Preconditioners for Generalized Saddle-Point Problems

Chris Siefert *and Eric de Sturler †

June 21, 2004

Abstract

We examine block-diagonal preconditioners and efficient variants of indefinite preconditioners for block two-by-two generalized saddle-point problems. We consider the general, nonsymmetric, nonsingular case. In particular, the (1,2) block need not equal the transposed (2,1) block. Our preconditioners arise from computationally efficient splittings of the (1,1) block. We provide analyses for the eigenvalue distributions and other properties of the preconditioned matrices. We extend the results of [de Sturler and Liesen 2003] to matrices with non-zero (2,2) block and to allow for the use of inexact Schur complements. To illustrate our eigenvalue bounds, we apply our analysis to a model Navier-Stokes problem, computing the bounds, comparing them to actual eigenvalue perturbations and examining the convergence behavior.

1 Introduction

We examine preconditioners for real systems of the form,

$$\mathcal{A}\begin{bmatrix} x \\ y \end{bmatrix} \equiv \begin{bmatrix} A & B^T \\ C & D \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix}, \tag{1}$$

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 such problems are referred to as generalized saddle point problems [23]. For other problems we consider, $D \neq 0$, but $\|D\|_2$ is small, so 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 [31], 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, 23] and [25, Sections 7.5 and 9.4], although many other problems have $B = C$. We consider all of these cases, which arise in many applications, ranging from stabilized formulations of the Navier-Stokes equations [4, 11, 25] to metal deformation [31] and interior point methods [12].

Problems of this type have been of recent interest [2, 8, 9, 17, 19, 22], as have their symmetric counterpart [7, 10, 13, 24, 28, 30], and the case where $D = 0$ [1, 5, 6, 8, 14, 18, 20, 22, 27]. However, preconditioners for the case $B \neq C$ have not received as much attention, although they are considered in [8, 17, 22]. 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 this analysis to the case where $D \neq 0$ and to allow for approximations to a Schur complement that arises

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

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

This work is supported in part by grant DOE LLNL B341494

in the preconditioner. We consider preconditioners for (1) that derive from a matrix 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 only that the matrix is non-singular or that the singularity can be easily removed, such as the constant pressure mode in the Oseen problem. For the splitting, we assume that F and $(D - CF^{-1}B^T)$ are nonsingular. We analyze a block-diagonal preconditioner that is a generalization of a preconditioner suggested in [17]. Further, we analyze a fixed-point method (cf. [13]) and its related system, which are reduced in size, following the discussion in [8]. For both types of preconditioners, we extend the analysis to the use of inexact Schur complements. Our analysis focuses on the $D \neq 0$ case, but we provide specializations to the $D = 0$ case as well. For the $D = 0$ case, the related system corresponds to an efficient implementation of a constraint preconditioner, see also [5, 6, 13, 24].

For all these preconditioners, we present bounds on the location and size of the eigenvalue clusters, and we explore their relation with actual eigenvalue clustering for a well-known model problem. This allows us to compare these preconditioners with respect to the bounds on their eigenvalue clusters, the actual eigenvalues, and the convergence of Krylov subspace methods applied to the preconditioned systems. We approach the derivation and analysis of these preconditioners from the algebraic point of view, without focusing on any particular application. Since we make few assumptions on our matrices, we cannot be as specific about the exact values of relevant parameters as if a given problem were analyzed. However, we identify and discuss the main factors that govern the bounds on and empirical behavior of the eigenvalues, which allows us to analyze the convergence of Krylov subspace methods applied to the preconditioned systems.

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 where F^{-1} and $(D - CF^{-1}B^T)^{-1}$ exist. Next, we introduce the block-diagonal preconditioner as a straightforward generalization of preconditioners in [8, 17],

$$\mathcal{P}(F) \;=\; \begin{bmatrix} F^{-1} & 0 \\ 0 & -(D - CF^{-1}B^T)^{-1} \end{bmatrix}. \tag{2}$$

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

$$\mathcal{B}(S) \begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix} \;=\; \begin{bmatrix} I - S & N \\ M & Q \end{bmatrix} \begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix} = \begin{bmatrix} \tilde{f} \\ \tilde{g} \end{bmatrix}, \tag{3}$$

where $\mathcal{B}(S)$ is either $\mathcal{P}\mathcal{A}$ or $\mathcal{A}\mathcal{P}$. For example, the matrix from the left-preconditioned system looks like

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

implicitly defining S, N, M and Q in (3) for the left-preconditioned case. Apart from the preconditioned (2,2) block Q, this is quite similar to the system arising from the zero (2,2) block case. While $MN = I$ for the $D = 0$ case [22, 8], for $D \neq 0$ we have

$$\begin{aligned} MN \;&=\; -(D - CF^{-1}B^T)^{-1}CF^{-1}B^T = -(D - CF^{-1}B^T)^{-1}(-D + CF^{-1}B^T + D) \\ &=\; I + Q. \end{aligned} \tag{4}$$

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 consider the eigendecomposition of the matrix $\mathcal{B}(0)$, that is $\mathcal{B}(S)$ in (3) with $S = 0$, where $I + Q$ (and thus B^T and C) are full-rank. Then in Section 2.2, we develop bounds for the eigenvalues of $\mathcal{B}(S)$ using perturbation theory. Finally, in Section 2.3, we discuss the consequences of the rank-deficient case.

2.1 Eigenvalue Analysis

Assume that $I + Q$ (and thus B^T and C) are full rank. We wish to find λ, u and v such that

$$\begin{aligned} u + Nv &= \lambda u \quad (5) \\ Mu + Qv &= \lambda v. \quad (6) \end{aligned}$$

First, we assume $\lambda = 1$. Substituting this into (5) and using $Q = MN - I$ in (6) yields

$$Nv = 0 \quad \text{and} \quad Mu = 2v. \quad (7)$$

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

$$\left(1, \begin{bmatrix} u \\ 0 \end{bmatrix}\right), \qquad \text{where } u \in \text{null}\,(M)\,. \quad (8)$$

If C is full rank, then so is M, and $\mathcal{B}(0)$ has precisely $n - m$ distinct eigenpairs of this type. Next, we consider the case where $\lambda \neq 1$. Solving (5) for u, and substituting into (6) yields

$$\lambda Q v_j = (\lambda^2 - \lambda - 1) v_j. \quad (9)$$

Hence, the v_j must be eigenvectors of Q. We assume that Q has a full set of eigenpairs, $Qv_j = \delta_j v_j$, for $j = 1 \ldots m$. We then then solve (9) for λ to yield:

$$\lambda_j^{\pm} = \frac{(1 + \delta_j) \pm \sqrt{4 + (1 + \delta_j)^2}}{2}, \quad (10)$$

cf. [11]. We then substitute $\delta_j v_j$ for Qv_j in (6), and solve for u. We finally rescale the eigenvector by $(\lambda_j^{\pm} - 1)$ to yield eigenpairs of the form

$$\left(\lambda_j^{\pm}, \begin{bmatrix} Nv_j \\ (\lambda_j^{\pm} - 1) v_j \end{bmatrix}\right). \quad (11)$$

Note that $\lambda_j^- \neq 1$ regardless of the choice of δ_j, and $\lambda_j^+ = 1$ only if $\delta_j = -1$. However, this 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 null (M), cf. (8), and let U_2 be the matrix with normalized columns $u_j = Nv_j$, where $Qv_j = \delta_j v_j$, cf. (11). Furthermore, let $\Lambda^+ = \text{diag}(\lambda_j^+)$ and $\Lambda^- = \text{diag}(\lambda_j^-)$, where diag$(\cdot)$ denotes a diagonal matrix with the arguments given. Then, the eigenvector matrix of $\mathcal{B}(0)$ is given by

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

For our perturbation results we also need

$$\mathcal{Z} = \mathcal{Y}^{-1} \;=\; \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix}. \tag{13}$$

The block-inversion formula in [16, Section 0.7.3] gives,

$$\begin{aligned} Z_{11} &= \left(Y_{11} - Y_{12}Y_{22}^{-1}Y_{21}\right)^{-1} = \begin{bmatrix} I_{n-m} & 0 \\ 0 & \Upsilon^+ \end{bmatrix} Y_{11}^{-1} = \hat{I}_n Y_{11}^{-1}, \\ Z_{21} &= -Y_{22}^{-1}Y_{21}Z_{11} \\ &= -\begin{bmatrix} 0 & (\Lambda^- - I)^{-1}(\Lambda^+ - I)\Upsilon^+ \end{bmatrix} Y_{11}^{-1} \\ &= -\begin{bmatrix} 0 & \Upsilon^- \end{bmatrix} Y_{11}^{-1}, \end{aligned} \tag{14}$$

with $\Upsilon^+ = \mathrm{diag}((\lambda_j^- - 1)/(\lambda_j^- - \lambda_j^+))$ and $\Upsilon^- = \mathrm{diag}((\lambda_j^+ - 1)/(\lambda_j^- - \lambda_j^+))$. Using $[U_1\ U_2]^{-1}NV = [0\ I]^T$ we also have

$$\begin{aligned} Z_{22} &= (Y_{22} - Y_{21}Y_{11}^{-1}Y_{12})^{-1} \\ &= \left(V(\Lambda^- - I) - \begin{bmatrix} 0 & V(\Lambda^+ - I) \end{bmatrix} \begin{bmatrix} U_1 & U_2 \end{bmatrix}^{-1} NV\right)^{-1} \\ &= \left(V(\Lambda^- - I) - \begin{bmatrix} 0 & V(\Lambda^+ - I) \end{bmatrix} \begin{bmatrix} 0 \\ I \end{bmatrix}\right)^{-1} = \left(V(\Lambda^- - \Lambda^+)\right)^{-1}, \end{aligned} \tag{15}$$

$$Z_{12} \;=\; Y_{11}^{-1}Y_{22}Z_{22} \;=\; -\begin{bmatrix} U_1 & U_2 \end{bmatrix}^{-1} U_2 Z_{22} \;=\; -\begin{bmatrix} 0 \\ (\Lambda^- - \Lambda^+)^{-1}V^{-1} \end{bmatrix}. \tag{16}$$

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 $\mathcal{B}(S)$

We are now ready to derive bounds on the eigenvalues of $\mathcal{B}(S)$. Note that throughout this paper $\|\cdot\|$ indicates the 2-norm.

Theorem 2.1. *Consider matrices $\mathcal{B}(S)$ of the form (3). Let $\mathcal{Y}$ be the eigenvector matrix of $\mathcal{B}$, as given by (12). Then for each eigenvalue λ_S of $\mathcal{B}(S)$, there exists an eigenvalue λ of $\mathcal{B}(0)$, such that*

$$|\lambda_S - \lambda| \;\le\; \left\| \mathcal{Y}^{-1} \begin{bmatrix} S & 0 \\ 0 & 0 \end{bmatrix} \mathcal{Y} \right\| \tag{17}$$

$$\le\; 2\max\left(1, \|\Upsilon^+\|, \|\Upsilon^-\|\right) \|Y_{11}^{-1}SY_{11}\|. \tag{18}$$

Proof: Since $\mathcal{B}(0)$ is diagonalizable, (17) follows from a classic result in perturbation theory [29, Theorem IV.1.12]. We expand (17) using (12)–(16) to get (see also [8])

$$\begin{aligned} |\lambda_S - \lambda| &\le \left\| \begin{bmatrix} \hat{I}_n Y_{11}^{-1}SY_{11} & \hat{I}_n Y_{11}^{-1}SY_{12} \\ -\begin{bmatrix} 0 & \Upsilon^- \end{bmatrix} Y_{11}^{-1}SY_{11} & -\begin{bmatrix} 0 & \Upsilon^- \end{bmatrix} Y_{11}^{-1}SY_{12} \end{bmatrix} \right\| \\ &\le \max(1, \|\Upsilon^+\|, \|\Upsilon^-\|) \left\| \begin{bmatrix} Y_{11}^{-1}SU_1 & Y_{11}^{-1}SU_2 & Y_{11}^{-1}SU_2 \\ -\begin{bmatrix} 0 & I \end{bmatrix} Y_{11}^{-1}SU_1 & -\begin{bmatrix} 0 & I \end{bmatrix} Y_{11}^{-1}SU_2 & -\begin{bmatrix} 0 & I \end{bmatrix} Y_{11}^{-1}SU_2 \end{bmatrix} \right\|. \end{aligned}$$

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

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

□

The $\Upsilon^{\pm}$ terms can only be large if $\delta_j \approx -1 \pm 2i$. Since $\|D\|$ is assumed to be small, this can only happen if $\|(D - CF^{-1}B^T)^{-1}\|$ is large, which typically means that the preconditioner is poorly 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 is true 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 $j = 1, \ldots, m$, then

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

Proof: Substituting $\lambda_j^{\pm}$ from (10) in $\Upsilon^+ = \mathrm{diag}(\lambda_j^- - 1)/(\lambda_j^- - \lambda_j^+)$ and $\Upsilon^- = \mathrm{diag}(\lambda_j^+ - 1)/(\lambda_j^- - \lambda_j^+)$ gives

$$\Upsilon^{\pm} = \mathrm{diag}\left(\frac{1-\delta_j}{2\sqrt{4+(1+\delta_j)^2}} \pm \frac{1}{2}\right) = \mathrm{diag}\left(\frac{1-\delta_j}{2\sqrt{p(\delta_j)}} \pm \frac{1}{2}\right). \tag{19}$$

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

For the complex case, note that let $p(\delta) = (\delta + 1 + 2i)(\delta + 1 - 2i)$. 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 $|p(\delta)| \geq 2\left(\sqrt{5} - \alpha\right)$. Using this inequality for $|p(\delta)|$ in after taking norms in (19) 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 $\left\|Y_{11}^{-1} S Y_{11}\right\|$ 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 range $(U_1) =$ null (NM) and range $(U_2) =$ range (NM).

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

$$\left\|Y_{11}^{-1} S Y_{11}\right\| \leq \kappa(\Theta)\left(\frac{1+\omega_1}{1-\omega_1}\right)^{1/2} \|S\|. \tag{20}$$

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

$$\|Y_{11}^{-1} S Y_{11}\| \leq \kappa(\Theta)\ \kappa\left(\begin{bmatrix} U_1 & V_2 \end{bmatrix}\right) \|S\| \leq \kappa(\Theta) \left(\frac{1+\omega_1}{1-\omega_1}\right)^{1/2} \|S\|, \tag{21}$$

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

Corollary 2.4. *Let Θ and ω_1 be defined as above.*
1. If $\delta_j \in \mathbb{R}$, for all j, then

$$|\lambda_s - \lambda| \quad \leq \quad (1+\sqrt{2})\kappa(\Theta) \left(\frac{1+\omega_1}{1-\omega_1}\right)^{1/2} \|S\|. \tag{22}$$

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

$$|\lambda_s - \lambda| \quad \leq \quad 2\max\left(1, \frac{1}{2} + \frac{1+\alpha}{2\sqrt{2\left(\sqrt{5}-\alpha\right)}}\right) \kappa(\Theta) \left(\frac{1+\omega_1}{1-\omega_1}\right)^{1/2} \|S\|. \tag{23}$$

Proof: Use Lemmas 2.2 and 2.3 in Theorem 2.1. □

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. The examples in Section 5 will illustrate this.

2.3 Rank-Deficiency in $I + Q$

In Section 2.1, we made the assumption that $I + Q$ is 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 range $\left(B^T\right)$ into 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 (10) that correspond to $\delta_j \neq -1$. It is also valid for the k eigenpairs where $\delta_j = -1$, corresponding 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 (8) equals $\dim(\text{null}\,(M)) = n - m + l_c$. This gives us a total of $n + m - k + l_c$ eigenpairs, leaving us to find $k - l_c$ eigenpairs.

From (7), 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, $\lambda_j^+ \to 1$ as $\delta_j =\to -1$. Therefore, we look for principal vectors of grade two (see [15]) for $\lambda = 1$. These vectors satisfy the equations

$$Nv = \tilde{u} \quad \text{and} \quad Mu = 2v, \tag{24}$$

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 (7) or (24).

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 $j = 1 \ldots k$, 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 range $(U_1) = \text{null}\,(M)$. Let $\widetilde{V}$ be such that $\widetilde{V}^T\widetilde{V} = I_{l_c}$ and range$(\widetilde{V}) = \text{null}\,(I + Q)$. Further, let the columns of $\widehat{V}$ be the eigenvectors of Q corresponding to the eigenvalues $\delta_j \neq -1$, scaled such that $U_2 = N\widehat{V}$ has unit columns. Finally, let the diagonal matrices $\widehat{\Lambda}^+$ and $\widehat{\Lambda}^-$ contain the eigenvalues λ_j^+ and λ_j^- corresponding to the eigenvalues $\delta_j \neq -1$ ordered consistently with the columns of $\widehat{V}$. Then the eigenvector matrix of $\mathcal{B}(0)$ is given by

$$\mathcal{Y} \;=\; \left[\begin{array}{cc|cc} U_1^{(n-m+l_c)} & U_2^{(m-l_c)} & N\tilde{V}^{(l_c)} & U_2^{(m-l_c)} \\ \hline 0 & \widehat{V}(\widehat{\Lambda}^+ - I) & -2\widetilde{V} & \widehat{V}(\widehat{\Lambda}^- - I) \end{array}\right], \tag{25}$$

where superscripts in the top row indicate the number of columns. The corresponding eigenvalues are those from (8) and (10). We can then use the eigenvector matrix of $\mathcal{B}(0)$ given in (25) 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 the 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 and Related System Solution Methods (exact Schur complement)

We now consider an alternative solution method, cf. [8]. In the $D = 0$ case this approach leads to an efficient implementation of so-called constraint preconditioners, cf. [5, 6, 13, 24]. We can derive the following splitting from (3) ,

$$\mathcal{B}(S)\begin{bmatrix} x \\ y \end{bmatrix} \;=\; \begin{bmatrix} I - S & N \\ M & Q \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \left(\mathcal{B}(0) - \begin{bmatrix} S & 0 \\ 0 & 0 \end{bmatrix}\right)\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \tilde{f} \\ \tilde{g} \end{bmatrix}. \tag{26}$$

Note that

$$\mathcal{B}(0)^{-1} = \begin{bmatrix} I - NM & N \\ M & -I \end{bmatrix}. \tag{27}$$

Left-multiplying (26) by $\mathcal{B}(0)^{-1}$ and splitting yields the fixed point iteration,

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} (I - NM)S & 0 \\ MS & 0 \end{bmatrix} \begin{bmatrix} x_k \\ y_k \end{bmatrix} + \begin{bmatrix} \hat{f} \\ \hat{g} \end{bmatrix}. \tag{28}$$

This iteration is essentially 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 (28) satisfies the so-called *related system* for the fixed-point iteration [15],

$$(I - (I - NM)S)x = \hat{f}. \tag{29}$$

The full-size related system (including the y component) and $D \neq 0$ has been examined elsewhere for special cases. In [24], A is symmetric positive definite and spectrally equivalent to the identity, and so a fixed splitting $F = I$ is used. In [13], 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 (12), $\Delta = \mathrm{diag}(\delta_j)$ and let $U_2 = V_2\Theta$, with $V_2^T V_2 = I$. Then, we have $NMU_1 = 0$, $NMU_2 = NMNV = NV(I + \Delta)$, and therefore

$$(I - NM) = \begin{bmatrix} U_1 & V_2 \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & -\Theta\Delta\Theta^{-1} \end{bmatrix} \begin{bmatrix} U_1 & V_2 \end{bmatrix}^{-1}. \tag{30}$$

In the rank-deficient case, we can use (25). 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),

$$U_1^T V_2 = \Phi\Omega\Psi^T, \text{ where } 1 > \omega_1 \geq \omega_2 \geq \ldots \geq \omega_m. \tag{31}$$

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

$$V_2 = U_1\Phi\Omega\Psi^T + W\Sigma\Psi^T. \tag{32}$$

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 (29). Then,*

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

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,

$$\begin{aligned} \|I-NM\| &= \left\| [U_1\, V_2] \begin{bmatrix} I & 0 \\ 0 & Z \end{bmatrix} [U_1\, V_2]^{-1} \right\| && (33) \\ &\leq \left\| [U_1\, V_2] \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} [U_1\, V_2]^{-1} \right\| + \left\| [U_1\, V_2] \begin{bmatrix} 0 & 0 \\ 0 & Z \end{bmatrix} [U_1\, V_2]^{-1} \right\| && (34) \\ &\leq (1-\omega_1^2)^{-1/2} + (1-\omega_1^2)^{-1/2}\|Z\| = (1-\omega_1^2)^{-1/2}(1+\|Z\|). && (35) \end{aligned}$$

The first term in (34) is the norm of an oblique projection. Given the SVD in (31), this norm equals $(1-\omega_1^2)^{-1/2}$ [21, Section 5.15]. We establish the bound for the second term as follows.

$$\left\| [U_1\, V_2] \begin{bmatrix} 0 & 0 \\ 0 & Z \end{bmatrix} [U_1\, V_2]^{-1} \right\| = \max_{U_1 a + V_2 b \neq 0} \frac{\|V_2 Z b\|}{\|U_1 a + V_2 b\|}. \quad (36)$$

Without loss of generality we may assume $\|b\| = 1$, so that $\|V_2 Z b\| \leq \|Z\|$. From (32) 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

$$\left\| [U_1\, V_2] \begin{bmatrix} 0 & 0 \\ 0 & Z \end{bmatrix} [U_1\, V_2]^{-1} \right\| = \max_{U_1 a + V_2 b \neq 0} \frac{\|V_2 Z b\|}{\|U_1 a + V_2 b\|} \leq (1-\omega_1^2)^{-1/2}\|Z\|. \quad (37)$$

So, using (33)–(37) we have $\rho((I-NM)S) \leq (1-\omega_1^2)^{-1/2}(1+\|\Theta\Delta\Theta\|)\|S\|$, which concludes our proof. □

The following lemma shows that the influence of $\kappa(\Theta)$ need not be large if the spread of the δ_j is small.

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

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

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

So, the eigenvalues of the related system cluster around 1, and the tightness of the clustering is controlled through $\|S\|$. Note that the ω_1 term in Corollary 3.2 is no larger than the the corresponding term for the block-diagonally preconditioned system (Corollary 2.4). Likewise, the influence of the $\kappa(\Theta)$ term is smaller for the related system if the spread of the values δ_j is small. This will generally give us a tighter bound for the related system than for the block-diagonally preconditioned system.

3.2 Satisfying 'Constraints'

In the $D = 0$ case, the second block of equations in (1) often represents a set of constraints. For the $D \neq 0$ case, this may or 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 (28) 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 $k = 1, 2, \ldots$, of (28) satisfy $Mx_k + Qy_k = \tilde{g}$ in (3) and $Cx_k + Dy_k = g$ in (1).*

Proof: From (26)–(28) and the equality $MN = I + Q$ we have

$$\begin{aligned} Mx_{k+1} + Qy_{k+1} &= M(I - NM)Sx_k + M(I - NM)\tilde{f} + MN\tilde{g} + QMSx_k + QM\tilde{f} - Q\tilde{g} \\ &= (M + QM - MNM)(Sx_k + \tilde{f}) + (MN - Q)\tilde{g} \\ &= \tilde{g}. \end{aligned}$$

Thus, the fixed-point method satisfies the second block of equations of (28) exactly after one step. Because the block diagonal preconditioner (2) is invertible, the second block of equations of (1) are also satisfied after one step. □

Corollary 3.4. *After the first iteration of (28), 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 prove 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). We will return to the nomenclature of (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 $Cx^{(0)} = d$.*

Proof: We have $CAx + Cb = d$ for any x. Taking $x = 0$ implies $Cb = d$, and hence $CAx = d - Cb = 0$ for any x. Hence, $CA = 0$ must hold. Next, let $x^{(0)}$ be the initial guess for a Krylov method, and $Cx^{(0)} = d$. Then the initial residual is given by $r^{(0)} = b - (I - A)x^{(0)}$, and $Cr^{(0)} = Cb - Cx^{(0)} + CAx^{(0)} = 0$. For $m \geq 1$, the iterates of a Krylov method applied to $(I - A)x = b$ satisfy

$$x^{(m)} = x^{(0)} + \sum_{i=0}^{m-1} \alpha_i (I - A)^i r^{(0)} = x^{(0)} + \gamma_0 r^{(0)} + A \sum_{i=1}^{m-1} \gamma_i A^{i-1} r^{(0)}. \tag{38}$$

Finally, we multiply (38) by C, and note that $Cx^{(0)} = d$, $Cr^{(0)} = 0$ and $CA = 0$. Therefore,

$$Cx^{(m)} = Cx^{(0)} + \gamma_0 Cr^{(0)} + CA \sum_{i=1}^{m-1} \gamma_i A^{i-1} r^{(0)} = d. \tag{39}$$

□

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 (28) 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 (28).*

Proof: Use Theorem 3.5, with A as fixed-point iteration matrix in (28), $b = [\hat{f}^T\ \hat{g}^T]^T$, $C = [M\ Q]$ and $d = \hat{g}$. □

4 Inexact 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 an 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. Let $S_1 = -(D - CF_1^{-1}B^T)$ denote the actual Schur complement and S_2^{-1} denote our approximation to its inverse. 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,

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

We multiply (1) from the left by $\mathcal{P}(F_1, S_2)$. We refer to the resulting preconditioned matrix as $\mathcal{B}(S, \mathcal{E})$. The system of equations with $\mathcal{B}(S, \mathcal{E})$ looks as follows,

$$\begin{bmatrix} I - S & N \\ M_2 & Q_2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \quad = \quad \left(\begin{bmatrix} I & N \\ M & Q \end{bmatrix} - \begin{bmatrix} S & 0 \\ -\mathcal{E}M & -\mathcal{E}Q \end{bmatrix} \right) \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \tilde{f} \\ \tilde{g} \end{bmatrix}, \tag{40}$$

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 (40), we can bound the eigenvalues of $\mathcal{B}(S, \mathcal{E})$ by considering the perturbation of the eigenvalues of $\mathcal{B}(0)$ analogously to our bounds in Section 2.2.

Theorem 4.1. *Let $\lambda_{S,\mathcal{E}}$ be an eigenvalue of $\mathcal{B}(S, \mathcal{E})$, λ be an eigenvalue of $\mathcal{B}(0)$ and $Qv_j = \delta_j v_j$.*
1. If $\delta_j \in \mathbb{R}$, for $j = 1, \ldots, m$, then

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

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

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

3. If $D = 0$, then

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

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 [29, Theorem IV.1.12]),

$$\begin{aligned} |\lambda_S - \lambda| \quad &\leq \quad \left\| \mathcal{Y}^{-1} \begin{bmatrix} S & 0 \\ -\mathcal{E}M & -\mathcal{E}Q \end{bmatrix} \mathcal{Y} \right\| \\ &\leq \quad \left\| \mathcal{Y}^{-1} \begin{bmatrix} S & 0 \\ 0 & 0 \end{bmatrix} \mathcal{Y} \right\| + \left\| \mathcal{Y}^{-1} \begin{bmatrix} 0 & 0 \\ \mathcal{E}M & \mathcal{E}Q \end{bmatrix} \mathcal{Y} \right\|. \end{aligned} \tag{41}$$

Corollary 2.4 gives bounds for the first term in (41). 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

$$\begin{bmatrix} 0 & 0 \\ \mathcal{E}M & \mathcal{E}Q \end{bmatrix}\mathcal{Y} \;=\; \left[\begin{array}{cc|c} 0 & 0 & 0 \\ 0 & -\mathcal{E}V(I+\Delta\Lambda^+) & -\mathcal{E}V(I+\Delta\Lambda^-) \end{array}\right]. \tag{42}$$

We then multiply (42) from the left by $\mathcal{Y}^{-1}$, see (13)–(16), and refactor to yield

$$\mathcal{X} \;=\; \begin{bmatrix} 0 & 0 & 0 \\ 0 & (\Lambda^- - \Lambda^+)^{-1} & 0 \\ 0 & 0 & -(\Lambda^- - \Lambda^+)^{-1} \end{bmatrix}\begin{bmatrix} 0 & 0 & 0 \\ 0 & V^{-1}\mathcal{E}V & V^{-1}\mathcal{E}V \\ 0 & V^{-1}\mathcal{E}V & V^{-1}\mathcal{E}V \end{bmatrix}\begin{bmatrix} 0 & 0 & 0 \\ 0 & I+\Delta\Lambda^+ & 0 \\ 0 & 0 & I+\Delta\Lambda^- \end{bmatrix}.$$

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

$$\|\mathcal{X}\| \;\leq\; 2\|(\Lambda^- - \Lambda^+)^{-1}\| \max_j \left\{|1+\delta_j\lambda_j^+|, |1+\delta_j\lambda_j^-|\right\} \kappa(V)\|\mathcal{E}\|. \tag{43}$$

The remainder of the proof concerns the bounds on the right hand side of (43) for each particular case.

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

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

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 (43) to complete the proof of the the first bound.

For the second part of the theorem, we assume $\exists \alpha > 0$ s.t. $|\delta_j| \leq \alpha < \sqrt{5}$, for $j = 1, \ldots, m$. 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 $\|(\Lambda^- - \Lambda^+)^{-1}\| \leq \left(2\left(\sqrt{5} - \alpha\right)\right)^{-1/2}$. 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

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

Substituting these bounds into (43) yields

$$\|\mathcal{X}\| \quad \leq \quad \frac{2+(1+\sqrt{5})\alpha+2\alpha^2}{\sqrt{2\left(\sqrt{5}-\alpha\right)}}\kappa(V)\|\mathcal{E}\|. \tag{44}$$

We can then substitute this result into (41) 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 (41) 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 (41), since $Q = 0$, $\delta_j = 0$, so $\lambda_j^- - \lambda_j^+ = -\sqrt{5}$, and we can choose $V = I$. We then substitute this into (43). □

As a side note, in the complex case the α term will generally be modest in practice. 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}. Hence, this is about as good as we can hope for. The bounds also demonstrate that in terms of (bounds on) eigenvalue clustering 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 (28), with S_1^{-1} instead of S_2^{-1} [8]. Instead, we use an alternative splitting of $\mathcal{B}(S)$,

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

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,

$$\begin{bmatrix} I-(I-NM_2)S & -N\mathcal{E} \\ -M_2S & I+\mathcal{E} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \quad = \quad \begin{bmatrix} \hat{f} \\ \hat{g} \end{bmatrix}. \tag{45}$$

For a problem in magnetostatics, a linear system similar to (45) was derived in [24]. If we use the choices for the splitting and approximations from [24], we obtain basically the same system to be solved. In [24], 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 (45),*

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

Proof: Note that the matrix in (45) 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 [29]) 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

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

we obtain

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

□

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) and the norms of the inverses of the splitting and inexact 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 inexact Schur complement are reasonably accurate, the norms of their inverses will not be large relative to the norm of (1), unless (1) is itself poorly conditioned.

It is important to note that, as for the block-diagonally preconditioned system, the eigenvalue perturbation of the related system is dependent on both $\|S\|$ and $\|\mathcal{E}\|$. Again, there is no advantage to be had by 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

The stabilized finite element discretization of the Navier-Stokes equations provides a good model problem on which to demonstrate our results. Using the software toolkit for a 2D leaky lid-driven cavity [11], 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 $B = C$.

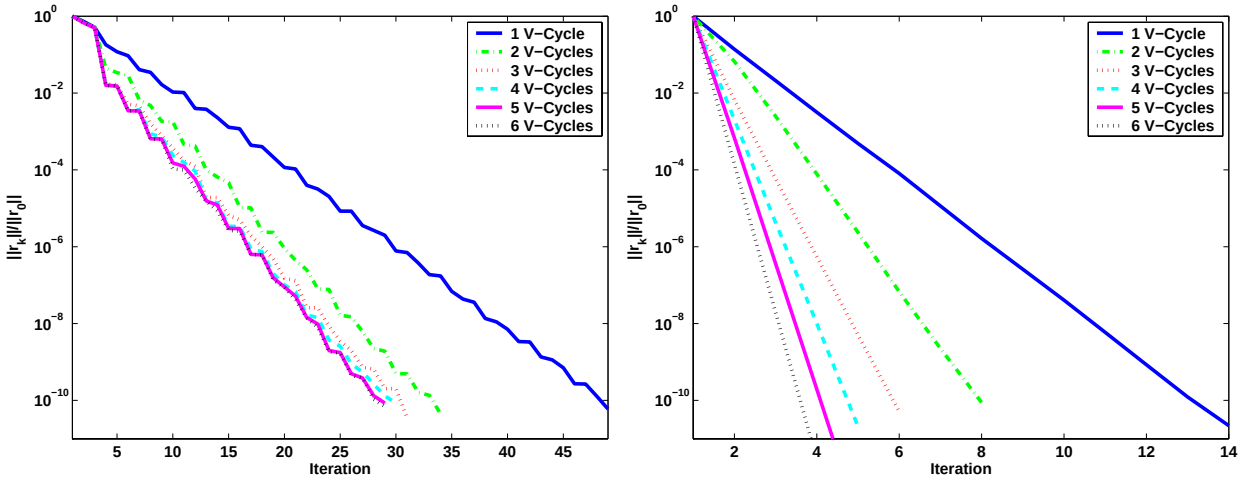
Plenty of excellent work has been done on preconditioners for this problem [11, 28, 30], which we do not intend to supplant. Rather, our goal is to illustrate the behavior of the preconditioners and bounds from this paper on a problem which is well-understood and easily accessible to the community. This is why we chose this problem rather than one where $B \neq C$ [3, 12] and [25, Sections 7.5 and 9.5], which might be less accessible.

In particular, we will show what happens to the eigenvalues, eigenvalue bounds and convergence of GMRES on the preconditioned problem, as we improve the splitting ($\|S\| \to 0$) and the inexact Schur

complement ($\|\mathcal{E}\| \to 0$). We will also compare the block-diagonally preconditioned system (3) and (40) with the related system (29) and (45), in terms of both eigenvalues and convergence. Finally, we shall provide examples to illustrate the importance of "balancing" the quality of the splitting and the Schur complement to avoid wasted effort.

For these experiments, we choose a 16×16 grid, viscosity parameter $\nu = .1$ and stabilization parameter $\beta = .25$. After removing the constant pressure mode, our final system is of size 705. Noting that multigrid cycles can be expressed as matrix splittings, we use a number of multigrid V-cycles for the splitting of our (1,1) block. For each V-cycle we use three SOR-Jacobi smoothing steps and relaxation parameter $\omega = .25$.

We start with the exact Schur complement, varying the number of V-cycles for the splitting from one to six. This demonstrates the relative performance difference between the block-diagonally preconditioned system and the related system. Figures 1(a) and 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.



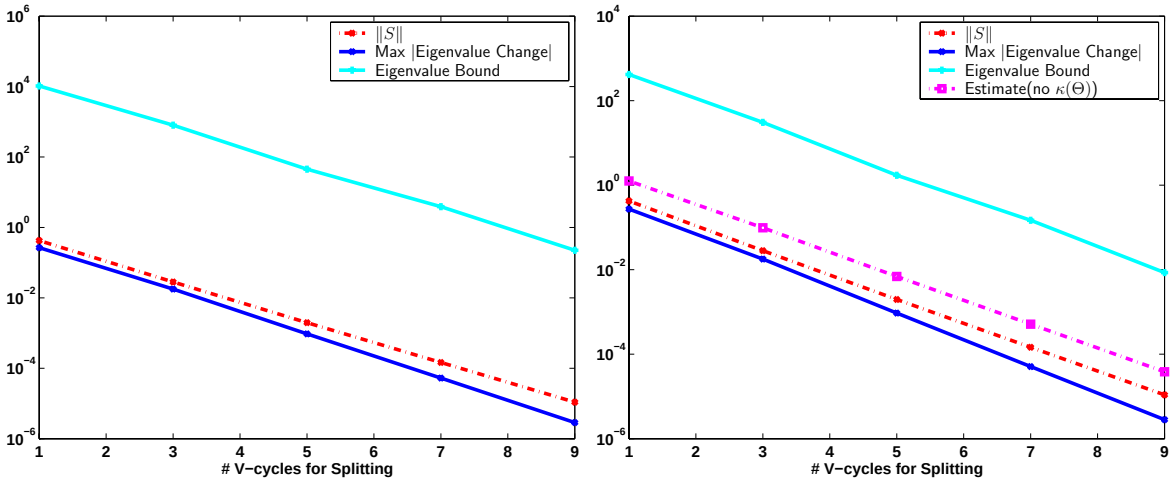
(a) Block-Diagonal Preconditioner (3)

(b) Related System (29)

Figure 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 bounds and the eigenvalue perturbations for both preconditioned systems, using up to nine V-cycles for the splitting, with the exact Schur complement. The results explain the difference in convergence between the block-diagonally preconditioned system and the related system. Figure 2(a) shows the maximum absolute eigenvalue perturbation from $\lambda \in \{1, \lambda_j^{\pm}\}$ for the block-diagonally preconditioned system, and Figure 2(b) shows the maximum absolute eigenvalue perturbation from 1 for the related system (29). Note how the eigenvalue perturbation follows the trend of both $\|S\|$ and the eigenvalue bound, although the bound is pessimistic.

As we use a better splitting for A (more V-cycles), we see that the eigenvalue perturbations decrease with approximately the same rate as the corresponding bound. Though the bound is pessimistic, this is mostly



(a) Block-Diagonal Preconditioner (3)

(b) Related System (29)

Figure 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

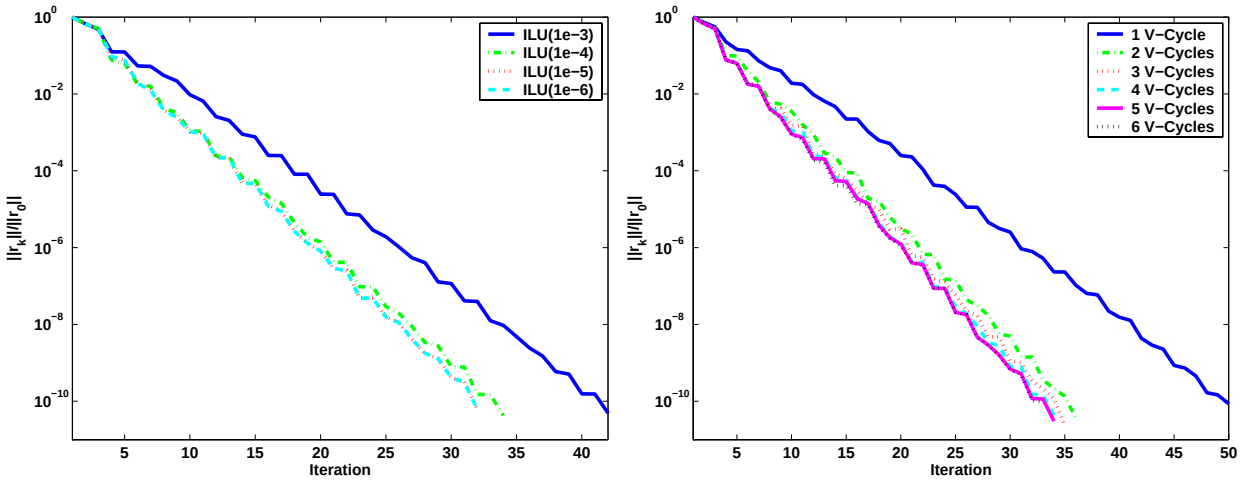
due to the $\kappa(\Theta)$ term. Figure 2(b) includes an "estimate" of the perturbation bound for the related system, which consists of the bound in Corollary 3.2 with $\kappa(\Theta)$ replaced by one. Both the bound and our "estimate" follow the trend in the actual eigenvalue perturbation well as the number of V-cycles increases. This shows that the bounds and the estimate give a good qualitative description of the behavior of the eigenvalues as the splitting improves.

In comparing the block-diagonally preconditioned system (3) with the related system (29) we note that the eigenvalue perturbation bound is about a factor five to ten smaller for the related system. This is largely because the bound for the related system has different (and smaller) terms involving ω_1 and $\kappa(\Theta)$. However, the actual maximum eigenvalue perturbation for both systems is about equal. For the related system, this represents a single eigenvalue cluster around 1. For the block-diagonally preconditioned system, this represents $2m+1$ (potentially) distinct clusters around 1 and $\lambda_j^{\pm}$, for $j = 1, \ldots, m$. The existence of multiple clusters in this 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 shown in Figure 1(a).

We choose to illustrate the convergence behavior for the preconditioner with an inexact Schur complement, as a function of the accuracy of the approximation, by using an ILU decomposition with a drop tolerance [26]. 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 $1e-6$.

We start by varying the drop tolerance for the inexact Schur complement and fix the number of V-cycles for the splitting at four. We will also vary the number of V-cycles for the splitting and fix the inexact Schur complement's drop tolerance to $1e-4$. This allows us to see the effects of improving the splitting and

the inexact Schur complement. Figures 3 and 4 show the convergence of GMRES for the block-diagonally preconditioned system (40) and the related system (45) respectively.



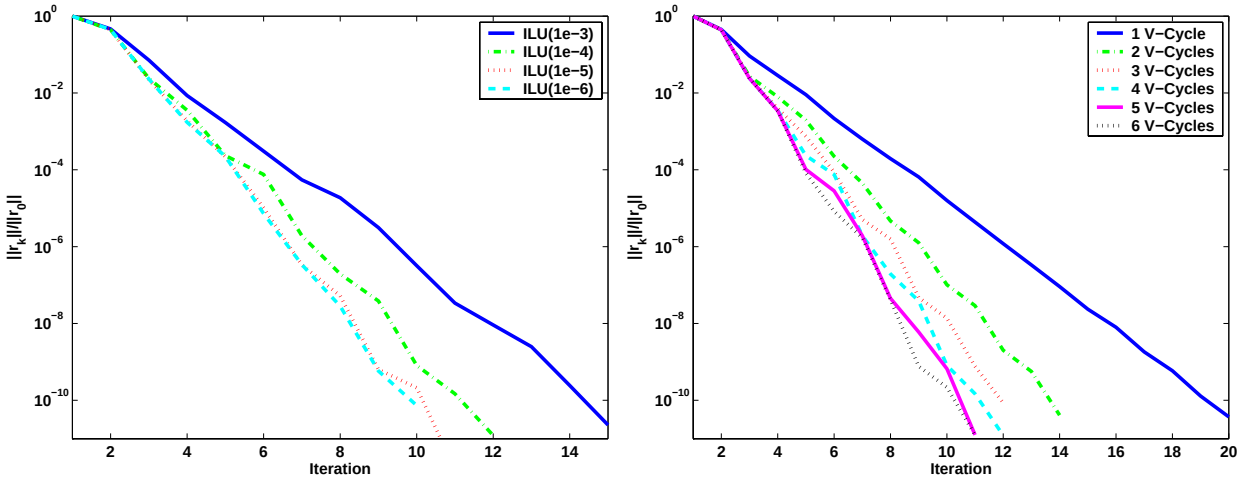
(a) Using four V-cycles for the splitting and varying the inexact Schur complement

(b) Using the inexact Schur complement with ILU($1e-4$) and varying the number of V-cycles for the splitting

Figure 3. Convergence results for the block-diagonal preconditioner using the inexact Schur complement

Like the results in Figure 1(a), the convergence rate in Figures 3 and 4 hits a point of diminishing returns, past which improving either the splitting or the inexact Schur complement while leaving the other unchanged does not improve convergence. To show why this occurs, we consider the eigenvalue perturbation using a very accurate splitting for the (1,1) block, A, namely nine V-cycles. For the block-diagonally preconditioned system Figure 5(a) shows the maximum absolute eigenvalue perturbation, the $\|\mathcal{E}\|$, the eigenvalue bound and $\|S\|$ (for reference purposes). Similar results for the related system are shown in Figure 5(b). Note that while pessimistic, the bound and $\|\mathcal{E}\|$ capture the general trend in the eigenvalue perturbation — with a good enough splitting, improving the accuracy of the inexact Schur complement will lead to better eigenvalue clustering.

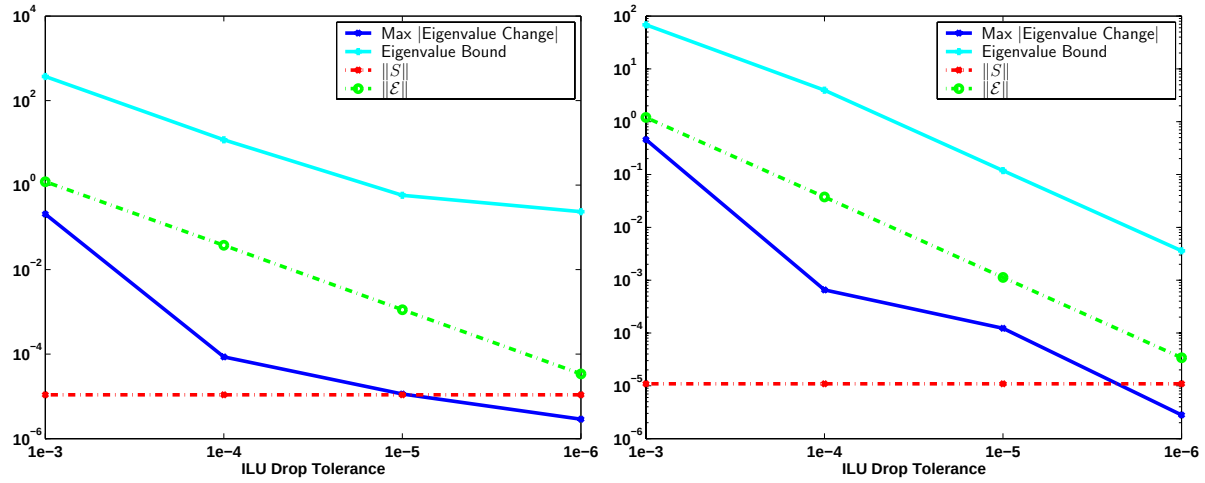
As we noted in Theorems 4.1 and 4.2, the bound for the eigenvalue perturbation of the related system depends on both $\|S\|$ and $\|\mathcal{E}\|$. Making one significantly smaller than the other is not effective. Figure 6 shows the bound for the block-diagonally preconditioned system, $\|S\|$ and $\|\mathcal{E}\|$, as well as the maximum absolute eigenvalue perturbation. Here we vary the ILU drop tolerance for the inexact Schur complement, fixing the number of V-Cycles at five (Figure 6(a)). We also vary the number of V-cycles for the splitting and fix the ILU tolerance at $1e-4$ (Figure 6(b)). Note that in both plots, once $\|S\|$ is less than $\|\mathcal{E}\|$, or vice versa, the eigenvalue perturbation ceases to decrease shortly thereafter. This suggests that the behavior of the bounds is indicative of the actual eigenvalue perturbation, and that undue attention to either the splitting or the Schur complement yields little additional benefit. Similar results for the related system are shown in Figure 7.



(a) Using four V-Cycles for the splitting and varying the inexact Schur complement

(b) Using the inexact Schur complement with ILU($1e-4$) and varying the number of V-cycles for the splitting

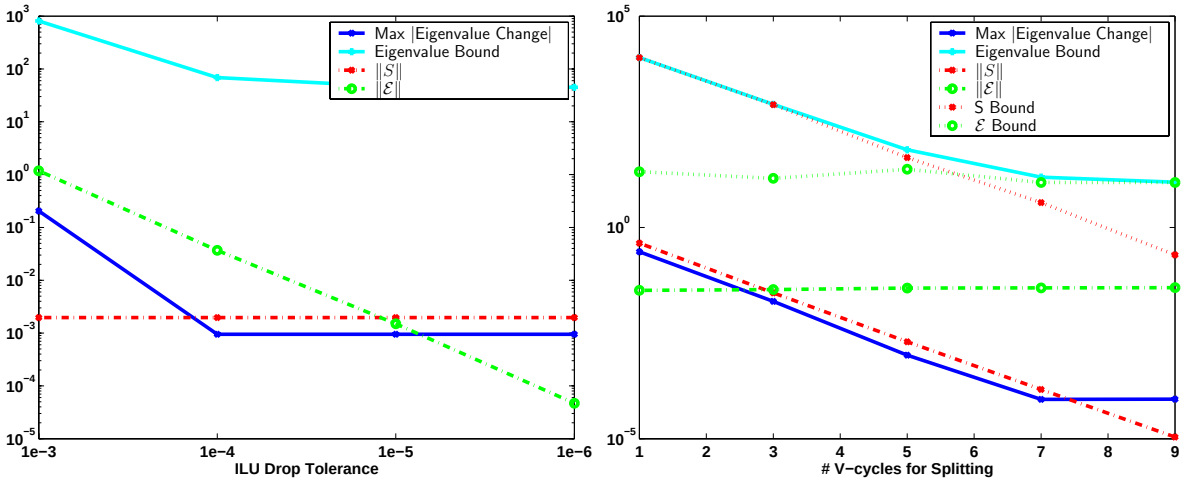
Figure 4. Convergence results for the related system using the inexact Schur complement



(a) Block-Diagonal Preconditioner (40)

(b) Related System (45)

Figure 5. Maximum absolute eigenvalue perturbation and perturbation bounds, for both types of preconditioners, using inexact Schur complements of varying accuracy and 9 V-cycles for the splitting



(a) Using five V-Cycles for the splitting and varying the inexact Schur complement

(b) Using the inexact Schur complement with ILU(1e − 4) and varying the number of V-cycles for the splitting

Figure 6. The effects of $\|S\|$ and $\|\mathcal{E}\|$ on eigenvalues and bounds for the block-diagonal preconditioner using the inexact Schur complement

6 Conclusions and Future Work

We have discussed and analyzed block-diagonal preconditioners and efficient variants of indefinite preconditioners for the $D \neq 0$ case, including the use of inexact Schur complements. We have illustrated this analysis using a well-known model problem and evaluated the bounds numerically. This allowed us to demonstrate the predictive power of the analysis in terms of actual eigenvalue and convergence behavior.

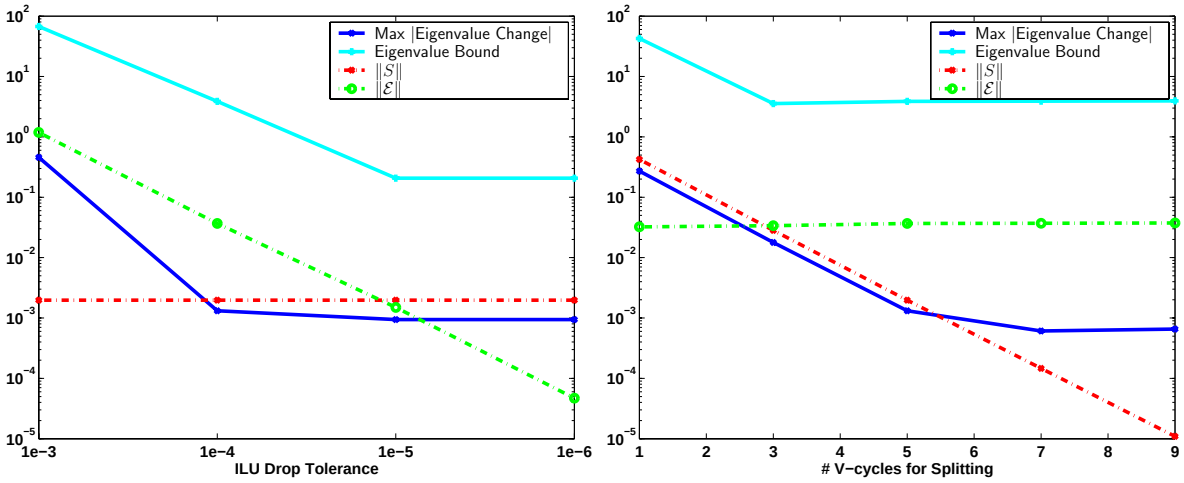
In this paper, we have focused on developing two classes of preconditioners and their analysis. While there is still work to be done in the realm of analysis, there is also the issue of specializing the methodology to particular problems. We are working on applications in areas such as metal deformation, optimization and electronics.

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. Technical report, Emory University, September 2003. To appear in SIAM Journal on Matrix Analysis and Applications.

[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.



(a) Using five V-Cycles for the splitting and varying the inexact Schur complement

(b) Using the inexact Schur complement with ILU(1e − 4) and varying the number of V-cycles for the splitting

Figure 7. The effects of $\|S\|$ and $\|\mathcal{E}\|$ on related system using the inexact Schur complement

[4] D. Braess. *Finite Elements: Theory, fast solvers and applications in solid mechanics*. Cambridge University Press, 2nd edition, 2001.

[5] D. Braess, P. Deuflhard, 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. *Mathematics of Computation*, 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. Technical Report 36-2003, Institute of Mathematics, Technical University of Berlin, September 2003. Accepted for publication in SIAM J. on Sci. Comput.

[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] A. Forsgren, P.E. Gill, and M.H. Wright. Interior methods for nonlinear optimization. *SIAM Review*, 44(4):525–597, 2002.

[13] 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.

[14] 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.

[15] L.A. Hageman and D.M. Young. *Applied Iterative Methods*. Academic Press, 1981.

[16] R.A. Horn and C.R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

[17] I.C.F. Ipsen. A note on preconditioning nonsymmetric matrices. *SIAM J. Sci. Comput.*, 23(3):1050–1051, 2001.

[18] C. Keller, N.I.M. Gould, and An.J. Wathen. Constraint preconditioning for indefinite linear systems. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1300–1317, 2000.

[19] Piotr Krzyżanowski. On block preconditioners for nonsymmetric saddle point problems. *SIAM J. Sci. Comput.*, 23(1):157–169, 2001.

[20] 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.

[21] C. Meyer. *Matrix Analysis and Applied Linear Algebra*. SIAM, 2001.

[22] 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.

[23] R. Nicolaides. Existence, uniqueness and approximation for generalized saddle point problems. *SIAM Journal on Numerical Analysis*, 19(2):349–357, 1982.

[24] I. Perugia and V. Simoncini. Block-diagonal and indefinite symmetric preconditioners for mixed finite element formulations. *Numer. Linear Algebra Appl.*, 7:585–616, 2000.

[25] A. Quarteroni and A. Valli. *Numerical Approximation of Partial Differential Equations*. Springer-Verlag, 2nd edition, 1997.

[26] Y. Saad. ILUT: a dual threshold incomplete ILU factorization. *Numerical Linear Algebra with Applications*, pages 387–402, 1994.

[27] 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.

[28] 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.

[29] G.W. Stewart and J.G. Sun. *Matrix perturbation theory*. Academic Press Inc., Boston, 1990.

[30] 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.

[31] 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.