A CLASS OF ITERATIVE METHODS FOR SOLVING SADDLE POINT PROBLEMS

RANDOLPH E. BANK*, BRUNO D. WELFERT†, AND HARRY YSERENTANT‡

Abstract. We consider the numerical solution of indefinite systems of linear equations arising in the calculation of saddle points. We are mainly concerned with sparse systems of this type resulting from certain discretizations of partial differential equations. We present an iterative method involving two levels of iteration, similar in some respects to the Uzawa algorithm. We relate the rates of convergence of the outer and inner iterations, proving that, under natural hypotheses, the outer iteration achieves the rate of convergence of the inner iteration. The technique is applied to finite element approximations of the Stokes equations.

Key words. Saddle point problems, iterative solvers, mixed finite element methods.

AMS subject classifications. 65F10, 65N20, 65N30.

1. Introduction.

In this paper, we consider the solution of the system of linear equations

$$\begin{bmatrix} A & B^t \\ B & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix} \tag{1.1}$$

Here A is a symmetric, positive definite $n \times n$ matrix, and B is an $m \times n$ matrix. We assume that the $(n+m) \times (n+m)$ coefficient matrix

$$M = \begin{bmatrix} A & B^t \\ B & 0 \end{bmatrix} \tag{1.2}$$

is nonsingular. Linear systems such as (1.1) correspond to saddle point problems, and can arise, for example, from mixed finite element formulations for second order elliptic problems or from Lagrange multiplier methods. Our main application will be to the numerical solution of the Stokes equations.

The linear system (1.1) may be reformulated as

$$Ax + B^t y = f \tag{1.3}$$

$$BA^{-1}B^t y = BA^{-1} f - g \tag{1.4}$$

This system is uniquely solvable and the matrix (1.2) is nonsingular if and only if the symmetric matrix

$$C = BA^{-1}B^t \tag{1.5}$$

---

*Department of Mathematics, University of California at San Diego, La Jolla, CA 92093, USA. The work of this author was supported by the Office of Naval Research under contract N00014-82K-0197, by Avions Marcel Dassault, 78 Quai Marcel Dassault, 92214 St Cloud, France, and by Direction des Recherches Etudes et Techniques, 26 boulevard Victor, 75996 Paris Armées, France.

†Department of Mathematics, University of California at San Diego, La Jolla, California 92093. The work of this author was supported by Avions Marcel Dassault - Breguet Aviation, 78 quai Marcel Dassault, 92214 St Cloud, France and by Direction des Recherches Etudes et Techniques, 26 boulevard Victor, 75996 Paris Armées, France.

‡Fachbereich Mathematik, Universität Dortmund, D-4600 Dortmund 50, Federal Republic of Germany. The work of this author was supported by Konrad-Zuse-Zentrum für Informationstechnik Berlin, West Germany.

is positive definite. Clearly $m \leq n$ is a necessary condition for C to be positive definite.

The matrix M can be factored as

$$M = \begin{bmatrix} A & 0 \\ B & I \end{bmatrix} \begin{bmatrix} A^{-1} & 0 \\ 0 & -C \end{bmatrix} \begin{bmatrix} A & B^t \\ 0 & I \end{bmatrix} \tag{1.6}$$

Sylvester's law of inertia then implies that M has n positive and m negative eigenvalues, and thus is strongly indefinite.

There are several approaches to the iterative solution of (1.1). First we mention multigrid methods; see for example, [11] [12] [13] [15] [16] [17], or for later developments using alternative splittings [18] [19]. A comprehensive bibliography is contained in [13]. Among the more classical methods, the most prominent is the Uzawa method [12], which is a Jacobi or gradient-type iteration for the solution of (1.4). This scheme requires that $A^{-1}x$ can be computed exactly. Verfürth [14] replaces the computation of $A^{-1}x$ by an approximate multigrid solution of the corresponding linear system, improving the accuracy of this inner iteration as the outer iteration proceeds to finally get a solution of (1.4). Bramble and Pasciak [5] discuss a procedure for directly preconditioning M given a preconditioning matrix for A. Other preconditionings were proposed by Axelsson [7] and by Dyn and Ferguson [8].

Our approach shares some features with several of these schemes. We replace the matrices A and C in (1.6) by the symmetric, positive definite matrices $\hat{A}$ and $\hat{C}$ corresponding to the iterations

$$x \leftarrow x + \hat{A}^{-1}(b - Ax) \tag{1.7}$$

for the solution of

$$Ax = b \tag{1.8}$$

and

$$y \leftarrow y + \hat{C}^{-1}(c - B\hat{A}^{-1}B^t y) \tag{1.9}$$

for the solution of

$$Hy = B\hat{A}^{-1}B^t y = c \tag{1.10}$$

Each iteration of (1.7) or (1.9) may in fact correspond to several steps of a given iterative method. This leads to the approximation of M given by

$$\hat{M} = \begin{bmatrix} \hat{A} & 0 \\ B & I \end{bmatrix} \begin{bmatrix} \hat{A}^{-1} & 0 \\ 0 & -\hat{C} \end{bmatrix} \begin{bmatrix} \hat{A} & B^t \\ 0 & I \end{bmatrix} \tag{1.11}$$

$$= \begin{bmatrix} \hat{A} & B^t \\ B & \hat{D} \end{bmatrix} \tag{1.12}$$

with

$$\hat{D} = B\hat{A}^{-1}B^t - \hat{C} \tag{1.13}$$

As with the original matrix M, $\hat{M}$ will have n positive and m negative eigenvalues. Our iterative scheme can be summarized as

$$\begin{bmatrix} x_{i+1} \\ y_{i+1} \end{bmatrix} = \begin{bmatrix} x_i \\ y_i \end{bmatrix} + \begin{bmatrix} \hat{A} & B^t \\ B & \hat{D} \end{bmatrix}^{-1} \left\{ \begin{bmatrix} f \\ g \end{bmatrix} - \begin{bmatrix} A & B^t \\ B & 0 \end{bmatrix} \begin{bmatrix} x_i \\ y_i \end{bmatrix} \right\} \tag{1.14}$$

Our iterative method involves two nested iterations, an outer iteration (1.14) for M and two inner iterations (1.7) and (1.9). In this respect it is similar to many other approaches to this problem. However, it is important to note that the iteration (1.9) is directed towards the solution of a linear system of the form (1.10) rather than one of the form (1.4). Had we chosen to solve (1.4), at each iteration step the required multiplication by C would typically introduce another level of iteration. Our use of H rather than C is an explicit recognition that such a third level of iteration would probably be very expensive, and thus unlikely to be iterated to completion. See [10] for a recent study of inner and outer iterations.

Typically, we expect that $\hat{A}$ will correspond to a small number of iterations (e.g., one) of some iterative method for solving (1.8). In the example of the Stokes equations, $\hat{A}$ might represent one cycle of the hierarchical basis multigrid method for the Poisson equation [3] [4] [20]. In this case one iteration is the most efficient choice, and we expect that this will most often be the case. However, since the choice of $\hat{A}$ has some secondary effects on the condition number of H, and hence on the choice of $\hat{C}$, it is possible that more than one iteration may be useful in some situations.

We expect that $\hat{C}$ will also correspond to the application of several iterations of some iterative method for solving (1.10). Unlike $\hat{A}$, where choosing one iteration is usually most efficient, controlling the number of inner iterations for $\hat{C}$ is much more subtle. From a purely theoretical point of view, we will see in sections 2 and 4 that if the convergence rate for (1.7) is α, then by taking sufficiently many inner iterations that the rate of convergence for $\hat{C}$ is bounded by $\alpha/(2+\alpha)$, we can guarantee that the overall convergence rate for (1.14) will be bounded by α. This is shown for a stationary iterative method for (1.10) in section 2. In section 4, we prove a similar result for the case where $\hat{C}$ corresponds to a preconditioned conjugate gradient iteration, and thus $\hat{C} = \hat{C}_i$ for each outer iteration.

In one of the more interesting twists in this algorithm, it turns out that forcing $\hat{C}$ to more closely approximate H (e.g., by taking more inner iterations) might dramatically increase the rate of convergence, even though (1.10) is the "wrong" linear system. Indeed, in section 3, we will give an example where solving (1.10) exactly guarantees convergence of (1.14) in 2 iterations, for *any* symmetric, positive definite $\hat{A}$ and independently of the rate of convergence α for (1.7). In this extreme case our algorithm essentially becomes a direct method.

In the practical example of the Stokes equations (section 5), we are not quite fortunate, but there we do encounter convergence rates significantly faster than those predicted by our norm estimates, although they apparently are not independent of α.

2. Analysis of the Convergence Rate. Let (x, y) denote the Euclidean inner product, $\|x\| = (x, x)^{1/2}$ the corresponding vector norm, and $\|F\|$ the spectral norm of the matrix F. If F is symmetric and positive definite, we will define the F norm of a vector by

$$\|x\|_F^2 = (x, Fx)$$

The rates of convergence of the basic iterations (1.7) and (1.9) with respect to the norms $\|x\|_A$ and $\|y\|_H$, respectively, are

$$\alpha = \|I - \bar{A}\| \tag{2.1}$$

$$\beta = \|I - \bar{B}\bar{B}^t\| \tag{2.2}$$

where the matrices $\bar{A}$ and $\bar{B}$ are defined by

$$\bar{A} = \hat{A}^{-1/2} A \hat{A}^{-1/2} \tag{2.3}$$

$$\bar{B} = \hat{C}^{-1/2} B \hat{A}^{-1/2} \tag{2.4}$$

Our aim in this section is to estimate the rate of convergence for the iteration (1.14) with respect to the norm

$$\|v\|_*^2 = \|x\|_A^2 + \|y\|_G^2 \qquad v = \begin{bmatrix} x \\ y \end{bmatrix} \in \mathcal{R}^n \times \mathcal{R}^m \tag{2.5}$$

in terms of α and β. Here G is a given symmetric positive definite $m \times m$ matrix. For finite element discretizations of the Stokes equation, $\|x\|_A$ can be interpreted as a scaled $\mathcal{H}^1$-norm of the finite element function corresponding to the coefficient vector x, and $\|y\|_G$ is a scaled $\mathcal{L}^2$-norm of the function represented by y.

We assume that there are positive constants μ_1 and μ_2 satisfying

$$\inf_{y \neq 0} \sup_{x \neq 0} \frac{(Bx, y)}{\|x\|_A \|y\|_G} \geq \sqrt{\mu_1} \tag{2.6}$$

and

$$|(Bx, y)| \leq \sqrt{\mu_2}\, \|x\|_A \|y\|_G \tag{2.7}$$

for all $x \in \mathcal{R}^n$ and all $y \in \mathcal{R}^m$.

Theorem 2.1. *Let $0 < \xi_1 \leq \xi_2 \leq \ldots \leq \xi_m$ be the eigenvalues of*

$$G^{-1/2} C G^{-1/2} = (G^{-1/2} B A^{-1/2})(G^{-1/2} B A^{-1/2})^t \tag{2.8}$$

Then the best possible constants in (2.6) and (2.7) are $\mu_1 = \xi_1$ and $\mu_2 = \xi_m$.

Proof. Consider the singular value decomposition

$$G^{-1/2} B A^{-1/2} = X S Y^t \tag{2.9}$$

Then

$$\begin{aligned} \inf_{y \neq 0} \sup_{x \neq 0} \frac{(Bx, y)}{\|x\|_A \|y\|_G} &= \inf_{y \neq 0} \sup_{x \neq 0} \frac{(x, A^{-1/2} B^t G^{-1/2} y)}{\|x\| \, \|y\|} \\ &= \inf_{y \neq 0} \frac{\|A^{-1/2} B^t G^{-1/2} y\|}{\|y\|} \\ &= \inf_{y \neq 0} \frac{\|Y S^t X^t y\|}{\|y\|} \\ &= \inf_{y \neq 0} \frac{\|S^t X^t y\|}{\|X^t y\|} \\ &= \sqrt{\xi_1} \end{aligned}$$

Inequality (2.7) holds if and only if for all x and y

$$(G^{-1/2} B A^{-1/2} x, y) \leq \sqrt{\mu_2}\, \|x\| \, \|y\|$$

Using the singular value decomposition in a fashion similar to the argument above, we see that the optimal choice for μ_2 is ξ_m. □

As a consequence of Theorem 1 we have $\mu_1 \leq \xi_i \leq \mu_2$ for $1 \leq i \leq m$, or equivalently

$$\mu_1 \|y\|_G^2 \leq \|y\|_C^2 \leq \mu_2 \|y\|_G^2 \tag{2.10}$$

for $y \in \mathcal{R}^m$. Inequalities (2.10) correspond to Lemma 3.1 of [14].

We remark that using the singular value decomposition (2.9), the matrix

$$\begin{bmatrix} A^{1/2} & 0 \\ 0 & G^{1/2} \end{bmatrix} \begin{bmatrix} A & B^t \\ B & 0 \end{bmatrix}^{-1} \begin{bmatrix} A^{1/2} & 0 \\ 0 & G^{1/2} \end{bmatrix} \tag{2.11}$$

is similar to

$$\begin{bmatrix} I & S^t \\ S & 0 \end{bmatrix}^{-1}$$

Therefore, its spectral norm is

$$\begin{cases} \max(1, 1/c) & \text{if } m < n \\ 1/c & \text{if } m = n \end{cases} \tag{2.12}$$

with

$$c = \frac{\sqrt{1 + 4\xi_1} - 1}{2} \geq \frac{\sqrt{1 + 4\mu_1} - 1}{2} \tag{2.13}$$

In the finite element context, the spectral norm (2.12) of the matrix (2.11) is the norm of the operator projecting the continuous solution onto the discrete solution. Thus (2.6) characterizes the stability of the discretization. This is the famous Babuška-Brezzi condition [2] [6] [9].

In our analysis, we ultimately will state convergence estimates using the $\|\cdot\|_*$ norm, since this is a natural norm associated with the problem, and in the finite element setting, it is the norm in which discretization errors are estimated. However, in deriving estimates for the rate of convergence, other norms will prove more useful. Let $|||\cdot|||$ be a norm on $\mathcal{R}^n \times \mathcal{R}^m$ which is comparable to the norm (2.5) in the sense that there are positive constants κ_1 and κ_2 of reasonable size with

$$\kappa_1 \|v\|_*^2 \leq |||v|||^2 \leq \kappa_2 \|v\|_*^2 \tag{2.14}$$

for $v \in \mathcal{R}^n \times \mathcal{R}^m$. Assume that the induced matrix norm of the error propagation matrix

$$I - \hat{M}^{-1} M = \begin{bmatrix} \hat{A} & B^t \\ B & \hat{D} \end{bmatrix}^{-1} \begin{bmatrix} \hat{A} - A & 0 \\ 0 & \hat{D} \end{bmatrix} \tag{2.15}$$

of the iterative method (1.14) satisfies the estimate

$$|||I - \hat{M}^{-1} M||| \leq \delta \tag{2.16}$$

Then with respect to the original norm (2.5) one gets the error reduction

$$\begin{aligned} \|e_i\|_* &\leq \kappa_1^{-1/2} \; |||e_i||| \\ &\leq \kappa_1^{-1/2} \, \delta^i \; |||e_0||| \\ &\leq (\kappa_2/\kappa_1)^{1/2} \, \delta^i \; \|e_0\|_* \end{aligned}$$

after i iterations. Good candidates for the $|||\cdot|||$ norm are norms behaving like

$$|||v|||^2 \sim \|x\|_{\hat{A}}^2 + \|y\|_H^2 \qquad v = \begin{bmatrix} x \\ y \end{bmatrix}$$

LEMMA 2.2. *For all $x \in \mathcal{R}^n$,*

$$(1+\alpha)^{-1}\|x\|_A^2 \le \|x\|_{\hat{A}}^2 \le (1-\alpha)^{-1}\|x\|_A^2 \tag{2.17}$$

and for all $y \in \mathcal{R}^m$,

$$(1-\alpha)\|y\|_C^2 \le \|y\|_H^2 \le (1+\alpha)\|y\|_C^2 \tag{2.18}$$

Proof. Because of (2.1), the eigenvalues of the matrix (2.3) range between $1-\alpha$ and $1+\alpha$. Inequalities (2.17) and

$$(1-\alpha)\|x\|_{A^{-1}}^2 \le \|x\|_{\hat{A}^{-1}}^2 \le (1+\alpha)\|x\|_{A^{-1}}^2 \tag{2.19}$$

follow. From (2.19),

$$\|(BA^{-1}B^t)^{1/2}y\| = \|A^{-1/2}B^ty\| = \|B^ty\|_{A^{-1}}$$

and

$$\|(B\hat{A}^{-1}B^t)^{1/2}y\| = \|\hat{A}^{-1/2}B^ty\| = \|B^ty\|_{\hat{A}^{-1}}$$

one obtains (2.18). $\square$

To construct vector norms for which an estimate like (2.16) can be proven, let

$$\bar{B} = U\Sigma V^t \tag{2.20}$$

be the singular value decomposition of the matrix (2.4). As usual, U is an orthogonal $m \times m$ matrix and V is an orthogonal $n \times n$ matrix. The diagonal elements $\sigma_i \ge 0$ of

$$\Sigma = \begin{bmatrix} \sigma_1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_m & 0 & \cdots & 0 \end{bmatrix} = \begin{bmatrix} \Sigma_0 & 0 \end{bmatrix} \tag{2.21}$$

are the singular values of $\bar{B}$. The diagonal matrix Σ_0 is $m \times m$. The eigenvalues of the error propagation matrix of the iteration (1.9)

$$I - \hat{C}^{-1}(B\hat{A}^{-1}B^t) = I - \hat{C}^{-1}H \tag{2.22}$$

can be expressed in terms of these singular values and are given by

$$\lambda_i = 1 - \sigma_i^2 \qquad 1 \le i \le m \tag{2.23}$$

The rate of convergence (2.2) is

$$\beta = \max_{1 \le i \le m} |\lambda_i| \tag{2.24}$$

THEOREM 2.3. *Let Γ be an $m \times m$ diagonal matrix with diagonal elements $\gamma_i > 0$, $1 \le i \le m$. Let the norm $||| \cdot |||$ be defined by*

$$|||v|||^2 = \|x\|_{\hat{A}}^2 + \|\Gamma U^t \hat{C}^{1/2} y\|^2 \qquad v = \begin{bmatrix} x \\ y \end{bmatrix} \tag{2.25}$$

Then the induced norm of the error propagation matrix (2.15) satisfies

$$|||I - \hat{M}^{-1} M||| \le \delta \tag{2.26}$$

with

$$\delta = \begin{cases} \max(\alpha, \rho) & \textit{if } m < n \\ \rho & \textit{if } m = n \end{cases} \tag{2.27}$$

where ρ is given by

$$\rho = \max_{1 \le i \le m} \left|\left|\left| \begin{bmatrix} \alpha\lambda_i & -\lambda_i \gamma_i^{-1} \sqrt{1-\lambda_i} \\ \alpha\gamma_i \sqrt{1-\lambda_i} & \lambda_i \end{bmatrix} \right|\right|\right| \tag{2.28}$$

Proof. Let

$$F = \begin{bmatrix} \hat{A}^{1/2} & 0 \\ 0 & \hat{C}^{1/2} \end{bmatrix} \begin{bmatrix} \hat{A} & B^t \\ B & \hat{D} \end{bmatrix}^{-1} \begin{bmatrix} \hat{A} - A & 0 \\ 0 & \hat{D} \end{bmatrix} \begin{bmatrix} \hat{A}^{-1/2} & 0 \\ 0 & \hat{C}^{-1/2} \end{bmatrix}$$

Because of

$$\begin{bmatrix} \hat{A}^{-1/2} & 0 \\ 0 & \hat{C}^{-1/2} \end{bmatrix} \begin{bmatrix} \hat{A} & B^t \\ B & \hat{D} \end{bmatrix} \begin{bmatrix} \hat{A}^{-1/2} & 0 \\ 0 & \hat{C}^{-1/2} \end{bmatrix} = \begin{bmatrix} I & \bar{B}^t \\ \bar{B} & \bar{B}\bar{B}^t - I \end{bmatrix}$$

$$\begin{bmatrix} \hat{A}^{-1/2} & 0 \\ 0 & \hat{C}^{-1/2} \end{bmatrix} \begin{bmatrix} \hat{A} - A & 0 \\ 0 & \hat{D} \end{bmatrix} \begin{bmatrix} \hat{A}^{-1/2} & 0 \\ 0 & \hat{C}^{-1/2} \end{bmatrix} = \begin{bmatrix} I - \bar{A} & 0 \\ 0 & \bar{B}\bar{B}^t - I \end{bmatrix}$$

and

$$\begin{bmatrix} I & \bar{B}^t \\ \bar{B} & \bar{B}\bar{B}^t - I \end{bmatrix}^{-1} = \begin{bmatrix} I - \bar{B}^t\bar{B} & \bar{B}^t \\ \bar{B} & -I \end{bmatrix}$$

the representation

$$F = \begin{bmatrix} I - \bar{B}^t\bar{B} & \bar{B}^t\bar{B}\bar{B}^t - \bar{B}^t \\ \bar{B} & I - \bar{B}\bar{B}^t \end{bmatrix} \begin{bmatrix} I - \bar{A} & 0 \\ 0 & I \end{bmatrix}$$

follows. Using the singular value decomposition (2.20), one gets

$$\begin{aligned} F &= \begin{bmatrix} V(I - \Sigma^t\Sigma)V^t & V(\Sigma^t\Sigma\Sigma^t - \Sigma^t)U^t \\ U\Sigma V^t & U(I - \Sigma\Sigma^t)U^t \end{bmatrix} \begin{bmatrix} I - \bar{A} & 0 \\ 0 & I \end{bmatrix} \\ &= \begin{bmatrix} V & 0 \\ 0 & U \end{bmatrix} \begin{bmatrix} I - \Sigma^t\Sigma & \Sigma^t\Sigma\Sigma^t - \Sigma^t \\ \Sigma & I - \Sigma\Sigma^t \end{bmatrix} \begin{bmatrix} V^t(I - \bar{A})V & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} V^t & 0 \\ 0 & U^t \end{bmatrix} \end{aligned}$$

Equivalently,

$$\begin{aligned} &\begin{bmatrix} V^t\hat{A}^{1/2} & 0 \\ 0 & \Gamma U^t \hat{C}^{1/2} \end{bmatrix} \begin{bmatrix} \hat{A} & B^t \\ B & \hat{D} \end{bmatrix}^{-1} \begin{bmatrix} \hat{A} - A & 0 \\ 0 & \hat{D} \end{bmatrix} \begin{bmatrix} \hat{A}^{-1/2}V & 0 \\ 0 & \hat{C}^{-1/2}U\Gamma^{-1} \end{bmatrix} \\ &\qquad = \begin{bmatrix} \alpha(I - \Sigma^t\Sigma) & (\Sigma^t\Sigma\Sigma^t - \Sigma^t)\Gamma^{-1} \\ \alpha\Gamma\Sigma & I - \Sigma\Sigma^t \end{bmatrix} \begin{bmatrix} \alpha^{-1}V^t(I - \bar{A})V & 0 \\ 0 & I \end{bmatrix} \end{aligned}$$

The spectral norm of the matrix on the left hand side of this equation is equal to $|||I - \hat{M}^{-1}M|||$. Using (2.1), one can easily see that

$$\left|\!\left|\!\left| \begin{bmatrix} \alpha^{-1}V^t(I - \bar{A})V & 0 \\ 0 & I \end{bmatrix} \right|\!\right|\!\right| = 1$$

The last equation leads to $|||I - \hat{M}^{-1}M||| \le \delta$ with

$$\delta = \left|\!\left|\!\left| \begin{bmatrix} \alpha(I - \Sigma^t\Sigma) & (\Sigma^t\Sigma\Sigma^t - \Sigma^t)\Gamma^{-1} \\ \alpha\Gamma\Sigma & I - \Sigma\Sigma^t \end{bmatrix} \right|\!\right|\!\right|$$

For $m < n$,it follows that δ is the spectral norm of the matrix

$$\begin{bmatrix} \alpha(I - \Sigma_0^2) & 0 & (\Sigma_0^3 - \Sigma_0)\Gamma^{-1} \\ 0 & \alpha I & 0 \\ \alpha\Gamma\Sigma_0 & 0 & I - \Sigma_0^2 \end{bmatrix}$$

where Σ_0 is defined in (2.21). Therefore, for $m < n$

$$\delta = \max(\alpha, \rho)$$

with

$$\rho = \max_{1 \le i \le m} \left|\!\left|\!\left| \begin{bmatrix} \alpha(1 - \sigma_i^2) & (\sigma_i^3 - \sigma_i)\gamma_i^{-1} \\ \alpha\sigma_i\gamma_i & 1 - \sigma_i^2 \end{bmatrix} \right|\!\right|\!\right|$$

For $m = n$, one has $\Sigma_0 = \Sigma$ and $\delta = \rho$. With $1 - \sigma_i^2 = \lambda_i$, the representation is proved. ☐

Note that for $\hat{A} = (1 - \alpha)^{-1}A$, we have $I - \bar{A} = \alpha I$ and

$$|||I - \hat{M}^{-1}M||| = \delta$$

Thus, for this choice of $\hat{A}$, our estimate is sharp.

To estimate the constant ρ in (2.28), the following lemma is useful

LEMMA 2.4. *For $\lambda < 1$ and $0 < \gamma^2 < 1 + \lambda$,*

$$\left|\!\left|\!\left| \begin{bmatrix} \alpha\lambda & -\lambda\gamma^{-1}\sqrt{1-\lambda} \\ \alpha\gamma\sqrt{1-\lambda} & \lambda \end{bmatrix} \right|\!\right|\!\right| \le \max\left(\alpha, \sqrt{\frac{\lambda^2}{\gamma^2(1 + \lambda - \gamma^2)}}\right) \tag{2.29}$$

Proof. For all $\xi, \eta \in \mathcal{R}$, and all $\epsilon \neq 0$, one has

$$\left|\!\left|\!\left| \begin{bmatrix} \alpha\lambda & -\lambda\gamma^{-1}\sqrt{1-\lambda} \\ \alpha\gamma\sqrt{1-\lambda} & \lambda \end{bmatrix} \begin{bmatrix} \xi \\ \eta \end{bmatrix} \right|\!\right|\!\right|^2$$

$$= (\lambda^2 - \gamma^2\lambda + \gamma^2)(\alpha\xi)^2 + 2\epsilon\alpha\xi\frac{(\gamma - \gamma^{-1}\lambda)\lambda\sqrt{1-\lambda}}{\epsilon}\eta + (\gamma^{-2}\lambda^2(1 - \lambda) + \lambda^2)\eta^2$$

$$\le (\lambda^2 - \gamma^2\lambda + \gamma^2 + \epsilon^2)(\alpha\xi)^2 + (\gamma^{-2}\lambda^2(1 - \lambda) + \lambda^2 + \frac{(\gamma - \gamma^{-1}\lambda)^2\lambda^2(1-\lambda)}{\epsilon^2})\eta^2$$

Because $\lambda < 1$ and $\gamma^2 < 1 + \lambda$, we can choose

$$\epsilon^2 = (1 + \lambda - \gamma^2)(1 - \lambda)$$

yielding the estimate

$$\left|\left|\left|\begin{bmatrix} \alpha\lambda & -\lambda\gamma^{-1}\sqrt{1-\lambda} \\ \alpha\gamma\sqrt{1-\lambda} & \lambda \end{bmatrix}\begin{bmatrix} \xi \\ \eta \end{bmatrix}\right|\right|\right|^2 \le \alpha^2\xi^2 + \frac{\lambda^2}{\gamma^2(1+\lambda-\gamma^2)}\,\eta^2$$

This proves the lemma. □

In the next theorem, we consider a particular choice for the norm (2.25).

THEOREM 2.5. *Let*

$$\theta = \frac{1}{2}\frac{1-\beta}{1+\beta} \tag{2.30}$$

$$|||v|||^2 = \|x\|_{\hat{A}}^2 + \theta\|y\|_H^2 \qquad\qquad v = \begin{bmatrix} x \\ y \end{bmatrix} \tag{2.31}$$

Then, with respect to the induced matrix norm

$$|||I - \hat{M}^{-1}M||| \le \delta \tag{2.32}$$

with

$$\delta = \max\left(\alpha,\ \frac{2\beta}{1-\beta}\right) \tag{2.33}$$

Furthermore, the norms (2.5) and (2.31) satisfy

$$\kappa_1\|v\|_*^2 \le |||v|||^2 \le \kappa_2\|v\|_*^2 \tag{2.34}$$

with

$$\frac{\kappa_2}{\kappa_1} = \frac{1+\alpha}{1-\alpha}\,\frac{\max(1,\theta(1-\alpha^2)\mu_2)}{\min(1,\theta(1-\alpha^2)\mu_1)} \tag{2.35}$$

Proof. Fix a constant $c > 0$ and set

$$\Gamma = (c\Sigma\Sigma^t)^{1/2} = c^{1/2}\Sigma_0$$

Then

$$(\Gamma U^t\hat{C}^{1/2})^t(\Gamma U^t\hat{C}^{1/2}) = cB\hat{A}^{-1}B^t = cH \tag{2.36}$$

so that for all $y \in \mathcal{R}^m$

$$c\|y\|_H^2 = \|\Gamma U^t\hat{C}^{1/2}y\|^2$$

Therefore the norm

$$|||v|||^2 = \|x\|_{\hat{A}}^2 + c\|y\|_H^2 \qquad\qquad v = \begin{bmatrix} x \\ y \end{bmatrix}$$

fits into the framework of Theorem 3. Let

$$\gamma(\lambda) = \sqrt{c(1-\lambda)}$$

Then $\gamma_i \equiv \gamma(\lambda_i)$, $1 \le i \le m$, are the diagonal elements of the matrix Γ introduced above.

It is easily checked that the hypothesis of Lemma 4,

$$1 + \lambda - \gamma(\lambda)^2 > 0$$

holds for $|\lambda| \le \beta$ if and only if

$$c < \frac{1-\beta}{1+\beta} \tag{2.37}$$

For these $c > 0$ one can apply Lemma 4 to estimate the spectral norms of the matrices

$$\begin{bmatrix} \alpha\lambda & -\lambda\gamma^{-1}(\lambda)\sqrt{1-\lambda} \\ \alpha\gamma(\lambda)\sqrt{1-\lambda} & \lambda \end{bmatrix}$$

for $|\lambda| \le \beta$. The derivative of the function

$$\begin{aligned} \psi(\lambda) &\equiv \frac{\lambda^2}{\gamma(\lambda)^2(1+\lambda-\gamma(\lambda)^2)} \\ &= \frac{\lambda^2}{(c-c^2)+2c^2\lambda-(c+c^2)\lambda^2} \end{aligned}$$

which is related to the right hand side of inequality (2.29), is negative on the interval

$$1 - c^{-1} < \lambda < 0$$

and nonnegative outside this interval (where it is defined). Therefore, for all $c > 0$ satisfying (2.37), $\psi(\lambda)$ is monotonely decreasing on the interval $-\beta \le \lambda \le 0$ and monotonely increasing on the interval $0 \le \lambda \le \beta$. Because

$$\begin{aligned} \psi(0) &= 0 \\ &< \psi(\beta) = \frac{\beta^2}{(c-c^2)+2c^2\beta-(c+c^2)\beta^2} \\ &< \psi(-\beta) = \frac{\beta^2}{(c-c^2)-2c^2\beta-(c+c^2)\beta^2} \end{aligned}$$

the estimate

$$\psi(\lambda) \le \psi(-\beta) = \frac{\beta^2}{\gamma(-\beta)^2(1-\beta-\gamma(-\beta)^2)} \qquad (|\lambda| \le \beta)$$

follows.

$\psi(-\beta)$ becomes minimal for $\gamma(-\beta)^2 = (1-\beta)/2$, corresponding to the choice $c = \theta$ in (2.30). For this constant c

$$\psi(-\beta) = \left(\frac{2\beta}{1-\beta}\right)^2$$

proving (2.32)-(2.33).

The estimate (2.34)-(2.35) is an immediate consequence of (2.10) and Lemma 2. $\square$

The estimate (2.32) guarantees that

$$\delta < 1 \quad \text{for } \alpha < 1 \text{ and } \beta < 1/3 \tag{2.38}$$

and that

$$\delta \leq \alpha \quad \text{for } \beta \leq \frac{\alpha}{2+\alpha} \tag{2.39}$$

For $\beta \leq \alpha/(2+\alpha)$, ℓ iterations (1.14) reduce the norm (2.5) of the initial error by at least the factor $\mathcal{K}\alpha^\ell$, with $\mathcal{K} = (\kappa_2/\kappa_1)^{1/2}$.

Note that the norm (2.31) is independent of the matrix $\hat{C}$ defining the iteration (1.9), a property which becomes essential in the next section. This was due to our special choice of Γ, which cancelled the apparent dependence of the norm (2.25) on $\hat{C}$, as was shown in (2.36).

In our final theorem of this section, we show that convergence can be forced whenever the inner iterations are convergent by damping the iteration (1.9).

THEOREM 2.6. *Assume that all the eigenvalues λ_i of the error propagation matrix (2.22) are nonnegative. Let the γ_i defining the norm (2.25) in Theorem 3 be given by*

$$\gamma_i = \sqrt{\frac{1+\lambda_i}{2}} \qquad 1 \leq i \leq m \tag{2.40}$$

Then the convergence rate (2.27) can be estimated by

$$\delta \leq \max\left(\alpha,\ \frac{2\beta}{1+\beta}\right) \tag{2.41}$$

The norm (2.25) corresponding to this choice of the γ_i and the norm (2.5) satisfy

$$\kappa_1 \|v\|_*^2 \leq |||v|||^2 \leq \kappa_2 \|v\|_*^2 \tag{2.42}$$

with

$$\frac{\kappa_2}{\kappa_1} = \frac{1+\alpha}{1-\alpha}\ \frac{\max(2, \phi(1-\alpha^2)\mu_2)}{\min(2, (1-\alpha^2)\mu_1)} \tag{2.43}$$

with

$$\phi = \frac{1+\beta}{1-\beta}$$

Proof. The square root on the right hand side of the estimate (2.29) becomes minimal for

$$\gamma = \sqrt{\frac{1+\lambda}{2}}$$

For this choice of γ

$$\frac{\lambda^2}{\gamma^2(1+\lambda-\gamma^2)} = \left(\frac{2\lambda}{1+\lambda}\right)^2$$

so that (2.41) follows from Theorem 3 and Lemma 4.

As shown in the proof of Theorem 5, in (2.36)

$$\|y\|_H = \|\Sigma_0 U^t \hat{C}^{1/2} y\|$$

for $y \in \mathcal{R}^m$. A straightforward calculation then shows

$$\|y\|_H^2 \leq 2\|\Gamma U^t \hat{C}^{1/2} y\|^2 \leq \phi \|y\|_H^2$$

This, along with (2.10) and Lemma 2, imply (2.43). □

Contrary to Theorem 5, Theorem 6 guarantees convergence for all $\beta < 1$; that is, for all converging iterative methods (1.7) and (1.9). With respect to the norm (2.25), with Γ defined as in (2.31), we have

$$\delta \leq \alpha \quad \text{for } \beta \leq \frac{\alpha}{2-\alpha} \tag{2.44}$$

which is an improvement compared with (2.39); however, one should remember that damping may reduce the rate of convergence (2.2), often reversing this effect.

3. Some Examples. In this section we consider the behavior of the iteration (1.14) and the associated norm estimates of Theorems 5 and 6 for some special cases.

We first consider the case $\hat{A} = (1-\alpha)^{-1}A$. This includes the limiting case $\alpha \to 0$. For this case the results of Theorem 5 are essentially sharp. We see this in

THEOREM 3.1. *If $\hat{A} = (1-\alpha)^{-1}A$ and if $\lambda = -\beta$ is an eigenvalue of the error propagation matrix (2.22), the spectral radius of the error propagation matrix (2.15) is greater than α for $\beta > \alpha/(2+\alpha)$ and greater than 1 for $\beta > 1/(1+2\alpha)$.*

Proof. For $\hat{A} = (1-\alpha)^{-1}A$, one has $I - \bar{A} = \alpha I$, so that it follows from the proof of Theorem 3 that the error propagation matrix (2.15) is similar to

$$\begin{bmatrix} \alpha(I - \Sigma^t\Sigma) & \Sigma^t\Sigma\Sigma^t - \Sigma^t \\ \alpha\Sigma & I - \Sigma\Sigma^t \end{bmatrix}$$

Therefore, for this choice of $\hat{A}$, the eigenvalues of the 2×2 matrices

$$\begin{bmatrix} \alpha(1-\sigma^2) & \sigma^3 - \sigma \\ \alpha\sigma & 1 - \sigma^2 \end{bmatrix}$$

for $\sigma = \sigma_1, \ldots, \sigma_m$ are eigenvalues of the matrix (2.15). With $\lambda = 1 - \sigma^2$, the eigenvalues of these 2×2 matrices are

$$\mu = \frac{(1+\alpha)\lambda \pm \sqrt{(1+\alpha)^2\lambda^2 - 4\alpha\lambda}}{2}$$

For $|\lambda| \leq \beta$, the maximum absolute value of μ is

$$\mu = \frac{(1+\alpha)\beta + \sqrt{(1+\alpha)^2\beta^2 + 4\alpha\beta}}{2}$$

attained for $\lambda = -\beta$. This value is bounded by α if and only if

$$\beta \leq \frac{\alpha}{2+\alpha}$$

and is bounded by 1 if and only if

$$\beta \leq \frac{1}{1+2\alpha}$$

□

As a second example, we consider the case $\hat{C} = H = B\hat{A}^{-1}B^t$. This is effectively the limiting case $\beta \to 0$. To analyze this case, we let

$$V^t(I - \bar{A})V = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{12}^t & Q_{22} \end{bmatrix} \tag{3.1}$$

(using the notation of section 2). Here Q_{11} is $m \times m$ and Q_{22} is $(n-m) \times (n-m)$. The error propagation matrix $I - \hat{M}^{-1}M$ is similar to

$$\begin{bmatrix} I - \Sigma_0^2 & 0 & \Sigma_0^3 - \Sigma_0 \\ 0 & \alpha I & 0 \\ \alpha\Sigma_0 & 0 & I - \Sigma_0^2 \end{bmatrix} \begin{bmatrix} Q_{11} & Q_{12} & 0 \\ Q_{12}^t & Q_{22} & 0 \\ 0 & 0 & I \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ Q_{12}^t & Q_{22} & 0 \\ Q_{11} & Q_{12} & 0 \end{bmatrix}$$

where we have used the fact that $\Sigma_0 = I$ for $\hat{C} = H$. Therefore for this case, $(I - \hat{M}^{-1}M)^k,\ k \geq 2$, is similar to

$$\begin{bmatrix} 0 & 0 & 0 \\ Q_{22}^{k-1}Q_{12}^t & Q_{22}^k & 0 \\ Q_{12}Q_{22}^{k-2}Q_{12}^t & Q_{12}Q_{22}^{k-1} & 0 \end{bmatrix}$$

and the rate of convergence is clearly governed by the matrix Q_{22}. Without further assumptions about A and $\hat{A}$, the best norm estimate that we can make is certainly $\|Q_{22}\| \leq \alpha$. On the other hand, since Q_{22} is a proper submatrix of $V^t(I - \bar{A})V$, it is possible that $\|Q_{22}\| < \alpha$. Indeed, we will find convergence rates significantly less than α for the case of the Stokes equations. Even more interesting is the case where $m = n$, hence $V^t(I - \bar{A})V = Q_{11}$. In this case it is easy to see that $(I - \hat{M}^{-1}M)^2 = 0$, so that the iteration (1.14) converges in 2 steps for *any* symmetric, positive definite $\hat{A}$, in effect becoming a direct method.

4. The Preconditioned Conjugate Gradient Method as Inner Iteration.

Every iteration step (1.14) requires the approximate solution of the linear system

$$Hy = c \tag{4.1}$$

Generally, the coefficient matrix $H = B\hat{A}^{-1}B^t$ of this system is only implicitly known. To construct iterative methods (1.9) for solving (4.1) fast, we show that the matrix G used in defining the norm (2.5) is not only a good preconditioner for $C = BA^{-1}B^t$ but also for H. The spectral condition number of $G^{-1/2}HG^{-1/2}$ can be estimated as follows.

Theorem 4.1.

$$\frac{1-\alpha}{1+\alpha}\mathcal{K}(G^{-1/2}CG^{-1/2}) \leq \mathcal{K}(G^{-1/2}HG^{-1/2}) \leq \frac{1+\alpha}{1-\alpha}\mathcal{K}(G^{-1/2}CG^{-1/2}) \tag{4.2}$$

Proof. The proof relies on the fact that for all nonsingular $m \times m$ matrices R and S

$$\mathcal{K}(RR^t) = \mathcal{K}(R)\,\mathcal{K}(R^t) = \mathcal{K}(R^tR)$$

$$\mathcal{K}(R^{-1}) = \mathcal{K}(R)$$

and

$$\mathcal{K}(RS) \leq \mathcal{K}(R)\,\mathcal{K}(S)$$

Therefore

$$\begin{aligned}\mathcal{K}(G^{-1/2}CG^{-1/2}) &\le \mathcal{K}(G^{-1/2}B\hat{A}^{-1/2})\,\mathcal{K}(\hat{A}^{1/2}A^{-1}\hat{A}^{1/2})\,\mathcal{K}(\hat{A}^{-1/2}B^tG^{-1/2})\\ &= \mathcal{K}(\bar{A})\,\mathcal{K}(G^{-1/2}HG^{-1/2})\end{aligned}$$

and

$$\begin{aligned}\mathcal{K}(G^{-1/2}HG^{-1/2}) &\le \mathcal{K}(G^{-1/2}BA^{-1/2})\,\mathcal{K}(A^{1/2}\hat{A}^{-1}A^{1/2})\,\mathcal{K}(A^{-1/2}B^tG^{-1/2})\\ &= \mathcal{K}(\bar{A})\,\mathcal{K}(G^{-1/2}CG^{-1/2})\end{aligned}$$

Since

$$\mathcal{K}(\bar{A}) \le \frac{1+\alpha}{1-\alpha}$$

the theorem follows. ◻

By (2.10),

$$\mathcal{K}(G^{-1/2}CG^{-1/2}) \le \frac{\mu_2}{\mu_1} \tag{4.3}$$

so that for reasonable sizes of α, G and every matrix $\hat{G}$ which is spectrally equivalent to G can serve as a preconditioner for H. For finite element approximations of the Stokes equations, G is associated with the $\mathcal{L}^2$ inner product, so that $\hat{G}$ can be a simple diagonal matrix.

This means that a feasible choice for the iteration (1.9) is

$$y \leftarrow y + \omega\hat{G}^{-1}(c - Hy) \tag{4.4}$$

with a properly chosen $\omega > 0$, or even better, a Chebyshev accelerated version of this iteration. The disadvantage of this approach is that it requires good estimates for the extremal eigenvalues of $\hat{G}^{-1}H$, which are often hard to obtain.

Therefore in this section, we propose to solve every system of the form (4.1) approximately by some steps of the conjugate gradient method, using G or a spectrally equivalent matrix $\hat{G}$ as preconditioner.

Because of

$$\hat{M}^{-1} = \begin{bmatrix} \hat{A}^{-1} & -\hat{A}^{-1}B^T \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & \hat{C}^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ B\hat{A}^{-1} & -I \end{bmatrix} \tag{4.5}$$

this leads to the following algorithm.

Step 1 : *Compute the components*

$$\begin{aligned} r &= f - (Ax + B^Ty)\\ s &= g - Bx \end{aligned}$$

of the current residual.

Step 2 : *Compute*

$$c = B\hat{A}^{-1}r - s$$

Step 3 : *Determine an approximate solution* $\hat{d}$ *of the linear system*

$$Hd = c \quad ; \quad H = B\hat{A}^{-1}B^T$$

Step 4 : *Compute*

$$v = \hat{A}^{-1}(r - B^T\hat{d})$$

and replace the old approximation by

$$\begin{aligned} x &\leftarrow x + v \\ y &\leftarrow y + \hat{d} \end{aligned}$$

In the next lemma, we show that independently of how $\hat{d}$ is computed (by conjugate gradients or some other method), one can view any such algorithm in terms of a single step of an iterative method with a particular choice of matrix $\hat{C} = \hat{C}_i$.

LEMMA 4.2. *Let* H *be a symmetric, positive definite* $m \times m$ *matrix and let* $d, \hat{d} \in \mathcal{R}^m$ *satisfy*

$$\|d - \hat{d}\|_H \leq \beta \|d\|_H \tag{4.6}$$

with $0 \leq \beta < 1$. *Then there exists a symmetric, positive definite matrix* $\hat{C}$ *with*

$$\hat{C}\hat{d} = Hd \tag{4.7}$$

and

$$\|I - \hat{C}^{-1/2}H\hat{C}^{-1/2}\| \leq \beta \tag{4.8}$$

Proof. Set $y = H^{1/2}d$, $\hat{y} = H^{1/2}\hat{d}$, and assume without restriction $\|y\| = 1$ and $m \geq 2$. Let $u_1, u_2, \ldots, u_m$ be an orthonormal basis of $\mathcal{R}^m$ with

$$y = u_1 \qquad \hat{y} = au_1 + bu_2$$

Let U be the $m \times m$ orthogonal matrix with columns $u_1, u_2, \ldots, u_m$. Define the symmetric matrix Q by

$$(U^tQU)_{ij} = q_{ij}$$

with

$$\begin{bmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{bmatrix} = \begin{bmatrix} a & b \\ b & 2-a \end{bmatrix} \tag{4.9}$$

and

$$q_{ij} = \delta_{ij} \quad \text{for } i > 2 \text{ or } j > 2$$

Then by construction

$$Qy = \hat{y} \tag{4.10}$$

Because of

$$(1-a)^2 + b^2 = \|y - \hat{y}\|^2 \le \beta^2 \|y\|^2 = \beta^2$$

and $\beta < 1$, one has $a > 0$, $2 - a > 0$, and $a(2-a) - b^2 > 0$. Therefore the matrices (4.9) and Q are positive definite. Because

$$\|I - Q\|^2 = (1-a)^2 + b^2$$

one obtains

$$\|I - Q\| \le \beta \tag{4.11}$$

Consider

$$\hat{C} \equiv H^{1/2} Q^{-1} H^{1/2}$$

By (4.10), $\hat{C}\hat{d} = Hd$ follows, and by (4.11) one gets

$$\|I - H^{1/2}\hat{C}^{-1}H^{1/2}\| \le \beta$$

As the matrices $H^{1/2}\hat{C}^{-1}H^{1/2}$ and $\hat{C}^{-1/2}H\hat{C}^{-1/2}$ are similar, this proves the lemma. ☐

Applying Theorem 5, the proof of our final result is completed

THEOREM 4.3. *Provided that in Step 3 of the algorithm described above*

$$\|d - \hat{d}\|_H \le \beta \|d\|_H \quad (\beta < 1/3) \tag{4.12}$$

is guaranteed, the norm of (2.31) of the error is reduced by at least the factor

$$\delta = \max\left(\alpha,\ \frac{2\beta}{1-\beta}\right) < 1 \tag{4.13}$$

As a consequence, for $\beta \le \alpha/(2+\alpha)$, ℓ iterations reduce the norm (2.5) of the initial error by at least the factor $\mathcal{K}\alpha^\ell$, with $\mathcal{K} = (\kappa_2/\kappa_1)^{1/2}$ as in (2.35). Therefore, the rate of convergence of the basic iteration (1.7) for solving (1.8) is achieved.

From the practical point of view the condition (4.12) is somewhat subtle because there is no immediate access to the quotient $\|d - \hat{d}\|_H / \|d\|_H$. Alternatively one could fix the number of conjugate gradient steps to determine $\hat{d}$, or one could use the preconditioned residual to test for convergence.

Finally we remark that with a proper rearrangement of the computations one can save the application of $\hat{A}^{-1}$ at the end of every iteration step. This can greatly improve the efficiency of the algorithm. Our operation counts in section 5 take into consideration this fact.

We finish this section with a work estimate. Assume that the application of $\hat{A}^{-1}$ consists of ℓ multigrid steps, for example. The corresponding rate of convergence is

$$\alpha = \eta^\ell$$

for some $\eta \in (0,1)$. By Theorem 8, for small η, the condition number of the corresponding matrix $G^{-1/2}HG^{-1/2}$ is roughly independent of ℓ, and one can assume that the number of preconditioned conjugate gradient steps necessary to compute $\hat{d}$ sufficiently accurate behaves like

$$\tilde{\gamma}|\log\alpha| = \tilde{\gamma}|\log\eta^{\ell}| = \gamma\ell$$

for some positive constants $\tilde{\gamma}$ and γ. Each conjugate gradient step requires ℓ applications of $\hat{A}^{-1}$, so that the total number of multigrid steps per iteration is

$$\gamma\ell^2 + \ell \tag{4.14}$$

The same error reduction by the factor α can be achieved by ℓ iteration cycles with *one* multigrid step to evaluate $A^{-1}x$ approximately. The corresponding number of multigrid steps is

$$\ell(\gamma + 1) \tag{4.15}$$

Obviously, (4.15) is superior to (4.14). Therefore, we conclude that $\ell = 1$ multigrid step is the optimal choice, and that one should not invest too much work to evaluate $A^{-1}x$ very precisely.

5. An Application to a Finite Element Discretization of the Stokes Equations. In this section we consider the mini-element approximation [1] [9] of the following Stokes problem :

$$-\Delta\boldsymbol{u} + \nabla p = \boldsymbol{f} \qquad \text{in } \Omega \tag{5.1}$$

$$\nabla\cdot\boldsymbol{u} = 0 \qquad \text{in } \Omega \tag{5.2}$$

$$\boldsymbol{u} = \boldsymbol{g} \qquad \text{on } \partial\Omega \tag{5.3}$$

$$\int_{\Omega} p\ dx = 0 \tag{5.4}$$

in a bounded domain $\Omega \subset \mathcal{R}^2$. Here $\boldsymbol{u^t} = (u\ v)$ denotes the velocity (two components) and p the pressure. $\boldsymbol{g}$ has to satisfy the compatibility condition $\int_{\partial\Omega}\boldsymbol{g}\cdot\boldsymbol{n}ds = 0$.

In this application the matrix A consists essentially of two copies of the discrete Laplacian corresponding to linear finite elements. Therefore we can choose the hierarchical basis multigrid method [4] to build up the preconditioner $\hat{A}$. G is the mass matrix, so that we can choose the diagonal of G as preconditioner $\hat{G}$ for H.

In our experimental code the constant α is estimated by performing a few iterations of the power method on $I - \hat{A}^{-1}A$ before the outer iteration is initiated. The inner conjugate gradient loop is terminated when the relative size of the residual in pressure has either decreased by a factor $\beta = \alpha/(2-\alpha)$ during the few inner iterations or reached the precision wanted for the global loop.

We now present a numerical example which confirms the theory developed in the previous sections and demonstrates the capability of our solver. We consider the classical case of the driven cavity problem. The system to be solved is the system (5.1)-(5.4) in a unit square. A unit tangential velocity at the top of the square (and 0 elsewhere) is considered. These boundary conditions satisfy the compatibility condition $\int_{\partial\Omega} \boldsymbol{g} \cdot \boldsymbol{n} ds = 0$. It is well known that in this example two singularities develop at the top corners of the square, due to the discontinuity in the boundary conditions.

The square is triangulated using a uniform mesh of isosceles right triangles. The tests were performed using a uniform 5×5 mesh as the level 1 mesh for the multigrid iteration. The number of triangles quadruples from one level to the next, so that the level 5 mesh contains 8192 triangles and 4225 nodes. After static condensation of the "bubble" unknowns this corresponds to 12163 unknowns.

Table 5.1 contains the values of the rates of convergence α and δ for different values of the number of inner iterations. In this table, N denotes the number of vertices in the mesh. The line corresponding to the level 1 mesh gives an idea of the influence of the conjugate gradient iteration since the systems with the matrix A are solved exactly on this level, except for the effects of roundoff error ($\alpha \approx 0$).

level	N	δ_1	δ_2	δ_3	δ_4	δ_∞	α	α/δ_∞
1	25	0.82	0.50	0.22	0.12	$1.2\ 10^{-7}$	$1.2\ 10^{-7}$	–
2	81	0.88	0.54	0.35	0.35	0.347	0.466	1.34
3	289	0.89	0.58	0.51	0.51	0.523	0.638	1.22
4	1089	0.89	0.62	0.61	0.61	0.622	0.731	1.17
5	4225	0.90	0.68	0.68	0.68	0.694	0.792	1.14

Table 5.1. Comparison between the rate of convergence α and the actual rate of convergence δ using a given number of inner iterations.

In table 5.1, δ_i ($i = 1, 2, 3, 4$) was computed using i CG iterations and 200 outer iterations, as δ_∞ was computed using 20 CG iterations and 1000 outer iterations (except for level 1). The outer iteration was the power method

$$v_{j+1} = (I - \hat{M}^{-1} M) v_j$$

(plus appropriate normalization). Because the conjugate gradient inner iteration is not stationary (properly speaking, $\hat{M} = \hat{M}_j$) the values of δ_i and δ_∞ appearing in the table are averages $(\|v_j\|/\|v_0\|)^{1/j}$. These results agree with our theory, i.e. the overall rate of convergence is effectively somewhat smaller than the rate of convergence of the iteration for solving a system with the matrix A. However the ratio α/δ_∞ seems to tend to 1 as the size of the problem increases. It is expected that for large scale problems, like those arising in the industrial environment, the rate of convergence of the global iteration may be close to α.

We can also note in that table that only a few conjugate gradient iterations suffice to achieve this rate of convergence. Also, as the problem size increases, fewer conjugate gradient iterations are needed to achieve the asymptotic rate.

In the next table we present a few results concerning the number of iterations required to achieve convergence under more realistic operating conditions. Here we solve the problem using nested iteration; that is, we sequentially solve all five problems, starting from the 5×5 mesh and working towards the 65×65 mesh. The level 1 problem is solved exactly (to the order of the machine epsilon); for $j > 1$, the level

$j-1$ solution is used as initial guess for the level j problem, and the initial error is reduced by a fixed factor ϵ. We chose $\epsilon = 10^{-2}$, although any factor less than .25 would have been possible. Included are the total number of inner iterations (inner iter), the maximum number of these iterations in all outer iterations (max inner) and the total number of outer iterations (outer iter), at each level. The average number of inner iterations per outer iteration is also listed. The inner conjugate gradient iteration was terminated when the residual norm was reduced by α, or to the precision required by the outer iteration, whichever was achieved first. Normally this results in α governing the termination in the early and middle stages of the calculation, with the precision of the outer iteration becoming dominant in the later stages.

level	inner iter	max inner	outer iter	average
1	31	31	1	31
2	9	3	4	2.25
3	10	3	6	1.67
4	10	3	6	1.67
5	11	3	7	1.57

Table 5.2. Number of inner and outer iterations for the Stokes problem on the 65×65 grid ($\epsilon = 10^{-2}$).

We note that for the level 1 problem $\alpha \approx 0$, so that only 1 outer iteration was used, with comparatively many inner iterations. After the first level, the total number of inner iterations and the number of outer iterations grow slowly, reflecting the dependence of α on $\log N$; see [20] [4].

Since the case of uniform refinement is the most unfavorable for the hierarchical basis multigrid method (in terms of its convergence rate α), we expect that this method will perform comparably or even better on an adaptive, locally refined grids. A posteriori error estimates and adaptively refined grids and solutions will be presented in a forthcoming paper.

REFERENCES

[1] D. N. Arnold, F. Brezzi, and M. Fortin, *A stable finite element for the Stokes equations*, Calcolo, 21 (1984), pp. 337–344.

[2] A. K. Aziz and I. Babuška, *Part I, 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, Academic Press, New York, 1972, pp. 1–362.

[3] R. E. Bank and T. F. Dupont, *Analysis of a Two Level Scheme for Solving Finite Element Equations*, Tech. Rep. CNA-159, Center for Numerical Analysis, University of Texas at Austin, 1980.

[4] R. E. Bank, T. F. Dupont, and H. Yserentant, *The hierarchical basis multigrid method*, Numer. Math., 52 (1988), pp. 427–458.

[5] J. H. Bramble and J. E. Pasciak, *A preconditioning technique for indefinite systems resulting from mixed approximations of elliptic problems*, Math. of Comp., 181 (1988), pp. 1–17.

[6] F. Brezzi, *On the existence, uniqueness and approximation of saddle-point problems arising from Lagrange multipliers*, R.A.I.R.O., 8 (1974), pp. 129–151.

[7] O. Axelsson, *Preconditioning indefinite problems by regularization* , SIAM J. Numer. Anal.,16 (1979), pp. 58–69.

[8] N. Dyn and W. E. Ferguson, *Numerical solution of equality constrained quadratic programming problems*, Math. of Comp., 41 (1983), pp. 165–170.

[9] V. Girault and P. Raviart, *Finite Element Methods for Navier-Stokes Equations*, Springer-Verlag, Berlin, 1986.

[10] G. H. Golub and M. L. Overton, *The convergence of inexact Chebyshev and Richardson iterative methods for solving linear systems*, Numer. Math., 53 (1988), pp. 571–593.

[11] W. Hackbusch, *Multigrid Methods and Applications*, Springer-Verlag, Berlin, 1985.

[12] J. F. Maitre, F. Musy, and P. Nigon, *A fast solver for the Stokes equations using multigrid with a Uzawa smoother*, in Notes on Numerical Fluid Mechanics, Volume 11, Vieweg, Braunschweig, 1985, pp. 77–83.

[13] S. McCormick, ed., *Multigrid Methods*, SIAM, Philadelphia, 1987.

[14] R. Verfürth, *A combined conjugate gradient-multigrid algorithm for the numerical solution of the Stokes problem*, IMA J. Numer. Anal., 4 (1984), pp. 441–455.

[15] ———, *Iterative Methods for the Numerical Solution of Mixed Finite Element Approximations of the Stokes Problem*, Tech. Rep. 379, Institut National de Recherche en Informatique et en Automatique, 1985.

[16] ———, *A multilevel algorithm for mixed problems*, SIAM J. Numer. Anal., 21 (1984), pp. 264–271.

[17] ———, *A multilevel algorithm for mixed problems. II. treatment of the mini-element*, SIAM J. Numer. Anal., 25 (1988), pp. 285–293.

[18] G. Wittum, *Multigrid Methods for Stokes and Navier Stokes Equations. Transforming Smoothers: Algorithms and Numerical Results*, Numer. Math., 54 (1989), pp. 543–563.

[19] ———, *On the Convergence of Multigrid Methods with Transforming Smoothers: Theory with Applications to the Navier-Stokes Equations*, Tech. Rep. 468, Sonderforschungsbereich 123, Universität Heidelberg, 1988.

[20] H. Yserentant, *On the multi-level splitting of finite element spaces*, Numer. Math., 49 (1986), pp. 379–412.