace operator,

$$(4.1) \quad \tilde{A}_k(u, v) = q_k D_k(u, v),$$

where q_k is a constant which is to be chosen below.

Let us first consider the problem of inverting (2.4). Let β denote the vector whose components $\{\beta_j\}$ are the values of W_P at the interior nodal points of the triangulation, ordered say successively along columns. Then the corresponding set of linear equations can be written as

$$(4.2) \quad M\beta = b,$$

where M is the $(m - 1) \times (m - 1)$ block tridiagonal matrix with block order $n - 1$ given by

$$M = q_k \begin{pmatrix} T & -I & & & \\ -I & T & -I & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \ddots \\ & & & -I & T & -I \\ & & & & -I & T \end{pmatrix}.$$

Here I is the $(n - 1) \times (n - 1)$ identity matrix and T is the $(n - 1) \times (n - 1)$ matrix

$$T = \begin{pmatrix} 4 & -1 & & & & \\ -1 & 4 & -1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & & & \ddots & \ddots & -1 \\ & & & & -1 & 4 & -1 \\ & & & & & -1 & 4 \end{pmatrix}.$$

The matrix M of course corresponds to the usual five-point centered difference approximation to $-\Delta$. It is well known that (4.2) may be solved by, for example, “fast” direct methods. An excellent discussion of some of these methods, which are also applicable to problems more general than (4.1), (4.2) on rectangles, may be found in Swarztrauber [18]. Let us just mention that using a discrete Fourier method or a cyclic reduction algorithm, the computational complexity of solving (4.2) on a serial machine is $O(mn \log(n))$. The Facr(1) algorithm which uses a combination of both of these is shown in [18] to have a computational complexity of $O(mn \log \log(n))$, where for convenience we have taken $n \leq m$.

Let us now turn to the choice of the constant q_k . As noted in Remark 2.3, any $q_k > 0$ will satisfy the hypothesis of Theorem 1. It is obvious, however, that \tilde{A} should model the original form A as closely as possible (with this simple choice of q_k). One prescription is as follows: Choose any point, say $\bar{x} \in R$, and let q_0 and q_1 be the smallest and largest eigenvalue of the matrix $\{a_{ij}(\bar{x})\}$ (which may be trivially calculated). We can choose q_k to be any number satisfying

$$(4.3) \quad q_0 \leq q_k \leq q_1.$$

Obviously, the problem (2.8) may be handled in exactly the same fashion since it also can be reduced to solving (4.2) with appropriate b .

We now turn to problem (2.6). Let Γ_{ij} be an edge which is on the common boundary segment for the subdomains say Ω_k and Ω_l . Since the mesh is equally spaced on Γ_{ij} , we take $a = 1$ (see (2.2)) on Γ_{ij} and hence, $\tilde{l}_0^{1/2} = l_0^{1/2}$. Without loss of generality, assume that there are $n - 1$ (as opposed to $(m - 1)$) interior nodes

on Γ_{ij} . Let β be the vector of length $n - 1$ whose components $\{\beta_p\}$ are the nodal values of W_E (the solution of (2.6)) on Γ_{ij} and let $\{\Phi_p\}$ be the nodal basis for $S_h^0(\Gamma_{ij})$. The problem of computing β is the same as the matrix problem $N\beta = \gamma$, where N is given by

$$(4.4) \quad N_{pq} = \alpha_{ij} \left\langle a^{-1} \tilde{l}_0^{1/2} \Phi_p, \Phi_q \right\rangle_{\Gamma_{ij}}.$$

Since the nodes are equally spaced, the eigenvectors of N are given by

$$(4.5) \quad \Psi_p = \begin{pmatrix} \sin(\pi p/n) \\ \sin(2\pi p/n) \\ \vdots \\ \sin((n-1)\pi p/n) \end{pmatrix}.$$

The eigenvalues for N are then given by

$$(4.6) \quad \lambda_p = \alpha_{ij} \sqrt{\frac{(2 - 2 \cos(\pi p/n))(4 + 2 \cos(\pi p/n))}{6}}.$$

Thus the computation of nodal values of W_E reduces to the expansion of γ in terms of the eigenvectors (4.5), the division of the resulting coefficients by the eigenvalues (4.6), and the evaluation of the resulting expansion (with the divided coefficients) at the nodal values. Both the expansion of γ in terms of the eigenvectors and the subsequent evaluation at the nodes reduce to discrete sine transforms. Using the Fast Fourier Transform, the sine transforms and hence the solution of (2.6) on each edge Γ_{ij} can be computed in computational work on the order of $O(n \log(n))$.

Finally, we come to the choice of the coefficients α_{ij} and solution of the difference equation (2.7). Again, let Γ_{ij} be an interior edge which is a common boundary segment of the two subdomains Ω_k and Ω_l . We take

$$(4.7) \quad \alpha_{ij} = q_k + q_l.$$

We solve the difference equations (2.7) by applying some standard method, for example sparse Gaussian elimination techniques [8], [9]. We emphasize that (2.7) may be solved in parallel with (2.6) and if the number of internal nodes is reasonable, the cost of solving (2.7) will be negligible.

We now turn our attention to the more general situation where the subregions may be quadrilaterals. For simplicity of presentation, let us indicate by an example how the subproblems (2.4), (2.6) and (2.8) may be set up so as to utilize the fast solution methods previously mentioned. Consider the region given in Figure 4.2. The key idea here is to set up a mesh on each subdomain Ω_k which is topologically equivalent to a regular mesh on R . Then, the coefficients a_{ij}^k defining \hat{A}_k can be chosen so that the matrix problem for (2.4) and (2.8) can be solved by fast direct methods. To make this example nontrivial we shall consider subspaces which are somewhat refined near the nonconvex corners of the domain.

We start by imposing a “radial like” rectangular refinement around these corners which is then extended to the rest of the domain as illustrated in Figure 4.3. The final triangular grid is shown in Figure 4.4 and is formed by subdividing each quadrilateral in Figure 4.3 into two triangles.

Obviously, the triangulation of each subdomain Ω_k in Figure 4.4 has the same topological nodal structure as that of R in Figure 4.1, and we can order the nodal points in precisely the same fashion. In fact, the triangulations are equivalent in

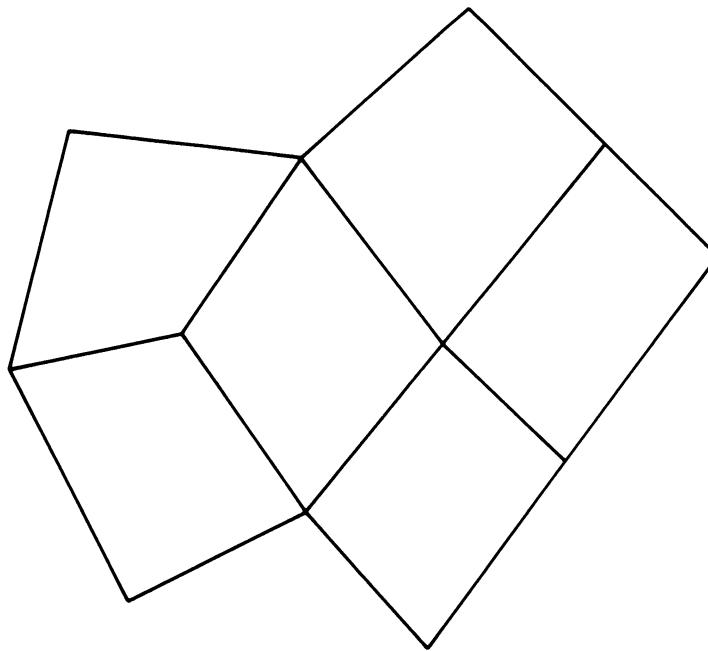


FIGURE 4.2. *A more general domain.*

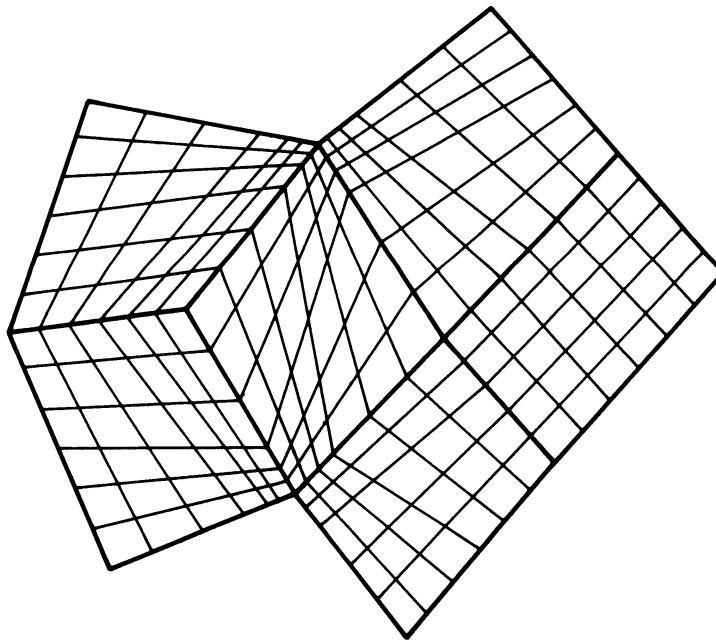
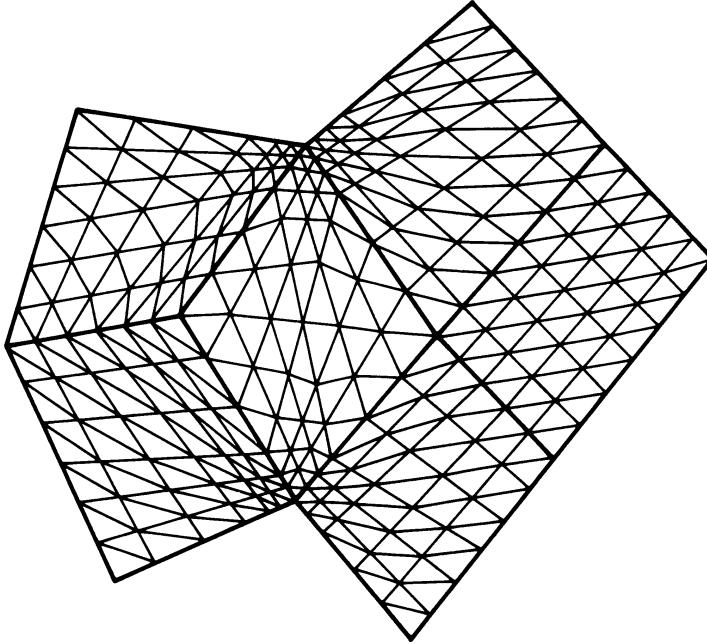


FIGURE 4.3. *The domain with rectangular subdivision.*

the sense that there exists a nondegenerate piecewise linear mapping of the triangulation of Figure 4.1 onto any of the subdomain triangulations of Figure 4.4. It can be shown that any simply connected piecewise smooth domain may be triangulated in an analogous way and hence our assumption that Ω_k is a quadrilateral is just for convenience of presentation.

We shall use the piecewise linear mapping mentioned above to define the coefficients a_{ij}^k . Fix k and let T be the corresponding piecewise linear mapping of R

FIGURE 4.4. *The domain with triangular subdivision.*

onto Ω_k . We clearly have, by a change of variable, that

$$(4.8) \quad q_k \int_R |\nabla v|^2 dx = \sum_{i,j=1}^2 \int_{\Omega_k} a_{ij}^k \frac{\partial v(T^{-1}(x))}{\partial x_i} \frac{\partial v(T^{-1}(x))}{\partial x_j} dx \\ \equiv \tilde{A}_k(v, v),$$

where $\{a_{ij}^k\}$ is a piecewise constant 2×2 matrix (here q_k is a positive constant to be defined later). We use (4.8) to define a_{ij}^k . If $\{\Phi_i\}$ denotes the usual nodal basis for $S_h(\Omega)$ restricted to Ω_k , then $\{\Psi_i \equiv \Phi_i \circ T\}$ is the usual nodal basis for the subspace $S_h(R)$. Thus, the form \tilde{A} applied to basis functions is given by

$$(4.9) \quad \tilde{A}_k(\Phi_i, \Phi_j) = q_k \int_R \nabla \Psi_i \cdot \nabla \Psi_j dx.$$

In light of (4.9), it is clear that the coefficients a_{ij}^k need never be computed. Furthermore, the matrix problem for the solution of (2.4) and (2.8) is given by (4.2) and hence can be fast solved.

The constant q_k may be chosen in the following manner: Let \bar{x} be any fixed point of Ω_k . By a change of variable, we clearly have

$$\int_{\Omega_k} a_{ij} \frac{\partial v}{\partial x_i} \frac{\partial v}{\partial x_j} dx = \int_R \bar{a}_{ij} \frac{\partial v(T(x))}{\partial x_i} \frac{\partial v(T(x))}{\partial x_j} dx.$$

Let q_1 and q_0 denote the largest and smallest eigenvalue of the matrix $\{\bar{a}_{ij}(\bar{x})\}$. Then we may choose q_k to be any number

$$(4.10) \quad q_0 \leq q_k \leq q_1.$$

This again is a trivial calculation. We remark that it is easily seen that \tilde{A}_k corresponds, in this case, to a form whose coefficients a_{ij}^k are piecewise constant functions on the triangles of the triangulation of Ω_k .

We now turn to defining \tilde{l}_0 and solving the problem (2.6). As in the definition of the coefficients a_{ij}^k above, we shall use a “mapping” technique to define the coefficient a in (2.2). Fix ij and assume that Γ_{ij} has $n - 1$ interior nodes and is a common boundary segment between subdomains Ω_k and Ω_l . Let L denote the line segment $[0, nh]$ which has an associated equally-spaced mesh with n equal segments. There is a piecewise linear mapping T which maps the mesh of L onto the nodes of Γ_{ij} . We clearly have, by a variable change,

$$\langle v \circ T, v \circ T \rangle_L = \langle a^{-1}v, v \rangle_{\Gamma_{ij}}$$

and

$$\left\langle \frac{dv(T(x))}{dx}, \frac{dv(T(x))}{dx} \right\rangle_L = \langle av', v' \rangle_{\Gamma_{ij}} = \langle a^{-1}\tilde{l}_0v, v \rangle_{\Gamma_{ij}}$$

for appropriate piecewise constant a . Let β be the vector whose components are the nodal values of W_E on Γ_{ij} . Then, β is the solution of

$$(4.11) \quad N\beta = \gamma$$

for an appropriate right-hand side vector γ , where the matrix N is given by (4.4). The discussion following (4.4) describes an efficient procedure for solving (4.11). Again, the coefficient a is a theoretical device and need never be actually computed in the implementation of Algorithm DD1.

Finally, the coefficients in (2.7) are chosen in exactly the same way as for the case of a rectangle, i.e., they are defined by (4.7) and (4.10).

5. Matrix Representation of the Operators. In this section we will describe the action of inverting the preconditioner B (given by Algorithm DD1) in terms of block matrices. It will be shown that B has a special structure and that the process for solving $B\alpha = \beta$ previously described may also be seen to be a block Gauss elimination process with an appropriate basis.

We consider the inversion of B in terms of basis functions of the following form:

1. $\{\Phi_P^i\}$ is the set of basis functions for $\bigcup S_h^0(\Omega_j)$. These functions correspond to the usual nodal basis of functions which are one on one of the nodes interior to some subdomain and zero on all of the remaining nodes.

2. $\{\Phi_E^i\}$ is the set of basis functions corresponding to the variables which lie on the edges of the subregions (excluding the corners). This basis consists of the usual nodal basis functions which are one on one of the edge nodes and zero on all of the remaining nodes of Ω^h .

3. $\{\Phi_V^i\}$ is a basis for the functions which are linear on the edges of the subregions. The function Φ_V^i is one on v_i , zero on all other vertices v_j with $j \neq i$, zero on all of the nodes which are interior to any of the subregions, and extended linearly along the edges of the subdomains.

It is easily seen that the above collection of functions give rise to a basis for $S_h^0(\Omega)$. As usual, we decompose functions in $S_h^0(\Omega)$ in terms of linear combinations of these basis functions and an arbitrary function in the subspace is represented by a vector of its coefficients. We order these vectors as follows:

$$\alpha = \begin{pmatrix} v_V \\ v_E \\ v_P \end{pmatrix}$$

where v_P , v_E , and v_V represent coefficients for basis functions of type 1, 2, and 3 respectively. In terms of block matrices the system corresponding to B is then

$$(5.1) \quad \begin{pmatrix} B_{VV} & B_{VE} & B_{VP} \\ B_{VE}^t & B_{EE} & B_{EP} \\ B_{VP}^t & B_{EP}^t & B_{PP} \end{pmatrix} \begin{pmatrix} v_V \\ v_E \\ v_P \end{pmatrix} = \begin{pmatrix} b_V \\ b_E \\ b_P \end{pmatrix}.$$

The first step of Algorithm DD1 corresponds to computing the solution of $B_{PP}^{-1}b_P$. Using this solution to calculate the data for the inversions of Steps 2 and 3 of Algorithm DD1 corresponds to eliminating two blocks in the third column of (5.1) to obtain

$$(5.2) \quad \begin{pmatrix} B_{VV} - B_{VP}B_{PP}^{-1}B_{VP}^t & B_{VE} - B_{VP}B_{PP}^{-1}B_{EP}^t & 0 \\ B_{VE}^t - B_{EP}B_{PP}^{-1}B_{VP}^t & B_{EE} - B_{EP}B_{PP}^{-1}B_{EP}^t & 0 \\ B_{VP}^t & B_{EP}^t & B_{PP} \end{pmatrix} \begin{pmatrix} v_V \\ v_E \\ v_P \end{pmatrix} = \begin{pmatrix} b_V - B_{VP}B_{PP}^{-1}b_P \\ b_E - B_{EP}B_{PP}^{-1}b_P \\ b_P \end{pmatrix}.$$

We note that the inversions of Steps 2 and 3 are problems which can be solved independently and involve the edge and vertex basis functions respectively. This fact means that the two blocks $B_{VE} - B_{VP}B_{PP}^{-1}B_{EP}^t$ and $B_{VE}^t - B_{EP}B_{PP}^{-1}B_{VP}^t$ in matrix (5.2) must be identically zero. Furthermore, the upper two diagonal blocks of (5.2) must correspond to

$$M_{kl} = \sum_{\Gamma_{ij}} \alpha_{ij} (\Phi_V^k(v_i) - \Phi_V^k(v_j)) (\Phi_V^l(v_i) - \Phi_V^l(v_j))$$

and

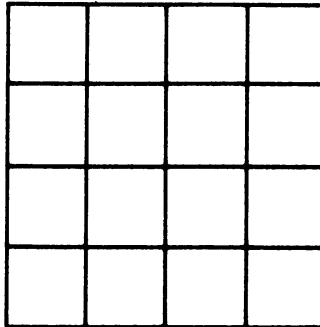
$$N_{kl} = \sum_{\Gamma_{ij}} \alpha_{ij} \left\langle \tilde{l}_0^{1/2} \Phi_E^k, \Phi_E^l \right\rangle_{\Gamma_{ij}},$$

respectively. Steps 4 and 5 of the algorithm correspond to backsolving (5.2) once the values of v_V and v_E are known.

6. Numerical Experiments. In this section, we shall present some results of numerical experiments which illustrate the convergence properties of the preconditioning algorithm using DD1 as a preconditioner discussed in Section 2, when used in conjunction with the conjugate gradient method. To this end we shall report a number of parameters which measure or effect the convergence of the scheme. We shall, for example, compute the condition number K of the preconditioned system. In some examples, we shall also report n , the number of iterations required to reduce the matrix norm $(Ax \cdot x)^{1/2}$ of the error $E_n = U - U_n$ below an indicated tolerance. Here U is a randomly generated solution of the matrix equations normalized so that $-1 \leq U \leq 1$ and U_n is the approximation to U obtained using n steps of the iterative algorithm.

The examples were chosen to illustrate the effectiveness of the algorithm on problems with both smooth and discontinuous coefficients on domains with different geometries. In all of these examples subspaces $S_h^0(\Omega)$ of piecewise linear functions defined on a quasi-uniform mesh of size h were used and the algorithm was applied to solve the finite element equations approximating the solution of an elliptic problem of the form

$$Lu = f \quad \text{in } \Omega, \quad u = 0 \quad \text{on } \partial\Omega.$$

FIGURE 6.1. *Subdivision of the square.*

The procedure discussed in Section 4 for choosing the coefficients of the preconditioning form and solving the related subproblems (in particular problems (2.4), (2.6), and (2.8) by fast methods) was used throughout this section.

Example 1. For our first example we take $L = -\Delta$, the Laplace operator (i.e., $a_{11} = a_{22} = 1$ and $a_{12} = a_{21} = 0$), Ω the unit square and $S_h^0(\Omega)$ the piecewise linear functions on a regular mesh of size h which vanish on $\partial\Omega$. Note that although, in this very simple case, the resulting equations may be fast solved on a serial machine by any of the methods discussed in Section 4, the algorithm used would be particularly appealing for a machine with parallel architecture. We will also use this example as a benchmark for the more complicated examples to follow. We subdivide the domain Ω into sixteen subregions as indicated in Figure 6.1.

TABLE 6.1. *Iterative convergence for Example 1.*

Iteration	A-error	A-error Average Reduction	Max-error	Max-error Average Reduction
1	9.5×10^{-2}	.095	6.6×10^{-1}	.66
2	5.5×10^{-2}	.23	5.4×10^{-1}	.74
3	2.4×10^{-2}	.29	1.8×10^{-1}	.56
4	4.8×10^{-3}	.26	4.2×10^{-2}	.45
5	1.2×10^{-3}	.26	9.9×10^{-3}	.40
6	6.7×10^{-4}	.30	9.6×10^{-3}	.46
7	3.6×10^{-4}	.32	3.2×10^{-3}	.44
8	9.5×10^{-5}	.31	7.5×10^{-4}	.41
9	1.6×10^{-5}	.29	1.2×10^{-4}	.37
10	5.0×10^{-6}	.30	5.6×10^{-5}	.38
11	3.3×10^{-6}	.32	4.2×10^{-5}	.40

Table 6.1 illustrates the iterative reduction rates for Example 1 when $h = 1/32$. The table lists the total reduction and average reduction rate as a function of the number of iterations in the matrix norm $(Ax \cdot x)^{1/2}$ and the maximum norm. These reductions are normalized so that the initial error is unity. We see, for example, that a reduction of .0001 in the A norm (resp. maximum norm) requires only 8 (resp. 10) iterations.

$\mu=300$	$\mu=0.0001$	$\mu=31400$	$\mu=5$
$\mu=0.05$	$\mu=8$	$\mu=0.07$	$\mu=2700$
$\mu=10^6$	$\mu=0.1$	$\mu=200$	$\mu=9$
$\mu=1$	$\mu=8000$	$\mu=4$	$\mu=140000$

FIGURE 6.2. *The coefficients for Example 2.*

To more fully illustrate the convergence behavior of the method on this problem we consider Table 6.2 which gives the condition number and theoretical reduction ρ^{**} for Example 1 as a function of the mesh size h . We note that the theoretical reduction gives a pessimistic bound on the worst-case convergence in the A norm. For example, the actual reduction rate given in Table 6.1 for 11 iterations was .32 which is considerably better than the theoretical rate of .45 given in Table 6.2 for $h = 1/32$. We also compare the condition number to the function $(\log_2 1/h)^2 / 3.5$ and hence demonstrate the log-squared growth in the condition number which suggests that Theorem 1 is sharp.

TABLE 6.2. *Condition number and theoretical reduction for Example 1.*

h	K	$(\log_2 1/h)^2 / 3.5$	ρ^{**}
1/8	3.0	2.6	.27
1/16	4.5	4.6	.36
1/32	7.0	7.1	.45
1/64	10.3	10.3	.52
1/128	14.0	14.0	.58
1/256	18.6	18.3	.62

Example 2. In this example, Ω is the unit square and the subdomains were taken as in Example 1 (see Figure 6.1). The operator L is taken to have coefficients which have jump discontinuities across the subdomain boundaries. More specifically, we take $a_{11} = a_{22} = \mu$ and $a_{12} = a_{21} = 0$, where μ is the randomly chosen piecewise constant function on the subdomains as indicated in Figure 6.2. Table 6.3 gives the results for the condition number of the preconditioned system and the theoretical reduction factors for this example as a function of h . Note that the results differ negligibly from those given for the Laplacian in Table 6.2. We remark that similar results were obtained in tests with other randomly chosen coefficients. This indicates that the iterative method DD1 will be extremely effective on interface

**It is well known (cf. [17]) that the error for preconditioned conjugate gradient iteration satisfies $(AE_n \cdot E_n) \leq 4\rho^{2n}(AE_0 \cdot E_0)$, where the reduction factor ρ is given by $\rho \equiv (\sqrt{K} - 1)/(\sqrt{K} + 1)$.

TABLE 6.3. *Condition number and theoretical reduction for Example 2.*

h	K	$(\log_2 1/h)^2 / 3.2$	ρ^\dagger
1/8	3.0	2.8	.27
1/16	5.0	5.0	.38
1/32	7.7	7.8	.47
1/64	11.2	11.3	.54
1/128	15.2	15.3	.59

problems, even when the coefficients change drastically across interfaces, as long as the subdomain boundaries align with the interface boundaries.

Example 3. Here we take L to be an operator with smoothly varying coefficients. The region Ω and the subdomains are taken exactly as in Example 1. The coefficients are defined by

$$(6.1) \quad \begin{aligned} a_{11} &= 1 + 4(x^2 + y^2), & a_{12} &= 3xy, \\ a_{21} &= 3xy, & a_{22} &= 1 + 11(x^2 + y^2). \end{aligned}$$

This example illustrates that the introduction of more subdomains allows the preconditioner to more closely model the differential operator. Table 6.4 gives convergence results for the above problem as a function of n_r , the number of subdomains used. The coefficients defining the preconditioner were chosen as in Section 4. More precisely, we set $q_k = \sqrt{q_0 q_1}$, where q_0 and q_1 are as in (4.3) and the point \bar{x} (see (4.3)) is chosen as the center of the subdomain. All computations in Table 6.4 were made for $h = 1/64$.

Note that if the Laplace operator on the original domain was used as a preconditioner for the variable coefficient problem (6.1), then the condition number would be larger than 55. In contrast, the results given in Table 6.4 show considerable improvement even when relatively few subdomains are used.

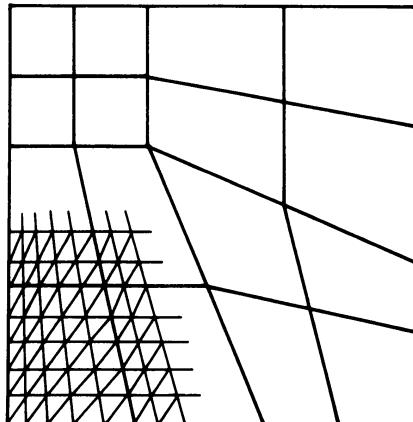
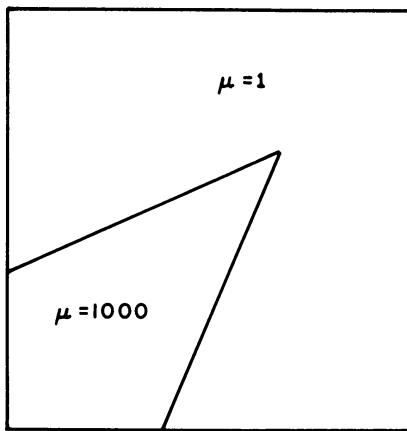
Table 6.4 also illustrates the fact that the theoretical reduction (computed from the eigenvalues of the discrete system) provides a useful bound for the actual rate of convergence. We finally included n , the number of iterations required to reduce the matrix norm $(Ax \cdot x)^{1/2}$ of the error $E_n = U - U_n$ by a factor of .0001.

TABLE 6.4. *Convergence results for Example 3.*

n_r	K	ρ^\dagger	Observed Reduction	n
4	42.3	.73	.56	17
16	17.5	.61	.51	14
64	11.1	.54	.45	12
256	7.4	.46	.43	11

Example 4. In this example, we consider an interface problem where the interface separates two domains with irregular geometries. The domain Ω is again the unit square subdivided into sixteen subdomains as illustrated in Figure 6.3. The space $S_h^0(\Omega)$ is taken to be piecewise linear functions defined on an irregular mesh. A

† See footnote **.

FIGURE 6.3. *The irregular geometry of Example 4.*FIGURE 6.4. *The coefficients of Example 4.*

portion of this mesh is illustrated by the fine triangulation in Figure 6.3. Again the coefficients of L are piecewise constant functions defined by $a_{11} = a_{22} = \mu$ and $a_{12} = a_{21} = 0$ where μ is given by Figure 6.4.

Results for this problem are given in Table 6.5. A comparison with Table 6.2 indicates that the irregular geometry of this example only increased the condition number by about 2.5. This results in less than a factor of two increase in the number of iterations required for a given accuracy. We again remark that fast methods were used to solve the subproblems (2.4), (2.6) and (2.8) required for the preconditioner.

TABLE 6.5. *Convergence results for Example 4.*

h	K	ρ^\dagger	Observed Reduction	n
1/8	5.6	.41	.32	9
1/16	10.8	.53	.45	13
1/32	17.6	.62	.51	15
1/64	25.4	.67	.55	16

† See footnote **.

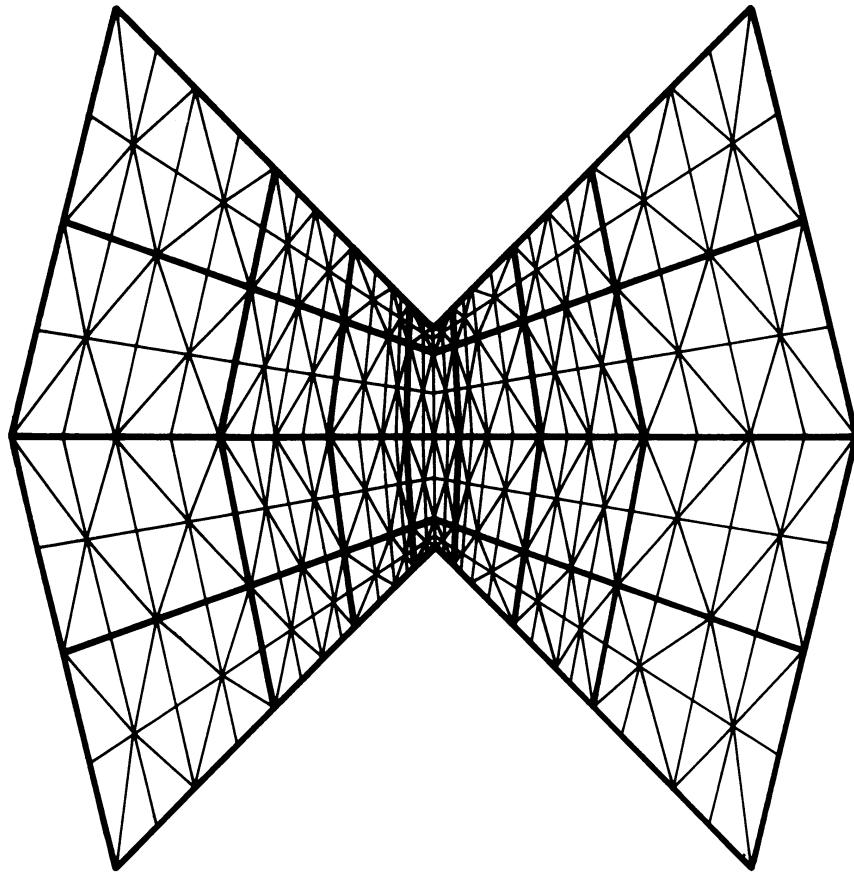


FIGURE 6.5. *The mesh and subdomain structure for Example 5.*

Example 5. In this example, we illustrate Algorithm DD1 applied to the solution of a problem on a polygonal domain with nonconvex corners. The mesh and subdomain structure were chosen as illustrated in Figure 6.5. Note the mild refinement near the nonconvex corners of the domain. For the operator L we use the Laplacian as in Example 1. The results for this case are given in Table 6.6. This example illustrates some of the power and flexibility of this algorithm which will be more fully developed in later papers.

TABLE 6.6. *Convergence results for Example 5.*

Number of Unknowns	K	ρ^\dagger	Observed Reduction	n
405	8.54	.49	.45	12
1705	14.4	.58	.50	14
6993	20.6	.64	.55	15

Example 6. As a final example, we compare the preconditioner DD1 with the somewhat simpler preconditioner discussed in Remark 2.6. In particular, we replace the global difference equation (2.7) by a weighted identity (2.9) on the coarse mesh points. For this example, we use the Laplace operator on the square as in Example 1. This example illustrates the effect that the diameter of the subdomains has on

† See footnote **.

the actual condition number observed in practice. Table 6.7 compares the condition numbers for DD1 and the preconditioner of Remark 2.6 as a function of d . Observe the clear superiority of DD1 in applications with many subdivisions.

TABLE 6.7. *Comparison of DD1 and the preconditioner of Remark 2.6.*

d	$K(\text{DD1})$	K (Remark 2.6)	h
1/2	6.3	6.3	1/16
1/4	7.0	10.5	1/32
1/8	7.5	26.6	1/64
1/16	7.5	96.9	1/128

Department of Mathematics
Cornell University
Ithaca, New York 14853

Department of Applied Mathematics
Brookhaven National Laboratory
Upton, New York 11973

Department of Mathematics
Cornell University
Ithaca, New York 14853

1. P. E. BJØRSTAD & O. B. WIDLUND, "Solving elliptic problems on regions partitioned into substructures," *Elliptic Problem Solvers II* (G. Birkhoff and A. Schoenstadt, eds.), Academic Press, New York, 1984, pp. 245–256.
2. P. E. BJØRSTAD & O. B. WIDLUND, "Iterative methods for the solution of elliptic problems on regions partitioned into substructures." (Preprint.)
3. J. H. BRAMBLE, "A second order finite difference analogue of the first biharmonic boundary value problem," *Numer. Math.*, v. 9, 1966, pp. 236–249.
4. J. H. BRAMBLE, J. E. PASCIAK & A. H. SCHATZ, "An iterative method for elliptic problems on regions partitioned into substructures," *Math. Comp.*, v. 46, 1986, pp. 361–369.
5. B. L. BUZBEE & F. W. DORR, "The direct solution of the biharmonic equation on rectangular regions and the Poisson equation on irregular regions," *SIAM J. Numer. Anal.*, v. 11, 1974, pp. 753–763.
6. B. L. BUZBEE, F. W. DORR, J. A. GEORGE & G. H. GOLUB, "The direct solution of the discrete Poisson equation on irregular regions," *SIAM J. Numer. Anal.*, v. 8, 1971, pp. 722–736.
7. Q. V. DIHN, R. GLOWINSKI & J. PÉRIAUX, "Solving elliptic problems by domain decomposition methods," *Elliptic Problem Solvers II* (G. Birkhoff and A. Schoenstadt, eds.), Academic Press, New York, 1984, pp. 395–426.
8. S. C. EISENSTADT, M. C. GURSKY, M. H. SCHULTZ & A. H. SHERMAN, *Yale Sparse Matrix Package, I. The Symmetric Codes*, Dept. of Computer Science Report No. 112, Yale University.
9. A. GEORGE & J. W. H. LIU, *User Guide for SPARSPAK*, Dept. of Computer Science Report No. CS-78-30, Waterloo University.
10. G. H. GOLUB & D. MEYERS, "The use of preconditioning over irregular regions," *Proc. Sixth Internat. Conf. on Computing Methods in Science and Engineering*. (Preprint.)
11. P. GRISVARD, *Elliptic Problems in Non Smooth Domains*, Pitman, Boston, 1985.
12. M. R. HESTENES, *The Conjugate Gradient Method for Solving Linear Systems*, Proc. Sympos. Appl. Math., vol. 6, Amer. Math. Soc., McGraw-Hill, 1956, pp. 83–102.
13. S. G. KREIN & Y. I. PETUNIN, "Scales of Banach spaces," *Russian Math. Surveys*, v. 21, 1966, pp. 85–160.
14. J. L. LIONS & E. MAGENES, *Problèmes aux Limites non Homogènes et Applications*, Dunod, Paris, 1968.

15. D. G. LUENBERGER, *Introduction to Linear and Nonlinear Programming*, Addison-Wesley, Reading, Mass., 1973.
16. J. NEČAS, *Les Méthodes Directes en Théorie des Équations Elliptiques*, Academia, Prague, 1967.
17. W. M. PATTERSON, 3RD, *Iterative Methods for the Solution of a Linear Operator Equation in Hilbert Space—A Survey*, Lecture Notes in Math., Vol. 394, Springer-Verlag, New York, 1974.
18. P. N. SWARZTRAUBER, "The methods of cyclic reduction, Fourier analysis and the FACR algorithm for the discrete solution of Poisson's equation on a rectangle," *SIAM Rev.*, v. 19, 1977, pp. 490–501.
19. V. THOMÉE, *Galerkin Finite Element Methods for Parabolic Problems*, Lecture Notes in Math., Vol. 1054, Springer-Verlag, New York, 1984.

*4.2. THE CONSTRUCTION OF PRECONDITIONERS FOR ELLIPTIC
PROBLEMS BY SUBSTRUCTURING. II*

**4.2 The construction of preconditioners for elliptic problems by
substructuring. II**

The construction of preconditioners for elliptic problems by substructuring. II[20]

The Construction of Preconditioners for Elliptic Problems by Substructuring. II

By J. H. Bramble*, J. E. Pasciak* and A. H. Schatz*

Abstract. We give a method for constructing preconditioners for the discrete systems arising in the approximation of solutions of elliptic boundary value problems. These preconditioners are based on domain decomposition techniques and lead to algorithms which are well suited for parallel computing environments. The method presented in this paper leads to a preconditioned system with condition number proportional to d/h where d is the subdomain size and h is the mesh size. These techniques are applied to singularly perturbed problems and problems in three dimensions. The results of numerical experiments illustrating the performance of the method on problems in two and three dimensions are given.

1. Introduction. The aim of this series of papers is to propose and analyze methods for efficiently solving the equations resulting from finite element discretizations of second-order elliptic boundary value problems on general domains in R^2 and R^3 . In particular we shall be concerned with constructing easily invertible and “effective” preconditioners for the resulting system of discrete equations which can be used in a preconditioned iterative algorithm to achieve a rapid solution method. The methods to be presented are well suited to parallel computing architectures.

For $N = 2$ or $N = 3$, let Ω be a bounded domain in R^N with a piecewise smooth boundary $\partial\Omega$. As a model problem for a second-order uniformly elliptic equation we shall consider the Dirichlet problem

$$(1.1) \quad \begin{aligned} Lu &= f && \text{in } \Omega, \\ u &= 0 && \text{on } \partial\Omega, \end{aligned}$$

where

$$(1.2) \quad Lv = - \sum_{i,j=1}^N \frac{\partial}{\partial x_i} \left(a_{ij} \frac{\partial v}{\partial x_j} \right) + av,$$

with a_{ij} symmetric, uniformly positive definite and bounded above on Ω . For ease of exposition, we assume that either $a \equiv 0$ or a is bounded above and below by

Received February 21, 1986.

1980 *Mathematics Subject Classification* (1985 Revision). Primary 65N30; Secondary 65F10.

*This manuscript has been authored under contract number DE-AC02-76CH00016 with the U.S. Department of Energy. Accordingly, the U.S. Government retains a non-exclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. This work was also supported in part under the National Science Foundation Grant No. DMS84-05352 and under the Air Force Office of Scientific Research, Contract No. ISSA86-0026.

©1987 American Mathematical Society
0025-5718/87 \$1.00 + \$.25 per page

positive constants. The generalized Dirichlet form is given by

$$(1.3) \quad A(v, \phi) = \sum_{i,j=1}^N \int_{\Omega} a_{ij} \frac{\partial v}{\partial x_i} \frac{\partial \phi}{\partial x_j} dx + \int_{\Omega} av\phi dx,$$

which is defined for all v and ϕ in the Sobolev space $H^1(\Omega)$ (the space of distributions with square-integrable first derivatives). The $L^2(\Omega)$ inner product is denoted

$$(v, \phi)_{\Omega} = \int_{\Omega} v\phi dx.$$

The subspace $H_0^1(\Omega)$ is the completion of the smooth functions with support in Ω with respect to the norm in $H^1(\Omega)$. The weak formulation of the problem defined by (1.1) is: Find $u \in H_0^1(\Omega)$ such that

$$(1.4) \quad A(u, \phi) = (f, \phi)_{\Omega}$$

for all $\phi \in H_0^1(\Omega)$. This leads immediately to the standard Galerkin approximation. Let $S_h^0(\Omega)$ be a finite-dimensional subspace of $H_0^1(\Omega)$. The Galerkin approximation is defined as the solution of the following problem: Find $U \in S_h^0(\Omega)$ such that

$$(1.5) \quad A(U, \Phi) = (f, \Phi)_{\Omega}$$

for all $\Phi \in S_h^0(\Omega)$.

We shall also be interested in solving (1.5) when the form A of (1.3) corresponds to the singularly perturbed operator

$$(1.6) \quad \tilde{L}v = v + \varepsilon Lv,$$

where L was defined by (1.2) and ε is a possibly small constant which in some applications depends upon h . The A form corresponding to (1.6) is then given by

$$(1.7) \quad A(v, \phi) = \varepsilon \left\{ \sum_{i,j=1}^N \int_{\Omega} a_{ij} \frac{\partial v}{\partial x_i} \frac{\partial \phi}{\partial x_j} dx + \int_{\Omega} av\phi dx \right\} + (v, \phi)_{\Omega}.$$

Singularly perturbed problems arise, for example, in time-stepping methods for the numerical approximation of parabolic problems.

Now it is easy to see that if ε is bounded away from zero, then any preconditioner for (1.5) gives a preconditioner for (1.7). Furthermore, if ε is of order h^2 , then the quadratic form $A(v, v)$ restricted to the subspace $S_h^0(\Omega)$ is equivalent to $(v, v)_{\Omega}$ and no preconditioner is necessary. We shall provide a preconditioner for (1.7) which has conditioning properties similar to those of the preconditioner developed for (1.5) independent of ε .

As illustrated in Part 1 [3], the preconditioning problem can be reduced to the problem of defining an appropriate form B on $S_h^0(\Omega) \times S_h^0(\Omega)$ satisfying the following criterion. Firstly, the problem of finding $W \in S_h^0(\Omega)$, given g , satisfying

$$(1.8) \quad B(W, \Phi) = (g, \Phi)_{\Omega} \quad \text{for all } \Phi \in S_h^0(\Omega)$$

should be easier to obtain than the solution of (1.5). Secondly, the forms B and A should be comparable in the sense that there are positive constants λ_0 and λ_1 satisfying

$$(1.9) \quad \lambda_0 B(V, V) \leq A(V, V) \leq \lambda_1 B(V, V) \quad \text{for all } V \in S_h^0(\Omega)$$

with λ_1/λ_0 “not too large.”

It should be noted that it is generally not possible to develop an effective preconditioner for (1.7) directly from a preconditioner for (1.5). If B is a preconditioner for (1.5), then a natural choice of a preconditioner for (1.7) would be the form given by

$$(1.10) \quad \varepsilon B(u, v) + (u, v)_\Omega.$$

Unfortunately, the problem corresponding to (1.8) using the form (1.10) cannot, in general, be efficiently solved.

In this paper we shall develop a particularly simple method for defining preconditioners by domain decomposition. As is typical with domain decomposition techniques, the given domain Ω is broken into a number of subdomains $\{\Omega_i\}$. Our preconditioner is defined so that the calculation of the solution of (1.8) involves solving in parallel related Galerkin equations on the subregions and some interconnecting equations. For the method to be developed, the number of unknowns involved in the interconnecting equations will be at most equal to the number of subdomains.

Other papers providing iterative methods involving domain decomposition for the solution of elliptic problems have appeared in the literature [1]–[8]. The earliest papers involved splitting the domain into subdomains without interior corner points [1], [2], [4]–[6], [8]. These methods became inefficient when many long thin subdomains were used. Consequently, it became natural to develop decomposition methods which use quasi-uniform subregions. In Part I, we defined and analyzed such a method for two-dimensional problems. That method was shown to have a condition number for the preconditioned system which was bounded by $c(1 + \log(d/h))^2$ (here d and h correspond, respectively, to the diameter of the subregions and the discretization size of the mesh).

The preconditioner defined and analyzed in this paper has the following advantages over that defined in Part I. Firstly, it is somewhat simpler, both conceptually and computationally. Secondly, it extends in a straightforward manner to three-dimensional problems. Thirdly, it applies to singularly perturbed systems without deterioration in the iterative convergence rates.

On the negative side, the preconditioner defined in this paper shows a somewhat faster asymptotic growth of the condition number for the preconditioned system than that of the Part I preconditioner. We will show that the condition number for the new method is bounded by cd/h in contrast to the $(1 + \log(d/h))^2$ growth for the preconditioned system of Part I. This is a reasonable growth for many rather large three-dimensional problems when d and h are judiciously chosen.

An important aspect of this paper involves the introduction of certain constants or ‘average values’ associated with discrete functions on the subdomains as part of the definition of the preconditioner B . A technique for computing these average values is presented. A future part in this series of papers will provide a three-dimensional preconditioner employing this averaging technique with a $(1 + \log(d/h))^2$ condition number growth for the preconditioned system.

The outline of the remainder of the paper is as follows. In Section 2 we describe the domain decomposition preconditioners and prove estimates for the growth of the condition numbers for the preconditioned system. In Section 3 we show how to compute the solution to (1.8). Numerical examples of the preconditioner applied to problems in two and three dimensions are given in Section 4.

We shall also let c and C , with or without subscript, denote generic positive constants. These constants will always be independent of the mesh and subdomain parameters h and d (see Section 2).

2. The Construction and Analysis of the Preconditioner. We will describe the preconditioner in this section and prove an estimate for the condition number of the preconditioned system. We start by giving some hypothesis on the domain and subdomain partitioning and the associated finite element subspaces.

For the sake of simplicity of exposition we shall proceed with the discussion only for the special case of polyhedral domains and piecewise linear approximations. Many generalities are possible and will be discussed in later papers.

More precisely we shall begin with the following assumptions with regard to Ω .

- (A.1) Ω is a polyhedral domain in R^2 or R^3 , which for each h , $0 < h < 1$, a parameter, has been given a triangulation Ω^h of maximal size h . That is, $\Omega^h = \bigcup_{j=1}^{m(h)} \tau_j^h$, where each τ_j^h is a simplex which is contained in a ball of radius h .

Any union of simplexes of Ω^h will be called a mesh subdomain, and the vertices of the simplexes in Ω^h will be denoted by x_i ordered in some fashion. We shall partition the domain Ω into a number of mesh subdomains $\{\Omega_k\}$.

- (A.2) We assume that the triangulation is quasi-uniform near the boundaries of the subdomains, i.e., if $\tau^h \in \Omega^h$ is a simplex such that $\tau^h \cap \partial\Omega_k \neq \emptyset$, then τ^h contains a ball of radius ch where c is independent of h .
- (A.3) The Ω_k are quasi-uniform of size d . This means there exists a positive constant c_1 which is independent of d and h such that each Ω_k contains a ball of radius $c_1 d$ and is contained in a ball of radius d . The number of domains n_d is proportional to d^{-N} .
- (A.4) Each Ω_k is uniformly star-shaped with respect to a point. This means that for each Ω_k there is a point \hat{x}_k and a constant $c_2 > 0$, independent of d and h , such that $(x - \hat{x}_k) \cdot n(x) \geq c_2 d$ for all $x \in \partial\Omega_k$. Here, $n(x)$ denotes the outward unit normal to $\partial\Omega_k$ at x .
- (A.5) Let $\tilde{\Omega}_k$ be the scaled domain defined by

$$\tilde{\Omega}_k \equiv \{x | dx \in \Omega_k\}.$$

We assume that $\tilde{\Omega}_k$ has a Lipschitz continuous boundary with Lipschitz constants which are independent of d .

Remark 2.1. Assumption (A.5) is a weak regularity hypothesis for the boundary of Ω_k . It guarantees that a Poincaré inequality of the form

$$(2.1) \quad \|V\|_{\Omega_k}^2 \leq Cd^2 D_k(V, V)$$

holds for functions V with zero mean value on Ω_k with a constant C independent of d and k . Here $D_k(\cdot, \cdot)$ denotes the Dirichlet inner product on Ω_k .

Remark 2.2. We note that Assumption (A.4) implies the inequality

$$(2.2) \quad |u|_{\partial\Omega_k}^2 \leq c\{d^{-1} \|u\|_{\Omega_k}^2 + dD_k(u, u)\}.$$

For each h , let $S_h(\Omega)$ be the space of continuous piecewise linear functions defined relative to the triangulation Ω^h and $S_h^0(\Omega)$ be the subspace of $S_h(\Omega)$ consisting of those functions which vanish on $\partial\Omega$. $S_h^0(\Omega_k)$ will denote the subspace of $S_h^0(\Omega)$ of functions whose supports are contained in $\tilde{\Omega}_k$ (in particular, they vanish on $\partial\Omega_k$

and outside $\bar{\Omega}_k$). Let Γ denote $\bigcup_k \partial\Omega_k$ and let $S_h(\Gamma)$ denote the functions which are restrictions to Γ of functions in $S_h^0(\Omega)$.

We shall need some additional notation. The $L^2(\partial\Omega_k)$ inner product shall be denoted

$$\langle v, w \rangle_{\partial\Omega_k} = \int_{\partial\Omega_k} vw \, ds$$

with corresponding norm

$$|v|_{\partial\Omega_k}^2 = \langle v, v \rangle_{\partial\Omega_k},$$

where ds is an element of arc length or surface area of $\partial\Omega_k$. The analogous discrete inner product is given by

$$\langle v, w \rangle_{\partial\Omega_k, h} = h^{N-1} \sum_{x_i \in \partial\Omega_k} v(x_i)w(x_i)$$

with corresponding discrete norm

$$|v|_{\partial\Omega_k, h}^2 = \langle v, v \rangle_{\partial\Omega_k, h}.$$

It follows from (A.2) that

$$(2.3) \quad c |v|_{\partial\Omega_k} \leq |v|_{\partial\Omega_k, h} \leq C |v|_{\partial\Omega_k}$$

holds for functions $v \in S_h(\Gamma)$. In addition, we have the following lemma.

LEMMA 2.1. *If $v \in S_h^0(\Omega)$ and vanishes at all interior nodes of Ω_k then .*

$$(2.4) \quad c_1 h |v|_{\partial\Omega_k, h}^2 \leq \|v\|_{\Omega_k}^2 \leq C_1 h |v|_{\partial\Omega_k, h}^2$$

and

$$(2.5) \quad c_1 h^{-1} |v|_{\partial\Omega_k, h}^2 \leq D_k(v, v) \leq C_1 h^{-1} |v|_{\partial\Omega_k, h}^2.$$

We next construct the bilinear form B corresponding to our preconditioner. We first introduce another form $\tilde{A}(\cdot, \cdot)$ on $S_h^0(\Omega)$. If A is given by (1.3), we define

$$\tilde{A}_k(v, \phi) = \sum_{i,j=1}^N \int_{\Omega_k} a_{ij}^k \frac{\partial v}{\partial x_i} \frac{\partial \phi}{\partial x_j} \, dx + \int_{\Omega_k} a^k v \phi \, dx.$$

Alternatively, if A is given by (1.7) then we define

$$(2.6) \quad \tilde{A}_k(v, \phi) = \varepsilon \left\{ \sum_{i,j=1}^N \int_{\Omega_k} a_{ij}^k \frac{\partial v}{\partial x_i} \frac{\partial \phi}{\partial x_j} \, dx + \int_{\Omega_k} a^k v \phi \, dx \right\} + \int_{\Omega_k} b^k v \phi \, dx.$$

We then define

$$(2.7) \quad \tilde{A}(U, V) = \sum_k \tilde{A}_k(U, V).$$

Here $a^k = 0$ if $a = 0$ or $a^k \geq c > 0$. Furthermore $b^k \geq c > 0$. These functions are piecewise smooth (possibly discontinuous) for each k . Finally, $a_{ij}^k(x)$ for $i, j = 1, \dots, N$ is a piecewise smooth (possibly discontinuous) uniformly positive definite matrix. The reason for the form of \tilde{A} was discussed in Part I ([3], Section 4). Basically, it allows for greater flexibility in the definition of the preconditioner and, for example, the use of constant coefficient fast solvers (even when L has variable coefficients).

We note that

$$(2.8) \quad C_0 \tilde{A}(U, U) \leq A(U, U) \leq C_1 \tilde{A}(U, U) \quad \text{for all } U \in S_h^0(\Omega)$$

holds. Thus, the problem of finding a preconditioner for A is the same as finding one for \tilde{A} .

We next decompose functions in $S_h^0(\Omega)$ as follows: Write $W = W_P + W_H$ where $W_P \in S_h^0(\Omega_1) \oplus \cdots \oplus S_h^0(\Omega_{n_d})$ and W_P restricted to Ω_k satisfies

$$\tilde{A}_k(W_P, \Phi) = \tilde{A}_k(W, \Phi) \quad \text{for all } \Phi \in S_h^0(\Omega_k)$$

for each k . Notice that W_P is determined on Ω_k by the values of W on Ω_k and that

$$(2.9) \quad \tilde{A}_k(W_H, \Phi) = 0 \quad \text{for all } \Phi \in S_h^0(\Omega_k).$$

Thus on each Ω_k , W is decomposed into a function W_P which vanishes on $\partial\Omega_k$ and a function W_H which satisfies the above homogeneous equation and has the same boundary values as W on $\partial\Omega_k$. We shall refer to such a function W_H as “discrete \tilde{A}_k -harmonic”. The subspace of discrete \tilde{A}_k -harmonic functions shall be denoted by $H(\Omega_k)$.

We note that the above decomposition is orthogonal in the \tilde{A} inner product and hence

$$(2.10) \quad \tilde{A}(W, W) = \tilde{A}(W_P, W_P) + \tilde{A}(W_H, W_H).$$

We shall define the preconditioning form B by replacing the $\tilde{A}(W_H, W_H)$ term in (2.10).

Note that a discrete \tilde{A}_k -harmonic function is completely determined by its values on the boundary. Accordingly, the form $\tilde{A}(W_H, W_H)$ can be replaced by a form which only involves the boundary values. The particular choice of the boundary form will depend on whether we are considering (1.3) or the singularly perturbed case (1.7).

Remark 2.3. It seems reasonable to consider replacing the $\tilde{A}(W_H, W_H)$ by the identity (or a weighted identity) on the subdomain boundary. This works reasonably well if A is given by (1.3), d is not too large, and the coefficients of A are smooth. The replacement forms to be described work better in more general situations.

We first consider the case when A is given by (1.3). To understand the motivation for the form to be defined, it is instructive to consider the case when $a = 0$. We would like to replace the form \tilde{A}_k (restricted to discrete harmonic functions) and define the replacement for $\tilde{A}(W_H, W_H)$ by summation. Note that if $a^k = 0$ then \tilde{A}_k is indefinite, and so its replacement should also be indefinite. Let α_k be a constant which will be chosen later; then

$$(2.11) \quad \tilde{A}_k(W_H, W_H) = \tilde{A}_k(W_H - \alpha_k, W_H - \alpha_k) \leq \tilde{A}_k(W, W),$$

where $W \in S_h(\Omega)|_{\Omega_k}$ is defined by the function which equals $W_H - \alpha_k$ on $\partial\Omega_k$ and vanishes on all interior nodes of Ω_k . Now

$$(2.12) \quad \tilde{A}_k(W, W) \leq c \tilde{a}_k D_k(W, W),$$

where \tilde{a}_k is (for example) the smallest eigenvalue of the matrix $\{a_{ij}^k(x)\}_{i,j=1}^N$ for some point $x \in \Omega_k$. It follows from (2.5), (2.11), and (2.12) that

$$(2.13) \quad \tilde{A}_k(W_H, W_H) \leq c \tilde{a}_k h^{-1} |W_H - \alpha_k|_{\partial\Omega_k, h}^2.$$

To make the right-hand side of (2.13) correspond to an indefinite form, we choose α_k to be the discrete mean value of W_H on $\partial\Omega_k$, i.e.,

$$\alpha_k = \bar{W}_H \equiv \frac{\langle W_H, 1 \rangle_{\partial\Omega_k, h}}{\langle 1, 1 \rangle_{\partial\Omega_k, h}}.$$

We replace $\tilde{A}(W_H, W_H)$ with

$$(2.14) \quad Q(W_H, W_H) = h^{-1} \sum_k \tilde{a}_k |W_H - \bar{W}_H|_{\partial\Omega_k, h}^2.$$

For more general A given by (1.3) with $a^k \neq 0$ we use

$$(2.15) \quad \begin{aligned} Q(W_H, W_H) &= \sum_k Q_k(W_H, W_H) \\ &\equiv \sum_k h^{-1} \left\{ (\tilde{a}_k + \bar{a}_k h^2) |W_H - (\bar{W}_H)_k|_{\partial\Omega_k, h}^2 + \bar{a}_k h d^N (\bar{W}_H)_k^2 \right\}. \end{aligned}$$

Finally, when A is given by (1.7) we use

$$(2.16) \quad \begin{aligned} Q(W_H, W_H) &= \sum_k Q_k(W_H, W_H) \\ &\equiv \sum_k h^{-1} \left\{ (\varepsilon \tilde{a}_k + (\bar{b}_k + \varepsilon \bar{a}_k) h^2) |W_H - (\bar{W}_H)_k|_{\partial\Omega_k, h}^2 \right. \\ &\quad \left. + (\bar{b}_k + \varepsilon \bar{a}_k) h d^N (\bar{W}_H)_k^2 \right\}. \end{aligned}$$

The constants \bar{b}_k and \bar{a}_k are defined as the average values of b^k and a^k over Ω_k . As before, it suffices to take \tilde{a}_k to be the minimal eigenvalue of the matrix $\{a_{ij}^k(x)\}$ for some point $x \in \Omega_k$.

We have the following theorem.

THEOREM 1. *Let A be given by (1.3) or (1.7), respectively. Let B be defined on $S_h^0(\Omega) \times S_h^0(\Omega)$ by*

$$(2.17) \quad B(W, W) \equiv \tilde{A}(W_P, W_P) + Q(W_H, W_H),$$

where Q is given by (2.15) or (2.16) respectively. We then have

$$(2.18) \quad \frac{ch}{d} B(W, W) \leq A(W, W) \leq C B(W, W) \quad \text{for all } W \in S_h^0(\Omega),$$

with c and C independent of h and d .

Proof. By (2.7), (2.8) and (2.10) it suffices to prove that

$$(2.19) \quad \frac{ch}{d} Q_k(V, V) \leq \tilde{A}_k(V, V) \leq C Q_k(V, V) \quad \text{for all } V \in H(\Omega_k).$$

We shall first prove the theorem when A is given by (1.3) and $a^k = 0$. In this case, (2.19) reduces to

$$(2.20) \quad \frac{c\tilde{a}_k}{d} |V - \bar{V}_k|_{\partial\Omega_k, h}^2 \leq \tilde{A}_k(V, V) \leq \frac{C\tilde{a}_k}{h} |V - \bar{V}_k|_{\partial\Omega_k, h}^2 \quad \text{for all } V \in H(\Omega_k).$$

The second inequality is just (2.13).

To prove the first inequality, let β_k denote the mean value of V on Ω_k . By (2.3) and the definition of \bar{V}_k , we have

$$|V - \bar{V}_k|_{\partial\Omega_k, h}^2 \leq |V - \beta_k|_{\partial\Omega_k, h}^2 \leq C |V - \beta_k|_{\partial\Omega_k}^2.$$

Applying (2.2), (2.1) and (2.3) gives

$$\begin{aligned} |V - \bar{V}_k|_{\partial\Omega_k, h}^2 &\leq c(d^{-1} \|V - \beta_k\|_{\Omega_k}^2 + dD_k(V - \beta_k, V - \beta_k)) \\ &\leq CdD_k(V - \beta_k, V - \beta_k) = CdD_k(V, V). \end{aligned}$$

Then, by the ellipticity assumptions on the coefficients defining \tilde{A}_k and A , we have

$$(2.21) \quad \tilde{a}_k |V - \bar{V}_k|_{\partial\Omega_k, h}^2 \leq Cd\tilde{A}_k(V, V).$$

Thus we have shown that the theorem holds for the case $a^k = 0$.

We next prove the theorem for the remaining cases. When $a \neq 0$ and A given by (1.3), we set $\bar{b}_k = 0$ and $\varepsilon = 1$. Hence (2.16) defines Q in either case. We first prove the second inequality of (2.19). Let \tilde{V}_k denote the discrete \tilde{A}_k -harmonic extension of \bar{V}_k . Note that in general, \tilde{V}_k is nonconstant even though \bar{V}_k is constant on $\partial\Omega_k$. Evidently,

$$(2.22) \quad \tilde{A}_k(V, V) \leq 2(\tilde{A}_k(V - \tilde{V}_k, V - \tilde{V}_k) + \tilde{A}_k(\tilde{V}_k, \tilde{V}_k)).$$

By the harmonicity of \tilde{V}_k ,

$$(2.23) \quad \tilde{A}_k(\tilde{V}_k, \tilde{V}_k) \leq \tilde{A}_k(\bar{V}_k, \bar{V}_k) \leq c(\bar{b}_k + \varepsilon\bar{a}_k) \|\bar{V}_k\|_{\Omega_k}^2 \leq C(\bar{b}_k + \varepsilon\bar{a}_k)d^N \bar{V}_k^2.$$

By the harmonicity of $V - \tilde{V}_k$,

$$(2.24) \quad \tilde{A}_k(V - \tilde{V}_k, V - \tilde{V}_k) \leq \tilde{A}_k(W, W),$$

where W is the function which equals $V - \tilde{V}_k$ on $\partial\Omega_k$ and vanishes on the interior nodes of Ω_k . The assumptions on the coefficients of the operator and the preconditioner and Lemma 2.1 give

$$\begin{aligned} (2.25) \quad \tilde{A}_k(W, W) &\leq c\{(\bar{b}_k + \varepsilon\bar{a}_k) \|W\|_{\Omega_k}^2 + \varepsilon\tilde{a}_k D_k(W, W)\} \\ &\leq C \left\{ \left(\frac{\varepsilon\tilde{a}_k}{h} + (\bar{b}_k + \varepsilon\bar{a}_k)h \right) |V - \bar{V}_k|_{\partial\Omega_k, h}^2 \right\}. \end{aligned}$$

Combining (2.22) through (2.25) proves the second inequality of (2.19).

We finally prove the first inequality of (2.19). Noting that (2.21) is also valid in the present case, it suffices to show that

$$(2.26) \quad h(\bar{b}_k + \varepsilon\bar{a}_k) \left\{ h |V - \bar{V}_k|_{\partial\Omega_k, h}^2 + d^N \bar{V}_k^2 \right\} \leq Cd\tilde{A}_k(V, V)$$

for all $V \in H(\Omega_k)$. By the arithmetic-geometric mean inequality,

$$(2.27) \quad h |V - \bar{V}_k|_{\partial\Omega_k, h}^2 \leq 2h(|V|_{\partial\Omega_k, h}^2 + |\bar{V}_k|_{\partial\Omega_k, h}^2).$$

Using (2.4), it follows trivially that

$$(2.28) \quad h |V|_{\partial\Omega_k, h}^2 \leq C \|V\|_{\Omega_k}^2 \leq \frac{C}{(\bar{b}_k + \varepsilon\bar{a}_k)} \tilde{A}_k(V, V).$$

Since d is larger than h , a straightforward computation using (A.3) gives

$$h |\bar{V}_k|_{\partial\Omega_k, h}^2 \leq d |\bar{V}_k|_{\partial\Omega_k, h}^2 \leq Cd^N \bar{V}_k^2.$$

Hence it remains to bound $d^N \bar{V}_k^2$. By the definition of \bar{V}_k and the Schwarz inequality,

$$\begin{aligned} d^N \bar{V}_k^2 &\leq cd^N (h/d)^{2N-2} \left(\sum_i V(x_i) \right)^2 \\ &\leq Cdh^{N-1} \sum_i V(x_i)^2 = Cd |V|_{\partial\Omega_k,h}^2, \end{aligned}$$

where the sum over i is taken over the set of nodes x_i on $\partial\Omega_k$. Hence by (2.4),

$$(2.29) \quad d^N \bar{V}_k^2 \leq C \frac{d}{h} \|V\|_{\Omega_k}^2.$$

Combining (2.28) and (2.29) proves (2.26) and hence completes the proof of the theorem. \square

Remark 2.4. The coefficients c and C appearing in the theorem depend on the local (with respect to the subdomains) behavior of the operator and preconditioner. Accordingly, the preconditioner will work well even in situations where there are large jumps in the coefficients defining L as long as these jumps only occur across the subdomain boundaries.

Remark 2.5. There is a fair amount of freedom in weighting the boundary form. For example, the Q form in the case $a = 0$ could have been defined by

$$Q(W_H, W_H) = \gamma^{-1} \sum_k \tilde{a}_k |W_H - \alpha_k|_{\partial\Omega_k,h}^2.$$

Then the condition number for the preconditioned system would remain unchanged as long as $h \leq \gamma \leq d$. The forms (2.15) and (2.16) could be similarly weighted.

3. The Solution of the Preconditioning Problem. In this section we describe an efficient algorithm for solving (1.8). In general, when B is of the form (2.17), we solve first for W_P , then for the values of W_H on Γ , and finally extend W_H to all of Ω .

We now give the details of a three-step algorithm for the solution of (1.8). As already mentioned, the problem of finding the solution W to (1.8) reduces to that of computing W_P and W_H . The first step is to compute W_P . By taking $\Phi \in S_h^0(\Omega_k)$ in (1.8) and using (2.9), we note that

$$(3.1) \quad \tilde{A}_k(W_P, \Phi) = (g, \Phi) \quad \text{for all } \Phi \in S_h^0(\Omega_k).$$

Equation (3.1) shows that W_P can be determined by solving independent discrete Dirichlet problems on the subregions. The second step involves the computation of the values of W_H on Γ . These values are determined as the solution of the following problem:

$$(3.2) \quad Q(W_H, \theta) = (g, \tilde{\theta})_\Omega - \tilde{A}(W_P, \tilde{\theta}) \quad \text{for all } \theta \in S_h(\Gamma).$$

Here $\tilde{\theta}$ denotes any extension of θ in $S_h^0(\Omega)$ and we note that by (3.1), the right-hand side of (3.2) is independent of the extension chosen. The development of an algorithm for solving (3.2) is an important part of this section and will be considered shortly. The third step is to compute the discrete \tilde{A}_k -harmonic extension of the boundary values of W_H computed in the previous step. This is done as follows: Let \tilde{W}_H be any extension of the boundary values of W_H in $S_h^0(\Omega)$, e.g., the extension

which is zero at all of the nodes not on Γ . Then $W_H = Y + \tilde{W}_H$, where Y vanishes on Γ and is the solution of

$$(3.3) \quad \tilde{A}_k(Y, \Phi) = -\tilde{A}_k(\tilde{W}_H, \Phi) \quad \text{for all } \Phi \in S_h^0(\Omega_k),$$

for $k = 1, \dots, n_d$. Equation (3.3), as in (3.1), requires independent discrete Dirichlet solves on the subdomains.

We now develop an algorithm for solving (3.2), that is, computing the values of W_H on Γ . For notational convenience we set $V = W_H|_\Gamma$. Then (3.2) reduces to

$$(3.4) \quad h^{-1} \sum_k \{ (\varepsilon \tilde{a}_k + h^2(\bar{b}_k + \varepsilon \bar{a}_k)) \langle V - \bar{V}_k, \chi \rangle_{\partial \Omega_k, h} + (\bar{b}_k + \varepsilon \bar{a}_k) h d^2 \bar{V}_k \bar{\chi}_k \} = F(\chi)$$

for all $\chi \in S_h(\Gamma)$, where F is a known linear functional on $S_h(\Gamma)$. We shall see that the problem of computing the solution V of (3.4) is straightforward if the values of \bar{V}_k are known. Indeed, V satisfies

$$(3.5) \quad h^{-1} \sum_k (\varepsilon \tilde{a}_k + h^2(\bar{b}_k + \varepsilon \bar{a}_k)) \langle V, \chi \rangle_{\partial \Omega_k, h} = G(\chi),$$

where the linear functional G in (3.5) depends upon F and the average values \bar{V}_k . Note that if we use the usual nodal basis for functions in $S_h(\Gamma)$, then the matrix corresponding to Problem (3.5) is diagonal and hence its solution can be trivially computed. Thus to solve (3.2), we need only demonstrate a technique for computing the average values \bar{V}_k .

We shall define another function $\tilde{V} \in S_h(\Gamma)$ which has the same average values as the solution V of (3.4). Obviously, the average values of V can then be computed by calculating the average values of \tilde{V} . Let $S_h^0(\Gamma)$ denote the collection of functions in $S_h(\Gamma)$ which have zero average value on every $\partial \Omega_k$. Clearly, $(V - \tilde{V}) \in S_h^0(\Gamma)$. To make the system of equations determining \tilde{V} of minimal size, we choose \tilde{V} in an orthogonal complement of $S_h^0(\Gamma)$. Specifically, let $S_h^\perp(\Gamma)$ be defined by

$$S_h^\perp(\Gamma) \equiv \{ \theta \in S_h(\Gamma) \mid Q(\theta, \omega) = 0 \text{ for all } \omega \in S_h^0(\Gamma) \},$$

and define \tilde{V} to be the unique function in $S_h^\perp(\Gamma)$ satisfying

$$(3.6) \quad Q(\tilde{V}, \theta) = F(\theta) \quad \text{for all } \theta \in S_h^\perp(\Gamma).$$

Note that \tilde{V} is the orthogonal projection of V into $S_h^\perp(\Gamma)$ and hence \tilde{V} has the same average values as V . In what follows, we shall derive a basis for $S_h^\perp(\Gamma)$. This basis will consist of functions with local (with respect to d) support, and hence \tilde{V} can be computed as the solution to a sparse, positive definite and symmetric “stiffness” matrix corresponding to (3.6). The number of unknowns in this system will always be less than or equal to n_d .

Before proceeding, we shall introduce some additional notation. Let $\nu_k \equiv h^{-1}(\varepsilon \tilde{a}_k + h^2(\bar{b}_k + \varepsilon \bar{a}_k))$ and define

$$Q_0(W, W) \equiv \sum_k \nu_k |W - \bar{W}|_{\partial \Omega_k, h}^2.$$

Note that Q and Q_0 only differ by terms involving the average values squared. Hence,

$$(3.7) \quad S_h^\perp(\Gamma) = \{ \theta \in S_h(\Gamma) \mid Q_0(\theta, \omega) = 0 \text{ for all } \omega \in S_h^0(\Gamma) \}.$$

We will first define functions $\phi_k \in S_h^\perp(\Gamma)$, for $k = 1, \dots, n_d$. Consider a fixed subregion with boundary $\partial\Omega_k$. The function ϕ_k is defined to be zero on all of the nodes on $\Gamma/\partial\Omega_k$, its values on $\partial\Omega_k$ are to be determined. Let W be in $S_h(\Gamma)$; then

$$(3.8) \quad Q_0(W, \phi_k) = \sum_i \gamma_i W(x_i) \phi_k(x_i) + \sum_j \kappa_{k,j} \bar{W}_j,$$

where

$$(3.9) \quad \gamma_i = h^{N-1} \sum_{\{j | x_i \in \partial\Omega_j \cap \partial\Omega_k\}} \nu_j$$

and

$$(3.10) \quad \kappa_{k,j} = -h^{N-1} \nu_j \sum_{x_l \in \partial\Omega_k \cap \partial\Omega_j} \phi_k(x_l).$$

The sum over i in (3.8) is taken over the nodes x_i on $\partial\Omega_k$ and the sum over j in (3.8) is taken over the subregions Ω_j with $\partial\Omega_j \cap \partial\Omega_k \neq \emptyset$. We define the nodal values of ϕ_k on $\partial\Omega_k$ by

$$(3.11) \quad \phi_k(x_i) = \frac{\nu_k}{\gamma_i}.$$

With the above choice for ϕ_k , it is evident that the first sum in (3.8) equals $\nu_k N_k \bar{W}_k$, where N_k is defined to be the number of nodes on $\partial\Omega_k$. Hence (3.8) becomes

$$(3.12) \quad Q_0(W, \phi_k) = \nu_k N_k \bar{W}_k + \sum_j \kappa_{k,j} \bar{W}_j.$$

By (3.7) and (3.12), $\phi_k \in S_h^\perp(\Gamma)$.

We will next show that

$$(3.13) \quad S_h^\perp(\Gamma) = \text{span}_{k=1, \dots, n_d} \phi_k.$$

It suffices to show that if

$$(3.14) \quad \theta \in S_h^\perp(\Gamma) \quad \text{and} \quad Q_0(\theta, \phi_k) = 0$$

for $k = 1, \dots, n_d$, then $\theta = 0$. Consider the matrix M defined by the right-hand side of (3.12), i.e.,

$$M_{i,j} = \delta_{i,j} \nu_i N_i + \kappa_{i,j},$$

where $\delta_{i,j}$ is the Kronecker Delta Function. Let θ satisfy (3.14) and $\bar{\theta}$ be the vector with components $\bar{\theta}_k$. Then by (3.12), (3.14) and the definition of M ,

$$(3.15) \quad M \bar{\theta} = 0.$$

To show that (3.13) holds, it suffices to show that M is invertible. Indeed, if M is invertible, then (3.15) implies that θ is also in $S_h^0(\Gamma)$, i.e., $\theta = 0$.

We will see that M is symmetric and positive definite and hence invertible. Indeed, the quantity γ_i in (3.9) depends upon the point x_i but not the subregion Ω_k . Consequently, $\kappa_{j,k} = \kappa_{k,j}$, i.e., M is symmetric. Furthermore, this system is sparse with positive diagonal entries and nonpositive off-diagonal entries. Also,

$$\sum_j \kappa_{k,j} = -\tilde{N}_k \nu_k,$$

where \tilde{N}_k is defined to be the number of nodes on $\partial\Omega_k$ which do not lie on $\partial\Omega$. Consequently, the matrix M is irreducibly diagonally dominant and hence positive definite; cf. [9].

In general, the functions $\phi_1, \dots, \phi_{n_d}$ may not be linearly independent. For example, in the case of the unit square with the checkerboard subdivision, the function

$$\sum_{\text{red squares}} \phi_k - \sum_{\text{black squares}} \phi_l = 0.$$

In this case it is easy to check that $\{\phi_1, \dots, \phi_{n_d-1}\}$ is linearly independent and hence forms a basis for $S_h^\perp(\Gamma)$. Bases for more complicated domains and subdivisions are also straightforward to derive.

For completeness, we restate the algorithm developed in this section for computing the solution W of (1.8).

Algorithm DD2.

- (1) Compute W_P by solving (3.1). This involves Dirichlet solves on the subdomains which can be done independently and in parallel.
- (2) Compute the values of W_H on Γ . First we compute the function \tilde{V} by solving (3.6) using the finite element basis $\{\phi_k\}$ described above. The average values of V are computed by calculating the average values of \tilde{V} . The values of W_H on Γ are then computed by solving the trivial equation (3.5).
- (3) Extend the boundary values of W_H by solving (3.3). As in Step 1, this involves Dirichlet solves on the subdomains which can be done independently and/or in parallel.
- (4) Set $W = W_P + W_H$.

Remark 3.1. When $\bar{b}_k = \bar{a}_k = 0$, the matrix M can be directly used to compute the average values of V . Indeed, if \bar{V} denotes the vector of average values of V , then by (3.12) and the definition of M ,

$$(3.16) \quad Q(V, \phi_k) = (M\bar{V})_k = F(\phi_k),$$

and hence the average values of V are given by

$$\bar{V} = M^{-1} \begin{pmatrix} F(\phi_1) \\ \vdots \\ F(\phi_{n_d}) \end{pmatrix}.$$

Remark 3.2. The method for computing the average values of V described in Remark 3.1 may not work well when either \bar{b}_k or \bar{a}_k are nonzero. In the general case, a matrix M satisfying (3.16) can be derived using similar techniques. In such cases, M may no longer be symmetric or diagonally dominant. Consequently, it may be difficult to obtain good numerical solutions for (3.16) when low-order terms are present. Note, however, that the algorithm described earlier for computing the average values by solving (3.6) always leads to numerically stable, sparse and symmetric positive definite systems.

4. Numerical Experiments. In this section we shall present some results of numerical experiments which illustrate the convergence properties of the preconditioned conjugate gradient algorithm using DD2 as a preconditioner. Two-dimensional examples where L is given by (1.2) will be considered first. These examples are taken from [3], so that a direct comparison between the preconditioners DD1 and DD2 can be made. Next, two-dimensional singularly perturbed

problems of the form (1.5), (1.7) will be studied. Finally, a model three-dimensional problem will be considered.

To avoid making this section too long, we shall not attempt to illustrate the full power and flexibility of the algorithm. Accordingly, other numerical examples on which we have tested the algorithm will not be included. These examples include applications to problems with discontinuous coefficients, problems with smoothly varying coefficients, and problems on domains with irregular geometry (see Examples 2, 3, 4, and 5 of Part I).

We shall define a number of parameters which will be introduced to study the convergence properties of the proposed preconditioning algorithm. The condition number of the preconditioned system is denoted by K . The integer n is defined to be the number of iterations required to reduce the A -norm (defined by $A(\cdot, \cdot)^{1/2}$) of the error $E_n = U - U_n$ by a factor of .0001. Here, U is a randomly generated solution of (1.5), normalized so that $-1 \leq U \leq 1$, and U_n is the approximation to U obtained using n steps of a conjugate gradient algorithm preconditioned by DD2. It is well known that the A -norm of the iteration error satisfies the bound

$$A(E_j, E_j) \leq 4\rho^{2j} A(E_0, E_0),$$

where

$$(4.1) \quad \rho \equiv \frac{\sqrt{K} - 1}{\sqrt{K} + 1}.$$

We shall sometimes compare ρ with the average observed reduction ρ_0 defined by

$$\rho_0 = \left(\frac{A(E_n, E_n)}{A(E_0, E_0)} \right)^{1/2n}.$$

Example 1. The first set of numerical experiments is applied to the standard model problem given by

$$(4.2) \quad Lu = f \quad \text{on } \Omega \quad \text{and} \quad u = 0 \quad \text{on } \partial\Omega,$$

where L is taken to be the Laplace operator $-\Delta$ and Ω is the unit square. There are many techniques available for solving this problem. However, this problem is interesting in that it illustrates many of the convergence properties of the proposed preconditioner. The square is partitioned into m^2 equal subsquares and hence $d = 1/m$.

The first table gives some indication of a typical run on an iteration-by-iteration basis. Here we break the square into sixteen subsquares and hence $d = 1/4$ and set $h = 1/32$. Table 4.1 gives the normalized error reduction as a function of the number of preconditioned steps in the A -norm and the maximum norm. Note that it takes 14 iterations to reduce the A -norm error by .0001. For this example, the average error reduction over 14 steps in the A -norm (resp. maximum norm) was .52 (resp. .60).

TABLE 4.1
Step by Step Iterative Convergence for Example 1.

Iteration	<i>A</i> -error	Max-error
1	8.0×10^{-1}	1.0
2	4.0×10^{-1}	1.4
3	1.4×10^{-1}	6.7×10^{-1}
4	5.7×10^{-2}	3.5×10^{-1}
5	2.3×10^{-2}	1.7×10^{-1}
6	1.1×10^{-2}	9.5×10^{-2}
7	5.4×10^{-3}	4.5×10^{-2}
8	2.3×10^{-3}	3.5×10^{-2}
9	1.8×10^{-3}	2.6×10^{-2}
10	1.2×10^{-3}	1.5×10^{-2}
11	6.8×10^{-4}	6.9×10^{-3}
12	3.5×10^{-4}	2.9×10^{-3}
13	1.9×10^{-4}	1.6×10^{-3}
14	9.3×10^{-5}	7.7×10^{-4}

The next two tables show that, in practice, the condition number of the preconditioned systems exhibit the growth rates predicted by the theory. In Table 4.2, we fix $d = 1/4$ and vary h . As predicted by Theorem 1, K grows like d/h . For Table 4.3, we fix $d/h = 4$ and vary h . In this case, the condition number for the precondition system remains bounded independent of h . Tables 4.2 and 4.3 also give the observed average reduction ρ_0 and n , the number of iterations required to reduce the *A*-norm error by a factor of .0001. The number of subregions m is also included in Table 4.3.

TABLE 4.2
Iterative Convergence Results for Example 1 when $d = 1/4$.

h	K	$\frac{15d}{8h}$	ρ_0	n
1/8	3.4	3.8	.21	7
1/16	7.2	7.5	.39	10
1/32	14	15	.52	14
1/64	30	30	.60	19
1/128	61	60	.68	24

TABLE 4.3
Iterative Convergence Results for Example 1 when $d/h = 4$.

h	K	ρ_0	n	m
1/8	6.6	.20	6	4
1/16	7.2	.40	10	16
1/32	7.5	.42	11	64
1/64	7.6	.42	11	256

Example 2. The next example illustrates the algorithm applied to singularly perturbed problems. We again consider the unit square with the same subdomain

subdivisions as in Example 1. For this problem, the form A is given by

$$(4.3) \quad A(v, \phi) = \varepsilon D(v, \phi) + (v, \phi).$$

Table 4.4 gives iterative convergence results when $h = 1/32$, $d = 1/4$, $\varepsilon = h^p$, and p varies between 0 and 2. This range of ε is typically that which occurs when time-stepping procedures are applied to parabolic problems and ε is essentially the size of the time step. Table 4.4 shows that the condition numbers for the preconditioned systems, as p varies, remain bounded by the condition number corresponding to the case $p = 0$. Similar results were obtained when h and d were varied.

TABLE 4.4
Iterative Convergence Results for Example 2 as p Varies.

p	K	ρ_0	n
0	15.1	.51	14
0.5	14.7	.51	14
1.0	12.4	.49	14
1.5	9.7	.45	12
2	6.6	.33	9

Example 3. For our final example we consider a model three-dimensional problem. Here we set Ω to be the unit cube and define the subregions by breaking Ω into 27 subcubes of equal size. We let L be an elliptic operator of the form

$$(4.4) \quad Lu = -\nabla \cdot \mu \nabla u \quad \text{in } \Omega,$$

where μ is a piecewise constant function on Ω and constant on the subdomains. Figure 4.1 gives the values of μ as a function of the x , y , z coordinates of the center of the subregions. These values were chosen to exhibit relatively large jumps across subregions, but are otherwise arbitrary. Table 4.5 gives iterative convergence results for the conjugate gradient method preconditioned by DD2 applied to the finite element equations corresponding to (4.4). Note that even though the coefficients of the operator have large jumps, the condition number K of the preconditioned system remains relatively small. In fact, the results reported do not differ significantly from results (not presented) for the case $\mu \equiv 1$. This is in agreement with Remark 2.4.

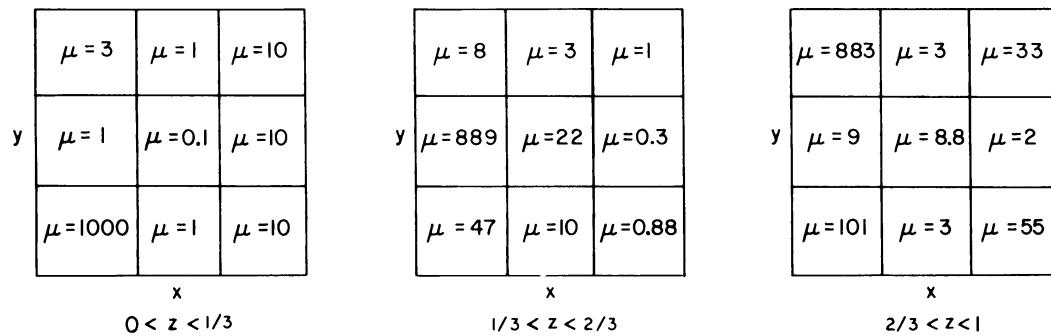


FIGURE 4.1
Coefficients for Example 3.

TABLE 4.5
Iterative Convergence Results for Example 3.

h	K	ρ_0	n
1/6	6.8	.39	11
1/12	17.4	.55	16
1/24	38	.64	21

Department of Mathematics
Cornell University
Ithaca, New York 14853

Brookhaven National Laboratory
Upton, New York 11973

Department of Mathematics
Cornell University
Ithaca, New York 14853

1. P. E. BJØRSTAD & O. B. WIDLUND, "Solving elliptic problems on regions partitioned into substructures," *Elliptic Problem Solvers II* (G. Birkhoff and A. Schoenstadt, eds.), Academic Press, New York, 1984, pp. 245–256.
2. P. E. BJØRSTAD & O. B. WIDLUND, "Iterative methods for the solution of elliptic problems on regions partitioned into substructures," *SIAM J. Numer. Anal.*, v. 23, 1986, pp. 1097–1120.
3. J. H. BRAMBLE, J. E. PASCIAK & A. H. SCHATZ, "The construction of preconditioners for elliptic problems by substructuring. I," *Math. Comp.*, v. 47, 1986, pp. 103–134.
4. J. H. BRAMBLE, J. E. PASCIAK & A. H. SCHATZ, "An iterative method for elliptic problems on regions partitioned into substructures," *Math. Comp.*, v. 46, 1986, pp. 361–369.
5. B. L. BUZBEE & F. W. DORR, "The direct solution of the biharmonic equation on rectangular regions and the Poisson equation on irregular regions," *SIAM J. Numer. Anal.*, v. 11, 1974, pp. 753–763.
6. B. L. BUZBEE, F. W. DORR, J. A. GEORGE & G. H. GOLUB, "The direct solution of the discrete Poisson equation on irregular regions," *SIAM J. Numer. Anal.*, v. 8, 1971, pp. 722–736.
7. Q. V. DIHN, R. GLOWINSKI & J. PÉRIAUX, "Solving elliptic problems by domain decomposition methods," *Elliptic Problem Solvers II* (G. Birkhoff and A. Schoenstadt, eds.), Academic Press, New York, 1984, pp. 395–426.
8. G. H. GOLUB & D. MEYERS, "The use of preconditioning over irregular regions," *Proc. 6th Internat. Conf. Comput. Methods in Sci. and Engng.*, Versailles, France, 1983.
9. R. S. VARGA, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, N.J., 1962.

**4.3 The construction of preconditioners for elliptic problems by
substructuring. III**

The construction of preconditioners for elliptic problems by substructuring. III
[21]

The Construction of Preconditioners for Elliptic Problems by Substructuring, III*

By James H. Bramble, Joseph E. Pasciak, and Alfred H. Schatz

Abstract. In earlier parts of this series of papers, we constructed preconditioners for the discrete systems of equations arising from the numerical approximation of elliptic boundary value problems. The resulting algorithms are well suited for implementation on computers with parallel architecture. In this paper, we will develop a technique which utilizes these earlier methods to derive even more efficient preconditioners. The iterative algorithms using these new preconditioners converge to the solution of the discrete equations with a rate that is independent of the number of unknowns. These preconditioners involve an incomplete Chebyshev iteration for boundary interface conditions which results in a negligible increase in the amount of computational work. Theoretical estimates and the results of numerical experiments are given which demonstrate the effectiveness of the methods.

1. Introduction. The aim of this series of papers is to propose and analyze methods for efficiently solving the equations resulting from finite element discretizations of second-order elliptic boundary value problems on general domains in R^2 and R^3 . In particular, we shall be concerned with constructing easily invertible and “effective” preconditioners for the resulting system of discrete equations which can be used in a preconditioned iterative algorithm to define a rapid solution method. The methods developed are well suited to parallel computing architectures.

In Parts I and II (references [4] and [5]), we described and analyzed methods for constructing preconditioners for elliptic boundary value problems on polygonal domains in R^2 and R^3 . The proposed methods were based on decomposing the domain into subdomains of size d and involved the solution of related problems on the subdomains and lower-order coupling systems on the subdomain boundaries. The condition number for the preconditioned system was shown to be on the order of $(1 + \ln(d/h))^2$ for the method of [4] and d/h for the method of [5]. Here h is the mesh size. In this paper, we describe a technique which can utilize such methods to develop more efficient preconditioners. The condition numbers for the resulting preconditioned systems will be made independent of d and h with only a slight increase in computational effort.

Received May 7, 1987.

1980 *Mathematics Subject Classification* (1985 Revision). Primary 65N30; Secondary 65F10.

*This manuscript has been authored under contract number DE-AC02-76CH00016 with the U.S. Department of Energy. Accordingly, the U.S. Government retains a non-exclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. This work was also supported in part under the National Science Foundation Grant No. DMS84-05352 and under the Air Force Office of Scientific Research, Contract No. ISSA86-0026 and by the U.S. Army Research Office through the Mathematical Science Institute, Cornell University.

©1988 American Mathematical Society
0025-5718/88 \$1.00 + \$.25 per page

Let Ω be a bounded domain in R^2 with a piecewise smooth boundary $\partial\Omega$. As a model problem for a second-order uniformly elliptic equation, we shall consider the Dirichlet problem

$$(1.1) \quad \begin{aligned} Lu &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega, \end{aligned}$$

where

$$Lv = - \sum_{i,j=1}^2 \frac{\partial}{\partial x_i} \left(a_{ij} \frac{\partial v}{\partial x_j} \right),$$

with a_{ij} symmetric and uniformly positive definite, bounded and piecewise smooth on Ω . The generalized Dirichlet form is given by

$$(1.2) \quad A(v, \phi) = \sum_{i,j=1}^2 \int_{\Omega} a_{ij} \frac{\partial v}{\partial x_i} \frac{\partial \phi}{\partial x_j} dx,$$

which is defined for all v and ϕ in the Sobolev space $H^1(\Omega)$ (the space of distributions with square-integrable first derivatives). The $L^2(\Omega)$ inner product is denoted

$$(v, \phi) = \int_{\Omega} v\phi dx.$$

The subspace $H_0^1(\Omega)$ is the completion of the smooth functions with support in Ω with respect to the norm in $H^1(\Omega)$. The weak formulation of the problem defined by (1.1) is: Find $u \in H_0^1(\Omega)$ such that

$$(1.3) \quad A(u, \phi) = (f, \phi) \quad \text{for all } \phi \in H_0^1(\Omega).$$

This leads immediately to the standard Galerkin approximation. Let $S_h^0(\Omega)$ be a finite-dimensional subspace of $H_0^1(\Omega)$. The Galerkin approximation is defined as the solution of the following problem: Find $U \in S_h^0(\Omega)$ such that

$$(1.4) \quad A(U, \Phi) = (f, \Phi) \quad \text{for all } \Phi \in S_h^0(\Omega).$$

The underlying method which we will consider is a preconditioned iterative method. As explained in Part I, the task of defining a preconditioner for the matrix problem corresponding to (1.4) is the same as that of defining another positive definite form $B(\cdot, \cdot)$ on $S_h^0(\Omega) \times S_h^0(\Omega)$. The importance of making a “good” choice for B is well known. The form B will define a good preconditioner provided it has two basic properties. First, the problem of finding the function $W \in S_h^0(\Omega)$ satisfying

$$(1.5) \quad B(W, \Phi) = G(\Phi) \quad \text{for all } \Phi \in S_h^0(\Omega),$$

for a given linear functional G , should be more economical to solve on a given computer architecture than (1.4). Secondly, B should be spectrally close to A in the sense that there are positive numbers β_0 and β_1 satisfying

$$(1.6) \quad \beta_0 B(V, V) \leq A(V, V) \leq \beta_1 B(V, V) \quad \text{for all } V \in S_h^0(\Omega),$$

where the ratio β_1/β_0 is not too large. These two properties will guarantee, firstly, that the work per iterative step in applying the preconditioned method will be small, and, secondly, that the number of steps to reduce the error to a given size will also be small, so that an efficient algorithm will result.

In Section 2 the form B is defined and the essential step in the iterative algorithm of solving (1.5) is described. This form is defined in terms of a polynomial P_m which is related to the classical Chebyshev polynomials. The relevant properties are given in Section 3. Section 4 discusses various computational aspects of the method in a more general setting. Finally, in Section 5, the results of numerical experiments are given. These computations show that the theoretical estimates are fully realized in practice.

For other works dealing with the numerical solution of boundary value problems via substructuring we refer the reader to [1], [2], [3], [6], [7], [8], [9]. We emphasize that a novel feature of our methods [4], [5] is that more than two subdomains can meet at an interior point of the original domain. For example, our methods apply to a checkerboard subdivision of a square. Using the technique of this paper, the condition number for the resulting system is shown to be bounded independently of the number of such points.

2. The Construction of $B(\cdot, \cdot)$ and the Preconditioning Algorithm. As mentioned in the introduction, the preconditioner which we will construct involves the solution of smaller related problems on subdomains and subdomain boundaries. As in Part I, for the sake of simplicity of exposition, we shall proceed with the discussion only for the special case of polygonal domains and piecewise linear approximations.

More precisely, we shall begin with the following assumptions with regard to Ω . These assumptions are the same as those given in Section 2 of Part I, and hence the results given in Section 3 of Part I apply.

- A.1: Ω is a polygonal domain.
- A.2: For each h , $0 < h < 1$ a parameter, Ω has been given a quasi-uniform triangulation Ω^h . By this we mean that there exists a positive constant c_1 independent of h such that each triangle $\tau^h \in \Omega^h$ contains a ball of radius $c_1 h$ and is contained in a ball of radius h .
- A.3: For each triangulation Ω^h , Ω may be written in terms of n_r disjoint regions, Ω_k , with $\bar{\Omega} = \bigcup \bar{\Omega}_k$, which are either quadrilaterals or triangles whose sides coincide with the mesh lines of the original triangulation and which are quasi-uniform of size $d \geq h$ with constants as above which are independent of d and h . If Ω_k is a quadrilateral, we require additionally that the lengths of each side be bounded from below by $c_1 d$ and that any interior angle α satisfy $0 < C_0 \leq \alpha \leq C_1 < \pi$. The collection of regions Ω_k will frequently be referred to as the subdomains (see Figure 2.1).

For each h , let $S_h(\Omega)$ be the space of continuous piecewise linear functions defined relative to the triangulation Ω^h and $S_h^0(\Omega)$ be the subspace of $S_h(\Omega)$ consisting of those functions which vanish on $\partial\Omega$. $S_h^0(\Omega_k)$ will denote the subspace of $S_h^0(\Omega)$ of functions whose supports are contained in $\bar{\Omega}_k$ (in particular, they vanish on $\partial\Omega_k$ and outside $\bar{\Omega}_k$). In addition, let $S_h(\Omega_k)$ be the set of functions which are restrictions of those in $S_h^0(\Omega)$ to $\bar{\Omega}_k$. $S_h(\partial\Omega_k)$ will denote the restrictions of $S_h(\Omega_k)$ to $\partial\Omega_k$. Let $\Gamma = \bigcup \partial\Omega_j$ and $S_h(\Gamma)$ be the restriction of functions in $S_h^0(\Omega)$ to Γ . In what follows, c and C (with or without subscript) will denote generic positive constants which are independent of h , d and the regions Ω_k .

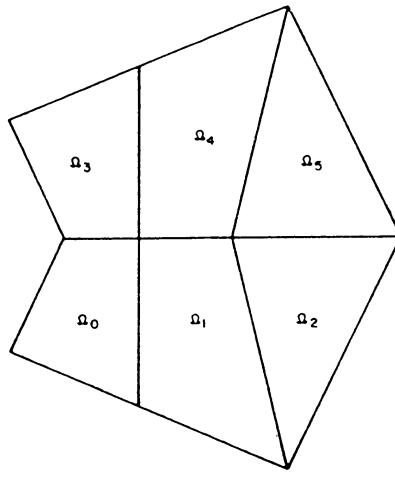


FIGURE 2.1
A typical domain with subdomains.

For simplicity of presentation, we shall restrict our development to the case in which each $\partial\Omega_k$ has a uniformly (with respect to arc length) spaced grid. This restriction will be removed in Section 4. We define

$$A_k(u, v) = \sum_{i,j=1}^2 \int_{\Omega_k} a_{ij} \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_j} dx,$$

and hence

$$(2.1) \quad A(u, v) = \sum_{i=1}^{n_r} A_i(u, v).$$

To define B , we first decompose functions in $S_h^0(\Omega)$ as follows: Write $W = W_P + W_H$ where $W_P \in S_h^0(\Omega_1) \oplus \cdots \oplus S_h^0(\Omega_{n_r})$ and satisfies, for $k = 1, \dots, n_r$,

$$(2.2) \quad A_k(W_P, \Phi) = A_k(W, \Phi) \quad \text{for all } \Phi \in S_h^0(\Omega_k).$$

Notice that W_P is determined on Ω_k by the values of W on Ω_k and that

$$(2.3) \quad A_k(W_H, \Phi) = 0 \quad \text{for all } \Phi \in S_h^0(\Omega_k).$$

Thus on each Ω_k , W is decomposed into a function W_P which vanishes on $\partial\Omega_k$ and a function $W_H \in S_h(\Omega_k)$ which satisfies the above homogeneous equations and has the same values as W on $\partial\Omega_k$. We shall refer to such a function W_H as “discrete A -harmonic”.

We note that the above decomposition is orthogonal in the inner product defined by A and hence

$$A(W, W) = A(W_P, W_P) + A(W_H, W_H).$$

We shall define $B(\cdot, \cdot)$ by replacing the $A(W_H, W_H)$ term in the above equation. To do this, we first note that by Lemma 3.2 of Part I [4],

$$(2.4) \quad c|W_H|_{1/2, \partial\Omega_k}^2 \leq A_k(W_H, W_H) \leq C|W_H|_{1/2, \partial\Omega_k}^2$$

for discrete A -harmonic functions W_H with zero mean value on Ω_k . The norm $|\cdot|_{1/2, \partial\Omega_k}$ is the weighted norm on $H^{1/2}(\partial\Omega_k)$ given by (see [10], [11])

$$|w|_{1/2, \partial\Omega_k} \equiv \left(\int_{\partial\Omega_k} \int_{\partial\Omega_k} \frac{(w(x) - w(y))^2}{|x - y|} ds(x) ds(y) + d^{-1} \langle w, w \rangle_{\partial\Omega_k} \right)^{1/2}.$$

Here, $\langle \cdot, \cdot \rangle_{\partial\Omega_k}$ denotes the L^2 inner product on $\partial\Omega_k$ (the corresponding norm will be denoted by $|\cdot|_{0,\partial\Omega_k}$). In turn, we shall replace the norm $|\cdot|_{1/2,\partial\Omega_k}$ by a more computationally convenient norm.

To this purpose, we define the operator l on $S_h(\partial\Omega_k)$ by

$$(2.5) \quad \langle lV, \Phi \rangle_{\partial\Omega_k} \equiv \langle V', \Phi' \rangle_{\partial\Omega_k} \quad \text{for all } \Phi \in S_h(\partial\Omega_k),$$

where the primes denote differentiation with respect to arc length along each side of $\partial\Omega_k$. Now l is a linear operator on $S_h(\partial\Omega_k)$ approximating the boundary operator $-\frac{\partial^2}{\partial s^2}$, and it can be shown that there are constants c and C , independent of d and h , such that

$$(2.6) \quad c|V|_{1/2,\partial\Omega_k}^2 \leq \langle l^{1/2}V, V \rangle_{\partial\Omega_k} + d^{-1}\langle V, V \rangle_{\partial\Omega_k} \leq C|V|_{1/2,\partial\Omega_k}^2$$

for all $V \in S_h(\partial\Omega_k)$. The following Poincaré inequality holds for all W with zero mean value on $\partial\Omega_k$,

$$d^{-1}|W|_{0,\partial\Omega_k}^2 \leq cd\langle lW, W \rangle_{\partial\Omega_k}.$$

It then follows by expansion in terms of eigenvectors of l that

$$|W - \bar{W}|_{1/2,\partial\Omega_k}^2 \leq c\langle l^{1/2}(W - \bar{W}), W - \bar{W} \rangle_{\partial\Omega_k} = c\langle l^{1/2}W, W \rangle_{\partial\Omega_k},$$

where \bar{W} is the mean value of W on $\partial\Omega_k$. Consequently, we may replace (2.4) by

$$(2.7) \quad c\langle l^{1/2}W_H, W_H \rangle_{\partial\Omega_k} \leq A_k(W_H, W_H) \leq C\langle l^{1/2}W_H, W_H \rangle_{\partial\Omega_k},$$

which holds for all discrete A -harmonic functions W_H . Summing the above inequality gives

$$(2.8) \quad c\langle QW_H, W_H \rangle_\Gamma \leq A(W_H, W_H) \leq C\langle QW_H, W_H \rangle_\Gamma,$$

where

$$\langle QW_H, W_H \rangle_\Gamma \equiv \sum_k \alpha_k \langle l^{1/2}W_H, W_H \rangle_{\partial\Omega_k}.$$

The constants α_k are scaling factors. One reasonable choice is to take $\alpha_k = (\lambda_1^k + \lambda_0^k)/2$ where λ_1^k and λ_0^k are respectively the largest and smallest eigenvalue of the 2×2 matrix $\{a_{ij}(x_0)\}$ at some point $x_0 \in \Omega_k$. By (2.8), the form $\langle QW_H, W_H \rangle_\Gamma$ is uniformly equivalent to $A(W_H, W_H)$. Consequently, the form \tilde{B} defined by

$$(2.9) \quad \tilde{B}(W, W) \equiv A(W_P, W_P) + \langle QW_H, W_H \rangle_\Gamma$$

is uniformly equivalent to A on $S_h^0(\Omega) \times S_h^0(\Omega)$. The difficulty with using \tilde{B} as our preconditioner is that the corresponding algorithm for solving (1.5) requires the solution of problems of the form: Find $V \in S_h(\Gamma)$ such that

$$(2.10) \quad \langle QV, \phi \rangle_\Gamma = F(\phi) \quad \text{for all } \phi \in S_h(\Gamma).$$

It is not easy to solve (2.10); consequently, the choice of $B = \tilde{B}$ will not lead to a good preconditioner.

Finally, we shall define our preconditioner for A by replacing Q in (2.9) by an operator \bar{Q} which is easier to invert. In fact, we define \bar{Q} from its inverse. Set

$$(2.11) \quad \bar{Q}^{-1} = P_m(\tilde{Q}^{-1}Q)\tilde{Q}^{-1},$$

where \tilde{Q} is some other positive definite symmetric operator on $S_h(\Gamma)$ and P_m is a polynomial of degree m . For our present discussion, we can think of \tilde{Q} as arbitrary. Some interesting examples, from a computational point of view, will result from the preconditioners constructed in Parts I and II. The possible choices for \tilde{Q} will be considered in more detail in later sections.

The polynomial in (2.11) is defined (complete details are given in Section 3) so that \bar{Q}^{-1} is positive definite and \bar{Q} is uniformly equivalent to Q on $S_h(\Gamma)$, i.e.,

$$(2.12) \quad c_0 \langle \bar{Q}V, V \rangle_{\Gamma} \leq \langle QV, V \rangle_{\Gamma} \leq c_1 \langle \bar{Q}V, V \rangle_{\Gamma} \quad \text{for all } V \in S_h(\Gamma).$$

Hence, we define our preconditioner B by

$$(2.13) \quad B(W, W) = A(W_P, W_P) + \langle \bar{Q}W_H, W_H \rangle_{\Gamma}.$$

An immediate consequence of (2.12) and (2.8) is that B is uniformly equivalent to A ; more precisely, we have the following:

THEOREM. *Let B be given by (2.13), where \bar{Q} , defined by (2.11), satisfies (2.12) with c_0 and c_1 independent of d and h . Then there exist positive constants c and C independent of d and h such that*

$$cB(W, W) \leq A(W, W) \leq CB(W, W) \quad \text{for all } W \in S_h^0(\Omega).$$

We shall describe a three-step algorithm to compute the solution $W = W_P + W_H$ of (1.5) (see [4] and [5]). The function W_P extended by zero outside of Ω_k is a function in $S_h^0(\Omega_k)$ which satisfies

$$(2.14) \quad A_k(W_P, \Phi) = G(\Phi) \quad \text{for all } \Phi \in S_h^0(\Omega_k).$$

Thus, for step one, the function W_P on Ω_k can be obtained by solving the corresponding Dirichlet problem (2.14). Note that the problems on different subdomains are independent of each other and hence can be solved in parallel.

Now with W_P known, we are left with the problem of finding W_H , the second step in the algorithm. It is not difficult to see that the boundary values of W_H satisfy the equation

$$(2.15) \quad \langle \bar{Q}W_H, \theta \rangle_{\Gamma} = G(\bar{\theta}) - A(W_P, \bar{\theta}) \quad \text{for all } \theta \in S_h(\Gamma),$$

where $\bar{\theta}$ is any extension of θ in $S_h^0(\Omega)$. Let us discuss the solution of (2.15) in more detail.

To solve (2.15), we must apply the polynomial in (2.11). Let V_H satisfy $\tilde{Q}V_H = \bar{Q}W_H$, i.e.,

$$(2.16) \quad \langle \tilde{Q}V_H, \theta \rangle_{\Gamma} = G(\bar{\theta}) - A(W_P, \bar{\theta}) \quad \text{for all } \theta \in S_h(\Gamma).$$

By the definition of \bar{Q} , W_H on Γ is given by

$$(2.17) \quad W_H = P_m(\tilde{Q}^{-1}Q)V_H.$$

In addition, we must evaluate $\tilde{Q}^{-1}Q$, i.e., given $\varsigma \in S_h(\Gamma)$ we must find $\eta = \tilde{Q}^{-1}Q\varsigma$ solving

$$(2.18) \quad \langle \tilde{Q}\eta, \theta \rangle_{\Gamma} = \langle Q\varsigma, \theta \rangle_{\Gamma} \quad \text{for all } \theta \in S_h(\Gamma).$$

Accordingly, the computation of W_H on Γ only requires evaluation of the form $\langle Q\varsigma, \cdot \rangle_{\Gamma}$ and the inversion of the $\langle \tilde{Q}\cdot, \cdot \rangle_{\Gamma}$ form. The evaluation of the right-hand side of (2.18) is discussed in Section 4.

Once the boundary values of W_H are known, the third step of the algorithm only requires the computation of the discrete harmonic extension to the interior of the subdomains. As described in [4], [5], this problem can be reduced to the solution of independent Dirichlet problems on the subdomains.

Remark 2.1. As will be seen in the following section, the degree of the polynomial P_m depends upon the relative condition number of the forms Q and \tilde{Q} . Indeed, if Q and \tilde{Q} satisfy inequalities of the form

$$(2.19) \quad \lambda_0 \langle \tilde{Q}v, v \rangle_\Gamma \leq \langle Qv, v \rangle_\Gamma \leq \lambda_1 \langle \tilde{Q}v, v \rangle_\Gamma \quad \text{for all } v \in S_h(\Gamma),$$

then it suffices to choose m proportional to $\sqrt{\lambda_1/\lambda_0}$.

Remark 2.2. Other examples of \tilde{Q} have been constructed in our earlier papers. If \tilde{Q} is chosen to be the identity, then (2.19) holds with $\lambda_1/\lambda_0 \leq c(dh)^{-1}$. Choosing \tilde{Q} corresponding to the boundary form constructed in [5], i.e.,

$$(2.20) \quad \langle \tilde{Q}W_H, W_H \rangle_\Gamma \equiv Q(W_H, W_H),$$

where Q is defined by (2.14) of [5], the results of [5] imply that (2.19) holds with $\lambda_1/\lambda_0 \leq cd/h$. Finally, choosing \tilde{Q} corresponding to the boundary form constructed in [4], i.e.,

$$(2.21) \quad \langle \tilde{Q}W_H, W_H \rangle_\Gamma \equiv B(W_H, W_H),$$

where B is defined by (2.3) of [4], the results of [4] show that (2.19) holds with $\lambda_1/\lambda_0 \leq c(1 + \ln(d/h)^2)$.

3. The Construction of the Polynomial P_m . In this section we shall construct and analyze the polynomial P_m which appears in (2.11). The ideas involved here are not new, but we will restate the relevant results and constructions for completeness.

We first observe that (2.12) is equivalent to

$$(3.1) \quad c \langle Q^{-1}V, V \rangle_\Gamma \leq \langle \tilde{Q}^{-1}V, V \rangle_\Gamma \leq C \langle Q^{-1}V, V \rangle_\Gamma \quad \text{for all } V \in S_h(\Gamma).$$

Now the operator $\tilde{Q}^{-1}Q$ is selfadjoint in the inner product given by

$$[u, v] \equiv \langle Qu, v \rangle_\Gamma,$$

and the change of variable $X = Q^{-1}V$ gives that (3.1) is equivalent to

$$(3.2) \quad c[X, X] \leq [P_m(\tilde{Q}^{-1}Q)\tilde{Q}^{-1}QX, X] \leq C[X, X] \quad \text{for all } X \in S_h(\Gamma).$$

A straightforward spectral argument gives that (3.2) holds (with $C = 1 + \varepsilon$ and $c = 1 - \varepsilon$) whenever the polynomial P_m satisfies

$$(3.3) \quad |1 - xP_m(x)| \leq \varepsilon \quad \text{for all } x \in [\lambda_0, \lambda_1],$$

where ε is any positive constant less than one not depending on d or h , and λ_0 and λ_1 are the constants appearing in (2.19).

We shall define P_m in terms of the Chebyshev polynomials. The Chebyshev polynomial $T_j(y)$ of degree j is given by

$$T_j(y) = \cos(j \arccos(y)).$$

Define P_m by

$$(3.4) \quad 1 - xP_m(x) = \frac{T_{m+1}(y(x))}{T_{m+1}(y(0))},$$

where y is the linear function which takes the interval (λ_0, λ_1) into $(-1, 1)$, i.e.,

$$y(x) = \frac{2}{\lambda_1 - \lambda_0}x - \frac{\lambda_1 + \lambda_0}{\lambda_1 - \lambda_0}.$$

Since $|T_{m+1}(y)| \leq 1$ for $y \in [-1, 1]$, we have that

$$(3.5) \quad |1 - xP_m(x)| \leq \frac{1}{|T_{m+1}(-\frac{\lambda_1+\lambda_0}{\lambda_1-\lambda_0})|} \leq 2 \left(\frac{\sqrt{\gamma} - 1}{\sqrt{\gamma} + 1} \right)^{m+1} \quad \text{for all } x \in [\lambda_0, \lambda_1],$$

where $\gamma \equiv \lambda_1/\lambda_0$. The second inequality in (3.5) follows from the identity

$$T_j(y) = \frac{1}{2}[(y + \sqrt{y^2 - 1})^j + (y - \sqrt{y^2 - 1})^j]$$

and elementary manipulations.

To satisfy (3.3), we must choose m large enough so that

$$(3.6) \quad \left(\frac{\sqrt{\gamma} - 1}{\sqrt{\gamma} + 1} \right)^{m+1} \leq \frac{\varepsilon}{2}.$$

Consequently, it suffices to choose m in proportion to $\sqrt{\gamma}$ as γ becomes large. Note however, that for any m , there is an $\varepsilon(m, \lambda_1, \lambda_0) < 1$ satisfying (3.3). This implies that \tilde{Q} will always be positive. The following proposition follows immediately.

PROPOSITION 3.1. *Let $0 < \varepsilon < 1$ be given. There exists a positive constant C_1 independent of d and h such that if P_m is given by (3.4) and*

- (i) *if \tilde{Q} corresponds to the identity operator on $S_h(\Gamma)$ and $m \geq C_1(1+(dh)^{-1/2})$,*
or
- (ii) *if \tilde{Q} is given by (2.20) and $m \geq C_1(1+(d/h)^{1/2})$, or*
- (iii) *if \tilde{Q} is given by (2.21) and $m \geq C_1(1+\ln(d/h))$,*

then (3.3) holds. Furthermore, (2.12) holds with $c_0 = 1 - \varepsilon$ and $c_1 = 1 + \varepsilon$.

Remark 3.1. To use the preconditioner defined by B with a given \tilde{Q} , one needs to know bounds λ_0 and λ_1 of (2.19). Excellent bounds can be obtained in practice by, for example, applying the power method for eigenvalue estimation. This involves the repeated evaluation of $\tilde{Q}^{-1}Q$, which is an essential ingredient in the polynomial evaluation (2.11). The cost of this calculation is minor compared to the overall cost of the algorithm, and no additional coding is necessary.

Remark 3.2. The most straightforward algorithm involves choosing \tilde{Q} to be the identity. For smooth problems, λ_1/λ_0 is not too large (see Example 5 of Section 5), in which case the degree of P_m grows like $(dh)^{-1/2}$. However, this algorithm requires an excessive number of terms in examples with large jumps in the coefficients across subregion interfaces. In contrast, if we use (2.21) or (2.20) to define \tilde{Q} , then the constant C_1 appearing in Proposition 3.1 is independent of the jumps in coefficients as long as the jumps occur at the subdomain boundaries (see Examples 2 and 4 of Section 5).

Remark 3.3. The coefficients of the polynomial P_m can easily be calculated by using well-known identities involving Chebyshev polynomials. However, when λ_1/λ_0 is large, the computation of $P_m(\tilde{Q}^{-1}Q)$ directly, using the coefficients of P_m , is somewhat unstable. We suggest the use of the following two-term recurrence relation for $R_m \equiv P_m(\tilde{Q}^{-1}Q)V$:

- (i) Define $\rho = \frac{\lambda_1 - \lambda_0}{\lambda_1 + \lambda_0}$ and $\alpha = \frac{2}{\lambda_1 + \lambda_0}$;

- (ii) set $\omega_0 = 2$ and $R_0 = \alpha V$;
- (iii) set $w_1 = (1 - \omega_0 \rho^2 / 4)^{-1}$ and $R_1 = \frac{4\alpha}{(2-\rho^2)} [V - \frac{\alpha}{2} \tilde{Q}^{-1} Q V]$;
- (iv) for $m > 1$, set $\omega_m = (1 - \omega_{m-1} \rho^2 / 4)^{-1}$ and

$$R_m = \omega_m (R_{m-1} - \alpha \tilde{Q}^{-1} Q R_{m-1}) + \alpha \omega_m V - (\omega_m - 1) R_{m-2}.$$

4. Computational Aspects and Generalizations. In this section we shall consider various computational aspects of the method as well as some extensions and generalizations. We first describe the computation of the right-hand side of (2.18) in the special case where the mesh points on $\partial\Omega_k$ are uniformly spaced. We next give a way of extending the techniques of Section 2 to variable coefficient problems on certain irregular mesh domains.

Assume first that the nodes on $\partial\Omega_j$ are equally spaced with respect to arc length. As discussed in Section 2, given a function $\zeta \in S_h(\Gamma)$, we must be able to compute the data

$$\langle Q\zeta, \cdot \rangle_\Gamma$$

appearing in (2.18). By the definition of Q , it obviously suffices to compute the data

$$(4.1) \quad \langle l^{1/2}\zeta, \cdot \rangle_{\partial\Omega_j}$$

for each subdomain Ω_j . We consider first the operator l from which $l^{1/2}$ is defined. Let r be the number of nodes on $\partial\Omega_j$ and $\{\Phi_p, p = 1, \dots, r\}$ denote the nodal basis for $S_h(\partial\Omega_j)$, where the nodes are listed in, for example, clockwise order. Given the nodal values

$$w \equiv \begin{pmatrix} w_1 \\ \vdots \\ w_r \end{pmatrix}$$

of a function $W \in S_h(\partial\Omega_j)$, the nodal values

$$v \equiv \begin{pmatrix} v_1 \\ \vdots \\ v_r \end{pmatrix}$$

of the function $V = lW$ satisfy

$$Mv = Nw,$$

where

$$(4.2) \quad N_{pq} = \langle l\Phi_p, \Phi_q \rangle_{\partial\Omega_k} \quad \text{and} \quad M_{pq} = \langle \Phi_p, \Phi_q \rangle_{\partial\Omega_k}.$$

In this case of equally spaced nodes, the matrices N and M are simultaneously diagonalizable. The eigenvectors are

$$(4.3) \quad \Psi_p = \begin{pmatrix} \exp\left(\frac{2\pi i p}{r}\right) \\ \exp\left(\frac{2\pi i 2p}{r}\right) \\ \vdots \\ \exp\left(\frac{2\pi i rp}{r}\right) \end{pmatrix} \quad \text{for } p = 1, \dots, r.$$

Here i is the square root of minus one. The corresponding eigenvalues are given by

$$\lambda_p^M = \left(4 + 2 \cos\left(\frac{2\pi p}{r}\right) \right) \frac{h}{6}$$

and

$$\lambda_p^N = \left(2 - 2 \cos\left(\frac{2\pi p}{r}\right) \right) / h.$$

It obviously follows that the eigenvalues for the matrix

$$(4.4) \quad L_{pq} = \langle l^{1/2} \Phi_p, \Phi_q \rangle_{\partial\Omega_k}$$

are given by

$$(4.5) \quad \lambda_p^L = \sqrt{\frac{(2 - 2 \cos(\frac{2\pi p}{r}))(4 + 2 \cos(\frac{2\pi p}{r}))}{6}}.$$

Thus one can compute (4.1) by first expanding the nodal values in the basis of eigenvectors (4.3), multiplying by the eigenvalues (4.5) and then computing the nodal values of the resulting expansion. Note that the transformation from nodal values to coordinates in the eigenvector basis (and vice versa) can be computed in $O(r \ln r)$ operations by use of the fast Fourier transform.

Remark 4.1. From the above discussion we see that the amount of work required to evaluate $\langle Q\zeta, \cdot \rangle_\gamma$ is $O(\ln(d/h)/dh)$. For reasonable domain subdivision strategies, the work involved in evaluating \tilde{Q}^{-1} is also $O(\ln(d/h)/dh)$. Thus, the amount of work required for evaluating \tilde{Q}^{-1} is $O(m \ln(d/h)/dh)$. This quantity is usually bounded by Ch^{-2} (see Proposition 3.1).

We next consider the extension of the techniques of Section 2 to the case where the nodes on $\partial\Omega_k$ are not uniformly spaced (with respect to arc length). Assume that there are r nodes on $\partial\Omega_k$. Let R_k be a rectangular mesh with a boundary which has an equally spaced mesh of r nodes. There exists a piecewise linear map $T_k : \partial\Omega_k \mapsto \partial R_k$ which takes the mesh of $\partial\Omega_k$ onto that of ∂R_k . We then define

$$(4.6) \quad \langle \tilde{l}^{1/2} V, V \rangle_{\partial\Omega_k} \equiv \langle l^{1/2} \tilde{V}, \tilde{V} \rangle_{\partial R_k} \quad \text{for all } V \in S_h(\partial\Omega_k),$$

where $\tilde{V} = V \circ T_k$. The form Q is defined by

$$(4.7) \quad \langle QV, V \rangle_\Gamma = \sum_k \alpha_k \langle \tilde{l}^{1/2} V, V \rangle_{\partial\Omega_k}.$$

All of the constructions of Section 2 now go through. Indeed, we decompose $W = W_P + W_H$ and define B by (2.13), (2.11) using Q given by (4.7). The algorithm for computing the solution of (1.5) is completely analogous to that described in Section 2. By (4.6), the evaluation of $\langle \tilde{l}^{1/2} \zeta, \cdot \rangle_{\partial\Omega_k}$ may be implemented exactly as described in the first part of this section, i.e., we use the procedure given immediately after (4.5).

Finally, we note that in order to get a preconditioner for A , we can apply our techniques to any other comparable form \tilde{A} . The form \tilde{A} is chosen for computational convenience. For example, \tilde{A} can be chosen so that it can be ‘fast solved’ even when A corresponds to a variable coefficient operator on a nonuniform mesh as described in Section 4 of [4].

5. Numerical Experiments. In this section we shall present some results of numerical experiments which illustrate the convergence properties of the preconditioning methods of this paper. We use (2.13) as a preconditioner in conjunction with the conjugate gradient method. To help illustrate the differences in performance

between the preconditioners of [4], [5] and that of this paper, we will consider the same basic set of examples as those given in Parts I and II. We shall report many of the same parameters as given in Parts I and II. We shall for example, compute the condition number K of the preconditioned system**. In some examples, we shall also report n , the number of iterations required to reduce the matrix norm $(Ax \cdot x)^{1/2}$ of the error $E_n = U - U_n$ by a specific factor. Here U is a randomly generated solution of the matrix equations normalized so that $-1 \leq U \leq 1$ and U_n is the approximation to U obtained using n steps of the iterative algorithm. In addition, we shall include spectral bounds K_b for the boundary operator $\tilde{Q}^{-1}Q$ and the degree m of the polynomial P_m .

The examples were chosen to illustrate the effectiveness of the algorithm on problems with both smooth and discontinuous coefficients on domains with different geometries. In all of these examples, subspaces $S_h^0(\Omega)$ of piecewise linear functions defined on a quasi-uniform mesh of size h were used and the algorithm was applied to solve the finite element equations approximating the solution of an elliptic problem of the form (1.1). The procedure discussed in Section 4 of Part I for choosing the coefficients of the preconditioning form and solving the related subproblems was used throughout this section. In all examples, estimates for the largest and smallest eigenvalue of $\tilde{Q}^{-1}Q$ were computed by the power method. These estimates were used for λ_0 and λ_1 in (2.19). The degree m of the polynomial P_m was usually taken to be the greatest integer less than or equal to $1 + \sqrt{K_b}$ where $K_b \equiv \sqrt{\lambda_1/\lambda_0}$.

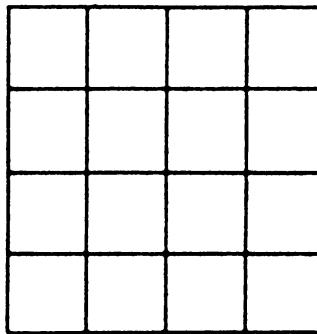


FIGURE 5.1
Subdivision of the square.

Example 1. For our first example we take $L = -\Delta$, the Laplace operator (i.e., $a_{11} = a_{22} = 1$ and $a_{12} = a_{21} = 0$), Ω the unit square and a regular rectangular mesh of size h . Note that, although in this very simple case the resulting equations may be solved rapidly on a serial machine by a variety of ‘fast’ methods, the algorithms of Part I and II would be particularly appealing for a machine with parallel architecture. We will also use this example as a benchmark for the more complicated examples to follow. We subdivide the domain Ω into sixteen subregions as indicated in Figure 5.1.

Table 5.1 illustrates the iterative reduction rates for Example 1 when $h = 1/32$. The largest and smallest eigenvalues for $\tilde{Q}^{-1}Q$ were .36 and 1.3, respectively, and m was taken to be equal to 2. The table lists the total reduction and average reduction

**The condition number K is defined to be β_1/β_0 where β_0 and β_1 are defined to be, respectively, the maximum and minimum constants satisfying (1.6).

rate as a function of the number of iterations in the matrix norm $(Ax \cdot x)^{1/2}$ and the maximum norm. These reductions are normalized so that the initial error is unity. We see, for example, that a reduction of .0001 in the A -norm (resp. maximum norm) requires only 7 (resp. 9) iterations.

TABLE 5.1
Iterative convergence for Example 1.

Iteration	A -error	Average Reduction	Max-error	Max-error Average Reduction
1	1.7×10^{-1}	.17	7.8×10^{-1}	.78
2	2.9×10^{-2}	.17	3.2×10^{-1}	.57
3	1.5×10^{-2}	.24	1.5×10^{-1}	.54
4	3.5×10^{-3}	.24	2.7×10^{-2}	.40
5	8.0×10^{-4}	.24	7.7×10^{-3}	.38
6	2.9×10^{-4}	.26	2.7×10^{-3}	.37
7	8.9×10^{-5}	.26	1.3×10^{-3}	.39
8	3.9×10^{-5}	.28	4.9×10^{-4}	.39
9	7.9×10^{-6}	.27	6.1×10^{-5}	.34
10	1.7×10^{-6}	.27	1.6×10^{-5}	.33
11	5.6×10^{-7}	.27	7.0×10^{-6}	.34

To more fully illustrate the convergence behavior of the method on this problem, we consider Table 5.2, which gives the condition number and theoretical reduction*** for Example 1 as a function of the mesh size h . We note that the theoretical reduction gives a pessimistic bound on the worst case convergence in the A -norm. For example, the actual reduction rate given in Table 5.1 for 11 iterations was .27, which is considerably better than the theoretical rate of .32 given in Table 5.2 for $h = 1/32$.

TABLE 5.2
Convergence for Example 1.

h	K_b	m	K	ρ	n
$1/8$	1.8	2	2.3	.21	6
$1/16$	2.6	2	3.0	.27	7
$1/32$	3.6	2	3.7	.32	7
$1/64$	5.0	3	3.2	.28	6
$1/128$	6.2	3	3.5	.30	6

In the next table, we consider the effect that the degree of the polynomial P_m has on the rate of convergence of the preconditioned algorithm. Table 5.3 gives the number of iterations n required to reduce the A -norm error by .0001 and the observed average reduction (in the A -norm) per iteration as a function of m . Clearly, as m tends to infinity, the operator \bar{Q} tends to Q . Table 5.3 suggests that the methods converge rapidly, even for small values of m , and shows that very little

***It is well known (cf. [12]) that the error for preconditioned conjugate gradient iteration satisfies $(AE_n \cdot E_n) \leq 4\rho^{2n}(AE_0 \cdot E_0)$, where the reduction factor ρ is given by $\rho \equiv (\sqrt{K} - 1)/(\sqrt{K} + 1)$.

$\mu=300$	$\mu=0.0001$	$\mu=31400$	$\mu=5$
$\mu=0.05$	$\mu=8$	$\mu=0.07$	$\mu=2700$
$\mu=10^{-6}$	$\mu=0.1$	$\mu=200$	$\mu=9$
$\mu=1$	$\mu=6000$	$\mu=4$	$\mu=140000$

FIGURE 5.2
The coefficients for Example 2.

improvement (in the convergence rate of the preconditioned algorithm) results from a more accurate approximation of Q^{-1} . The results given in the table correspond to $h = 1/32$; similar results were obtained for other values of h .

TABLE 5.3
*Convergence of the preconditioned algorithm
as a function of m for Example 1.*

m	K	Observed Reduction	n
1	7.5	.33	9
2	3.7	.27	7
3	2.8	.21	6
4	2.9	.21	6
8	2.8	.21	6

Example 2. In this example, Ω is the unit square and the subdomains were taken as in Example 1 (see Figure 5.1). The operator L is taken to have coefficients which have discontinuities across the subdomain boundaries. More specifically, we take $a_{11} = a_{22} = \mu$ and $a_{12} = a_{21} = 0$, where μ is the randomly chosen piecewise constant function on the subdomains as indicated in Figure 5.2. Table 5.4 gives the results for the condition number of the preconditioned system and the theoretical reduction factors for this example as a function of h .

TABLE 5.4
Convergence results for Example 2.

h	K_b	m	K	ρ	n
1/8	1.9	2	2.3	.21	6
1/16	2.7	2	3.1	.27	6
1/32	3.9	2	3.8	.32	6
1/64	5.0	3	3.4	.29	6
1/128	6.4	3	3.8	.32	6

Note that the results differ only slightly from those given for the Laplacian in Table 5.2. We remark that similar results were obtained in tests with other

randomly chosen coefficients. This indicates that the iterative method of this paper will be extremely effective on interface problems, even when the coefficients change drastically across interfaces, as long as the subdomain boundaries align with the interface boundaries.

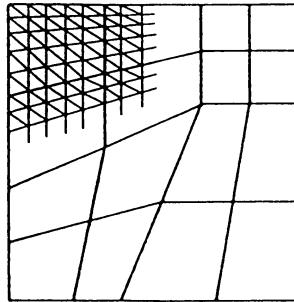


FIGURE 5.3
The irregular geometry of Example 3.

Example 3. In this example, we consider an interface problem where the interface separates two domains with irregular geometries. The domain Ω is again the unit square subdivided into sixteen subdomains as illustrated in Figure 5.3. The space $S_h^0(\Omega)$ is taken to be the piecewise linear functions defined on the irregular mesh roughly exemplified by the lighter lines. Again the coefficients of L are piecewise constant functions defined by $a_{11} = a_{22} = \mu$ and $a_{12} = a_{21} = 0$, where μ is given by Figure 5.4.

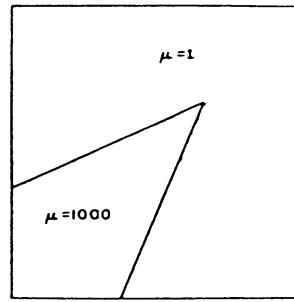


FIGURE 5.4
The coefficients of Example 3.

Results for this problem are given in Table 5.5. A comparison with Table 5.2 indicates that the irregular geometry of this example increased the condition number only by at most a factor of three. This results in less than a factor of two increase in the number of iterations required for a given accuracy. Here again, m was equal to two or three.

TABLE 5.5
Convergence results for Example 3.

h	K_b	K	ρ	Observed Reduction	n
1/8	1.8	4.9	.38	.29	9
1/16	2.6	7.6	.47	.40	11
1/32	3.6	9.9	.52	.45	12
1/64	4.9	8.9	.50	.42	11

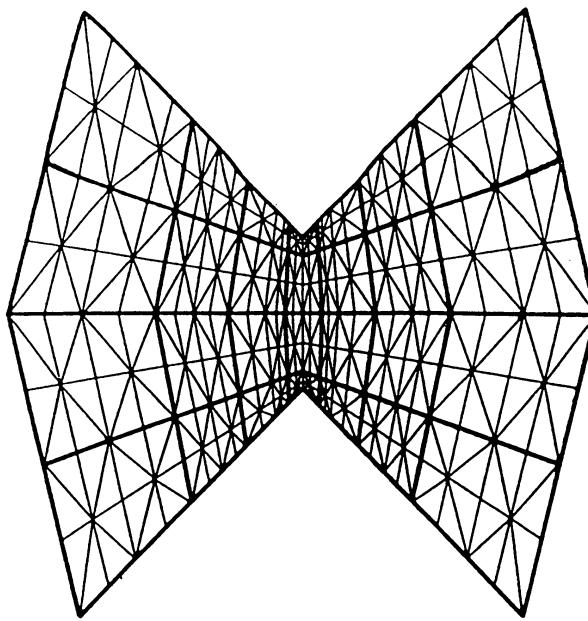


FIGURE 5.5
The mesh and subdomain structure for Example 4.

Example 4. In this example, we illustrate the present algorithm applied to the solution of a problem on a polygonal domain with nonconvex corners. The mesh and subdomain structure were chosen as illustrated in Figure 5.5. Note the mild refinement near the nonconvex corners of the domain. For the operator L we use the Laplacian as in Example 1. The results for this case are given in Table 5.6.

TABLE 5.6
Convergence results for Example 4.

Number of Unknowns	K_b	K	ρ	Observed Reduction	n
405	2.8	4.4	.35	.35	9
1705	3.8	5.8	.41	.40	10
6993	5.2	5.3	.40	.37	10

Example 5. As a final example, we illustrate the algorithm described in Remark 3.2, i.e., we consider the case where $\tilde{Q} = I$. We consider the problem and domain decomposition of Example 2. Table 5.7 gives the condition number K_b of \bar{Q} , n , K , and the observed reduction in the A -norm as a function of h . In this case, we increased m as suggested by Proposition 3.1 (i).

TABLE 5.7
Convergence results for Example 5.

h	K_b	m	K	Observed Reduction	n
1/8	11.5	4	2.4	.21	6
1/16	25	5	3.2	.25	7
1/32	52	8	3.3	.31	7
1/64	105	11	4.3	.31	8

Department of Mathematics
Cornell University
Ithaca, New York 14853
E-mail: bramble@mathvax.msi.cornell.edu

Brookhaven National Laboratory
Upton, New York 11973
E-mail: pasciak@bnl.apa

Department of Mathematics
Cornell University
Ithaca, New York 14853

1. P. E. BJØRSTAD & O. B. WIDLUND, "Solving elliptic problems on regions partitioned into substructures," *Elliptic Problem Solvers II* (G. Birkhoff and A. Schoenstadt, eds.), Academic Press, New York, 1984, pp. 245–256.
2. P. E. BJØRSTAD & O. B. WIDLUND, "Iterative methods for the solution of elliptic problems on regions partitioned into substructures," *SIAM J. Numer. Anal.*, v. 23, 1986, pp. 1097–1120.
3. J. H. BRAMBLE, J. E. PASCIAK & A. H. SCHATZ, "An iterative method for elliptic problems on regions partitioned into substructures," *Math. Comp.*, v. 46, 1986, pp. 361–369.
4. J. H. BRAMBLE, J. E. PASCIAK & A. H. SCHATZ, "The construction of preconditioners for elliptic problems by substructuring, I," *Math. Comp.*, v. 47, 1986, pp. 103–134.
5. J. H. BRAMBLE, J. E. PASCIAK & A. H. SCHATZ, "The construction of preconditioners for elliptic problems by substructuring, II," *Math. Comp.*, v. 49, 1987, pp. 1–16.
6. B. L. BUZZBEE & F. W. DORR, "The direct solution of the biharmonic equation on rectangular regions and the Poisson equation on irregular regions," *SIAM J. Numer. Anal.*, v. 11, 1974, pp. 753–763.
7. B. L. BUZZBEE, F. W. DORR, J. A. GEORGE & G. H. GOLUB, "The direct solution of the discrete Poisson equation on irregular regions," *SIAM J. Numer. Anal.*, v. 8, 1971, pp. 722–736.
8. Q. V. DIHN, R. GLOWINSKI & J. PÉRIAUX, "Solving elliptic problems by domain decomposition methods," in *Elliptic Problem Solvers II* (G. Birkhoff and A. Schoenstadt, eds.), Academic Press, New York, 1984, pp. 395–426.
9. G. H. GOLUB & D. MEYERS, *The Use of Preconditioning Over Irregular Regions*, Proc. 6th Internat. Conf. Comput. Meth. Sci. and Engrg., Versailles, 1983.
10. J. L. LIONS & E. MAGENES, *Problèmes aux Limites non Homogènes et Applications*, Dunod, Paris, 1968.
11. J. NEČAS, *Les Méthodes Directes en Théorie des Équations Elliptiques*, Academia, Prague, 1967.
12. W. M. PATTERSON, 3rd, *Iterative Methods for the Solution of a Linear Operator Equation in Hilbert Space—A Survey*, Lecture Notes in Math., vol. 394, Springer-Verlag, New York, 1974.

*4.4. THE CONSTRUCTION OF PRECONDITIONERS FOR ELLIPTIC
PROBLEMS BY SUBSTRUCTURING. IV*

**4.4 The construction of preconditioners for elliptic problems by
substructuring. IV**

The construction of preconditioners for elliptic problems by substructuring. IV
[22]

The Construction of Preconditioners for Elliptic Problems by Substructuring, IV*

By James H. Bramble, Joseph E. Pasciak, and Alfred H. Schatz

Abstract. We consider the problem of solving the algebraic system of equations which result from the discretization of elliptic boundary value problems defined on three-dimensional Euclidean space. We develop preconditioners for such systems based on substructuring (also known as domain decomposition). The resulting algorithms are well suited to emerging parallel computing architectures. We describe two techniques for developing these preconditioners. A theory for the analysis of the condition number for the resulting preconditioned system is given and the results of supporting numerical experiments are presented.

1. Introduction. The aim of this series of papers is to propose and analyze methods for efficiently solving the equations resulting from finite element discretizations of second-order elliptic boundary value problems on general domains in R^2 and R^3 . In particular, we shall be concerned with constructing computationally “effective” preconditioners for these discrete equations which can be used in a preconditioned iterative algorithm to define a rapid solution method. The methods developed are well suited to parallel computing architectures.

In Part I, [6], a flexible domain decomposition algorithm for the two-dimensional problems was developed and analyzed. This algorithm had the novel feature that it enabled subdivision into an arbitrary number of subdomains without the deterioration of the resulting iterative convergence rates. This property has important implications in parallel applications since for this type of algorithm, the number of subdomains is proportional to the number of parallel tasks.

In Parts II, [7] and III, [8], we extended the domain decomposition techniques along two directions. In Part II, we developed some simplified domain decomposition strategies for two- and three-dimensional problems, including a class of singularly perturbed systems which occur in parabolic timestepping applications. In Part III, we introduced a technique for two-dimensional problems which gave rise to domain decomposition strategies whose convergence rates stayed bounded independently of both the subdomain size d and the mesh size h .

Received September 9, 1987; revised April 20, 1988.

1980 *Mathematics Subject Classification* (1985 Revision). Primary 65N30, 65F10.

*This manuscript has been authored under contract number DE-AC02-76CH00016 with the U.S. Department of Energy. Accordingly, the U.S. Government retains a non-exclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. This work was also supported in part under the National Science Foundation Grant No. DMS84-05352 and under the Air Force Office of Scientific Research, Contract No. ISSA86-0026 and by the U.S. Army Research Office through the Mathematical Science Institute, Cornell University.

©1989 American Mathematical Society
0025-5718/89 \$1.00 + \$.25 per page

In this part of the series, we will develop two domain decomposition algorithms for problems in three dimensions. We shall present a general theoretical approach for the analysis of such methods. These methods lead to preconditioned systems with condition number bounded by $c(1 + \ln^2(d/h))$. In contrast, the simplified strategies of Part II give rise to a condition number bounded by cd/h .

Let Ω be a bounded domain in R^3 with boundary $\partial\Omega$. As a model problem for a second-order uniformly elliptic equation, we shall consider the Dirichlet problem

$$(1.1) \quad \begin{aligned} Lu &= f && \text{in } \Omega, \\ u &= 0 && \text{on } \partial\Omega, \end{aligned}$$

where

$$Lv = - \sum_{i,j=1}^3 \frac{\partial}{\partial x_i} \left(a_{ij} \frac{\partial v}{\partial x_j} \right)$$

with $\{a_{ij}\}$ uniformly positive definite, bounded and piecewise smooth on Ω .

In this paper, we shall develop and analyze preconditioners for the matrices which result from finite element and finite difference discretization of (1.1). This is most naturally carried out from the finite element point of view. Accordingly, we shall proceed with the general finite element framework with a detailed formulation of both cases considered in Section 2.

The generalized Dirichlet form corresponding to (1.1) is given by

$$(1.2) \quad A(v, w) = \sum_{i,j=1}^3 \int_{\Omega} a_{ij} \frac{\partial v}{\partial x_i} \frac{\partial w}{\partial x_j} dx,$$

which is defined for all v and w in the Sobolev space $H^1(\Omega)$ (the space of distributions with square-integrable first derivatives). The $L^2(\Omega)$ inner product is denoted

$$(v, w) = \int_{\Omega} vw dx.$$

The subspace $H_0^1(\Omega)$ is the completion of the smooth functions with support in Ω , with respect to the norm in $H^1(\Omega)$. The weak formulation of the problem defined by (1.1) is: Find $u \in H_0^1(\Omega)$ such that

$$(1.3) \quad A(u, w) = (f, w) \quad \text{for all } w \in H_0^1(\Omega).$$

This leads immediately to the standard Galerkin approximation. Let $S_h(\Omega)$ be a finite-dimensional subspace of $H_0^1(\Omega)$. The Galerkin approximation is defined as the solution of the following problem: Find $U \in S_h(\Omega)$ such that

$$(1.4) \quad A(U, \Phi) = (f, \Phi) \quad \text{for all } \Phi \in S_h(\Omega).$$

The underlying method which we will consider is a preconditioned iterative method. As explained in Part I, the task of defining a preconditioner for the matrix problem corresponding to (1.4) is the same as that of defining another positive definite form $B(\cdot, \cdot)$ on $S_h(\Omega) \times S_h(\Omega)$. The importance of making a “good” choice for B is well known. The form B will define a good preconditioner provided it has two basic properties. First, the problem of finding the function $W \in S_h(\Omega)$ satisfying

$$(1.5) \quad B(W, \Phi) = G(\Phi) \quad \text{for all } \Phi \in S_h(\Omega),$$

for a given linear functional G , should be more economical to solve on a given computer architecture than (1.4). Secondly, B should be spectrally close to A in the sense that there are positive numbers β_0 and β_1 satisfying

$$(1.6) \quad \beta_0 B(V, V) \leq A(V, V) \leq \beta_1 B(V, V) \quad \text{for all } V \in S_h(\Omega),$$

where the ratio β_1/β_0 is not too large.

We will define the preconditioning form B using domain decomposition and ‘mapping’ techniques. The domain $\bar{\Omega}$ is written as a union of subdomains $\bigcup \bar{\Omega}_i$. The mesh on each subdomain is assumed to be related to the mesh on the reference cube $\hat{\Omega}$ under a transformation T_i . The framework developed in this paper reduces the task of defining domain decomposition preconditioners on Ω to a problem of defining an appropriate form Q acting on subspaces of functions defined on the boundary of the reference domain. A consequence of this approach is that the most significant part of the analysis need only be carried out on the reference domain in conjunction with a reference subspace.

The outline of the remainder of the paper is as follows. In Section 2, we describe the finite element and finite difference discretizations. We also give the assumptions on the subspaces on Ω and $\hat{\Omega}$. In Section 3, we show how the construction of the preconditioner B can be reduced to the definition of an appropriate form Q . The preconditioner B is described in terms of Q in this section. In Section 4, we develop two forms $Q = Q_1$ and $Q = Q_2$ which lead to different domain decomposition algorithms. It is shown that (1.6) holds with

$$\beta_1/\beta_0 \leq c(1 + \ln^2(d/h))$$

for the domain decomposition form B resulting from either of these two forms. Here, d is roughly the domain size and h is the mesh size. The most computationally effective preconditioner results from the form Q_1 . We describe the algorithm for the solution of (1.5) in Section 5. Finally, in Section 6, we give the results of numerical experiments for some three-dimensional problems.

For earlier papers dealing with domain decomposition techniques applied to the solution of the linear systems resulting from numerical approximation of boundary value problems see [2], [5], [9], [10]. The obvious generalizations of these methods lead to preconditioned systems whose condition number increases with the number of subdomains. Thus, these methods may not lead to effective algorithms on parallel computers. For some numerical results for domain decomposition methods on parallel computers see [11]. Additional papers and references for recent work on domain decomposition can be found in the proceedings to be published by SIAM of the ‘First International Symposium on Domain Decomposition Methods for Partial Differential Equations’ held in Paris 1987.

A domain decomposition technique which is well suited to applications with refinements is developed in [4]. The resulting algorithms are sometimes the same as those developed with the FAC approach of [15] which represent yet another technique for developing domain decomposition-like algorithms for refinement problems.

Before proceeding, we give some notation. In what follows, edges, faces, and subdomains will be open sets in R^1 , R^2 , and R^3 , respectively. Let $\tilde{\Omega}$ be a generic domain in R^j for $j = 1, 2, 3$. For nonnegative s , the Sobolev space of order s on

$\tilde{\Omega}$ will be denoted $H^s(\tilde{\Omega})$ (cf. [14], [16]). The norm on $H^s(\tilde{\Omega})$ will be denoted by $\|\cdot\|_{s,\tilde{\Omega}}$ when $j = 3$ and $|\cdot|_{s,\tilde{\Omega}}$ when $j = 1, 2$. The L^2 inner products and norms will be denoted

$$(u, v)_{\tilde{\Omega}} = \int_{\tilde{\Omega}} uv \, dx$$

and $\|u\|_{\tilde{\Omega}} = (u, u)_{\tilde{\Omega}}^{1/2}$ when $j = 3$, and

$$\langle u, v \rangle_{\tilde{\Omega}} = \int_{\tilde{\Omega}} uv \, dx$$

and $|u|_{\tilde{\Omega}} = \langle u, u \rangle_{\tilde{\Omega}}^{1/2}$ when $j = 1, 2$.

Throughout this paper, c and C , with or without subscripts, will denote positive constants which are independent of the subdivision, d and h . These constants may take on different values in different places.

2. Discretization of (1.1). In this section, we shall formulate the finite element and finite difference methods to be considered. In order to do so, we shall first describe our assumptions on the domain Ω and its decomposition $\bar{\Omega} = \bigcup \bar{\Omega}_i$. We next use this decomposition to define the finite element and finite difference approximations. Finally, we describe some norms which will play an essential role in the analysis of the preconditioners to be developed in later sections.

An important aspect of this paper is to reduce the problem of defining domain decomposition algorithms on the union of subdomains to a problem on the unit cube $\hat{\Omega}$ with respect to a reference subspace. The faces of $\hat{\Omega}$ will be denoted $\hat{\Gamma}_i^f$ for $i = 1, \dots, 6$. In addition, the union of the closures of the edges of $\hat{\Omega}$ will be denoted by $\hat{\Gamma}^e$.

We make the following assumptions with respect to the domain Ω :

- (A.1) Ω can be subdivided into m subdomains $\bar{\Omega} = \bigcup \bar{\Omega}_i$ with $\Omega_i \cap \Omega_j = \emptyset$ for $i \neq j$.
- (A.2) These subdomains are related to the unit cube in that for each i there is an orientation-preserving trilinear mapping T_i which takes $\hat{\Omega}$ onto Ω_i . We assume that there exists a positive constant d such that

$$(2.1) \quad d^{-1} |DT_i(x)| \leq C \quad \text{for all } x \in \hat{\Omega}$$

and

$$(2.2) \quad d |DT_i^{-1}(x)| \leq C \quad \text{for all } x \in \Omega_i.$$

Here, $DT_i(x)$ is the Jacobian matrix of T_i at x . In (2.1) and (2.2), $|\cdot|$ denotes the matrix norm. Note that the subdomains are roughly of size d .

- (A.3) The set of faces of Ω_i is denoted by $\{\Gamma_{ij}^f\}$, where Γ_{ij}^f is defined to be the image of the j th face of $\partial\hat{\Omega}$ under T_i . Furthermore, we require that if two faces, Γ_{ij}^f and Γ_{kl}^f , share a common point x , then $\Gamma_{ij}^f = \Gamma_{kl}^f$ and

$$T_i^{-1}(x) = T_{ik}^r \circ T_k^{-1}(x) \quad \text{for all } x \in \Gamma_{ij}^f,$$

where T_{ik}^r is a rigid body rotation of $\hat{\Omega}$.

We next consider the definition of the approximation procedures. For simplicity of presentation, we shall consider particular finite element and finite difference

applications. Many generalizations are possible. For either procedure, we partition the unit cube into $m_1 \times m_2 \times m_3$ regular rectangular parallelepipeds and define $S_{\hat{h}}(\hat{\Omega})$ to be the functions which are continuous on $\hat{\Omega}$ and piecewise trilinear with respect to this partition. The reference mesh size \hat{h} is defined to be

$$\hat{h} = \max(1/m_1, 1/m_2, 1/m_3).$$

We assume that

$$\min(1/m_1, 1/m_2, 1/m_3) \geq C\hat{h}.$$

We first consider the finite element case. Let $h = d\hat{h}$ and define

$$S_h(\Omega_i) = \{\phi = \psi \circ T_i^{-1} | \phi = 0 \text{ on } \partial\Omega \text{ and } \psi \in S_{\hat{h}}(\hat{\Omega})\}.$$

Define the map $I_i: S_h(\Omega_i) \mapsto S_{\hat{h}}(\hat{\Omega})$ by $I_i V = V \circ T_i$. Define the space $S_h(\Omega)$ to be the set of continuous functions on $\bar{\Omega}$ whose restrictions on each subdomain Ω_i are functions in $S_h(\Omega_i)$. We assume that

$$(A.4) \quad S_h(\Omega_i) = \{\phi|_{\Omega_i} \text{ for } \phi \in S_h(\Omega)\}.$$

This implies that the boundary nodes of $S_h(\Omega_i)$ and $S_h(\Omega_j)$ coincide on common faces.

Note that we obviously have

$$A(V, W) = \sum_{i=1}^m A_i(V, W)$$

where

$$A_i(V, W) = \sum_{j,k=1}^3 \int_{\Omega_i} a_{jk} \frac{\partial V}{\partial x_j} \frac{\partial W}{\partial x_k} dx,$$

and m is defined in (A.1). The finite element approximation to the solution u of (1.1) is the function $U \in S_h(\Omega)$ satisfying (1.4). We will derive preconditioners for this problem.

We next define the finite difference approximation. We only consider the case when L is given by

$$Lv = -\nabla \cdot a \nabla v.$$

Assume that (A.1), (A.2) and (A.3) hold and furthermore that the mappings T_i are simply dilatation and translation with respect to the coordinate axes. Also assume that we have a regular grid of nodal points $\{p_j\}_{j=1}^N$ on a mesh of size h defined on $\bar{\Omega}$. We label these nodes so that $\{p_j\}_{j=1}^N$ are the nodes in Ω and assume that the nodal points of $\bar{\Omega}_i$ coincide with the image of the nodal points (corresponding to the subspace $S_{\hat{h}}(\hat{\Omega})$ defined above) of $\hat{\Omega}$ under T_i . The space $S_h(\Omega)$ consists of N -dimensional vectors of nodal values at the nodes of Ω . The subspace $S_h(\Omega_i)$ consists of the nodal values at nodes in $\Omega \cap \bar{\Omega}_i$. The map $I_i: S_h(\Omega_i) \mapsto S_{\hat{h}}(\hat{\Omega})$ is defined by interpolation, i.e., $I_i V$ is the function in $S_{\hat{h}}(\hat{\Omega})$ defined by

$$I_i V(p) = \begin{cases} V(T_i(p)) & \text{when } T_i(p) \text{ is a node of } \Omega \cap \bar{\Omega}_i, \\ 0 & \text{for the remaining nodes of } \hat{\Omega}. \end{cases}$$

Note that $S_{\hat{h}}(\hat{\Omega})$ is a finite element subspace of trilinear functions, even when we are using the finite difference approximation on Ω .

Let \mathcal{N}_i be the list of neighbors for $S_h(\Omega_i)$, i.e., $(k, l) \in \mathcal{N}_i$ if and only if p_k, p_l are nodal points in $\bar{\Omega}_i$ which are a distance of h apart. Let $p_{k,l}$ be the midpoint between p_k and p_l and set

$$A_i(V, W) = h^3 \sum_{(k,l) \in \mathcal{N}_i} w_{kl} a(p_{k,l}) \frac{(V(p_k) - V(p_l))(W(p_k) - W(p_l))}{h^2},$$

where we set $V(p_l) \equiv W(p_l) \equiv 0$ for nodes p_l on $\partial\Omega$. Here, w_{kl} is the weight function defined by

$$w_{kl} = \begin{cases} 1 & \text{if the line segment between } p_k \text{ and } p_l \text{ is in } \Omega_i, \\ 1/2 & \text{if the line segment between } p_k \text{ and } p_l \text{ is in some face of } \Omega_i, \\ 1/4 & \text{if the line segment between } p_k \text{ and } p_l \text{ is in an edge of } \partial\Omega_i. \end{cases}$$

For functions $V, W \in S_h(\Omega)$ define

$$(2.3) \quad A(V, W) = \sum_{i=1}^m A_i(V, W).$$

The finite difference approximation to the solution u of (1.1) at the nodes is the function $U \in S_h(\Omega)$ satisfying

$$(2.4) \quad A(U, \Phi) = F \cdot \Phi \quad \text{for all } \Phi \in S_h(\Omega),$$

where F is the vector $\{h^3 f(p_k)\}$ and \cdot denotes the usual Euclidean inner product.

Remark 2.1. By summation by parts, it is not difficult to see that the form A can be written

$$A(V, W) = (L_h V) \cdot W,$$

where L_h is the usual second-order 7-point difference operator multiplied by h^3 . Thus, the solution U of (2.4) is the standard finite difference approximation to the solution of (1.1). We have taken the above approach for developing these equations because it naturally gives rise to the decomposition of the form given by (2.3).

Remark 2.2. For both finite element and finite difference discretizations, we allow for the case when each subdomain $S_h(\Omega_i)$ has a different number of nodes. Accordingly, the reference subspace may differ with i . We have suppressed this dependence in the notation for convenience.

We finish this section with some additional notation. Let $\Gamma = \bigcup \partial\Omega_i$ and $S_h(\Gamma)$ be the space of functions which are restrictions of those in $S_h(\Omega)$ to Γ . Let $S_h^0(\Omega_i)$ be the subspace of $S_h(\Omega_i)$ of functions which vanish on $\partial\Omega_i$. Finally, let $S_{\hat{h}}(\partial\hat{\Omega})$ denote the space of restrictions of the functions of $S_{\hat{h}}(\hat{\Omega})$ to $\partial\hat{\Omega}$ and $S_{\hat{h}}^0(\hat{\Omega})$ denote the space of functions in $S_{\hat{h}}(\hat{\Omega})$ which vanish on $\partial\hat{\Omega}$.

3. A General Construction of $B(\cdot, \cdot)$. We will define our domain decomposition form by replacing the terms $A_i(V, W)$ in (2.3). To do this, we decompose an arbitrary function $W \in S_h(\Omega_i)$ into $W = W_P + W_H$, where $W_P \in S_h^0(\Omega_i)$ and

$$(3.1) \quad A_i(W_H, \Phi) = 0 \quad \text{for all } \Phi \in S_h^0(\Omega_i).$$

W_H is the unique function in $S_h(\Omega_i)$ which equals W on $\partial\Omega_i$ and satisfies (3.1). Such a function will be called ‘discrete A_i -harmonic.’ A consequence of (3.1) is that

$$(3.2) \quad A_i(W, W) = A_i(W_P, W_P) + A_i(W_H, W_H).$$

To define our preconditioner B , we shall replace the term $A_i(W_H, W_H)$ above.

We note that assumptions (2.1) and (2.2) imply

$$(3.3) \quad cA_i(V, V) \leq d\delta_i D(I_i V, I_i V) \leq CA_i(V, V) \quad \text{for all } V \in S_h(\Omega_i),$$

in the finite element case. Here $D(\cdot, \cdot)$ denotes the Dirichlet integral and is defined by

$$D(v, w) \equiv \int_{\hat{\Omega}} \nabla v \cdot \nabla w \, dx.$$

The constant δ_i appearing in (3.3) is a scaling factor. One reasonable choice is to take $\delta_i = (\lambda_1^i + \lambda_0^i)/2$, where λ_1^i and λ_0^i are respectively the largest and smallest eigenvalue of the 3×3 matrix $\{a_{ij}(x_0)\}$ at some point $x_0 \in \Omega_i$. Then the values of c and C appearing in (3.3) only depend on the local variation of the coefficients $\{a_{ij}\}$ on the subregions. It is straightforward to show that (3.3) also holds in the case of finite differences.

The problem of defining a replacement for $A_i(W_H, W_H)$ is thus the same as that of finding one for $dD(I_i W_H, I_i W_H)$. Note that $I_i W_H$ depends only on its boundary values. Accordingly, the form $dD(I_i W_H, I_i W_H)$ can be replaced by a form which explicitly depends only on the boundary values of $I_i W_H$.

To this end, we introduce a bilinear form Q on $S_{\hat{h}}(\partial\hat{\Omega}) \times S_{\hat{h}}(\partial\hat{\Omega})$ and define the form B by

$$(3.4) \quad B(W, W) = \sum_{i=1}^m \{ A_i(W_P, W_P) + d\delta_i Q(I_i W - \gamma_i(W), I_i W - \gamma_i(W)) \},$$

where $\gamma_i(W)$, for each i and W , is the constant function on $\hat{\Omega}$ whose value is determined by

$$(3.5) \quad Q(I_i W - \gamma_i(W), 1) = 0.$$

Notice that the function W_P depends upon i in (3.4). For convenience we have suppressed this dependence in the notation.

Two constructions of Q which lead to effective domain decomposition algorithms will be given in the next section. For the remainder of this section, we assume that such a form has been given which satisfies

$$(3.6) \quad \alpha_0(\hat{h})Q(V, V) \leq |V|_{1/2, \partial\hat{\Omega}}^2 \leq \alpha_1(\hat{h})Q(V, V) \quad \text{for all } V \in S_{\hat{h}}(\partial\hat{\Omega}).$$

For the Q to be defined, the constants $\alpha_0(\hat{h})$ and $\alpha_1(\hat{h})$ can be estimated in terms of \hat{h} (see Proposition 2).

We then have the following proposition.

PROPOSITION 1. *Assume that (3.6) holds. There are constants β_2 and β_3 which do not depend on d or h satisfying*

$$(3.7) \quad \beta_2 \alpha_0(d/h) B(W, W) \leq A(W, W) \leq \beta_3 \alpha_1(d/h) B(W, W) \quad \text{for all } W \in S_h(\Omega).$$

Proof. By (2.3) and (3.4), it suffices to consider a fixed subdomain Ω_i . Let $W \in S_h(\Omega_i)$ be decomposed into $W = W_P + W_H$ as in (3.1). By the definition of B and (3.2), it suffices to show

$$(3.8) \quad d\delta_i \alpha_0(d/h) Q(I_i W - \gamma_i(W), I_i W - \gamma_i(W)) \leq cA_i(W, W)$$

and

$$(3.9) \quad A_i(W_H, W_H) \leq C d \delta_i \alpha_1(d/h) Q(I_i W - \gamma_i(W), I_i W - \gamma_i(W)),$$

where $\gamma_i(W)$ is the constant appearing in (3.5). Let $(I_i W)_H$ be the discrete harmonic function in $S_{\hat{h}}(\hat{\Omega})$ which equals $I_i W$ on $\partial\hat{\Omega}$, i.e., $(I_i W)_H$ is the unique function in $S_{\hat{h}}(\hat{\Omega})$ which equals $I_i W$ on $\partial\hat{\Omega}$ and satisfies the homogeneous equation

$$(3.10) \quad D((I_i W)_H, \Phi) = 0 \quad \text{for all } \Phi \in S_{\hat{h}}^0(\hat{\Omega}).$$

For any constant γ on $\hat{\Omega}$, (3.5) implies

$$Q(I_i W - \gamma_i(W), I_i W - \gamma_i(W)) \leq Q(I_i W - \gamma, I_i W - \gamma).$$

In particular, taking γ to be the average value of $(I_i W)_H$ on $\hat{\Omega}$, it follows from (3.6), the trace theorem and Poincaré's inequality that

$$\begin{aligned} Q(I_i W - \gamma_i(W), I_i W - \gamma_i(W)) &\leq c \alpha_0^{-1}(\hat{h}) D((I_i W)_H, (I_i W)_H) \\ &\leq c \alpha_0^{-1}(\hat{h}) D(I_i W, I_i W). \end{aligned}$$

Inequality (3.8) then follows from (3.3).

We next prove (3.9). We first show that for discrete harmonic functions V ,

$$(3.11) \quad \|V\|_{H^1(\hat{\Omega})}^2 \leq C_1 |V|_{1/2, \partial\hat{\Omega}}^2.$$

By the Poincaré inequality and the minimization property of discrete harmonic functions, (3.11) will follow if we can construct a function $\tilde{W} \in S_{\hat{h}}(\hat{\Omega})$ with the same boundary values as V satisfying

$$(3.12) \quad \|\tilde{W}\|_{H^1(\hat{\Omega})}^2 \leq C_1 |V|_{1/2, \partial\hat{\Omega}}^2.$$

To do this, we use a variation of an argument given in [1]. Let v be the harmonic function on $\hat{\Omega}$ which is equal to V on $\partial\hat{\Omega}$. There exists a function $W \in S_{\hat{h}}(\hat{\Omega})$ (which may differ from V on $\partial\hat{\Omega}$) satisfying

$$(3.13) \quad \|v - W\|_{\hat{\Omega}}^2 + \hat{h}^2 \|v - W\|_{H^1(\hat{\Omega})}^2 \leq C \|v\|_{H^1(\hat{\Omega})}^2.$$

The function W in (3.13) can be taken to be, for example, the $L^2(\hat{\Omega})$ projection of v . We define \tilde{W} to be the function in $S_{\hat{h}}(\hat{\Omega})$ which is equal to V on the nodes of $S_{\hat{h}}(\hat{\Omega})$ which lie on $\partial\hat{\Omega}$ and is equal to W on the nodes of $S_{\hat{h}}(\hat{\Omega})$ which are in the interior of $\hat{\Omega}$. Clearly,

$$\|\tilde{W}\|_{H^1(\hat{\Omega})} \leq \|\tilde{W} - W\|_{H^1(\hat{\Omega})} + \|W\|_{H^1(\hat{\Omega})} \leq c \hat{h}^{-1/2} |v - W|_{\partial\hat{\Omega}} + \|W\|_{H^1(\hat{\Omega})}.$$

By a well-known trace inequality,

$$|v - W|_{\partial\hat{\Omega}}^2 \leq C(\hat{h}^{-1} \|v - W\|_{\hat{\Omega}}^2 + \hat{h} \|v - W\|_{H^1(\hat{\Omega})}^2).$$

Combining the above inequalities with the well-known inequality for harmonic functions,

$$\|v\|_{H^1(\hat{\Omega})} \leq C |V|_{1/2, \partial\hat{\Omega}},$$

completes the proof of (3.12).

Let X be the function in $S_h(\Omega_i)$ satisfying $I_i X = (I_i W)_H$. Note that

$$X = W \quad \text{on } \partial\Omega_i,$$

and hence by (3.1), (3.3),

$$\begin{aligned} A_i(W_H, W_H) &\leq A_i(X, X) \leq Cd\delta_i D((I_i W)_H, (I_i W)_H) \\ &= Cd\delta_i D((I_i W)_H - \gamma_i(W), (I_i W)_H - \gamma_i(W)). \end{aligned}$$

Applying (3.11) gives

$$A_i(W_H, W_H) \leq cd\delta_i |I_i W - \gamma_i(W)|_{1/2, \partial\hat{\Omega}}^2.$$

Inequality (3.9) now follows from (3.6), which completes the proof of Proposition 1.

4. The Construction and Analysis of Q . In this section, we construct and analyze two forms Q_j which give rise to effective domain decomposition preconditioners for three-dimensional problems. It will be shown that for each of these forms, (3.6) holds with $\alpha_1(\hat{h}) \leq C$ and $\alpha_0(\hat{h}) \geq c/(1 + \ln^2(\hat{h}^{-1}))$. Thus, by Proposition 1, the preconditioner B defined by (3.4) using these Q_j will give rise to preconditioned systems with condition number growth bounded by $C_0(1 + \ln^2(d/h))$. As will be demonstrated in Section 5, the first form Q_1 gives rise to a more efficient computational strategy and is hence the preferred method. We include the second form since it is, in some sense, the natural extension of the method of Part I to three dimensions.

We want to derive replacement forms for the norm $|\cdot|_{1/2, \partial\hat{\Omega}}$ on $S_{\hat{h}}(\partial\hat{\Omega})$. As in [14], [16], this norm is given by

$$(4.1) \quad |w|_{1/2, \partial\hat{\Omega}} = \left(\int_{\partial\hat{\Omega}} \int_{\partial\hat{\Omega}} \frac{(w(x) - w(y))^2}{|x - y|^3} ds(x) ds(y) + |w|_{\partial\hat{\Omega}}^2 \right)^{1/2},$$

where s denotes area on $\partial\hat{\Omega}$. Let $\hat{\Gamma}_i^f$ be a face of $\hat{\Omega}$. The space $\dot{H}^{1/2}(\hat{\Gamma}_i^f)$ is defined to be the completion of the smooth functions defined on $\partial\hat{\Omega}$ with support in $\hat{\Gamma}_i^f$ with respect to the norm given by (4.1).

Remark 4.1. It is well known that the space $\dot{H}^{1/2}(\hat{\Gamma}_i^f)$ is the interpolation space which is halfway between $H_0^1(\hat{\Gamma}_i^f)$ and $L^2(\hat{\Gamma}_i^f)$ [14], [16]. For smooth functions with support in $\hat{\Gamma}_i^f$, $\langle -\Delta u, u \rangle_{\hat{\Gamma}_i^f}^{1/2}$ is equivalent to the norm on $H_0^1(\hat{\Gamma}_i^f)$ (here, Δ denotes the two-dimensional Laplacian on the face). Consequently, the completion of the norm given by

$$(4.2) \quad \left(\left\langle (-\Delta)^{1/2} w, w \right\rangle_{\hat{\Gamma}_i^f} \right)^{1/2} \quad \text{for } w \in H_0^1(\hat{\Gamma}_i^f)$$

is equivalent to the norm on $\dot{H}^{1/2}(\hat{\Gamma}_i^f)$.

We shall use a discrete operator $l_0^{1/2}$ which approximates $(-\Delta)^{1/2}$ in the definitions of the computational forms Q_1 and Q_2 . Let

$$S_{\hat{h}}^0(\hat{\Gamma}_i^f) \equiv \{ \phi|_{\hat{\Gamma}_i^f} \text{ such that } \phi \in S_{\hat{h}}(\partial\hat{\Omega}) \text{ and } \phi = 0 \text{ on the edges of } \hat{\Omega} \}.$$

The discrete operator $l_0: S_{\hat{h}}^0(\hat{\Gamma}_i^f) \mapsto S_{\hat{h}}^0(\hat{\Gamma}_i^f)$ is defined by

$$(4.3) \quad \langle l_0 \Psi, \Phi \rangle_{\hat{\Gamma}_i^f} = \int_{\hat{\Gamma}_i^f} \nabla \Psi \cdot \nabla \Phi \, ds \quad \text{for all } \Phi \in S_{\hat{h}}^0(\hat{\Gamma}_i^f).$$

Here, ∇ denotes the two-dimensional gradient on $\hat{\Gamma}_i^f$. The operator l_0 is symmetric positive definite on $S_h^0(\hat{\Gamma}_i^f)$ and $l_0^{1/2}$ is defined to be its positive square root.

Remark 4.2. Note that the discrete operator l_0 is a finite-dimensional approximation to $-\Delta$. It can be shown by interpolation [13, Theorem 9.1] and the inverse assumptions on $S_h^0(\hat{\Gamma}_i^f)$ that

$$(4.4) \quad c |V|_{H^{1/2}(\hat{\Gamma}_i^f)}^2 \leq \left\langle l_0^{1/2} V, V \right\rangle_{\hat{\Gamma}_i^f} \leq C |V|_{H^{1/2}(\hat{\Gamma}_i^f)}^2 \quad \text{for all } V \in S_h^0(\hat{\Gamma}_i^f).$$

The constants c and C in (4.4) can be chosen to be independent of \hat{h} .

We now construct the first form Q_1 .

Method 1. We decompose functions $V \in S_h(\partial\hat{\Omega})$ by $V = V_{e,1} + V_{f,1}$, where $V_{e,1}$ and $V_{f,1}$ satisfy:

- (1) $V_{f,1} = 0$ on $\hat{\Gamma}^e \equiv \bigcup_{i=1}^6 \partial\hat{\Gamma}_i^f$.
- (2) $V_{e,1} = 0$ on all nodes on the faces of $\hat{\Omega}$.

Define

$$(4.5) \quad Q_1(V, V) \equiv \hat{h} \sum_{x_i \in \hat{\Gamma}^e} V(x_i)^2 + \sum_{i=1}^6 \left\langle l_0^{1/2} V_{f,1}, V_{f,1} \right\rangle_{\hat{\Gamma}_i^f}.$$

The first sum in (4.5) is over the nodes x_i on $\hat{\Gamma}^e$.

Remark 4.3. The quasi-uniformity of the mesh defined on $\hat{\Gamma}^e$ implies that

$$c \langle V, V \rangle_{\hat{\Gamma}^e} \leq \hat{h} \sum_{x_i \in \hat{\Gamma}^e} V(x_i)^2 \leq C \langle V, V \rangle_{\hat{\Gamma}^e}.$$

The construction of the second form Q_2 differs from the first only in the way that V is decomposed.

Method 2. We define $V_{e,2}$ on $\partial\hat{\Omega}$ to be the function which equals V on $\hat{\Gamma}^e$ and is discrete harmonic in the faces, i.e., for each face $\hat{\Gamma}_i^f$,

$$(4.6) \quad \int_{\hat{\Gamma}_i^f} \nabla V_{e,2} \cdot \nabla \Phi \, ds = 0 \quad \text{for all } \Phi \in S_h^0(\hat{\Gamma}_i^f).$$

Again, we set $V = V_{e,2} + V_{f,2}$ and define

$$(4.7) \quad Q_2(V, V) \equiv \hat{h} \sum_{x_i \in \hat{\Gamma}^e} V(x_i)^2 + \sum_{i=1}^6 \left\langle l_0^{1/2} V_{f,2}, V_{f,2} \right\rangle_{\hat{\Gamma}_i^f}.$$

Note that the definitions (4.5) and (4.7) only differ in their respective use of $V_{f,1}$ and $V_{f,2}$. These constructions lead to completely different quadratic forms.

The following proposition provides estimates for α_0 and α_1 for the forms Q_1 and Q_2 . Its proof will be given later in this section.

PROPOSITION 2. *For $j = 1, 2$, there are positive constants c and C which are independent of \hat{h} and satisfy*

$$(4.8) \quad c(1 + \ln^2(\hat{h}^{-1}))^{-1} Q_j(V, V) \leq |V|_{1/2, \partial\hat{\Omega}}^2 \leq C Q_j(V, V) \quad \text{for all } V \in S_h(\partial\hat{\Omega}).$$

Combining Propositions 1 and 2 gives the following theorem.

THEOREM. *Let B be defined by (3.4) with $Q = Q_1$ or $Q = Q_2$. Then there are constants c and C which are independent of d and h satisfying*

$$c(1 + \ln^2(d/h))^{-1}B(W, W) \leq A(W, W) \leq CB(W, W) \quad \text{for all } W \in S_h(\Omega).$$

Remark 4.4. A construction analogous to that used in Method 1 can be carried out in the two-dimensional case. This leads to a preconditioned system with condition number on the order of $\ln^2(d/h)$. Instead of the corner problem of the preconditioner of [6], this method requires the solution of a sparse system with the number of variables equal to the number of subdomains.

We next give a proof of Proposition 2. We shall start by stating some lemmas which are used in the proof. Two of these lemmas (Lemmas 4.2 and 4.3) represent a fundamental part of the analysis. We prove the proposition assuming the lemmas and then devote the remainder of the section to the proof of the lemmas.

LEMMA 4.1. *Let $V \in S_{\hat{h}}(\partial\hat{\Omega})$ and $V_e = V_{e,1}$ or $V_e = V_{e,2}$; then*

$$|V_e|_{1/2, \partial\hat{\Omega}}^2 \leq C \langle V, V \rangle_{\hat{\Gamma}^e}.$$

LEMMA 4.2. *Let $V \in S_{\hat{h}}(\partial\hat{\Omega})$; then*

$$\langle V, V \rangle_{\hat{\Gamma}^e} \leq C(1 + \ln(\hat{h}^{-1}))|V|_{1/2, \partial\hat{\Omega}}^2.$$

LEMMA 4.3. *Let $V \in S_{\hat{h}}(\partial\hat{\Omega})$ and $V_f = V_{f,1}$ or $V_f = V_{f,2}$. Then*

$$\left\langle l_0^{1/2}V_f, V_f \right\rangle_{\hat{\Gamma}_i^f} \leq C(1 + \ln^2(\hat{h}^{-1}))|V|_{1/2, \partial\hat{\Omega}}^2$$

holds for every face $\hat{\Gamma}_i^f$ of $\partial\hat{\Omega}$.

Assuming Lemmas 4.1–4.3, we can prove Proposition 2.

Proof of Proposition 2. Let V in $S_{\hat{h}}(\partial\hat{\Omega})$ be decomposed into $V = V_e + V_f$. Then

$$(4.9) \quad |V|_{1/2, \partial\hat{\Omega}}^2 \leq 7 \left(|V_e|_{1/2, \partial\hat{\Omega}}^2 + \sum_{i=1}^6 |V_f^i|_{1/2, \partial\hat{\Omega}}^2 \right),$$

where V_f^i is the function defined on $\partial\hat{\Omega}$ which equals V_f on $\hat{\Gamma}_i^f$ and is zero on $\partial\hat{\Omega}/\hat{\Gamma}_i^f$. The second inequality of the proposition follows from (4.9), Lemma 4.1, Remarks 4.3, 4.2 and the definition of $\dot{H}^{1/2}(\hat{\Gamma}_i^f)$. The first inequality of the proposition follows from Remark 4.3 and Lemmas 4.2 and 4.3.

We now proceed with the proof of the lemmas. Some of the details for the proofs in the case of Method 2 are somewhat technical. So as not to disturb the flow of the domain decomposition analysis, these details will be given in the Appendix.

Proof of Lemma 4.1 for Method 1. By convexity,

$$(4.10) \quad |V_{e,1}|_{1/2, \partial\hat{\Omega}}^2 \leq c |V_{e,1}|_{\partial\hat{\Omega}} |V_{e,1}|_{1, \partial\hat{\Omega}}.$$

Using the fact that $V_{e,1}$ vanishes on the nodes of $\partial\hat{\Omega}$ which are not on $\hat{\Gamma}^e$, a straightforward computation gives

$$|V_{e,1}|_{\partial\hat{\Omega}}^2 \leq C\hat{h}^2 \sum_{x_i \in \hat{\Gamma}^e} V_{e,1}(x_i)^2$$

and

$$|V_{e,1}|_{1,\partial\hat{\Omega}}^2 \leq C \sum_{x_i \in \hat{\Gamma}^e} V_{e,1}(x_i)^2.$$

The lemma for Method 1 then follows from Remark 4.3.

The proof of Lemma 4.1 in the case of Method 2 involves some technical estimates for functions which are discrete harmonic on the faces of $\partial\hat{\Omega}$ and will be given in the Appendix.

In preparation for the proofs of the remaining two lemmas, we state a certain type of two-dimensional discrete Sobolev inequality whose proof can be found in [3], [6]. Let \tilde{S}_h be a subspace of approximating functions with mesh size \tilde{h} defined on a two-dimensional domain $\tilde{\Omega}$ (in our applications, \tilde{S}_h will be $S_{\hat{h}}(\hat{\Omega})$ restricted to some two-dimensional slice of $\hat{\Omega}$). We assume that $\tilde{\Omega}$ satisfies a cone condition of size d and angle α bounded away from zero and that \tilde{S}_h satisfies the following inverse inequality:

$$(4.11) \quad |\nabla V|_{L^\infty(\tilde{\Omega})} \leq C_1 \tilde{h}^{-1} |V|_{L^\infty(\tilde{\Omega})} \quad \text{for all } V \in \tilde{S}_h.$$

Then there exists a positive constant C independent of \tilde{h} such that

$$(4.12) \quad |V|_{L^\infty(\tilde{\Omega})}^2 \leq C(1 + \ln(\tilde{h}^{-1})) |V|_{1,\tilde{\Omega}}^2 \quad \text{for all } V \in \tilde{S}_h.$$

Proof of Lemma 4.2. Let $V \in S_{\hat{h}}(\partial\hat{\Omega})$. Define \tilde{V} to be the discrete harmonic extension of V (into the interior of $\hat{\Omega}$). By (3.11), it suffices to show that

$$\langle V, V \rangle_{\hat{\Gamma}^e} \leq c(1 + \ln(\hat{h}^{-1})) \|\tilde{V}\|_{1,\hat{\Omega}}^2.$$

Without loss of generality, we consider the integral over that part of the edge which corresponds to $x = z = 0$. Then by (4.12),

$$(4.13) \quad \int_0^1 V^2(0, y, 0) dy \leq c(1 + \ln(\hat{h}^{-1})) \int_0^1 |\tilde{V}(\cdot, y, \cdot)|_{H^1}^2 dy.$$

The H^1 norm in the last integral of (4.13) is over the intersection of $\hat{\Omega}$ with the plane at the given y -value. This integral is clearly bounded by $\|\tilde{V}\|_{1,\hat{\Omega}}^2$, and hence the proof is complete.

Proof of Lemma 4.3. Much of the proof of the lemma is the same whether we are considering Method 1 or Method 2. Accordingly, let $V \in S_{\hat{h}}(\partial\hat{\Omega})$ be decomposed into $V = V_e + V_f$, where V_e and V_f are given by either Method 1 or Method 2. By Remark 4.2, it suffices to prove

$$(4.14) \quad |V_f|_{H^{1/2}(\hat{\Gamma}_i^f)}^2 \leq C(1 + \ln^2(\hat{h}^{-1})) |V|_{1/2,\partial\hat{\Omega}}^2.$$

Let w be the function defined on $\partial\hat{\Omega}$ which equals V_f on $\hat{\Gamma}_i^f$ and is zero on $\partial\hat{\Omega}/\hat{\Gamma}_i^f$. Then the $\dot{H}^{1/2}(\hat{\Gamma}_i^f)$ norm of V_f is given by (4.1). The corresponding integral term in (4.1) reduces to

$$\int_{\hat{\Gamma}_i^f} \int_{\hat{\Gamma}_i^f} \frac{(V_f(x) - V_f(y))^2}{|x - y|^3} ds(x) ds(y) + 2 \int_{\hat{\Gamma}_i^f} \int_{\partial\hat{\Omega}/\hat{\Gamma}_i^f} \frac{V_f(x)^2}{|x - y|^3} ds(x) ds(y).$$

Let the four edges of $\hat{\Gamma}_i^f$ be denoted $\Gamma_{i,j}^{f,e}$ for $j = 1, 2, 3, 4$. A straightforward computation gives that

$$c \int_{\partial\hat{\Omega}/\hat{\Gamma}_i^f} |x - y|^{-3} ds(x) \leq \sum_{j=1}^4 \text{Dist}(y, \Gamma_{i,j}^{f,e})^{-1} \leq C \int_{\partial\hat{\Omega}/\hat{\Gamma}_i^f} |x - y|^{-3} ds(x),$$

where $\text{Dist}(y, \Gamma_{i,j}^{f,e})$ denotes the distance from y to $\Gamma_{i,j}^{f,e}$. Thus, the quantity $|V_f|_{H^{1/2}(\hat{\Gamma}_i^f)}^2$ is equivalent to

$$(4.15) \quad \int_{\hat{\Gamma}_i^f} \int_{\hat{\Gamma}_i^f} \frac{(V_f(x) - V_f(y))^2}{|x - y|^3} ds(x) ds(y) + \sum_{j=1}^4 \int_{\hat{\Gamma}_i^f} \frac{V_f(y)^2}{\text{Dist}(y, \Gamma_{i,j}^{f,e})} ds(y).$$

To bound the double integral term in (4.15), it suffices to bound the square of the $H^{1/2}(\partial\hat{\Omega})$ norm of V_f . Lemmas 4.1 and 4.2 give

$$(4.16) \quad |V_f|_{1/2, \partial\hat{\Omega}}^2 \leq 2(|V|_{1/2, \partial\hat{\Omega}}^2 + |V_e|_{1/2, \partial\hat{\Omega}}^2) \leq C(1 + \ln(\hat{h}^{-1}))|V|_{1/2, \partial\hat{\Omega}}^2.$$

Thus, to complete the proof of the theorem, we need only bound the single integral terms in (4.15).

Without loss of generality, it suffices to consider the face $\hat{\Gamma}_i^f$ in the plane $z = 0$ and a typical term, for example

$$(4.17) \quad \int_0^1 \int_0^1 \frac{V_f(x, y, 0)^2}{x} dx dy.$$

Thus it suffices to prove

$$(4.18) \quad \begin{aligned} & \int_0^1 \int_0^{\hat{h}} \frac{V_f(x, y, 0)^2}{x} dx dy + \int_0^1 \int_{\hat{h}}^1 \frac{V_f(x, y, 0)^2}{x} dx dy \\ & \leq c(1 + \ln^2(\hat{h}^{-1}))|V|_{1/2, \partial\hat{\Omega}}^2. \end{aligned}$$

For the first term in (4.18), we have

$$\int_0^1 \int_0^{\hat{h}} \frac{V_f(x, y, 0)^2}{x} dx dy \leq c\hat{h}^2 \int_0^1 \left| \frac{\partial V_f(\cdot, y, 0)}{\partial x} \right|_{L^\infty}^2 dy.$$

Let \tilde{V}_f be the discrete harmonic extension of V_f into $\hat{\Omega}$. By an inverse property of the subspace $S_{\hat{h}}(\hat{\Omega})$ restricted to the plane $y = \text{constant}$ and (4.12),

$$(4.19) \quad \int_0^1 \int_0^{\hat{h}} \frac{V_f(x, y, 0)^2}{x} dx dy \leq c(1 + \ln(\hat{h}^{-1})) \int_0^1 \left| \tilde{V}_f(\cdot, y, \cdot) \right|_{H^1}^2 dy.$$

The integral term in the right-hand side of (4.19) is clearly bounded by $\|\tilde{V}_f\|_{1, \hat{\Omega}}^2$. But \tilde{V}_f is discrete harmonic and hence by (3.11) and (4.16),

$$(4.20) \quad \begin{aligned} & \int_0^1 \int_0^{\hat{h}} \frac{V_f(x, y, 0)^2}{x} dx dy \leq c(1 + \ln(\hat{h}^{-1}))|V_f|_{1/2, \partial\hat{\Omega}}^2 \\ & \leq C(1 + \ln^2(\hat{h}^{-1}))|V|_{1/2, \partial\hat{\Omega}}^2. \end{aligned}$$

We next consider the second term of (4.18). For the case of Method 1, we have

$$(4.21) \quad \begin{aligned} & \int_0^1 \int_{\hat{h}}^1 \frac{V_{f,1}(x, y, 0)^2}{x} dx dy \\ & \leq \ln(\hat{h}^{-1}) \int_0^1 |V_{f,1}(\cdot, y, 0)|_{L^\infty}^2 dy \\ & \leq 2\ln(\hat{h}^{-1}) \left(\int_0^1 |V(\cdot, y, 0)|_{L^\infty}^2 dy + \int_0^1 |V_{e,1}(\cdot, y, 0)|_{L^\infty}^2 dy \right). \end{aligned}$$

Let \tilde{V} denote the discrete harmonic extension of V into $\hat{\Omega}$. By (4.12) and (3.11),

$$(4.22) \quad \int_0^1 |V(\cdot, y, 0)|_{L^\infty}^2 dy \leq c(1 + \ln(\hat{h}^{-1})) \|\tilde{V}\|_{1,\hat{\Omega}}^2 \leq c(1 + \ln(\hat{h}^{-1})) |V|_{1/2,\partial\hat{\Omega}}^2.$$

Since $V_{e,1}$ vanishes on the interior nodes of the faces,

$$|V_{e,1}(\cdot, y, 0)|_{L^\infty}^2 \leq V_{e,1}(0, y, 0)^2 + V_{e,1}(1, y, 0)^2.$$

Hence by Lemma 4.2,

$$(4.23) \quad \begin{aligned} \int_0^1 |V_{e,1}(\cdot, y, 0)|_{L^\infty}^2 dy &\leq C \langle V_{e,1}, V_{e,1} \rangle_{\hat{\Gamma}^e} \\ &\leq c(1 + \ln(\hat{h}^{-1})) |V|_{1/2,\partial\hat{\Omega}}^2. \end{aligned}$$

Inequality (4.18) follows from (4.20), (4.21), (4.22), and (4.23). This completes the proof of the lemma in the case of Method 1.

To complete the proof of the lemma, we have only to bound the second integral term in (4.18) in the case of Method 2. Applying the arithmetic geometric mean inequality, and changing the order of integration, gives

$$(4.24) \quad \begin{aligned} &\int_0^1 \int_{\hat{h}}^1 \frac{V_{f,2}(x, y, 0)^2}{x} dx dy \\ &\leq 2 \int_0^1 \int_{\hat{h}}^1 \frac{V(x, y, 0)^2}{x} dx dy + 2 \int_{\hat{h}}^1 \int_0^1 \frac{V_{e,2}(x, y, 0)^2}{x} dy dx \\ &\leq 2 \ln(\hat{h}^{-1}) \left(\int_0^1 |V(\cdot, y, 0)|_{L^\infty}^2 dy + \sup_{x \in [0, 1]} \int_0^1 V_{e,2}(x, y, 0)^2 dy \right). \end{aligned}$$

In the Appendix, we shall show that

$$(4.25) \quad \sup_{x \in [0, 1]} \int_0^1 V_{e,2}(x, y, 0)^2 dy \leq c \langle V_{e,2}, V_{e,2} \rangle_{\hat{\Gamma}^e}.$$

Applying (4.22) to the first term in (4.24) and (4.25) and Lemma 4.2 to the second gives the desired bound for the second term of (4.18). This completes the proof of the lemma.

5. The Solution of the Preconditioning Problem (1.5). In this section, we give an efficient algorithm for solving (1.5) in the case of Method 1. A similar algorithm can be developed for Method 2. Because of the discrete harmonic extensions on the faces, the algorithm for Method 2 is somewhat less efficient than that corresponding to Method 1 (see Remark 5.2).

In general, when B is of the form (3.4), we solve first for W_P on each subdomain, then for the values of W_H on Γ , and finally extend W_H to all of Ω . Most of the ideas described in this section have appeared in our earlier papers. However, the application of these techniques is not transparent and hence we include a discussion here.

The solution of (1.5) involves a three-step procedure. As already mentioned, the problem of finding the solution W to (1.5) reduces to that of computing W_P and

W_H on each subdomain. The first step is to compute W_P . By taking $\Phi \in S_h^0(\Omega_i)$ in (1.5) and using (3.4), it follows that

$$(5.1) \quad A_i(W_P, \Phi) = G(\Phi) \quad \text{for all } \Phi \in S_h^0(\Omega_i).$$

Thus, W_P can be determined by solving independent discrete Dirichlet problems on the subregions. The second step involves the computation of the values of W_H on Γ . These values are determined as the solution of the problem

$$(5.2) \quad \begin{aligned} d \sum_{i=1}^m \delta_i Q(I_i W - \gamma_i(W), \theta) \\ = G(\tilde{\theta}) - A(W_P, \tilde{\theta}) \equiv F(\theta) \quad \text{for all } \theta \in S_h(\Gamma). \end{aligned}$$

Here, $\tilde{\theta}$ denotes any extension of θ in $S_h(\Omega)$, and $\{\gamma_i(W)\}$ are the constants defined by (3.5). Note that by (5.1), the right-hand side of (5.2) is independent of the particular extension $\tilde{\theta}$. The development of an algorithm for solving (5.2) is an important part of this section and will be considered shortly. Assuming the values of W_H on Γ have been computed, the third step is to compute the ‘discrete A_i -harmonic’ extension into the interior of the subdomains. This is done separately on each subdomain as follows: Let \tilde{W}_H be any extension of the boundary values of W_H in $S_h(\Omega_i)$, e.g., the extension which is zero at all of the nodes not on $\partial\Omega_i$. Then on Ω_i , $W_H = Y + \tilde{W}_H$, where $Y \in S_h^0(\Omega_i)$ is the solution of

$$(5.3) \quad A_i(Y, \Phi) = -A_i(\tilde{W}_H, \Phi) \quad \text{for all } \Phi \in S_h^0(\Omega_i).$$

Thus, the computation of W_H (once its values on Γ are known) reduces to the solution of independent discrete Dirichlet problems on the subdomains.

To complete the description of the algorithm, we provide an efficient way to compute the function W_H on Γ , i.e., the solution of (5.2). This involves a two-step procedure. The first step requires the computation of the average values $\{\gamma_i(W)\}$ appearing in (5.2). We shall use a technique described in [7] to derive a sparse matrix problem for these values. We will only consider the case when all of the δ_i ’s are equal to one; the more general case is similar (cf. [7]). This matrix is derived using a special choice of test functions, $\phi_i \in S_h(\Gamma)$, for $i = 1, \dots, m$. Consider a fixed i . We define the function ϕ_i to vanish on the nodes which are not on $\partial\Omega_i$. Its values on the nodes of $\partial\Omega_i$ are to be determined.

Let X and Y be in $S_h(\Omega)$ and $\Gamma_{ij}^f = \Gamma_{kl}^f$; then by (A.3) and (A.4),

$$(5.4) \quad \left\langle l_0^{1/2}(I_i X)_f, (I_i Y)_f \right\rangle_{\hat{\Gamma}_j^f} = \left\langle l_0^{1/2}(I_k X)_f, (I_k Y)_f \right\rangle_{\hat{\Gamma}_l^f}.$$

Here, $(\cdot)_f$ denotes the face component in the decomposition of Method 1. Let $\mathcal{N}(x_j)$ be the indices of the subregions which share a boundary node x_j and $\mathcal{N}(\Gamma_{ij}^f)$ be the indices of the two subregions which share a boundary face Γ_{ij}^f . The number of indices in $\mathcal{N}(x_j)$ will be denoted $|\mathcal{N}(x_j)|$. Then (5.4) and the properties of ϕ_i

imply

$$\begin{aligned}
d \sum_{j=1}^m Q(I_j W - \gamma_j(W), \phi_i) &= d\hat{h} \sum_{x_j \in \Gamma_i^e} |\mathcal{N}(x_j)| W(x_j) \phi_i(x_j) \\
&\quad + 2d \sum_{j=1}^6 \left\langle l_0^{1/2}(I_j W)_f, (I_j \phi_i)_f \right\rangle_{\hat{\Gamma}_j^f} \\
(5.5) \quad &\quad - d\hat{h} \sum_{x_j \in \Gamma_i^e} \left(\sum_{k \in \mathcal{N}(x_j)} \gamma_k(W) \right) \phi_i(x_j) \\
&\quad - d \sum_{j=1}^6 \left(\sum_{l \in \mathcal{N}(\Gamma_{ij}^f)} \gamma_l(W) \right) \left\langle l_0^{1/2}(1)_f, (I_j \phi_i)_f \right\rangle_{\hat{\Gamma}_j^f} \\
&\equiv S_1 + S_2 + S_3 + S_4,
\end{aligned}$$

where $\{x_j\}$ are the nodal values on $\Gamma_i^e = \bigcup_{j=1}^6 \partial\Gamma_{ij}^f$.

We define ϕ_i at the nodal values on $\partial\Omega_i$ by

$$(5.6) \quad \phi_i(x_j) = \begin{cases} 1/|\mathcal{N}(x_j)| & \text{when } x_j \in \Gamma_i^e, \\ 1/2 & \text{when } x_j \in \Gamma_{ik}^f. \end{cases}$$

Then, by (3.5), the first two sums of (5.5) can be written

$$(5.7) \quad S_1 + S_2 = dQ(I_i W, 1) = \gamma_i(W) dQ(1, 1).$$

Combining (5.5) and (5.7) gives

$$(5.8) \quad dQ(1, 1)\gamma_i(W) - M_{ik}\gamma_k(W) = F(\phi_i),$$

where

$$M_{ik} = d\hat{h} \sum_{x_j \in \Gamma_i^e \cap \Gamma_k^e} \frac{1}{|\mathcal{N}(x_j)|} + d \sum_{\substack{j=1, \dots, 6 \\ \Gamma_{ij}^f \cap \partial\Omega_k \neq \emptyset}} \frac{\left\langle l_0^{1/2}(1)_f, (1)_f \right\rangle_{\hat{\Gamma}_j^f}}{2}.$$

It is straightforward to check that M is symmetric with nonnegative entries. Furthermore, the row sum of M for any row is less than or equal to $dQ(1, 1)$, with strict inequality when the row corresponds to a domain Ω_i with $\partial\Omega_i \cap \partial\Omega \neq \emptyset$. This means that $dQ(1, 1)I - M$ (where I denotes the $m \times m$ identity matrix) is an M -matrix [17] and resembles matrices arising in standard finite difference methods. We compute the values of $\{\gamma_i(W)\}$ by solving this $m \times m$ system.

Remark 5.1. In the case of many subdomains, the matrix $dQ(1, 1)I - M$ is sparse since the i th equation only involves the values of $\gamma_k(W)$ for subdomains Ω_k with $\partial\Omega_i \cap \partial\Omega_j \neq \emptyset$.

Once the values of $\{\gamma_k(W)\}$ are known, we compute the values of W_H on Γ as follows. We are left to solve

$$(5.9) \quad d \sum_{i=1}^m Q(I_i W, \theta) = F(\theta) + d \sum_{i=1}^m Q(\gamma_i(W), \theta) \equiv \tilde{F}(\theta) \quad \text{for all } \theta \in S_h(\Gamma).$$

By (5.4), we have

$$(5.10) \quad \begin{aligned} d \sum_{i=1}^m Q(I_i W, \theta) &= d\hat{h} \sum_{x_j \in \Gamma^e} |\mathcal{N}(x_j)| W(x_j) \theta(x_j) \\ &\quad + d \sum_{k,l} \left\langle l_0^{1/2} (I_k W)_f, (I_k \theta)_f \right\rangle_{\hat{\Gamma}_l^f} \quad \text{for all } \theta \in S_h(\Gamma). \end{aligned}$$

Here, Γ^e is the union of the closures of the edges of the subdomains. For functions θ with support contained on the j th face of the i th subregion, (5.9)–(5.10) reduces to

$$(5.11) \quad 2d \left\langle l_0^{1/2} (I_i W)_f, (I_i \theta)_f \right\rangle_{\hat{\Gamma}_j^f} = \tilde{F}(\theta).$$

Equation (5.11) completely determines the values of W on the nodes of Γ_{ij}^f . For functions θ which vanish on the face nodes, (5.9)–(5.10) reduces to

$$(5.12) \quad d\hat{h} \sum_{x_j \in \Gamma^e} |\mathcal{N}(x_j)| W(x_j) \theta(x_j) = \tilde{F}(\theta).$$

The nodal values of W on Γ^e are trivially computed from (5.12) using θ corresponding to nodal basis functions.

Remark 5.2. An algorithm similar to that described above could be developed for the solution of the preconditioning form B defined using Q_2 (i.e., Method 2). In fact, if $I_i \phi_i$ is discrete harmonic on the faces, (5.5) gets replaced by

$$\begin{aligned} d \sum_{i=1}^m Q(I_i W - \gamma_i(W), \phi_i) &= d\hat{h} \sum_{x_j \in \Gamma_i^e} |\mathcal{N}(x_j)| W(x_j) \phi_i(x_j) \\ &\quad - d\hat{h} \sum_{x_j \in \Gamma_i^e} \left(\sum_{k \in \mathcal{N}(x_j)} \gamma_k(W) \right) \phi_i(x_j). \end{aligned}$$

For this case, $\phi_i(x_j)$ is defined by (5.6) for $x_j \in \Gamma_i^e$ and extended discrete harmonically into the faces. An equation for the values of $\gamma_i(W)$ analogous to (5.8) can then be derived. As above, once the values of $\gamma_i(W)$ have been computed, we are left to solve (5.9). (5.10) gets replaced by

$$(5.13) \quad \begin{aligned} d \sum_{i=1}^m Q(I_i W, \theta) &= d\hat{h} \sum_{x_j \in \Gamma^e} |\mathcal{N}(x_j)| W(x_j) \theta(x_j) \\ &\quad + d \sum_{k,l} \left\langle l_0^{1/2} (I_k W)_{f,2}, (I_k \theta)_{f,2} \right\rangle_{\hat{\Gamma}_l^f} \quad \text{for all } \theta \in S_h(\Gamma). \end{aligned}$$

The values of $(I_k W)_{f,2}$ can then be computed on the faces using equations similar to (5.11). In the case of Method 2, (5.12) is only valid for functions θ for which $I_i \theta$ is discrete harmonic on all faces of $\partial\hat{\Omega}$ for all i . Thus the discrete harmonic extension into the faces must be computed for each edge nodal function (i.e., a function which is one on one of the edge nodes and vanishes on the remaining edge nodes). Even if these extensions are preprocessed, one must compute \tilde{F} applied to each of these for each inversion of B . This results in a work increase of $O(N)$ operations and substantially complicates the coding.

We conclude this section with a review of the procedure developed here for solving (1.5) when B is given by (3.4) and Q is given by Method 1.

Algorithm for Solving (1.5).

- (1) Compute W_P by solving (5.1). This involves the solution of Dirichlet problems on the subdomains, which can be done independently and in parallel.
- (2) Compute the values of W_H on Γ solving (5.2). First, we compute the values $\{\gamma_i(W)\}$ by solving the matrix problem (5.8). The values of W_H on Γ are then computed by (5.11) and (5.12).
- (3) Extend the boundary values of W_H by solving (5.3). As in Step 1, this involves the solution of Dirichlet problems on the subdomains, which can be done independently and in parallel.
- (4) Set $W = W_P + W_H$.

6. Numerical Experiments. In this section, we present the results of numerical experiments using the preconditioners developed earlier. We shall only report results for the more computationally effective algorithms resulting from Method 1. We have made no attempt to develop a general code, and consequently our results will be for model applications. These computations are designed to illustrate the theory developed in the earlier sections.

The domain Ω will be the unit cube partitioned into $m = m_0 \times m_0 \times m_0$ subdomains which are subcubes of side length $1/m_0$. We will use a finite difference approximation on a grid of size $k \times k \times k$. Let $h = 1/(k+1)$ and $J = (j_1, j_2, j_3)$ be a multi-integer. Then the nodes of the grid are the points $x_J = (j_1 h, j_2 h, j_3 h)$ for $1 \leq j_1, j_2, j_3 \leq k$.

Example 1. For the first example, we consider the model problem

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

Here, Δ denotes the Laplace operator. The finite difference approximation to u is the nodal function U which satisfies (2.4). In this case

$$A(V, W) = (L_h V) \cdot W,$$

where L_h is the seven-point difference operator given by

$$(6.1) \quad \begin{aligned} (L_h V)_J &= 6V_{j_1, j_2, j_3} - V_{j_1+1, j_2, j_3} - V_{j_1-1, j_2, j_3} - V_{j_1, j_2+1, j_3} \\ &\quad - V_{j_1, j_2-1, j_3} - V_{j_1, j_2, j_3+1} - V_{j_1, j_2, j_3-1}. \end{aligned}$$

We define $V_K = 0$ for indices K appearing on the right-hand side of (6.1) which are not in Ω .

For this example, the nodes on the faces of the subdomains are regularly spaced. We note that the definition of Method 1 and Method 2 only requires the computation of $l_0^{1/2}$ on the reference element with respect to the reference subspace. Because of the uniformly spaced grid on the faces, $l_0^{1/2}$ can be economically computed by use of the discrete Fourier transform. In addition, it is possible to replace $l_0^{1/2}$ on this subspace by any uniformly spectrally equivalent operator. For example, $l_0^{1/2}$ could be replaced by $\tilde{l}_0^{1/2}$, where $\tilde{l}_0^{1/2}$ is \hat{h} times the square root of the five-point operator on the face. We use $\tilde{l}_0^{1/2}$ in the numerical examples of this section.

Table 6.1 gives computational results for Example 1. In this case, the cube was broken up into eight subcubes ($m=2$). We report the condition number K for the preconditioned system as a function of h . For comparison, we provide the function

$$f(d/h) = 10.9 + .76 \log_2^2(d/h).$$

The close correlation between K and $f(1/2h)$ suggests that the growth of the condition number of the preconditioned system is in agreement with the theorem in Section 4. We have also included the number of nodes, N , and the number of iterations, N_i , of preconditioned conjugate gradient required to reduce the A -norm error of a typical example by a factor of .001.

TABLE 6.1
Iterative convergence for Example 1.

h	K	$f(1/2h)$	N_i	N
1/4	10.5	11.7	7	27
1/8	13.9	13.9	8	343
1/16	17.7	17.7	8	3375
1/32	23	23	7	29791

Example 2. In this example, we consider a variable coefficient problem which has large jumps in the coefficients across the subdomain boundaries. Specifically, we consider the problem

$$-\nabla \cdot \mu \nabla u = f \quad \text{in } \Omega,$$

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

For this example, we consider the unit cube broken down into twenty-seven subdomains. The function μ is piecewise constant on the subregions with values given by Figure 6.1. Table 6.2 gives the results of computational experiments for this example. Note that the results for the condition number K of the preconditioned system

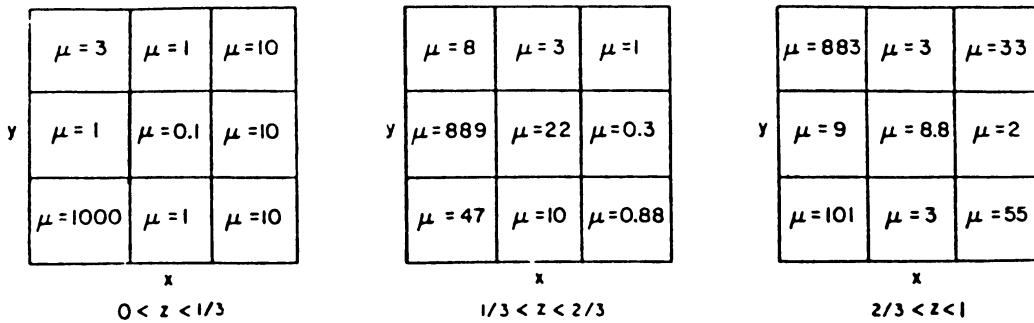


FIGURE 6.1
Coefficients for Example 2.

are of the same magnitude as those of Table 6.1. This suggests that the method gives rise to convergence rates which are independent of jumps in coefficients across the subregions. This is in agreement with the analysis since, with an appropriate choice of $\{\delta_i\}$, the constants c and C in (3.3) can be chosen independent of such jumps.

TABLE 6.2
Iterative convergence for Example 2.

h	K	$f(1/3h)$	N_i	N
1/6	11.6	11.7	11	125
1/12	14.1	13.9	10	1331
1/24	18.3	17.7	10	12167

7. Appendix. We first prove Lemma 4.1 in the case of Method 2. To this end, we prove auxiliary lemmas involving harmonic functions. Then the proof of Lemma 4.1 will follow from approximation.

We will use the integral representation given in (4.1) for estimating the $H^{1/2}(\partial\tilde{\Omega})$ norm. Let θ_1 and θ_2 be two-dimensional domains and u be defined on $\theta_1 \cup \theta_2$. We define

$$(7.1) \quad I(\theta_1, \theta_2, u) = \int_{\theta_1} \int_{\theta_2} \frac{(u(x) - u(y))^2}{|x - y|^3} dx dy.$$

The first auxiliary lemma will involve the domain $\tilde{\Omega} = [-1, 1] \times [0, 1]$. Let $\tilde{\Omega}_1 = [-1, 0] \times [0, 1]$, $\tilde{\Omega}_2 = [0, 1] \times [0, 1]$ and $\tilde{\Gamma} = \partial\tilde{\Omega}_1 \cup \partial\tilde{\Omega}_2$.

LEMMA 7.1. *Let $u \in H^{1/2}(\partial\tilde{\Omega})$ be harmonic in $\tilde{\Omega}_i$ for $i = 1, 2$. Then*

$$(7.2) \quad |u|_{1/2, \tilde{\Omega}}^2 \leq c |u|_{\tilde{\Gamma}}^2.$$

Proof. We define

$$u_e(x, y) = (u(x, y) + u(-x, y))/2$$

and

$$u_o(x, y) = (u(x, y) - u(-x, y))/2.$$

Now $u = u_e + u_o$ gives an orthogonal decomposition of u in the $L^2(\tilde{\Gamma})$ -inner product. Consequently, it suffices to prove (7.2) for $u = u_e$ and $u = u_o$. By the Schwarz reflection principle, u_o is harmonic in $\tilde{\Omega}$ and hence

$$|u_o|_{1/2, \tilde{\Omega}}^2 \leq c |u_o|_{\partial\tilde{\Omega}}^2 \leq c |u_o|_{\tilde{\Gamma}}^2.$$

By a representation analogous to (4.1),

$$\begin{aligned} |u_e|_{1/2, \tilde{\Omega}}^2 &= I(\tilde{\Omega}, \tilde{\Omega}, u_e) + |u_e|_{\tilde{\Omega}}^2 \\ &= 2I(\tilde{\Omega}_1, \tilde{\Omega}_1, u_e) + 2I(\tilde{\Omega}_1, \tilde{\Omega}_2, u_e) + |u_e|_{\tilde{\Omega}}^2. \end{aligned}$$

Since $|x - y| \geq |(-x_1, x_2) - y|$ holds when $x \in \tilde{\Omega}_1$ and $y \in \tilde{\Omega}_2$,

$$I(\tilde{\Omega}_1, \tilde{\Omega}_2, u_e) \leq I(\tilde{\Omega}_1, \tilde{\Omega}_1, u_e).$$

Hence,

$$|u_e|_{1/2, \tilde{\Omega}}^2 \leq c |u_e|_{1/2, \tilde{\Omega}_1}^2 \leq c |u_e|_{\partial\tilde{\Omega}_1}^2 \leq c |u_e|_{\tilde{\Gamma}}^2.$$

This completes the proof of the lemma.

The following lemma gives the result corresponding to Lemma 4.1 for functions which are harmonic on the faces of $\partial\hat{\Omega}$.

LEMMA 7.2. *Let $w \in H^{1/2}(\partial\Omega)$ be a function which is harmonic on each face of $\partial\hat{\Omega}$. Then*

$$|w|_{1/2,\partial\hat{\Omega}}^2 \leq c \langle w, w \rangle_{\hat{\Gamma}^e}.$$

Proof. We again use (4.1) to bound the $H^{1/2}(\partial\hat{\Omega})$ norm. The integral term of (4.1) is given by

$$I(\partial\hat{\Omega}, \partial\hat{\Omega}, w) = \sum_{i,j=1}^6 I(\hat{\Gamma}_i^f, \hat{\Gamma}_j^f, w).$$

Let ω_i be the union of the two faces adjacent to the i th edge. If two faces $\hat{\Gamma}_i^f$ and $\hat{\Gamma}_j^f$ do not share an edge, then

$$I(\hat{\Gamma}_i^f, \hat{\Gamma}_j^f, w) \leq c |w|_{\partial\hat{\Omega}}^2$$

and hence

$$(7.3) \quad |w|_{1/2,\partial\hat{\Omega}}^2 \leq c \left(\sum_{i=1}^{12} I(\omega_i, \omega_i, w) + |w|_{\partial\hat{\Omega}}^2 \right).$$

The lemma follows from (7.3) and Lemma 7.1.

Proof of Lemma 4.1 for Method 2. Let V be a function which is discrete harmonic on the faces of $\partial\hat{\Omega}$. Let v be the function which equals V on $\hat{\Gamma}^e$ and is harmonic on the faces of $\partial\hat{\Omega}$. By Lemma 7.2, it obviously suffices to show that

$$|V - v|_{1/2,\partial\hat{\Omega}}^2 \leq c |V|_{\hat{\Gamma}^e}^2.$$

By convexity,

$$(7.4) \quad |V - v|_{1/2,\partial\hat{\Omega}}^2 \leq c |V - v|_{\partial\hat{\Omega}} |V - v|_{1,\partial\hat{\Omega}}.$$

Applying well-known finite element techniques to estimate $|V - v|_{\partial\hat{\Omega}}$ and the Poincaré inequality gives

$$(7.5) \quad |V - v|_{1/2,\partial\hat{\Omega}}^2 \leq c \hat{h} D_{\partial\hat{\Omega}}(V - v, V - v) \leq c \hat{h} D_{\partial\hat{\Omega}}(\tilde{V} - v, \tilde{V} - v),$$

where \tilde{V} is the function in $S_{\hat{h}}(\hat{\Omega})$ which equals V on $\hat{\Gamma}^e$ and vanishes on the face nodes and $D_{\partial\hat{\Omega}}(\cdot, \cdot)$ denotes the Dirichlet inner product on $\partial\hat{\Omega}$. By the arithmetic geometric mean inequality, an inequality similar to (3.11), inverse assumptions and an obvious computation,

$$|V - v|_{1/2,\partial\hat{\Omega}}^2 \leq c \hat{h} \sum_{i=1}^6 \{ |V|_{1/2,\partial\hat{\Gamma}_i^f}^2 + D_{\hat{\Gamma}_i^f}(\tilde{V}, \tilde{V}) \} \leq C |V|_{\hat{\Gamma}^e}^2.$$

This completes the proof of the lemma.

We next prove (4.25). To do this, we first prove the analogous result for harmonic functions. The discrete result will then be derived by approximation.

LEMMA 7.3. *Let u be harmonic on the face $\hat{\Gamma}_i^f$ in the plane $z = 0$. Then*

$$(7.6) \quad \sup_{x \in [0,1]} \int_0^1 u^2(x, y, 0) dy \leq C \int_{\partial \hat{\Gamma}_i^f} u^2 ds.$$

Proof. Let $u = \sum_{j=1}^4 u_j$, where u_j is the harmonic function which equals u on $\Gamma_{i,j}^{f,e}$ and vanishes on the remaining three edges. By the arithmetic geometric mean inequality and obvious properties of the integral, it suffices to prove that

$$\sup_{x \in [0,1]} \int_0^1 u_j^2(x, y, 0) dy \leq C \int_{\Gamma_{i,j}^{f,e}} u_j^2 ds$$

holds for $j = 1, 2, 3, 4$. By obvious symmetries involving the $\{u_j\}$, it suffices to show that

$$(7.7) \quad \sup_{x \in [0,1]} \int_0^1 u_j^2(x, y, 0) dy + \sup_{y \in [0,1]} \int_0^1 u_j^2(x, y, 0) dx \leq C \int_{\Gamma_{i,j}^{f,e}} u_j^2 ds$$

holds for any j . Without loss of generality, we consider u_1 , the function which is nonzero on the line $x = 1$. Expanding u_1 in a sine series gives

$$(7.8) \quad u_1(x, y, 0) = \sum_{k=1}^{\infty} \alpha_k \sin(\pi k y) \left(\frac{e^{\pi k x} - e^{-\pi k x}}{e^{\pi k} - e^{-\pi k}} \right).$$

We consider the two terms of (7.7) separately. For the first, we use (7.8) and the Plancherel Theorem to get

$$\begin{aligned} \sup_{x \in [0,1]} \int_0^1 u_1^2(x, y, 0) dy &= 1/2 \sup_{x \in [0,1]} \sum_{k=1}^{\infty} \alpha_k^2 \left(\frac{e^{\pi k x} - e^{-\pi k x}}{e^{\pi k} - e^{-\pi k}} \right)^2 \\ &\leq 1/2 \sum_{k=1}^{\infty} \alpha_k^2 = \int_0^1 u_1(1, y, 0)^2 dy. \end{aligned}$$

For the second term of (7.7), using (7.8) and changing the order of summation and integration gives

$$\begin{aligned} (7.9) \quad &\sup_{y \in [0,1]} \int_0^1 u_1^2(x, y, 0) dx \\ &\leq \sum_{k,l=1}^{\infty} |\alpha_k| |\alpha_l| \int_0^1 \left(\frac{e^{\pi k x} - e^{-\pi k x}}{e^{\pi k} - e^{-\pi k}} \right) \left(\frac{e^{\pi l x} - e^{-\pi l x}}{e^{\pi l} - e^{-\pi l}} \right) dx. \end{aligned}$$

We clearly have that

$$\left(\frac{e^{\pi k x} - e^{-\pi k x}}{e^{\pi k} - e^{-\pi k}} \right) \leq \frac{2}{1 - e^{-2\pi}} e^{\pi k(x-1)}$$

and hence

$$(7.10) \quad \int_0^1 \left(\frac{e^{\pi k x} - e^{-\pi k x}}{e^{\pi k} - e^{-\pi k}} \right) \left(\frac{e^{\pi l x} - e^{-\pi l x}}{e^{\pi l} - e^{-\pi l}} \right) dx \leq \frac{C}{k+l}.$$

Combining the above inequalities gives

$$\sup_{y \in [0,1]} \int_0^1 u_1^2(x, y, 0) dx \leq c \sum_{k,l=1}^{\infty} \frac{|\alpha_k| |\alpha_l|}{k+l}.$$

Applying Hilbert's Double Series Theorem (cf. [12]) finally gives

$$(7.11) \quad \sup_{y \in [0,1]} \int_0^1 u_1^2(x, y, 0) dx \leq C \sum_{k=1}^{\infty} \alpha_k^2 \leq C \int_0^1 u_1(1, y, 0)^2 dy.$$

This completes the proof of the lemma.

Proof of (4.25). We must prove that (7.6) holds for functions $U \in S_{\hat{h}}(\partial\hat{\Omega})$ which are discrete harmonic on the face $\hat{\Gamma}_i^f$. Let u be the function which equals U on $\partial\hat{\Gamma}_i^f$ and is harmonic on $\hat{\Gamma}_i^f$. By the arithmetic geometric mean inequality and Lemma 7.3, it suffices to show that

$$\int_0^1 (u(x, y, 0) - U(x, y, 0))^2 dy \leq c \int_{\partial\hat{\Gamma}_i^f} U^2 ds$$

holds for any $x \in [0, 1]$. By standard finite element techniques, the Poincaré and trace inequalities,

$$\begin{aligned} \int_0^1 (u(x, y, 0) - U(x, y, 0))^2 dy &\leq \hat{c} h |u - U|_{1/2, \partial\hat{\Gamma}_i^f}^2 \\ &\leq \hat{c} h |u - U|_{1, \partial\hat{\Omega}}^2 \leq C \hat{h} D_{\partial\hat{\Omega}}(u - U, u - U). \end{aligned}$$

Inequality (4.25) follows from the argument used to bound (7.5) in the proof of Lemma 4.1 for Method 2.

Department of Mathematics
Cornell University
Ithaca, New York 14853
E-mail: bramble@mssun7.msi.cornell.edu

Brookhaven National Laboratory
Upton, New York 11973
E-mail: pasciak@bnl.gov

Department of Mathematics
Cornell University
Ithaca, New York 14853
E-mail: schatz@mssun7.msi.cornell.edu

1. G. P. ASTRAKHANTSEV, "Method for fictitious domains for a second-order elliptic equation with natural boundary conditions," *U.S.S.R. Comput. Math. and Math. Phys.*, v. 18, 1978, pp. 114–121.
2. P. E. BJØRSTAD & O. B. WIDLUND, "Iterative methods for the solution of elliptic problems on regions partitioned into substructures," *SIAM J. Numer. Anal.*, v. 23, 1986, pp. 1097–1120.
3. J. H. BRAMBLE, "A second order finite difference analogue of the first biharmonic boundary value problem," *Numer. Math.*, v. 9, 1966, pp. 236–249.
4. J. H. BRAMBLE, R. E. EWING, J. E. PASCIAK & A. H. SCHATZ, "A preconditioning technique for the efficient solution of problems with local grid refinement," *Comput. Methods Appl. Mech. Engrg.*, v. 67, 1988, pp. 149–159.
5. J. H. BRAMBLE, J. E. PASCIAK & A. H. SCHATZ, "An iterative method for elliptic problems on regions partitioned into substructures," *Math. Comp.*, v. 46, 1986, pp. 361–369.
6. J. H. BRAMBLE, J. E. PASCIAK & A. H. SCHATZ, "The construction of preconditioners for elliptic problems by substructuring, I," *Math. Comp.*, v. 47, 1986, pp. 103–134.
7. J. H. BRAMBLE, J. E. PASCIAK & A. H. SCHATZ, "The construction of preconditioners for elliptic problems by substructuring, II," *Math. Comp.*, v. 49, 1987, pp. 1–16.

8. J. H. BRAMBLE, J. E. PASCIAK & A. H. SCHATZ, "The construction of preconditioners for elliptic problems by substructuring, III," *Math. Comp.*, v. 51, 1988, pp. 415–430.
9. Q. V. DIHN, R. GLOWINSKI & J. PÉRIAUX, "Solving elliptic problems by domain decomposition methods," in *Elliptic Problem Solvers II* (G. Birkhoff and A. Schoenstadt, eds.), Academic Press, New York, 1984, pp. 395–426.
10. G. H. GOLUB & D. MEYERS, *The Use of Preconditioning Over Irregular Regions*, Proc. 6th Internat. Conf. Comput. Meth. Sci. and Engrg., Versailles, France, 1983.
11. W. D. GROPP & D. E. KEYES, *A Comparison on Domain Decomposition Techniques for Elliptic Partial Differential Equations and the Parallel Implementation*, Research Report YALEU/DCS/RR-448, 1985.
12. G. H. HARDY, J. E. LITTLEWOOD & G. PÓLYA, *Inequalities*, Cambridge Univ. Press, New York, 1952.
13. S. G. KREIN & Y. I. PETUNIN, *Scales of Banach Spaces*, Russian Math. Surveys, vol. 21, 1966, pp. 85–160.
14. J. L. LIONS & E. MAGENES, *Problèmes aux Limites non Homogènes et Applications*, vol. 1, Dunod, Paris, 1968.
15. S. MCCORMICK & J. THOMAS, "The fast adaptive composite grid (FAC) method for elliptic equations," *Math. Comp.*, v. 46, 1986, pp. 439–456.
16. J. NEČAS, *Les Méthodes Directes en Théorie des Équations Elliptiques*, Academia, Prague, 1967.
17. R. S. VARGA, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, N.J., 1962.

4.5 An iterative method for elliptic problems on regions partitioned into substructures

An iterative method for elliptic problems on regions partitioned into substructures
[19]

An Iterative Method for Elliptic Problems on Regions Partitioned into Substructures*

By J. H. Bramble, J. E. Pasciak and A. H. Schatz

Abstract. Some new preconditioners for discretizations of elliptic boundary problems are studied. With these preconditioners, the domain under consideration is broken into subdomains and preconditioners are defined which only require the solution of matrix problems on the subdomains. Analytic estimates are given which guarantee that under appropriate hypotheses, the preconditioned iterative procedure converges to the solution of the discrete equations with a rate per iteration that is independent of the number of unknowns. Numerical examples are presented which illustrate the theoretically predicted iterative convergence rates.

1. Introduction. In this paper we will consider as a model problem the Dirichlet problem for a second-order uniformly elliptic equation in two dimensions. Let Ω be a bounded domain in R^2 , with boundary $\partial\Omega$, which, for the sake of exposition, is the union of two regions Ω_1 and Ω_2 and a common boundary Γ . Examples of such splittings are given in Figure 1.

Thus we shall consider the problem

$$(1) \quad \begin{aligned} Lu &= f && \text{in } \Omega, \\ u &= 0 && \text{on } \partial\Omega, \end{aligned}$$

where

$$Lv = - \sum_{i,j=1}^2 \frac{\partial}{\partial x_i} \left(a_{ij} \frac{\partial v}{\partial x_j} \right)$$

with a_{ij} symmetric, uniformly positive definite, bounded and piecewise smooth on Ω . The generalized Dirichlet form is given by

$$A(v, \phi) = \sum_{i,j=1}^2 \int_{\Omega} a_{ij} \frac{\partial v}{\partial x_i} \frac{\partial \phi}{\partial x_j} dx$$

defined for all v and ϕ in the Sobolev space $H^1(\Omega)$ (the space of distributions with square-integrable first derivatives). The $L^2(\Omega)$ inner product is denoted by

$$(v, \phi) = \int_{\Omega} v \phi dx.$$

The subspace $H_0^1(\Omega)$ of $H^1(\Omega)$ is the completion of the smooth functions with support in Ω with respect to the norm in $H^1(\Omega)$. By integration by parts, the

Received September 25, 1984.

1980 *Mathematics Subject Classification*. Primary 65N30, 65F10.

*This work was supported in part under the National Science Foundation Grant No. DMS84-05352, under the Applied Mathematical Sciences subprogram of the Office of Energy Research, U.S. Department of Energy, Contract No. DE-AC02-76CH00016, and under the Air Force Office of Scientific Research, Contract No. ISSA86-0026.

©1986 American Mathematical Society
 0025-5718/86 \$1.00 + \$.25 per page

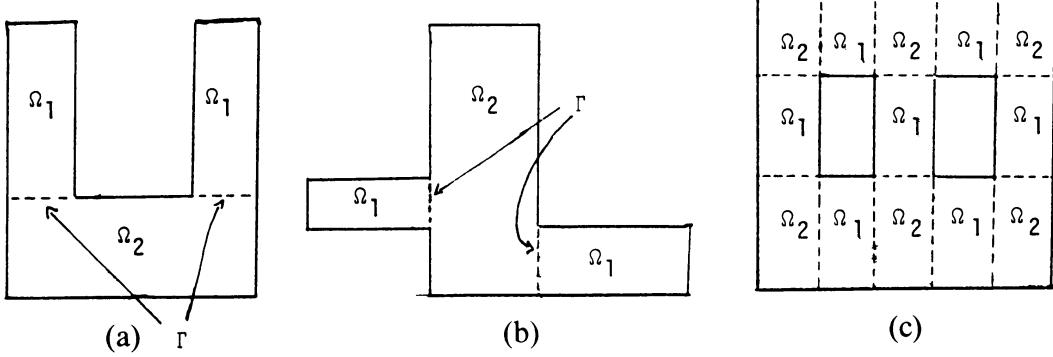


FIGURE 1

problem defined by (1) may be written in weak form: Find $u \in H_0^1(\Omega)$, such that

$$(2) \quad A(u, \phi) = (f, \phi)$$

for all $\phi \in H_0^1(\Omega)$. This leads immediately to the standard Galerkin approximation. Let S_h^0 be a finite-dimensional subspace of $H_0^1(\Omega)$. The Galerkin approximation is defined as the solution of the following problem: Find $U \in S_h^0$ such that

$$(3) \quad A(U, \chi) = (f, \chi)$$

for all $\chi \in S_h^0$. Once a basis $\{\chi_i\}_{i=1}^N$ for S_h^0 is chosen, (3) leads to a system of linear algebraic equations. Write

$$U = \sum_{i=1}^N \alpha_i \chi_i;$$

then (3) becomes

$$(4) \quad \sum_{i=1}^N \alpha_i A(\chi_i, \chi_j) = (f, \chi_j)$$

$j = 1, \dots, N$, which is a linear system for the determination of the coefficients α_i , $i = 1, \dots, N$.

It is well known that for a wide class of approximation spaces, S_h^0 , U will be a good approximation to u . We shall consider certain spaces S_h^0 for which we may also develop efficient algorithms for the solution of the underlying linear system (4).

The underlying method which we will consider is a preconditioned iterative method. The choice of particular iterative method within a certain class is not essential, but for the purpose of this paper we may think of the well-known conjugate gradient method which is often used in practice (cf. [1], [6], [9], [10]).

Let A be the $N \times N$ matrix with entries $A(\chi_i, \chi_j)$, $\alpha = (\alpha_1, \dots, \alpha_N)$ and F the vector with components (f, χ_j) . Then (4) may be written

$$(5) \quad A\alpha = F.$$

Generally, the matrix A is not well-conditioned so that a direct application of the conjugate gradient method to the symmetric positive definite system (5) will not be a very efficient algorithm. A preconditioned conjugate gradient method can be derived as follows. Let B be a positive definite symmetric matrix and write

$$(6) \quad B^{-1}A\alpha = B^{-1}F.$$

In the context of this paper the matrix B will be associated with another bilinear form $B(\cdot, \cdot)$ defined on $S_h^0 \times S_h^0$. The system (6) is symmetric with respect to the inner product defined by

$$(7) \quad [\alpha, \beta] \equiv \sum_{i,j=1}^N B_{ij} \alpha_i \beta_j.$$

Thus the conjugate gradient method may be applied to (6) with respect to (7). The importance of making a “good” choice for B is well known. The matrix B should have two main properties. First, the solution of the problem

$$(8) \quad B\beta = b$$

should be easy to obtain. This is tantamount to applying the operator B^{-1} to the vector b . Secondly, B should be spectrally close to A in the sense that the condition number K of $B^{-1}A$ should not be large. Clearly $K \leq \lambda_1/\lambda_0$, where λ_0 and λ_1 are constants such that

$$\lambda_0[\beta, \beta] \leq [B^{-1}A\beta, \beta] \leq \lambda_1[\beta, \beta].$$

In terms of the form $B(\cdot, \cdot)$ the first property means that the solution W of

$$B(W, \chi) = (g, \chi), \quad \forall \chi \in S_h^0$$

for a given function g should be easier to obtain than the solution of (3). The spectral condition, in terms of the forms, is

$$\lambda_0 B(V, V) \leq A(V, V) \leq \lambda_1 B(V, V)$$

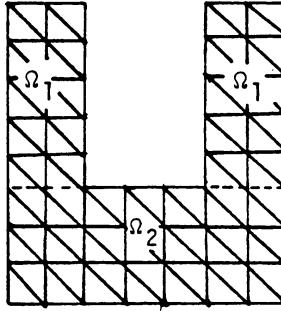
for all $V \in S_h^0$.

These two properties will guarantee, firstly, that the amount of work per step in applying the conjugate gradient method (as an iterative method) will be small, and, secondly, that the number of steps to reduce the error to a given size will be also small so that an efficient algorithm will result.

In this paper we will describe and analyze a technique for constructing the bilinear form $B(\cdot, \cdot)$ (and hence the preconditioner B^{-1}) which only involves solving related Galerkin (or matrix) equations on the subregions Ω_1 and Ω_2 . For other works dealing with the solution of boundary value problems via substructuring cf. [4], [5], [7], [13], [14].

2. The Preconditioning Algorithm. In order to present the ideas clearly, we shall specifically consider the case in which the endpoints of Γ lie on $\partial\Omega$. This is exemplified by case (a) in Figure 1. The approximation subspace S_h^0 of $H_0^1(\Omega)$ is defined by first triangulating Ω , for example as in Figure 2. Then S_h^0 is defined to be the collection of functions which are piecewise linear on the triangles, continuous on Ω and vanish on $\partial\Omega$. Notice that Γ is a “mesh line”.

In order to construct our form $B(\cdot, \cdot)$, we shall need to define two finite element spaces, related to S_h^0 . Let $S_h(\Omega_1)$ be the restrictions to Ω_1 of elements in S_h^0 and let $S_h^0(\Omega_2)$ consist of those elements of S_h^0 which vanish in Ω_1 and in particular on Γ . We shall also need some related bilinear forms defined on $H_0^1(\Omega) \times H_0^1(\Omega)$.



Domain of case (a) with triangulation

FIGURE 2

Let

$$(9) \quad \tilde{A}_k(V, \chi) \equiv \sum_{i,j=1}^2 \int_{\Omega_k} a_{ij}^k \frac{\partial V}{\partial x_i} \cdot \frac{\partial \chi}{\partial x_j} dx, \quad k = 1, 2.$$

Here a_{ij}^k is a positive definite matrix for each k , which may differ from a_{ij} . For example, if the coefficients a_{ij} are variable we may want to choose $a_{ij}^k = a_{ij}(x_k)$, where x_k is some point of Ω_k . In this case the resulting subproblems may be efficiently solved. Set $\tilde{A}(V, \chi) = \tilde{A}_1(V, \chi) + \tilde{A}_2(V, \chi)$ on $S_h^0 \times S_h^0$. Let us now consider an arbitrary function $V \in S_h^0$. We decompose V on Ω_2 as follows. Let $V = V_H + V_P$, where $V_P \in S_h^0(\Omega_2)$ and satisfies

$$\tilde{A}_2(V_P, \chi) = \tilde{A}_2(V, \chi), \quad \forall \chi \in S_h^0(\Omega_2).$$

Notice that V_P is determined on Ω_2 by the values of V on Ω_2 and that

$$\tilde{A}_2(V_H, \chi) = 0, \quad \forall \chi \in S_h^0(\Omega_2).$$

Thus, on Ω_2 , V is decomposed into a function V_P which vanishes on $\partial\Omega_2$ and a function V_H which satisfies the above homogeneous equations. With a slight abuse of terminology we shall refer to such a function V_H as “discrete harmonic”. We now define the bilinear form $B(\cdot, \cdot)$ on $S_h^0 \times S_h^0$ by

$$B(V, \phi) = \tilde{A}_1(V, \phi) + \tilde{A}_2(V_P, \phi_P).$$

We shall show that the corresponding equation

$$(10) \quad B(W, \chi) = (g, \chi), \quad \forall \chi \in S_h^0$$

can be solved by solving related Galerkin equations on Ω_1 and Ω_2 . This is done as follows: Consider $\chi \in S_h^0(\Omega_2)$. Then (10) reduces to

$$\tilde{A}_2(W_P, \chi) = (g, \chi), \quad \forall \chi \in S_h^0(\Omega_2).$$

Since $W_P \in S_h^0(\Omega_2)$ this is just the solution of a discrete Dirichlet problem on Ω_2 . With W_P now known, we write (10) as

$$(11) \quad \tilde{A}_1(W, \chi) = (g, \chi) - \tilde{A}_2(W_P, \chi_P) = (g, \chi) - \tilde{A}_2(W_P, \chi).$$

The last equality follows since $\tilde{A}_2(W_P, \chi_H) = 0$. The equations (11) uniquely determine $W \in S_h(\Omega_1)$. In fact, W is the discrete solution of a mixed Neumann-Dirichlet problem on Ω_1 . Having now determined W on Ω_1 and, in particular, on Γ , we determine W_H as the discrete harmonic function on Ω_2 with values W on Γ and zero on the rest of $\partial\Omega_2$. This involves solving another discrete Dirichlet problem on Ω_2 .

The spectral equivalence of the form $A(\cdot, \cdot)$ and $B(\cdot, \cdot)$ (and hence of the matrices A and B) will now be demonstrated. We shall show that the condition number K is bounded independent of the dimension of S_h^0 . In particular, we shall prove the following

THEOREM. *Let $A(\cdot, \cdot)$, S_h^0 and $B(\cdot, \cdot)$ be defined as above. Then there are positive constants λ_0 and λ_1 independent of h such that*

$$\lambda_0 B(V, V) \leq A(V, V) \leq \lambda_1 B(V, V), \quad \forall V \in S_h^0.$$

Proof. Because of the uniform positive definiteness of the 2×2 matrix $\{a_{ij}\}$ and the positive definiteness of the constant matrices $\{a_{ij}^k\}$, $k = 1, 2$, there are positive constants α_0 and α_1 such that

$$\alpha_0 \tilde{A}(v, v) \leq A(v, v) \leq \alpha_1 \tilde{A}(v, v)$$

for all $v \in H_0^1(\Omega)$. Thus, it remains to be shown that there exist positive constants β_0 and β_1 such that

$$\beta_0 B(V, V) \leq \tilde{A}(V, V) \leq \beta_1 B(V, V)$$

for all $V \in S_h^0$. Clearly,

$$B(V, V) \leq \tilde{A}(V, V),$$

so that $\beta_0 = 1$. Thus, we need to prove that

$$\tilde{A}(V, V) \leq \beta_1 B(V, V),$$

which will obviously follow from the inequality

$$(12) \quad \tilde{A}_2(V_H, V_H) \leq \gamma \tilde{A}_1(V, V)$$

with γ independent of h . The inequality (12) is proved as follows: Let v_H be the restriction to Ω_2 of a function in $H_0^1(\Omega)$ which satisfies

$$v_H = V \quad \text{on } \Gamma$$

and

$$\tilde{A}_2(v_H, \phi) = 0 \quad \forall \phi \in H_0^1(\Omega_2).$$

This is the “harmonic” function in Ω_2 taking the values V on Γ . To estimate V_H , we compare it with v_H and use known estimates. Clearly,

$$\tilde{A}_2(V_H, V_H) \leq 2\tilde{A}_2(V_H - v_H, V_H - v_H) + 2\tilde{A}_2(v_H, v_H).$$

Since v_H vanishes on $\partial\Omega_2/\Gamma$, we have the a priori estimate (cf. [12], [8])

$$\begin{aligned} \tilde{A}_2(v_H, v_H) &\leq C|v_H|_{H^{1/2}(\partial\Omega_2)}^2 \\ &\leq C|V|_{H^{1/2}(\Gamma)}^2 \leq C|V|_{H^{1/2}(\partial\Omega_1)}^2. \end{aligned}$$

The last two inequalities follow from the definition of $\overset{\circ}{H}{}^{1/2}(\Gamma)$ (cf. [11]) and the fact that V vanishes on $\partial\Omega$.

Now from the definition of V_H and v_H it follows easily that

$$\tilde{A}_2(V_H - v_H, V_H - v_H) \leq \inf \tilde{A}_2(\chi - v_H, \chi - v_H)$$

with the infimum taken over functions $\chi \in S_h^0$ with $\chi = V_H$ on Γ . By well-known properties of S_h^0 we see that for $0 < \epsilon < 1/2$,

$$\inf \tilde{A}_2(\chi - v_H, \chi - v_H) \leq Ch^{2\epsilon} \|v_H\|_{H^{1+\epsilon}(\Omega_2)}^2.$$

Now, using a well-known a priori inequality (cf. [12], [8]) and an “inverse property” of S_h^0 (cf. [2]), we see that

$$h^{2\epsilon} \|v_H\|_{H^{1+\epsilon}(\Omega_2)}^2 \leq C h^{2\epsilon} |v_H|_{H^{1/2+\epsilon}(\partial\Omega_2)}^2 \leq C |V|_{H^{1/2}(\partial\Omega_1)}^2.$$

Combining the above estimates yields

$$\tilde{A}_2(V_H, V_H) \leq C |V|_{H^{1/2}(\Omega_1)}^2 \leq C \tilde{A}_1(V, V),$$

the last inequality following by a trace theorem (cf. [12]). This proves (12) which in turn completes the proof of the theorem.

3. Matrix Representation of the Operators. In this section we will describe the preconditioner B in terms of block matrices. It will be shown that B has a special structure and that the process for solving $B\alpha = b$ previously described may also be seen to be a block Gauss elimination process.

We shall suppose that we have the usual nodal basis for S_h^0 and that the nodes are partitioned into three subsets corresponding to those in Γ , Ω_1 and Ω_2 . We shall order the corresponding vectors as follows:

$$\alpha = \begin{pmatrix} v_0 \\ v_1 \\ v_2 \end{pmatrix}$$

with v_0 , v_1 and v_2 corresponding to the nodes on Γ , Ω_1 and Ω_2 , respectively. In terms of block matrices the system corresponding to B is

$$\begin{pmatrix} B_{00} & B_{01} & B_{02} \\ B_{01}^T & B_{11} & 0 \\ B_{02}^T & 0 & B_{22} \end{pmatrix} \begin{pmatrix} v_0 \\ v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix}.$$

Now the first step of the solution process described in Section 2 consists of writing the equivalent system

$$(13) \quad \begin{pmatrix} B_{00} - B_{02}B_{22}^{-1}B_{02}^T & B_{01} & 0 \\ B_{01}^T & B_{11} & 0 \\ B_{02}^T & 0 & B_{22} \end{pmatrix} \begin{pmatrix} v_0 \\ v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} b_0 - B_{02}B_{22}^{-1}b_2 \\ b_1 \\ b_2 \end{pmatrix}.$$

The second step in the process corresponds to the solution of

$$\begin{pmatrix} B_{00}^{(1)} & B_{01} \\ B_{01}^T & B_{11} \end{pmatrix} \begin{pmatrix} v_0 \\ v_1 \end{pmatrix} = \begin{pmatrix} b_0 - B_{02}B_{22}^{-1}b_2 \\ b_1 \end{pmatrix},$$

where the entries of $B_{00}^{(1)}$ are given by $\tilde{A}_1(\phi_i, \phi_j)$ with i and j corresponding to nodes on Γ . Consequently, $B_{00} = B_{00}^{(1)} + B_{02}B_{22}^{-1}B_{02}^T$. Thus, we see that the B has the form

$$B = \begin{pmatrix} B_{00}^{(1)} + B_{02}B_{22}^{-1}B_{02}^T & B_{01} & B_{02} \\ B_{01}^T & B_{11} & 0 \\ B_{02}^T & 0 & B_{22} \end{pmatrix}$$

and the process is just that of block Gauss elimination. The final step in the process corresponds to backsolving (13) for v_2 , once v_0 and v_1 are known.

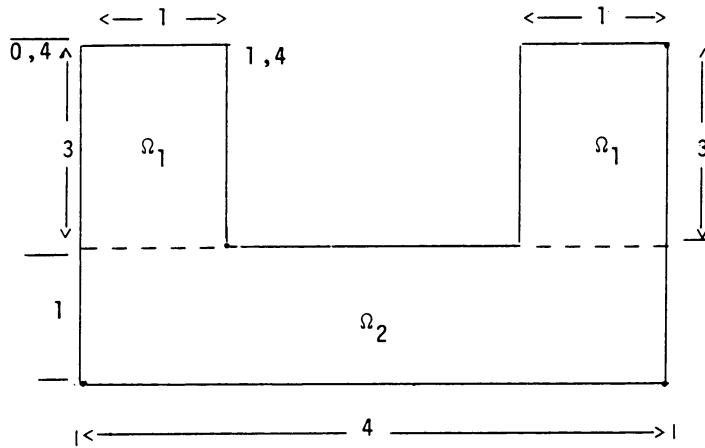
Remark. In the case that the forms $A(\cdot, \cdot)$ and $\tilde{A}(\cdot, \cdot)$ coincide, so that the subproblems for the original form are solved exactly, the present method coincides with that recently studied by Bjørstad and Widlund [3]. In contrast, the general method presented here only assumes the solvability of the subproblems corresponding to the form $\tilde{A}(\cdot, \cdot)$ which we are free to choose. We emphasize that the purpose here is to construct a preconditioner for the original problem which in our case is more general.

4. Applications and Numerical Experiments. In this section we shall present some results of numerical experiments which illustrate the convergence of the iterative algorithm discussed in Section 2. To this end we shall measure K (the condition number of the preconditioned system), the number of iterations required to reduce the iteration error in the L^2 -norm of the residual by some factor ϵ and the average reduction per iteration.

In the two examples considered in this section we shall use subspaces $\{S_h^0\}$ of piecewise linear functions on a rectangular grid of size h . In both examples we shall use the algorithm to solve the finite element equations approximating an elliptic problem of the form

$$\begin{aligned} -\nabla \cdot (a(x, y)\nabla u) &= f \quad \text{in } \Omega, \\ u &= g \quad \text{on } \partial\Omega, \end{aligned}$$

where Ω is the “U” shaped domain given in the following figure.



Example 1. For our first example we chose

$$a(x, y) = 1 + x/2 + y/3.$$

The functions f and g were taken to correspond to the solution $u = \sin x \sin y$. The domain Ω was split into two domains Ω_1 and Ω_2 and the algorithm of Section 2 was applied where the coefficients of the preconditioning form (9) were taken to be piecewise constant in Ω_1 and constant in Ω_2 .

Table 1 gives the average error reduction per iteration and the number of iterations necessary to reduce the initial error by a factor of 10^{-4} . The table clearly indicates that the error reduction is independent of the mesh parameters as theoretically predicted.

TABLE 1

h	Number of Iterations	Average Reduction per Iteration	$K = \frac{\lambda_1}{\lambda_0}$	Number of Unknowns
1/4	7	.23	3.6	108
1/8	7	.22	4.0	532
1/12	7	.22	4.0	1276

Example 2. In this example we study the condition number of the preconditioned system of equations for a problem with discontinuous coefficients. More precisely, we consider solving the above problem with

$$a(x, y) = \begin{cases} 1 & \text{in } \Omega_1, \\ \gamma & \text{in } \Omega_2, \end{cases}$$

where γ is a constant. The functions f and g are chosen so that the solution u is given by

$$\begin{aligned} (x + y)(1 - y)^2 + 3\gamma xy + 3(1 - \gamma)x &\quad \text{in } \Omega_1, \\ (x^2 + y^2)(1 - y)^2 + 3xy &\quad \text{in } \Omega_2. \end{aligned}$$

Table 2 lists the condition number of the preconditioned system for various values of γ . The results are given for $h = 1/12$; almost identical results were obtained for $h = 1/3$ and $h = 1/6$. Note the improved condition number as γ becomes small.

TABLE 2

γ	$K = \frac{\lambda_1}{\lambda_0}$
1	2
.5	1.5
.1	1.1
.05	1.05

Department of Mathematics
White Hall
Cornell University
Ithaca, New York 14853

Applied Mathematics Department
Brookhaven National Laboratory
Upton, New York 11973

Department of Mathematics
White Hall
Cornell University
Ithaca, New York 14853

1. O. AXELSSON, "A class of iterative methods for finite element equations," *Comput. Methods Appl. Mech. Engrg.*, v. 9, 1976, pp. 123-137.

2. I. BABUŠKA & A. K. AZIZ, "Part I. Survey lectures on the mathematical foundations of the finite element method," *The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations* (A. K. Aziz, ed.), Academic Press, New York, 1972.

3. P. E. BJØRSTAD & O. B. WIDLUND, "Solving elliptic problems on regions partitioned into substructures," *Elliptic Problem Solvers II* (G. Birkhoff and A. Schoenstadt, eds.), Academic Press, New York, 1984.
4. B. L. BUZBEE & F. W. DORR, "The direct solution of the biharmonic equation on rectangular regions and the Poisson equation on irregular regions," *SIAM J. Numer. Anal.*, v. 11, 1974, pp. 753–763.
5. B. L. BUZBEE, F. W. DORR, J. A. GEORGE & G. H. GOLUB, "The direct solution of the discrete Poisson equation on irregular regions," *SIAM J. Numer. Anal.*, v. 8, 1971, pp. 722–736.
6. R. CHANDRA, *Conjugate Gradient Methods for Partial Differential Equations*, Yale Univ. Dept. Comp. Sci. Report No. 129, 1978.
7. M. DRYJA, "A capacitance matrix method for the Dirichlet problem on a polygonal region," *Numer. Math.*, v. 39, 1982, pp. 51–64.
8. P. GRISVARD, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.
9. P. CONCUS, G. GOLUB & D. O'LEARY, "A generalized conjugate gradient method for the numerical solution of elliptic partial differential equations," *Sparse Matrix Computations* (J. Bunch and D. Rose, eds.), Academic Press, New York, 1976, pp. 309–322.
10. H. C. ELMAN, *Iterative Methods for Large, Sparse, Nonsymmetric Systems of Linear Equations*, Yale Univ. Dept. Comp. Sci. Report No. 229, 1978.
11. J. L. LIONS & E. MAGENES, *Problèmes aux Limites non Homogènes et Applications*, Vol. 1, Dunod, Paris, 1968.
12. J. NEČAS, *Les Méthodes Directes en Théorie des Équations Elliptiques*, Academia, Prague, 1967.
13. W. PROSKUROWSKI & O. WIDLUND, "A finite element-capacitance matrix method for the Neumann problem for Laplace's equation," *SIAM J. Sci. Statist. Comput.*, v. 1, 1980, pp. 410–426.
14. W. PROSKUROWSKI & O. WIDLUND, "On the numerical solution of Helmholtz's equation by the capacitance matrix method," *Math. Comp.*, v. 20, 1976, pp. 433–468.

4.6 A domain decomposition technique for Stokes problems

A domain decomposition technique for Stokes problems [17]

A DOMAIN DECOMPOSITION TECHNIQUE FOR STOKES PROBLEMS *

James H. BRAMBLE

Cornell University, Ithaca, NY 14850, USA

Joseph E. PASCIAK

Brookhaven National Laboratory, Upton NY 11973, USA

In this paper, we give an analysis for a domain decomposition technique for Stokes problems. The technique involves the application of domain decomposition directly to the Stokes problem and gives rise to an indefinite system for the velocity nodes on the subdomain boundaries and the mean values of the pressure on the subdomains. We analyze the resulting system and show how it can be efficiently solved.

1. Introduction

In this paper, we analyze a domain decomposition technique discussed in [21] for the solution of the discrete systems which arise in finite element approximation to Stokes problems. Specifically, we consider the velocity-pressure formulation of the Stokes equations where the divergence constraint is treated by a Lagrange multiplier technique and the pressure variable corresponds to the multiplier (cf. [15]). The discrete systems which arise are indefinite systems of a special form. In Section 2, we review some properties of these systems and discuss the construction of effective iteration schemes for their solution. The rate of convergence of these iterative techniques will be related to the corresponding “inf-sup” condition.

Section 3 defines the model Stokes problem and gives the corresponding weak formulation. The finite element approximation is then defined in terms of this formulation.

In Section 4, we develop and analyze iterative algorithms for Stokes problems by directly applying domain decomposition to the discrete Stokes systems. We develop iterative algorithms for the solution of the original Stokes system which require the solution of discrete Stokes problems on subdomains at each iterative step. The work in [17] provided insight for the development of this technique. Alternative algorithms for domain decomposition applied to Stokes problems can be found in the proceedings of the first and second international symposium on domain decomposition methods for partial differential equations [12,16].

For clarity of presentation we will only consider the simplest applications and approximation techniques. Many generalizations are possible but will not be addressed here.

* This manuscript has been authored under contract number DE-AC02-76CH00016 with the U.S. Department of Energy. Accordingly, the US Government retains a non-exclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for US Government purposes.

2. Iterative methods for multiplier systems

In this section, we consider two techniques used to develop iterative methods for multiplier systems. The first method is well known and the second is a preconditioning technique discussed in [6]. We included this discussion for completeness and continuity of exposition since it explains how the estimates derived in later sections relate to iterative convergence rates of the resulting algorithms.

Let H^1 and H^2 be finite-dimensional Hilbert spaces and consider the problem

$$M \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} F \\ G \end{pmatrix}, \quad (2.1)$$

where $X, F \in H^1$ and $Y, G \in H^2$. We study operators M of the form,

$$M = \begin{pmatrix} A & B^* \\ B & 0 \end{pmatrix}. \quad (2.2)$$

We assume that A is a positive-definite, symmetric operator on H^1 and that B and B^* are adjoints with respect to the inner products in H^1 and H^2 . We shall use the notation (\cdot, \cdot) and $\|\cdot\|$ to denote the inner products and norms on H^1 and H^2 .

Multiplier systems of the form (2.1) arise in many applications. For example, such systems must be solved for finite element Lagrange multiplier approximations to Dirichlet and interface problems [3,4], velocity-pressure formulations of the equations of Stokes and elasticity [15], and mixed finite element methods [22].

Applying block Gaussian elimination to (2.1) implies that the solution of (2.1) satisfies

$$\begin{pmatrix} A & B^* \\ 0 & BA^{-1}B^* \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} F \\ BA^{-1}F - G \end{pmatrix}. \quad (2.3)$$

Thus, (2.1) is nonsingular if and only if $BA^{-1}B^*$ is invertible. But $BA^{-1}B^*$ is symmetric and nonnegative. Hence, $BA^{-1}B^*$ is invertible if and only if it is definite. A straightforward computation gives

$$(BA^{-1}B^*U, U) = \sup_{\Theta \in H^1} \frac{(B^*U, \Theta)^2}{(A\Theta, \Theta)} \quad \text{for all } U \in H^2, \quad (2.4)$$

and hence solvability of (2.1) will follow if we can verify

$$\sup_{\Theta \in H^1} \frac{(B^*U, \Theta)^2}{(A\Theta, \Theta)} \geq c_0 \|U\|^2 \quad \text{for all } U \in H^2, \quad (2.5)$$

holds for some positive constant c_0 . Inequality (2.5) is equivalent to the classical LBB (Ladyzhenskaya–Babuška–Brezzi) condition (cf. [15]). In addition to being a sufficient condition for the solvability of (2.1), the constant c_0 in (2.5) will be an ingredient in determining convergence rates for the iterative methods to be subsequently discussed.

By (2.3), we see that the solution of (2.1) can be computed by first solving for Y from

$$BA^{-1}B^*Y = BA^{-1}F - G \quad (2.6)$$

and then backsolving (2.3) for X , i.e., $X = A^{-1}(F - B^*Y)$. For our applications, $BA^{-1}B^*$ is a full matrix and expensive to compute. One alternative is to iteratively solve (2.6), e.g., apply

conjugate gradient iteration. The rate of convergence for this iteration is related to the condition number K of $BA^{-1}B^*$. From the above discussion, we clearly have that $K \leq c_1/c_0$ where c_0 satisfies (2.5) and c_1 satisfies the reverse inequality,

$$\sup_{\Theta \in H^1} \frac{(B^*U, \Theta)^2}{(A\Theta, \Theta)} \leq c_1 \|U\|^2 \quad \text{for all } U \in H^2. \quad (2.7)$$

One gets a rapidly convergent algorithm for the computation of Y if the condition number K is not too large. This is the first iterative technique for solving (2.1) to be considered.

One problem with the iterative technique just developed is that it requires the evaluation of the action of A^{-1} at each step in the iteration. In many applications, the action of A^{-1} is more expensive to compute than that of a suitable preconditioner. We next consider a natural preconditioned conjugate gradient technique for solving (2.1) which does not require the evaluation of the action of A^{-1} . An alternative technique (developed and analyzed in [5]) has similar properties but will not be fully discussed here.

We first consider the block operator

$$M_1 = \begin{pmatrix} A & B^* \\ B & 2BA^{-1}B^* \end{pmatrix} = M \begin{pmatrix} I & 2A^{-1}B^* \\ 0 & -I \end{pmatrix}.$$

Clearly, $MM_1^{-1}M = M_1$. Assume that we are given another symmetric positive-definite operator of the form

$$M_0 = \begin{pmatrix} A_0 & 0 \\ 0 & \mathcal{K} \end{pmatrix},$$

with the action of M_0^{-1} easy to obtain. We further assume that

$$\begin{aligned} \alpha_0(M_0\Phi, \Phi) &\leq \left(\begin{pmatrix} A & 0 \\ 0 & BA^{-1}B^* \end{pmatrix} \Phi, \Phi \right) \\ &\leq \alpha_1(M_0\Phi, \Phi) \quad \text{for all } \Phi \in H = H^1 \times H^2, \end{aligned} \quad (2.8)$$

with α_1/α_0 not too large. Here (\cdot, \cdot) denotes the sum of the componentwise inner products. The inequalities (2.8) immediately imply that M_0 is comparable to M_1 . It then follows from the identity $MM_1^{-1}M = M_1$ and the Schwarz inequality that

$$C_0(M_0\Phi, \Phi) \leq (MM_0^{-1}M\Phi, \Phi) \leq C_1(M_0\Phi, \Phi) \quad (2.9)$$

for all $\Phi \in H$. The constants C_0 and C_1 are proportional to α_0^2 and α_1^2 respectively. Note that $M_0^{-1}MM_0^{-1}M$ is a symmetric operator in the inner product $(M_0 \cdot, \cdot)$. Moreover, by (2.9), it is positive-definite and well-conditioned provided that α_1/α_0 is not too large. Thus, applying the conjugate gradient method in the $(M_0 \cdot, \cdot)$ inner product to the problem

$$M_0^{-1}MM_0^{-1}M \begin{pmatrix} X \\ Y \end{pmatrix} = M_0^{-1}MM_0^{-1} \begin{pmatrix} F \\ G \end{pmatrix} \quad (2.10)$$

leads to a rapidly convergent iterative algorithm for solving (2.1).

Remark 2.1. The condition number of the operator on the left-hand side of (2.10) is proportional to the square of $K = \alpha_1/\alpha_0$. The condition number of the reformulation of (2.1) developed in [5] is linear in K however requires estimation of the smallest eigenvalue of $A_0^{-1}A$. Accordingly, if K is not too large, the above method seems simpler.

3. The model Stokes problem

In this section, we describe the model Stokes problem and its finite element discretization. Let Ω be a domain in N -dimensional Euclidean space for $N = 2$ or $N = 3$. The velocity-pressure formulation of the steady-state Stokes problem is: Find \mathbf{u} and P satisfying

$$\begin{aligned} -\Delta \mathbf{u} - \nabla P &= \mathbf{F} && \text{in } \Omega, \\ \nabla \cdot \mathbf{u} &= 0 && \text{in } \Omega, \\ \mathbf{u} &= 0 && \text{on } \partial\Omega, \\ \int_{\Omega} P &= 0. \end{aligned} \tag{3.1}$$

Here, \mathbf{u} is a vector-valued function and P is a scalar-valued function defined on Ω . The first equation is, of course, a vector equality at each $x \in \Omega$ and Δ denotes the componentwise Laplace operator.

We restrict ourselves to the model problem (3.1) for simplicity. Applications to problems with variable coefficients and the equations of linear elasticity are similar.

We consider a weak formulation of problem (3.1). Let (\cdot, \cdot) denote the $L^2(\Omega)$ inner product and $\|\cdot\|$ denote the corresponding norm applied either to scalar or vector functions. Let $H_0^1(\Omega)$ be the Sobolev space of scalar-valued functions defined on Ω which vanish (in an appropriate sense) on $\partial\Omega$ and which along with their first derivatives are square integrable on Ω . Define $\mathbf{H} \equiv H_0^1(\Omega) \times H_0^1(\Omega)$ and let $\|\cdot\|_1$ denote the corresponding norm. Let $\Pi = L^2(\Omega)$ and $\Pi/1$ denote the functions in Π with zero mean value on Ω . Multiplying (3.1) by functions in \mathbf{H} and Π and integrating by parts when appropriate, it is easy to see that the solution (\mathbf{u}, P) satisfies

$$\begin{aligned} D(\mathbf{u}, \mathbf{v}) + (P, \nabla \cdot \mathbf{v}) &= (\mathbf{F}, \mathbf{v}) && \text{for all } \mathbf{v} \in \mathbf{H}, \\ (\nabla \cdot \mathbf{u}, q) &= 0 && \text{for all } q \in \Pi/1. \end{aligned} \tag{3.2}$$

Here, D is the Dirichlet form on Ω defined by

$$D(\mathbf{w}, \mathbf{v}) \equiv \sum_{i=1}^N \int_{\Omega} \nabla w_i \cdot \nabla v_i \, dx.$$

Clearly, (3.2) has the same form as (2.1). The corresponding operator A is unbounded but has a bounded inverse. Moreover, it is well known that the corresponding inf-sup condition:

$$\sup_{\theta \in \mathbf{H}} \frac{(p, \nabla \cdot \theta)^2}{D(\theta, \theta)} \geq C_0 \|p\|^2 \quad \text{for all } p \in \Pi/1 \tag{3.3}$$

holds for some positive constant C_0 (cf. [15]). It then follows that there is a unique solution (\mathbf{u}, P) in $\mathbf{H} \times \Pi/1$ to (3.2).

To approximately solve (3.2), we introduce a collection of pairs of approximation subspaces $\mathbf{H}_h \subset \mathbf{H}$ and $\Pi_h \subset \Pi$ indexed by h in the interval $0 < h < 1$. We will assume that the inf-sup condition holds for the pair of spaces; i.e. we assume that there is a constant c_0 which does not depend upon h such that

$$\sup_{\theta \in \mathbf{H}_h} \frac{(p, \nabla \cdot \theta)^2}{D(\theta, \theta)} \geq c_0 \|p\|^2 \quad \text{for all } p \in \Pi_h/1. \tag{3.4}$$

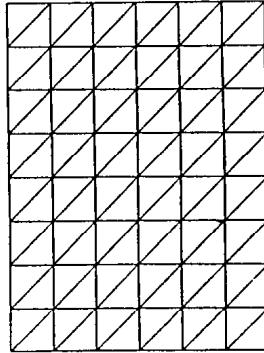
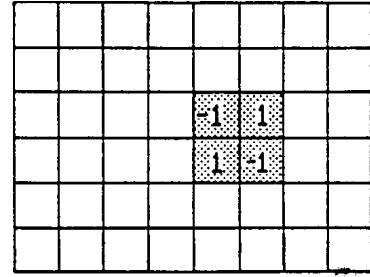


Fig. 1. The triangular mesh.

Fig. 2. The rectangular mesh used for $\tilde{\Pi}_h$; the support (shaded) and values for a typical ϕ_{ij} .

Many subspace pairs satisfying (3.4) have been studied and their approximation properties are well known [15,20,23].

The approximations to the functions (\mathbf{u}, P) are defined by replacing the spaces in (3.2) by their discrete counterparts. Specifically, the approximations are defined as the functions $\mathbf{u}_h \in \mathbf{H}_h$ and $P_h \in \Pi_h/1$ satisfying

$$\begin{aligned} D(\mathbf{u}_h, \mathbf{v}) + (P_h, \nabla \cdot \mathbf{v}) &= (\mathbf{F}, \mathbf{v}) \quad \text{for all } \mathbf{v} \in \mathbf{H}_h, \\ (\nabla \cdot \mathbf{u}_h, q) &= 0 \quad \text{for all } q \in \Pi_h/1. \end{aligned} \quad (3.5)$$

Existence and uniqueness for the solution of (3.5) follows from (3.4) and the discussion in Section 2.

We conclude this section with an example of a pair of approximation subspaces. For simplicity of exposition, we shall only describe these spaces when Ω is the unit square. Generalizations to certain more complex domains are possible.

Let $n > 0$ be given. We start by breaking the square into $2n \times 2n$ square subregions and define $h = 1/2n$ (see Fig. 1). Let $x_i \equiv ih$ and $y_j \equiv jh$ for $i, j = 1, \dots, 2n$. We partition the square subregions into pairs of triangles using one of the square subregions diagonals (for example, the diagonal going from the bottom right corner to the upper left corner of the square subregion). Let H_h be the collection of functions which vanish on the boundary of the square and are piecewise linear and continuous on this triangulation. The subspace \mathbf{H}_h is defined to be $H_h \times H_h$.

To define the space Π_h , we first consider the space $\tilde{\Pi}_h$ which is defined to be the space of functions which are piecewise constant on the square subregions (see Fig. 2). It is interesting to note [18] that the subspace pair $\{\mathbf{H}_h, \tilde{\Pi}_h/1\}$ is not stable in L^2 , i.e., (3.4) fails to hold with c_0 independent of h for the subspace pair. To get a stable pair, we shall consider a somewhat smaller subspace of $\tilde{\Pi}_h$. Let θ_{kl} for $k, l = 1, \dots, 2n$ be the function which is one on the square subregion $[x_{k-1}, x_k] \times [y_{l-1}, y_l]$ and vanishes elsewhere. We define the functions $\phi_{ij} \in \tilde{\Pi}_h$ for $i, j = 1, \dots, n$ by (see also, Fig. 2)

$$\phi_{ij} \equiv \theta_{2i-1, 2j-1} - \theta_{2i, 2j-1} - \theta_{2i-1, 2j} + \theta_{2i, 2j}. \quad (3.6)$$

We then define Π_h by

$$\Pi_h \equiv \{Q \in \tilde{\Pi}_h : (Q, \phi_{ij}) = 0 \text{ for } i, j = 1, \dots, n\}.$$

An estimate of the form of (3.4) holds with c_0 independent of h for the subspace pair $\{\mathbf{H}_h, \Pi_h\}$ [18]. Furthermore, the exclusion of the functions of the form (3.6) does not result in a change in the order of approximation for the space (we obviously still have the subspace of constants on the mesh of size $2h$).

Remark 3.1. The exclusion of functions of the form (3.6) poses no difficulty in practice. In fact, it only affects the definition of the corresponding B in a trivial way. By definition, $B\mathbf{v} \equiv Q$ where $Q \in \Pi_h/1$ solves

$$(Q, R) = (\nabla \cdot \mathbf{v}, R) \quad \text{for all } R \in \Pi_h/1.$$

It is easy to see that Q is the L^2 orthogonal projection (onto $\Pi_h/1$) of the function $\tilde{Q} \in \tilde{\Pi}_h$ satisfying

$$(\tilde{Q}, R) = (\nabla \cdot \mathbf{v}, R) \quad \text{for all } R \in \tilde{\Pi}_h. \quad (3.7)$$

This projection is a trivial local operation since the supports of the functions $\{\phi_{i,j}\}$ are essentially disjoint. Furthermore, the computation of \tilde{Q} is straightforward since the gram matrix for (3.7) is diagonal (with the obvious choice of basis).

The discrete Stokes problem can be cast into the form of (2.1). To see this, we introduce the following notation. Let $A : \mathbf{H}_h \mapsto \mathbf{H}_h$ be defined by

$$(A\mathbf{v}, \mathbf{w}) = D(\mathbf{v}, \mathbf{w}) \quad \text{for all } \mathbf{w} \in \mathbf{H}_h. \quad (3.8)$$

Clearly, (3.8) defines a symmetric positive-definite operator on \mathbf{H}_h . We define $B : \mathbf{H}_h \mapsto \Pi_h/1$ by

$$(B\mathbf{w}, q) = (\nabla \cdot \mathbf{w}, q) \quad \text{for all } q \in \Pi_h/1,$$

which is nothing more than the divergence followed by L^2 projection into Π_h . Its adjoint, $B^* : \Pi_h/1 \mapsto \mathbf{H}_h$ is then defined by

$$(B^*\mathbf{p}, \mathbf{w}) = (p, \nabla \cdot \mathbf{w}) \quad \text{for all } \mathbf{w} \in \mathbf{H}_h.$$

The discrete solution pair (\mathbf{u}_h, P_h) satisfies (2.1).

Remark 3.2. Since the “inf-sup” condition holds for this subspace pair, the second iterative technique of Section 2 (or the technique described in [5]) can be applied by taking

$$M_0 = \begin{pmatrix} A_0 & 0 \\ 0 & I \end{pmatrix},$$

where A_0 is a preconditioner for A . Domain decomposition preconditioners for the general second-order problems have been given in [7–11]. In the remainder of this paper, we consider domain decomposition applied directly to the solution of the discrete Stokes system.

4. A direct domain decomposition approach

In this section, we shall directly apply domain decomposition to the discrete system (3.5). We shall develop algorithms for solving the discrete system (3.5) which only require the solution of

smaller discrete Stokes systems on the subdomains and another reduced system. In this case, the reduced system will be of the form of (2.1) and involve the values of \mathbf{u}_h on the boundary of the subdomains and the mean value of the pressure on the subdomains.

We assume that $\bar{\Omega}$ has been partitioned into a number of subdomains $\bar{\Omega} = \bigcup_{i=1}^m \bar{\Omega}_i$. We require that the boundary of the subdomains ($\Gamma \equiv \bigcup_{i=1}^m \partial\Omega_i$) align with the mesh in \mathbf{H}_h and Π_h . We then define

$$\mathbf{H}_h^i = \{ \phi \in \mathbf{H}_h : \text{support}(\phi) \subset \bar{\Omega}_i \} \quad (4.1)$$

and

$$\Pi_h^i = \{ \phi \in \Pi_h : \text{support}(\phi) \subset \bar{\Omega}_i \}. \quad (4.2)$$

We shall assume that the inf-sup condition holds for each subspace pair, i.e.,

$$\sup_{\theta \in \mathbf{H}_h^i} \frac{(q, \nabla \cdot \theta)^2}{D(\theta, \theta)} \geq c_0 \| q \|_{\Omega_i}^2 \quad \text{for all } q \in \Pi_h^i/1, \quad (4.3)$$

and that the function which is one on Ω_i and vanishes in the remainder of Ω is an element in Π_h . Note that, since the functions in \mathbf{H}_h are continuous, the subspace pair $(\mathbf{H}_h^i, \Pi_h^i)$ can be used to approximate the Stokes problem with zero boundary conditions on the subdomains.

Because of (4.3), local Stokes problems on the subdomains are solvable. The first step is to solve these local problems and reduce the problem to one which implicitly involves fewer degrees of freedom. To do this, we let $(\mathbf{v}_h^i, Q_h^i) \in \mathbf{H}_h^i \times \Pi_h^i/1$ be the solution of

$$\begin{aligned} D(\mathbf{v}_h^i, \mathbf{w}) + (Q_h^i, \nabla \cdot \mathbf{w}) &= (\mathbf{F}, \mathbf{w}) \quad \text{for all } \mathbf{w} \in \mathbf{H}_h^i, \\ (\nabla \cdot \mathbf{v}_{h,q}^i) &= 0 \quad \text{for all } q \in \Pi_h^i/1. \end{aligned} \quad (4.4)$$

We set $\mathbf{v}_h = \sum \mathbf{v}_h^i$, $Q_h = \sum Q_h^i$ and define $\mathbf{w}_h = \mathbf{u}_h - \mathbf{v}_h$ and $R_h = P_h - Q_h$. Then, \mathbf{w}_h and R_h satisfy

$$\begin{aligned} D(\mathbf{w}_h, \mathbf{v}) + (R_h, \nabla \cdot \mathbf{v}) &= F(\mathbf{v}) \quad \text{for all } \mathbf{v} \in \mathbf{H}_h, \\ (\nabla \cdot \mathbf{w}_h, q) &= G(q) \quad \text{for all } q \in \Pi_h/1. \end{aligned} \quad (4.5)$$

The functionals F and G vanish for functions in \mathbf{H}_h^i and $\Pi_h^i/1$ respectively. Thus, the functions \mathbf{w}_h and R_h lie in a subspace of $\mathbf{H}_h \times \Pi_h/1$ with significantly lower dimension. We shall parameterize this subspace and then derive equations for the parameters corresponding to the solution \mathbf{w}_h and R_h .

We shall parameterize the solution (\mathbf{w}_h, R_h) in terms of parameters $\sigma \in \mathbf{H}_h(\Gamma)$ and $\lambda \in \Pi_0$ where

$$\mathbf{H}_h(\Gamma) \equiv \{ \phi|_\Gamma : \phi \in \mathbf{H}_h \}$$

and

$$\Pi_0 \equiv \{ \phi \in \Pi_h/1 : \phi \text{ is constant on } \Omega_i \text{ for each } i \}.$$

To do this, we define the operators $S: \mathbf{H}_h(\Gamma) \mapsto \Pi_h$ and $T: \mathbf{H}_h(\Gamma) \mapsto \mathbf{H}_h$ satisfying the following:

- (1) $S(\gamma)|_{\Omega_i} \in \Pi_h^i/1$,
- (2) $T(\gamma)|_\Gamma = \gamma$,
- (3) $D(T(\gamma), \phi) + (S(\gamma), \nabla \cdot \phi) = 0 \quad \text{for all } \phi \in \mathbf{H}_h^i$,
- (4) $(\nabla \cdot T(\gamma), q) = 0 \quad \text{for all } q \in \Pi_h^i/1$.

It is not difficult to show that the above conditions uniquely define S and T . Moreover, if $\sigma = \mathbf{w}_h|_{\Gamma}$ and $\lambda \in \Pi_0$ is the function which has the same mean values on the subdomains as R_h then it follows directly from the definitions that

$$\mathbf{w}_h = T(\sigma), \quad R_h = S(\sigma) + \lambda. \quad (4.6)$$

Thus, (4.6) gives a parameterization of \mathbf{w}_h and R_h in terms of the parameters (σ, λ) in $\mathbf{H}_h(\Gamma) \times \Pi_0$. Note that given a value of γ , the evaluation of $S(\gamma)$ and $T(\gamma)$ essentially only involves the solution of discrete Stokes problems on the subdomains.

We next give equations for the determination of σ and λ satisfying (4.6). To do this, we define a quadratic form $E : (\mathbf{H}_h(\Gamma) \times \Pi_0)^2 \rightarrow \mathbb{R}^1$ given by

$$E((\gamma_1, \delta_1), (\gamma_2, \delta_2)) = D(T(\gamma_1), T(\gamma_2)) + (\delta_1, \nabla \cdot T(\gamma_2)) + (\nabla \cdot T(\gamma_1), \delta_2). \quad (4.7)$$

Using the definition of T , it is not difficult to see that

$$E((\sigma, \lambda), (\phi, \psi)) = \tilde{F}(\phi, \psi) = F(\bar{\phi}) + G(\psi), \quad (4.8)$$

where $\bar{\phi}$ is any extension of ϕ in \mathbf{H}_h . Thus, given local bases for $\mathbf{H}_h(\Gamma)$ and Π_0 , we can compute the data \tilde{F} satisfying (4.8) using a few operations per basis function.

From the definition of E , it is clear that (4.8) gives rise to a symmetric indefinite system of the form (2.1) which can be used to compute (σ, λ) satisfying (4.6). The form $D(T(\gamma_1), T(\gamma_2))$ corresponds to the operator A in (2.1). The form $(\delta_1, \nabla \cdot T(\gamma_2))$ corresponds to B^* , etc.

Stability properties for the above system are given in the following theorem. We make the further assumption that the velocity subspaces $\mathbf{H}_h|_{\Omega_i}$ satisfy a standard extension property: Given a function $\mathbf{v} \in \mathbf{H}_h(\Gamma)$, there exists $\mathbf{w} \in \mathbf{H}_h$ which equals \mathbf{v} on $\partial\Omega_i$ and satisfies

$$\|\mathbf{w}\|_{H^1(\Omega_i)} \leq c_2 \|\mathbf{v}\|_{1/2, \partial\Omega_i}. \quad (4.9)$$

Here $\|\cdot\|_{1/2, \partial\Omega_i}$ denotes the Sobolev norm of order $\frac{1}{2}$ on $\partial\Omega_i$ and $H^1(\Omega_i)$ denotes the Sobolev norm of order one on Ω_i (cf. [19]). Property (4.9) is known for finite element subspaces defined on quasi-uniform triangulations (cf. [2, 7, 11]). The constant C depends upon the shape of the subdomains but not on h .

Theorem 4.1. *Assume the extension property holds on the subdomains (see (4.9)). Then, there are positive constants α_0, α_1, C_0 such that*

$$\alpha_0 D(T(\gamma), T(\gamma)) \leq \sum_{i=1}^m |\gamma|_{1/2, \partial\Omega_i}^2 \leq \alpha_1 D(T(\gamma), T(\gamma)), \quad (4.10)$$

for all $\gamma \in \mathbf{H}_h(\Gamma)$ and

$$C_0 \|\delta\|^2 \leq \sup_{\gamma \in \mathbf{H}_h(\Gamma)} \frac{(\delta, \nabla \cdot T(\gamma))^2}{D(T(\gamma), T(\gamma))} \leq \|\delta\|^2, \quad (4.11)$$

for all $\delta \in \Pi_0$. Here $|\cdot|_{1/2, \partial\Omega_i}$ denotes the Sobolev seminorm of order $\frac{1}{2}$ on $\partial\Omega_i$. These constants only depend on c_0 in (3.4) and (4.3) and c_2 in (4.9), i.e., not on h or the number of subdomains.

Proof. We first prove (4.10). Given $\gamma \in \mathbf{H}_h(\Gamma)$, let u_γ denote its discrete harmonic extension, i.e., u_γ is the unique function in \mathbf{H}_h which equals γ on Γ and satisfies

$$D(u_\gamma, \phi) = 0 \quad (4.12)$$

for all $\phi \in \mathbf{H}_h$ which vanish on Γ . It is well known [2,7] that if (4.9) holds then on each subdomain

$$cD_i(u_\gamma, u_\gamma) \leq |\gamma|_{1/2, \Omega_i}^2 \leq CD_i(u_\gamma, u_\gamma), \quad (4.13)$$

where D_i denotes the Dirichlet form on Ω_i . Moreover, since u_γ is discrete harmonic,

$$D_i(u_\gamma, u_\gamma) \leq D_i(T(\gamma), T(\gamma)). \quad (4.14)$$

Combining (4.13), (4.14) and summing gives the second inequality of (4.10). For the first inequality, we note that, by the definition of S and T ,

$$D(T(\gamma), T(\gamma)) = D(T(\gamma), u_\gamma) + (S(\gamma), \nabla \cdot u_\gamma). \quad (4.15)$$

By (4.3),

$$\begin{aligned} \|S(\gamma)\|_{\Omega_i}^2 &\leq c_0^{-1} \sup_{\xi \in \mathbf{H}_h} \frac{(S(\gamma), \nabla \cdot \xi)^2}{D_i(\xi, \xi)} \\ &= c_0^{-1} \sup_{\xi \in \mathbf{H}_h} \frac{D_i(T(\gamma), \xi)^2}{D_i(\xi, \xi)} \leq c_0^{-1} D_i(T(\gamma), T(\gamma)). \end{aligned} \quad (4.16)$$

Applying the Schwarz inequality to (4.15) and using (4.16) gives

$$D(T(\gamma), T(\gamma)) \leq cD(u_\gamma, u_\gamma).$$

The first inequality of (4.10) then follows from (4.13).

We next prove (4.11). The second inequality follows immediately from the Schwarz inequality. For the first, by (3.4),

$$\|\delta\|^2 \leq c_0^{-1} \sup_{\xi \in \mathbf{H}_h} \frac{(\delta, \nabla \cdot \xi)^2}{D(\xi, \xi)}. \quad (4.17)$$

Moreover, the above inequalities imply that for $\gamma = \xi|_\Gamma$,

$$D(T(\gamma), T(\gamma)) \leq cD(u_\gamma, u_\gamma) \leq cD(\xi, \xi). \quad (4.18)$$

In addition, since δ is constant on the subdomains

$$(\delta, \nabla \cdot \xi) = (\delta, \nabla \cdot T(\gamma)). \quad (4.19)$$

Combining (4.17)–(4.19) proves the first inequality of (4.11). This completes the proof of (4.11). \square

Inequalities (4.11) imply that the operator $BA^{-1}B^*$ corresponding to E on the subspace Π_0 is well conditioned independently of h . The boundary form $D(T(\gamma), T(\gamma))$ is not well conditioned but is equivalent to a sum of seminorms on the boundaries of the subdomains. The corresponding form,

$$\langle\langle \gamma_1, \gamma_2 \rangle\rangle_{1/2} \equiv \sum_{i=1}^m \langle \gamma_1, \gamma_2 \rangle_{1/2, \partial\Omega_i} \quad \text{for all } \gamma_1, \gamma_2 \in \mathbf{H}_h(\Gamma),$$

has been studied in the development of domain decomposition preconditioners for second-order

problems. Each domain decomposition technique developed in [1,2,7–11,13,14] gives rise to a computationally effective domain decomposition preconditioner for $\langle\langle \cdot, \cdot \rangle\rangle_{1/2}$. Thus, we can solve (4.8) by using the second iterative technique of Section 2, with preconditioner

$$M_0^{-1} = \begin{pmatrix} A_0 & 0 \\ 0 & I \end{pmatrix}^{-1},$$

where A_0 corresponds componentwise to the boundary part of a second-order method developed in [1,2,7–11,13,14]. For example, we can use the technique presented in [8]. This means that the preconditioner for the boundary velocities will involve inverting the $l_0^{1/2}$ operator on the edge segments and the solution of a coarse grid problem with the number of unknowns equal to the number of “cross-points” in the subdomain subdivision. The resulting symmetric positive-definite system (2.10) will have a condition number bounded by $C(1 + \ln^4(d/h))$.

It is possible to implement the above technique in such a way that each Stokes subdomain problem need be solved only once per step in the iterative algorithm for the solution of σ , λ . Once these parameters are solved to satisfactory accuracy, w_h and R^h can be computed with one more set of subdomain solves. Moreover, the action of A_0 appearing in the inner product $(M_0 \cdot, \cdot)$ need never be explicitly computed in the conjugate gradient algorithm (e.g. see [5, Appendix]).

References

- [1] V.I. Agoshkov, Poincaré-Steklov's operators and domain decomposition method in finite dimensional spaces, in: R. Glowinski, G.H. Golub, G.A. Meurant and J. Périaux, eds., *Proceedings First International Symposium on Domain Decomposition Methods for Partial Differential Equations* (SIAM, Philadelphia, PA, 1988) 73–112.
- [2] P.E. Bjørstad and O.B. Widlund, Iterative methods for the solution of elliptic problems on regions partitioned into substructures, *SIAM J. Numer. Anal.* 23 (1986) 1097–1120.
- [3] J.H. Bramble, The Lagrange multiplier method for Dirichlet's problem, *Math. Comp.* 37 (1981) 1–12.
- [4] J.H. Bramble and J.E. Pasciak, A boundary parametric approximation to the linearized scalar potential magnetostatic field problem, *Appl. Numer. Math.* 1 (1985) 493–514.
- [5] J.H. Bramble and J.E. Pasciak, A preconditioning technique for indefinite systems resulting from mixed approximations of elliptic problems, *Math. Comp.* 50 (1988) 1–18.
- [6] J.H. Bramble and J.E. Pasciak, A preconditioning technique for multiplier problems (in preparation).
- [7] J.H. Bramble, J.E. Pasciak and A.H. Schatz, An iterative method for elliptic problems on regions partitioned into substructures, *Math. Comp.* 46 (1986) 361–369.
- [8] J.H. Bramble, J.E. Pasciak and A.H. Schatz, The construction of preconditioners for elliptic problems by substructuring, I, *Math. Comp.* 47 (1986) 103–134.
- [9] J.H. Bramble, J.E. Pasciak and A.H. Schatz, The construction of preconditioners for elliptic problems by substructuring, II, *Math. Comp.* 49 (1987) 1–16.
- [10] J.H. Bramble, J.E. Pasciak and A.H. Schatz, The construction of preconditioners for elliptic problems by substructuring, III, *Math. Comp.* 51 (1988) 415–430.
- [11] J.H. Bramble, J.E. Pasciak and A.H. Schatz, The construction of preconditioners for elliptic problems by substructuring, IV, *Math. Comp.* 53 (1989) 1–24.
- [12] T.F. Chan, R. Glowinski, J. Périaux and O.B. Widlund, *Domain Decomposition Methods* (SIAM, Philadelphia, PA, 1989).
- [13] M. Dryja, A capacitance matrix method for Dirichlet problems on polygonal domains, *Numer. Math.* 39 (1982) 51–64.
- [14] M. Dryja, A finite element-capacitance matrix method for elliptic problems in regions partitioned into subregions, *Numer. Math.* 44 (1984) 153–168.

- [15] V. Girault and P. Raviart, *Finite Element Approximation of the Navier–Stokes Equations*, Lecture Notes in Mathematics 749 (Springer, New York, 1981).
- [16] R. Glowinski, G.H. Golub, G.A. Meurant and J. Périaux eds., *Proceedings First International Symposium on Domain Decomposition Methods for Partial Differential Equations* (SIAM, Philadelphia, PA, 1988).
- [17] R. Glowinski and M.F. Wheeler, Domain decomposition methods for mixed finite element approximation, in: R. Glowinski, G.H. Golub, G.A. Meurant and J. Périaux, eds. *Proceedings First International Symposium on Domain Decomposition Methods for Partial Differential Equations* (SIAM, Philadelphia, PA, 1988) 144–172.
- [18] C. Johnson and J. Pitkäranta, Analysis of some mixed finite element methods related to reduced integration, *Math. Comp.* 38 (1982) 375–400.
- [19] J.L. Lions and E. Magenes, *Problèmes aux Limites non Homogènes et Applications* (Dunod, Paris, 1968).
- [20] J.C. Nedelec, Elements finis mixtes incompressibles pour l'équation de Stokes dans R^3 , *Numer Math.* 39 (1982) 97–112.
- [21] J.E. Pasciak, Two domain decomposition techniques for Stokes problems, in: T.F. Chan, R. Glowinski, J. Périaux and O.B. Widlund, eds., *Domain Decomposition Methods* (SIAM, Philadelphia, PA, 1989) 419–430.
- [22] P.A. Raviart and J.M. Thomas, A mixed finite element method for 2nd order elliptic problems, in: I. Galligani and E. Magenes, eds., *Mathematical Aspects of Finite Element Methods*, Lecture Notes in Mathematics 606 (Springer, New York, 1977) 292–315.
- [23] L.R. Scott and M. Vogelius, Conforming finite element methods for incompressible and nearly incompressible continua, Tech. Rept. BN-1018, Institute for Physical Science and Technology, University of Maryland, College Park, MD (1984).

4.7 Convergence estimates for product iterative methods with applications to domain decomposition

Convergence estimates for product iterative methods with applications to domain decomposition [24]

CONVERGENCE ESTIMATES FOR PRODUCT ITERATIVE METHODS WITH APPLICATIONS TO DOMAIN DECOMPOSITION

JAMES H. BRAMBLE, JOSEPH E. PASCIAK, JUNPING WANG, AND JINCHAO XU

ABSTRACT. In this paper, we consider iterative methods for the solution of symmetric positive definite problems on a space \mathcal{V} which are defined in terms of products of operators defined with respect to a number of subspaces. The simplest algorithm of this sort has an error-reducing operator which is the product of orthogonal projections onto the complement of the subspaces. New norm-reduction estimates for these iterative techniques will be presented in an abstract setting. Applications are given for overlapping Schwarz algorithms with many subregions for finite element approximation of second-order elliptic problems.

1. INTRODUCTION

In this paper, we shall be concerned with solving problems in an abstract Hilbert space \mathcal{V} with inner product $A(\cdot, \cdot)$. Denote by \mathcal{V}' the dual of the Hilbert space \mathcal{V} with $\langle \chi, \cdot \rangle$ being the action of $\chi \in \mathcal{V}'$ on \mathcal{V} . Given a function $f \in \mathcal{V}'$, we seek the solution $u \in \mathcal{V}$ of the equation

$$(1.1) \quad A(u, \phi) = \langle f, \phi \rangle \quad \text{for all } \phi \in \mathcal{V}.$$

We will consider iterative methods for (1.1) based on a sequence of subspaces. To this end, let $\mathcal{V}_1, \dots, \mathcal{V}_J$ denote closed subspaces of \mathcal{V} and define $A_i: \mathcal{V}_i \mapsto \mathcal{V}'$ by

$$\langle A_i v, \phi \rangle = A(v, \phi) \quad \text{for all } v, \phi \in \mathcal{V}_i.$$

Denote by A the corresponding operator defined on the whole space \mathcal{V} . Thus, the problem (1.1) is equivalent to the problem of finding $u \in \mathcal{V}$ such that

$$(1.2) \quad Au = f.$$

Received February 5, 1990.

1980 *Mathematics Subject Classification* (1985 Revision). Primary 65N30; Secondary 65F10.

Key words and phrases. Second-order elliptic equation, domain decomposition.

This manuscript has been authored under contract number DE-AC02-76CH00016 with the U.S. Department of Energy. Accordingly, the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. This work was also supported in part under the National Science Foundation Grant No. DMS84-05352, DMS88-05311-04 and by the U.S. Army Research Office through the Mathematical Science Institute, Cornell University. The third author's research was supported by Office of Naval Research Contract No. 0014-88-K-0370 and by the Institute for Scientific Computation of the University of Wyoming through National Science Foundation Grant No. RII-8610680.

©1991 American Mathematical Society
0025-5718/91 \$1.00 + \$.25 per page

We assume that linear operators $R_i : \mathcal{V}'_i \mapsto \mathcal{V}_i$ for $i = 1, \dots, J$ are given. In practice, the operator R_i will in some sense “approximately invert” A_i . We will consider iterative methods of the following type for solving (1.1).

Algorithm 1. Given $u^l \in \mathcal{V}$, an approximation to the solution u of (1.2), we define the next iterate $u^{l+1} \in \mathcal{V}$ as follows:

- (1) Set $Y_0 = u^l$.
- (2) For $i = 1, \dots, J$ define Y_i by

$$Y_i = Y_{i-1} + R_i Q_i (f - AY_{i-1}),$$

where Q_i denotes the projection onto the subspace \mathcal{V}'_i defined by

$$\langle \chi - Q_i \chi, \phi \rangle = 0 \quad \text{for all } \phi \in \mathcal{V}_i.$$

- (3) Set $u^{l+1} = Y_J$.

Let $e_0 = u - u^l$ and $e_i = u - Y_i$ for $i = 1, \dots, J$. Let P_i denote the orthogonal projection into the subspace \mathcal{V}_i , i.e., $P_i v = w$ where w is the unique function in \mathcal{V}_i satisfying

$$A(w, \phi) = A(v, \phi) \quad \text{for all } \phi \in \mathcal{V}_i.$$

A simple computation shows that $Q_i A = A_i P_i$ and hence $e_i = (I - R_i A_i P_i) e_{i-1}$. Consequently,

$$(1.3) \quad u - u^{l+1} = (I - T_J)(I - T_{J-1}) \cdots (I - T_1)(u - u^l),$$

where $T_i = R_i A_i P_i$. Thus, estimates for the rate of convergence for the iterative method follow directly from norm bounds on the product in (1.3). The purpose of this paper is to provide new techniques for estimating norms of products of operators of the form appearing on the right-hand side of (1.3).

One natural example of a choice of R_i is $R_i = A_i^{-1}$. In this case the above product reduces to a product of orthogonal projections onto the complements of the subspaces. Note, however, that the action of the inverse of A_i must be computed as part of the iterative procedure. For this reason, it is often more efficient to use various preconditioners for A_i . This will be clearly illustrated in the applications to be presented.

The most important application of the results of this paper is to the computation of the solutions of the discrete equations which result from the numerical approximation of elliptic boundary value problems. In the case of finite element approximation, the inner product $A(\cdot, \cdot)$ is the form corresponding to the differential operator and \mathcal{V} is the finite element approximation space. The subspaces \mathcal{V}_i are either associated with subdomains in domain decomposition applications or coarser grids in multilevel applications. We shall discuss primarily the applications to domain decomposition in this paper, even though the theorems of this paper could be directly applied to provide new estimates for multigrid algorithms. However, it is possible to modify the analysis presented

here and develop sharper estimates for multigrid algorithms. This is done in [10] and provides a theory for multigrid algorithms which does not require regularity estimates for the underlying boundary value problem.

Iterative algorithms involving a product of projectors have been studied by other researchers (cf. [15], [17], [19]). The case of two subspaces has been thoroughly understood (cf. [15], [16]). The results for greater than two projectors only gave either that the methods were convergent (without any estimate on the rate of convergence [15]) or provided a convergence rate which approached 1 faster than exponentially with the number of projectors [17], [19]. Our analysis provides much better bounds for the convergence rate of these methods. In fact, there are applications where our analysis shows that the convergence rate remains bounded away from one even though the number of projectors becomes large (cf. §4).

Another important aspect of our analysis is that it shows that the projectors can be replaced by properly scaled preconditioners without significant deterioration in the convergence rates. Our result applies to the many-level case and to other applications as well.

The paper is organized in the following way. We will provide an abstract analysis for estimating the norms of operators of the form of (1.3) in §2. In §3, we provide estimates for the sum of projectors and define the additive variant of Algorithm 1. In §4, we apply the abstract results of §2 to the multiplicative iterative methods resulting from overlapping domain decomposition. The methods discussed there are extensions of the classical Schwarz alternating algorithm. Finally, the results of numerical experiments illustrating the rapid convergence of the product algorithms will be given in §5.

2. ABSTRACT ANALYSIS OF PRODUCT ALGORITHMS

In this section, we shall present an abstract analysis of products of operators of the form appearing in (1.3). To this end, we assume that we are given a Hilbert space \mathcal{V} with inner product (\cdot, \cdot) and a sequence of linear operators $\{T_i\}$ mapping \mathcal{V} into \mathcal{V} for $i = 1, \dots, J$ which are selfadjoint, positive semidefinite and of norm bounded by a constant $\omega < 2$. The main results of this section are Theorems 2.1 and 2.2, which give explicit bounds for the norm of the product operator in (1.3) in terms of a number of assumptions to be described.

To begin our analysis, let

$$(2.1) \quad E_i = (I - T_i)(I - T_{i-1}) \cdots (I - T_1)$$

for $i = 1, \dots, J$. For convenience, we let $E_0 = I$, the identity operator on \mathcal{V} , and $E = E_J$. We clearly have for $i = 1, \dots, J$

$$(2.2) \quad E_{i-1} - E_i = T_i E_{i-1},$$

from which it follows that

$$(2.3) \quad I - E_i = \sum_{j=1}^i T_j E_{j-1}.$$

The following lemma will be a fundamental part of the analysis of this section.

Lemma 2.1. *Let T_i and E_i , $i = 1, \dots, J$, be as above. Then*

$$(2.4) \quad (2 - \omega) \sum_{i=1}^J (T_i E_{i-1} v, E_{i-1} v) \leq \|v\|^2 - \|Ev\|^2,$$

where $\|\cdot\| = (\cdot, \cdot)^{1/2}$.

Proof. It is obvious from (2.2) that

$$(2.5) \quad \|E_{i-1} v\|^2 - \|E_i v\|^2 = \|T_i E_{i-1} v\|^2 + 2(T_i E_{i-1} v, E_i v).$$

We note that

$$(T_i E_{i-1} v, E_i v) = (T_i (I - T_i) E_{i-1} v, E_{i-1} v),$$

and hence the right-hand side of (2.5) can be bounded by

$$(2.6) \quad \begin{aligned} \|T_i - E_{i-1} v\|^2 + 2(T_i E_{i-1} v, E_i v) &= ((2I - T_i) T_i E_{i-1} v, E_{i-1} v) \\ &\geq (2 - \omega)(T_i E_{i-1} v, E_{i-1} v). \end{aligned}$$

Combining (2.5), (2.6) and summing gives (2.4). This completes the proof of the lemma. \square

A fundamental assumption for the analysis to be presented in this section involves an inequality regarding the sum of the operators $\{T_i\}$. Specifically, we assume that there is a positive constant C_0 satisfying

$$(2.7) \quad \|v\|^2 \leq C_0 \sum_{i=1}^J (T_i v, v) \quad \text{for all } v \in \mathcal{V}.$$

We can now state and prove the first theorem of this section.

Theorem 2.1. *Assume that (2.7) holds. Then*

$$(2.8) \quad \|Ev\|^2 \leq \gamma \|v\|^2$$

for

$$(2.9) \quad \gamma = 1 - \frac{2 - \omega}{C_0(J + \omega^2 J(J - 1)/2)}$$

or

$$(2.10) \quad \gamma = 1 - \frac{2 - \omega}{2C_0(1 + \omega^2 J(J - 1)/2)}.$$

Proof. We note that it clearly suffices to prove that for $v \in \mathcal{V}$

$$\|v\|^2 \leq C_0(J + \omega^2 J(J - 1)/2)/(2 - \omega)(\|v\|^2 - \|Ev\|^2)$$

and

$$\|v\|^2 \leq 2C_0(1 + \omega^2 J(J-1)/2)/(2 - \omega)(\|v\|^2 - \|Ev\|^2).$$

Applying Lemma 2.1 and (2.7), we see that the theorem will be proved if we can show

$$(2.11) \quad \sum_{i=1}^J (T_i v, v) \leq (J + \omega^2 J(J-1)/2) \sum_{i=1}^J (T_i E_{i-1} v, E_{i-1} v)$$

and

$$(2.12) \quad \sum_{i=1}^J (T_i v, v) \leq 2(1 + \omega^2 J(J-1)/2) \sum_{i=1}^J (T_i E_{i-1} v, E_{i-1} v).$$

By (2.3),

$$(2.13) \quad (T_i v, v) = (T_i v, E_{i-1} v) + \sum_{j=1}^{i-1} (T_i v, T_j E_{j-1} v).$$

Using the bound for the operators $\{T_i\}$ and the Schwarz inequality gives

$$(2.14) \quad \begin{aligned} (T_i v, v) &\leq (T_i v, v)^{1/2} (T_i E_{i-1} v, E_{i-1} v)^{1/2} \\ &\quad + \omega(T_i v, v)^{1/2} \sum_{j=1}^{i-1} (T_j E_{j-1} v, E_{j-1} v)^{1/2} \\ &\leq (T_i v, v)^{1/2} \left((T_i E_{i-1} v, E_{i-1} v)^{1/2} \right. \\ &\quad \left. + \omega \sqrt{i-1} \left(\sum_{j=1}^{i-1} (T_j E_{j-1} v, E_{j-1} v) \right)^{1/2} \right). \end{aligned}$$

Thus, by first eliminating $(T_i v, v)$ on the right-hand side and then applying the Schwarz inequality, (2.14) can be bounded as follows:

$$(2.15) \quad \begin{aligned} (T_i v, v) &\leq \left((T_i E_{i-1} v, E_{i-1} v)^{1/2} \right. \\ &\quad \left. + \omega \sqrt{i-1} \left(\sum_{j=1}^{i-1} (T_j E_{j-1} v, E_{j-1} v) \right)^{1/2} \right)^2 \\ &\leq (1 + \omega^2(i-1)) \sum_{j=1}^J (T_j E_{j-1} v, E_{j-1} v). \end{aligned}$$

Alternatively,

$$(2.16) \quad (T_i v, v) \leq 2 \left((T_i E_{i-1} v, E_{i-1} v) + \omega^2(i-1) \left(\sum_{j=1}^{i-1} (T_j E_{j-1} v, E_{j-1} v) \right) \right).$$

Summing (2.15) and (2.16) over i proves (2.11) and (2.12). This completes the proof of the theorem. \square

Remark 2.1. Note that (2.9) provides a better estimate when, for example, $\omega \geq 1$. In contrast, (2.10) provides a better estimate when ω is small and $j > 2$. In fact, the form of estimate (2.10) suggests that it may be possible to accelerate the convergence of the algorithms by scaling the T_i 's. Assume that we are given a sequence of symmetric positive semidefinite operators \tilde{T}_i , $i = 1, \dots, J$, with norm bounded by ω and satisfying

$$\|v\|^2 \leq \tilde{C}_0 \sum_{i=1}^J (\tilde{T}_i v, v) \quad \text{for all } v \in \mathcal{V}.$$

Defining $T_i = \alpha \tilde{T}_i$ and applying (2.10) gives that the reduction corresponding to the algorithm with $\{T_i\}$ is bounded, for example, by

$$\gamma = 1 - \frac{\alpha(2 - \omega\alpha)}{2\tilde{C}_0(1 + \alpha^2\omega^2 J(J-1)/2)}.$$

Taking $\alpha = 1/(\omega\sqrt{J(J-1)})$ gives that the reduction rate is bounded, for example, by

$$\gamma = 1 - \frac{1}{3\tilde{C}_0\omega J}.$$

Remark 2.2. One important application of the theorems of this section is the case where the operator $T_i = P_i$, the orthogonal projection onto a subspace \mathcal{V}_i of \mathcal{V} . In this case, the proof of (2.7) reduces to the construction of a decomposition of $v \in \mathcal{V}$ of the form $v = \sum_{i=1}^J v_i$ with $v_i \in \mathcal{V}_i$ satisfying

$$(2.17) \quad \sum_{i=1}^J \|v_i\|^2 \leq C_0 \|v\|^2.$$

This was observed in [15]. In fact, if (2.17) holds, then

$$\begin{aligned} \|v\|^2 &= \sum_{i=1}^J (v, v_i) = \sum_{i=1}^J (P_i v, v_i) \\ &\leq \left(\sum_{i=1}^J \|v_i\|^2 \right)^{1/2} \left(\sum_{i=1}^J (T_i v, v) \right)^{1/2} \leq C_0 \sum_{i=1}^J (T_i v, v). \end{aligned}$$

Theorem 2.1 provides good bounds for a number of applications. However, there is an important class of applications (see §4) where $\{T_i\}$ satisfies an interaction property. To describe this property, we first define

$$\kappa_{ij} = \begin{cases} 0 & \text{if } T_i T_j = 0, \\ 1 & \text{otherwise.} \end{cases}$$

We consider a set $I_0 \subseteq [1, \dots, J]$ and define J_0 to be the number of integers in I_0 and

$$N_0 = \max_{j \notin I_0} \sum_{i \notin I_0} \kappa_{ij}.$$

It can happen that the numbers J_0 and N_0 remain small even when J becomes large. Basically, the indices in I_0 correspond to T_i 's which interact with many of the other T_i 's. The remaining T_i 's interact with at most N_0 of the T_i 's with indices not in I_0 . We shall also let \tilde{I}_0 denote the integers in $[1, \dots, J]$ which are not in I_0 . Under these assumptions, we have the following theorem.

Theorem 2.2. *Let N_0 and J_0 be defined as above. If (2.7) holds, then*

$$\|Ev\|^2 \leq \gamma \|v\|^2$$

for

$$\gamma = 1 - \frac{2 - \omega}{3C_0(1 + \omega^2(J_0 + N_0)K_0)},$$

where $K_0 = \max(J_0, N_0)$.

Before proving the theorem, we prove the following lemma.

Lemma 2.2. *Let J_0 and N_0 be defined as above and $u_i, v_i \in \mathcal{M}$ for $i = 1, \dots, J$. If S_1 is a subset of $\tilde{I}_0 \times \tilde{I}_0$, then*

$$(2.18) \quad \left(\sum_{(i,j) \in S_1} (T_i u_i, T_j v_j) \right)^2 \leq \omega^2 N_0^2 \sum_{i \in \tilde{I}_0} (T_i u_i, u_i) \sum_{j \in \tilde{I}_0} (T_j v_j, v_j).$$

If S_2 is a subset of $I_0 \times I_0$, then

$$(2.19) \quad \left(\sum_{(i,j) \in S_2} (T_i u_i, T_j v_j) \right)^2 \leq \omega^2 J_0^2 \sum_{i \in I_0} (T_i u_i, u_i) \sum_{j \in I_0} (T_j v_j, v_j).$$

Finally, if S_3 is a subset of $I_0 \times \tilde{I}_0$, then

$$(2.20) \quad \left(\sum_{(i,j) \in S_3} (T_i u_i, T_j v_j) \right)^2 \leq \omega^2 J_0 N_0 \sum_{i \in I_0} (T_i u_i, u_i) \sum_{j \in \tilde{I}_0} (T_j v_j, v_j).$$

Proof. The bound on the $\{T_i\}$ implies that

$$|(T_i u_i, T_j v_j)| \leq \omega \kappa_{ij} (T_i u_i, u_i)^{1/2} (T_j v_j, v_j)^{1/2}.$$

Consequently,

$$\begin{aligned} \left(\sum_{(i,j) \in S_1} (T_i u_i, T_j v_j) \right)^2 &\leq \omega^2 \sum_{i \in \tilde{I}_0} (T_i u_i, u_i) \sum_{i \in \tilde{I}_0} \left(\sum_{j \in \tilde{I}_0} \kappa_{ij} (T_j v_j, v_j)^{1/2} \right)^2 \\ &\leq \omega^2 N_0 \sum_{i \in \tilde{I}_0} (T_i u_i, u_i) \sum_{i \in \tilde{I}_0} \sum_{j \in \tilde{I}_0} \kappa_{ij} (T_j v_j, v_j) \\ &\leq \omega^2 N_0^2 \sum_{i \in \tilde{I}_0} (T_i u_i, u_i) \sum_{j \in \tilde{I}_0} (T_j v_j, v_j). \end{aligned}$$

This proves (2.18). The proof of (2.19) is similar.

Let $\tilde{I}_0(i)$ denote the possibly empty set of indices $j \in \tilde{I}_0$ such that $(i, j) \in S_3$. To prove (2.20),

$$\begin{aligned}
\left(\sum_{(i,j) \in S_3} (T_i u_i, T_j v_j) \right)^2 &= \left(\sum_{i \in I_0} \left(T_i u_i, \sum_{j \in \tilde{I}_0(i)} T_j v_j \right) \right)^2 \\
(2.21) \quad &\leq J_0 \sum_{i \in I_0} \|T_i u_i\|^2 \left\| \sum_{j \in \tilde{I}_0(i)} T_j v_j \right\|^2 \\
&\leq \omega J_0 \sum_{i \in I_0} (T_i u_i, u_i) \left\| \sum_{j \in \tilde{I}_0(i)} T_j v_j \right\|^2.
\end{aligned}$$

We clearly have that for any i , $\tilde{I}_0(i) \subseteq \tilde{I}_0$, and hence (2.18) implies

$$(2.22) \quad \left\| \sum_{j \in \tilde{I}_0(i)} T_j v_j \right\|^2 \leq \omega N_0 \sum_{j \in \tilde{I}_0} (T_j v_j, v_j).$$

Combining (2.21) and (2.22) proves (2.20). \square

Proof of Theorem 2.2. As in the proof of Theorem 2.1 (compare with (2.11)), it suffices to prove that for $v \in \mathcal{V}$

$$(2.23) \quad \sum_{i=1}^J (T_i v, v) \leq 3(1 + \omega^2(J_0 + N_0)K_0) \sum_{i=1}^J (T_i E_{i-1} v, E_{i-1} v).$$

Using (2.13) and partitioning the (i, j) indices appearing in the double sum into sets S_2 and S_3 yield

$$\begin{aligned}
\sum_{i \in I_0} (T_i v, v) &= \sum_{i \in I_0} (T_i v, E_{i-1} v) + \sum_{i \in I_0} \sum_{j=1}^{i-1} (T_i v, T_j E_{j-1} v) \\
&= \sum_{i \in I_0} (T_i v, E_{i-1} v) + \sum_{(i,j) \in S_2} (T_i v, T_j E_{j-1} v) \\
&\quad + \sum_{(i,j) \in S_3} (T_i v, T_j E_{j-1} v),
\end{aligned}$$

where

$$S_2 \subseteq I_0 \times I_0 \quad \text{and} \quad S_3 \subseteq I_0 \times \tilde{I}_0.$$

Thus, by the arithmetic-geometric mean inequality and Lemma 2.2,

$$\begin{aligned}
 & \left(\sum_{i \in I_0} (T_i v, v) \right)^2 \leq 3 \left\{ \sum_{i \in I_0} (T_i v, v) \sum_{i \in I_0} (T_i E_{i-1} v, E_{i-1} v) \right. \\
 & \quad + \left[\sum_{(i,j) \in S_2} (T_i v, T_j E_{j-1} v) \right]^2 \\
 & \quad \left. + \left[\sum_{(i,j) \in S_3} (T_i v, T_j E_{j-1} v) \right]^2 \right\} \\
 (2.24) \quad & \leq 3 \sum_{i \in I_0} (T_i v, v) \left\{ \sum_{i \in I_0} (T_i E_{i-1} v, E_{i-1} v) \right. \\
 & \quad + \omega^2 J_0^2 \sum_{i \in I_0} (T_i E_{i-1} v, E_{i-1} v) \\
 & \quad \left. + \omega^2 J_0 N_0 \sum_{i \in \tilde{I}_0} (T_i E_{i-1} v, E_{i-1} v) \right\}.
 \end{aligned}$$

It follows from (2.24) that

$$\begin{aligned}
 \sum_{i \in I_0} (T_i v, v) \leq 3 \left\{ \sum_{i \in I_0} (T_i E_{i-1} v, E_{i-1} v) + \omega^2 J_0^2 \sum_{i \in I_0} (T_i E_{i-1} v, E_{i-1} v) \right. \\
 (2.25) \quad \left. + \omega^2 J_0 N_0 \sum_{i \in \tilde{I}_0} (T_i E_{i-1} v, E_{i-1} v) \right\}.
 \end{aligned}$$

Similarly,

$$\begin{aligned}
 \sum_{i \in \tilde{I}_0} (T_i v, v) \leq 3 \left\{ \sum_{i \in \tilde{I}_0} (T_i E_{i-1} v, E_{i-1} v) + \omega^2 N_0^2 \sum_{i \in \tilde{I}_0} (T_i E_{i-1} v, E_{i-1} v) \right. \\
 (2.26) \quad \left. + \omega^2 J_0 N_0 \sum_{i \in I_0} (T_i E_{i-1} v, E_{i-1} v) \right\}.
 \end{aligned}$$

Combining (2.25) with (2.26) proves (2.23). This completes the proof of the theorem. \square

Remark 2.3. Theorem 2.2 suggests that a smaller K_0 results in a better convergence estimate. Since the numbers J_0 and N_0 merely characterize the interaction between T_i 's, $i = 1, \dots, J$, one might think that less interaction yields

a better convergence rate. However, less interaction between subdomains may result in a larger constant C_0 in (2.7).

3. ADDITIVE PRECONDITIONING ALGORITHMS

Section 2 developed a theory which related the constant C_0 in (2.7) to the norms of products of the form (1.3). In this section, we shall provide converse estimates, i.e., assuming that the norms of the product is bounded by some constant $\gamma < 1$, we shall give a simple bound for C_0 . We will then discuss the implications of these results to a natural additive variant of Algorithm 1.

Theorem 3.1. *Assume that the norm of the operators $\{T_i\}$ are bounded by a positive constant ω less than 2 and there exists a constant $\gamma \in (0, 1)$ such that*

$$\|Ev\|^2 \leq \gamma \|v\|^2 \quad \text{for all } v \in \mathcal{V}.$$

Then (2.7) holds with $C_0 \leq 4(1 - \gamma)^{-1}(2 - \omega)^{-1}$.

Proof. Let $v \in \mathcal{V}$. We first note that by Lemma 2.1,

$$\begin{aligned} (2 - \omega) \sum_{i=1}^J (T_i E_{i-1} v, E_{i-1} v) \\ \leq \|v\|^2 - \|Ev\|^2 = 2(v, (I - E)v) - \|(I - E)v\|^2 \\ \leq 2(v, (I - E)v) = 2 \sum_{i=1}^J (v, T_i E_{i-1} v) \\ \leq 2 \left(\sum_{i=1}^J (T_i v, v) \right)^{1/2} \left(\sum_{i=1}^J (T_i E_{i-1} v, E_{i-1} v) \right)^{1/2}. \end{aligned} \tag{3.1}$$

Hence,

$$\left(\sum_{i=1}^J (T_i E_{i-1} v, E_{i-1} v) \right)^{1/2} \leq 2(2 - \omega)^{-1} \left(\sum_{i=1}^J (T_i v, v) \right)^{1/2}.$$

By (3.1) and the definition of γ ,

$$\begin{aligned} (1 - \gamma) \|v\|^2 &\leq \|v\|^2 - \|Ev\|^2 \\ &\leq 2 \left(\sum_{i=1}^J (T_i v, v) \right)^{1/2} \left(\sum_{i=1}^J (T_i E_{i-1} v, E_{i-1} v) \right)^{1/2} \\ &\leq 4(2 - \omega)^{-1} \sum_{i=1}^J (T_i v, v). \end{aligned}$$

This completes the proof of the theorem. \square

Theorem 3.1 provides a lower bound for the smallest eigenvalue of the operator $\sum_i T_i$. To estimate rates of convergence for the additive preconditioned

algorithms to be described, we must also bound the largest eigenvalue. We obviously have that

$$\sum_{i=1}^J (T_i v, v) \leq \omega J \|v\|^2.$$

A somewhat better bound may be derived when one assumes the interaction property of §2. Such a result is given by the following proposition.

Proposition 3.1. *Let N_0 and J_0 be as in §2. Assume that the norms of the operators T_i are bounded by a positive constant ω (which may be greater than or equal to 2). Then*

$$(3.2) \quad \left(\sum_{i=1}^J T_i v, v \right) \leq \omega (J_0 + N_0) \|v\|^2 \quad \text{for all } v \in \mathcal{V}.$$

Proof. Note that

$$\left(\sum_{i=1}^J T_i v, v \right) \leq \left\| \sum_{i=1}^J T_i v \right\| \|v\|.$$

Define the sets

$$\begin{aligned} S_1 &= \tilde{I}_0 \times \tilde{I}_0, & S_2 &= I_0 \times I_0, \\ S_3 &= I_0 \times \tilde{I}_0, & S_4 &= \tilde{I}_0 \times I_0. \end{aligned}$$

Then

$$\begin{aligned} \left\| \sum_{i=1}^J T_i v \right\|^2 &= \sum_{(i,j) \in S_1} (T_i v, T_j v) + \sum_{(i,j) \in S_2} (T_i v, T_j v) \\ &\quad + \sum_{(i,j) \in S_3} (T_i v, T_j v) + \sum_{(i,j) \in S_4} (T_i v, T_j v). \end{aligned}$$

Applying Lemma 2.2 gives

$$\begin{aligned} \left\| \sum_{i=1}^J T_i v \right\|^2 &\leq \omega N_0 \sum_{\tilde{I}_0} (T_i v, v) + \omega J_0 \sum_{I_0} (T_i v, v) \\ &\quad + 2\omega \sqrt{N_0 J_0} \left(\sum_{\tilde{I}_0} (T_i v, v) \right)^{1/2} \left(\sum_{I_0} (T_i v, v) \right)^{1/2} \\ &\leq \omega (J_0 + N_0) \sum_{i=1}^J (T_i v, v). \end{aligned}$$

The proposition follows by combining the above inequalities. \square

In the remainder of this section, we apply the above results to an additive variant of the algorithm described in the Introduction. We include this discussion for completeness, since the algorithms are so closely related to those discussed earlier. We note that the additive algorithms may have an advantage

when used on a computer with a parallel architecture, since the terms can always be evaluated concurrently. However, compared with the corresponding multiplicative variants, these algorithms often produce a somewhat slower rate of convergence in practice, mainly because of a larger upper eigenvalue (see §6).

The additive algorithms are defined in terms of a preconditioner B for the operator A in (1.2). Specifically, we set

$$(3.3) \quad B = \sum_{i=1}^J R_i Q_i$$

and note that

$$BA = \sum_{i=1}^J T_i,$$

where $T_i = R_i A_i P_i$ as defined in the Introduction. Effective algorithms for the solution of (1.1) are obtained by iterative methods applied to the preconditioned equations

$$BAu = Bf.$$

Clearly, BA is a symmetric operator in the inner product $(\cdot, \cdot) \equiv A(\cdot, \cdot)$, and one can apply, for example, the conjugate gradient iterative procedure. Alternatively, one could use the simplest linear iterative scheme as in the following algorithm.

Algorithm 2. Given $u^l \in \mathcal{V}$, define

$$(3.4) \quad u^{l+1} = u^l + \tau B(f - Au^l).$$

In the above algorithm, τ is a positive iteration parameter. The sequence of iterates converges to u provided that τ times the maximum eigenvalue of BA is less than two. The optimal rate of convergence is obtained by taking $\tau = 2/(\lambda_0 + \lambda_1)$ and results in a reduction of $(K(BA) - 1)/(K(BA) + 1)$ per iteration. Here, λ_0 and λ_1 denote respectively the smallest and largest eigenvalue of BA , and $K(BA) = \lambda_1/\lambda_0$ is the condition number. The results in this section can be used to provide bounds for $K(BA) = K(\sum_i T_i)$.

From (3.4), it is clear that the correction due to any subspace \mathcal{V}_i in the additive algorithm is computed by applying T_i to the original error $u - u^l = A^{-1}(f - Au^l)$ and adding the resulting contributions. In contrast, the product algorithm (Algorithm 1) involves applying T_i consecutively to the latest error.

4. OVERLAPPING DOMAIN DECOMPOSITION METHODS

In this section, the abstract results developed earlier will be applied to examples where the subspaces have been defined by overlapping domain decomposition. We shall consider both the continuous case as well as the case of finite element approximation. The results of §2 show that in order to bound the rate of convergence for the product algorithms, it suffices to prove inequality (2.7) and apply Theorem 2.1 or 2.2. For the applications of this section, the

main part of the proof of (2.7) was given in [13]. We first consider the case of overlapping subdomains of quasi-uniform size d without the use of a “coarse” problem. The convergence rate for this method deteriorates as the number of subdomains increases. Finally, to develop a method with a uniform rate of convergence (not depending upon the mesh size or the number of subdomains), the sequence of subspaces is augmented with a coarse subspace, i.e., a subspace of functions on a mesh of size d .

We will consider the following elliptic boundary value problem. For simplicity, we shall restrict to problems in two-dimensional Euclidean space R^2 . Extensions to higher dimensions are straightforward. Let Ω be a polygonal domain in R^2 and consider the problem

$$(4.1) \quad - \sum_{i,j=1}^2 \partial_i(a_{ij}\partial_j u) + au = f \quad \text{in } \Omega, \\ u = 0 \quad \text{on } \partial\Omega.$$

Here, $\partial_i = \frac{\partial}{\partial x_i}$, and we assume that the matrix $(a_{ij})_{2 \times 2}$ is symmetric for each $x \in \Omega$ and uniformly positive definite. We further assume that the coefficients of $(a_{ij})_{2 \times 2}$ and a are continuously differentiable and that $a \geq 0$. The solution u of (4.1) is in $H_0^1(\Omega)$ (the functions defined on Ω which vanish on $\partial\Omega$ in the appropriate sense and together with their first derivatives are L^2 integrable) and satisfies

$$(4.2) \quad A(u, \chi) = (f, \chi)_0 \quad \text{for all } \chi \in H_0^1(\Omega).$$

Here, $A(\cdot, \cdot)$ denotes the bilinear form associated with the operator in (4.1) defined by

$$A(u, v) = \sum_{i,j=1}^2 \int_{\Omega} a_{ij} \partial_j u \partial_i v \, dx + \int_{\Omega} a u v \, dx$$

and $(\cdot, \cdot)_0$ denotes the L^2 -inner product on Ω .

We will consider overlapping domain decomposition applied to both the continuous problem (i.e., $\mathcal{V} = H_0^1(\Omega)$) and its corresponding finite element approximations. Most of our development will be directed to the finite element exposition, the exposition for the continuous application is similar and will be discussed in accompanying remarks. The continuous case is only of theoretical interest, since actual computer implementation requires the use of finite-dimensional approximation spaces.

We first define the finite element approximation scheme. For simplicity of presentation, we use the simplest finite element spaces. Extensions to more complex elements are possible. Assume that Ω has been triangulated, $\Omega = \bigcup_i \tau_i$, where the triangles $\{\tau_i\}$ are of quasi-uniform size h with $h \in (0, 1]$. By this we mean that there are constants C_0, C_1 not depending on h such that each triangle τ_i is contained in (respectively, contains) a ball of radius $C_1 h$ (respectively $C_0 h$). The finite element space \mathcal{V} is defined to be the functions

which are continuous on Ω , piecewise linear with respect to the triangulation $\{\tau_i\}$, and vanish on $\partial\Omega$. The finite element approximation to the solution of (4.1) is the function $U \in \mathcal{V}$ satisfying

$$(4.3) \quad A(U, \chi) = (f, \chi)_0 \quad \text{for all } \chi \in \mathcal{V}.$$

Estimates for the error between u and U are well known (cf. [1], [2], [12]). We shall be concerned with iterative processes of the form of Algorithm 1 for the computation of U .

To define the overlapping domain decomposition method, we start by assuming that we are given a set of overlapping subdomains $\{\Omega_i\}_{i=1}^J$ whose boundaries align with the mesh triangulation defining \mathcal{V} . Associated with these subdomains, we assume that there is a partition of unity $\sum_{i=1}^J \rho_i = 1$ defined on Ω satisfying

$$(4.4) \quad \text{supp } \rho_i \subseteq \Omega_i \cup \partial\Omega,$$

$$(4.5) \quad \|\rho_i\|_{\infty, \Omega_i} \leq C,$$

$$(4.6) \quad \|\nabla \rho_i\|_{\infty, \Omega_i} \leq Cd^{-1},$$

for $i = 1, \dots, J$. Here, $\|\cdot\|_{\infty, D}$ denotes the L^∞ norm of a function defined on a domain D . One way of defining the subdomains and the associated partition is by starting with disjoint open sets $\{\Omega_i^0\}_{i=1}^J$ with $\overline{\Omega} = \bigcup_{i=1}^J \overline{\Omega}_i^0$ and $\{\Omega_i^0\}_{i=1}^J$ quasi-uniform of size d . The subdomain Ω_i is defined to be a mesh subdomain containing Ω_i^0 with the distance from $\partial\Omega_i \cap \Omega$ to Ω_i^0 greater than or equal to Cd for some prescribed constant C . The construction of functions ρ_i satisfying (4.4)–(4.6) is then straightforward.

Remark 4.1. To define the overlapping domain decomposition method for the continuous problem, we start by partitioning Ω into a sequence of overlapping subdomains $\Omega = \bigcup_{i=1}^J \Omega_i$. We assume that these subdomains have Lipschitz continuous boundaries and that there is a corresponding partition of unity $\sum_{i=1}^J \rho_i = 1$ satisfying (4.4)–(4.6). As above, the subdomains can be defined by starting with disjoint open sets $\{\Omega_i^0\}$ and expanding by Cd .

In either case, the subspace \mathcal{V}_i is defined by

$$\mathcal{V}_i = \{\varphi \mid \varphi \in \mathcal{V}, \text{ supp } \varphi \subseteq \Omega_i\}.$$

The first product method associated with these subspaces is when T_i of Algorithm 1 is taken to be P_i , the elliptic projection into \mathcal{V}_i . For the two-domain case, this is the classical Schwarz method (cf. [3], [15], [18]). The inner product (\cdot, \cdot) used in §2 is defined by $(u, v) \equiv A(u, v)$ for all $u, v \in \mathcal{V}$. By Remark 2.2, estimate (2.7) for this case will follow from the existence of a decomposition $v = \sum_{i=1}^J v_i$, with $v_i \in \mathcal{V}_i$ satisfying

$$(4.7) \quad \sum_{i=1}^J A(v_i, v_i) \leq C_0 A(v, v).$$

The decomposition used in [13], $v = \sum_{i=1}^J I_h(\rho_i v)$, where I_h denotes the nodal interpolation operator onto \mathcal{V} , satisfies (4.7) with $C_0 = Cd^{-2}$. In fact (cf. [13]), using (4.4)–(4.6), one shows that

$$(4.8) \quad \sum_{i=1}^J A(I_h(\rho_i v), I_h(\rho_i v)) \leq c\{d^{-2}\|v\|_0^2 + A(v, v)\} \quad \text{for all } v \in \mathcal{V}.$$

In the continuous case, one simply takes I_h to be the identity. We now give the following theorem.

Theorem 4.1. *Let n_i denote the number of subdomains Ω_j with $\Omega_j \cap \Omega_i \neq \emptyset$ and assume that $n_i \leq n$, $i = 1, \dots, J$. In addition, assume that (4.4)–(4.6) hold. Then*

$$\|Ev\|^2 \leq \left(1 - \frac{d^2}{Cn^2}\right) \|v\|^2 \quad \text{for all } v \in \mathcal{V},$$

where the constant C does not depend on d or h . This holds for overlapping domain decomposition in the continuous as well as the finite element application.

Proof. We have already proved estimate (2.7). To apply Theorem 2.2, we need only identify N_0 and I_0 . We set $I_0 = \emptyset$. Note that κ_{ij} is nonzero only if $\Omega_i \cap \Omega_j \neq \emptyset$. Consequently, we can take $N_0 = n$. This completes the proof of the theorem. \square

Remark 4.2. The convergence estimate for the product algorithm in the case of two subdomains was given in [15]. It was also demonstrated in [15] that the product algorithm described above gives rise to the error propagation matrix associated with the classical Schwarz overlapping method [18]. For the case of more than two subdomains, convergence was proved in [15] but no estimate of the rate was given. Weak estimates were given in [17], [19]. Theorem 4.1 provides new estimates in the case of more than two projectors.

Remark 4.3. It is possible to view point-Gauss-Seidel iteration as a product iteration of the above form. For each node x_m of the triangulation, we define the subdomain Ω_m to be the union of the triangles $\tau_j \in \{\tau_k\}$ which have x_m as a vertex. In this case, the subspaces \mathcal{V}_i are one-dimensional and the projectors P_i are trivial to compute. It is then easy to see that the product algorithms with $R_i = A_i^{-1}$ correspond to the Gauss-Seidel method. Inequality (2.7) is a direct consequence of inverse properties of \mathcal{V} , since the decomposition $v = \sum v_i$ is unique and satisfies

$$\sum_i A(v_i, v_i) \leq C \sum_i v(x_i)^2 \leq Ch^{-2} A(v, v).$$

Thus, Theorem 2.2 provides an alternative proof of the convergence rate of Gauss-Seidel iteration.

In the finite element case, this algorithm requires the computation of functions $u_i \in \mathcal{V}_i$ satisfying

$$(4.9) \quad A(u_i, \varphi) = G_i(\varphi) \quad \text{for all } \varphi \in \mathcal{V}_i$$

for appropriate linear functionals G_i . To avoid solving systems of the form (4.9), we introduce T_i by preconditioning. Specifically, let R_i be a scaled preconditioner for A_i , i.e., $R_i A_i$ is a symmetric positive definite operator on \mathcal{V}_i satisfying

$$(4.10) \quad C_R A(w, w) \leq A(R_i A_i w, w) \leq A(w, w) \quad \text{for all } w \in \mathcal{V}_i.$$

Define $T_i = R_i A_i P_i$. Then (4.7) implies

$$(4.11) \quad \begin{aligned} A(w, w) &\leq C_0 \sum_{i=1}^J A(P_i w, w) \\ &\leq C_0 / C_R \sum_{i=1}^J A(T_i w, w) \quad \text{for all } w \in \mathcal{V}, \end{aligned}$$

i.e., (2.7) holds for $\{T_i\}$. The development of preconditioners for (4.9) has been subject to intensive research [4]–[9], [11], [13], [14], etc. Computational results for this algorithm with multigrid V-cycle preconditioning are given in §5. Combining (4.11) and Theorem 2.2 implies the following theorem.

Theorem 4.2. *Assume that the hypotheses for Theorem 4.1 hold. Define $T_i = R_i A_i P_i$ for $i = 1, \dots, J$ and assume that R_i satisfies (4.10) with C_R independent of d and h . Then*

$$\|Ev\|^2 \leq \left(1 - \frac{d^2}{Cn^2}\right) \|v\|^2 \quad \text{for all } v \in \mathcal{V},$$

where the constant C does not depend on d or h . This holds for overlapping domain decomposition in the continuous as well as the finite element application.

To obtain algorithms which converge with rates that are independent of d , one can add a coarse subspace. Let \mathcal{V}_0 be a finite element subspace of \mathcal{V} defined from a quasi-uniform triangulation of Ω of size d . Let Q denote the $L^2(\Omega)$ orthogonal projection onto \mathcal{V}_0 . Then, for $v \in \mathcal{V}$,

$$v = Qv + \sum_{i=1}^J I_h(\rho_i(v - Qv)) = \sum_{i=0}^J v_i$$

is a decomposition of v into $\{\mathcal{V}_i\}_{i=0}^J$ which, by (4.8), satisfies

$$\sum_{i=0}^J A(v_i, v_i) \leq C \{A(Qv, Qv) + d^{-2} \|v - Qv\|_0^2 + A(v - Qv, v - Qv)\}.$$

It is known that (cf. [20])

$$\|v - Qv\|_0 \leq Cd\|v\|_1, \quad \|Qv\|_1 \leq C\|v\|_1,$$

which implies

$$(4.12) \quad \sum_{i=0}^J A(v_i, v_i) \leq CA(v, v),$$

i.e., (2.7) holds with C_0 independent of d . Inequality (4.12) has been given in [14]. We can apply Theorem 2.2 to this formulation. In this case, I_0 contains the one integer corresponding to the coarse subdomain. We have the following theorem.

Theorem 4.3. *Assume that the hypotheses for Theorem 4.2 hold. For the algorithm which includes a coarse grid term $j = 0$, we have*

$$\|Ev\| \leq \gamma \|v\| \quad \text{for all } v \in \mathcal{V},$$

where γ is a constant which is less than one and does not depend on d or h . This holds for overlapping domain decomposition in the continuous as well as the finite element application.

5. NUMERICAL RESULTS

In this section, we provide the results of numerical examples illustrating the theory developed in earlier sections. We shall consider the model problem

$$(5.1) \quad \begin{aligned} -\Delta u &= f && \text{in } \Omega, \\ u &= 0 && \text{on } \partial\Omega, \end{aligned}$$

where Δ denotes the Laplacian and Ω is the unit square $[0, 1] \times [0, 1]$. Problem (5.1) is discretized by the finite element method. Specifically, the domain Ω is first partitioned into $m \times m$ square subdomains of side length $1/m$. Each smaller square is then divided into two triangles by one of the diagonals (e.g., the diagonal which goes from the bottom left to the upper right-hand corners of the square). The approximation space \mathcal{V} is defined to be the set of functions which are continuous on Ω , piecewise linear with respect to the triangulation, and vanish on $\partial\Omega$. We seek the Galerkin solution $U \in \mathcal{V}$ satisfying

$$(5.2) \quad D(U, \phi) = (f, \phi)_0 \quad \text{for all } \phi \in \mathcal{V}.$$

We shall consider a number of overlapping algorithms for the solution of the discrete equations determining U . The first algorithm uses overlapping subdomains defined in terms of strips. The remaining algorithms are based on overlapping subdomains Ω_i of quasi-uniform size as described in §4.

To define the strip overlapping subspaces, we first define $j - 1$ subregions $\Omega_k = [(k - 1)d, (k + 1)d]$ for $k = 1, \dots, j - 1$, where $d = 1/j$. We assume that the mesh aligns with the subdomain boundaries, i.e., j divides m , and set

- (1) $\mathcal{V}_k = \{\phi \in \mathcal{M} \mid \text{supp } \phi \subseteq \Omega_k\}$.
- (2) $A_k: \mathcal{V}_k \mapsto \mathcal{V}_k$ by $A_k V = \Psi$ where Ψ is the unique function in \mathcal{V}_k satisfying

$$(\Psi, \theta) = D(V, \theta) \quad \text{for all } \theta \in \mathcal{V}_k.$$

Note that here we identify \mathcal{V}'_k with \mathcal{V}_k . We define A as above but with \mathcal{V}_k replaced by \mathcal{V} . Note that Problem (5.2) can be rewritten as $AU = F$, where F is the L^2 -projection of f into \mathcal{V} . For this example, we shall use $R_k = A_k^{-1}$ and hence $T_k = P_k$.

TABLE 5.1
The values of γ for overlapping strips

$m = 1/h$	$\gamma (j = 4)$	$\gamma (j = 8)$	$\gamma (j = 16)$	$\gamma (j = 32)$
16	.21	.59	—	—
32	.21	.59	.86	—
64	.21	.59	.86	.96
128	.21	.59	.86	.96

Table 5.1 gives the numerically computed value for the norm-squared reduction rate for Algorithm 1 with T_i defined in terms of overlapping strip domain decomposition as a function of j and m . Note that the norm-squared reduction rate (the minimal value γ satisfying (2.8)) is the largest eigenvalue of the symmetric iteration operator E^*E . Here, E^* is the adjoint of E with respect to the inner product $A(\cdot, \cdot)$ and is computed by reversing the order of the factors in (2.1).

It is not difficult to prove that the inequality (2.7) is satisfied with $C_0 \leq Cj^2$, and thus Theorem 2.1 guarantees that $\gamma \leq (1 - C/j^2)$. We note that the expected asymptotic behavior is seen comparing the $k = 16$ with the $k = 32$ results where $1 - \gamma$ is reduced by almost a factor of 4. The values of γ in Table 5.1 do not appear to depend upon $h = 1/m$.

For the remaining examples, we consider the above problem and discretization but define overlapping subdomains of quasi-uniform size. More precisely, let $d = 1/j$ and for each $i, l = 1, \dots, j-1$ define the subdomain $\Omega_{il} = [(i-1)d, (i+1)d] \times [(l-1)d, (l+1)d]$. We again assume that j divides m and define subspaces

$$\mathcal{V}_{il} = \{\phi \in \mathcal{V} \mid \text{supp } \phi \subseteq \Omega_{il}\}.$$

As seen in §4, (2.7) holds for the sequence of spaces $\{\mathcal{V}_{il} \mid i, l = 1, \dots, j-1\}$ with constant $C_0 \leq Cj^2$. As discussed in §4, the dependence on j can be removed by using a coarse space, e.g., the space \mathcal{V}_0 defined to be the continuous piecewise linear functions on the mesh of size d which vanish on $\partial\Omega$. We assume that the triangles of this coarse grid mesh are defined so that they align with those of \mathcal{V} and hence $\mathcal{V}_0 \subseteq \mathcal{V}$.

We first consider the product algorithm using the coarse subspace \mathcal{V}_0 and the $(j-1)^2$ overlapping subdomains of quasi-uniform size described above. For this example, we again use $R_k = A_k^{-1}$ and hence $T_k = P_k$. This algorithm converges in relatively few iterations but requires the exact solution of the subspace problems on each step of the iteration. The computed values of the reduction rates ($\sqrt{\gamma}$) for Algorithm 1 were less than .2 for combinations of $m = 16, 32, 64, 128$ and $j = 4, 8, 16$ (see Table 5.2). For comparison, we computed the condition numbers for the corresponding additive algorithm. For the same values of m and j , the additive algorithms gave condition numbers of at most 5.3, which corresponds to a reduction of 0.68 for Algorithm 2.

TABLE 5.2
Reduction rates for the $(K - 1)^2$ overlapping domain decomposition algorithm

$m = 1/h$	$\sqrt{\gamma}: K = 4$	$\sqrt{\gamma}: K = 8$	$\sqrt{\gamma}: K = 16$
16	.17(.53)	—	—
32	.17(.52)	.17(.83)	—
64	.17(.52)	.2(.82)	.2(.95)
128	.17(.52)	.2(.82)	.2(.95)

To illustrate the improvement resulting from including the coarse problem, Table 5.2 reports values for $\sqrt{\gamma}$ for the overlapping method without the coarse problem. These are the values given in parentheses and are always larger than the reduction rates for the algorithm with the coarse grid subspace. These results clearly indicate that the use of a coarse grid problem results in a significant improvement in the rate of convergence.

In almost all realistic applications, the direct solution of subproblems is much more expensive than the evaluation of a suitable preconditioner. To illustrate the effect on the convergence rate of the Algorithm 1, we next consider the previous example but with the direct solves on the subspaces replaced by multigrid preconditioners. Specifically, we employ the V-cycle multigrid algorithm (cf. [5]) using one pre and post Jacobi smoothing on each grid level. This leads to a preconditioning operator $R_{il} : \mathcal{V}_{il} \rightarrow \mathcal{V}_{il}$ which satisfies

$$(5.3) \quad 0.4A(v, v) \leq A(R_{il}A_{il}v, v) \leq A(v, v) \quad \text{for all } v \in \mathcal{V}_{il}.$$

The constant 0.4 above was computed numerically and holds for all of the subspace problems which are required for this application, including \mathcal{V}_0 . For this example, the reduction rates for the preconditioned product algorithm (Algorithm 1) were all between .63 and .50 for combinations of $m = 16, 32, 64, 128$ and $k = 4, 8, 16$ (see Figure 5.3). Again, we computed condition numbers for the corresponding additive algorithms and found them to be bounded by 9.8 for the same range of k and m . This corresponds to a reduction of .81 for Algorithm 2 with an appropriate choice of iteration parameter τ . Finally, we computed the reduction rates for the product algorithm with preconditioning but without the coarse grid solve (the reductions are the numbers in parentheses in Table 5.3). As in the case of direct solves, the algorithm without the coarse problem shows a significant loss of efficiency.

In all of the above runs, Algorithm 1 with direct solves replaced by multigrid preconditioning would require, at most, four times as many iterations as the direct solve version. On the other hand, our coding of the simple multigrid V-cycle algorithm actually ran more than four times as fast as the corresponding direct (FFT-based) solving version. Hence, even in the special case where fast algorithms are available for the subproblems, the efficient multigrid preconditioning remains competitive. In more complex problems where fast direct

TABLE 5.3
Reduction rates for the domain decomposition with preconditioning

$m = 1/h$	$\sqrt{\gamma}: K = 4$	$\sqrt{\gamma}: K = 8$	$\sqrt{\gamma}: K = 16$
16	.50(.62)	—	—
32	.56(.62)	.52(.87)	—
64	.56(.62)	.55(.87)	.63(.96)
128	.57(.62)	.57(.87)	.63(.96)

methods are not available, the product algorithms using preconditioning will show a significant improvement in computational efficiency.

It is interesting to note that in the above two examples, the product algorithm converges somewhat faster than the corresponding additive algorithm. On the other hand, the additive algorithms will generally be somewhat easier to implement on a parallel machine. This leads one to the question of when an additive algorithm gives rise to a more efficient algorithm even in a parallel environment. Note that if the subspace problems are relatively large and reasonably parallelizable on the given architecture, then little computational gain will be obtained from executing them concurrently. This seems to suggest that the additive algorithm will be more effective than the product only if the computer has a very large number of processors.

BIBLIOGRAPHY

1. J. P. Aubin, *Approximation of elliptic boundary-value problems*, Wiley-Interscience, New York, 1972.
2. A. K. Aziz and I. Babuška, *Part I, Survey lectures on the mathematical foundations of the finite element method*, The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations (A. K. Aziz, ed.), Academic Press, New York, 1972, pp. 1–362.
3. I. Babuška, *On the Schwarz algorithm in the theory of differential equations of mathematical physics*, Czechoslovak Math. J. **8** (1958), 328–342 (Russian).
4. G. Birkhoff and A. Schoenstadt, Editors, *Elliptic problem solvers*, II (G. Birkhoff and A. Schoenstadt, eds.), Academic Press, Orlando, Florida, 1984.
5. J. H. Bramble and J. E. Pasciak, *New convergence estimates for multigrid algorithms*, Math. Comp. **49** (1987), 311–329.
6. J. H. Bramble, J. E. Pasciak, and A. H. Schatz, *The construction of preconditioners for elliptic problems by substructuring*, I, Math. Comp. **47** (1986), 103–134.
7. ———, *The construction of preconditioners for elliptic problems by substructuring*, II, Math. Comp. **49** (1987), 1–16.
8. ———, *The construction of preconditioners for elliptic problems by substructuring*, III, Math. Comp. **51** (1988), 415–430.
9. ———, *The construction of preconditioners for elliptic problems by substructuring*, IV, Math. Comp. **53** (1989), 1–24.
10. J. H. Bramble, J. E. Pasciak, J. Wang, and J. Xu, *Multigrid results which do not depend upon elliptic regularity assumptions* (in preparation).
11. J. H. Bramble, J. E. Pasciak and J. Xu, *Parallel multilevel preconditioners*, Math. Comp. **55** (1990), 1–22.

12. P. G. Ciarlet, *The finite element method for elliptic problems*, North-Holland, New York, 1978.
13. M. Dryja and O. Widlund, *An additive variant of the Schwarz alternating method for the case of many subregions*, Technical Report 339, Courant Institute of Mathematical Sciences, 1987.
14. ———, *Some domain decomposition algorithms for elliptic problems*, Technical Report 438, Courant Institute of Mathematical Sciences, 1989.
15. P. L. Lions, *On the Schwarz alternating method*, Proc. First Internat. Sympos. on Domain Decomposition Methods for Partial Differential Equations (R. Glowinski, G. H. Golub, G. A. Meurant, and J. Périaux, eds.), SIAM, Philadelphia, PA, 1988.
16. J. Mandel and S. F. McCormick, *Iterative solution of elliptic equations with refinement: The two-level case*, Domain Decomposition Methods (T. F. Chan, R. Glowinski, J. Périaux, and O. B. Widlund, eds.), SIAM, Philadelphia, PA, 1989, pp. 81–92.
17. T. P. Mathew, *Domain decomposition and iterative refinement methods for mixed finite element discretizations of elliptic problems*, Thesis, New York University, 1989.
18. H. A. Schwarz, *Ueber einige Abbildungsaufgaben*, J. Reine Angew. Math. **70** (1869), 105–120. [Ges. Math. Abh., vol.2, 65-83].
19. O. Widlund, *A comparison of some domain decomposition and iterative refinement algorithms for elliptic finite element problems*, Technical Report BSC 88/15, IBM Bergen Scientific Centre, Bergen, Norway, 1988.
20. J. Xu, *Theory of multilevel methods*, Dept. Math. Rep. AM-48, Penn. State University, 1989.

DEPARTMENT OF MATHEMATICS, CORNELL UNIVERSITY, ITHACA, NEW YORK 14853
E-mail address: bramble@mathvax.msi.cornell.edu

DEPARTMENT OF APPLIED SCIENCE, BROOKHAVEN NATIONAL LABORATORY, UPTON, NEW YORK 11973
E-mail address: pasciak@bnl.gov

MATHEMATICS DEPARTMENT, THE UNIVERSITY OF WYOMING, LARAMIE, WYOMING 82071

DEPARTMENT OF MATHEMATICS, PENNSYLVANIA STATE UNIVERSITY, UNIVERSITY PARK, PENNSYLVANIA 16802

4.8 Domain decomposition methods for problems with uniform local refinement in two dimensions

Domain decomposition methods for problems with uniform local refinement in two dimensions [3]

CHAPTER 9

Domain Decomposition Methods for Problems with Uniform Local Refinement in Two Dimensions*

James H. Bramble†
Richard E. Ewing‡
Rossen R. Parashkevov‡
Joseph E. Pasciak§

Abstract. In this talk, we first present a flexible mesh refinement strategy for the approximation of solutions of elliptic boundary value problems in two dimensional domains. Coupled with this approximation scheme, we shall describe preconditioners for the resulting discrete system of algebraic equations. These techniques lead to efficient computational procedures in serial as well as parallel computing environments. The preconditioners are based on overlapping domain decomposition and involve solving (or preconditioning) subproblems on regular subregions. These techniques are analyzed in a forthcoming paper [2]. We present the results of numerical experiments illustrating the preconditioning algorithms.

INTRODUCTION

To provide the required accuracy in many applications involving large scale scientific computation, it becomes necessary to use local mesh refinement techniques. These techniques allow the use of finer meshes in regions of the computational domain where the solution exhibits large gradients. This remains practical only if efficient techniques for the solution of the resulting discrete systems are available. In this talk, we will give a flexible scheme for refinement as well as develop effective iterative methods for the solution of the resulting systems of discrete equations. The analysis for the methods discussed in this talk is given in [2].

* This manuscript has been authored under contract number DE-AC02-76CH00016 with the U.S. Department of Energy. Accordingly, the U.S. Government retains a non-exclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. This work was also supported in part under the National Science Foundation Grant No. DMS84-05352 and by the U.S. Army Research Office through the Mathematical Science Institute, Cornell University. Additional supporters of this work include the Office of Naval Research under contract No. 0014-88-K-0370 and by the Institute for Scientific Computation at the University of Wyoming though National Science Foundation grant No. RII-8610680.

†Department of Mathematics, Cornell University, Ithaca, NY 14853.

‡Mathematics Department, The University of Wyoming, Laramie, WY 82071.

§Department of Applied Science, Brookhaven National Laboratory, Upton, NY 11973.

We shall be interested in techniques for problems with refinements which are not quite local. As an example, one might consider a front passing through a two dimensional domain. In this case, it might be necessary to refine in the neighborhood of the front.

There are a number of ways of developing preconditioned iterative schemes for the discrete systems resulting from local mesh refinement in the literature. Techniques based on nested multilevel spaces are given in [1],[10],[11]. Techniques based on domain decomposition are given in [3],[14],[15]. The analysis presented there implicitly depends on the shape of the the refinement domain, and hence the resulting algorithms may not be as effective with irregularly shaped refinement regions. These algorithms also require the solution of a subproblem or preconditioner on the refinement regions. This talk will provide alternative preconditioned iterative techniques for these problems based on overlapping domain decomposition. Our algorithms are simpler and possibly more effective when implemented since they often lead to preconditioning subproblems defined on either regular subregions or topologically ‘nice’ meshes. The refinement region is the union of the subregions.

The proposed mesh refinement strategy is important in that it provides a basic approach for implementing dynamic local grid refinement. An example of a refinement strategy involves starting with a uniform coarse-grid and refining in small subregions associated with a selected set of coarse-grid vertices. These subregions are allowed to overlap and there are no theoretical restrictions on the resulting refinement region (the union of the subregions). Dynamic refinement is achieved by simply dynamically changing the selected set of coarse-grid vertices.

In addition, the technique can be integrated into existing large scale simulators without a complete redesign of the code. This is because most of the computation involves tasks on either the global coarse grid or the refinement grids associated with the refinement subregions. Choosing the coarse and refinement grid structure to be that already used in the code saves considerable development costs. For example, if one uses regularly structured meshes in the coarse and refinement grids, a substantial part of the resulting algorithm only requires operations on regular grids even though the resulting final approximation space is not regular.

The outline of the remainder of the talk is as follows. In Section 2, we define some preliminaries and describe the second-order elliptic problems which will be considered. The overlapping domain decomposition algorithms for grids with partial refinement is defined in Section 3. The theoretical estimates for the resulting preconditioned systems (from [2]) are also given there. Finally, computational aspects and the results of numerical experiments using these preconditioning techniques are discussed in Section 4.

2. THE ELLIPTIC PROBLEM AND PRELIMINARIES

We shall be concerned with the efficient solution of discrete equations resulting from approximation of second-order elliptic boundary value problems in a polygonal domain Ω contained in two dimensional Euclidean space R^2 . We consider the problem of approximating the solution u of

$$(2.1) \quad \begin{aligned} Lv &= f && \text{in } \Omega, \\ u &= 0 && \text{on } \partial\Omega. \end{aligned}$$

Here L is given by

$$Lv = - \sum_{i,j=1}^2 \frac{\partial}{\partial x_i} a_{ij} \frac{\partial v}{\partial x_j},$$

and $\{a_{ij}(x)\}$ is a uniformly positive definite, bounded, piecewise smooth coefficient matrix on Ω . The corresponding bilinear form is denoted by $A(\cdot, \cdot)$ and is given by

$$(2.2) \quad A(v, w) = \sum_{i,j=1}^2 \int_{\Omega} a_{ij} \frac{\partial v}{\partial x_i} \frac{\partial w}{\partial x_j} dx,$$

and is defined for functions $v, w \in H^1(\Omega)$. Here $H^1(\Omega)$ is the Sobolev space of order one on Ω . We denote the $L^2(\Omega)$ inner product by (\cdot, \cdot) . The weak solution u of (2.1) is the function $u \in H_0^1(\Omega)$ satisfying

$$A(u, \varphi) = (f, \varphi) \quad \text{for all } \varphi \in H_0^1(\Omega).$$

Here, $H_0^1(\Omega)$ is the subspace of functions in $H^1(\Omega)$ whose traces vanish on $\partial\Omega$.

We consider the above model problem for convenience. Many extensions of the techniques to be presented are possible; for example, one could consider equations with lower-order terms and different boundary conditions.

In this talk, we shall deal with various domains. These domains will always be open.

3. THE OVERLAPPING ALGORITHMS

In this section, we shall define iterative methods for problems with partial refinement based on overlapping domain decomposition. We start with a coarse mesh $\cup \tau_H^i$ consisting of triangles of quasi-uniform size H . The associated finite element space M_0 is defined to be the set of continuous piecewise linear functions on the coarse mesh which vanish on $\partial\Omega$. The interior nodes of this mesh will be denoted $\{x_i\}$, for $i = 1, \dots, N_c$. The mesh refinement is defined in terms of a number of coarse grid subdomains $\{\Omega_i\}$ for $i = 1, \dots, K$. By convention, Ω_i is defined to be the interior of the union of the closures of the coarse grid triangles. The refinement regions will also be referred to as “the subdomains.” We assume that they have limited overlap in that any point of Ω is contained in at most a fixed number (not depending on H) of the subdomains. We define the domain of refinement Ω^r to be the union of the subdomains, $\Omega^r = \cup_{i=1}^K \Omega_i$. There are no theoretical restrictions concerning the definition of the refinement subregions except that they are defined in terms of the coarse grid triangles and satisfy the overlap property as described above.

We provide two examples of this construction. For both examples, the subregions are associated with coarse grid nodes. For the first example, we define the region associated with a coarse-grid node x_i as the subdomain Ω_i which contains the coarse-grid triangles having x_i as a vertex. For the second example, we consider a mesh which is topologically equivalent to a regular rectangular mesh (see Figure 3.1). In this case, we define Ω_i to be the four quadrilaterals which share the vertex x_i . Some reasons for such a choice will be explained later. In either case, an index set $I \subseteq [1, \dots, N_c]$ is selected and the domains $\{\Omega_i\}$ with $i \in I$ are used to define the refinement region. By possibly changing the numbering of the coarse grid nodes, we assume, without loss of generality, that $I = [1, 2, \dots, K]$. There are no additional restrictions concerning this set I and hence rather complex refinement regions are possible.

The composite space is defined in terms of a quasi-uniform mesh $\{\tau_h^i\}$ on Ω of size $h < H$ which satisfies

$$\cup_i \partial \tau_H^i \subseteq \cup_i \partial \tau_h^i.$$

The space of continuous piecewise linear functions with respect to this triangulation (which vanish on $\partial\Omega$) will be denoted by M . Note that this space is introduced for the construction and analysis of the composite grid space. It is not used in actual computation since it has too many degrees of freedom in Ω/Ω^r . The subspace M_i associated with the subdomain Ω_i is defined by

$$(3.1) \quad M_i = \{\phi \in \tilde{M} \mid \text{support } \phi \subseteq \Omega_i\}.$$

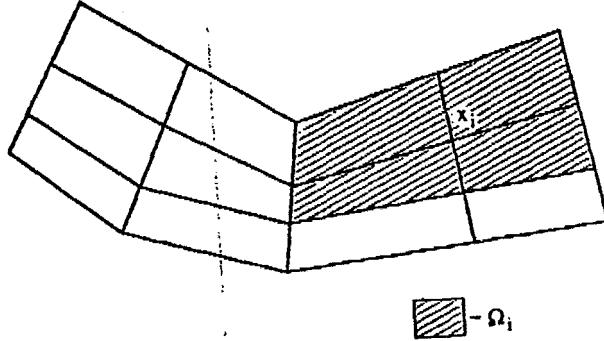


Figure 3.1
A distorted rectangular mesh.

The composite finite element space is then defined to be

$$M = \sum_{i=0}^K M_i.$$

Note that the space M provides finer grid approximation in the refinement region Ω^r . An illustrative example of a mesh so generated is given in Figure 3.2. The nodes on the boundary of the refinement region which are not coarse-grid nodes are slave nodes since, by continuity, the values of functions in M on these points are completely determined by their values on neighboring coarse-grid nodes. The operator $A_i : M_i \mapsto M_i$ is defined for $v \in M_i$ by

$$(A_i v, \phi) = A(v, \phi) \quad \text{for all } \phi \in M_i.$$

Our goal is to efficiently solve the composite grid problem: Given a function $f \in L^2(\Omega)$, find $U \in M$ satisfying

$$(3.2) \quad A(U, \phi) = (f, \phi) \quad \text{for all } \phi \in M.$$

As above, we define $A : M \mapsto M$ by

$$(Av, \phi) = A(v, \phi) \quad \text{for all } \phi \in M.$$

Problem (3.2) can then be rewritten as

$$(3.3) \quad AU = F,$$

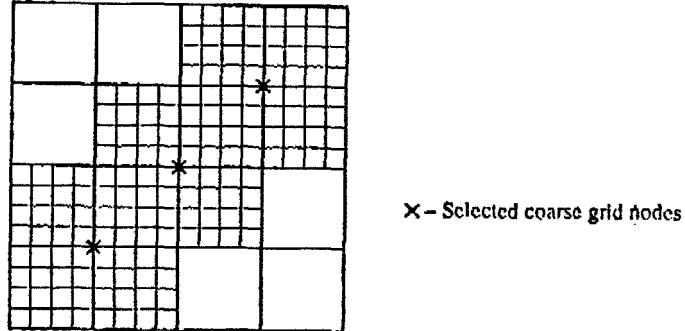


Figure 3.2
A composite grid.

for appropriate $F \in M$. We will develop preconditioners for (3.3) by using overlapping domain decomposition.

There are basically two classes of these preconditioners, the additive and multiplicative. The additive version defines the preconditioner B_a for A of (3.3) by

$$B_a = \sum_{i=0}^K R_i Q_i.$$

Here, Q_i denotes the $L^2(\Omega)$ projection operator onto M_i and R_i is a symmetric positive definite operator on M_i . Explicit choices for R_i will be discussed later; however, we note that it suffices to take R_i to be a preconditioner for A_i .

The multiplicative version is defined by applying the R_i consecutively. The multiplicative preconditioner B_m applied to a function $W \in M$ is defined as follows:

- (1) Set $Y_0 = 0$.
- (2) For $i = 1, \dots, K+1$, define Y_i by

$$(3.4) \quad Y_i = Y_{i-1} + R_{i-1} Q_{i-1} (W - AY_{i-1}).$$

- (3) For $i = K+2, \dots, 2K+2$, define Y_i by

$$(3.5) \quad Y_i = Y_{i-1} + R_{2K+2-i} Q_{2K+2-i} (W - AY_{i-1}).$$

- (4) Set $B_m W = Y_{2K+2}$.

It is not difficult to see that B_m is a symmetric linear operator on M .

The operators B_a and B_m defined above will be effective as preconditioners A if they satisfy the following:

- (1) They are relatively inexpensive to evaluate.
- (2) They lead to well conditioned linear systems.

The first criterion involves implementation issues and will be discussed later in more detail. The second criterion requires that the condition numbers $K(B_a A)$ and $K(B_m A)$ be small. In the case of the additive algorithms, this is equivalent to the existence of positive constants c_0, c_1 satisfying

$$(3.6) \quad c_0 A(v, v) \leq A(B_a A v, v) \leq c_1 A(v, v) \quad \text{for all } v \in M,$$

with c_1/c_0 small. A similar statement holds for the product algorithm.

The analysis presented in [2] requires the following hypotheses. It is assumed that there are positive constants C_0 and ω which do not depend on h , H or the subdomains and satisfy

$$(3.7) \quad C_0 A(w, w) \leq A(R_i A_i w, w) \leq \omega A(w, w) \quad \text{for all } w \in M_i.$$

This means that the operators R_i are spectrally good preconditioners for A_i . For the product algorithm, we also assume that $0 < \omega < 2$. The following theorem is proved in [2].

THEOREM 3.1. *Assume that there are no isolated points on the boundary of Ω^r . Then the condition numbers $K(B_a A)$ and $K(B_m A)$ remain bounded independently of h , H and the choice of subdomains $\{\Omega_i\}$.*

REMARK 3.1: The analysis given in [2] uses techniques from both the theory of overlapping domain decomposition [12],[13] as well as the standard domain decomposition theory [5]-[8] to provide the result for the additive algorithms. The result for the multiplicative version follows from that for the additive and the application of a general theory given in [9].

REMARK 3.2: The hypothesis concerning isolated points on the boundary of Ω^r is included to provide a uniform estimate for the preconditioned systems. If $\partial\Omega^r$ contains isolated points then it is possible to show (cf. Remark 4.2 of [2]) that the condition number grows at most on the order of $\ln^2(H/h)$. This sort of decay is actually seen in the last numerical example in Section 6.

REMARK 3.3: There is very little restriction concerning the way that the domains Ω_i are defined. Note that if only one refinement domain is used, then Theorem 3.1 provides a result for the imbedded space case proposed in [3]. Alternatively, one can consider the case where Ω^r is all of Ω and hence $M = \hat{M}$. In this case, Theorem 3.1 guarantees uniform bounds for the condition numbers without putting restrictions on the shapes of the subdomains $\{\Omega_i\}$. Thus, for example, the subdomains can be taken to be strips as long as the coarse problem is included.

4. COMPUTATIONAL ASPECTS AND NUMERICAL EXAMPLES

In this section, we discuss some of the computational properties associated with the method. In particular, we shall consider its feasibility for use in dynamic refinement strategies. We shall also see that with this type of method, it is possible to develop highly vectorizable and parallelizable code. Finally, we provide the results of numerical examples illustrating the condition numbers for the preconditioned systems described earlier.

We consider the earlier discussed examples where the domain of refinement is defined by simply selecting coarse-grid nodes and a rule for defining the refinement region associated with a coarse node. Specifically, we consider the example where the coarse mesh is defined from quadrilaterals and the refinement region associated with a coarse-grid vertex is defined to be the four quadrilaterals which share the vertex. An easy way to implement this refinement involves using vectors of unknowns with some redundancy. Associated with each quadrilateral, we associate a vector which contains the fine-grid unknowns in the quadrilateral and its boundary. The program is designed to operate on a data structure which contains a coarse-grid vector and a list of fine-grid vectors corresponding to the quadrilaterals appearing in the refinement regions. This process is controlled by a list of pointers which connect the location of quadrilateral fine-grid vectors in this data structure

to the coarse grid node refinement regions in which they appear. A simple control structure is also developed to handle the redundancy in the data vectors. These control structures can be easily derived from the list of coarse-grid refinement nodes and the coarse mesh geometry. Thus, a dynamic change in the refinement region only requires a simple (and of negligible cost) computation of some control pointers associated with the coarse grid.

An advantage of the proposed approach is that it can be used to invoke refinement without the use of the general data structures associated with meshes which are not regular. One assigns a regular mesh topology to the coarse mesh and to the meshes in the refinement subregions. This means that even though the composite mesh is highly irregular, all of the problems (on M_i , $i \in I_0$) which need to be solved or preconditioned will be on regular rectangular meshes. Similarly, it is possible to decompose the evaluation of the composite grid operator into pieces which involve operator evaluation on the topologically rectangular mesh parts. For these topologically rectangular meshes, highly efficient modules for preconditioning and operator evaluation are available for both vector and parallel computing architectures.

We shall consider the model problem

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

where Δ denote the Laplacian and Ω is the unit square $[0, 1] \times [0, 1]$. To define the coarse mesh, the domain Ω is first partitioned into $m \times m$ square subdomains of side length $H = 1/m$. Each smaller square is then divided into two triangles by one of the diagonals (e.g. the diagonal which goes from the bottom left to the upper right hand corner of the square). The coarse-grid approximation space M_0 is defined to be the set of functions which are continuous on Ω , are piecewise linear with respect to the triangulation, and vanish on $\partial\Omega$. The space M is defined from a similar finer mesh of size $h = H/l$ for some integer $l > 1$.

For our first two examples, we consider an application where it is required to refine along the diagonal connecting the origin with the point $(1, 1)$. Such a refinement might be necessary if the function f has large gradients near this diagonal but is well behaved in the remainder of Ω . Accordingly, we select the coarse-grid nodes on the diagonal for refinement. We define the refinement region associated with a refinement node to be the four coarse mesh squares which have that node as a corner. Note that the refinement region is highly irregular even though the coarse problem and the refinement subproblems involve regular rectangular meshes.

We will illustrate the rate of convergence of preconditioned algorithms for solving (3.2) where $A(\cdot, \cdot)$ is given by the Dirichlet form. To do this, we shall numerically compute the largest and smallest eigenvalue (λ_1 and λ_0 respectively) of the preconditioned operator $B_a A$. As is well known, the rate of convergence of the resulting preconditioned algorithms can be bounded in terms of the condition number $K(B_a A) = \lambda_1/\lambda_0$. We shall not report results for preconditioning with the product operator B_m , although our previous experience [9] suggests that the product version will converge somewhat faster than the additive.

Table 4.1 gives the largest and smallest eigenvalue and the condition number of the system $B_a A$ as a function of h . In this example, we took $R_i = A_i^{-1}$; i.e., we solved exactly on the subspaces $\{M_i\}$. For Table 4.1, $m = 4$ and there are three refinement subdomains $(0, 1/2) \times (0, 1/2)$, $(1/4, 3/4) \times (1/4, 3/4)$, and $(1/2, 1) \times (1/2, 1)$. Note that both the upper and lower eigenvalues appear to be tending to a limit as the ratio $h/H \mapsto 0$. Similar behavior is seen in Table 4.2, which corresponds to $m = 8$ and uses seven smaller refinement subregions.

Table 4.1
Condition numbers for 3 overlapping subregions

h	λ_1	λ_0	$K(B_a A)$
1/8	2.44	0.50	4.9
1/16	2.50	0.41	6.1
1/32	2.51	0.38	6.6
1/64	2.52	0.36	6.9
1/128	2.52	0.35	7.1

In almost all realistic applications, the direct solution of subproblems is much more expensive than the evaluation of a suitable preconditioner. To illustrate the effect on the convergence rate of the preconditioned iteration, we next consider the previous example but with the direct solves on the subspaces replaced by multigrid preconditioners. Specifically, we employ the V-cycle multigrid algorithm (cf. [4]) using one pre- and post-smoothing Jacobi iteration on each grid level. This leads to a preconditioning operator $R_i : M_i \mapsto M_i$ which satisfies

$$(4.2) \quad 0.4A(v, v) \leq A(R_i A_i v, v) \leq A(v, v) \quad \text{for all } v \in M_i.$$

Table 4.2
Condition numbers for 7 overlapping subregions

h	λ_1	λ_0	$K(B_a A)$
1/16	2.46	0.47	5.2
1/32	2.52	0.39	6.5
1/64	2.54	0.35	7.2
1/128	2.54	0.34	7.5

The constant 0.4 above was computed numerically and holds for all of the subspace problems which are required for this application, including M_0 .

Tables 4.3 and 4.4 provide the eigenvalues and condition numbers for the above examples when direct solves were replaced by multigrid preconditioners. Note that in all of the reported runs, the condition number with multigrid preconditioners was at most 5/4 times as large as that corresponding to exact solves. Such an increase in condition number is negligible in a preconditioned iteration. In contrast, the computational time required for the multigrid sweep is considerably less than that needed for a direct solve (especially in more general problems with variable coefficients).

Table 4.3
Preconditioned subproblems, 3 overlapping subregions

h	λ_1	λ_0	$K(B_a A)$
1/8	2.37	0.53	4.5
1/16	2.12	0.33	6.4
1/32	2.07	0.27	7.6
1/64	2.04	0.25	8.2
1/128	2.02	0.24	8.4

Table 4.4
Preconditioned subproblems, 7 overlapping subregions

h	λ_1	λ_0	$K(B_a A)$
1/16	2.36	0.40	5.9
1/32	2.11	0.28	7.5
1/64	2.06	0.24	8.8
1/128	2.03	0.22	9.4

As a final example, we consider a case where the isolated point hypothesis of Theorem 3.1 is not satisfied. Specifically, we consider a coarse mesh of size $H = 1/4$ and select the four nodes with (x, y) values $(1/4, 1/2)$, $(3/4, 1/2)$, $(1/2, 1/4)$, and $(1/2, 3/4)$. The refinement region is everything but the subsquares $[0, 1/4] \times [0, 1/4]$, $[0, 1/4] \times [3/4, 1]$, $[3/4, 1] \times [0, 1/4]$, and $[3/4, 1] \times [3/4, 1]$. Note that, to satisfy the hypotheses of the theorem, it would be necessary to include a refinement region centered at the coarse-grid node $(1/2, 1/2)$. Table 4.5 gives the smallest eigenvalue for the operator $B_a A$ as a function of h . The function $(.32 + .36 \log_2(h^{-1}))^{-2}$ is also provided for comparison. These results suggest that smallest eigenvalue λ_0 decays as predicted by the theoretical bound $C / \ln(H/h)^2$ (see Remark 3.2).

Table 4.5
A “bad” example in two dimensions.

h	λ_0	$(.32 + .36 \log_2(h^{-1}))^{-2}$
1/8	.50	.51
1/16	.32	.32
1/32	.22	.22
1/64	.16	.16
1/128	.12	.12

REFERENCES

1. R.E. Bank, T. Dupont, and H. Yserentant, *The hierarchical basis multigrid method*, Num. Math. **52** (1988), 427–458.
2. J.H. Bramble, R.E. Ewing, R.R. Parashkevov, and J.E. Pasciak, *Domain decomposition methods for problems with partial refinement*, Proceedings of the Copper Mountain Meeting on Iterative Methods, April, 1990 (submitted).
3. J.H. Bramble, R.E. Ewing, J.E. Pasciak and A.H. Schatz, *A preconditioning technique for the efficient solution of problems with local grid refinement*, Comp. Meth. Appl. Mech. Eng. **67** (1988), 149–159.
4. J.H. Bramble and J.E. Pasciak, *New convergence estimates for multigrid algorithms*, Math. Comp. **49** (1987), 311–329.
5. J.H. Bramble, J.E. Pasciak and A.H. Schatz, *The construction of preconditioners for elliptic problems by substructuring, I*, Math. Comp. **47** (1986), 103–134.
6. J.H. Bramble, J.E. Pasciak and A.H. Schatz, *The construction of preconditioners for elliptic problems by substructuring, II*, Math. Comp. **49** (1987), 1–16.
7. J.H. Bramble, J.E. Pasciak and A.H. Schatz, *The construction of preconditioners for elliptic problems by substructuring, III*, Math. Comp. **51** (1988), 415–430.
8. J.H. Bramble, J.E. Pasciak and A.H. Schatz, *The construction of preconditioners for elliptic problems by substructuring, IV*, Math. Comp. **53** (1989), 1–24.
9. J.H. Bramble, J.E. Pasciak, J. Wang, and J. Xu, *Convergence estimates for product iterative methods with applications to domain decomposition and multigrid*, (preprint).
10. J.H. Bramble, J.E. Pasciak, J. Wang, and J. Xu, *Multigrid results which do not depend upon elliptic regularity assumptions*, (in preparation).
11. J.H. Bramble, J.E. Pasciak and J. Xu, *Parallel multilevel preconditioners*, Math. Comp., (in print).
12. M. Dryja and O. Widlund, *An additive variant of the Schwarz alternating method for the case of many subregions*, Technical Report, Courant Institute of Mathematical Sciences **339** (1987).
13. P.L. Lions, *On the Schwarz alternating method*, In the Proceedings of the First International Symposium on Domain Decomposition Methods for Partial Differential Equations, R. Glowinski, G.H. Golub, G.A. Meurant, and J. Periaux (eds.), 1987.
14. S. McCormick and J. Thomas, *The fast adaptive composite grid (FAC) method for elliptic equations*, Math. Comp. **46** (1986), 439–456.
15. O. Widlund, *Optimal iterative refinement methods*, Technical Report, Courant Institute of Mathematical Sciences **391** (1988).

*4.9. DOMAIN DECOMPOSITION METHODS FOR PROBLEMS WITH
PARTIAL REFINEMENT*

4.9 Domain decomposition methods for problems with partial refinement

Domain decomposition methods for problems with partial refinement [9]

DOMAIN DECOMPOSITION METHODS FOR PROBLEMS WITH PARTIAL REFINEMENT*

JAMES H. BRAMBLE[†], RICHARD E. EWING[‡],
ROSSEN R. PARASHKEVOV[‡], AND JOSEPH E. PASCIAK[§]

Abstract. In this paper, a flexible mesh refinement strategy for the approximation of solutions of elliptic boundary value problems is considered. The main purpose of the paper is the development of preconditioners for the resulting discrete system of algebraic equations. These techniques lead to efficient computational procedures in serial as well as parallel computing environments. The preconditioners are based on overlapping domain decomposition and involve solving (or preconditioning) subproblems on regular subregions. It is proven that the iteration schemes converge to the discrete solution at a rate which is independent of the mesh parameters in the case of two spatial dimensions. The estimates proved for the iterative convergence rate in three dimensions are somewhat weaker. The results of numerical experiments illustrating the theory are also presented.

Key words. second-order elliptic equation, domain decomposition, overlapping domain decomposition, local mesh refinement, partial refinement, overlapping Schwarz methods, preconditioners

AMS(MOS) subject classifications. 65N30, 65F10

1. Introduction. To provide the required accuracy in many applications involving large scale scientific computation, it becomes necessary to use local mesh refinement techniques. These techniques allow the use of finer meshes in regions of the computational domain where the solution exhibits large gradients. This remains practical only if efficient techniques for the solution of the resulting discrete systems are available. It is the purpose of this paper to provide such techniques. We will give a flexible scheme for refinement as well as develop and analyze effective iterative methods for the solution of the resulting systems of discrete equations.

In this paper, we shall be interested in techniques for problems with refinements which are not quite local. As an example, one might consider a front passing through a two-dimensional domain. In this case, it might be necessary to refine in the neighborhood of the front.

There are a number of ways of developing preconditioned iterative schemes for the discrete systems resulting from local mesh refinement in the literature. Techniques based on nested multilevel spaces are given in [1], [7], [8], [12]. Techniques based on domain decomposition are given in [2], [10], [13], [14]. The analysis presented there implicitly depends on the shape of the the refinement domain, and hence the resulting algorithms may not be as effective with irregularly shaped refinement regions.

*Received by the editors April 5, 1990; accepted for publication (in revised form) October 5, 1990. This manuscript has been authored under contract number DE-AC02-76CH00016 with the U.S. Department of Energy. Accordingly, the U.S. Government retains a non-exclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. This work was also supported in part under the National Science Foundation grant DMS84-05352 and by the U.S. Army Research Office through the Mathematical Science Institute, Cornell University. Additional supporters of this work include the Office of Naval Research under contract 0014-88-K-0370 and by the Institute for Scientific Computation at the University of Wyoming through National Science Foundation grant RII-8610680.

[†]Department of Mathematics, Cornell University, Ithaca, New York 14853.

[‡]Mathematics Department, University of Wyoming, Laramie, Wyoming 82071.

[§]Department of Applied Science, Brookhaven National Laboratory, Upton, New York 11973.

These algorithms also require the solution of a subproblem or preconditioner on the refinement regions. We shall provide alternative preconditioned iterative techniques for these problems based on overlapping domain decomposition. Our algorithms are simpler and possibly more effective when implemented since they often lead to preconditioning subproblems defined on either regular subregions or topologically “nice” meshes. The refinement region is the union of the subregions and may be irregularly shaped.

The proposed mesh refinement strategy is important in that it provides a basic approach for implementing dynamic local grid refinement. An example of a refinement strategy involves starting with a uniform coarse grid and refining in small subregions associated with a selected set of coarse-grid vertices. These subregions are allowed to overlap and there are no theoretical restrictions on the resulting refinement region (the union of the subregions). Dynamic refinement is achieved by simply dynamically changing the selected set of coarse-grid vertices.

In addition, the technique can be integrated into existing large scale simulators without a complete redesign of the code. This is because most of the computation involves tasks on either the global coarse grid or the refinement grids associated with the refinement subregions. Choosing the coarse and refinement grid structure to be that already used in the code saves considerable development costs. For example, if one uses regularly structured meshes in the coarse and refinement grids, a substantial part of the resulting algorithm only requires operations on regular grids even though the resulting final approximation space is not regular.

The outline of the remainder of the paper is as follows. In §2, we define some preliminaries and describe the second-order elliptic problems that will be considered. The overlapping domain decomposition algorithms for grids with partial refinement are given in §3. An analysis of the resulting preconditioned algorithms is given in §4. It is shown that the condition number of the preconditioned systems is bounded independently of the mesh parameters for many two-dimensional applications. The results for three dimensions are somewhat weaker and involve logarithms of the mesh parameters. Finally, the results of numerical experiments using these preconditioning techniques are given in §5.

2. The elliptic problem and preliminaries. We shall be concerned with the efficient solution of discrete equations resulting from approximation of second-order elliptic boundary value problems in a polygonal or polyhedral domain Ω contained in Euclidean space R^d , for $d = 2, 3$. We consider the problem of approximating the solution u of

$$(2.1) \quad \begin{aligned} Lu &= f && \text{in } \Omega, \\ u &= 0 && \text{on } \partial\Omega. \end{aligned}$$

Here L is given by

$$Lv = - \sum_{i,j=1}^d \frac{\partial}{\partial x_i} a_{ij} \frac{\partial v}{\partial x_j},$$

and $\{a_{ij}(x)\}$ is a uniformly positive definite, bounded, piecewise smooth coefficient matrix on Ω . The corresponding bilinear form is denoted by $A(\cdot, \cdot)$ and is given by

$$(2.2) \quad A(v, w) = \sum_{i,j=1}^d \int_{\Omega} a_{ij} \frac{\partial v}{\partial x_i} \frac{\partial w}{\partial x_j} dx,$$

and is defined for functions $v, w \in H^1(\Omega)$. Here $H^1(\Omega)$ is the Sobolev space of order one on Ω . We denote the $L^2(\Omega)$ inner product by (\cdot, \cdot) . The weak solution u of (2.1) is the function $u \in H_0^1(\Omega)$ satisfying

$$A(u, \varphi) = (f, \varphi) \quad \text{for all } \varphi \in H_0^1(\Omega).$$

Here, $H_0^1(\Omega)$ is the subspace of functions in $H^1(\Omega)$ whose trace vanishes on $\partial\Omega$.

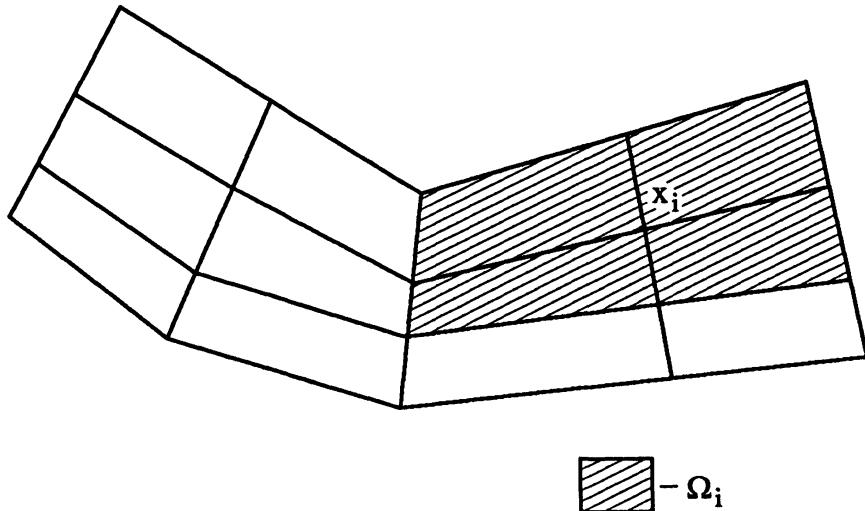
We consider the above model problem for convenience. Many extensions of the techniques to be presented are possible; for example, one could consider equations with lower-order terms and different boundary conditions.

In this paper, we shall deal with various domains. These domains will always be open. The closure of a domain θ will be denoted $\bar{\theta}$. In addition, we shall use various positive constants. These will be denoted by the character C , which will take on different values in different places. However, this constant shall always be independent of the mesh parameters in the approximation schemes.

3. The overlapping algorithms. In this section, we shall define iterative methods for problems with partial refinement based on overlapping domain decomposition. Our goal is to illustrate the technique and analysis and hence, for simplicity, we shall not attempt to provide the most general theorems. Many extensions are possible and can be inferred from the analysis presented.

The analysis given in the following section requires the application of techniques from both the theory of overlapping domain decomposition [9], [11] as well as the standard domain decomposition theory [4], [5]. We first give the setup in the two-dimensional case. We start with a coarse mesh $\cup \tau_H^i$ consisting of triangles of quasi-uniform size H . The associated finite element space M_0 is defined to be the set of those continuous piecewise linear functions on the coarse mesh that vanish on $\partial\Omega$. The mesh refinement is defined in terms of a number of coarse grid subdomains $\{\Omega_i\}$, for $i = 1, \dots, K$. By convention, Ω_i is defined to be the interior of the union of the closures of the coarse grid triangles. The refinement regions will also be referred to as “the subdomains.” We assume that they have limited overlap in that any point of Ω is contained in at most a fixed number (not depending on H) of the subdomains. We define the domain of refinement Ω^r to be the union of the subdomains, $\Omega^r = \cup_{i=1}^K \Omega_i$. There are no theoretical restrictions concerning the definition of the refinement subregions except that they are defined in terms of the coarse-grid triangles and satisfy the overlap property as described above.

We provide two examples of the construction in the two-dimensional case. For both examples, the subregions are associated with coarse-grid nodes. The interior and boundary nodes of this mesh will be denoted $\{x_i\}$, for $i = 1, \dots, N_c$. For the first example, we define the region associated with a coarse-grid node x_i as the subdomain Ω_i that contains the coarse-grid triangles having x_i as a vertex. For the second example, we consider a mesh that is topologically equivalent to a regular rectangular mesh (see Fig. 3.1). In this case, we define Ω_i to be the four quadrilaterals that share the vertex x_i . Some reasons for such a choice will be explained later. In either case, an index set $I \subseteq [1, \dots, N_c]$ is selected and only those subdomains $\{\Omega_i\}$ with $i \in I$ are used to define the refinement region. By possibly changing the numbering of the coarse-grid nodes, we assume, without loss of generality, that $I = 1, 2, \dots, K$. There are no additional restrictions concerning this set I and hence rather complex refinement regions are possible.

FIG. 3.1. *A distorted rectangular mesh.*

The composite space is defined in terms of a quasi-uniform mesh $\{\tau_h^i\}$ on Ω of size $h < H$ that satisfies

$$\cup_i \partial\tau_H^i \subseteq \cup_i \partial\tau_h^i.$$

The space of continuous piecewise linear functions with respect to this triangulation (which vanish on $\partial\Omega$) will be denoted by \tilde{M} . Note that this space is introduced for the construction and analysis of the composite grid space. It is not used in actual computation since it has too many degrees of freedom in Ω/Ω^r . The subspace M_i associated with the subdomain Ω_i is defined by

$$(3.1) \quad M_i = \{\phi \in \tilde{M} \mid \text{supp } \phi \subseteq \Omega_i\}.$$

The composite finite element space is then defined to be

$$M = \sum_{i=0}^K M_i.$$

Note that the space M provides finer grid approximation in the refinement region Ω^r . An illustrative example of a mesh so generated is given in Fig. 3.2. The nodes on the boundary of the refinement region that are not coarse-grid nodes are slave nodes since, by continuity, the values of functions in M on these points are completely determined by their values on neighboring coarse-grid nodes. The operator $A_i : M_i \mapsto M_i$ is defined for $v \in M_i$ by

$$(A_i v, \phi) = A(v, \phi) \quad \text{for all } \phi \in M_i.$$

Our goal is to efficiently solve the composite grid problem: Given a function $f \in L^2(\Omega)$, find $U \in M$ satisfying

$$(3.2) \quad A(U, \phi) = (f, \phi) \quad \text{for all } \phi \in M.$$

As above, we define $A : M \mapsto M$ by

$$(Av, \phi) = A(v, \phi) \quad \text{for all } \phi \in M.$$

Problem (3.2) can then be rewritten as

$$(3.3) \quad AU = F,$$

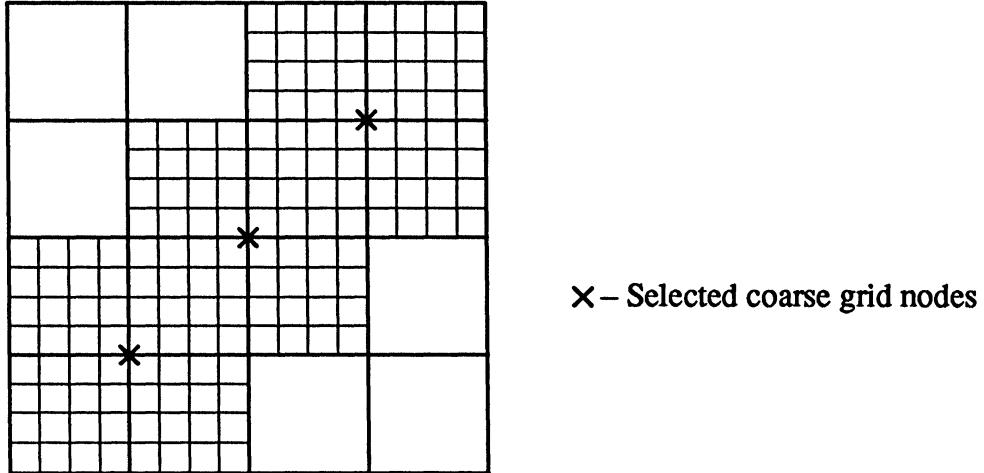


FIG. 3.2. A composite grid.

for appropriate $F \in M$. We will develop preconditioners for (3.3) by using overlapping domain decomposition.

There are basically two classes of these preconditioners, the additive and the multiplicative. The additive version defines the preconditioner B_a for A of (3.3) by

$$B_a = \sum_{i=0}^K R_i Q_i.$$

Here, Q_i denotes the $L^2(\Omega)$ projection operator onto M_i and R_i is a symmetric positive definite operator on M_i . Explicit choices for R_i will be discussed later; however, we note that it suffices to take R_i to be a preconditioner for A_i .

The multiplicative version is defined by applying the R_i consecutively. The multiplicative preconditioner B_m applied to a function $W \in M$ is defined as follows:

- (1) Set $Y_0 = 0$.
- (2) For $i = 1, \dots, K + 1$, define Y_i by

$$(3.4) \quad Y_i = Y_{i-1} + R_{i-1} Q_{i-1} (W - AY_{i-1}).$$

- (3) For $i = K + 2, \dots, 2K + 2$, define Y_i by

$$(3.5) \quad Y_i = Y_{i-1} + R_{2K+2-i} Q_{2K+2-i} (W - AY_{i-1}).$$

- (4) Set $B_m W = Y_{2K+2}$.

It is not difficult to see that B_m is a symmetric linear operator on M .

The operators B_a and B_m defined above will be effective as preconditioners A if they satisfy the following:

- (1) They are relatively inexpensive to evaluate.
- (2) They lead to well-conditioned linear systems.

The first criterion involves implementation issues. The second criterion requires that the condition numbers $K(B_a A)$ and $K(B_m A)$ be small. In the case of the additive algorithms, this is equivalent to the existence of positive constants c_0, c_1 satisfying

$$(3.6) \quad c_0 A(v, v) \leq A(B_a A v, v) \leq c_1 A(v, v) \quad \text{for all } v \in M,$$

with c_1/c_0 small. A similar statement holds for the product algorithm. The goal of the analysis to be presented is to provide estimates for c_0 and c_1 .

We note that the subdomains have limited overlap. This immediately implies that $c_1 \leq CN_0$, where C depends only on the preconditioning properties of R_i , and N_0 is the maximum number of subdomains overlapping any point $x \in \Omega$. Thus, to analyze the additive algorithm, we are left to estimate the size of c_0 . This will be done in the following section.

To analyze the product algorithm, we apply Theorem 2.2 of [6]. Assume that R_i is scaled so that for a fixed $\omega \in (0, 2)$,

$$(3.7) \quad A(R_i A_i v, v) \leq \omega A(v, v) \quad \text{for all } v \in M_i.$$

Then the product operator defined above satisfies

$$(3.8) \quad c_2 A(v, v) \leq A(B_m A v, v) \leq A(v, v) \quad \text{for all } v \in M$$

where $c_2 \geq Cc_0$. Here c_0 is the constant in (3.6) and C is a positive constant which depends on N_0 and the preconditioning properties of R_i but not on h or H . Thus, to analyze the product algorithm, we are once again left to estimate the size of c_0 in (3.6).

Remark 3.1. It is easy to extend the above ideas to three-dimensional calculations. We consider the case where Ω is the union of rectangular parallelopipeds and the coarse-grid functions are piecewise trilinear on the rectangular parallelopipeds. The refinement subregions are defined to be the interior of the closures of coarse grid parallelopipeds. The composite mesh is defined in terms of a quasi-uniform mesh of parallelopipeds of size $h < H$. This mesh is assumed to be a refinement of the coarse-grid mesh. The space \tilde{M} is defined to be the set of functions which are continuous on Ω , are trilinear with respect to the finer mesh, and vanish on $\partial\Omega$. The construction then proceeds exactly as described above for the two-dimensional case.

4. Convergence analysis. In this section, we provide an analysis for the overlapping domain decomposition preconditioners described in the previous section. As discussed earlier, we are to provide estimates for the constant c_0 appearing in (3.6). The analysis to be presented uses tools from both the theory of overlapping domain decomposition and the standard domain decomposition theory. We shall prove that under suitable hypotheses, the condition numbers $K(B_a A)$ and $K(B_m A)$ remain bounded independently of h and H in the two-dimensional case. The theorem for three dimensions guarantees that the condition numbers grow at most proportional to $(1 + \ln^2(H/h))$.

The first hypothesis for the theorems of this section provides control of the condition number $K(R_i A_i)$. Specifically, we assume that

$$(4.1) \quad C_0 A(w, w) \leq A(R_i A_i w, w) \leq \omega A(w, w) \quad \text{for all } w \in M_i,$$

where the constants C_0 and ω remain fixed independent of h and H . For the product algorithm, we also assume that $0 < \omega < 2$. We have the following theorem.

THEOREM 4.1. *Let $d = 2$ and assume that there are no isolated points on the boundary of Ω^r . Then the condition numbers $K(B_a A)$ and $K(B_m A)$ remain bounded independently of h , H and the choice of subdomains $\{\Omega_i\}$.*

Before proving Theorem 4.1, we review some results from the theory of overlapping and nonoverlapping domain decomposition. These results will play a major role in the subsequent proof.

Let P_i denote the elliptic projection into the subspace M_i , i.e., $P_i v = w$ where w is the unique function in M_i satisfying

$$A(w, \phi) = A(v, \phi) \quad \text{for all } \phi \in M_i.$$

It immediately follows from the definitions that $Q_i A = A_i P_i$ and hence (4.1) implies that, for $v \in M$,

$$\begin{aligned} A(BAv, v) &= \sum_{i=0}^K A(R_i Q_i Av, v) \\ &= \sum_{i=0}^K A(R_i A_i P_i v, P_i v) \geq C_0 \sum_{i=0}^K A(P_i v, v). \end{aligned}$$

Thus, $c_0 \geq \tilde{c}_0 C_0$ for any constant \tilde{c}_0 satisfying the inequality

$$(4.2) \quad \tilde{c}_0 A(v, v) \leq \sum_{i=0}^K A(P_i v, v) \quad \text{for all } v \in M.$$

It is known (cf. [11]) that (4.2) follows provided that \tilde{c}_0 is a constant such that for any $v \in M$ there is a decomposition $v = \sum_{i=0}^K v_i$, with $v_i \in M_i$, satisfying

$$(4.3) \quad \sum_{i=0}^K A(v_i, v_i) \leq \tilde{c}_0^{-1} A(v, v).$$

We remark that it is easy to prove that statements (4.2) and (4.3) are equivalent.

We shall require a known result concerning overlapping domain decomposition [9], [11]. For each coarse-grid node x_i , we let $\tilde{\Omega}_i$ denote the interior of the union of the closures of the coarse-grid triangles that have x_i as a vertex. We define \tilde{M}_i in terms of $\tilde{\Omega}_i$ as in (3.1). Given $w \in \tilde{M}$, there exists a decomposition $w = \sum_{i=1}^{N_c} \tilde{w}_i$, with $\tilde{w}_i \in \tilde{M}_i$, satisfying

$$(4.4) \quad \sum_{i=1}^{N_c} A(\tilde{w}_i, \tilde{w}_i) \leq C \left(A(w, w) + H^{-2} \|w\|^2 \right).$$

Here C is a constant not depending on h or H . The functions $\{\tilde{w}_i\}$ are defined in terms of a partition of unity with respect to the subdomains $\{\tilde{\Omega}_i\}$. An explicit partition can be defined from the coarse-grid nodal basis functions. These decompositions preserve support in the sense that if w vanishes at a node then every \tilde{w}_i vanishes there also.

We will also need results from the standard domain decomposition theory which we introduce as the following lemma. The proof of this lemma is essentially given in [4].

LEMMA 4.1. *Let $y \in M \cap \tilde{M}_i$ be discrete harmonic on each coarse grid triangle. By this we mean that*

$$A(y, \phi) = 0$$

for all functions $\phi \in M$ which vanish on the coarse-grid mesh. Assume that y vanishes on at least one coarse-grid edge connecting $\partial\tilde{\Omega}_i$ and x_i . Let $\{\tilde{\Gamma}_j\}$ denote the remaining coarse-grid edges connecting $\partial\tilde{\Omega}_i$ and x_i and define y_j to be the function which is

discrete harmonic on the coarse-grid triangles, equals y on $\tilde{\Gamma}_j$ and vanishes on the remaining coarse-grid edges. Then $y = \sum_j y_j$ and

$$(4.5) \quad \sum_j A(y_j, y_j) \leq CA(y, y).$$

Remark 4.1. If the function y in Lemma 4.1 vanishes only on the point x_i (instead of a line from $\partial\tilde{\Omega}_i$ to x_i), then the above decomposition is still defined. However, in such cases, (4.5) only holds with the constant C replaced by $c \ln^2(H/h)$.

Proof of Theorem 4.1. As discussed in the previous section, it suffices to estimate the constant c_0 in (3.6). This in turn follows from the construction of a decomposition $v = \sum_{i=0}^K v_i$ with $v_i \in M_i$ satisfying (4.3).

Let Q denote the $L^2(\Omega)$ projection operator onto the subspace M_0 .

We note that

$$(4.6) \quad \|(I - Q)w\|^2 \leq CH^2 A(w, w)$$

and

$$(4.7) \quad A(Qw, Qw) \leq CA(w, w)$$

hold for all $w \in H_0^1(\Omega)$.

We define v_0 in terms of Q by

$$v_0(x_i) = \begin{cases} v(x_i) & \text{for coarse-grid nodes } x_i \notin \Omega^r, \\ Qv(x_i) & \text{for coarse-grid nodes } x_i \in \Omega^r. \end{cases}$$

Clearly, we have that

$$(4.8) \quad \begin{aligned} A(v - v_0, v - v_0) &= A_r(v - v_0, v - v_0) \\ &\leq 2[A((I - Q)v, (I - Q)v) + A_r(Qv - v_0, Qv - v_0)]. \end{aligned}$$

Here $A_r(\cdot, \cdot)$ is given by (2.2) but with integration only over the domain Ω^r . Note that $Qv - v_0$ is a function in M_0 which vanishes at all coarse-grid nodes in Ω^r and is equal to $Qv - v$ on the remaining coarse-grid nodes. Consequently,

$$(4.9) \quad A_r(Qv - v_0, Qv - v_0) \leq C \sum_{x_i \in \partial\Omega^r} |(Qv - v)(x_i)|^2.$$

Here, the sum is taken over coarse grid nodes $x_i \in \partial\Omega^r$. There are no isolated points on $\partial\Omega^r$ and hence for each $x_i \in \partial\Omega^r$, there is a coarse grid edge Γ_i contained in $\partial\Omega^r$ ending at x_i . Both functions Qv and v are linear on Γ_i and hence

$$(4.10) \quad \begin{aligned} |(Qv - v)(x_i)|^2 &\leq cH^{-1} \|(Qv - v)\|_{\Gamma_i}^2 \\ &\leq C[H^{-2} \|Qv - v\|_{\tau_H^i}^2 + A_{\tau_H^i}(Qv - v, Qv - v)]. \end{aligned}$$

Here τ_H^i denotes a coarse-grid triangle containing the edge Γ_i and $A_{\tau_H^i}$ denotes the form defined by (2.2) but with integration only over the region τ_H^i . The last inequality in (4.10) is a simple consequence of the divergence theorem and is well known. Combining (4.6)–(4.10) proves that

$$(4.11) \quad A(v - v_0, v - v_0) \leq CA(v, v).$$

A similar argument gives that

$$(4.12) \quad \|v - v_0\|^2 \leq CH^2 A(v, v).$$

We next apply the overlapping domain decomposition result to $w = v - v_0$. Specifically, we decompose $w = \sum_{i=1}^{N_c} \tilde{w}_i$, with $\tilde{w}_i \in \tilde{M}_i$ and satisfying (4.4). Note that this is clearly not the desired decomposition into the refinement subspaces $\{M_i\}$. We will distribute the functions \tilde{w}_j , $j = 1, \dots, N_c$, into these subspaces. We start assigning each coarse grid-node $x_j \in \Omega^r$ to a unique subdomain $\Omega_{J(j)}$ which contains x_j . We then define

$$w_i = \sum_{J(j)=i} \tilde{w}_j.$$

We need to decompose the remaining functions \tilde{w}_i for $x_i \notin \Omega^r$. Note that by the support property, \tilde{w}_i vanishes unless $x_i \in \partial\Omega^r$. Consider a fixed function \tilde{w}_i with $x_i \in \partial\Omega^r$. We write $\tilde{w}_i = y + y_i^0$ where $y = \tilde{w}_i$ on the boundaries of the coarse-grid triangulation and is discrete harmonic on the coarse-grid triangles (as in Lemma 4.1). The function y_i^0 vanishes on the boundaries of the coarse-grid triangulation and is orthogonal (in $A(\cdot, \cdot)$) to y . Note that the function y_i^0 is nonzero only on triangles contained in the refinement region. Thus, we can assign each of these triangles uniquely to a subdomain Ω_j and add the restriction of y_i^0 to the corresponding function w_j . The result of these modifications will still be denoted $\{w_j\}$.

We finally distribute the function y . There are no isolated points in $\partial\Omega^r$ and hence there must be a coarse-grid edge ending at x_i contained in $\partial\Omega^r$. Note, in addition, that both \tilde{w}_i and y vanish on this edge. Thus, by Lemma 4.1,

$$(4.13) \quad \sum_j A(y_j, y_j) \leq C A(y, y).$$

The functions y_j are defined in Lemma 4.1 and the sum over j is taken over the coarse-grid indices corresponding to coarse-grid neighbors of x_i in Ω^r . The functions y_j are assigned to subdomains Ω_k which contain the corresponding edge (where y_j is nonzero) and the functions y_j are added into the corresponding w_k , producing a result which is still denoted w_k . It follows immediately that $w = \sum_{i=1}^K w_i$ and

$$(4.14) \quad \sum_{i=1}^K A(w_i, w_i) \leq C(A(w, w) + H^{-2} \|w\|^2).$$

Combining (4.11), (4.12), and (4.14) shows that the decomposition $v = \sum_{i=0}^K v_i$ with $v_i = w_i$ for $i = 1, \dots, K$ satisfies (4.3). This completes the proof of the theorem. \square

Remark 4.2. The hypothesis concerning isolated points on the boundary of Ω^r is included to provide a uniform bound for c_0 . It is possible to show (using of [4], Thm. 1) that the constant c_0 only deteriorates like $C/\ln^2(H/h)$ if the isolated point hypothesis is not satisfied. This sort of decay is actually seen in the last numerical example in §5 where this assumption is violated.

Remark 4.3. There is very little restriction concerning the way that the domains Ω_i are defined. Note that if only one refinement domain is used, then Theorem 4.1 provides a result for the imbedded space case proposed in [2]. Alternatively, one can consider the case where Ω^r is all of Ω and hence $M = \tilde{M}$. In this case, Theorem 4.1 guarantees uniform bounds for the condition numbers without putting restrictions on the shapes of the subdomains $\{\Omega_i\}$. Thus, for example, the subdomains can be taken to be strips as long as the coarse problem is included.

Remark 4.4. There are numerous ways of modifying the above algorithm. One possibility is to include additional subspaces corresponding to the coarse-grid nodes on $\partial\Omega^r$. For such a node x_k , the space M_k would be defined to be the functions in M^r

with support in Ω_k . The same result holds with a somewhat simpler analysis since the standard domain decomposition theory is avoided. However, this algorithm has some practical disadvantages. There are more subproblems and many of them correspond to grids on irregularly shaped domains.

We next provide the result for three-dimensional applications.

THEOREM 4.2. *Let Ω be a domain in R^3 and let the mesh and approximation space be as discussed in Remark 3.1. Assume that the hypotheses preceding Theorem 4.1 hold and that Ω^r is the interior of its closure. Then*

$$K(B_a A) \leq C(1 + \ln^2(H/h))$$

and

$$K(B_m A) \leq C(1 + \ln^2(H/h)).$$

The constant C above does not depend on H or h .

Proof. The major part of the proof follows the proof of Theorem 4.1. We seek a decomposition of $v \in M$ satisfying (4.3) with $\tilde{c}_0^{-1} \leq C(1 + \ln^2(H/h))$. The construction of v_0 is exactly the same as in Theorem 4.1 and still satisfies (4.11) and (4.12). Here we used the assumption that Ω^r was the interior of its closure.

The overlapping domain decomposition $w = \sum_{i=1}^{N_c} \tilde{w}_i$ satisfying (4.4) is also valid in three dimensions. Once again we reduce to the problem of decomposing functions \tilde{w}_k corresponding to coarse-grid nodes $x_k \in \partial\Omega^r$. As in the proof of Theorem 4.1, we write $\tilde{w}_k = y + y_k^0$, where y is discrete harmonic (with respect to the refined mesh) in the interior of the coarse parallelopipeds. The y_k^0 part is added into $\{w_i\}_{i=1}^K$.

Finally, we must take care of the function y . We write

$$(4.15) \quad y = \sum \bar{y}_{ij} + \sum \tilde{y}_l,$$

where:

- (1) The functions $\{\bar{y}_{ij}\}$ are discrete harmonic (with respect to the refined mesh) in the interior of the coarse parallelopipeds.
- (2) $\bar{y}_{ij} = y$ on the interior nodes on the face between coarse regions τ_H^i and τ_H^j and vanishes on the remaining nodes of $\cup \partial\tau_H^l$.
- (3) \tilde{y}_l equals y on the nodes of an edge of $\{\tau_H^l\}$ which is in $\tilde{\Omega}_k \cap \Omega_l$ and vanishes on all of the remaining nodes of the composite grid.
- (4) The sums in (4.15) are taken over the appropriate faces and edges.

Applying Lemma 4.3 of [5] gives that

$$A(\bar{y}_{ij}, \bar{y}_{ij}) \leq C(1 + \ln^2(H/h)) \{ |\tilde{w}_k|_{1/2, \partial\tau_H^i}^2 + H^{-1} |\tilde{w}_k|_{\tau_H^i}^2 \}.$$

Here $|\cdot|_{1/2, \partial\tau_H^i}$ denotes the Sobolev seminorm of order 1/2 on $\partial\tau_H^i$. We clearly have

$$A(\bar{y}_{ij}, \bar{y}_{ij}) \leq C(1 + \ln^2(H/h))(A(w_k, w_k) + H^{-2} \|w_k\|^2).$$

Similar arguments using Lemmas 4.1–4.2 of [5] give

$$A(\tilde{y}_l, \tilde{y}_l) \leq C(1 + \ln^2(H/h))(A(w_k, w_k) + H^{-2} \|w_k\|^2).$$

The desired bound for \tilde{c}_0^{-1} follows as in the proof of Theorem 4.1.

5. Numerical results. In this section, we provide the results of numerical examples illustrating the theory developed earlier. We shall consider the model problem

$$(5.1) \quad \begin{aligned} -\Delta u &= f && \text{in } \Omega, \\ u &= 0 && \text{on } \partial\Omega, \end{aligned}$$

where Δ denotes the Laplacian and Ω is the unit square $[0, 1] \times [0, 1]$. To define the coarse mesh, the domain Ω is first partitioned into $m \times m$ square subdomains of side length $H = 1/m$. Each smaller square is then divided into two triangles by one of the diagonals (e.g., the diagonal which goes from the bottom left to the upper right-hand corner of the square). The coarse-grid approximation space M_0 is defined to be the set of functions which are continuous on Ω , are piecewise linear with respect to the triangulation, and vanish on $\partial\Omega$. The space \tilde{M} is defined from a similar finer mesh of size $h = H/l$ for some integer $l > 1$.

For our first two examples, we consider an application where it is required to refine along the diagonal connecting the origin with the point $(1, 1)$. Such a refinement might be necessary if the function f has large gradients near this diagonal but is well behaved in the remainder of Ω . Accordingly, we select the coarse-grid nodes on the diagonal for refinement. We define the refinement region associated with a refinement node to be the four coarse mesh squares which have that node as a corner.

Note that the refinement region is highly irregular even though the coarse problem and the refinement subproblems involve regular rectangular meshes.

We will illustrate the rate of convergence of preconditioned algorithms for solving (3.2) where $A(\cdot, \cdot)$ is given by the Dirichlet form. To do this, we shall numerically compute the largest and smallest eigenvalue (λ_1 and λ_0 , respectively) of the preconditioned operator $B_a A$. As is well known, the rate of convergence of the resulting preconditioned algorithms can be bounded in terms of the condition number $K(B_a A) = \lambda_1/\lambda_0$. We shall not report results for preconditioning with the product operator B_m , although our previous experience [6] suggests that the product version will converge somewhat faster than the additive.

Table 5.1 gives the largest and smallest eigenvalue and the condition number of the system $B_a A$ as a function of h . In this example, we took $R_i = A_i^{-1}$; i.e., we solved exactly on the subspaces $\{M_i\}$. For Table 5.1, $m = 4$, and there are three refinement subdomains $(0, 1/2) \times (0, 1/2)$, $(1/4, 3/4) \times (1/4, 3/4)$, and $(1/2, 1) \times (1/2, 1)$. Note that both the upper and lower eigenvalues appear to be tending to a limit as the ratio $h/H \mapsto 0$. Similar behavior is seen in Table 5.2, which corresponds to $m = 8$ and uses seven smaller refinement subregions.

TABLE 5.1
Condition numbers for 3 overlapping subregions.

h	λ_1	λ_0	$K(B_a A)$
1/8	2.44	0.50	4.9
1/16	2.50	0.41	6.1
1/32	2.51	0.38	6.6
1/64	2.52	0.36	6.9
1/128	2.52	0.35	7.1

In almost all realistic applications, the direct solution of subproblems is much more expensive than the evaluation of a suitable preconditioner. To illustrate the effect on the convergence rate of the preconditioned iteration, we next consider the previous example but with the direct solves on the subspaces replaced by multigrid preconditioners. Specifically, we employ the V-cycle multigrid algorithm (cf. [3]) using one pre- and post-smoothing Jacobi iteration on each grid level. This leads to

TABLE 5.2
Condition numbers for 7 overlapping subregions.

h	λ_1	λ_0	$K(B_a A)$
1/16	2.46	0.47	5.2
1/32	2.52	0.39	6.5
1/64	2.54	0.35	7.2
1/128	2.54	0.34	7.5

a preconditioning operator $R_i : M_i \mapsto M_i$, which satisfies

$$(5.2) \quad 0.4A(v, v) \leq A(R_i A_i v, v) \leq A(v, v) \quad \text{for all } v \in M_i.$$

The constant 0.4 above was computed numerically and holds for all of the subspace problems that are required for this application, including M_0 .

Tables 5.3 and 5.4 provide the eigenvalues and condition numbers for the above examples when direct solves were replaced by multigrid preconditioners. Note that in all of the reported runs, the condition number with multigrid preconditioners was at most 5/4 times as large as that corresponding to exact solves. Such an increase in condition number is negligible in a preconditioned iteration. In contrast, the computational time required for the multigrid sweep is considerably less than that needed for a direct solve (especially in more general problems with variable coefficients).

TABLE 5.3
Preconditioned subproblems, 3 overlapping subregions.

h	λ_1	λ_0	$K(B_a A)$
1/8	2.37	0.53	4.5
1/16	2.12	0.33	6.4
1/32	2.07	0.27	7.6
1/64	2.04	0.25	8.2
1/128	2.02	0.24	8.4

TABLE 5.4
Preconditioned subproblems, 7 overlapping subregions.

h	λ_1	λ_0	$K(B_a A)$
1/16	2.36	0.40	5.9
1/32	2.11	0.28	7.5
1/64	2.06	0.24	8.8
1/128	2.03	0.22	9.4

As a final example, we consider a case where the isolated point hypothesis of Theorem 4.1 is not satisfied. Specifically, we consider a coarse mesh of size $H = 1/4$ and select the four nodes with (x, y) values $(1/4, 1/2)$, $(3/4, 1/2)$, $(1/2, 1/4)$, and

$(1/2, 3/4)$. The refinement region is everything but the subsquares $[0, 1/4] \times [0, 1/4]$, $[0, 1/4] \times [3/4, 1]$, $[3/4, 1] \times [0, 1/4]$, and $[3/4, 1] \times [3/4, 1]$. Note that, to satisfy the hypotheses of the theorem, it would be necessary to include a refinement region centered at the coarse-grid node $(1/2, 1/2)$. Table 5.5 gives the smallest eigenvalue for the operator $B_a A$ as a function of h . The function $(.32 + .36 \log_2(h^{-1}))^{-2}$ is also provided for comparison. These results suggest that smallest eigenvalue λ_0 decays as predicted by the theoretical bound $C/\ln(H/h)^2$ (see Remark 4.2).

TABLE 5.5
A “bad” example in two dimensions.

h	λ_0	$(.32 + .36 \log_2(h^{-1}))^{-2}$
1/8	.50	.51
1/16	.32	.32
1/32	.22	.22
1/64	.16	.16
1/128	.12	.12

REFERENCES

- [1] R. E. BANK, T. DUPONT, AND H. YSERENTANT, *The hierarchical basis multigrid method*, Numer. Math., 52 (1988), pp. 427–458.
- [2] J. H. BRAMBLE, R. E. EWING, J. E. PASCIAK, AND A. H. SCHATZ, *A preconditioning technique for the efficient solution of problems with local grid refinement*, Comput. Methods Appl. Mech. Engrg., 67 (1988), pp. 149–159.
- [3] J. H. BRAMBLE AND J. E. PASCIAK, *New convergence estimates for multigrid algorithms*, Math. Comp., 49 (1987), pp. 311–329.
- [4] J. H. BRAMBLE, J. E. PASCIAK, AND A. H. SCHATZ, *The construction of preconditioners for elliptic problems by substructuring*, I, Math. Comp., 47 (1986), pp. 103–134.
- [5] ———, *The construction of preconditioners for elliptic problems by substructuring*, IV, Math. Comp., 53 (1989), pp. 1–24.
- [6] J. H. BRAMBLE, J. E. PASCIAK, J. WANG, AND J. XU, *Convergence estimates for product iterative methods with applications to domain decomposition*, Math. Comp., 57 (1991), pp. 1–21.
- [7] ———, *Convergence estimates for multigrid algorithms without regularity assumptions*, Math. Comp., 57 (1991), pp. 23–45.
- [8] J. H. BRAMBLE, J. E. PASCIAK, AND J. XU, *Parallel multilevel preconditioners*, Math. Comp., 55 (1990), pp. 1–22.
- [9] M. DRYJA AND O. WIDLUND, *Some domain decomposition algorithms for elliptic problems*, Iterative Methods for Large Linear Systems, L. Hayes and D. Kincaid, eds., Academic Press, New York, 1989.
- [10] R.E. EWING, R.D. LAZAROV, AND P.S. VASSILEVSKI, *Local refinement techniques for elliptic problems on cell-centered grids*, II: Two-grid iterative methods, Math. Comp., submitted.
- [11] P.L. LIONS, *On the Schwarz alternating method*, in Proceedings of the First International Symposium on Domain Decomposition Methods for Partial Differential Equations, R. Glowinski, G. H. Golub, G. A. Meurant, and J. Périaux, eds., Society for Industrial Applied Mathematics, Philadelphia, PA, 1988.
- [12] S.F. MCCORMICK, *Multilevel Adaptive Methods for Partial Differential Equations*, Society for Industrial Applied Mathematics, Philadelphia, PA, 1989.

- [13] S. MCCORMICK AND J. THOMAS, *The fast adaptive composite grid (FAC) method for elliptic equations*, Math. Comp., 46 (1986), pp. 439–456.
- [14] O. WIDLUND, *Optimal iterative refinement methods*, Proceedings of the Second International Symposium on Domain Decomposition Methods, T. F. Chan, R. Glowinski, J. Périaux and O.B. Widlund, eds., Society for Industrial and Applied Mathematics, Philadelphia, PA, 1989, pp. 114–125.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

*4.10. ANALYSIS OF NON-OVERLAPPING DOMAIN DECOMPOSITION
ALGORITHMS WITH INEXACT SOLVES*

**4.10 Analysis of non-overlapping domain decomposition
algorithms with inexact solves**

Analysis of non-overlapping domain decomposition algorithms with inexact solves
[5]

ANALYSIS OF NON-OVERLAPPING DOMAIN DECOMPOSITION ALGORITHMS WITH INEXACT SOLVES

JAMES H. BRAMBLE, JOSEPH E. PASCIAK, AND APOSTOL T. VASSILEV

ABSTRACT. In this paper we construct and analyze new non-overlapping domain decomposition preconditioners for the solution of second-order elliptic and parabolic boundary value problems. The preconditioners are developed using uniform preconditioners on the subdomains instead of exact solves. They exhibit the same asymptotic condition number growth as the corresponding preconditioners with exact subdomain solves and are much more efficient computationally. Moreover, this asymptotic condition number growth is bounded independently of jumps in the operator coefficients across subdomain boundaries. We also show that our preconditioners fit into the additive Schwarz framework with appropriately chosen subspace decompositions. Condition numbers associated with the new algorithms are computed numerically in several cases and compared with those of the corresponding algorithms in which exact subdomain solves are used.

1. INTRODUCTION

In this paper, we consider the solution of the discrete systems of equations which result from finite element or finite difference approximation of second order elliptic and parabolic boundary problems. To effectively take advantage of modern parallel computing environments, algorithms must involve a large number of tasks which can be executed concurrently. Domain decomposition preconditioning techniques represent a very effective way of developing such algorithms. The parallelizable tasks are associated with subdomain solves.

There are two basic approaches to the development of domain decomposition preconditioners. The first is the so-called non-overlapping approach and is characterized by the need to solve subproblems on disjoint subdomains. Early work was applicable to domains partitioned into subdomains without internal cross-points [1], [4], [14]. To handle the case of cross-points, Bramble, Pasciak and Schatz introduced in [5] algorithms involving a coarse grid problem and provided analytic techniques for estimating the conditioning of the domain decomposition boundary preconditioner, a central issue in the subject. Various extensions of these ideas were provided in [23] including a Neumann-Dirichlet checkerboard like preconditioner.

Received by the editor February 21, 1996 and, in revised form, September 6, 1996.

1991 *Mathematics Subject Classification*. Primary 65N30, 65F10.

This manuscript has been authored under contract number DE-AC02-76CH00016 with the U.S. Department of Energy. Accordingly, the U.S. Government retains a non-exclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. This work was also supported in part under the National Science Foundation Grant No. DMS-9007185 and by the PICS ground water research initiative under contract AS-413-ASD. .

©1998 American Mathematical Society

Subsequently, these techniques were extended to problems in three dimensions in [8] and [15]. A critical ingredient in the three dimensional algorithms was a coarse grid problem involving the solution averages developed in [6]. Related work is contained in [13], [20], [21].

The papers [4], [5], [6], [7], and [8] developed domain decomposition preconditioners for the original discrete system. The alternative approach, to reduce to an iteration involving only the unknowns on the boundary, was taken in [1], [11], [13], and [21]. The difference in the two techniques is important in that for the first, it is at least feasible to consider replacing the subproblem solves by preconditioners.

The second approach for developing domain decomposition preconditioners involves the solution of subproblems on overlapping subdomains. For such methods it is always possible to replace the subproblem solution with a preconditioning evaluation [9]. However, in parallel implementations, the amount of inter-processor communication is proportional to the amount of overlap. These methods lose some efficiency as the overlap becomes smaller [17]. Theoretically, they are much worse in the case when there are jumps in coefficients (see, Remark 3.3 below). In contrast, the convergence estimates for correctly designed non-overlapping domain decomposition algorithms are the same as those for smooth coefficients as long as the jumps align with subdomain boundaries.

Thus, it is natural to investigate the effect of inexact solves on non-overlapping domain decomposition algorithms. Early computational results showing that inexact non-overlapping algorithms can perform well were reported in [18]. References to other experimental work can be found in [16]. Analysis and numerical experiments with inexact algorithms of Neumann–Dirichlet and Dirichlet types, under the additional assumption of high accuracy of the inexact solves, were given in [2] and [19]. Their analysis suggests that the inexact preconditioners do not, in general, preserve the asymptotic condition number behavior of the corresponding exact method, even when the forms providing the inexact interior solves are uniformly equivalent to the original.

In this paper, we construct and analyze new non-overlapping domain decomposition preconditioners with inexact solves. We provide variations of the exact algorithm considered in [6]. We develop algorithms based only on the assumption that the interior solves are provided by uniform preconditioning forms. The inexact methods exhibit the same asymptotic condition number growth as the one in [6] and are much more efficient computationally. Our algorithms are alternatives to and in many applications less restrictive than the preconditioners in [2] and [19]. The convergence estimates developed here are independent of jumps of the operator coefficients across subdomain boundaries. The results of this paper were reported by the second author at the Seventh Copper Mountain Multigrid Conference in April of 1995 and by the third author at the Ninth Conference on Domain Decomposition Methods in June of 1996.

An important aspect of the analysis provided in this paper is that the non-overlapping preconditioners are shown to be of additive Schwarz type. Even though the new methods are inspired by and implemented according to the classical non-overlapping methodology, they can be reformulated as additive Schwarz algorithms with appropriately chosen subspace decompositions.

The paper is organized as follows. In Section 2, we formulate the problem and introduce notation. In Section 3, we construct an inexact non-overlapping domain decomposition preconditioner and investigate its properties. Section 4 provides an

application of our preconditioning approach to discretizations of parabolic problems. Computational considerations concerning the preconditioners are given in Section 5. Section 6 considers alternative inexact preconditioners. Finally, the condition number of the preconditioners developed in Section 3 and Section 6 are computed in several cases and presented in Section 7.

2. PRELIMINARIES AND NOTATION

In this section we formulate a model elliptic problem and introduce the corresponding finite element discretization. We also outline the guiding principles in constructing our preconditioner.

We consider the Dirichlet problem

$$(2.1) \quad \begin{aligned} \mathcal{L}u &= f && \text{in } \Omega, \\ u &= 0 && \text{on } \partial\Omega, \end{aligned}$$

where Ω is a bounded polyhedral domain in \mathbb{R}^n for $n = 2, 3$ and

$$(2.2) \quad \mathcal{L}v = - \sum_{i,j=1}^n \frac{\partial}{\partial x_i} \left(a_{ij} \frac{\partial v}{\partial x_j} \right).$$

Here the $n \times n$ coefficient matrix $\{a_{ij}\}$ is symmetric, uniformly positive definite, and bounded above on Ω . This is a classical model problem for a second order uniformly elliptic equation. Generalizations of (2.2) which are needed for time stepping schemes approximating parabolic problems will be discussed in Section 4.

The generalized Dirichlet form on Ω is given by

$$(2.3) \quad \mathcal{A}(v, w) = \sum_{i,j=1}^n \int_{\Omega} a_{ij} \frac{\partial v}{\partial x_i} \frac{\partial w}{\partial x_j} dx.$$

Clearly, this form is well defined for functions v and w in the Sobolev space $H^1(\Omega)$.

The $L^2(\Omega)$ -inner product and the related norm are defined by

$$(v, w)_{\Omega} = \int_{\Omega} vw dx$$

and

$$\|v\|_{\Omega}^2 = (v, v)_{\Omega}.$$

$H_0^1(\Omega)$ is the Sobolev space obtained by the completion of smooth functions with support in Ω with respect to the norm in $H^1(\Omega)$. The weak formulation of (2.1) in $H_0^1(\Omega)$ is then given by the following.

Find $u \in H_0^1(\Omega)$ such that

$$(2.4) \quad \mathcal{A}(u, \varphi) = (f, \varphi)_{\Omega} \quad \text{for all } \varphi \in H_0^1(\Omega).$$

Given a finite dimensional subspace $S_h^0(\Omega)$ of $H_0^1(\Omega)$, the standard Galerkin approximation to (2.4) is defined by:

Find $U \in S_h^0(\Omega)$ such that

$$(2.5) \quad \mathcal{A}(U, \varphi) = (f, \varphi)_{\Omega} \quad \text{for all } \varphi \in S_h^0(\Omega).$$

To define $S_h^0(\Omega)$, we partition Ω into triangles $\{\tau_i^h\}$ (or tetrahedra) in the usual way. Here h is the mesh parameter and is defined to be the maximal diameter of all such triangles. By definition, these triangles are closed sets. We assume that the

triangulation is quasi-uniform. The collection of simplex vertices will be denoted by $\{x_i\}$.

By convention, any union of elements τ_j^h in a given triangulation will be called a mesh subdomain. In the sequel Ω is assumed partitioned into n_d mesh subdomains $\{\Omega_k\}_{k=1}^{n_d}$ of diameter less than or equal to d . The notation Ω_k will be used for the set of all points of a subdomain including the boundary $\partial\Omega_k$.

We now define the finite element spaces. Let $S_h^0(\Omega)$ be the space of continuous piecewise linear (with respect to the triangulation) functions that vanish on $\partial\Omega$. Correspondingly, $S_h^0(\Omega_k)$ will be the space of functions whose supports are contained in Ω_k and hence each function in $S_h^0(\Omega_k)$ vanishes on $\partial\Omega_k$. $S_h(\Omega_k)$ will consist of restrictions to Ω_k of functions in $S_h^0(\Omega)$. Let Γ denote $\bigcup_k \partial\Omega_k$ and let $S_h(\Gamma)$ and $S_h(\partial\Omega_k)$ be the spaces of functions that are restrictions to Γ and $\partial\Omega_k$, respectively, of functions in $S_h^0(\Omega)$. We consider piecewise linear functions for convenience since the results and algorithms to be developed extend to higher order elements without difficulty. However, application to h - p methods is beyond the scope of this paper.

The following additional notation will be used. Let the $L^2(\partial\Omega_k)$ -inner product be denoted by

$$\langle u, v \rangle_{\partial\Omega_k} = \int_{\partial\Omega_k} uv \, ds$$

and the corresponding norm by

$$|v|_{\partial\Omega_k} = \langle v, v \rangle_{\partial\Omega_k}^{1/2}.$$

On $S_h(\partial\Omega_k)$, the discrete inner product and norm are defined by

$$\langle u, v \rangle_{\partial\Omega_k, h} = h^{n-1} \sum_{x_i \in \partial\Omega_k} u(x_i)v(x_i)$$

and

$$|v|_{\partial\Omega_k, h} = \langle v, v \rangle_{\partial\Omega_k, h}^{1/2}.$$

Because of the mesh quasi-uniformity, the norm equivalence

$$(2.6) \quad c |v|_{\partial\Omega_k}^2 \leq |v|_{\partial\Omega_k, h}^2 \leq C |v|_{\partial\Omega_k}^2$$

holds for function $v \in S_h(\partial\Omega_k)$.

Here and in the remainder of the paper, we shall use c and C to denote generic positive constants independent of discretization and subdivision parameters such as h , n_d , and subdomain index k . The actual values of these constants will not necessarily be the same in any two instances.

Finally, $\mathcal{D}_k(\cdot, \cdot)$ denotes the Dirichlet inner product on Ω_k defined by

$$(2.7) \quad \mathcal{D}_k(v, w) = \sum_{i=1}^n \int_{\Omega_k} \frac{\partial v}{\partial x_i} \frac{\partial w}{\partial x_i} \, dx, \quad \text{for all } v, w \in H^1(\Omega_k).$$

The development of a method for efficient iterative solution of (2.5) is the subject of this paper. In particular, using the above described decomposition of Ω , we shall define a bilinear form $\mathcal{B}(\cdot, \cdot)$ on $S_h^0(\Omega) \times S_h^0(\Omega)$ which satisfies the following two basic requirements. First, the solution $W \in S_h^0(\Omega)$ of

$$(2.8) \quad \mathcal{B}(W, \varphi) = (g, \varphi)_\Omega \quad \text{for all } \varphi \in S_h^0(\Omega),$$

with g given, should be more efficient to compute than the solution of (2.5). Second, the two forms should be equivalent in the sense that

$$(2.9) \quad \lambda_1 \mathcal{B}(V, V) \leq \mathcal{A}(V, V) \leq \lambda_2 \mathcal{B}(V, V) \quad \text{for all } V \in S_h^0(\Omega),$$

for some positive constants λ_1 and λ_2 with λ_2/λ_1 not too large. These conditions, though somewhat vague, serve as guidelines for our construction.

3. THE PRECONDITIONER $\mathcal{B}(\cdot, \cdot)$ AND ITS ANALYSIS

In this section we construct an inexact non-overlapping domain decomposition preconditioner and prove an estimate for the condition number of the preconditioned system. We also show that our preconditioner is of additive Schwarz type with appropriately defined subspace decomposition.

3.1. The preconditioner. To define our domain decomposition preconditioner, we will need boundary extension operators. For each k , let us define linear extension operators $\mathcal{E}_k : S_h(\partial\Omega_k) \rightarrow S_h(\Omega_k)$ by

$$\mathcal{E}_k \phi(x_i) = \begin{cases} \phi(x_i) & \text{for } x_i \in \partial\Omega_k, \\ 0 & \text{for } x_i \in \Omega_k \setminus \partial\Omega_k. \end{cases}$$

Correspondingly, let $\mathcal{E} : S_h^0(\Omega) \mapsto S_h^0(\Omega)$ be defined by

$$(3.1) \quad \mathcal{E} \phi(x_i) = \begin{cases} \phi(x_i) & \text{for } x_i \in \Gamma, \\ 0 & \text{for } x_i \in \Omega \setminus \Gamma. \end{cases}$$

Clearly, defining these operators at the mesh vertices defines them everywhere.

For each k , let $\mathcal{B}_k(\cdot, \cdot)$ be a bilinear form on $S_h^0(\Omega_k) \times S_h^0(\Omega_k)$ which is uniformly equivalent to $\mathcal{A}_k(\cdot, \cdot)$ where $\mathcal{A}_k(\cdot, \cdot)$ is defined as in (2.3) but with integration only over Ω_k . By this we mean that for each k there are constants c_k and C_k with C_k/c_k bounded independently of h and d such that

$$(3.2) \quad c_k \mathcal{B}_k(V, V) \leq \mathcal{A}_k(V, V) \leq C_k \mathcal{B}_k(V, V) \quad \text{for all } V \in S_h^0(\Omega_k).$$

The preconditioning form is given by

$$(3.3) \quad \begin{aligned} \mathcal{B}(U, V) &= \sum_{k=1}^{n_d} \mathcal{B}_k(U - \bar{U}_k - \mathcal{E}_k(U - \bar{U}_k), V - \bar{V}_k - \mathcal{E}_k(V - \bar{V}_k)) \\ &\quad + h^{-1} \sum_{k=1}^{n_d} \tilde{a}_k \langle U - \bar{U}_k, V - \bar{V}_k \rangle_{\partial\Omega_k, h}. \end{aligned}$$

Here, \bar{U}_k denotes the discrete mean value of U on $\partial\Omega_k$, i.e.,

$$\bar{U}_k \equiv \frac{\langle U, 1 \rangle_{\partial\Omega_k, h}}{\langle 1, 1 \rangle_{\partial\Omega_k, h}}.$$

In (3.3), \tilde{a}_k , $k = 1, \dots, n_d$, are parameters associated with the coefficients a_{ij} in Ω_k . For example, if \tilde{a}_k is taken to be the smallest eigenvalue of $\{a_{i,j}\}$ at some point $x \in \Omega_k$, then

$$(3.4) \quad C_k^{-1} \tilde{a}_k \mathcal{D}_k(v, v) \leq \mathcal{A}_k(v, v) \leq C_k \tilde{a}_k \mathcal{D}_k(v, v) \quad \text{for all } v \in S_h(\Omega_k).$$

The constant C_k only depends on the local variation of the coefficients $\{a_{ij}\}$ on the subdomain Ω_k . Consequently, we will assume that (3.4) holds with C_k bounded independently of d , h , and k .

3.2. Analysis of the preconditioning form $\mathcal{B}(\cdot, \cdot)$. We introduce some standard assumptions about the domain Ω , the subdomain splitting and the associated finite element spaces which are needed for the analysis.

We start by requiring that the collection $\{\Omega_k\}$ be quasi-uniform of size d . Also, we shall assume that

$$(3.5) \quad |u|_{\partial\Omega_k}^2 \leq C\{\epsilon^{-1}\|u\|_{\Omega_k}^2 + \epsilon\mathcal{D}_k(u, u)\}$$

holds for any ϵ in $(0, d]$ and all k . Finally, we assume that a Poincaré inequality of the form

$$(3.6) \quad \|v\|_{\Omega_k}^2 \leq Cd^2\mathcal{D}_k(v, v)$$

holds for functions v with zero mean value on Ω_k .

The inequalities (3.5) and (3.6) hold for all but pathological subdomains. A sufficient but by no means necessary condition for the above two inequalities is given in the following assumption.

Each Ω_k is star-shaped with respect to a point. This means that for each Ω_k there is a point \hat{x}_k and a constant $c_k > 0$ such that $(x - \hat{x}_k) \cdot \mathbf{n}(x) \geq c_k d$ for all $x \in \partial\Omega_k$ which are not mesh vertices. We further assume that $c_k \geq c$ for some constant c not depending on d, k or h . Here $\mathbf{n}(x)$ denotes the outward unit normal vector to $\partial\Omega_k$ at a nonvertex point x .

The following lemma will be used in the derivation of our results.

Lemma 3.1. *If $v \in S_h(\Omega_k)$ and vanishes at all interior nodes of Ω_k , then*

$$(3.7) \quad \mathcal{D}_k(v, v) \leq Ch^{-1}|v|_{\partial\Omega_k, h}^2.$$

This lemma is obvious from the local properties of the functions in finite element spaces and we shall omit its proof.

The main result of this paper is contained in the following theorem.

Theorem 3.1. *Let $\mathcal{A}(\cdot, \cdot)$ and $\mathcal{B}(\cdot, \cdot)$ be given by (2.3) and (3.3) respectively. Assume that (3.2), (3.4), (3.5) and (3.6) hold. Then there exist positive constants c and C not depending on d or h such that*

$$(3.8) \quad c\mathcal{A}(V, V) \leq \mathcal{B}(V, V) \leq C\frac{d}{h}\mathcal{A}(V, V),$$

for all $V \in S_h^0(\Omega)$.

Proof. Because of (3.2), it suffices to prove the theorem for

$$(3.9) \quad \begin{aligned} \mathcal{B}(U, V) &= \sum_{k=1}^{n_d} \mathcal{A}_k(U - \bar{U}_k - \mathcal{E}_k(U - \bar{U}_k), V - \bar{V}_k - \mathcal{E}_k(V - \bar{V}_k)) \\ &\quad + h^{-1} \sum_{k=1}^{n_d} \tilde{a}_k \langle U - \bar{U}_k, V - \bar{V}_k \rangle_{\partial\Omega_k, h}. \end{aligned}$$

We first prove the left inequality in (3.8). The arithmetic–geometric mean inequality shows that for any constant α we have

$$(3.10) \quad \begin{aligned} \frac{1}{2}\mathcal{A}_k(V, V) &= \frac{1}{2}\mathcal{A}_k(V - \alpha, V - \alpha) \\ &\leq \mathcal{A}_k(V - \alpha - \mathcal{E}_k(V - \alpha), V - \alpha - \mathcal{E}_k(V - \alpha)) \\ &\quad + \mathcal{A}_k(\mathcal{E}_k(V - \alpha), \mathcal{E}_k(V - \alpha)). \end{aligned}$$

The left inequality in (3.8) is a simple consequence of (3.7), (3.4), (3.10), and the definition of \mathcal{E}_k with $\alpha = \bar{V}_k$.

In order to prove the right inequality, we apply the arithmetic–geometric mean inequality to the terms in the first sum in (3.9) and get

$$(3.11) \quad \begin{aligned} \mathcal{B}(V, V) &\leq 2\mathcal{A}(V, V) + 2 \sum_{k=1}^{n_d} \mathcal{A}_k(\mathcal{E}_k(V - \bar{V}_k), \mathcal{E}_k(V - \bar{V}_k)) \\ &+ h^{-1} \sum_{k=1}^{n_d} \tilde{a}_k \langle V - \bar{V}_k, V - \bar{V}_k \rangle_{\partial\Omega_k, h}. \end{aligned}$$

By (3.4) and (3.7), we obtain

$$(3.12) \quad \mathcal{B}(V, V) \leq 2\mathcal{A}(V, V) + Ch^{-1} \sum_{k=1}^{n_d} \tilde{a}_k \langle V - \bar{V}_k, V - \bar{V}_k \rangle_{\partial\Omega_k, h}.$$

Let $\bar{\bar{V}}_k$ be the mean value of V on Ω_k . Using the definition of \bar{V}_k yields

$$\langle V - \bar{V}_k, V - \bar{V}_k \rangle_{\partial\Omega_k, h} \leq \langle V - \bar{\bar{V}}_k, V - \bar{\bar{V}}_k \rangle_{\partial\Omega_k, h}.$$

We combine the above inequality with (2.6) and obtain

$$|V - \bar{V}_k|_{\partial\Omega_k, h}^2 \leq C |V - \bar{\bar{V}}_k|_{\partial\Omega_k}^2.$$

Applying (3.5) with $\epsilon = d$ and (3.6) to the right hand side of the last inequality gives

$$(3.13) \quad |V - \bar{V}_k|_{\partial\Omega_k, h}^2 \leq Cd\mathcal{A}_k(V, V).$$

Using this estimate in (3.12) proves (3.8). \square

Remark 3.1. The preconditioning form $\mathcal{B}(\cdot, \cdot)$ defined above is not uniformly equivalent to $\mathcal{A}(\cdot, \cdot)$. Nevertheless, its preconditioning effect is very close to that of a uniform preconditioner for many practical problems, particularly in three space dimensions. The number of subdomains often equals the number of processors in a parallel implementation and it is feasible to keep d on the order of $h^{1/2}$. Applying a conjugate gradient method preconditioned by $\mathcal{B}(\cdot, \cdot)$ for solving (2.5) would result in a number of iterations proportional to $h^{-1/4}$. In \mathbb{R}^3 , $h = 10^{-2}$ corresponds to a very large computational problem whereas $10^{1/2} \approx 3.2$.

Remark 3.2. The constants c and C in (3.8) depend on the local (with respect to the subdomains) behavior of the operator and the preconditioner. Clearly, one of the most influential factors on the local properties of $\mathcal{A}(\cdot, \cdot)$ and $\mathcal{B}(\cdot, \cdot)$ is the coefficient matrix $\{a_{i,j}\}|_{\Omega_k}$. In fact, the constants C_k in (3.4) depend on the local lower and upper bounds for the eigenvalues of $\{a_{i,j}\}|_{\Omega_k}$ and in general so do the constants c_k and C_k in (3.2). Therefore, in applications to problems with large jumps in the coefficients, it is desirable to align the subdomain boundaries with the locations of the jumps. In this case the preconditioner (3.3) will be independent of these jumps.

Remark 3.3. It is well known that classical overlapping domain decomposition algorithms with small overlap exhibit the same condition number growth but in contrast to our method the overlapping preconditioners are adversely sensitive to large jumps in the operator coefficients. The utilization of the averages \bar{U}_k plays the role of a coarse problem especially designed to take into account cases with interior

subdomains and also applications with large jumps in the operator coefficients provided that the locations of the jumps are aligned with the subdomain boundaries. The numerical calculations in Section 7 indicate the effectiveness of our preconditioner when such problems are solved. To illustrate that the role of the averages is essential in overcoming difficulties coming from large jumps of the coefficients, we consider a conventional additive Schwarz preconditioner with minimal overlap [17]. The asymptotic condition number bound provided in [17] is the same as that of our theorem in the case of smooth coefficients. However, because of the deterioration in the approximation and boundedness properties of the weighted L^2 projection into the coarse subspace [12], the condition number of the preconditioned system for the minimal overlap algorithm when $n = 3$ can only be bounded by $(d/h)^2$.

Remark 3.4. It is possible to apply the above preconditioner to the discrete systems which arise from other types of numerical approximation. For example, it is straightforward to apply the technique to finite difference approximations. In addition, its application to non-conforming finite element discretizations as well as mixed finite element approximations is given in [22].

Our preconditioner is very economical computationally. In fact, it allows the use of efficient subdomain preconditioners such as one multigrid V-cycle. The use of the simple extension \mathcal{E} also results in enhanced efficiency. We shall discuss the computational aspects of this algorithm in detail in the Section 5.

3.3. An additive Schwarz reformulation of the domain decomposition algorithm. We shall demonstrate that the preconditioner $\mathcal{B}(\cdot, \cdot)$ can be viewed as an additive subspace correction method (cf. [10] and [24]) with judiciously chosen subspaces. Let the linear operator $\tilde{\mathcal{E}} : S_h^0(\Omega) \mapsto S_h^0(\Omega)$ be defined by

$$\tilde{\mathcal{E}}V = \mathcal{E}V + \sum_{k=1}^{n_d} (\bar{V}_k - \mathcal{E}_k \bar{V}_k).$$

Furthermore, define

$$\hat{S}_h^0(\Omega) = \{v \in S_h^0(\Omega) \mid v = 0 \text{ on } \Gamma\}$$

and

$$S_\Gamma(\Omega) = \{\tilde{\mathcal{E}}v \mid v \in S_h^0(\Omega)\}.$$

It is immediate that $\hat{S}_h^0(\Omega)$ and $S_\Gamma(\Omega)$ provide a direct sum decomposition of $S_h^0(\Omega)$.

The additive Schwarz preconditioner applied to $g \in S_h^0(\Omega)$ based on the above two spaces results in a function $Y = Y_0 + Y_\Gamma$ where $Y_0 \in \hat{S}_h^0(\Omega)$ satisfies

$$(3.14) \quad \mathcal{B}_0(Y_0, \phi) = (g, \phi), \text{ for all } \phi \in \hat{S}_h^0(\Omega)$$

and $Y_\Gamma \in S_\Gamma(\Omega)$ satisfies

$$(3.15) \quad \mathcal{B}_\Gamma(Y_\Gamma, \phi) = (g, \phi), \text{ for all } \phi \in S_\Gamma(\Omega).$$

Here $\mathcal{B}_0(\cdot, \cdot)$ and $\mathcal{B}_\Gamma(\cdot, \cdot)$ are symmetric and positive definite bilinear forms.

We shall see that the preconditioner in (3.3) is equivalent to the additive Schwarz method above when

$$(3.16) \quad \mathcal{B}_0(\varphi, \phi) = \sum_{k=1}^{n_d} \mathcal{B}_k(\varphi, \phi)$$

and

$$(3.17) \quad \mathcal{B}_\Gamma(\varphi, \phi) = h^{-1} \sum_{k=1}^{n_d} \tilde{a}_k \langle \varphi - \bar{\varphi}_k, \phi - \bar{\phi}_k \rangle_{\partial\Omega_k, h}.$$

Indeed, let W be the solution of (2.8). Then

$$(3.18) \quad \mathcal{B}(W, \varphi) = \mathcal{B}_k(W^{(k)}, \varphi) = (g, \varphi)_\Omega, \text{ for all } \varphi \in S_h^0(\Omega_k),$$

where $W^{(k)} \equiv W - \bar{W}_k - \mathcal{E}_k(W - \bar{W}_k)$. It follows that the function Y_0 satisfying (3.14) (with $\mathcal{B}_0(\cdot, \cdot)$ given by (3.16)) can be written

$$Y_0 = W - \tilde{\mathcal{E}}W \quad \text{on } \Omega_k.$$

Taking $\varphi = \tilde{\mathcal{E}}V$ in (2.8) shows that for $\mathcal{B}_\Gamma(\cdot, \cdot)$ given by (3.17),

$$\mathcal{B}_\Gamma(W, \tilde{\mathcal{E}}V) = (g, \tilde{\mathcal{E}}V).$$

Thus, $Y_\Gamma = W$ on Γ . From the definition of $S_\Gamma(\Omega)$ it follows that $Y_\Gamma = \tilde{\mathcal{E}}W$, i.e., $W = Y_0 + Y_\Gamma$. Thus, the solution W of (2.8) is the result of the additive Schwarz algorithm with subspace decomposition given by $\hat{S}_h^0(\Omega)$ and $S_\Gamma(\Omega)$ with forms defined by (3.16) and (3.17).

4. APPLICATION TO PARABOLIC PROBLEMS

Our preconditioning approach can be extended to more general bilinear forms of the type

$$\mathcal{A}(v, w) = \delta \sum_{i,j=1}^n \int_\Omega a_{ij} \frac{\partial v}{\partial x_i} \frac{\partial w}{\partial x_j} dx + (v, w)_\Omega.$$

Such forms arise from implicit time-stepping numerical approximations of parabolic problems. In such settings δ is related to the time step and is usually small. We shall consider the case when $ch^2 \leq \delta \leq Cd^2$.

We define our preconditioner $\mathcal{B}(\cdot, \cdot)$ by

$$\mathcal{B}(v, w) = \sum_{k=1}^{n_d} \mathcal{B}_k(v - \mathcal{E}_k v, w - \mathcal{E}_k w) + \frac{\delta}{h} \sum_{k=1}^{n_d} \langle w, v \rangle_{\partial\Omega_k, h},$$

where $\mathcal{B}_k(\cdot, \cdot)$ are the subdomain preconditioning forms satisfying (3.2). Note that the above form no longer includes the average values on the subdomain boundaries.

It is easy to see that

$$(4.1) \quad \begin{aligned} \mathcal{A}(v, v) &\leq 2[\mathcal{A}(v - \mathcal{E}v, v - \mathcal{E}v) + \mathcal{A}(\mathcal{E}v, \mathcal{E}v)] \\ &\leq C \left\{ \sum_{k=1}^{n_d} \mathcal{B}_k(v - \mathcal{E}_k v, v - \mathcal{E}_k v) + (h + \frac{\delta}{h}) \sum_{k=1}^{n_d} \langle v, v \rangle_{\partial\Omega_k, h} \right\} \\ &\leq C\mathcal{B}(v, v). \end{aligned}$$

Moreover, applying (3.5) gives

$$\frac{\delta}{h} \langle v, v \rangle_{\partial\Omega_k, h} \leq C \frac{\delta}{h} \left(\frac{1}{\epsilon} (v, v)_{\Omega_k} + \epsilon \mathcal{D}_k(v, v) \right).$$

Choosing $\epsilon = \max(\delta^{1/2}, d)$ in the last inequality yields

$$(4.2) \quad \frac{\delta}{h} \langle v, v \rangle_{\partial\Omega_k, h} \leq C \frac{\delta^{1/2}}{h} \mathcal{A}_k(v, v).$$

Using (4.2) for each k as in the proof of Theorem 3.1, we obtain

$$(4.3) \quad \mathcal{B}(v, v) \leq C \frac{\delta^{1/2}}{h} \mathcal{A}(v, v),$$

Combining (4.1) and (4.3) shows that

$$(4.4) \quad c\mathcal{A}(v, v) \leq \mathcal{B}(v, v) \leq C \frac{\delta^{1/2}}{h} \mathcal{A}(v, v) \quad \text{for all } v \in S_h^0(\Omega).$$

The resulting condition number depends on δ in a natural way. Smaller time steps correspond to better conditioning. Obviously, the preconditioner would be uniform if $\delta = h^2$ but such time stepping is too restrictive for the vast majority of applications. On the other hand, $\delta = h$ corresponds to a very reasonable time-stepping scheme whose condition number is governed by $h^{-1/2}$. Again, although not uniform, such rate of growth is often acceptable in practice for reasons already mentioned.

5. COMPUTATIONAL ASPECTS OF THE PRECONDITIONING PROBLEM

In this section, we provide an algorithm for applying the preconditioning operator corresponding to the form $\mathcal{B}(\cdot, \cdot)$. This consists of two main steps, solution of the approximate subdomain problems and inversion of the boundary form. As we shall see, these steps are independent and can be carried out in parallel.

5.1. The domain decomposition algorithm. The action of the preconditioner corresponding to $\mathcal{B}(\cdot, \cdot)$ is obtained by computing the solution of (2.8) for a given g . The first step involves the computation of $W^{(k)} \in S_h^0(\Omega_k)$ satisfying (3.18) and reduces to the solution of subdomain preconditioning problems which can be performed in parallel.

The second step involves the solution of a problem on Γ which we shall now describe. For $\psi \in S_h^0(\Omega)$ set $\varphi = \mathcal{E}\psi$. Notice that $\varphi = \mathcal{E}_k\psi$ on Ω_k , $(\bar{\mathcal{E}}\psi)_k = \bar{\psi}_k$, and $\mathcal{E}_k^2\psi = \mathcal{E}_k\psi$. For this choice of φ , (2.8) becomes

$$(5.1) \quad \begin{aligned} \mathcal{B}(W, \varphi) &= \sum_{k=1}^{n_d} \mathcal{B}_k(W - \bar{W}_k - \mathcal{E}_k(W - \bar{W}_k), \mathcal{E}_k\bar{\psi}_k - \bar{\psi}_k) \\ &\quad + h^{-1} \sum_{k=1}^{n_d} \tilde{a}_k \langle W - \bar{W}_k, \psi \rangle_{\partial\Omega_k, h} = (g, \mathcal{E}\psi)_\Omega. \end{aligned}$$

Here we have used the fact the $W - \bar{W}_k$ has zero discrete mean value on $\partial\Omega_k$ and therefore is orthogonal to constants with respect to the inner product on $\partial\Omega_k$.

Since $\mathcal{E}_k\bar{\psi}_k - \bar{\psi}_k$ vanishes on $\partial\Omega_k$,

$$\sum_{k=1}^{n_d} \mathcal{B}_k(W - \bar{W}_k - \mathcal{E}_k(W - \bar{W}_k), \mathcal{E}_k\bar{\psi}_k - \bar{\psi}_k) = \sum_{k=1}^{n_d} (g, \mathcal{E}_k\bar{\psi}_k - \bar{\psi}_k)_{\Omega_k}$$

and hence

$$(5.2) \quad h^{-1} \sum_{k=1}^{n_d} \tilde{a}_k \langle W - \bar{W}_k, \psi \rangle_{\partial\Omega_k, h} = (g, \mathcal{E}\psi)_\Omega - \sum_{k=1}^{n_d} (g, \mathcal{E}_k\bar{\psi}_k - \bar{\psi}_k)_{\Omega_k}.$$

Notice that because of the explicit extensions used in the definition of $\mathcal{B}(\cdot, \cdot)$, the setup of the right hand side in (5.2) involves minimal computational cost. Clearly, this step is independent of the previous one and thus the procedure solving the preconditioning problem with $\mathcal{B}(\cdot, \cdot)$ given by (3.3) decouples into two independent

tasks. Once $W^{(k)}$ and $W|_{\Gamma}$ are known then the assembly of the solution in Ω is easy. The actual implementation of the solution procedure for (5.2) is an important issue for the overall computational efficiency of the proposed preconditioner. We shall give a detailed description how to solve this problem in the next subsection.

The above discussion can be summarized in the following algorithm.

Algorithm 5.1. *Solve the preconditioning problem (2.8) by*

- (1) *Compute the solution $W^{(k)}$ of (3.18) for each k .*
- (2) *Compute the trace of W on Γ from (5.2).*
- (3) *The solution of (2.8) is given by*

$$W = \mathcal{E}W + \sum_{k=1}^{n_d} (W^{(k)} + \bar{W}_k - \mathcal{E}_k \bar{W}_k).$$

5.2. The algorithm for inverting the boundary form. In this subsection we describe the algorithm for solving (5.2). As it was observed in the previous subsection (this applies also to Section 5 below), the algorithm for inverting the preconditioner (3.3) requires an efficient method for determining the averages \bar{W}_k and finding the solution to (5.2). The implementation details of this method are described below. The algorithm for solving (5.2) was originally developed in [6] and it is included here for completeness.

We start by observing that the solution of (5.2) is trivial provided that \bar{W}_k is known for each k . In fact, the resulting matrix is diagonal using the usual nodal basis for $S_h(\Gamma)$ and thus inverting it is straightforward. Therefore, we only have to describe how to solve for \bar{W}_k .

For $\ell = 1, \dots, n_d$, let ψ^ℓ be the unique function in $S_h(\Gamma)$ which satisfies

$$(5.3) \quad h^{-1} \sum_{k=1}^{n_d} \tilde{a}_k \langle V, \psi^\ell \rangle_{\partial\Omega_k, h} = \bar{V}_\ell, \quad \text{for all } V \in S_h(\Gamma).$$

That such functions are uniquely defined follows from the Riesz Representation Theorem. Substituting ψ^ℓ in (5.2) gives

$$(5.4) \quad h^{-1} \left\{ \bar{W}_\ell - \sum_{k=1}^{n_d} \tilde{a}_k \bar{W}_k \langle 1, \psi^\ell \rangle_{\partial\Omega_k} \right\} = (g, \mathcal{E}\psi^\ell)_\Omega - \sum_{k=1}^{n_d} (g, \mathcal{E}_k \bar{\psi}_k^\ell - \bar{\psi}_k^\ell)_{\Omega_k}.$$

Setting $\bar{W} = [\bar{W}_1, \dots, \bar{W}_{n_d}]^T$, (5.4) can be rewritten in a matrix form as

$$(5.5) \quad \mathbf{M}\bar{W} = G.$$

It was observed in [6] that the matrix \mathbf{M} is symmetric and irreducibly diagonally dominant and hence positive definite. Thus (5.4) is solvable. One efficiently implements the above algorithm by explicitly computing the functions $\{\psi^\ell\}$. We illustrate this construction in the case when the operator \mathcal{L} from (2.2) is the Laplacian in two spatial dimensions and $\tilde{a}_k = 1$, $k = 1, \dots, n_d$. To do this we need to define some additional notation. The nodes on $\Gamma \setminus \partial\Omega$ that are shared by exactly m subdomains will be called **m -edge** nodes. For example, in the picture shown in Fig. 5.1, all nodes on $\Gamma \setminus \partial\Omega$ but node E are **2-edge** nodes. Node E of this example is a **4-edge** node. With this terminology in mind, we define ψ^ℓ by

$$(5.6) \quad \psi^\ell(x_i) = \begin{cases} \frac{1}{mN_\ell}, & \text{if } x_i \in \partial\Omega_\ell \setminus \partial\Omega \text{ and } x_i \text{ is an } m\text{-edge,} \\ 0, & \text{elsewhere.} \end{cases}$$

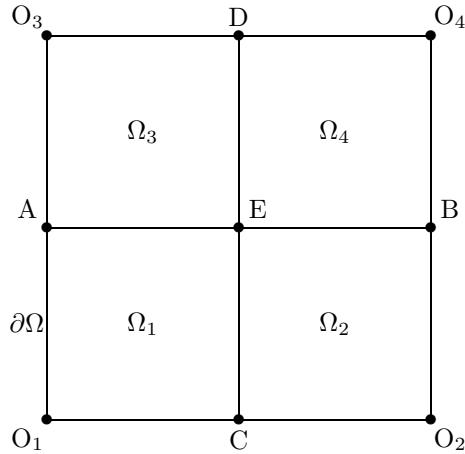


FIGURE 5.1. A simple example with four subdomains.

Here N_ℓ is the number of nodes on $\partial\Omega_\ell$. For the example shown in Fig. 5.1, there will be four such basis functions. The function associated with the first subdomain is ψ^1 such that $\psi^1(E) = 1/(4N_1)$; $\psi^1 \equiv 1/(2N_1)$ at all remaining nodes on the edges AE and EC, the points A and C excluded; $\psi^1 \equiv 0$ on the edges AO₁, O₁C, and everywhere in the exterior and interior of Ω_1 .

This approach to solving the problem for the averages extends to three dimensional problems as well as the case when $\tilde{a}_k \neq 1$. The reader is referred to [6] for further details.

6. ALTERNATIVE ADDITIVE PRECONDITIONERS WITH INEXACT SOLVES

In this section, we consider a classical technique for developing non-overlapping domain decomposition preconditioners. The behavior of such methods has been investigated in the case when the boundary form is uniformly equivalent to the corresponding Schur complement subsystem [2], [19]. Here, we show that this method also reduces to an additive Schwarz preconditioner. In addition, we show that the inexact solve technique combined with the boundary form discussed earlier provides an effective preconditioner. Indeed, our results are much better than what would be expected from the analysis of [2], [19].

6.1. Matrix form of the inexact solve domain decomposition algorithm. The classical inexact domain decomposition preconditioners are easily understood from the matrix point of view. In this case, one orders the unknowns so that the stiffness matrix corresponding to $\mathcal{A}(\cdot, \cdot)$ can be written in a block form as

$$\begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}.$$

Here \mathbf{A}_{22} corresponds to the nodes on Γ and \mathbf{A}_{11} to the remaining nodes. With this ordering, the form corresponding to a typical domain decomposition preconditioner

(e.g., [5],[6],[7],[8]) has a stiffness matrix of the form

$$\hat{\mathbf{A}} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{Z} \end{pmatrix},$$

where $\mathbf{Z} = \mathbf{B}_{22} + \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}$ and \mathbf{B}_{22} is the domain decomposition boundary preconditioning matrix. Computing the inverse of $\hat{\mathbf{A}}$ applied to a vector reduces to a three step block Gaussian elimination procedure.

The classical inexact method is defined by replacing \mathbf{A}_{11} with \mathbf{B}_{11} where \mathbf{B}_{11} is another symmetric and positive definite matrix. This defines a new preconditioning operator \mathbf{B} given by

$$(6.1) \quad \mathbf{B} = \begin{pmatrix} \mathbf{B}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \tilde{\mathbf{Z}} \end{pmatrix}.$$

Here $\tilde{\mathbf{Z}}$ is given by $\tilde{\mathbf{Z}} = \mathbf{B}_{22} + \mathbf{A}_{21}\mathbf{B}_{11}^{-1}\mathbf{A}_{12}$.

Generally, the inexact algorithm may not converge as well as the exact version. Even if one takes \mathbf{B}_{22} to be the Schur complement, $\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{B}_{11}^{-1}\mathbf{A}_{12}$, the inexact preconditioner may perform poorly unless the difference between the two matrices \mathbf{B}_{11} and \mathbf{A}_{11} is sufficiently small in an appropriate sense (see Theorem 6.1).

6.2. The inexact algorithm as a two level additive Schwarz procedure.

We now show that the inexact preconditioners correspond to additive Schwarz methods. The first subspace in this decomposition is $\hat{S}_h^0(\Omega)$. Let $\mathcal{B}_0(\cdot, \cdot)$ be the form on $\hat{S}_h^0(\Omega) \times \hat{S}_h^0(\Omega)$ with stiffness matrix \mathbf{B}_{11} . The second subspace is given by

$$(6.2) \quad \begin{aligned} \hat{S}_h(\Gamma) &= \left\{ \mathcal{E}\varphi + \varphi_0 \mid \varphi \in S_h^0(\Omega) \text{ and } \varphi_0 \in \hat{S}_h^0(\Omega) \text{ such that} \right. \\ &\quad \left. \mathcal{B}_0(\varphi_0, \phi) = -\mathcal{A}(\mathcal{E}\varphi, \phi), \text{ for all } \phi \in \hat{S}_h^0(\Omega) \right\}. \end{aligned}$$

Clearly, the functions in $\hat{S}_h(\Gamma)$ are completely determined by their traces on Γ . Let $\mathcal{B}_\Gamma(\cdot, \cdot)$ be the form on $\hat{S}_h(\Gamma) \times \hat{S}_h(\Gamma)$ with stiffness matrix \mathbf{B}_{22} . $\mathcal{B}_\Gamma(u, v)$ only depends on the boundary nodal values of u and v and is thus defined on $S_h^0(\Omega) \times S_h^0(\Omega)$ by restriction.

Clearly, $\hat{S}_h^0(\Omega)$ and $\hat{S}_h(\Gamma)$ provide a direct sum decomposition of $S_h^0(\Omega)$. This decomposition is tied strongly to the bilinear form $\mathcal{B}_0(\cdot, \cdot)$. In particular, if $\mathcal{B}_0(\cdot, \cdot) \equiv \mathcal{A}(\cdot, \cdot)$ on $\hat{S}_h^0(\Omega) \times \hat{S}_h^0(\Omega)$, then the space $\hat{S}_h(\Gamma)$ consists of discrete harmonic functions and the decomposition is $\mathcal{A}(\cdot, \cdot)$ -orthogonal. In general, the decomposition is not $\mathcal{A}(\cdot, \cdot)$ -orthogonal.

6.3. Conditioning estimates for the inexact algorithms. The preconditioner defined by (6.1) can be restated as an operator $\mathbf{B} : S_h^0(\Omega) \mapsto S_h^0(\Omega)$. In fact, it is a straightforward exercise to check that the block Gaussian elimination procedure applied to the matrix \mathbf{B} of (6.1) corresponds to the preconditioning operator defined in the following algorithm.

Algorithm 6.1. *Given $g \in S_h^0(\Omega)$ we define $\mathbf{B}^{-1}g = U$ where U is computed as follows:*

- (1) *Compute $U_0 \in \hat{S}_h^0(\Omega)$ by solving*

$$(6.3) \quad \mathcal{B}_0(U_0, \varphi) = (g, \varphi) \quad \text{for all } \varphi \in \hat{S}_h^0(\Omega).$$

(2) Compute the trace U_Γ on Γ by solving

$$\mathcal{B}_\Gamma(U_\Gamma, \mathcal{E}\phi) = (g, \mathcal{E}\phi) - \mathcal{A}(U_0, \mathcal{E}\phi) \quad \text{for all } \phi \in \hat{S}_h(\Gamma).$$

(3) Compute U_{Γ_0} by solving

$$\mathcal{B}_0(U_{\Gamma_0}, \varphi) = -\mathcal{A}(\mathcal{E}U_\Gamma, \varphi) \quad \text{for all } \varphi \in \hat{S}_h^0(\Omega).$$

(4) Set $U = U_0 + \mathcal{E}U_\Gamma + U_{\Gamma_0}$.

Although the above algorithm appears as a multiplicative procedure, we shall now demonstrate that it is equivalent to an additive Schwarz method. The problem solved in Step (2) of Algorithm 6.1 is independent of U_0 . Indeed, for any $\phi \in \hat{S}_h(\Gamma)$, we decompose $\phi = \mathcal{E}\phi + \phi_0$ as in (6.2) and observe

$$-\mathcal{A}(\mathcal{E}\phi, U_0) = \mathcal{B}(\phi_0, U_0) = (g, \phi_0).$$

Thus, Steps (2) and (3) of the above algorithm reduce to finding $U_\Gamma = \mathcal{E}U_\Gamma + U_{\Gamma_0} \in \hat{S}_h(\Gamma)$ such that

$$(6.4) \quad \mathcal{B}_\Gamma(U_\Gamma, \phi) = (g, \phi) \text{ for all } \phi \in \hat{S}_h(\Gamma).$$

Hence, $\mathbf{B}^{-1}g = U = U_0 + U_\Gamma$ where U_0 and U_Γ satisfy (6.3) and (6.4) respectively, i.e., Algorithm 6.1 is an implementation of an additive Schwarz procedure.

Notice that Algorithm 6.1 avoids the need of knowing explicitly a basis for the space $\hat{S}_h(\Gamma)$ which could be either a computationally expensive problem or a significant complication of the overall algorithm. Obviously this procedure provides inexact variants of the methods given in [5], [6], [7], [8], [14] and [23].

Since $\hat{S}_h^0(\Omega)$ and $\hat{S}_h(\Gamma)$ give a direct sum decomposition of $S_h^0(\Omega)$, the preconditioning form $\mathcal{B}(\cdot, \cdot)$ corresponding to the operator defined in Algorithm 6.1 is given by

$$(6.5) \quad \mathcal{B}(V, V) = \mathcal{B}_0(V_0, V_0) + \mathcal{B}_\Gamma(V_\Gamma, V_\Gamma).$$

Here $V = V_0 + V_\Gamma$ with $V_0 \in \hat{S}_h^0(\Omega)$ and $V_\Gamma \in \hat{S}_h(\Gamma)$. In the remainder of this section we analyze the above preconditioner by providing bounds for (6.5). We take

$$\mathcal{B}_0(u, v) = \sum_{k=1}^{n_d} \mathcal{B}_k(u, v)$$

where $\mathcal{B}_k(\cdot, \cdot)$ is defined as in Section 3.

The first theorem in this section was given by Börgers [2] and Haase et al. [19] and provides a result when \mathbf{B}_{22} is uniformly equivalent to the Schur complement $\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}$. This is the same as assuming that the quadratic form $\mathcal{B}_\Gamma(\cdot, \cdot)$ is equivalent to the boundary form with diagonal

$$(6.6) \quad \inf_{\phi \in \hat{S}_h^0(\Omega)} \mathcal{A}(u + \phi, u + \phi), \quad \text{for all } u \in \hat{S}_h(\Gamma).$$

Theorem 6.1. *Let $\mathcal{A}(\cdot, \cdot)$ be given by (2.3) and $\mathcal{B}(\cdot, \cdot)$ by (6.5) respectively. Assume that the quadratic form $\mathcal{B}_\Gamma(\cdot, \cdot)$ is uniformly equivalent to the quadratic form induced by (6.6). In addition, let γ be the smallest positive constant such that*

$$(6.7) \quad |\mathcal{A}(\varphi, \varphi) - \mathcal{B}(\varphi, \varphi)| \leq \gamma \mathcal{A}(\varphi, \varphi) \quad \text{for all } \varphi \in \hat{S}_h^0(\Omega).$$

Then

$$(6.8) \quad c \left(\frac{\gamma^2}{h} \right)^{-1} \mathcal{A}(U, U) \leq \mathcal{B}(U, U) \leq C \frac{\gamma^2}{h} \mathcal{A}(U, U)$$

holds for all $U \in S_h^0(\Omega)$ with constants c and C independent of d and h .

Remark 6.1. Condition (6.7) requires $\mathcal{B}_0(\cdot, \cdot)$ to be a good approximation to $\mathcal{A}(\cdot, \cdot)$ for the preconditioner (6.5) to be efficient. The result of the theorem shows that if (6.7) holds with γ on the order of $h^{1/2}$, then the preconditioner $\mathcal{B}(\cdot, \cdot)$ is uniform. However, the development of a form $\mathcal{B}_0(\cdot, \cdot)$ satisfying (6.7) usually involves significant additional computational work since γ must tend to zero as h becomes small. Alternatively keeping γ fixed independent of h may result in a rather ill-conditioned method when h is small. There are examples of reasonably accurate preconditioners $\mathcal{B}_0(\cdot, \cdot)$, e.g. multigrid V- or W-cycles, which appear to perform well when h is not very small (cf. [2]) due to the fact that the corresponding γ 's are comparable to $h^{1/2}$.

The main result of this section is given in the next theorem. It is for the case when

$$(6.9) \quad \mathcal{B}_\Gamma(u, v) = h^{-1} \sum_{k=1}^{n_d} \tilde{a}_k \langle u - \bar{u}_k, v - \bar{v}_k \rangle_{\partial\Omega_k, h}, \quad \text{for all } u, v \in \hat{S}_h(\Gamma).$$

Theorem 6.2. *Let $\mathcal{A}(\cdot, \cdot)$ be given by (2.3), $\mathcal{B}(\cdot, \cdot)$ be given by (6.5), and $\mathcal{B}_\Gamma(\cdot, \cdot)$ defined by (6.9). Then*

$$(6.10) \quad c\mathcal{A}(U, U) \leq \mathcal{B}(U, U) \leq C \frac{d}{h} \mathcal{A}(U, U)$$

holds for all $U \in S_h^0(\Omega)$ with constants c and C independent of d and h .

Remark 6.2. The result of Theorem 6.2 shows that introducing inexact solves in the interior of the subdomains does not degrade the overall preconditioning effect of the corresponding exact method analyzed in [6]. As we have pointed out in Section 3, the adverse effect on the condition number of h approaching zero can be compensated easily by adjusting the parameter d . This balance is an alternative to (6.7) and could be a better choice when h is small relative to γ . In fact, the utilization of the bilinear form (6.9) leads to computationally efficient algorithms, unconstrained by accuracy conditions like (6.7). We shall see in Section 7 that for this boundary form the differences in the preconditioning effect of the inexact (Algorithm 6.1) and exact (cf. [6]) methods are negligible. However, the saving of computational time is significantly in favor of Algorithm 6.1.

We conclude this section with the proof of Theorem 6.2.

Proof of Theorem 6.2. Because of (6.5), the technique for establishing (6.10) is similar to the one used in the proof of Theorem 3.1.

Let $U_\Gamma = \mathcal{E}U_\Gamma + U_{\Gamma 0}$ as in (6.2) and write $U = U_0 + U_\Gamma$. The first inequality in (6.10) follows from the arithmetic–geometric mean inequality and the assumptions on $\{\mathcal{B}_k(\cdot, \cdot)\}$. Indeed, we have

$$(6.11) \quad \begin{aligned} \mathcal{A}(U, U) &= \mathcal{A}(U_0 + U_\Gamma, U_0 + U_\Gamma) \\ &\leq C (\mathcal{B}_0(U_0, U_0) + \mathcal{B}_0(U_{\Gamma 0}, U_{\Gamma 0}) + \mathcal{A}(\mathcal{E}U_\Gamma, \mathcal{E}U_\Gamma)). \end{aligned}$$

If follows from the definition of $U_{\Gamma 0}$ that

$$(6.12) \quad \mathcal{B}_0(U_{\Gamma 0}, U_{\Gamma 0}) \leq C\mathcal{A}(\mathcal{E}U_\Gamma, \mathcal{E}U_\Gamma).$$

Using (6.12) together with (3.7) and (3.4) in (6.11) yields

$$\mathcal{A}(U, U) \leq C\mathcal{B}(U, U).$$

To prove the right hand inequality in (6.10), we use again the decomposition of U . Thus,

$$(6.13) \quad \begin{aligned} \mathcal{B}_0(U_0, U_0) &\leq C\mathcal{A}(U - U_\Gamma, U - U_\Gamma) \leq C(\mathcal{A}(U, U) + \mathcal{A}(\mathcal{E}U_\Gamma, \mathcal{E}U_\Gamma)) \\ &\leq C(\mathcal{A}(U, U) + \mathcal{B}_\Gamma(U_\Gamma, U_\Gamma)). \end{aligned}$$

Hence, we need to estimate $\mathcal{B}_\Gamma(U_\Gamma, U_\Gamma)$ from above by $\mathcal{A}(U, U)$. Applying the reasoning used to show (3.13) in (6.13) gives the desired bound. \square

7. NUMERICAL EXAMPLES

In this section we present numerical calculations involving the non-overlapping domain decomposition preconditioners developed in Section 3 and Section 5. We report results obtained from examples with Algorithm 5.1 and Algorithm 6.1 with boundary form given by (6.9). We tested two main aspects of these preconditioners, namely the computational efficiency of the method, in terms of the condition numbers obtained, and the independence of the jumps in the operator coefficients $\{a_{ij}\}$. Comparisons between the inexact algorithms and the corresponding exact methods are included as well.

The numerical results presented in this section are applied to

$$(7.1) \quad \mathcal{L} = -\nabla \cdot a \nabla,$$

where a is a piecewise constant function in Ω and constant on each subdomain. In all of our calculations Ω is the unit cube in three spatial dimensions. The subdomains are obtained by subdividing Ω into regions by slicing it parallel to the coordinate axes. Here we shall consider only cases where the unit cube is split into m^3 equal sub-cubes, which implies $d = 1/m$. In the examples below, $S_h^0(\Omega)$ is the space of piecewise linear functions with respect to a uniform mesh of size h . Also, the action of one multigrid V-cycle is used as an inexact solver in the interior of the subdomains.

The multigrid algorithm is variational and based on a trilinear finite element approximation. A nested sequence of approximation subspaces is defined by successively doubling the mesh size. For computational efficiency, the fine grid form is defined by numerical quadrature utilizing a quadrature which gives rise to a seven point operator. The operators on the coarser grids are twenty seven point and determined variationally from the fine grid operator. The analysis of variational multigrid procedures based on a fine grid operator defined by numerical quadrature can be found in [3]. Pointwise forward and backward Gauss-Seidel sweeps are used as pre- and post-smoothing iterations respectively. On the coarsest level we apply five pairs of forward and backward Gauss-Seidel sweeps. Obviously, if we have only one degree of freedom on the coarsest level, then this is equivalent to an exact solve on that level. This multigrid procedure results in a symmetric and positive definite operator whose action provides an inexact interior solve. The corresponding $\mathcal{B}_k(\cdot, \cdot)$ satisfies (3.2) with uniform constants c_k and C_k for each k . Also, the evaluation of the action of this operator is proportional to the number of grid points on the mesh used for the discretization of Ω_k .

The first cases which we report are intended to confirm numerically the d/h -like behavior of the condition number K , established in Theorem 3.1. We consider the model problem (2.1) with $\mathcal{L} \equiv -\Delta$. The results are presented in Table 7.1. According to our theory, the condition number K should be bounded if d/h is fixed. This is clearly indicated in the computational results of Table 7.1.

TABLE 7.1. Condition numbers with the inexact preconditioner (3.3).

h	$d = 1/3$	$d = 1/6$
1/12	21.46	8.12
1/24	55.70	23.20
1/48	131.19	59.33

TABLE 7.2. Condition numbers with the inexact preconditioner (3.3); $d = 1/4$.

h	Variable a	$a \equiv 1$
1/12	15.71	13.87
1/24	42.94	39.79
1/48	106.76	95.38

TABLE 7.3. Comparison of the inexact and the exact methods; $d = 1/3$.

h	K_{exact}	K -Algorithm 5.1	K -Algorithm 6.1
1/6	6.27	6.73	6.27
1/12	15.23	21.96	15.40
1/24	32.55	57.01	33.83
1/48	66.12	130.88	70.76

The second set of calculations illustrate that the condition number for the preconditioner defined in (3.3) can be bounded independently of large jumps in the operator coefficients. The data in Table 7.2 represent experimental results where Ω is split into $4 \times 4 \times 4$ subdomains. The coefficient a in (7.1) is defined as follows: $a_{222} = a_{333} = 10^5$, a is a constant in the interval $[0.1, 21.1]$ for the remaining subdomains. Here a_{ijk} is the operator coefficient in the subdomain with integer coordinates i, j, k . The largest jump in the operator coefficient between two neighboring subdomains in this case is 10^6 . For comparison, we have included the corresponding condition numbers for the case when $a \equiv 1$ in Ω . Clearly, the results in Table 7.2 are in good agreement with Remark 3.2.

Our final numerical example is a comparison of the performance of the inexact preconditioners (3.3) and (6.5) with $\mathcal{B}_\Gamma(\cdot, \cdot)$ given by (6.9), and the exact method analyzed in [6]. The piecewise constant coefficient a in this case is defined according to the data for μ in Example 3 in [6]. We note that the condition numbers for the exact method reported in Table 7.3 are better than the ones reported in Table 4.5 in [6] due to the different scaling of the boundary form (cf. Remark 2.5, [6]). The data in Table 4.5, [6] are obtained when the boundary form is scaled by d^{-1} whereas the results in Table 7.3 are obtained with the scaling h^{-1} . Clearly, the exact preconditioner and the inexact method implemented by Algorithm 6.1 exhibit almost the same condition numbers which is in good agreement with Remark 6.2.

Although the condition numbers reported for these two methods are better than those for Algorithm 5.1, one application of the inexact preconditioner (3.3) requires substantially less computer time thus resulting in a more efficient computational algorithm. We illustrate this with some timing statistics made on a SUN Sparc 20/502 workstation. For mesh sizes between 1/12 and 1/48, the inexact preconditioner (Algorithm 5.1) was more than 4.5 times faster to evaluate than the

exact method which used the Fast Fourier transform to diagonalize the stiffness matrix. This translates into an overall factor of three reduction in computing time in the case of the grid with $h = 1/48$ and a problem solved by a preconditioned conjugate gradient iteration. A similar comparison of Algorithm 5.1 and Algorithm 6.1 indicates that Algorithm 5.1 is about 25 percent more efficient.

REFERENCES

1. P.E. Bjørstad and O.B. Widlund, *Iterative methods for the solution of elliptic problems on regions partitioned into substructures*, SIAM J. Numer. Anal. **23** (1986), 1097–1120. MR **88h**:65188
2. C. Börgers, *The Neumann–Dirichlet Domain Decomposition Method with Inexact Solvers on the Subdomains*, Numerische Mathematik **55** (1989), 123–136. MR **90f**:65191
3. J.H. Bramble, C.I. Goldstein, and J.E. Pasciak, *Analysis of V -cycle multigrid algorithms for forms defined by numerical quadrature*, SIAM Sci. Stat. Comp. **15** (1994), 566–576. MR **95b**:65047
4. J.H. Bramble, J.E. Pasciak and A.H. Schatz , *An iterative method for elliptic problems on regions partitioned into substructures*, Math. Comp. **46** (1986), 361–369. MR **88a**:65123
5. J.H. Bramble, J.E. Pasciak and A.H. Schatz , *The construction of preconditioners for elliptic problems by substructuring, I*, Math. Comp. **47** (1986), 103–134. MR **87m**:65174
6. J.H. Bramble, J.E. Pasciak and A.H. Schatz , *The construction of preconditioners for elliptic problems by substructuring, II*, Math. Comp. **49** (1987), 1–16. MR **88j**:65248
7. J.H. Bramble, J.E. Pasciak and A.H. Schatz , *The construction of preconditioners for elliptic problems by substructuring, III*, Math. Comp. **51** (1988), 415–430. MR **89e**:65118
8. J.H. Bramble, J.E. Pasciak and A.H. Schatz , *The construction of preconditioners for elliptic problems by substructuring, IV*, Math. Comp. **53** (1989), 1–24. MR **89m**:65098
9. J.H. Bramble, J.E. Pasciak, J. Wang, and J. Xu, *Convergence estimates for product iterative methods with applications to domain decomposition*, Math. Comp. **57** (1991), 1–21. MR **92d**:65094
10. J.H. Bramble, J.E. Pasciak and J. Xu, *Parallel multilevel preconditioners*, Math. Comp. **55** (1990), 1–22. MR **90k**:65170
11. J.H. Bramble, J.E. Pasciak and J. Xu, *A multilevel preconditioner for domain decomposition boundary systems*, Proceedings of the 10'th Inter. Conf. on Comput. Meth. in Appl. Sci. and Engr., Nova Sciences, New York, 1992.
12. J.H. Bramble and J. Xu, *Some estimates for weighted L^2 projections*, Math. Comp. **56** (1991), 463–476. MR **91k**:65140
13. L.C. Cowsar, J. Mandel, and M.F. Wheeler, *Balancing Domain Decomposition for Mixed Finite Elements*, Math. Comp. **64** (1995), 989–1015. MR **95j**:65161
14. M. Dryja, *A capacitance matrix method for the Dirichlet problem on a polygonal region*, Numer. Math. **39** (1982), 51–64. MR **83g**:65102
15. M. Dryja, *A method of domain decomposition for three-dimensional finite element elliptic problems*, First International Symposium on Domain Decomposition Methods for Partial Differential Equations, (eds, R. Glowinski, G.H. Golub, G.A. Meurant, and J. Périault) SIAM, Phil. PN, 1988, pp. 43–61. MR **90b**:65200
16. M. Dryja, B.F. Smith, and O.B. Widlund, *Schwarz analysis of iterative substructuring algorithms for elliptic problems in three dimensions*, SIAM J. Num. Anal. **31** (1994), 1662–1694. MR **95m**:65211
17. M. Dryja and O.B. Widlund, *Domain decomposition algorithms with small overlap*, SIAM J. Sci. Comput. **15** (1994), 604–620. MR **95d**:65102
18. R. Gonzalez and M.F. Wheeler, *Domain decomposition for elliptic partial differential equations with Neumann boundary conditions*, Proceedings, International conference on vector and parallel computing—issues in applied research and development, Loen, 1986., vol. 5, Parallel Comput., 1987, pp. 257–263. MR **88d**:65148
19. G. Haase, U. Langer, and A. Meyer, *The Approximate Dirichlet Domain Decomposition Method. Part I: An Algebraic Approach*, Computing **47** (1991), 137–151. MR **93e**:65146a
20. S.V. Nepomnyaschikh, *Application of domain decomposition to elliptic problems with discontinuous coefficients*, Fourth International Symposium on Domain Decomposition Methods for

- Partial Differential Equations, (eds, R. Glowinski, Y.A. Kuznetsov, G.A. Meurant, and J. Périaux) SIAM, Phil. PN, 1991, pp. 242–251. CMP 91:12
21. B.F. Smith, *Domain Decomposition Algorithms for the Partial Differential Equations of Linear Elasticity*, Ph.D. Thesis, Courant Institute of Mathematical Sciences, Dept. of Computer Science Tech. Rep. 517, New York, 1990.
 22. A.T. Vassilev, *On Discretization and Iterative Techniques for Second-Order Problems with Applications to Multiphase Flow in Porous Media*, Ph.D. Thesis, Texas A&M University, College Station, Texas, 1996.
 23. O. Widlund, *Iterative substructuring methods: algorithms and theory for elliptic problems in the plane*, First International Symposium on Domain Decomposition Methods for Partial Differential Equations (R. Glowinski, G.H. Golub, G.A. Meurant, and J. Périaux, eds.), SIAM, Phil. PN, 1988, pp. 113–128. MR 90c:65138
 24. J. Xu, *Iterative methods by space decomposition and subspace correction*, vol. 34, SIAM Review, 1992, pp. 581–613. MR 93k:65029

DEPARTMENT OF MATHEMATICS, TEXAS A&M UNIVERSITY, COLLEGE STATION, TEXAS 77843
E-mail address: bramble@math.tamu.edu

DEPARTMENT OF MATHEMATICS, TEXAS A&M UNIVERSITY, COLLEGE STATION, TEXAS 77843
E-mail address: pasciak@math.tamu.edu

SCHLUMBERGER, 8311 N. FM 620 RD., AUSTIN, TEXAS 78726
E-mail address: vassilev@slb.com

*4.10. ANALYSIS OF NON-OVERLAPPING DOMAIN DECOMPOSITION
ALGORITHMS WITH INEXACT SOLVES*

666666 c2dc0ab402d56a302fcaa5bc098dd2ea0060ad1f

5

PML Methods

5.1 Analysis of a finite PML approximation for the three dimensional time-harmonic Maxwell and acoustic scattering problems

Analysis of a finite PML approximation for the three dimensional time-harmonic Maxwell and acoustic scattering problems[4]

ANALYSIS OF A FINITE PML APPROXIMATION FOR THE THREE DIMENSIONAL TIME-HARMONIC MAXWELL AND ACOUSTIC SCATTERING PROBLEMS

JAMES H. BRAMBLE AND JOSEPH E. PASCIAK

ABSTRACT. We consider the approximation of the frequency domain three-dimensional Maxwell scattering problem using a truncated domain perfectly matched layer (PML). We also treat the time-harmonic PML approximation to the acoustic scattering problem. Following work of Lassas and Somersalo in 1998, a transitional layer based on spherical geometry is defined, which results in a constant coefficient problem outside the transition. A truncated (computational) domain is then defined, which covers the transition region. The truncated domain need only have a minimally smooth outer boundary (e.g., Lipschitz continuous). We consider the truncated PML problem which results when a perfectly conducting boundary condition is imposed on the outer boundary of the truncated domain. The existence and uniqueness of solutions to the truncated PML problem will be shown provided that the truncated domain is sufficiently large, e.g., contains a sphere of radius R_t . We also show exponential (in the parameter R_t) convergence of the truncated PML solution to the solution of the original scattering problem inside the transition layer.

Our results are important in that they are the first to show that the truncated PML problem can be posed on a domain with nonsmooth outer boundary. This allows the use of approximation based on polygonal meshes. In addition, even though the transition coefficients depend on spherical geometry, they can be made arbitrarily smooth and hence the resulting problems are amenable to numerical quadrature. Approximation schemes based on our analysis are the focus of future research.

1. INTRODUCTION

In this paper, we consider the acoustic and electromagnetic scattering problem in three spatial dimensions. Simulations involving these problems are inherently difficult for a number of reasons. First, although the problems are symmetric, they are indefinite. Second, the problems have a scale related to the wavenumber k , so standard discretizations require mesh sizes proportional to k^{-1} . Third, the problems are posed on infinite domains.

The focus of this paper is on the third issue above, i.e., how to deal with the boundary condition at infinity in a computationally effective way. Specifically, we

Received by the editor August 9, 2005 and, in revised form, April 20, 2006.

2000 *Mathematics Subject Classification*. Primary 78M10, 65F10, 65N30.

Key words and phrases. Maxwell's equations, Helmholtz equation, time-harmonic acoustic and electromagnetic scattering, div-curl systems, perfectly matched layer, PML.

This work was supported in part by the National Science Foundation through grant No. 0311902.

©2006 American Mathematical Society
Reverts to public domain 28 years from publication

shall study perfectly matched layer (PML) approximations to acoustic and electromagnetic problems. The goal is to demonstrate both the solvability of the continuous PML approximations and the convergence of the resulting solutions to the solutions of the original acoustic/electromagnetic problem.

Recently, there has been intensive computational and theoretical research toward understanding the properties of PML approximations. The research into the computational aspects of these methods is the subject of many papers in the engineering literature and we shall not attempt to discuss them here. There is evidence to suggest that this approach is very competitive with standard techniques for computational domain truncation. In the paper by Petropoulos [14] he refers to recent numerical results in [9], which “indicate our reflectionless sponge layer provides levels of numerical reflection from $\partial\Omega^c$ that are comparable to those obtained with the exact ABC [8] for vector spherical waves scattering from a dielectric sphere but at a substantial savings in computational cost.”

The original PML method was suggested by Bérenger in [3] and [2]. The observation that a PML method could be considered as a complex change of variable was made by Chew and Weedon [4]. Using this technique, Collino and Monk [5] derived PML equations based on rectangular and polar coordinates. There, they also showed the existence and uniqueness of solutions of the truncated acoustic PML except for a countable number of wave numbers. The formulation of PML equations for (2.1) in spherical coordinates can be found in [12]. Lassas and Somersalo [10] proved the existence and uniqueness of the PML acoustic approximation on a truncated domain where the outer boundary was circular. In a later paper [11], they extended these results to smooth convex domains in \mathbb{R}^n .

To date, there has been relatively little analysis of the truncated electromagnetic PML equations. Techniques for the acoustic problem do not carry over directly to the electromagnetic problem. This stems from the fact that the acoustic problem is strongly elliptic (up to perturbation) while the electromagnetic operator has an infinite dimensional kernel consisting of functions which are gradients. For example, Collino and Monk [5] use a perturbation analysis to derive their existence result. Carrying this argument over to the electromagnetic PML poses significant analytical difficulties requiring the analysis of vector decompositions involving the complex-valued PML coefficient. We will present a new analytical approach for the study of the electromagnetic PML equation in this paper.

Let Ω (the scatterer) be a domain in \mathbb{R}^3 . We shall first consider the acoustic scattering problem with a *sound-soft* obstacle. This involves a scalar function u defined on Ω^c , the complement of $\bar{\Omega}$, satisfying

$$(1.1) \quad \begin{aligned} \Delta u + k^2 u &= 0 \text{ in } \Omega^c, \\ u &= g \text{ on } \partial\Omega, \\ \lim_{\rho \rightarrow \infty} \rho(\nabla u \cdot \hat{x} - iku) &= 0. \end{aligned}$$

Here $\rho = |\mathbf{x}|$, $\hat{\mathbf{x}} = \mathbf{x}/\rho$, and k is a real positive constant. We have absorbed the medium properties into the constant k .

We will also consider the time-harmonic electromagnetic scattering problem. In this case, we shall assume, for convenience, that Ω is simply connected with only

one boundary component. We seek vector fields \mathbf{E} and \mathbf{H} defined on Ω^c satisfying

$$(1.2) \quad \begin{aligned} -ik\mu\mathbf{H} + \nabla \times \mathbf{E} &= \mathbf{0}, \text{ in } \Omega^c, \\ -ik\epsilon\mathbf{E} - \nabla \times \mathbf{H} &= \mathbf{0}, \text{ in } \Omega^c, \\ \mathbf{n} \times \mathbf{E} &= \mathbf{n} \times \mathbf{g}, \text{ on } \partial\Omega, \\ \lim_{\rho \rightarrow \infty} \rho(\mu\mathbf{H} \times \hat{\mathbf{x}} - \mathbf{E}) &= \mathbf{0}. \end{aligned}$$

Here \mathbf{g} results from a given incidence field, μ is the magnetic permeability, ϵ is the electric permittivity, and \mathbf{n} is the outward unit normal on $\partial\Omega$. The last line corresponds to the Silver-Müller condition at infinity. We assume that the coefficients μ and ϵ are real valued, bounded away from zero and constant outside of some ball.

We introduce some notation that will be used in the remainder of the paper. For a domain D , let $L^2(D)$ be the space of (complex-valued) square integrable functions on D and $\mathbf{L}^2(D) = (L^2(D))^3$ the space of vector-valued L^2 -functions. We shall use $(\cdot, \cdot)_\Omega$ to denote the (vector or scalar Hermitian) $L^2(\Omega)$ inner product and $\langle \cdot, \cdot \rangle_\Gamma$ to denote the (vector or scalar Hermitian) $L^2(\Gamma)$ boundary inner product. When the inner product is on all of \mathbb{R}^3 , we will use the notation (\cdot, \cdot) . The scalar and vector Sobolev spaces on D will be denoted $H^s(D)$ and $\mathbf{H}^s(D)$, respectively. Let $\mathbf{H}(\mathbf{curl}; D)$ be the set of vector-valued functions, which along with their curls, are in $\mathbf{L}^2(D)$. $\mathbf{H}_0(\mathbf{curl}; D)$ denotes the functions \mathbf{f} in $\mathbf{H}(\mathbf{curl}; D)$ satisfying $\mathbf{n} \times \mathbf{f} = \mathbf{0}$ on $\partial\Omega$. We assume that $\mathbf{n} \times \mathbf{g}$ above is the trace $\mathbf{n} \times \hat{\mathbf{g}}$ of a function $\hat{\mathbf{g}} \in \mathbf{H}(\mathbf{curl}; \Omega^c)$ supported close to $\partial\Omega$.

For a subdomain $D \subset \Omega_\infty$, by extension by zero, we identify $H_0^1(D)$ (respectively, $\mathbf{H}_0(\mathbf{curl}; D)$) with $\{v \in H_0^1(\Omega^c) \text{ (respectively, } \mathbf{H}_0(\mathbf{curl}; \Omega^c)) : \text{supp}(v) \subseteq \bar{D}\}$.

2. THE BÉRENGER LAYER

For convenience, we shall take $\mu = \epsilon = 1$ in (1.2) as all of our results extend to the more general case as long as the coefficients are constant outside of a ball of radius r_0 . We can reduce to a single equation involving \mathbf{E} by eliminating \mathbf{H} in (1.2). This gives

$$(2.1) \quad \begin{aligned} -\nabla \times \nabla \times \mathbf{E} + k^2 \mathbf{E} &= \mathbf{0} \text{ in } \Omega^c, \\ \mathbf{n} \times \mathbf{E} &= \mathbf{n} \times \mathbf{g} \text{ on } \partial\Omega, \\ \lim_{\rho \rightarrow \infty} \rho((\nabla \times \mathbf{E}) \times \hat{\mathbf{x}} - ik\mathbf{E}) &= \mathbf{0}. \end{aligned}$$

Throughout this paper, we shall use a sequence of finite subdomains of Ω^c with spherical outer boundaries. Let $r_0 < r_1 < \dots < r_4$ be an increasing sequence of real numbers and let Ω_i denote (interior of) the open ball B_i of radius r_i excluding $\bar{\Omega}$ (we assume that r_0 is large enough so that the corresponding ball contains $\bar{\Omega}$ and that the origin is contained in Ω). We denote the outer boundary of Ω_i by Γ_i . The values of r_0, r_1, \dots, r_4 are independent of the computational outer boundary scaling parameter R_t (introduced below).

As discussed in [5], the PML problem can be viewed as a complex coordinate transformation. Following [10], a transitional layer based on spherical geometry is defined, which results in a constant coefficient problem outside the transition.

Given σ_0 , r_1 , and r_2 , we start with a function $\tilde{\sigma} \in C^2(\mathbb{R}^+)$ satisfying

$$\begin{aligned}\tilde{\sigma}(\rho) &= 0 \quad \text{for } 0 \leq \rho \leq r_1, \\ \tilde{\sigma}(\rho) &= \sigma_0 \quad \text{for } \rho \geq r_2, \\ \tilde{\sigma}(\rho) &\text{ increasing for } \rho \in (r_1, r_2).\end{aligned}$$

We define

$$\tilde{\rho} = \rho(1 + i\tilde{\sigma}) \equiv \rho\tilde{d}.$$

One obvious construction of such a function $\tilde{\sigma}$ in the transition layer $r_1 \leq \rho \leq r_2$ with the above properties is given by the fifth order polynomial,

$$\tilde{\sigma}(\rho) = \sigma_0 \left(\int_{r_1}^{\rho} (t - r_1)^2 (r_2 - t)^2 dt \right) \left(\int_{r_1}^{r_2} (t - r_1)^2 (r_2 - t)^2 dt \right)^{-1} \quad \text{for } r_1 \leq \rho \leq r_2.$$

A smoother $\tilde{\sigma}$ can be constructed by increasing the exponents in the above formula.

Each component of the solution \mathbf{E} of (2.1) satisfies the Helmholtz equation with Sommerfeld radiation condition, i.e.,

$$(2.2) \quad \begin{aligned}\Delta u + k^2 u &= 0 \quad \text{for } \rho > r_0, \\ \lim_{\rho \rightarrow \infty} \rho(\nabla u \cdot \hat{\mathbf{x}} - iku) &= 0.\end{aligned}$$

Of course, this also holds for the acoustic problem (1.1). It follows that the solution of (2.1) can be expanded

$$(2.3) \quad \mathbf{E} = \sum_{n=1}^{\infty} \sum_{m=-n}^n \mathbf{a}_{n,m} h_n^1(k\rho) Y_n^m(\theta, \phi) \quad \text{for } \rho \geq r_0.$$

Here $h_n^1(r)$ are spherical Bessel functions of the third kind (Hankel functions), Y_n^m are spherical harmonics (see, e.g., [12] for details) and $\mathbf{a}_{n,m}$ are vector-valued constants. The solution of the acoustic scattering problem satisfies (2.3) as well with \mathbf{E} replaced by u and the vector coefficients $\{\mathbf{a}_{n,m}\}$ replaced by scalar coefficients $\{a_{n,m}\}$.

The PML solution in either case is developed in a similar fashion. We illustrate the development in the case of Maxwell's equations. The (infinite domain) PML solution is defined by

$$\tilde{\mathbf{E}} = \begin{cases} \mathbf{E}(\mathbf{x}) & \text{for } |\mathbf{x}| \leq r_1, \\ \sum_{n=1}^{\infty} \sum_{m=-n}^n \mathbf{a}_{n,m} h_n^1(k\tilde{\rho}) Y_n^m(\theta, \phi) & \text{for } \rho = |\mathbf{x}| \geq r_1. \end{cases}$$

By construction $\tilde{\mathbf{E}}$ and \mathbf{E} coincide on Ω_1 . Furthermore, the complex shift in the argument of h_n^1 above guarantees exponential decay of $\tilde{\mathbf{E}}$.

The PML solution defined above satisfies a differential equation involving $\tilde{\rho}$ and $\frac{d\tilde{\rho}}{d\rho}$. A simple computation shows that

$$\frac{d\tilde{\rho}}{d\rho} = (1 + i\sigma(\rho)) \equiv d$$

where

$$\sigma(\rho) = \tilde{\sigma}(\rho) + \rho\tilde{\sigma}'(\rho).$$

It follows that σ is in $C^1(\mathbb{R}^+)$ and satisfies

$$\begin{aligned}\sigma(\rho) &= 0 \quad \text{for } 0 \leq \rho \leq r_1, \\ \sigma(\rho) &> \tilde{\sigma}(\rho) \quad \text{for } \rho \in (r_1 r_2), \\ \sigma(\rho) &= \sigma_0 \quad \text{for } \rho \geq r_2\end{aligned}$$

The solution $\tilde{\mathbf{E}}$ satisfies Maxwell's equations using the spherical coordinates $(\tilde{\rho}, \theta, \phi)$ [12]. More precisely,

$$\begin{aligned}(2.4) \quad -\tilde{\nabla} \times \tilde{\nabla} \times \tilde{\mathbf{E}} + k^2 \tilde{\mathbf{E}} &= \mathbf{0} \text{ in } \Omega^c, \\ \mathbf{n} \times \tilde{\mathbf{E}} &= \mathbf{n} \times \mathbf{g} \text{ on } \partial\Omega, \\ \tilde{\mathbf{E}} &\text{ bounded at } \infty.\end{aligned}$$

For $\tilde{\mathbf{E}}$ expanded in spherical coordinates,

$$\tilde{\mathbf{E}} = \tilde{\mathbf{E}}_\rho \mathbf{e}_\rho + \tilde{\mathbf{E}}_\theta \mathbf{e}_\theta + \tilde{\mathbf{E}}_\phi \mathbf{e}_\phi,$$

we have

$$\begin{aligned}(2.5) \quad \tilde{\nabla} \times \tilde{\mathbf{E}} &= \frac{1}{\tilde{d}\rho \sin \theta} \left(\frac{\partial}{\partial \theta} (\sin \theta \tilde{\mathbf{E}}_\phi) - \frac{\partial \tilde{\mathbf{E}}_\theta}{\partial \phi} \right) \mathbf{e}_\rho \\ &+ \frac{1}{\rho \tilde{d}} \left(\frac{1}{\sin \theta} \frac{\partial \tilde{\mathbf{E}}_\rho}{\partial \phi} - \frac{1}{d} \frac{\partial}{\partial \rho} (\tilde{d}\rho \tilde{\mathbf{E}}_\phi) \right) \mathbf{e}_\theta \\ &+ \frac{1}{\tilde{d}\rho} \left(\frac{1}{d} \frac{\partial}{\partial \rho} (\tilde{d}\rho \tilde{\mathbf{E}}_\theta) - \frac{\partial \tilde{\mathbf{E}}_\rho}{\partial \theta} \right) \mathbf{e}_\phi.\end{aligned}$$

The PML approximation in the acoustic case is given by

$$\begin{aligned}(2.6) \quad \tilde{\Delta} \tilde{u} + k^2 \tilde{u} &= 0 \text{ in } \Omega^c, \\ \tilde{u} &= g \text{ on } \partial\Omega, \\ \tilde{u} &\text{ bounded at } \infty.\end{aligned}$$

In polar coordinates (ρ, θ, ϕ) ,

$$(2.7) \quad \tilde{\Delta} v = \frac{1}{\tilde{d}^2 d \rho^2} \frac{\partial}{\partial \rho} \left(\frac{\tilde{d}^2 \rho^2}{d} \frac{\partial v}{\partial \rho} \right) + \frac{1}{\tilde{d}^2 \rho^2 \sin \theta} \frac{\partial}{\partial \theta} \left(\sin \theta \frac{\partial v}{\partial \theta} \right) + \frac{1}{\tilde{d}^2 \rho^2 \sin^2 \theta} \frac{\partial^2 v}{\partial \phi^2}.$$

Since the solutions of (2.4) and (2.6) coincide with those of (2.1) and (1.1), respectively, on Ω_1 while rapidly decaying as ρ tends to infinity, it is natural to truncate to a finite computational domain Ω_∞ and impose convenient boundary conditions on the outer boundary of Ω_∞ (which we denote by Γ_∞). We shall always require that the transitional region is contained in Ω_∞ , i.e., $\bar{\Omega}_2 \subset \Omega_\infty$. We introduce the parameter R_t and assume that Ω_∞ contains the sphere of radius R_t . Our analysis will require only a fixed (Lipshitz continuous) outer boundary shape (but one that we enlarge by our dilation parameter R_t). Of course, in practice, it is often convenient to take a polyhedral outer boundary.

It will be critical to keep track of the relation between constants appearing in the inequalities and the scaling parameter R_t . Our constants are independent of R_t and will be denoted generically with the letter C .

The truncated PML approximations are then given as follows. In the case of Maxwell's problem, we consider the truncated PML problem involving a vector function $\tilde{\mathbf{E}}_t$ defined on Ω_∞ and satisfying

$$(2.8) \quad \begin{aligned} -\tilde{\nabla} \times \tilde{\nabla} \times \tilde{\mathbf{E}}_t + k^2 \tilde{\mathbf{E}}_t &= \mathbf{0} \text{ in } \Omega_\infty, \\ \mathbf{n} \times \tilde{\mathbf{E}}_t &= \mathbf{n} \times \mathbf{g} \text{ on } \partial\Omega, \\ \mathbf{n} \times \tilde{\mathbf{E}}_t &= \mathbf{0} \text{ on } \Gamma_\infty. \end{aligned}$$

Analogously, for the acoustics problem, we consider \tilde{u}_t defined on Ω_∞ satisfying

$$(2.9) \quad \begin{aligned} \tilde{\Delta} \tilde{u}_t + k^2 \tilde{u}_t &= 0 \text{ in } \Omega_\infty, \\ \tilde{u}_t &= g \text{ on } \partial\Omega, \\ \tilde{u}_t &= 0 \text{ on } \Gamma_\infty. \end{aligned}$$

Remark 2.1. It is possible to use and analyze other conditions on the outer boundary. We choose Dirichlet conditions for convenience.

3. ANALYSIS OF THE TRUNCATED ACOUSTIC PML (2.9)

In this section, we will prove that the truncated PML acoustic problem (2.9) has a unique weak solution which converges exponentially to the solution of (1.1) near the obstacle. We will first prove uniqueness for (2.9). To do this we will use a duality argument. A similar technique was used in [7] for the exterior Helmholtz problem to estimate the effect of truncating the infinite domain and imposing an approximate absorbing boundary condition.

We first consider a weak formulation of (2.6). Define the sesquilinear form,

$$(3.1) \quad \begin{aligned} b(v, \chi) &= k^2 (\tilde{d}^2 v, \chi)_{\Omega^c} - \left(\frac{\tilde{d}^2}{d} \frac{\partial v}{\partial \rho}, \frac{\partial}{\partial \rho} \left(\frac{\chi}{\tilde{d}} \right) \right)_{\Omega^c} \\ &\quad - \left(\frac{1}{\rho^2} \frac{\partial v}{\partial \theta}, \frac{\partial \chi}{\partial \theta} \right)_{\Omega^c} - \left(\frac{1}{\rho^2 \sin^2 \theta} \frac{\partial v}{\partial \phi}, \frac{\partial \chi}{\partial \phi} \right)_{\Omega^c}. \end{aligned}$$

This form is well defined for $v \in H^1(\Omega^c)$ and $\chi \in H^1(\Omega^c)$ and results from (2.6) and integration by parts. For $g \in H^{1/2}(\partial\Omega)$, let \hat{g} be an $H^1(\Omega^c)$ bounded extension of g supported in Ω_0 . The weak solution of (2.6) is the function $\tilde{u} = \hat{g} - w$ where $w \in H_0^1(\Omega^c)$ satisfies

$$(3.2) \quad b(w, \phi) = b(\hat{g}, \phi) \quad \text{for all } \phi \in H_0^1(\Omega^c).$$

We will subsequently show that the variational problem (3.2) is well posed and that \tilde{u} is well defined and independent of the particular extension \hat{g} .

To employ the duality technique, we need to consider the adjoint source problem on the infinite domain. For $\Phi \in L^2(\Omega^c)$, let $\hat{z} \in H_0^1(\Omega^c)$ satisfy

$$(3.3) \quad b(\chi, \hat{z}) = (\chi, \Phi)_{\Omega^c} \quad \text{for all } \chi \in H_0^1(\Omega^c).$$

It is immediate that $\hat{z} = \bar{d}\bar{z}$ (\bar{z} denotes the complex conjugate of z), where z satisfies

$$(3.4) \quad b(z, \chi) = (\bar{\Phi}/d, \chi)_{\Omega^c} \quad \text{for all } \chi \in H_0^1(\Omega^c).$$

The above problems are well posed as is shown in the following theorem.

Theorem 3.1. *Let Φ be in $L^2(\Omega^c)$. Problems (3.4) and (3.3) have unique solutions $z, \hat{z} \in H_0^1(\Omega^c)$ satisfying*

$$(3.5) \quad \|z\|_{H^1(\Omega^c)} \leq C\|\Phi\|_{L^2(\Omega^c)} \text{ and } \|\hat{z}\|_{H^1(\Omega^c)} \leq C\|\Phi\|_{L^2(\Omega^c)}.$$

To prove the above theorem and subsequent results, we shall require the following theorem which follows easily from a theorem due to Peetre [13] and Tartar [15] (see, e.g. Theorem 2.1 of [6]).

Theorem 3.2. *Let $A_0(\cdot, \cdot), I(\cdot, \cdot)$ be bounded sesquilinear forms on a complex Hilbert space V with norm $\|\cdot\|_V$. Let W be another Hilbert space with V compactly imbedded in W . Suppose that*

$$|I(v, v)| \leq C_1\|v\|_V\|v\|_W \quad \text{for all } v \in V$$

and

$$\|v\|_V^2 \leq C_2|A_0(v, v)| \quad \text{for all } v \in V.$$

Set $A = A_0 + I$ and assume that the only $u \in V$ satisfying

$$A(u, v) = 0 \quad \text{for all } v \in V$$

is $u = 0$. Then, there exists $C_3 > 0$ such that for all $u \in V$,

$$\|u\|_V \leq C_3 \sup_{v \in V} \frac{|A(u, v)|}{\|v\|_V}.$$

Proof of Theorem 3.1. We will use Theorem 3.2 to show that the form (3.1) satisfies an inf-sup condition on $H_0^1(\Omega^c)$. To this end we break the form into two parts as follows:

$$(3.6) \quad b(v, \chi) = b_1(v, \chi) + I(v, \chi)$$

where

$$(3.7)$$

$$b_1(v, \chi) = k^2(d_0^2 u, \chi)_{\Omega^c} - \left(\frac{\tilde{d}^2}{d^2} \frac{\partial v}{\partial \rho}, \frac{\partial \chi}{\partial \rho} \right)_{\Omega^c} - \left(\frac{1}{\rho^2} \frac{\partial v}{\partial \theta}, \frac{\partial \chi}{\partial \theta} \right)_{\Omega^c} - \left(\frac{1}{\rho^2 \sin^2 \theta} \frac{\partial v}{\partial \phi}, \frac{\partial \chi}{\partial \phi} \right)_{\Omega^c},$$

and

$$(3.8) \quad I(v, \chi) = \left(\frac{\tilde{d}^2 d'}{d^3} \frac{\partial v}{\partial \rho}, \chi \right)_{\Omega^c} + k^2((\tilde{d}^2 - d_0^2)v, \chi)_{\Omega^c}.$$

Notice that d' and $(\tilde{d}^2 - d_0^2)$ both vanish for $\rho \geq r_2$. Hence

$$(3.9) \quad |I(v, v)| \leq C\|v\|_{L^2(\Omega^c)}\|v\|_{H^1(\Omega^c)}.$$

Recall that $d_0 = 1 + i\sigma_0$, $\tilde{d} = 1 + i\tilde{\sigma}$, $d = 1 + i\sigma$ and $\sigma \geq \tilde{\sigma}$. It follows easily that there is a positive real number α such that

$$(3.10) \quad \operatorname{Re}[d_0^2(1 + i\alpha)] \leq -C_1 < 0 \text{ and } \operatorname{Re}\left[\frac{\tilde{d}^2}{d^2}(1 + i\alpha)\right] \geq C_2 > 0,$$

for α large enough. In fact, it suffices to choose $\alpha > \max[(1 - \sigma_0^2)/2\sigma_0, \sigma_M]$, where σ_M is the maximum of σ . It follows from (3.7) and (3.10) that

$$(3.11) \quad (1 + \alpha^2)^{1/2}|b_1(v, v)| \geq |\operatorname{Re}[(1 + i\alpha)b_1(v, v)]| \geq C\|v\|_{H^1(\Omega^c)}^2.$$

Now, using the argument in [5] we have the uniqueness property that if $v \in H_0^1(\Omega^c)$ and $b(v, \phi) = 0$ for all $\phi \in C_0^\infty(\Omega^c)$, then $v = 0$. Theorem 3.2 then gives the inf-sup condition

$$(3.12) \quad \|v\|_{H^1(\Omega^c)} \leq C \sup_{\phi \in C_0^\infty(\Omega^c)} \frac{|b(v, \phi)|}{\|\phi\|_{H^1(\Omega^c)}} \quad \text{for all } v \in H_0^1(\Omega^c).$$

The corresponding inf-sup condition for the adjoint problem follows from the identity

$$(3.13) \quad b(\phi, v) = b(\bar{v}/d, \bar{d}\phi).$$

Hence, by the generalized Lax-Milgram Lemma, there exists a unique $z \in H_0^1(\Omega^c)$ satisfying (3.4). Moreover,

$$\|z\|_{H^1(\Omega^c)} \leq C \sup_{\phi \in C_0^\infty(\Omega^c)} \frac{|b(z, \phi)|}{\|\phi\|_{H^1(\Omega^c)}} \leq C \|\Phi\|_{L^2(\Omega^c)}.$$

This completes the proof of the theorem. \square

Remark 3.1. Applying a standard trace estimate, the proof of the above theorem implies that the solution \tilde{u} of (2.6) satisfies

$$\|\tilde{u}\|_{H^1(\Omega^c)} \leq C\|\hat{g}\|_{H^1(\Omega_0)} \leq C\|g\|_{H^{1/2}(\partial\Omega)}.$$

In addition, the inf-sup condition proved above immediately implies that \tilde{u} is independent of the choice of extension \hat{g} .

We will first prove uniqueness for (2.9). In order to do this we will need the following two propositions. These propositions will be used extensively throughout the remainder of this paper. The first is a classical interior estimate for the solution of an elliptic equation. The proof is elementary.

Proposition 3.1. *Suppose that w satisfies the Helmholtz equation*

$$(3.14) \quad \Delta w + \beta w = 0$$

in a domain D with a (possibly complex) constant β . If D_1 is a subdomain, whose closure is contained in D , then

$$(3.15) \quad \|w\|_{H^2(D_1)} \leq C\|w\|_{L^2(D)}.$$

We also need the following proposition.

Proposition 3.2. *Assume that w is bounded at infinity and satisfies (3.14) in $\Omega^c \setminus \bar{\Omega}_2$ with $\beta = k^2 d_0^2$. Set $S_\gamma = \{\mathbf{x} : \text{dist}(\mathbf{x}, \Gamma_\infty) < \gamma\}$ with γ fixed independent of $R_t > r_4$ and small enough such that \bar{S}_γ is in $\Omega^c \setminus \bar{\Omega}_4$. Then*

$$\|w\|_{L^2(S_\gamma)} \leq C e^{-\sigma_0 k R_t} \|w\|_{L^2(\Omega_4)}.$$

Proof. The fundamental solution of (3.14) with $\beta = k^2 d_0^2$ is

$$\psi(\mathbf{x}, \mathbf{y}) = -\frac{\exp(ikd_0|\mathbf{x} - \mathbf{y}|)}{4\pi|\mathbf{x} - \mathbf{y}|}.$$

For any point \mathbf{x} in S_γ

$$(3.16) \quad w(\mathbf{x}) = \int_{\Gamma_3} w(\mathbf{y}) \frac{\partial \psi(\mathbf{x}, \mathbf{y})}{\partial r_y} dS_y - \int_{\Gamma_3} \frac{\partial w(\mathbf{y})}{\partial r_y} \psi(\mathbf{x}, \mathbf{y}) dS_y.$$

Note that there is no contribution above from infinity. Indeed, since w is bounded at infinity, it can be written as a (scalar) expansion of the form of (2.3) with k replaced by $d_0 k$. In addition, ψ decays rapidly at infinity since d_0 has a positive imaginary part and so the outer boundary contribution limits to zero.

Using Schwarz's inequality and the properties of ψ it is easy to see that

$$(3.17) \quad |w(\mathbf{x})|^2 \leq C e^{-2\sigma_0 k R_t} \left(\|w\|_{L^2(\Gamma_3)}^2 + \left\| \frac{\partial w}{\partial r} \right\|_{L^2(\Gamma_3)}^2 \right) \left(\int_{\Gamma_3} \frac{dS_y}{|\mathbf{x} - \mathbf{y}|^2} \right).$$

Integrating over S_γ , using a standard trace inequality and Proposition 3.1 we obtain Proposition 3.2. \square

We can now prove the following theorem.

Theorem 3.3. *Let u be in $H_0^1(\Omega_\infty)$ and satisfy (2.9) with $g = 0$. Then, for R_t large enough, $u = 0$. That is to say, if $u \in H_0^1(\Omega_\infty)$ satisfies $b(u, \psi) = 0$ for all $\psi \in H_0^1(\Omega_\infty)$, then $u = 0$.*

Proof. Let u be in $H_0^1(\Omega_\infty)$ satisfy (2.9) with $g = 0$, let Φ be in $L^2(\Omega^c)$ with support in Ω_4 , and let $\hat{z} \in H_0^1(\Omega^c)$ be the solution of (3.3). Then

$$(3.18) \quad (u, \Phi)_{\Omega_4} = b(u, \hat{z}) = \left\langle \frac{\partial u}{\partial \mathbf{n}}, \hat{z} \right\rangle_{\Gamma_\infty}.$$

Let $\tilde{H}^1(\Omega_\infty \setminus \bar{\Omega}_3)$ denote the set of functions in $H^1(\Omega_\infty \setminus \bar{\Omega}_3)$ which vanish on Γ_3 . Define the norm

$$\|w\|_{H^{-1/2}(\Gamma_\infty)} = \sup_{\phi \in \tilde{H}^1(\Omega_\infty \setminus \bar{\Omega}_3)} \frac{|\langle w, \phi \rangle_{\Gamma_\infty}|}{\|\phi\|_{H^1(\Omega_\infty \setminus \bar{\Omega}_3)}}.$$

Let χ be a smooth cutoff function with support \bar{D}_1 in S_γ (of Proposition 3.2), which is one on Γ_∞ . Applying Propositions 3.1 and 3.2 to z and (3.5), it follows that

$$(3.19) \quad \begin{aligned} |(u, \Phi)_{\Omega_4}| &\leq C \frac{|\langle \frac{\partial u}{\partial \mathbf{n}}, \chi \hat{z} \rangle_{\Gamma_\infty}|}{\|\chi \hat{z}\|_{H^1(S_\gamma)}} \|\hat{z}\|_{H^1(D_1)} \\ &\leq C \frac{|\langle \frac{\partial u}{\partial \mathbf{n}}, \chi \hat{z} \rangle_{\Gamma_\infty}|}{\|\chi \hat{z}\|_{H^1(S_\gamma)}} \|\hat{z}\|_{L^2(S_\gamma)} \\ &\leq C e^{-\sigma_0 k R_t} \|\Phi\|_{L^2(\Omega_4)} \left\| \frac{\partial u}{\partial \mathbf{n}} \right\|_{H^{-1/2}(\Gamma_\infty)}. \end{aligned}$$

We next estimate the negative norm on the right hand side above. Let \hat{h} be in $\tilde{H}^1(\Omega_\infty \setminus \bar{\Omega}_2)$ and be equal to zero in $\Omega_3 \setminus \bar{\Omega}_2$. Let $\psi \in H_0^1(\Omega_\infty \setminus \bar{\Omega}_2)$ satisfy

$$(\nabla \psi, \nabla \theta)_{\Omega_\infty \setminus \bar{\Omega}_2} - k^2 d_0^2(\psi, \theta)_{\Omega_\infty \setminus \bar{\Omega}_2} = (\nabla \hat{h}, \nabla \theta)_{\Omega_\infty \setminus \bar{\Omega}_2} - k^2 d_0^2(\hat{h}, \theta)_{\Omega_\infty \setminus \bar{\Omega}_2}$$

for all $\theta \in H_0^1(\Omega_\infty \setminus \bar{\Omega}_2)$. This problem is well posed since d_0^2 has a nonzero imaginary part. It follows that

$$\|\psi\|_{H^1(\Omega_\infty \setminus \bar{\Omega}_2)} \leq C \|\hat{h}\|_{H^1(\Omega_\infty \setminus \bar{\Omega}_2)}.$$

We set $h = \hat{h} - \psi$. Note that both u and h satisfy homogeneous equations in $\Omega_\infty \setminus \bar{\Omega}_2$, i.e.,

$$(3.20) \quad \Delta u + k^2 d_0^2 u = 0, \quad \Delta h + k^2 d_0^2 h = 0.$$

Now, using Green's identity,

$$(3.21) \quad \langle \frac{\partial u}{\partial \mathbf{n}}, \hat{h} \rangle_{\Gamma_\infty} = \langle \frac{\partial u}{\partial \mathbf{n}}, h \rangle_{\Gamma_\infty} = -\langle \frac{\partial u}{\partial \mathbf{n}}, h \rangle_{\Gamma_3} + \langle u, \frac{\partial h}{\partial \mathbf{n}} \rangle_{\Gamma_3}.$$

Finally, using Proposition 3.1 (with D_1 a domain containing Γ_3 whose closure is in $\Omega_4 \setminus \bar{\Omega}_2$), we get

$$(3.22) \quad \begin{aligned} |\langle \frac{\partial u}{\partial \mathbf{n}}, h \rangle_{\Gamma_3} - \langle u, \frac{\partial h}{\partial \mathbf{n}} \rangle_{\Gamma_3}| &\leq C \|u\|_{H^2(D_1)} \|h\|_{H^2(D_1)} \\ &\leq C \|u\|_{L^2(\Omega_4)} \|h\|_{L^2(\Omega_4)} \\ &\leq C \|u\|_{L^2(\Omega_4)} \|\hat{h}\|_{H^1(\Omega_\infty \setminus \bar{\Omega}_3)}. \end{aligned}$$

Combining the above results shows that

$$\left\| \frac{\partial u}{\partial \mathbf{n}} \right\|_{H^{-1/2}(\Gamma_\infty)} \leq C \|u\|_{L^2(\Omega_4)}$$

and hence, using (3.19), we have

$$(3.23) \quad \|u\|_{L^2(\Omega_4)} \leq C e^{-\sigma_0 k R_t} \|u\|_{L^2(\Omega_4)},$$

i.e., u vanishes on Ω_4 provided R_t is taken large enough. It follows by unique continuation that u vanishes on all of Ω_∞ . This completes the proof of the uniqueness theorem. \square

We next give a weak form of problem (2.9). For $g \in H^{1/2}(\partial\Omega)$ let \hat{g} be an $H^1(\Omega^c)$ bounded extension of g with support in Ω_0 . The weak solution of (2.9) is the function $\tilde{u}_t = \hat{g} - w$ where $w \in H_0^1(\Omega_\infty)$ satisfies

$$(3.24) \quad b(w, \phi) = b(\hat{g}, \phi) \quad \text{for all } \phi \in H_0^1(\Omega_\infty).$$

The next theorem shows existence and gives error estimates for the weak solution.

Theorem 3.4. *The variational problem (3.24), for R_t sufficiently large, has a unique solution. The resulting weak solution \tilde{u}_t of (2.9) is well defined and independent of the extension \hat{g} . Finally,*

$$\|\tilde{u} - \tilde{u}_t\|_{L^2(\Omega_4)} \leq C e^{-2\sigma_0 k R_t} \|g\|_{H^{1/2}(\partial\Omega)}.$$

Here \tilde{u} is the solution of (2.6).

Remark 3.2. The above theorem shows that \tilde{u}_t converges exponentially on Ω_4 to \tilde{u} as $R_t \rightarrow \infty$. It follows that \tilde{u}_t converges exponentially to u on Ω_1 .

Proof. We note that (3.9) and (3.11) hold on the restricted space $H_0^1(\Omega_\infty)$ so the uniqueness result of the previous theorem and Theorem 3.2 implies the inf-sup condition

$$(3.25) \quad \|v\|_{H^1(\Omega_\infty)} \leq C \sup_{\phi \in C_0^\infty(\Omega_\infty)} \frac{|b(v, \phi)|}{\|\phi\|_{H^1(\Omega_\infty)}} \quad \text{for all } v \in H_0^1(\Omega_\infty).$$

Uniqueness for the adjoint problem on $H_0^1(\Omega_\infty)$ follows from Theorem 3.3 and (3.13). This implies the existence and uniqueness of solutions to (3.24). It is easy to see that the resulting function \tilde{u}_t is independent of extension \hat{g} .

To finish the proof, we need to show that \tilde{u}_t converges to \tilde{u} , the solution of (2.6), in $L^2(\Omega_4)$ and that the convergence is exponential as R_t increases beyond some threshold. To see this set $\tilde{e} = \tilde{u} - \tilde{u}_t$. As in the proof of Theorem 3.3, let Φ be in $L^2(\Omega^c)$ with support in Ω_4 and let $\hat{z} \in H_0^1(\Omega^c)$ be the solution of (3.3). Let \mathcal{L}

denote the formal adjoint of the operator $\tilde{d}^2\tilde{\Delta}$. Since $\tilde{\Delta}$ is a multiple of Δ except on the transition layer $r_1 < \rho < r_2$, it follows that

$$(3.26) \quad \begin{aligned} (\tilde{e}, \Phi)_{\Omega_4} &= -(\tilde{e}, \mathcal{L}\hat{z})_{\Omega_\infty} = -\langle \tilde{u}, \frac{\partial \hat{z}}{\partial \mathbf{n}} \rangle_{\Gamma_\infty} + b_\infty(\tilde{e}, \hat{z}) \\ &= \langle \frac{\partial \tilde{e}}{\partial \mathbf{n}}, \hat{z} \rangle_{\Gamma_\infty} - \langle \tilde{u}, \frac{\partial \hat{z}}{\partial \mathbf{n}} \rangle_{\Gamma_\infty}. \end{aligned}$$

Here $b_\infty(\cdot, \cdot)$ denotes the form on $H^1(\Omega_\infty) \times H^1(\Omega_\infty)$, which results from replacing the domain of integration Ω^c in (3.1) by Ω_∞ . To handle the first term on the right hand side of (3.26), we shall use estimates in the proof of Theorem 3.3 (with u replaced by \tilde{e}). As in (3.19),

$$|\langle \frac{\partial \tilde{e}}{\partial \mathbf{n}}, \hat{z} \rangle_{\Gamma_\infty}| \leq C e^{-\sigma_0 k R_t} \|\Phi\|_{L^2(\Omega_4)} \left\| \frac{\partial \tilde{e}}{\partial \mathbf{n}} \right\|_{H^{-1/2}(\Gamma_\infty)}.$$

We estimate the negative norm again following the proof of Theorem 3.3, but replace (3.21) with

$$\begin{aligned} \langle \frac{\partial \tilde{e}}{\partial \mathbf{n}}, \hat{h} \rangle_{\Gamma_\infty} &= \langle \frac{\partial \tilde{e}}{\partial \mathbf{n}}, h \rangle_{\Gamma_\infty} \\ &= -\langle \frac{\partial \tilde{e}}{\partial \mathbf{n}}, h \rangle_{\Gamma_3} + \langle \tilde{e}, \frac{\partial h}{\partial \mathbf{n}} \rangle_{\Gamma_3} + \langle \tilde{u}, \frac{\partial h}{\partial \mathbf{n}} \rangle_{\Gamma_\infty}. \end{aligned}$$

The first two terms on the right hand side above are estimated exactly as in (3.22). For the last one, we note that because h satisfies (3.20),

$$\left\| \frac{\partial h}{\partial \mathbf{n}} \right\|_{H^{-1/2}(\Gamma_\infty)} \leq C \|h\|_{H^1(\Omega_\infty \setminus \bar{\Omega}_2)}.$$

Thus,

$$|\langle \tilde{u}, \frac{\partial h}{\partial \mathbf{n}} \rangle_{\Gamma_\infty}| \leq C \|\tilde{u}\|_{H^1(D_1)} \|\hat{h}\|_{H^1(\Omega_\infty \setminus \bar{\Omega}_2)}$$

where $\bar{D}_1 \subset S_\gamma$. Applying Propositions 3.1 and 3.2 gives

$$|\langle \tilde{u}, \frac{\partial h}{\partial \mathbf{n}} \rangle_{\Gamma_\infty}| \leq C e^{-\sigma_0 k R_t} \|\tilde{u}\|_{L^2(\Omega_4)} \|\hat{h}\|_{H^1(\Omega_\infty \setminus \bar{\Omega}_2)}.$$

Combining the above gives

$$(3.27) \quad \begin{aligned} |(\tilde{e}, \Phi)_{\Omega_4}| &\leq C e^{-\sigma_0 k R_t} \|\Phi\|_{L^2(\Omega_4)} (\|\tilde{e}\|_{L^2(\Omega_4)} \\ &\quad + e^{-\sigma_0 k R_t} \|\tilde{u}\|_{L^2(\Omega_4)}) + |\langle \tilde{u}, \frac{\partial \hat{z}}{\partial \mathbf{n}} \rangle_{\Gamma_\infty}|. \end{aligned}$$

Now, using a standard trace inequality, Theorem 3.1, Proposition 3.1, and Proposition 3.2, we obtain

$$(3.28) \quad |\langle \tilde{u}, \frac{\partial \hat{z}}{\partial \mathbf{n}} \rangle_{\Gamma_\infty}| \leq C e^{-2\sigma_0 k R_t} \|\tilde{u}\|_{L^2(\Omega_4)} \|\Phi\|_{L^2(\Omega_4)}.$$

Thus we have

$$|(\tilde{e}, \Phi)_{\Omega_4}| \leq C(e^{-\sigma_0 k R_t} \|\tilde{e}\|_{L^2(\Omega_4)} + e^{-2\sigma_0 k R_t} \|\tilde{u}\|_{L^2(\Omega_4)}) \|\Phi\|_{L^2(\Omega_4)}.$$

From this and Remark 3.1, it follows that

$$\|\tilde{e}\|_{L^2(\Omega_4)} \leq C e^{-\sigma_0 k R_t} \|\tilde{e}\|_{L^2(\Omega_4)} + C e^{-2\sigma_0 k R_t} \|g\|_{H^{1/2}(\partial\Omega)}.$$

Hence, for R_t large enough, we obtain the convergence estimate

$$\|\tilde{u} - \tilde{u}_t\|_{L^2(\Omega_4)} \leq C e^{-2\sigma_0 k R_t} \|g\|_{H^{1/2}(\partial\Omega)}.$$

This completes the proof of the theorem. \square

4. UNIQUENESS FOR THE TRUNCATED ELECTROMAGNETIC PML PROBLEM

Following [12], we define the diagonal matrices (in spherical coordinates),

$$\mathbf{A}\mathbf{v} = \tilde{d}^{-2}v_\rho\mathbf{e}_\rho + (\tilde{d}d)^{-1}(v_\theta\mathbf{e}_\theta + v_\phi\mathbf{e}_\phi)$$

and

$$\mathbf{B}\mathbf{v} = dv_\rho\mathbf{e}_\rho + \tilde{d}(v_\theta\mathbf{e}_\theta + v_\phi\mathbf{e}_\phi).$$

Then, $\tilde{\nabla} \times \tilde{\mathbf{E}} = \mathbf{A} \nabla \times (\mathbf{B}\tilde{\mathbf{E}})$.

We first define a weak form of the PML problem (2.8) by setting $\tilde{\mathbf{E}}_t = \hat{\mathbf{g}} - \mathbf{w}$ and setting up a variational problem for $\boldsymbol{\Xi} = \mathbf{B}\mathbf{w} \in \mathbf{H}_0(\mathbf{curl}; \Omega_\infty)$, i.e.,

$$(4.1) \quad \mathcal{A}(\boldsymbol{\Xi}, \Psi) = \mathcal{A}(\mathbf{B}\hat{\mathbf{g}}, \Psi) \quad \text{for all } \Psi \in \mathbf{H}_0(\mathbf{curl}; \Omega_\infty).$$

Here the sesquilinear form \mathcal{A} is given by

$$\mathcal{A}(\Theta, \Psi) \equiv -(\mu^{-1} \nabla \times \Theta, \nabla \times \Psi)_{\Omega^c} + k^2(\mu\Theta, \Psi)_{\Omega^c} \quad \text{for all } \Theta, \Psi \in \mathbf{H}(\mathbf{curl}; \Omega^c)$$

and μ is the three by three matrix which corresponds to the diagonal matrix $(\mathbf{AB})^{-1}$ in spherical coordinates. As usual, we define the form on the larger space and consider the space $\mathbf{H}_0(\mathbf{curl}; \Omega_\infty)$ as the subspace of $\mathbf{H}_0(\mathbf{curl}; \Omega^c)$ defined by extension by zero.

Our first task is to show uniqueness when R_t is sufficiently large. That is, if Θ is in $\mathbf{H}_0(\mathbf{curl}; \Omega_\infty)$ and

$$(4.2) \quad \mathcal{A}(\Theta, \Psi) = 0 \quad \text{for all } \Psi \in \mathbf{H}_0(\mathbf{curl}; \Omega_\infty),$$

then $\Theta = \mathbf{0}$.

As was done in the analysis of the acoustic problem, we will again use a duality argument. We consider the adjoint source problem: For $\Phi \in \mathbf{L}^2(\Omega^c)$, find $\hat{\mathbf{z}} \in \mathbf{H}_0(\mathbf{curl}; \Omega^c)$ satisfying

$$(4.3) \quad \mathcal{A}(\Theta, \hat{\mathbf{z}}) = (\Theta, \Phi)_{\Omega^c} \quad \text{for all } \Theta \in \mathbf{H}_0(\mathbf{curl}; \Omega^c).$$

It is immediate that $\hat{\mathbf{z}} = \bar{\mathbf{z}}$, where \mathbf{z} is the solution of

$$(4.4) \quad \mathcal{A}(\mathbf{z}, \Theta) = (\bar{\Phi}, \Theta)_{\Omega^c} \quad \text{for all } \Theta \in \mathbf{H}_0(\mathbf{curl}; \Omega^c).$$

We need the following theorem.

Theorem 4.1. *Let Φ be in $\mathbf{L}^2(\Omega^c)$. Problems (4.4) and (4.3) have unique solutions $\mathbf{z}, \hat{\mathbf{z}} \in \mathbf{H}_0(\mathbf{curl}; \Omega^c)$ satisfying*

$$(4.5) \quad \|\mathbf{z}\|_{\mathbf{H}(\mathbf{curl}; \Omega^c)} = \|\hat{\mathbf{z}}\|_{\mathbf{H}(\mathbf{curl}; \Omega^c)} \leq C\|\Phi\|_{\mathbf{L}^2(\Omega^c)}.$$

For the proof of this theorem, we require the following lemma whose proof appears in the appendix.

Lemma 4.1. *Let D be either Ω_∞ or Ω^c , respectively. Set*

$$\mathbf{X}(D) = \mathbf{H}_0(\mathbf{curl}; D) \cap \mathbf{H}^0(\mathbf{div}; \mu, D),$$

where $\mathbf{H}^0(\mathbf{div}; \mu, D) = \{\mathbf{U} \in \mathbf{L}^2(D) : \nabla \cdot (\mu \mathbf{U}) = 0\}$. Then $\mathbf{X}(D) \subset \mathbf{H}^s(\omega)$ for some $s > 1/2$ where $\omega = \Omega_\infty$ or $\omega = \Omega_2$, respectively.

Proof of Theorem 4.1. We will prove that, for $\mathbf{W} \in \mathbf{X}(\Omega^c)$,

$$(4.6) \quad \|\mathbf{W}\|_{\mathbf{H}(\mathbf{curl}; \Omega^c)} \leq C \sup_{\mathbf{V} \in \mathbf{X}(\Omega^c)} \frac{|\mathcal{A}(\mathbf{W}, \mathbf{V})|}{\|\mathbf{V}\|_{\mathbf{H}(\mathbf{curl}; \Omega^c)}}.$$

To this end, set $\mathcal{B}(\mathbf{U}, \mathbf{V}) = \mathcal{A}(\mathbf{U}, \bar{\eta}\mathbf{V})$, where $\eta = \tilde{d}^2/d$. Then

$$\begin{aligned} \mathcal{B}(\mathbf{U}, \mathbf{V}) &= -(\boldsymbol{\mu}^{-1} \nabla \times \mathbf{U}, \nabla \times \bar{\eta}\mathbf{V})_{\Omega^c} + k^2(\eta\boldsymbol{\mu}\mathbf{U}, \mathbf{V})_{\Omega^c} \\ (4.7) \quad &= -(\eta\boldsymbol{\mu}^{-1} \nabla \times \mathbf{U}, \nabla \times \mathbf{V})_{\Omega^c} - (\boldsymbol{\mu}^{-1} \nabla \times \mathbf{U}, (\nabla \bar{\eta}) \times \mathbf{V})_{\Omega^c} \\ &\quad + k^2(\eta\boldsymbol{\mu}\mathbf{U}, \mathbf{V})_{\Omega^c} \\ &= \mathcal{B}_1(\mathbf{U}, \mathbf{V}) + \mathcal{I}(\mathbf{U}, \mathbf{V}). \end{aligned}$$

Here

$$\mathcal{B}_1(\mathbf{U}, \mathbf{V}) = -(\eta\boldsymbol{\mu}^{-1} \nabla \times \mathbf{U}, \nabla \times \mathbf{V})_{\Omega^c} + k^2(d_0^2 \mathbf{U}, \mathbf{V})_{\Omega^c}$$

and

$$\mathcal{I}(\mathbf{U}, \mathbf{V}) = -(\boldsymbol{\mu}^{-1} \nabla \times \mathbf{U}, (\nabla \bar{\eta}) \times \mathbf{V})_{\Omega_2} - k^2((d_0^2 \mathbf{I} - \eta\boldsymbol{\mu})\mathbf{U}, \mathbf{V})_{\Omega_2}.$$

The last two integrations are over Ω_2 since both $\nabla \bar{\eta}$ and $(d_0^2 \mathbf{I} - \eta\boldsymbol{\mu})$ vanish for $\rho > r_2$. We obviously have

$$(4.8) \quad |\mathcal{I}(\mathbf{V}, \mathbf{V})| \leq C\|\mathbf{V}\|_{\mathbf{H}(\mathbf{curl}; \Omega_2)}\|\mathbf{V}\|_{L^2(\Omega_2)}.$$

Choosing α as in (3.10) we obtain for $\mathbf{V} \in \mathbf{H}_0(\mathbf{curl}; \Omega^c)$,

$$(4.9) \quad (1 + \alpha^2)^{1/2}|\mathcal{B}_1(\mathbf{V}, \mathbf{V})| \geq |Re[(1 + i\alpha)\mathcal{B}_1(\mathbf{V}, \mathbf{V})]| \geq C\|\mathbf{V}\|_{\mathbf{H}(\mathbf{curl}; \Omega^c)}^2.$$

Multiplication by $\bar{\eta}$ and $\bar{\eta}^{-1}$ are bounded operators on $\mathbf{H}(\mathbf{curl}; \Omega^c)$ and $\mathbf{H}^s(\Omega_2)$. Set

$$\mathbf{X}^\eta(\Omega^c) = \{\bar{\eta}^{-1}\boldsymbol{\theta} : \boldsymbol{\theta} \in \mathbf{X}(\Omega^c)\}.$$

If follows from Lemma 4.1 that $\mathbf{X}^\eta(\Omega^c)$ is compactly contained in $L^2(\Omega_2)$. Moreover,

$$\sup_{\mathbf{V} \in \mathbf{X}(\Omega^c)} \frac{|\mathcal{A}(\mathbf{W}, \mathbf{V})|}{\|\mathbf{V}\|_{\mathbf{H}(\mathbf{curl}; \Omega^c)}} = \sup_{\mathbf{V} \in \mathbf{X}^\eta(\Omega^c)} \frac{|\mathcal{B}(\mathbf{W}, \mathbf{V})|}{\|\bar{\eta}\mathbf{V}\|_{\mathbf{H}(\mathbf{curl}; \Omega^c)}} \geq C \sup_{\mathbf{V} \in \mathbf{X}^\eta(\Omega^c)} \frac{|\mathcal{B}(\mathbf{W}, \mathbf{V})|}{\|\mathbf{V}\|_{\mathbf{H}(\mathbf{curl}; \Omega^c)}}.$$

Thus, (4.6) will follow if we show that for $\mathbf{W} \in \mathbf{X}^\eta(\Omega^c)$,

$$(4.10) \quad \|\mathbf{W}\|_{\mathbf{H}(\mathbf{curl}; \Omega^c)} \leq C \sup_{\mathbf{V} \in \mathbf{X}^\eta(\Omega^c)} \frac{|\mathcal{B}(\mathbf{W}, \mathbf{V})|}{\|\mathbf{V}\|_{\mathbf{H}(\mathbf{curl}; \Omega^c)}}.$$

We apply Theorem 3.2 to prove (4.10). By the above discussion, we need only check the uniqueness property. It is immediate that this is equivalent to: If $\mathbf{w} \in \mathbf{X}(\Omega^c)$ satisfies

$$\mathcal{A}(\mathbf{w}, \mathbf{v}) = 0 \quad \text{for all } \mathbf{v} \in \mathbf{X}(\Omega^c),$$

then $\mathbf{w} = \mathbf{0}$.

We note that we have the uniqueness properties for \mathcal{A} and its adjoint in $\mathbf{H}_0(\mathbf{curl}; \Omega^c)$ (cf. [12]), specifically, if $\mathbf{u} \in \mathbf{H}_0(\mathbf{curl}; \Omega^c)$ satisfies

$$(4.11) \quad \mathcal{A}(\mathbf{u}, \mathbf{w}) = 0 \quad \text{for all } \mathbf{w} \in \mathbf{H}_0(\mathbf{curl}; \Omega^c),$$

then $\mathbf{u} = \mathbf{0}$, and if $\mathbf{u} \in \mathbf{H}_0(\mathbf{curl}; \Omega^c)$ satisfies

$$(4.12) \quad \mathcal{A}(\mathbf{w}, \mathbf{u}) = 0 \quad \text{for all } \mathbf{w} \in \mathbf{H}_0(\mathbf{curl}; \Omega^c),$$

then $\mathbf{u} = \mathbf{0}$.

We show the above uniqueness property on the restricted space $\mathbf{X}(\Omega^c)$. Suppose that \mathbf{W} is in $\mathbf{X}(\Omega^c)$ and satisfies

$$(4.13) \quad \mathcal{A}(\mathbf{W}, \mathbf{v}) = 0 \quad \text{for all } \mathbf{v} \in \mathbf{X}(\Omega^c).$$

Let $\tilde{H}_0^1(\Omega^c)$ denote the completion $C_0^\infty(\Omega^c)$ in the norm $\|\nabla u\|_{L^2(\Omega^c)}$. For any $\mathbf{V} \in \mathbf{H}_0(\mathbf{curl}; \Omega^c)$, we can decompose \mathbf{V} as $\mathbf{V} = \mathbf{v} + \nabla\psi$ with $\mathbf{v} \in \mathbf{X}(\Omega^c)$ and $\psi \in \tilde{H}_0^1(\Omega^c)$. Indeed, ψ is the solution of

$$(\boldsymbol{\mu}\nabla\psi, \nabla\theta) = (\boldsymbol{\mu}\mathbf{V}, \nabla\theta) \quad \text{for all } \theta \in \tilde{H}_0^1(\Omega^c).$$

This problem is uniquely solvable since the real part of $\boldsymbol{\mu}$ is uniformly positive definite. Since $\mathbf{W} \in \mathbf{X}(\Omega^c)$, $\mathcal{A}(\mathbf{W}, \mathbf{V}) = \mathcal{A}(\mathbf{W}, \mathbf{v}) = 0$ by (4.13) and $\mathbf{W} = \mathbf{0}$ follows from (4.11). The inequality (4.10) and hence (4.6) follows from Theorem 3.2.

We can now complete the proof of the theorem. For $\Phi \in L^2(\Omega^c)$, define $\phi \in \tilde{H}_0^1(\Omega^c)$ by

$$\mathcal{A}(\nabla\phi, \nabla\psi) = k^2(\boldsymbol{\mu}\nabla\phi, \nabla\psi) = (\bar{\Phi}, \nabla\psi) \quad \text{for all } \psi \in \tilde{H}_0^1(\Omega^c).$$

Next define $\mathbf{W} \in \mathbf{X}(\Omega^c)$ to be the solution of

$$(4.14) \quad \mathcal{A}(\mathbf{W}, \mathbf{v}) = (\bar{\Phi}, \mathbf{v}) - k^2(\boldsymbol{\mu}\nabla\phi, \mathbf{v}) \quad \text{for all } \mathbf{v} \in \mathbf{X}(\Omega^c).$$

As above, the uniqueness property for the adjoint on $\mathbf{H}_0(\mathbf{curl}; \Omega^c)$ implies (4.12) on the restricted space $\mathbf{X}(\Omega^c)$. Thus, the existence of \mathbf{W} satisfying (4.14) follows from (4.6) and the generalized Lax-Milgram Lemma.

The solution of (4.4) is then given by $\mathbf{z} = \mathbf{W} + \nabla\phi$. Indeed, for any $\Theta \in \mathbf{H}_0(\mathbf{curl}; \Omega^c)$, we decompose $\Theta = \mathbf{v} + \nabla\psi$ with $\mathbf{v} \in \mathbf{X}(\Omega^c)$. Then

$$\begin{aligned} \mathcal{A}(\mathbf{z}, \Theta) &= \mathcal{A}(\mathbf{W}, \mathbf{v}) + \mathcal{A}(\mathbf{W}, \nabla\psi) + \mathcal{A}(\nabla\phi, \mathbf{v}) + \mathcal{A}(\nabla\phi, \nabla\psi) \\ &= (\bar{\Phi}, \mathbf{v}) - k^2(\boldsymbol{\mu}\nabla\phi, \mathbf{v}) + k^2(\boldsymbol{\mu}\nabla\phi, \mathbf{v}) + (\bar{\Phi}, \nabla\psi) = (\bar{\Phi}, \Theta). \end{aligned}$$

Evidently,

$$\|\mathbf{z}\|_{\mathbf{H}(\mathbf{curl}; \Omega^c)} \leq C\|\Phi\|_{L^2(\Omega^c)},$$

which concludes the proof. \square

Corollary 4.1. *Let $\tilde{\mathbf{E}}$ be the solution of (2.4) and $\hat{\mathbf{g}}$ an H -curl extension of \mathbf{g} with support in Ω_0 . Then*

$$\|\tilde{\mathbf{E}}\|_{\mathbf{H}(\mathbf{curl}; \Omega^c)} \leq C\|\hat{\mathbf{g}}\|_{\mathbf{H}(\mathbf{curl}; \Omega_0)}.$$

We can now prove the uniqueness theorem for the problem on Ω_∞ .

Theorem 4.2. *For R_t sufficiently large, the only solution $\Theta \in \mathbf{H}_0(\mathbf{curl}; \Omega_\infty)$ satisfying (4.2) is $\Theta = \mathbf{0}$.*

Proof. Suppose that Θ satisfies (4.2). For $\Phi \in L^2(\Omega^c)$ with support in Ω_4 let $\hat{\mathbf{z}}$ satisfy (4.3). Both $\mathbf{n} \times \Theta$ and $\mathbf{n} \times \hat{\mathbf{z}}$ vanish on $\partial\Omega$. Also $\mathbf{n} \times \Theta$ vanishes on Γ_∞ . In addition, the components of $\hat{\mathbf{z}}$ satisfy (3.14) with $\beta = k^2d_0^2$ outside of Ω_2 so $\hat{\mathbf{z}}$ is in \mathbf{H}^2 near Γ_∞ . Thus,

$$(4.15) \quad (\Theta, \Phi)_{\Omega_4} = \mathcal{A}(\Theta, \hat{\mathbf{z}}) = d_0^{-1}\langle \mathbf{n} \times \nabla \times \Theta, \hat{\mathbf{z}} \rangle_{\Gamma_\infty}.$$

Let $\widetilde{\mathbf{H}}^1(\Omega_\infty \setminus \bar{\Omega}_3)$ denote the set of functions in $\mathbf{H}^1(\Omega_\infty \setminus \bar{\Omega}_3)$ which vanish on Γ_3 . Set

$$\|\mathbf{w}\|_{\mathbf{H}^{-1/2}(\Gamma_\infty)} = \sup_{\phi \in \widetilde{\mathbf{H}}^1(\Omega_\infty \setminus \bar{\Omega}_3)} \frac{|\langle \mathbf{w}, \phi \rangle_{\Gamma_\infty}|}{\|\phi\|_{\mathbf{H}^1(\Omega_\infty \setminus \bar{\Omega}_3)}}.$$

Let χ be a smooth cutoff function with support \bar{D}_1 in S_γ (of Proposition 3.2), which is one on Γ_∞ . Since each component of $\hat{\mathbf{z}}$ satisfies (3.14) with $\beta = k^2 d_0^2$ and applying Propositions 3.1 and 3.2 and Theorem 4.1, it follows that

$$\begin{aligned} |(\Theta, \Phi)_{\Omega_4}| &\leq C \frac{|\langle \mathbf{n} \times \nabla \times \Theta, \chi \hat{\mathbf{z}} \rangle_{\Gamma_\infty}|}{\|\chi \hat{\mathbf{z}}\|_{\mathbf{H}^1(S_\gamma)}} \|\hat{\mathbf{z}}\|_{\mathbf{H}^1(D_1)} \\ (4.16) \quad &\leq C \frac{|\langle \mathbf{n} \times \nabla \times \Theta, \chi \hat{\mathbf{z}} \rangle_{\Gamma_\infty}|}{\|\chi \hat{\mathbf{z}}\|_{\mathbf{H}^1(S_\gamma)}} \|\hat{\mathbf{z}}\|_{L^2(S_\gamma)} \\ &\leq C e^{-\sigma_0 k R_t} \|\Phi\|_{L^2(\Omega_4)} \|\mathbf{n} \times (\nabla \times \Theta)\|_{\mathbf{H}^{-1/2}(\Gamma_\infty)}. \end{aligned}$$

We next estimate the negative norm on the right hand side above. Let $\hat{\mathbf{h}}$ be in $\widetilde{\mathbf{H}}^1(\Omega_\infty \setminus \bar{\Omega}_3)$ and let $\psi \in \mathbf{H}_0(\mathbf{curl}; \Omega_\infty \setminus \bar{\Omega}_3)$ satisfy

$$\begin{aligned} &-(\nabla \times \psi, \nabla \times \theta)_{\Omega_\infty \setminus \bar{\Omega}_3} + k^2 d_0^2 (\psi, \theta)_{\Omega_\infty \setminus \bar{\Omega}_3} \\ &= -(\nabla \times \hat{\mathbf{h}}, \nabla \times \theta)_{\Omega_\infty \setminus \bar{\Omega}_3} + k^2 d_0^2 (\hat{\mathbf{h}}, \theta)_{\Omega_\infty \setminus \bar{\Omega}_3} \end{aligned}$$

for all $\theta \in \mathbf{H}_0(\mathbf{curl}; \Omega_\infty \setminus \bar{\Omega}_3)$. This problem is well posed since d_0^2 has a nonzero imaginary part. We set $\mathbf{h} = \hat{\mathbf{h}} - \psi$. It follows that

$$\|\mathbf{h}\|_{\mathbf{H}(\mathbf{curl}; \Omega_\infty \setminus \bar{\Omega}_3)} \leq C \|\hat{\mathbf{h}}\|_{\mathbf{H}(\mathbf{curl}; \Omega_\infty \setminus \bar{\Omega}_3)}.$$

Note that both Θ and \mathbf{h} satisfy homogeneous equations in $\Omega_\infty \setminus \bar{\Omega}_3$,

$$\begin{aligned} (4.17) \quad &-\nabla \times \nabla \times \Theta + k^2 d_0^2 \Theta = \Delta \Theta + k^2 d_0^2 \Theta = \mathbf{0}, \\ &-\nabla \times \nabla \times \mathbf{h} + k^2 d_0^2 \mathbf{h} = \mathbf{0}. \end{aligned}$$

It follows that $\nabla \times \Theta$ and $\nabla \times \mathbf{h}$ are also in $\mathbf{H}(\mathbf{curl}; \Omega_\infty \setminus \bar{\Omega}_3)$. Now, integrating by parts we get

$$\begin{aligned} \langle \mathbf{n} \times \nabla \times \Theta, \hat{\mathbf{h}} \rangle_{\Gamma_\infty} &= \langle \nabla \times \Theta, \hat{\mathbf{h}} \times \mathbf{n} \rangle_{\Gamma_\infty} \\ &= -(\nabla \times \Theta, \nabla \times \mathbf{h})_{\Omega_\infty \setminus \bar{\Omega}_3} + (\nabla \times \nabla \times \Theta, \mathbf{h})_{\Omega_\infty \setminus \bar{\Omega}_3} \\ (4.18) \quad &= \langle \Theta, \mathbf{n} \times \nabla \times \mathbf{h} \rangle_{\Gamma_3} - (\Theta, \nabla \times \nabla \times \mathbf{h})_{\Omega_\infty \setminus \bar{\Omega}_3} \\ &\quad + (\nabla \times \nabla \times \Theta, \mathbf{h})_{\Omega_\infty \setminus \bar{\Omega}_3} \\ &= \langle \Theta, \mathbf{n} \times \nabla \times \mathbf{h} \rangle_{\Gamma_3}. \end{aligned}$$

The first integration by parts formula is justified as $\mathbf{h} = \hat{\mathbf{h}} + \psi$ and the formula holds for both terms. The second integration by parts above is justified because $\mathbf{n} \times \Theta$ vanishes on Γ_∞ and Θ is smooth in a neighborhood of Γ_3 since it satisfies (3.14) with $\beta = k^2 d_0^2$ there. Finally, using (3.15) (with D_1 a domain containing Γ_3 and whose closure is in $\Omega_4 \setminus \bar{\Omega}_2$), we get

$$\begin{aligned} |\langle \Theta, \mathbf{n} \times \nabla \times \mathbf{h} \rangle_{\Gamma_3}| &\leq C \|\Theta\|_{\mathbf{H}^1(D_1)} \|\nabla \times \mathbf{h}\|_{\mathbf{H}(\mathbf{curl}; \Omega_\infty \setminus \bar{\Omega}_3)} \\ &\leq C \|\Theta\|_{L^2(\Omega_4)} \|\hat{\mathbf{h}}\|_{\mathbf{H}^1(\Omega_\infty \setminus \bar{\Omega}_3)}. \end{aligned}$$

Combining the above results shows that

$$\|\mathbf{n} \times \nabla \times \Theta\|_{\mathbf{H}^{-1/2}(\Gamma_\infty)} \leq C \|\Theta\|_{L^2(\Omega_4)}$$

and hence, using (4.16),

$$\|\Theta\|_{L^2(\Omega_4)} \leq C e^{-\sigma_0 k R_t} \|\Theta\|_{L^2(\Omega_4)}.$$

It follows that Θ vanishes on Ω_4 for R_t sufficiently large. In this case, unique continuation implies that Θ vanishes on all of Ω_∞ . This completes the proof of the theorem. \square

5. EXISTENCE AND CONVERGENCE OF SOLUTIONS OF THE TRUNCATED ELECTROMAGNETIC PML PROBLEM (2.8)

The existence and convergence of solutions to the PML problem depend on the uniqueness result of the previous section. Accordingly, we shall assume that the hypotheses of Theorem 4.2 are satisfied throughout this section.

We are now in position to prove the existence theorem.

Theorem 5.1. *Let \mathbf{g} admit an $\mathbf{H}(\mathbf{curl}; \Omega_0)$ -extension $\widehat{\mathbf{g}}$ supported in Ω_0 . Then for R_t sufficiently large, the truncated PML problem (2.8) has a unique solution $\widetilde{\mathbf{E}_t}$.*

Proof. The theorem will follow if we show the existence of a solution to (4.1). Now that we have proved uniqueness of (4.2), we follow the proof of Theorem 4.1 with Ω^c replaced by Ω_∞ . In exactly the same way we arrive at the analogous inf-sup condition,

$$(5.1) \quad \|\mathbf{W}\|_{\mathbf{H}(\mathbf{curl}; \Omega_\infty)} \leq C \left(\sup_{\mathbf{V} \in \mathbf{X}(\Omega_\infty)} \frac{|\mathcal{A}(\mathbf{W}, \mathbf{V})|}{\|\mathbf{V}\|_{\mathbf{H}(\mathbf{curl}; \Omega_\infty)}} \right) \quad \text{for all } \mathbf{v} \in \mathbf{X}(\Omega_\infty).$$

Following the proof of Theorem 4.1, we define $\phi \in H_0^1(\Omega_\infty)$ by

$$\mathcal{A}(\nabla \phi, \nabla \psi) = \mathcal{A}(\mathbf{B}\widehat{\mathbf{g}}, \nabla \psi) \quad \text{for all } \nabla \psi \in H_0^1(\Omega_\infty).$$

Clearly, $\mathcal{A}(\boldsymbol{\theta}, \mathbf{u}) = 0$ for all $\boldsymbol{\theta} \in \mathbf{H}_0(\mathbf{curl}; \Omega_\infty)$ is the same as $\mathcal{A}(\bar{\mathbf{u}}, \boldsymbol{\theta}) = 0$ for all $\boldsymbol{\theta} \in \mathbf{H}_0(\mathbf{curl}; \Omega_\infty)$. As in the proof of Theorem 4.1, if $\mathbf{u} \in \mathbf{X}(\Omega_\infty)$ and $\mathcal{A}(\boldsymbol{\theta}, \mathbf{u}) = 0$ for all $\boldsymbol{\theta} \in \mathbf{H}_0(\mathbf{curl}; \Omega_\infty)$, then $\mathbf{u} = \mathbf{0}$. The generalized Lax-Millgram Theorem shows that there is a unique $\mathbf{W} \in \mathbf{X}(\Omega_\infty)$ satisfying

$$\mathcal{A}(\mathbf{W}, \mathbf{v}) = \mathcal{A}(\mathbf{B}\widehat{\mathbf{g}}, \mathbf{v}) - k^2(\boldsymbol{\mu} \nabla \phi, \mathbf{v}) \quad \text{for all } \mathbf{v} \in \mathbf{X}(\Omega_\infty).$$

Then $\boldsymbol{\Xi} = \mathbf{W} + \nabla \phi$ satisfies (4.1). Setting $\widetilde{\mathbf{E}_t} = \widehat{\mathbf{g}} - \mathbf{B}^{-1} \boldsymbol{\Xi}$ concludes the proof. \square

Finally, we want to estimate the error created in replacing the solution of the scattering problem (2.1) by the solution of the truncated PML problem (2.8). To do this, we compare the solutions of (2.8) and (2.4), since the solutions of (2.4) and (2.1) coincide in Ω_1 . The proof follows the arguments in the proof of uniqueness. We have the following convergence theorem.

Theorem 5.2. *Let $\widetilde{\mathbf{E}}$ be the solution of (2.4) and $\widetilde{\mathbf{E}_t}$ the solution of (2.8). For R_t sufficiently large,*

$$(5.2) \quad \|\widetilde{\mathbf{E}_t} - \widetilde{\mathbf{E}}\|_{L^2(\Omega_4)} \leq C e^{-2\sigma_0 k R_t} \|\widehat{\mathbf{g}}\|_{\mathbf{H}(\mathbf{curl}; \Omega_0)}.$$

Proof. Let $\widehat{\mathbf{F}} = \mathbf{B}(\widetilde{\mathbf{E}_t} - \widetilde{\mathbf{E}})$. We follow the proof of Theorem 4.2 replacing Θ by $\widehat{\mathbf{F}}$. The main difference is that $\mathbf{n} \times \widehat{\mathbf{F}}$ does not vanish on Γ_∞ .

Let Φ and $\hat{\mathbf{z}}$ be as in the proof of Theorem 4.2. Then, as in (4.15),

$$\begin{aligned}
 (\hat{\mathbf{F}}, \Phi)_{\Omega_4} &= -(\hat{\mathbf{F}}, \nabla \times \bar{\boldsymbol{\mu}}^{-1} \nabla \times \hat{\mathbf{z}})_{\Omega_\infty} + k^2 (\hat{\mathbf{F}}, \bar{\boldsymbol{\mu}} \hat{\mathbf{z}})_{\Omega_\infty} \\
 (5.3) \quad &= -(\boldsymbol{\mu}^{-1} \nabla \times \hat{\mathbf{F}}, \nabla \times \hat{\mathbf{z}})_{\Omega_\infty} + k^2 (\boldsymbol{\mu} \hat{\mathbf{F}}, \hat{\mathbf{z}})_{\Omega_\infty} \\
 &\quad - d_0^{-1} \langle \mathbf{n} \times \tilde{\mathbf{E}}, \nabla \times \hat{\mathbf{z}} \rangle_{\Gamma_\infty} \\
 &= d_0^{-1} \langle \mathbf{n} \times \nabla \times \hat{\mathbf{F}}, \hat{\mathbf{z}} \rangle_{\Gamma_\infty} - d_0^{-1} \langle \mathbf{n} \times \tilde{\mathbf{E}}, \nabla \times \hat{\mathbf{z}} \rangle_{\Gamma_\infty}.
 \end{aligned}$$

To bound the first term on the right hand side of (5.3), we follow the proof of Theorem 4.2. The integration by parts on (4.18) gives an extra term, i.e.,

$$(5.4) \quad \langle \mathbf{n} \times \nabla \times \hat{\mathbf{F}}, \hat{\mathbf{h}} \rangle_{\Gamma_\infty} = \langle \hat{\mathbf{F}}, \mathbf{n} \times \nabla \times \mathbf{h} \rangle_{\Gamma_3} - \langle \tilde{\mathbf{E}}, \mathbf{n} \times \nabla \times \mathbf{h} \rangle_{\Gamma_\infty}.$$

For the second term of (5.4), we note that since \mathbf{h} satisfies the homogeneous equation (4.17),

$$\| \mathbf{n} \times \nabla \times \mathbf{h} \|_{\mathbf{H}^{-1/2}(\Gamma_\infty)} \leq C \| \mathbf{h} \|_{\mathbf{H}(\mathbf{curl}; \Omega_\infty \setminus \bar{\Omega}_3)};$$

so by Propositions 3.1 and 3.2,

$$(5.5) \quad | \langle \tilde{\mathbf{E}}, \mathbf{n} \times \nabla \times \mathbf{h} \rangle_{\Gamma_\infty} | \leq C e^{-\sigma_0 k R_t} \| \tilde{\mathbf{E}} \|_{\mathbf{L}^2(\Omega^c)} \| \mathbf{h} \|_{\mathbf{H}(\mathbf{curl}; \Omega_\infty \setminus \bar{\Omega}_3)}.$$

Using (5.4), (5.5), Propositions 3.1 and 3.2, and following the proof of Theorem 4.2 (below (4.18)) gives

$$(5.6) \quad | \langle \mathbf{n} \times \nabla \times \hat{\mathbf{F}}, \hat{\mathbf{z}} \rangle_{\Gamma_\infty} | \leq C \| \Phi \|_{\mathbf{L}^2(\Omega_4)} (e^{-\sigma_0 k R_t} \| \hat{\mathbf{F}} \|_{\mathbf{L}^2(\Omega_4)} + e^{-2\sigma_0 k R_t} \| \tilde{\mathbf{E}} \|_{\mathbf{L}^2(\Omega_4)}).$$

Finally, we bound the second term on the right hand side of (5.3). Using a trace inequality we have that

$$| \langle \mathbf{n} \times \tilde{\mathbf{E}}, \nabla \times \hat{\mathbf{z}} \rangle_{\Gamma_\infty} | \leq C \| \tilde{\mathbf{E}} \|_{\mathbf{H}^1(S_\gamma)} \| \hat{\mathbf{z}} \|_{\mathbf{H}^2(S_\gamma)}.$$

Note that the components $\tilde{\mathbf{E}}$ and $\hat{\mathbf{z}}$ satisfy (3.14) with $\beta = k^2 d_0^2$ in S_γ . Thus from Proposition 3.1, Theorem 4.1, and Proposition 3.2, we obtain

$$| \langle \mathbf{n} \times \tilde{\mathbf{E}}, \nabla \times \hat{\mathbf{z}} \rangle_{\Gamma_\infty} | \leq C e^{-2\sigma_0 k R_t} \| \tilde{\mathbf{E}} \|_{\mathbf{L}^2(\Omega^c)} \| \Phi \|_{\mathbf{L}^2(\Omega_4)}.$$

Combining the above gives

$$| (\hat{\mathbf{F}}, \Phi)_{\Omega_4} | \leq C \| \Phi \|_{\mathbf{L}^2(\Omega_4)} (e^{-2\sigma_0 k R_t} \| \tilde{\mathbf{E}} \|_{\mathbf{L}^2(\Omega^c)} + e^{-\sigma_0 k R_t} \| \hat{\mathbf{F}} \|_{\mathbf{L}^2(\Omega_4)}).$$

The theorem easily follows from the above inequality and Corollary 4.1. \square

6. APPENDIX

We now provide a proof of Lemma 4.1. We first consider the case of Ω^c and $\omega = \Omega_2$. Let χ be a smooth cutoff function which is one on $\Omega_2 \setminus \Omega_1$ and supported in $\Omega_3 \setminus \Omega_0$. Let \mathbf{W} be in $\mathbf{X}(\Omega^c)$ and set $\mathbf{W}_1 = (1 - \chi)\mathbf{W}$. Then, \mathbf{W}_1 is in $\mathbf{H}_0(\mathbf{curl}; \Omega_2) \cap \mathbf{H}(\mathbf{div}; \Omega_2)$ and therefore is in $\mathbf{H}^s(\Omega_2)$ (see [1]).

The proof in this case will be complete if we show that \mathbf{W} is in $\mathbf{H}^s(D)$ where $D = \Omega_3 \setminus \Omega_0$. Let \check{D} and \tilde{D} extend D ($D \subset \check{D} \subset \tilde{D}$) with the closure of \check{D} contained in Ω^c and let χ_1 be a cutoff function, which is supported on \check{D} and is one on \tilde{D} . Let $\phi \in H_0^1(\tilde{D})$ be the solution of

$$(\nabla \phi, \nabla \theta)_{\tilde{D}} = (\chi_1 \mathbf{W}, \nabla \theta)_{\tilde{D}} \quad \text{for all } \theta \in H_0^1(\tilde{D}).$$

Then $\widetilde{\mathbf{W}} = \chi_1 \mathbf{W} - \nabla \phi$ is in $\mathbf{H}_0(\mathbf{curl}; \check{D}) \cap \mathbf{H}(\mathbf{div}; \tilde{D})$, i.e., it is in $\mathbf{H}^s(\tilde{D})$. We note that ϕ also satisfies

$$(6.1) \quad (\boldsymbol{\mu} \nabla \phi, \nabla \theta)_{\tilde{D}} = -(\boldsymbol{\mu} \widetilde{\mathbf{W}}, \nabla \theta)_{\tilde{D}} \quad \text{for all } \theta \in H_0^1(\tilde{D}).$$

Now, $\widetilde{\mathbf{W}}$ is in $\mathbf{H}^1(\tilde{D})$ (see Corollary 2.10 of [6]), so the right hand side above coincides with a bounded functional on L^2 . Since the coefficients in $\boldsymbol{\mu}$ are in $W_\infty^2(\tilde{D})$, the solution ϕ is in $H^2(D)$, i.e., $\nabla\phi$ is in $\mathbf{H}^1(D)$. Thus, $\mathbf{W} = \widetilde{\mathbf{W}} + \nabla\theta$ is in $\mathbf{H}^1(D)$.

The proof in the case of Ω_∞ is similar. The only difference is that one uses the constant coefficient operator in the neighborhood of both the inner and outer boundary (of Ω_∞) to reduce to regularity on an overlapping interior domain.

REFERENCES

- [1] C. Amrouche, C. Bernardi, M. Dauge, and V. Girault. Vector potentials in three-dimensional non-smooth domains. *Math. Methods Appl. Sci.*, 21(9):823–864, 1998. MR1626990 (99e:35037)
- [2] J.-P. Bérenger. A perfectly matched layer for the absorption of electromagnetic waves. *J. Comput. Phys.*, 114(2):185–200, 1994. MR1294924 (95e:78002)
- [3] J.-P. Bérenger. Three-dimensional perfectly matched layer for the absorption of electromagnetic waves. *J. Comput. Phys.*, 127(2):363–379, 1996. MR1412240 (97h:78001)
- [4] W. Chew and W. Weedon. A 3d perfectly matched medium for modified Maxwell's equations with stretched coordinates. *Microwave Opt. Technol. Lett.*, 13(7):599–604, 1994.
- [5] F. Collino and P. Monk. The perfectly matched layer in curvilinear coordinates. *SIAM J. Sci. Comp.*, 19(6):2061–2090, 1998. MR1638033 (99e:78029)
- [6] V. Girault and P. Raviart. *Finite Element Approximation of the Navier-Stokes Equations*. Lecture Notes in Math. 749, Springer-Verlag, New York, 1981. MR0548867 (83b:65122)
- [7] C. I. Goldstein. The finite element method with nonuniform mesh sizes applied to the exterior Helmholtz problem. *Numer. Math.*, 38(1):61–82, 1981/82. MR0634753 (82k:65062)
- [8] M. Grote and J. Keller. Nonreflecting boundary conditions for Maxwell's equations. *J. Comput. Phys.*, 139:327–342, 1998. MR1614098 (99a:78026)
- [9] N. Kantartzis, P. Petropoulos, and T. Tsiboukis. A comparison of the Grote-Keller exact ABC and the well posed PML for Maxwell's equations in spherical coordinates. *IEEE Trans. on Magnetics*, 35:1418–1422, 1999.
- [10] M. Lassas and E. Somersalo. On the existence and convergence of the solution of PML equations. *Computing*, 60(3):229–241, 1998. MR1621305 (99a:65133)
- [11] M. Lassas and E. Somersalo. Analysis of the PML equations in general convex geometry. *Proc. Roy. Soc. Edinburgh Sect. A*, 131(5):1183–1207, 2001. MR1862449 (2002k:35020)
- [12] P. Monk. *Finite Element Methods for Maxwell's Equations*. Numerical Mathematics and Scientific Computation. Oxford University Press, Oxford, UK, 2003. MR2059447 (2005d:65003)
- [13] J. Peetre. Espaces d'interpolation et théorème de Soboleff. *Ann. Inst. Fourier*, 16:279–317, 1966. MR0221282 (36:4334)
- [14] P. G. Petropoulos. Reflectionless sponge layers as absorbing boundary conditions for the numerical solution of Maxwell equations in rectangular, cylindrical, and spherical coordinates. *SIAM J. Appl. Math.*, 60(3):1037–1058 (electronic), 2000. MR1750090 (2001c:35230)
- [15] L. Tartar. *Topics in Nonlinear Analysis*. Math. d'Orsay, Univ. Paris-Sud, 1978. MR0532371 (81b:35001)

DEPARTMENT OF MATHEMATICS, TEXAS A&M UNIVERSITY, COLLEGE STATION, TEXAS 77843-3368

E-mail address: bramble@math.tamu.edu

DEPARTMENT OF MATHEMATICS, TEXAS A&M UNIVERSITY, COLLEGE STATION, TEXAS 77843-3368

E-mail address: pasciak@math.tamu.edu

Others

References

1. Garth A Baker and James H Bramble. Semidiscrete and single step fully discrete approximations for second order hyperbolic equations. *RAIRO. Analyse numérique*, 13(2):75–100, 1979.
2. Garth A Baker, James H Bramble, and Vidar Thomée. Single step galerkin approximations for parabolic problems. *Mathematics of Computation*, 31(140):818–847, 1977.
3. J Bramble, R Ewing, R Parashkevov, and J Pasciak. Domain decomposition methods for problems with uniform local refinement in two dimensions. In *Fourth International Symposium on Domain Decomposition Methods for Partial Differential Equations, Philadelphia*, 1991.
4. James Bramble and Joseph Pasciak. Analysis of a finite pml approximation for the three dimensional time-harmonic maxwell and acoustic scattering problems. *Mathematics of Computation*, 76(258):597–614, 2007.
5. James Bramble, Joseph Pasciak, and Apostol Vassilev. Analysis of non-overlapping domain decomposition algorithms with inexact solves. *Mathematics of Computation*, 67(221):1–19, 1998.
6. James H Bramble. Fourth-order finite difference analogues of the dirichlet problem for poisson’s equation in three and four dimensions. *Mathematics of Computation*, 17(83):217–222, 1963.
7. James H Bramble. A second order finite difference analog of the first biharmonic boundary value problem. *Numerische Mathematik*, 9(3):236–249, 1966.
8. James H Bramble. A proof of the inf–sup condition for the stokes equations on lipschitz domains. *Mathematical Models and Methods in Applied Sciences*, 13(03):361–371, 2003.
9. James H Bramble, Richard E Ewing, Rossen R Parashkevov, and Joseph E Pasciak. Domain decomposition methods for problems with partial refinement. *SIAM Journal on Scientific and Statistical Computing*, 13(1):397–410, 1992.
10. James H Bramble and SR Hilbert. Estimation of linear functionals on sobolev spaces with application to fourier transforms and spline interpolation. *SIAM Journal on Numerical Analysis*, 7(1):112–124, 1970.
11. James H Bramble and Bert E Hubbard. New monotone type approximations for elliptic problems. *Mathematics of Computation*, 18(87):349–367, 1964.
12. James H Bramble and Bert E Hubbard. On a finite difference analogue of an elliptic boundary problem which is neither diagonally dominant nor of non-negative type. *Journal of Mathematics and Physics*, 43(1-4):117–132, 1964.

References

13. James H Bramble and J Thomas King. A finite element method for interface problems in domains with smooth boundaries and interfaces. *Advances in Computational Mathematics*, 6(1):109–138, 1996.
14. James H Bramble, Do Y Kwak, and Joseph E Pasciak. Uniform convergence of multigrid v-cycle iterations for indefinite and nonsymmetric problems. *SIAM journal on numerical analysis*, 31(6):1746–1763, 1994.
15. James H Bramble, Zbigniew Leyk, and Joseph E Pasciak. The analysis of multigrid algorithms for pseudodifferential operators of order minus one. *Mathematics of computation*, 63(208):461–478, 1994.
16. James H Bramble and Joseph E Pasciak. New convergence estimates for multigrid algorithms. *Mathematics of computation*, 49(180):311–329, 1987.
17. James H Bramble and Joseph E Pasciak. A domain decomposition technique for stokes problems. *Applied Numerical Mathematics*, 6(4):251–261, 1990.
18. James H Bramble and Joseph E Pasciak. The analysis of smoothers for multigrid algorithms. *mathematics of computation*, 58(198):467–488, 1992.
19. James H Bramble, Joseph E Pasciak, and Alfred H Schatz. The construction of preconditioners for elliptic problems by substructuring. i. *Mathematics of Computation*, 47(175):103–134, 1986.
20. James H Bramble, Joseph E Pasciak, and Alfred H Schatz. The construction of preconditioners for elliptic problems by substructuring. ii. *Mathematics of Computation*, 49(179):1–16, 1987.
21. James H Bramble, Joseph E Pasciak, and Alfred H Schatz. The construction of preconditioners for elliptic problems by substructuring. iii. *Mathematics of Computation*, 51(184):415–430, 1988.
22. James H Bramble, Joseph E Pasciak, and Alfred H Schatz. The construction of preconditioners for elliptic problems by substructuring. iv. *Mathematics of Computation*, 53(187):1–24, 1989.
23. James H Bramble, Joseph E Pasciak, Jun Ping Wang, and Jinchao Xu. Convergence estimates for multigrid algorithms without regularity assumptions. *Mathematics of Computation*, 57(195):23–45, 1991.
24. James H Bramble, Joseph E Pasciak, Jun Ping Wang, and Jinchao Xu. Convergence estimates for product iterative methods with applications to domain decomposition. *Mathematics of Computation*, 57(195):1–21, 1991.
25. James H Bramble, Joseph E Pasciak, and Jinchao Xu. The analysis of multigrid algorithms for nonsymmetric and indefinite elliptic problems. *Mathematics of Computation*, 51(184):389–414, 1988.
26. James H Bramble, Joseph E Pasciak, and Jinchao Xu. Parallel multilevel preconditioners. *Mathematics of computation*, 55(191):1–22, 1990.
27. James H Bramble, Joseph E Pasciak, and Jinchao Xu. The analysis of multigrid algorithms with nonnested spaces or noninherited quadratic forms. *Mathematics of Computation*, 56(193):1–34, 1991.
28. James H Bramble and Alfred H Schatz. Rayleigh-ritz-galerkin methods for dirichlet’s problem using subspaces without boundary conditions. *Communications on Pure and Applied Mathematics*, 23(4):653–675, 1970.
29. James H Bramble and Alfred H Schatz. Higher order local accuracy by averaging in the finite element method. *Mathematics of Computation*, 31(137):94–111, 1977.
30. James H Bramble and Xuejun Zhang. The analysis of multigrid methods. *Handbook of numerical analysis*, 7:173–415, 2000.
31. James H Bramble and Miloš Zlámal. Triangular elements in the finite element method. *Mathematics of Computation*, 24(112):809–820, 1970.

References

32. JH Bramble. Error estimates for difference methods in forced vibration problems. *SIAM Journal on Numerical Analysis*, 3(1):1–12, 1966.
33. JH Bramble. On the convergence of difference approximations to weak solutions of dirichlet's problem. *Numerische Mathematik*, 13(2):101–111, 1969.
34. JH Bramble and BE Hubbard. On the formulation of finite difference analogues of the dirichlet problem for poisson's equation. *Numerische Mathematik*, 4(1):313–327, 1962.
35. JH Bramble and BE Hubbard. Approximation of solutions of mixed boundary value problems for poisson's equation by finite differences. *Journal of the ACM (JACM)*, 12(1):114–123, 1965.
36. JH Bramble and BE Hubbard. A finite difference analog of the neumann problem for Poisson's equation. *Journal of the Society for Industrial and Applied Mathematics, Series B: Numerical Analysis*, 2(1):1–14, 1965.
37. JH Bramble, AH Schatz, V Thomée, and LB Wahlbin. Some convergence estimates for semidiscrete galerkin type approximations for parabolic equations. *SIAM Journal on Numerical Analysis*, 14(2):218–241, 1977.