

PRECONDITIONERS FOR GENERALIZED SADDLE-POINT PROBLEMS*

CHRIS SIEFERT[†] AND ERIC DE STURLER[‡]

Abstract. We propose and examine block-diagonal preconditioners and variants of indefinite preconditioners for block two-by-two generalized saddle-point problems. That is, we consider the nonsymmetric, nonsingular case where the (2,2) block is small in norm, and we are particularly concerned with the case where the (1,2) block is different from the transposed (2,1) block. We provide theoretical and experimental analyses of the convergence and eigenvalue distributions of the preconditioned matrices. We also extend the results of [de Sturler and Liesen 2005] to matrices with non-zero (2,2) block and to the use of approximate Schur complements. To demonstrate the effectiveness of these preconditioners we show convergence results, spectra and eigenvalue bounds for two model Navier-Stokes problems.

Key words. saddle point problems, generalized saddle point problems, iterative methods, preconditioning, Krylov subspace methods, eigenvalue bounds

AMS subject classifications. 65F10

1. Introduction. We examine preconditioners for real systems of the form,

$$\mathcal{A} \begin{bmatrix} x \\ y \end{bmatrix} \equiv \begin{bmatrix} A & B^T \\ C & D \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix}, \quad (1.1)$$

where $A \in \mathbb{R}^{n \times n}$, $D \in \mathbb{R}^{m \times m}$, and $n > m$. For many relevant problems, $D = 0$ and $B \neq C$, and such problems are referred to as generalized saddle point problems [24]. For other problems we consider, $D \neq 0$, but $\|D\|_2$ is small enough that the problem retains the characteristics of a generalized saddle-point problem. In many such problems, the non-zero (2,2) block arises from a stabilization term. However, this is not always the case. In a problem involving metal deformation [34], for example, it derives from very slight compressibility. In addition, we note that certain approaches to stabilization lead to systems where $B \neq C$ [3, 24, 26, Sections 7.5 and 9.4], although many other problems have $B = C$. Finally, our preconditioners allow A to be singular. We consider all of these cases, which arise in many applications, ranging from stabilized formulations of the Navier-Stokes equations [4, 11, 26] to metal deformation [34] and interior point methods [13].

Problems of this type have been of recent interest [2, 8, 9, 18, 20, 23], as have their symmetric counterpart [7, 10, 14, 25, 31, 33], and the case where $D = 0$ [1, 5, 6, 8, 15, 19, 21, 23, 30]. However, preconditioners for the case where $B \neq C$ have not received as much attention. Though they are considered in [8, 18, 23], these papers do not provide numerical experiments for such problems. We will do this in the present paper. In [8], a detailed analysis is provided for two classes of preconditioners for the case where $B \neq C$ and $D = 0$. Here, we extend these preconditioners to the case where $D \neq 0$ and to allow for approximations to the Schur complement matrix that arises in the preconditioner. Our preconditioners for (1.1) derive from a matrix

*This work was supported, in part, by the U.S. Department of Energy under grant DOE LLNL B341494 through the Center for the Simulation of Advanced Rockets.

[†]Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801 (siefert@uiuc.edu).

[‡]Department of Mathematics, 460 McBryde, Virginia Tech, Blacksburg, VA 24061-0123 (sturler@vt.edu).

splitting, $A = F - E$. Our purpose is to derive preconditioners that result in tightly clustered eigenvalues. In general, this leads to fast convergence for Krylov subspace methods, although in the nonsymmetric case the eigenvectors may play a role as well.

In this paper we assume that the matrix is non-singular or that the singularity can be easily removed, such as the constant pressure mode in the Oseen problem [12]. For the splitting, we assume that F and $(D - CF^{-1}B^T)$ are nonsingular. In Section 2, we propose a block-diagonal preconditioner that is a generalization of the preconditioners discussed in [18] and [8]. In Section 3, we use this preconditioner to derive a second preconditioned system, which is a generalization of the *related system* presented in [8]. For the $D = 0$ case, the related system corresponds to an efficient implementation of a constraint preconditioner, see also [5, 6, 14, 25]. In Section 4, we extend both types of preconditioner to the use of approximate Schur complements. Our analysis focuses on the $D \neq 0$ case, but we provide specializations to the $D = 0$ case as well. While the block-diagonally preconditioned system may be very effective or more convenient in certain situations, the related system is generally the better preconditioner, offering much faster convergence for a modest increase in the computational cost per iteration. Therefore, in Section 5 on numerical experiments we focus on the related system.

We propose preconditioners with exact (Sections 2 and 3) and with approximate Schur complements (Section 4), and we discuss the convergence for the preconditioned systems and the clustering of the eigenvalues. We explore two model problems in Section 5. The first, which arises from a finite element discretization of the Navier-Stokes equations, has $D \neq 0$ and $A \neq A^T$. The second, which arises from a spectral collocation approach for an incompressible Stokes problem, has $B \neq C$ and $D = 0$. We use eigenvalue bounds and numerical experiments to illustrate that reasonable choices for splittings and approximate Schur complements yield good convergence. Our analysis also illustrates the issues involved in choosing splittings and approximate Schur complements to achieve effective preconditioning. Although eigenvalue bounds are often wide, they nevertheless indicate good eigenvalue clustering for reasonable choices for splittings and approximate Schur complements.

2. Block-Diagonal Preconditioners (exact Schur complement). We consider a splitting of the (1,1) block, $A = F - E$, where F is easy to solve with and $(D - CF^{-1}B^T)^{-1}$ exists. Note that $-(D - CF^{-1}B^T)$ is the Schur complement of the matrix

$$\begin{bmatrix} F & B^T \\ C & D \end{bmatrix}, \quad (2.1)$$

and we will use the phrase *exact Schur complement* to refer to $-(D - CF^{-1}B^T)$. Next, we introduce the following block-diagonal preconditioner as a straightforward generalization of preconditioners in [8, 18],

$$\mathcal{P}(F) = \begin{bmatrix} F^{-1} & 0 \\ 0 & -(D - CF^{-1}B^T)^{-1} \end{bmatrix}. \quad (2.2)$$

Preconditioning from the left or the right with \mathcal{P} yields a system of the form

$$\mathcal{B}(F) \begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix} = \begin{bmatrix} I - S & N \\ M & Q \end{bmatrix} \begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix} = \begin{bmatrix} \tilde{f} \\ \tilde{g} \end{bmatrix}, \quad (2.3)$$

where $\mathcal{B}(F)$ is either \mathcal{PA} or \mathcal{AP} . For example, the matrix from the left-preconditioned system is

$$\mathcal{P}(F)\mathcal{A} = \begin{bmatrix} I - F^{-1}E & F^{-1}B^T \\ -(D - CF^{-1}B^T)^{-1}C & -(D - CF^{-1}B^T)^{-1}D \end{bmatrix},$$

implicitly defining S , N , M and Q in (2.3) for the left-preconditioned case. Apart from the preconditioned (2,2) block Q , this resembles the system arising from the zero (2,2) block case. For the rest of this paper, we assume that Q is diagonalizable. While $MN = I$ for the $D = 0$ case [23, 8], for $D \neq 0$ we have

$$\begin{aligned} MN &= -(D - CF^{-1}B^T)^{-1}CF^{-1}B^T = -(D - CF^{-1}B^T)^{-1}(-D + CF^{-1}B^T + D) \\ &= I + Q. \end{aligned} \tag{2.4}$$

This is true for both the left and right-preconditioned cases. In the $D = 0$ case NM is a projector [8]. For the $D \neq 0$ case, it is not, as $(NM)^2 = NM + NQM$.

In Section 2.1 we derive the eigendecomposition of the matrix

$$\mathcal{B}_0 = \begin{bmatrix} I & N \\ M & Q \end{bmatrix}, \tag{2.5}$$

when $I + Q$ (and thus B^T and C) have full-rank. We use this in Section 2.2 to develop bounds for the eigenvalues of $\mathcal{B}(F)$ using perturbation theory. Finally, in Section 2.3, we discuss the case when $I + Q$ is rank-deficient.

2.1. Eigenvalues and Eigenvectors of \mathcal{B}_0 . Assume that $I + Q$ (and thus B^T and C) have full rank. We wish to find λ , u and v such that

$$u + Nv = \lambda u \tag{2.6}$$

$$Mu + Qv = \lambda v. \tag{2.7}$$

First, we assume $\lambda = 1$. Substituting this into (2.6) and using $Q = MN - I$ in (2.7) yields

$$Nv = 0 \text{ and } Mu = 2v. \tag{2.8}$$

Since B^T has full column rank by assumption, this implies that $v = 0$, and that \mathcal{B}_0 has only eigenpairs of the form

$$\left(1, \begin{bmatrix} u \\ 0 \end{bmatrix}\right), \quad \text{where } u \in \text{null}(M). \tag{2.9}$$

Since C has full row rank, so does M , and \mathcal{B}_0 has precisely $n - m$ distinct eigenpairs of this type. Next, we consider the case where $\lambda \neq 1$. Solving (2.6) for u , and substituting into (2.7) yields

$$\lambda Qv_j = (\lambda^2 - \lambda - 1)v_j. \tag{2.10}$$

Hence, the v_j must be eigenvectors of Q . We have assumed that Q has a full set of eigenpairs, $Qv_j = \delta_j v_j$, for $j = 1 \dots m$. We then solve (2.10) for λ to yield:

$$\lambda_j^\pm = \frac{(1 + \delta_j) \pm \sqrt{4 + (1 + \delta_j)^2}}{2}, \tag{2.11}$$

cf. [11]. Using (2.6) with the eigenvectors of Q for v yields the vectors u . We finally rescale the eigenvector by $(\lambda_j^\pm - 1)$ to yield eigenpairs of the form

$$\left(\lambda_j^\pm, \begin{bmatrix} Nv_j \\ (\lambda_j^\pm - 1)v_j \end{bmatrix} \right). \quad (2.12)$$

Note that $\lambda_j^- \neq 1$ regardless of the choice of δ_j , and $\lambda_j^+ = 1$ only if $\delta_j = -1$. However, the latter would contradict the assumption that $I + Q$ has full rank. Thus, \mathcal{B}_0 has $2m$ eigenpairs corresponding to $\lambda \neq 1$. This completes a full set of eigenpairs for \mathcal{B}_0 . Let U_1 be a matrix whose columns form an orthonormal basis for $\text{null}(M)$, cf. (2.9), and let U_2 be the matrix with normalized columns $u_j = Nv_j$, where $Qv_j = \delta_j v_j$, cf. (2.12). Furthermore, let $\Lambda^+ = \text{diag}(\lambda_j^+)$ and $\Lambda^- = \text{diag}(\lambda_j^-)$, where $\text{diag}(\cdot)$ denotes the diagonal matrix with the given arguments. Then, the following matrix, \mathcal{Y} , is an eigenvector matrix of \mathcal{B}_0 .

$$\mathcal{Y} \equiv \left[\begin{array}{c|c} Y_{11} & Y_{12} \\ \hline Y_{21} & Y_{22} \end{array} \right] = \left[\begin{array}{c|c|c} U_1 & U_2 & U_2 \\ \hline 0 & V(\Lambda^+ - I) & V(\Lambda^- - I) \end{array} \right]. \quad (2.13)$$

For our perturbation results we also need

$$\mathcal{Z} = \mathcal{Y}^{-1} = \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix}. \quad (2.14)$$

Using the block-inversion formula in [17, Section 0.7.3] we obtain [28, 29],

$$Z_{11} = \begin{bmatrix} I_{n-m} & 0 \\ 0 & \Upsilon^+ \end{bmatrix} Y_{11}^{-1} = \hat{I}_n Y_{11}^{-1}, \quad (2.15)$$

$$Z_{21} = - \begin{bmatrix} 0 & \Upsilon^- \end{bmatrix} Y_{11}^{-1} \quad (2.16)$$

$$Z_{12} = - \begin{bmatrix} 0 \\ (\Lambda^- - \Lambda^+)^{-1} V^{-1} \end{bmatrix} \quad (2.17)$$

$$Z_{22} = (V(\Lambda^- - \Lambda^+))^{-1}, \quad (2.18)$$

with $\Upsilon^+ = \text{diag}((\lambda_j^- - 1)/(\lambda_j^- - \lambda_j^+))$ and $\Upsilon^- = \text{diag}((\lambda_j^+ - 1)/(\lambda_j^- - \lambda_j^+))$. For $Q = 0$ (because $D = 0$), the eigendecomposition of \mathcal{B}_0 reduces to the case discussed in [8].

2.2. Perturbation Bounds on the Eigenvalues of $\mathcal{B}(F)$. We are now ready to consider the eigenvalues of $\mathcal{B}(F)$ and derive bounds on the spectrum. Throughout this paper $\|\cdot\|$ indicates the 2-norm.

THEOREM 2.1. *Consider matrices $\mathcal{B}(F)$ of the form (2.3). Let \mathcal{Y} be the eigenvector matrix of \mathcal{B}_0 , as given by (2.13). Then for each eigenvalue $\lambda_{\mathcal{B}}$ of $\mathcal{B}(F)$, there exists an eigenvalue λ of \mathcal{B}_0 , such that*

$$|\lambda_{\mathcal{B}} - \lambda| \leq \left\| \mathcal{Y}^{-1} \begin{bmatrix} S & 0 \\ 0 & 0 \end{bmatrix} \mathcal{Y} \right\| \quad (2.19)$$

$$\leq 2 \max(1, \|\Upsilon^+\|, \|\Upsilon^-\|) \|Y_{11}^{-1} S Y_{11}\|. \quad (2.20)$$

Proof. Since \mathcal{B}_0 is diagonalizable, (2.19) follows from a classic result in perturbation theory [32, Theorem IV.1.12]. We expand the right-hand-side of (2.19) using

(2.13)–(2.17) to get (see also [8])

$$\begin{aligned} |\lambda_{\mathcal{B}} - \lambda| &\leq \left\| \begin{bmatrix} \hat{I}_n Y_{11}^{-1} S Y_{11} & \hat{I}_n Y_{11}^{-1} S Y_{12} \\ -[0 \ \Upsilon^-] Y_{11}^{-1} S Y_{11} & -[0 \ \Upsilon^-] Y_{11}^{-1} S Y_{12} \end{bmatrix} \right\| \\ &\leq \max(1, \|\Upsilon^+\|, \|\Upsilon^-\|) \cdot \\ &\quad \left\| \begin{bmatrix} Y_{11}^{-1} S U_1 & Y_{11}^{-1} S U_2 & Y_{11}^{-1} S U_2 \\ -[0 \ I] Y_{11}^{-1} S U_1 & -[0 \ I] Y_{11}^{-1} S U_2 & -[0 \ I] Y_{11}^{-1} S U_2 \end{bmatrix} \right\|. \end{aligned}$$

Using the consistency of the 2-norm we can simplify this to (see also [8]):

$$\begin{aligned} |\lambda_{\mathcal{B}} - \lambda| &\leq \sqrt{2} \max(1, \|\Upsilon^+\|, \|\Upsilon^-\|) \left\| \begin{bmatrix} Y_{11}^{-1} S Y_{11} \\ -[0 \ I] Y_{11}^{-1} S Y_{11} \end{bmatrix} \right\| \\ &\leq 2 \max(1, \|\Upsilon^+\|, \|\Upsilon^-\|) \|Y_{11}^{-1} S Y_{11}\|. \end{aligned}$$

□

The Υ^\pm terms can only be large if $\delta_j \approx -1 \pm 2i$. For the problems discussed in Section 5, the δ_j 's are well-separated from this value, because $\|D\|$ is small and the problem and preconditioner are relatively well-conditioned. The following lemma provides bounds on the $\|\Upsilon^\pm\|$. We explicitly consider the special case where the δ_j 's are real (and thus bounded away from $-1 \pm 2i$). This occurs in the important case that D is symmetric and the Schur complement is definite. For the following proof and subsequent discussions, we define the function $p(z) = 4 + (1 + z)^2$.

LEMMA 2.2. *Let Υ^+ and Υ^- be defined as above.*

1. *If $\delta_j \in \mathbb{R}$, for all j , then*

$$\max(1, \|\Upsilon^+\|, \|\Upsilon^-\|) \leq \frac{1 + \sqrt{2}}{2}.$$

Moreover, if $\delta_j \geq -1$, for all j , then $\max(1, \|\Upsilon^+\|, \|\Upsilon^-\|) = 1$.

2. *If $\delta_j \in \mathbb{C}$ and $\exists \alpha : |\delta_j| \leq \alpha < \sqrt{5}$, for $j = 1, \dots, m$, then*

$$\max(1, \|\Upsilon^+\|, \|\Upsilon^-\|) \leq \max \left(1, \frac{1}{2} + \frac{1 + \alpha}{2\sqrt{2(\sqrt{5} - \alpha)}} \right).$$

Proof. Substituting λ_j^\pm from (2.11) in $\Upsilon^+ = \text{diag}(\lambda_j^- - 1)/(\lambda_j^- - \lambda_j^+)$ and $\Upsilon^- = \text{diag}(\lambda_j^+ - 1)/(\lambda_j^- - \lambda_j^+)$ gives

$$\Upsilon^\pm = \text{diag} \left(\frac{1 - \delta_j}{2\sqrt{4 + (1 + \delta_j)^2}} \pm \frac{1}{2} \right) = \text{diag} \left(\frac{1 - \delta_j}{2\sqrt{p(\delta_j)}} \pm \frac{1}{2} \right). \quad (2.21)$$

The proof for the real case now follows from basic calculus.

For the complex case, note that $p(\delta) = (\delta + 1 + 2i)(\delta + 1 - 2i)$. Any δ must be at least distance 2 from one of the roots of $p(\delta)$. We assume without loss of generality that δ is near $-1 + 2i$. The value $\delta_* = (-1 + 2i)\alpha/\sqrt{5}$ minimizes $|\delta + 1 - 2i|$ subject to $|\delta| \leq \alpha$, and we have $|\delta_* + 1 - 2i| = \sqrt{5} - \alpha$. So, we have $|p(\delta)| \geq 2(\sqrt{5} - \alpha)$. Using this inequality for $|p(\delta)|$ after taking norms in (2.21) completes the proof. □

In practice, the bound for the complex case is quite modest. For example, if $|\delta_j| \leq 1$, for all j , then our bound on $\max(1, \|\Upsilon^+\|, \|\Upsilon^-\|)$ is about 1.136. Likewise, if $|\delta_j| \leq 2$, for all j , the bound is about 1.470.

We derive a bound on $\|Y_{11}^{-1}SY_{11}\|$ following the approach in [8]. Recall that $Y_{11} = [U_1 \ U_2]$, where $U_1^T U_1 = I$, and $U_2 = NV$ with unit columns. Let $U_2 = V_2 \Theta$, where $V_2^T V_2 = I$. Furthermore, let $\omega_1 = \|U_1^T V_2\|$, which is the cosine of the smallest principal angle between $\text{range}(U_1) = \text{null}(NM)$ and $\text{range}(U_2) = \text{range}(NM)$.

LEMMA 2.3. *Define Y_{11} , S , U_1 , U_2 , V_2 , Θ , and ω_1 as above, and let $\kappa(\cdot)$ denote the 2-norm condition number. Then,*

$$\|Y_{11}^{-1}SY_{11}\| \leq \kappa(\Theta) \left(\frac{1 + \omega_1}{1 - \omega_1} \right)^{1/2} \|S\|. \quad (2.22)$$

Proof. We have $\|Y_{11}^{-1}SY_{11}\| \leq \kappa(Y_{11})\|S\|$, where

$$Y_{11} = \begin{bmatrix} U_1 & V_2 \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & \Theta \end{bmatrix}.$$

Since U_2 has unit columns, $\|\Theta\| \geq 1$ and $\|\Theta^{-1}\| \geq 1$. So, our bound simplifies to

$$\|Y_{11}^{-1}SY_{11}\| \leq \kappa(\Theta) \kappa \left(\begin{bmatrix} U_1 & V_2 \end{bmatrix} \right) \|S\| \leq \kappa(\Theta) \left(\frac{1 + \omega_1}{1 - \omega_1} \right)^{1/2} \|S\|, \quad (2.23)$$

where the second inequality follows from the bound on $\kappa([U_1 \ V_2])$ from Lemma 3.6 in [8]. \square

COROLLARY 2.4. *Let Θ and ω_1 be defined as above.*

1. *If $\delta_j \in \mathbb{R}$, for all j , then*

$$|\lambda_B - \lambda| \leq (1 + \sqrt{2})\kappa(\Theta) \left(\frac{1 + \omega_1}{1 - \omega_1} \right)^{1/2} \|S\|. \quad (2.24)$$

2. *If $\delta_j \in \mathbb{C}$ and $\exists \alpha : |\delta_j| \leq \alpha < \sqrt{5}$, for $j = 1, \dots, m$, then*

$$|\lambda_B - \lambda| \leq 2 \max \left(1, \frac{1}{2} + \frac{1 + \alpha}{2\sqrt{2}(\sqrt{5} - \alpha)} \right) \kappa(\Theta) \left(\frac{1 + \omega_1}{1 - \omega_1} \right)^{1/2} \|S\|. \quad (2.25)$$

Proof. Use Lemmas 2.2 and 2.3 in Theorem 2.1. \square

We see that the clustering of the eigenvalues depends mainly on $\|S\|$ and the size of the δ_j , unless $\omega_1 \approx 1$ or $\kappa(\Theta)$ large. This implies that the block-diagonally preconditioned system can have as many as $2m + 1$ eigenvalue clusters, one for $\lambda = 1$ and one for each λ_j^\pm . Hence, the convergence of Krylov methods may not be very good for the block-diagonally preconditioned system, even if $\|S\|$ is small. Examples in Section 5 will illustrate this. However, when the δ_j and $\|S\|$ are small, the block-diagonal preconditioner will give good convergence. This typically happens for small mesh width when D and Q are h -dependent; see Table 5.1. In addition, the block-diagonal preconditioner provides an intermediate step to a better preconditioner described in Section 3.

2.3. Rank-Deficiency of $I + Q$. In Section 2.1, we made the assumption that $I + Q$ has full rank (for $D = 0$, this is always true). We now briefly discuss the rank-deficient case.

There are three sources of potential rank-deficiency in $I + Q$. The first two are rank-deficiency in C and B^T . The third is when there are vectors v such that $Nv \neq 0$

and $Nv \in \text{null}(M)$. This implies that $MNv = (I + Q)v = 0$ and v is an eigenvector of Q . This case occurs when F^{-1} (for left preconditioning) or $-(D - CF^{-1}B^T)^{-1}$ (for right preconditioning) maps a non-trivial vector from $\text{range}(B^T)$ into $\text{null}(C)$.

Assume that $I + Q$, C and B^T are rank deficient by k , l_c and l_b respectively. Note that $k \geq \max(l_b, l_c)$, since $I + Q = -(D - CF^{-1}B^T)^{-1}CF^{-1}B^T$ and the product of matrices cannot be of higher rank than any of its factors.

Our previous analysis remains valid for the $2(m - k)$ eigenpairs (2.11) that correspond to $\delta_j \neq -1$. It is also valid for the k eigenpairs where $\delta_j = -1$ that correspond to λ_j^- . Since the Schur complement is invertible, M must also be rank deficient by l_c . Thus, the number of eigenpairs of the form (2.9) equals $\dim(\text{null}(M)) = n - m + l_c$. This gives a total of $n + m - k + l_c$ eigenpairs, leaving us to find $k - l_c$ eigenpairs.

From (2.8), we have that all eigenvectors corresponding to $\lambda = 1$ must satisfy $Nv = 0$ and $Mu = 2v$. Since $\dim(\text{null}(N)) = l_b$, there are l_b independent vectors v that satisfy $Nv = 0$. Unfortunately, there may be as many as l_c independent vectors v where $Mu = 2v$ has no solution. If we do not have $k - l_c$ independent vectors v such that $Mu = 2v$ has a solution, then \mathcal{B}_0 is defective. The analysis of Section 2.1 does not permit any other eigenvectors.

For the missing eigenpairs we have that $\lambda_j^+ \rightarrow 1$ as $\delta_j \rightarrow -1$. Therefore, we look for principal vectors of grade two (see [16]) for $\lambda = 1$. These vectors satisfy the equations

$$Nv = \tilde{u} \quad \text{and} \quad Mu = 2v, \quad (2.26)$$

where $\tilde{u} \neq 0$ and $\tilde{u} \in \text{null}(M)$. We note that there are k independent vectors v such that $(I + Q)v = 0$. Since there are precisely l_b independent vectors v such that $Nv = 0$, there must be $k - l_b$ such vectors v that satisfy $Nv = \tilde{u}$ with $\tilde{u} \neq 0$ and $M\tilde{u} = 0$. This gives k independent vectors v that satisfy the first equation of either (2.8) or (2.26).

There exists a space of dimension l_c , such that $Mu = 2v$ has no solution. However, since we have k independent v 's to propose, we are guaranteed to find $k - l_c$ independent vectors v 's that satisfy this equation. This gives us either our remaining eigenvectors or principal vectors of grade two. This also guarantees us that we have Jordan blocks of size at most two.

In the special case when $k = l_b = l_c$, we have $k - l_c = 0$, so we have a full set of eigenvectors. We can apply the analysis described in the full rank case with k additional eigenpairs $(1, [\tilde{u}_{n-m+j}^T, 0^T]^T)$, for $j = 1 \dots k$, replacing the corresponding eigenpairs $(\lambda_j^+, [(Nv_j)^T, (\lambda_j^+ - 1)v_j^T]^T)$ for which $\delta_j = -1$. Let U_1 be such that $U_1^T U_1 = I_{n-m+l_c}$ and $\text{range}(U_1) = \text{null}(M)$. Let \tilde{V} be such that $\tilde{V}^T \tilde{V} = I_{l_c}$ and $\text{range}(\tilde{V}) = \text{null}(I + Q)$. Further, let the columns of \hat{V} be the eigenvectors of Q corresponding to the eigenvalues $\delta_j \neq -1$, scaled such that $U_2 = N\hat{V}$ has unit columns. Finally, let the diagonal matrices $\hat{\Lambda}^+$ and $\hat{\Lambda}^-$ contain the eigenvalues λ_j^+ and λ_j^- corresponding to the eigenvalues $\delta_j \neq -1$ ordered consistently with the columns of \hat{V} . Then the eigenvector matrix of \mathcal{B}_0 is given by

$$\mathcal{Y} = \left[\begin{array}{cc|cc} U_1^{(n-m+l_c)} & U_2^{(m-l_c)} & N\tilde{V}^{(l_c)} & U_2^{(m-l_c)} \\ 0 & \hat{V}(\hat{\Lambda}^+ - I) & -2\tilde{V} & \hat{V}(\hat{\Lambda}^- - I) \end{array} \right], \quad (2.27)$$

where superscripts in the top row indicate the number of columns. The corresponding eigenvalues are those from (2.9) and (2.11). We can then use the eigenvector matrix

of \mathcal{B}_0 given in (2.27) to derive bounds on the eigenvalues, as for the full rank case. The reduction in the number of columns of U_2 may in fact reduce the factor $\kappa(\Theta)$ in Corollary 2.4. An important example of this case is the stabilized Navier-Stokes (Oseen) problem [11], where $C = B$ and F is positive definite.

3. Fixed Point Method and its Related System (exact Schur complement). We now consider an alternative solution method that leads to faster convergence in general, cf. [8]. In the $D = 0$ case this approach leads to an efficient implementation of so-called constraint preconditioners, cf. [5, 6, 14, 25]. We can derive the following splitting from (2.3),

$$\mathcal{B}(F) \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} I - S & N \\ M & Q \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \left(\mathcal{B}_0 - \begin{bmatrix} S & 0 \\ 0 & 0 \end{bmatrix} \right) \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \tilde{f} \\ \tilde{g} \end{bmatrix}. \quad (3.1)$$

Note that

$$\mathcal{B}_0^{-1} = \begin{bmatrix} I - NM & N \\ M & -I \end{bmatrix}. \quad (3.2)$$

We left-multiply (3.1) by \mathcal{B}_0^{-1} to yield the fixed point iteration,

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} (I - NM)S & 0 \\ MS & 0 \end{bmatrix} \begin{bmatrix} x_k \\ y_k \end{bmatrix} + \begin{bmatrix} \hat{f} \\ \hat{g} \end{bmatrix}. \quad (3.3)$$

Note that this iteration is formally the same as for the $D = 0$ case in [6, 8]. Since x_{k+1} and y_{k+1} depend only on x_k , we need to iterate only on the x_k variables; see also [4, pp. 214–215] and [8]. The x -component of the fixed point of (3.3) satisfies the so-called *related system* for the fixed-point iteration [16],

$$(I - (I - NM)S)x = \hat{f}.^1 \quad (3.4)$$

The full-size related system (that is, with the y component) and $D \neq 0$ has been examined elsewhere for special cases. In [25], A is symmetric positive definite and spectrally equivalent to the identity, and so the splitting $F = I$ is used. In [14], F is symmetric positive definite. In both of these cases $B = C$.

3.1. Eigenvalue Bounds for Fixed Point Matrix and Related System. In this section we assume $n - m \geq m$, but equivalent results are obtained for $m > n - m$. Let U_1 and U_2 be defined as in (2.13), $\Delta = \text{diag}(\delta_j)$ and let $U_2 = V_2\Theta$, with $V_2^T V_2 = I$. Then, we have $NMU_1 = 0$, $NMU_2 = NNMNV = NV(I + \Delta)$, and therefore

$$(I - NM) = \begin{bmatrix} U_1 & V_2 \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & -\Theta\Delta\Theta^{-1} \end{bmatrix} \begin{bmatrix} U_1 & V_2 \end{bmatrix}^{-1}. \quad (3.5)$$

In the rank-deficient case, we can use (2.27). So, for this approach rank-deficiency has a potential advantage in terms of the conditioning of Θ . To analyze $\|I - NM\|$ we need the following singular value decomposition (SVD),

$$U_1^T V_2 = \Phi\Omega\Psi^T, \text{ where } 1 > \omega_1 \geq \omega_2 \geq \dots \geq \omega_m. \quad (3.6)$$

¹The full-size related system derives from using (2.1) as a left preconditioner; see also [8].

Following [8], we define W by $W\Sigma = V_2\Psi - U_1\Phi\Omega$, where the diagonal matrix $\Sigma = \text{diag}((1 - \omega_j^2)^{1/2})$ contains the sines of the principal angles between $\text{range}(U_1)$ and $\text{range}(V_2)$. Then, $[U_1 \ W]$ is orthogonal, and we can decompose V_2 as follows,

$$V_2 = U_1\Phi\Omega\Psi^T + W\Sigma\Psi^T. \quad (3.7)$$

THEOREM 3.1. *Let U_1, V_2 and ω_1 be defined as above. Let λ_R be an eigenvalue of the related system matrix in (3.4). Then,*

$$\left. \begin{array}{l} \rho((I - NM)S) \\ |1 - \lambda_R| \end{array} \right\} \leq (1 - \omega_1^2)^{-1/2}(1 + \|\Theta\Delta\Theta^{-1}\|)\|S\|.$$

where $\rho(\cdot)$ designates the spectral radius.

Proof. The proof of this theorem largely follows [8]. Note that the result for $\rho((I - NM)S)$ immediately implies the result for $|1 - \lambda_R|$. We have $\rho((I - NM)S) \leq \|I - NM\|\|S\|$. Let $Z = -\Theta\Delta\Theta^{-1}$. Then,

$$\|I - NM\| = \left\| [U_1 \ V_2] \begin{bmatrix} I & 0 \\ 0 & Z \end{bmatrix} [U_1 \ V_2]^{-1} \right\| \quad (3.8)$$

$$\leq \left\| [U_1 \ V_2] \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} [U_1 \ V_2]^{-1} \right\| + \left\| [U_1 \ V_2] \begin{bmatrix} 0 & 0 \\ 0 & Z \end{bmatrix} [U_1 \ V_2]^{-1} \right\| \quad (3.9)$$

$$\leq (1 - \omega_1^2)^{-1/2} + (1 - \omega_1^2)^{-1/2}\|Z\| = (1 - \omega_1^2)^{-1/2}(1 + \|Z\|). \quad (3.10)$$

The first term in (3.9) is the norm of an oblique projection. Given the SVD in (3.6), this norm equals $(1 - \omega_1^2)^{-1/2}$ [22, Section 5.15]. We establish the bound for the second term as follows.

$$\left\| [U_1 \ V_2] \begin{bmatrix} 0 & 0 \\ 0 & Z \end{bmatrix} [U_1 \ V_2]^{-1} \right\| = \max_{U_1 a + V_2 b \neq 0} \frac{\|V_2 Z b\|}{\|U_1 a + V_2 b\|}. \quad (3.11)$$

Without loss of generality we may assume $\|b\| = 1$, so that $\|V_2 Z b\| \leq \|Z\|$. From (3.7) we see that $\|U_1 a + V_2 b\| = \|U_1 a + U_1\Phi\Omega\Psi^T b + W\Sigma\Psi^T b\|$, which for any given b is minimized by $a = -\Phi\Omega\Psi^T b$. This gives $\|U_1 a + V_2 b\| = \|W\Sigma\Psi^T b\|$, which in turn is minimized for $b = \psi_1$. Hence, we have

$$\left\| [U_1 \ V_2] \begin{bmatrix} 0 & 0 \\ 0 & Z \end{bmatrix} [U_1 \ V_2]^{-1} \right\| = \max_{U_1 a + V_2 b \neq 0} \frac{\|V_2 Z b\|}{\|U_1 a + V_2 b\|} \leq (1 - \omega_1^2)^{-1/2}\|Z\|. \quad (3.12)$$

So, using (3.8)–(3.12) we have $\rho((I - NM)S) \leq (1 - \omega_1^2)^{-1/2}(1 + \|\Theta\Delta\Theta^{-1}\|)\|S\|$. \square

If the δ_j are clustered, the influence of $\kappa(\Theta)$ is small.

COROLLARY 3.2. *Let $\hat{\delta} = \arg \min_{z \in \mathbb{C}} \max_j |z - \delta_j|$ and $\tilde{\delta}_j = \delta_j - \hat{\delta}$. Then*

$$\left. \begin{array}{l} \rho((I - NM)S) \\ |1 - \lambda_R| \end{array} \right\} \leq (1 - \omega_1^2)^{-1/2}(1 + \hat{\delta} + \kappa(\Theta) \max |\tilde{\delta}_j|)\|S\|.$$

Proof. Note that $\Delta = \hat{\delta}I + \text{diag}(\tilde{\delta}_j)$, so $\Theta\Delta\Theta^{-1} = \hat{\delta}I + \Theta \text{diag}(\tilde{\delta}_j) \Theta^{-1}$. \square

So, the eigenvalues of the related system cluster around 1, and the tightness of the clustering is controlled through $\|S\|$. Note that the factor containing ω_1 in Corollary 3.2 is no larger than the corresponding factor for the block-diagonally preconditioned system in Corollary 2.4. In addition, the influence of the $\kappa(\Theta)$ term is smaller for the related system if the δ_j are clustered. This generally leads to better clustering and tighter bounds for the related system than for the block-diagonally preconditioned system. Because of these advantages, the related system will generally have faster convergence than the block-diagonally preconditioned system.

3.2. Satisfying ‘Constraints’. In the $D = 0$ case, the second block of equations in (1.1) often represents a set of constraints. For the $D \neq 0$ case, this may or may not be the case. So-called constraint preconditioners in the $D = 0$ case have the advantage that each iterate of a Krylov subspace method for the preconditioned system satisfies the constraints, if the initial guess is chosen appropriately. Fixed point methods such as (3.3) often satisfy the constraints after a single step. This is the case for the fixed-point method proposed in [8] for $D = 0$. It turns out that we can prove an analogous property for the $D \neq 0$ case.

LEMMA 3.3. *For any initial guess $[x_0^T, y_0^T]^T$, the iterates, $[x_k^T, y_k^T]^T$, for $k = 1, 2, \dots$, of (3.3) satisfy $Mx_k + Qy_k = \tilde{g}$ in (2.3) and $Cx_k + Dy_k = g$ in (1.1).*

The proof can be found in [28, 29].

COROLLARY 3.4. *After the first iteration of (3.3), all fixed-point updates are in the null space of $[M \ Q]$. This follows trivially from Lemma 3.3.*

We can also show that the iterates of a Krylov subspace method will satisfy the constraints if the initial guess satisfies the constraints (cf. [8]). We first give a general result and then specialize it to our problem. For the remainder of this section, A and C are arbitrary matrices not the matrices referred to in (1.1). We return to the nomenclature of (1.1) in the next section.

THEOREM 3.5. *Let $A \in \mathbb{R}^{n \times n}$, $b \in \mathbb{R}^n$, $C \in \mathbb{R}^{m \times n}$, and $d \in \mathbb{R}^m$, and define the iteration $x_{k+1} = Ax_k + b$. Further, let the iterates x_k satisfy $Cx_k = d$ for $k \geq 1$ and any starting vector x_0 . Then, the iterates $x^{(m)}$, $m \geq 0$, of a Krylov method applied to the (related) system, $(I - A)x = b$, will satisfy $Cx^{(m)} = d$ if $Cx^{(0)} = d$.*

The proof can be found in [28, 29].

COROLLARY 3.6. *The iterates, $[x^{(m)T}, y^{(m)T}]^T$, of any Krylov method applied to the full $n+m$ related system for (3.3) satisfy $Mx^{(m)} + Qy^{(m)} = \tilde{g}$ and $Cx^{(m)} + Dy^{(m)} = g$ if the initial guess is the result of at least one step of fixed point iteration (3.3).*

Proof. Use Theorem 3.5, with A as fixed-point iteration matrix in (3.3), $b = [\hat{f}^T \ \hat{g}^T]^T$, $C = [M \ Q]$ and $d = \hat{g}$. \square

4. Approximate Schur Complement. It may be expensive to compute the Schur complement matrix $(D - CF^{-1}B^T)$ or to compute and apply its inverse (or factors). So, we would like to use a cheap approximation to the inverse of the Schur complement. We now consider the effect of such an approximation on the eigenvalue clustering of the preconditioned matrices and on the resulting convergence. Let $S_1 = -(D - CF^{-1}B^T)$ denote the actual Schur complement and S_2^{-1} denote our approximation to its inverse. As we only need to apply S_2^{-1} , no explicit representation of S_2 is needed. Finally, let $S_2^{-1}S_1 = I + \mathcal{E}$.

4.1. Eigenvalue Analysis of the Block-Diagonally Preconditioned System. Now, the block diagonal preconditioner looks as follows,

$$\mathcal{P}(F, S_2) = \begin{bmatrix} F^{-1} & 0 \\ 0 & S_2^{-1} \end{bmatrix}.$$

We multiply (1.1) from the left by $\mathcal{P}(F, S_2)$. We refer to the resulting preconditioned matrix as $\mathcal{B}(F, S_2)$. The system of equations with $\mathcal{B}(F, S_2)$ looks as follows,

$$\begin{bmatrix} I - S & N \\ M_2 & Q_2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \left(\begin{bmatrix} I & N \\ M & Q \end{bmatrix} - \begin{bmatrix} S & 0 \\ -\mathcal{E}M & -\mathcal{E}Q \end{bmatrix} \right) \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \tilde{f} \\ \tilde{g} \end{bmatrix}, \quad (4.1)$$

where M , N and Q are defined as in Section 2, $M_2 = S_2^{-1}C$ and $Q_2 = S_2^{-1}D$. Note also that $M_2 = S_2^{-1}S_1S_1^{-1}C = (I + \mathcal{E})M$ and analogously $Q_2 = (I + \mathcal{E})Q$. Using

(4.1), we can bound the eigenvalues of $\mathcal{B}(F, S_2)$ by considering the perturbation of the eigenvalues of \mathcal{B}_0 analogously to our bounds in Section 2.2.

THEOREM 4.1. *Let $\lambda_{\mathcal{B}}$ be an eigenvalue of $\mathcal{B}(F, S_2)$, λ be an eigenvalue of \mathcal{B}_0 and $Qv_j = \delta_j v_j$.*

1. *If $\delta_j \in \mathbb{R}$, for $j = 1, \dots, m$, then*

$$|\lambda_{\mathcal{B}} - \lambda| \leq (1 + \sqrt{2})\kappa(\Theta) \left(\frac{1 + \omega_1}{1 - \omega_1} \right)^{1/2} \|S\| + \max_j \{ |1 + \delta_j \lambda_j^+|, |1 + \delta_j \lambda_j^-| \} \kappa(V) \|\mathcal{E}\|.$$

2. *If $\delta_j \in \mathbb{C}$ and $\exists \alpha > 0$ s.t. $|\delta_j| \leq \alpha < \sqrt{5}$, for $j = 1, \dots, m$, then*

$$\begin{aligned} |\lambda_{\mathcal{B}} - \lambda| \leq & 2 \max \left(1, \frac{1}{2} + \frac{1 + \alpha}{2\sqrt{2}(\sqrt{5} - \alpha)} \right) \kappa(\Theta) \left(\frac{1 + \omega_1}{1 - \omega_1} \right)^{1/2} \|S\| \\ & + \frac{2 + (1 + \sqrt{5})\alpha + 2\alpha^2}{\sqrt{2}(\sqrt{5} - \alpha)} \kappa(V) \|\mathcal{E}\|. \end{aligned}$$

3. *If $D = 0$, then*

$$|\lambda_{\mathcal{B}} - \lambda| \leq 2 \left(\frac{1 + \omega_1}{1 - \omega_1} \right)^{1/2} \|S\| + \frac{2\sqrt{5}}{5} \|\mathcal{E}\|.$$

Proof. In Section 2.1 we have already derived the eigendecomposition of \mathcal{B}_0 . From this decomposition we get the following perturbation bound (see [32, Theorem IV.1.12]),

$$\begin{aligned} |\lambda_{\mathcal{B}} - \lambda| & \leq \left\| \mathcal{Y}^{-1} \begin{bmatrix} S & 0 \\ -\mathcal{E}M & -\mathcal{E}Q \end{bmatrix} \mathcal{Y} \right\| \\ & \leq \left\| \mathcal{Y}^{-1} \begin{bmatrix} S & 0 \\ 0 & 0 \end{bmatrix} \mathcal{Y} \right\| + \left\| \mathcal{Y}^{-1} \begin{bmatrix} 0 & 0 \\ \mathcal{E}M & \mathcal{E}Q \end{bmatrix} \mathcal{Y} \right\|. \end{aligned} \quad (4.2)$$

Corollary 2.4 gives bounds for the first term in (4.2). So, we only need bounds for the second term. Define \mathcal{X} such that

$$\mathcal{X} = \mathcal{Y}^{-1} \begin{bmatrix} 0 & 0 \\ \mathcal{E}M & \mathcal{E}Q \end{bmatrix} \mathcal{Y}.$$

We have

$$\begin{bmatrix} 0 & 0 \\ \mathcal{E}M & \mathcal{E}Q \end{bmatrix} \mathcal{Y} = \begin{bmatrix} 0 & 0 \\ -\mathcal{E}(MY_{11} + QY_{21}) & -\mathcal{E}(MY_{12} + QY_{22}) \end{bmatrix},$$

where $MU_1 = 0$ and $MU_2 = MNV = (I + Q)V = V(I + \Delta)$. This gives $MY_{12} = MU_2 = V(I + \Delta)$, $MY_{11} = [0 \ V(I + \Delta)]$, $QY_{22} = V\Delta(\Lambda^- - I)$ and $QY_{21} = [0 \ V\Delta(\Lambda^+ - I)]$. So, the previous equation reduces to

$$\begin{bmatrix} 0 & 0 \\ \mathcal{E}M & \mathcal{E}Q \end{bmatrix} \mathcal{Y} = \left[\begin{array}{c|c} 0 & 0 \\ \hline 0 & -\mathcal{E}V(I + \Delta\Lambda^+) \end{array} \middle| \begin{array}{c} 0 \\ -\mathcal{E}V(I + \Delta\Lambda^-) \end{array} \right]. \quad (4.3)$$

We then multiply (4.3) from the left by \mathcal{Y}^{-1} , see (2.14)–(2.17), and refactor to yield

$$\mathcal{X} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & (\Lambda^- - \Lambda^+)^{-1} & 0 \\ 0 & 0 & -(\Lambda^- - \Lambda^+)^{-1} \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ 0 & V^{-1}\mathcal{E}V & V^{-1}\mathcal{E}V \\ 0 & V^{-1}\mathcal{E}V & V^{-1}\mathcal{E}V \end{bmatrix} \mathcal{W},$$

where

$$\mathcal{W} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & I + \Delta\Lambda^+ & 0 \\ 0 & 0 & I + \Delta\Lambda^- \end{bmatrix}.$$

Using the consistency of the 2-norm we have the following bound on $\|\mathcal{X}\|$.

$$\|\mathcal{X}\| \leq 2\|(\Lambda^- - \Lambda^+)^{-1}\| \max_j \{|1 + \delta_j \lambda_j^+|, |1 + \delta_j \lambda_j^-|\} \kappa(V) \|\mathcal{E}\|. \quad (4.4)$$

The remainder of the proof concerns the bounds on the right hand side of (4.4) for each particular case.

For the first part of the theorem, assume $\delta_j \in \mathbb{R}$, for $j = 1, \dots, m$. We have

$$\begin{aligned} \lambda_j^- - \lambda_j^+ &= \frac{1 + \delta_j - \sqrt{4 + (1 + \delta)^2}}{2} - \frac{1 + \delta_j + \sqrt{4 + (1 + \delta)^2}}{2} = -\sqrt{4 + (1 + \delta_j)^2} \\ &= -\sqrt{p(\delta)}. \end{aligned}$$

Clearly, $|1/(\lambda_j^- - \lambda_j^+)|$ obtains its maximum at $\delta_j = -1$. This yields $|1/(\lambda_j^- - \lambda_j^+)| \leq 1/2$. We can use this in (4.4) to complete the proof of the first bound.

For the second part of the theorem, we assume $\exists \alpha > 0$ s.t. $|\delta_j| \leq \alpha < \sqrt{5}$, for $j = 1, \dots, m$. First we derive a bound for $\|(\Lambda^- - \Lambda^+)^{-1}\|$. Recall the lower bound on $p(\delta)$ in the proof of Lemma 2.2 and note that $|1/(\lambda_j^- - \lambda_j^+)| = 2/\sqrt{|p(\delta_j)|}$. So, we have $\|(\Lambda^- - \Lambda^+)^{-1}\| \leq (2(\sqrt{5} - \alpha))^{-1/2}$. Furthermore, we have

$$|1 + \delta_j \lambda_j^\pm| = \left| 1 + \delta_j \frac{1 + \delta_j \pm \sqrt{4 + (1 + \delta_j)^2}}{2} \right| \leq 1 + \frac{|\delta_j| |1 + \delta_j| + |\delta_j| \sqrt{4 + (1 + \delta_j)^2}}{2}.$$

We can bound $|\delta + 1 - 2i|$ and $|\delta + 1 + 2i|$ from above by $\sqrt{5} + \alpha$; so, $\sqrt{|4 + (1 + \delta_j)^2|} \leq \sqrt{5} + \alpha$. Thus, we have

$$|1 + \delta_j \lambda_j^\pm| \leq 1 + \frac{\alpha(1 + \alpha) + \alpha(\sqrt{5} + \alpha)}{2} = 1 + \frac{1 + \sqrt{5}}{2} \alpha + \alpha^2.$$

Substituting these bounds into (4.4) yields

$$\|\mathcal{X}\| \leq \frac{2 + (1 + \sqrt{5})\alpha + 2\alpha^2}{\sqrt{2}(\sqrt{5} - \alpha)} \kappa(V) \|\mathcal{E}\|. \quad (4.5)$$

We can then substitute this result into (4.2) to prove the second part of the theorem.

For the third part of the theorem, we assume $D = 0$. We bound the first term in (4.2) using Theorem 2.1, Lemma 2.2 for $\delta \geq -1$ and Lemma 2.3 where $\kappa(\Theta) = 1$. This follows from the fact that U_2 can be chosen to be orthogonal (see [8]).

For the second term in (4.2), since $Q = 0$, $\delta_j = 0$, so $\lambda_j^- - \lambda_j^+ = -\sqrt{5}$, and we can choose $V = I$. We then substitute this into (4.4). \square

In practice, in the complex case the term involving α will generally be modest. For example, if $\alpha = 1$, it is about 4.6022, and for $\alpha = 2$, it is about 23.9727.

If we compare the bounds from Theorem 4.1 with those from Corollary 2.4 for the block-diagonal preconditioner with the exact Schur complement, $(D - CF^{-1}B^T)$, we see that the deterioration of the bounds is $O(\|\mathcal{E}\|)$. Note that the factors that

multiply the $\|\mathcal{E}\|$ are all constants with respect to the choice of the approximate Schur complement, S_2^{-1} . This is about as good as we can hope for. The bounds also demonstrate that there is no point in investing in a really good splitting when a poor approximation to the Schur complement is used or vice versa. Rather, we should be equally attentive to both if we want good eigenvalue clustering.

4.2. Eigenvalue Analysis of the Related System. If we follow the approach in Section 3 to generate the related system for this problem, we would generate precisely the related system derived from (3.3), with S_1^{-1} instead of S_2^{-1} [8]. Therefore, we use an alternative splitting of $\mathcal{B}(F)$,

$$\mathcal{B}(F) = \begin{bmatrix} I & N \\ M_2 & Q_2 + \mathcal{E} \end{bmatrix} - \begin{bmatrix} S & 0 \\ 0 & \mathcal{E} \end{bmatrix},$$

and derive the related system for this splitting. Due to the \mathcal{E} term in the splitting, however, we cannot reduce the size of our system. Instead, we get,

$$\begin{bmatrix} I - (I - NM_2)S & -N\mathcal{E} \\ -M_2S & I + \mathcal{E} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \hat{f} \\ \hat{g} \end{bmatrix}. \quad (4.6)$$

For a special problem in magnetostatics, a linear system similar to (4.6) was derived in [25]. If we use the choices for the splitting and approximations from [25], we obtain basically the same system to be solved. In [25], the authors only outline the qualitative behavior of the eigenvalues in the case that \mathcal{E} is sufficiently small.

THEOREM 4.2. *For any eigenvalue, λ_R , of the related system matrix (4.6),*

$$|1 - \lambda_R| \leq \sqrt{1 + \|N\|^2} \sqrt{1 + \|M_2\|^2} \max(\|S\|, \|\mathcal{E}\|).$$

Proof. Note that the matrix in (4.6) can be split as follows,

$$\begin{aligned} \begin{bmatrix} I - (I - NM_2)S & -N\mathcal{E} \\ -M_2S & I + \mathcal{E} \end{bmatrix} &= I - \begin{bmatrix} I - NM_2 & N \\ M_2 & -I \end{bmatrix} \begin{bmatrix} S & 0 \\ 0 & \mathcal{E} \end{bmatrix} \\ &= I - \begin{bmatrix} I & -N \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ M_2 & -I \end{bmatrix} \begin{bmatrix} S & 0 \\ 0 & \mathcal{E} \end{bmatrix}. \end{aligned}$$

Expressing our matrix as a perturbation of the identity and using a classic perturbation bound (see [32]) yields

$$|1 - \lambda_R| \leq \left\| \begin{bmatrix} I & -N \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ M_2 & -I \end{bmatrix} \begin{bmatrix} S & 0 \\ 0 & \mathcal{E} \end{bmatrix} \right\|.$$

Noting that

$$\left\| \begin{bmatrix} I & -N \\ 0 & I \end{bmatrix} \right\| \leq \sqrt{1 + \|N\|^2} \quad \text{and} \quad \left\| \begin{bmatrix} I & 0 \\ M_2 & -I \end{bmatrix} \right\| \leq \sqrt{1 + \|M_2\|^2},$$

we obtain

$$|1 - \lambda_R| \leq \sqrt{1 + \|N\|^2} \sqrt{1 + \|M_2\|^2} \max(\|S\|, \|\mathcal{E}\|).$$

□

The terms $\|N\|$ and $\|M_2\|$ in the bound from Theorem 4.2 are fairly benign. They are bounded by the norms of the off-diagonal blocks of the un-preconditioned matrix (1.1) and the norms of the inverses of the splitting and approximate Schur complement. Note that the latter two are chosen by the user. Moreover, if we use a good preconditioner for this problem and therefore both our splitting and approximate Schur complement are reasonably accurate, the norms of their inverses will not be large relative to the norm of (1.1), unless (1.1) is itself poorly conditioned.

Just as for the block-diagonally preconditioned system, the eigenvalue perturbation of the related system depends on both $\|S\|$ and $\|\mathcal{E}\|$. Again, there is no advantage in making one significantly smaller than the other. Thus, we should be equally attentive to both $\|S\|$ and $\|\mathcal{E}\|$ in order to achieve tight clustering and fast convergence.

5. Numerical Experiments. We present numerical experiments for two model problems, both arising from the Navier-Stokes equations.

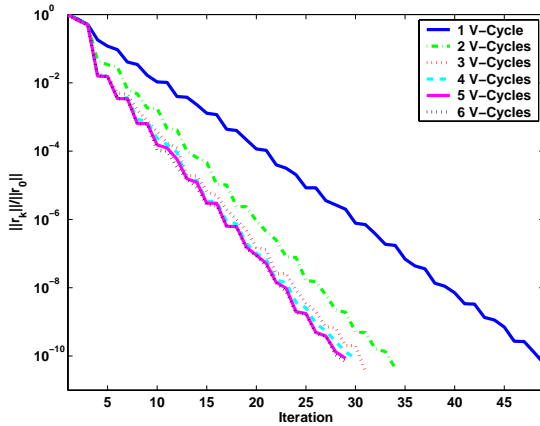
The first model problem involves a stabilized finite element discretization of the Navier-Stokes equations. We use the software toolkit for a 2D leaky lid-driven cavity problem developed for the Winter School in Scientific Computing and Iterative Methods hosted by the Chinese University of Hong Kong in December 1995 and made available by David Silvester [11]. Using this toolkit, we can easily apply the preconditioners and analysis from this paper to the stabilized Navier-Stokes problem (Oseen case). This problem is non-symmetric but has $B = C$. Excellent work has been done by others on preconditioners for this specific problem [11, 31, 33], which we do not intend to supplant. Rather, our goal is to illustrate the effect of the preconditioners proposed in this paper on the convergence behavior and the eigenvalue distributions for a problem which is well-understood and easily accessible to the community.

In particular, we show what happens to the convergence of GMRES, the eigenvalues, and our eigenvalue bounds as we improve the splitting ($\|S\| \rightarrow 0$) and the approximate Schur complement ($\|\mathcal{E}\| \rightarrow 0$). We also succinctly compare the block-diagonally preconditioned systems (2.3) and (4.1) with the related systems (3.4) and (4.6), in terms of both eigenvalues and convergence. We also illustrate the importance of *balancing* the quality of the splitting and the Schur complement to avoid wasted effort. Finally, we study the influence of the mesh width on the convergence of the related system.

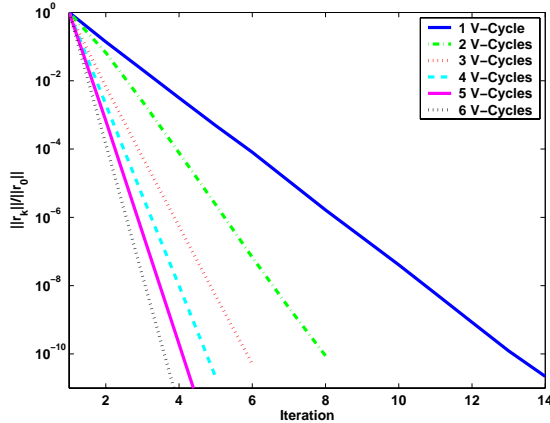
The second model problem involves a spectral collocation discretization for the incompressible Stokes equations on a square [3, 26]. This application has $B \neq C$ and $D = 0$, and this particular formulation uses the Chebyshev nodes for the collocation sites to allow the rapid computation of Gauss-Lobatto quadrature. To our knowledge, this is the first presentation of convergence and eigenvalue results in the literature for preconditioners for generalized saddle-point problems with $B \neq C$. For this application, we present GMRES convergence results as well as the locations of the eigenvalues of the preconditioned system.

5.1. Navier-Stokes with Finite Elements. For our first experiments, we choose a 16×16 grid, viscosity parameter $\nu = 0.1$ and stabilization parameter $\beta = 0.25$. After removing the constant pressure mode, the system has 705 unknowns. Since multigrid cycles are actually matrix splittings, we use a number of multigrid V-cycles to define the splitting of the (1,1) block. For each V-cycle we use three SOR-Jacobi pre- and post-smoothing steps with relaxation parameter $\omega = 0.25$. As a purely algebraic alternative, we also employ an ILUT factorization of the (1,1) block and vary the drop tolerance to change the accuracy of our splitting [27].

We start with the exact Schur complement, varying the number of V-cycles for the splitting from one to six. Figures 5.1(a) and 5.1(b) show the convergence history for preconditioned GMRES for the block-diagonally preconditioned system and the related system, respectively. Note that the related system converges in significantly fewer iterations, for any choice of the number of V-cycles, demonstrating the performance difference between the two preconditioned systems.



(a) Block-Diagonal Preconditioner (2.3).

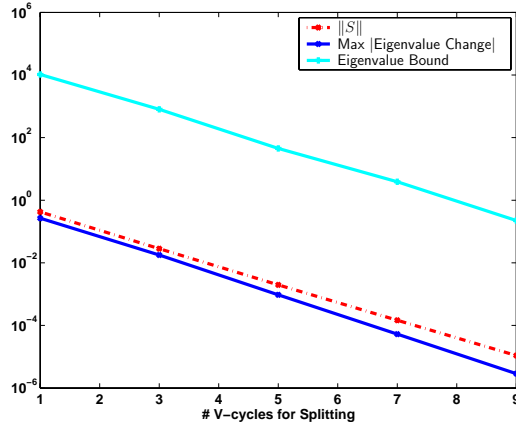


(b) Related System (3.4).

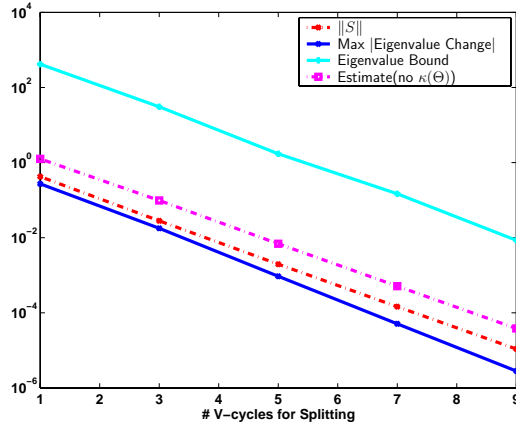
FIG. 5.1. Convergence of GMRES for both types of preconditioners, using the exact Schur complement and varying the number of V-cycles for the splitting.

We have also computed the eigenvalue perturbation and the eigenvalue bounds for both preconditioned systems, using up to nine V-cycles for the splitting, with the exact Schur complement. Figure 5.2(a) shows the maximum absolute eigenvalue perturbation from $\lambda \in \{1, \lambda_j^\pm\}$ for the block-diagonally preconditioned system (2.3), and Figure 5.2(b) shows the maximum absolute eigenvalue perturbation from 1 for the related system (3.4).

As we use a better splitting for A (more V-cycles), we see that the eigenvalue bound decreases with approximately the same rate as the corresponding eigenvalue perturbations, although the bound is pessimistic. This pessimism is mostly due to



(a) Block-Diagonal Preconditioner (2.3).

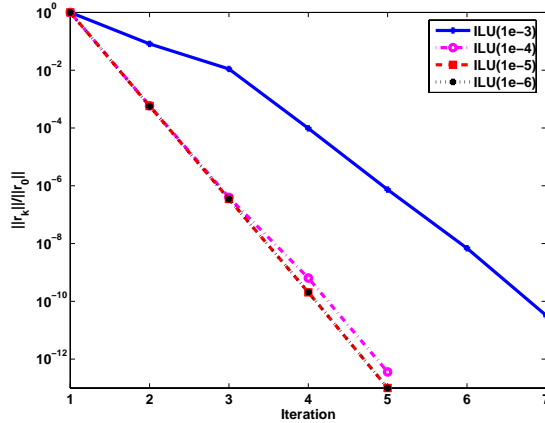


(b) Related System (3.4).

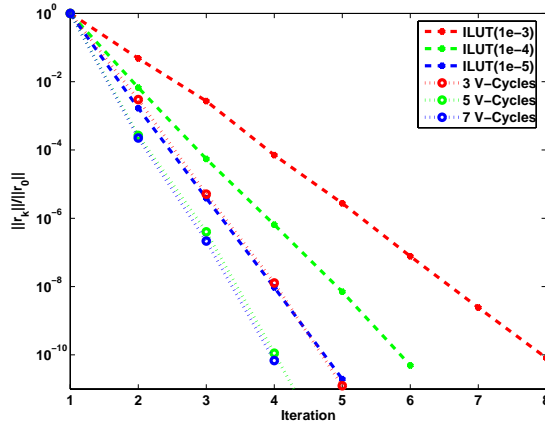
FIG. 5.2. Maximum absolute eigenvalue perturbation and perturbation bounds, for both types of preconditioners, using the exact Schur complement and varying the number of V-cycles for the splitting.

the $\kappa(\Theta)$ factor. Figure 5.2(b) includes an *estimate* of the perturbation for the related system, which consists of the bound in Corollary 3.2 with $\kappa(\Theta)$ replaced by one. Both the bound and our estimate follow the trend in the actual eigenvalue perturbation well as the number of V-cycles increases. The figure shows that the bounds and the estimate give good qualitative respectively quantitative descriptions of the eigenvalue perturbation as the splitting improves.

The eigenvalue perturbation bound for the related system (3.4) is much smaller than for the block-diagonally preconditioned system (2.3). However, the actual maximum eigenvalue perturbation for both systems is about equal. For the related system, this represents a single eigenvalue cluster around 1, which means that the bound proves fast convergence for about 6 V-cycles or more, and the actual (max) perturbation indicates good convergence already for 1 V-cycle. On the other hand, for the block-diagonally preconditioned system, this represents $2m + 1$ (potentially) distinct clusters around 1 and λ_j^\pm , for $j = 1, \dots, m$. The existence of multiple clusters in this



(a) Using five V-cycles for the splitting of the (1,1) block and varying the approximate Schur complement.



(b) Using the approximate Schur complement with ILUT(1e - 5) and varying the splitting of the (1,1) block (# V-cycles and ILUT tolerance).

FIG. 5.3. Convergence results for the related system using an approximate Schur complement.

case, compared with the single cluster for the related system, explains the difference in their convergence behavior. These multiple clusters also explain the diminishing returns of improving the splitting for the block-diagonal preconditioner shown in Figure 5.1(a). As we see similar differences between the preconditioners for the other test cases, we only show results for the related system for the remainder of this section.

We illustrate the convergence behavior for the preconditioner with an approximate Schur complement as a function of the accuracy of the approximation by using an ILUT decomposition [27]. While this may not be a practical choice, it serves our purposes for this paper, because it allows us to progressively increase the accuracy of the approximation to the inverse of the Schur complement. We use drop tolerances ranging from $1e - 3$ to $5e - 8$.

Figures 5.3(a) and 5.3(b) show the effects of improving the splitting (for multi-grid and ILUT) and the approximation to the Schur complement on the convergence of GMRES for the related system (4.6). First, in Figure 5.3(a), we vary the drop

tolerance for the approximate Schur complement and fix the number of V-cycles for the splitting at five. Then, in Figure 5.3(b), we demonstrate a number of splittings using V-cycles and ILUT, and fix the drop tolerance at $1e - 5$ for the approximate Schur complement. The convergence results are quite good, regardless of the choice of splitting.

The convergence rates in Figures 5.3(a) and 5.3(b) hit a point of diminishing returns, past which improving either the splitting or the approximate Schur complement while leaving the other unchanged does not improve convergence. To explain this, we show the eigenvalue perturbations from 1 and the perturbation bound for the same example in Figure 5.4. In both plots, the eigenvalue perturbation (and bound) cease to decrease shortly after $\|S\|$ is less than $\|\mathcal{E}\|$ or vice versa. This demonstrates that the eigenvalue bound from Theorem 4.2 is indicative of the actual eigenvalue perturbation and the resulting convergence behavior, and that using a significantly more accurate splitting than approximate Schur complement, or vice versa, yields little additional benefit. Finally, note that for reasonable choices of splitting and approximation to the Schur complement, the bounds are less than 1 indicating that the eigenvalues are clustered away from the origin. This should lead to rapid convergence for Krylov methods.

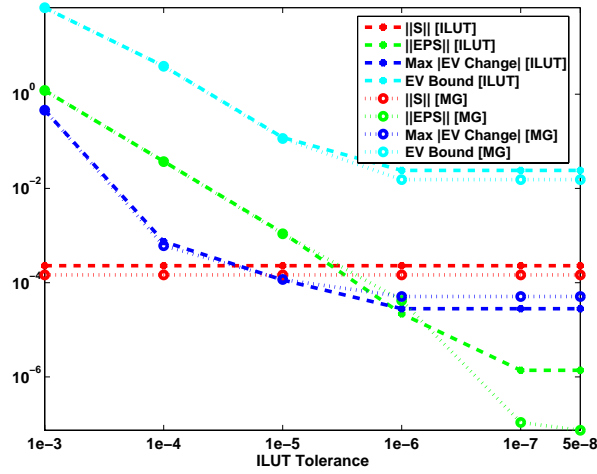
Varying the number of grid points per dimension, $n = 1/h$, gives some insight how the convergence of the related system (4.6) depends on h . Table 5.1 summarizes these results. First, note that $|\delta_j|$ decreases with h . This leads to significant reductions of the factors involving δ_j in the theorems of Section 2, 3 and 4. In particular, with respect to Corrolary 3.2 for the related system and Corrolary 2.4 and Theorem 4.1 for the block-diagonal preconditioner, note that for small h the δ_j are nearly real. Moreover, note that the convergence of GMRES for the related system (4.6) depends only mildly on h . A good splitting and a reasonably accurate approximate Schur complement seem to lead to h -independent convergence.

n	$\max \delta_j $	Number of GMRES iterations			
		ILUT(1e-3)	ILUT(1e-4)	ILUT(1e-5)	ILUT(1e-6)
4	1.72e+00	5	5	5	5
8	5.92e-01	5	4	4	4
16	1.60e-01	7	5	5	5
32	4.07e-02	13	6	5	5

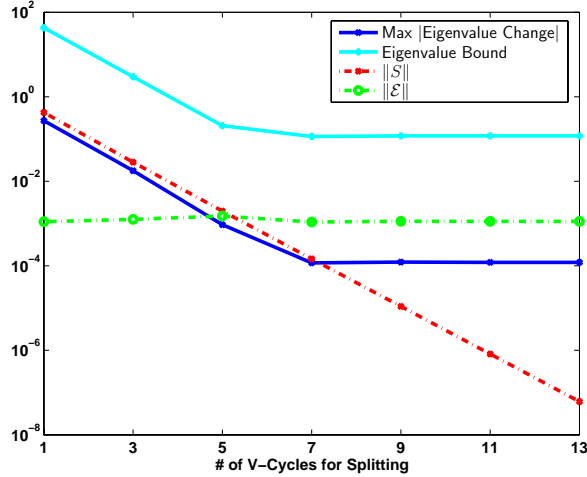
TABLE 5.1

Effect of the number of grid points per dimension (n) on $\max_j |\delta_j|$ and the number of GMRES iterations for the related system (4.6) using a splitting of 5 V-cycles and various approximate Schur complements.

5.2. Incompressible Stokes with Spectral Collocation. We will build discretizations with polynomials of degree up to 22 for this problem. The largest system will be of size 1241. We use an odd-even ordering for the velocity unknowns to exploit the orthogonality properties of Chebyshev polynomials and put the the (1,1) block in block-diagonal form. We use ILUT with a drop tolerance of $1e - 4$ for the splitting of the (1,1) block, and for the approximate Schur complement we use ILUT with a drop tolerance between $1e - 3$ and $1e - 5$. Figure 5.5(a) shows the eigenvalues of the related system for the largest problem, $N = 22$. Except for a single eigenvalue of $O(1e - 2)$, the eigenvalues are tightly clustered around one. As expected, this leads to rapid convergence, as shown in Figure 5.5(b). Moreover, the GMRES iteration count



(a) Using seven V-cycles or ILUT($1e-5$) for the splitting and varying the approximate Schur complement.



(b) Using the approximate Schur complement with ILUT($1e-5$) and varying the number of V-cycles for the splitting.

FIG. 5.4. The effects of $\|S\|$ and $\|\mathcal{E}\|$ on related system using the approximate Schur complement.

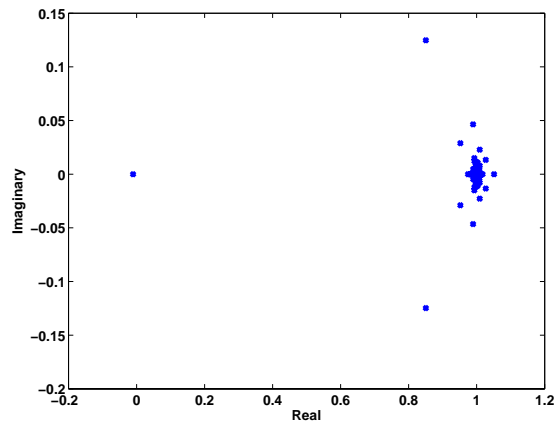
for the related system with an approximate Schur complement (with the exception of ILUT($1e-3$)) shows only modest dependence on the maximum polynomial degree N . Hence, even for fully asymmetric problems, our preconditioners are effective, and show the potential of scaling well to larger problems.

6. Conclusions and Future Work. We have proposed and analyzed variants of indefinite preconditioners (the related system) and block-diagonal preconditioners for the $D \neq 0$ case, including the use of approximate Schur complements. We have illustrated their performance in terms of convergence, eigenvalue perturbations and eigenvalue bounds using well-known model problems. Further analysis should help tighten the eigenvalue bounds, in particular using the consistency property of

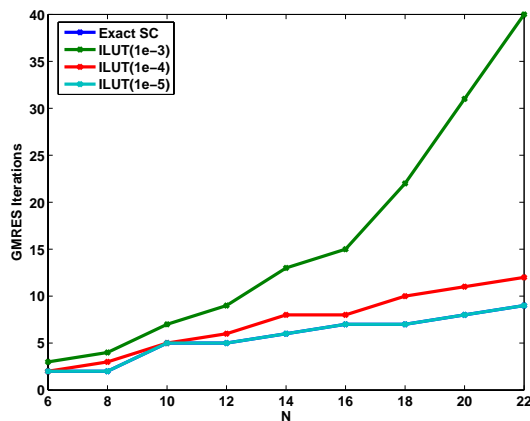
matrix norms less. We also aim to specialize our methods to particular problems. We are currently exploring applications from metal deformation, porous media flow, optimization and electromagnetics.

Acknowledgements. We gratefully acknowledge the use of the software toolkit for a 2D leaky lid-driven cavity problem developed by David Silvester in collaboration with Howard Elman, Bernd Fischer, Alison Ramage, and Andy Wathen.

We also gratefully acknowledge the reviewers for many suggestions that helped improve this paper.



(a) Eigenvalues of related system for polynomial degree $N = 22$ using an approximate Schur complement with $\text{ILUT}(1e-4)$.



(b) GMRES iteration count versus maximum polynomial degree (N) for various approximate Schur complements. The iteration counts for the exact Schur complement coincide with those for the approximate Schur complement with $\text{ILUT}(1e-5)$.

FIG. 5.5. *Eigenvalues and iteration counts for the related system (4.6) from spectral discretization of the incompressible Stokes equations with an $\text{ILUT}(1e-4)$ splitting and an approximate Schur complement.*

REFERENCES

- [1] M. Benzi, M.J. Gander, and G.H. Golub. Optimization of the Hermitian and skew-Hermitian splitting iteration for saddle-point problems. *BIT*, 43:881–900, 2003.
- [2] M. Benzi and G.H. Golub. A preconditioner for generalized saddle point problems. *SIAM J. Matrix Anal. Appl.*, 26:20–41, 2004.
- [3] C. Bernardi, C. Canuto, and Y. Maday. Generalized inf-sup conditions for Chebyshev spectral approximation of the Stokes problem. *SIAM J. on Numer. Anal.*, 25(6):1237–1271, 1988.
- [4] D. Braess. *Finite Elements: Theory, fast solvers and applications in solid mechanics*. Cambridge University Press, 2nd edition, 2001.
- [5] D. Braess, P. Deuffhard, and K. Lipnikov. A subspace cascadic multigrid method for mortar elements. *Computing*, 69:205–225, 2002.
- [6] D. Braess and R. Sarazin. An efficient smoother for the Stokes problem. *Applied Numerical Mathematics*, 23:3–19, 1997.
- [7] J.H. Bramble and J.E. Pasciak. A preconditioning technique for indefinite systems resulting from mixed approximations of elliptic problems. *Math. Comp.*, 50(181):1–17, January 1988.
- [8] E. de Sturler and J. Liesen. Block-diagonal and constraint preconditioners for nonsymmetric indefinite linear systems. Part I: Theory. *SIAM J. Sci. Comput.*, 26(5):1598–1619, 2005.
- [9] H.C. Elman. Preconditioning for the steady-state Navier-Stokes equations with low viscosity. *SIAM J. Sci. Comput.*, 20(4):1299–1316, 1999.
- [10] H.C. Elman and G.H. Golub. Inexact and preconditioned Uzawa algorithms for saddle point problems. *SIAM J. Numer. Anal.*, 31(6):1645–1661, 1994.
- [11] H.C. Elman, D.J. Silvester, and A.J. Wathen. Iterative methods for problems in computational fluid dynamics. In *Winter School on Iterative Methods in Scientific Computing and Applications*. Chinese University of Hong Kong, 1996.
- [12] H.C. Elman, D.J. Silvester, and A.J. Wathen. *Finite Elements and Fast Iterative Solvers*. Oxford University Press, 2005.
- [13] A. Forsgren, P.E. Gill, and M.H. Wright. Interior methods for nonlinear optimization. *SIAM Review*, 44(4):525–597, 2002.
- [14] G.H. Golub and A.J. Wathen. An iteration for indefinite systems and its application to the Navier-Stokes equations. *SIAM J. Sci. Comput.*, 19(2):530–539, 1998.
- [15] N.I.M. Gould, M.E. Hribar, and J. Nocedal. On the solution of equality constrained quadratic programming problems arising in optimization. *SIAM J. Sci. Comput.*, 23(4):1376–1395, 2001.
- [16] L.A. Hageman and D.M. Young. *Applied Iterative Methods*. Academic Press, 1981.
- [17] R.A. Horn and C.R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.
- [18] I.C.F. Ipsen. A note on preconditioning nonsymmetric matrices. *SIAM J. Sci. Comput.*, 23(3):1050–1051, 2001.
- [19] C. Keller, N.I.M. Gould, and A.J. Wathen. Constraint preconditioning for indefinite linear systems. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1300–1317, 2000.
- [20] P. Krzyżanowski. On block preconditioners for nonsymmetric saddle point problems. *SIAM J. Sci. Comput.*, 23(1):157–169, 2001.
- [21] L. Lukšan and J. Vlček. Indefinitely preconditioned inexact Newton method for large sparse equality constrained non-linear programming problems. *Numer. Linear Algebra Appl.*, 5:219–247, 1998.
- [22] C. Meyer. *Matrix Analysis and Applied Linear Algebra*. SIAM, 2001.
- [23] M.F. Murphy, G.H. Golub, and A.J. Wathen. A note on preconditioning for indefinite linear systems. *SIAM J. Sci. Comput.*, 21(6):2969–1972, 2000.
- [24] R. Nicolaides. Existence, uniqueness and approximation for generalized saddle point problems. *SIAM Journal on Numerical Analysis*, 19(2):349–357, 1982.
- [25] I. Perugia and V. Simoncini. Block-diagonal and indefinite symmetric preconditioners for mixed finite element formulations. *Numer. Linear Algebra Appl.*, 7:585–616, 2000.
- [26] A. Quarteroni and A. Valli. *Numerical Approximation of Partial Differential Equations*. Springer-Verlag, 2nd edition, 1997.
- [27] Y. Saad. ILUT: a dual threshold incomplete ILU factorization. *Numerical Linear Algebra with Applications*, pages 387–402, 1994.
- [28] C. Siefert. *Preconditioners for Generalized Saddle-Point Problems*. PhD thesis, University of Illinois at Urbana-Champaign, 2005.
- [29] C. Siefert and E. de Sturler. Preconditioners for generalized saddle-point problems. Technical Report UIUCDCS-R-2004-2448, Department of Computer Science, University of Illinois at Urbana-Champaign, June 2004.
- [30] D. Silvester, H. Elman, D. Kay, and A. Wathen. Efficient preconditioning of the linearized

- Navier-Stokes equations for incompressible flow. *J. Comput. Appl. Math.*, 128(1-2):261–279, 2001. Numerical analysis 2000, Vol. VII, Partial differential equations.
- [31] D. Silvester and A. Wathen. Fast iterative solution of stabilised Stokes systems Part II: Using general block preconditioners. *SIAM J. Numer. Anal.*, 31:1352–1367, October 1994.
 - [32] G.W. Stewart and J.G. Sun. *Matrix perturbation theory*. Academic Press Inc., Boston, 1990.
 - [33] A. Wathen and D. Silvester. Fast iterative solution of stabilised Stokes systems Part I: Using simple diagonal preconditioners. *SIAM J. Numer. Anal.*, 30:630–649, June 1993.
 - [34] L. Zhu, A.J. Beaudoin, and S.R. MacEwan. A study of kinetics in stress relaxation of AA 5182. In *Proceedings of TMS Fall 2001: Microstructural Modeling and Prediction During Thermomechanical Processing*, pages 189–199, 2001.