Field of values analysis that includes zero for preconditioned nonsymmetric saddle-point systems

Hao Chen and Chen Greif Department of Computer Science The University of British Columbia

Abstract

We present a field-of-values (FOV) analysis for preconditioned nonsymmetric saddlepoint linear systems, where zero is included in the field of values of the matrix. We rely on recent results of Crouzeix and Greenbaum [Spectral sets: numerical range and beyond. SIAM Journal on Matrix Analysis and Applications, 40(3):1087-1001, 2019], showing that a convex region with a circular hole is a spectral set. Sufficient conditions are derived for convergence independent of the matrix dimensions. We apply our results to preconditioned nonsymmetric saddle-point systems, and show their applicability to families of block preconditioners that have not been previously covered by existing FOV analysis. A limitation of our theory is that the preconditioned matrix is required to have a small skew-symmetric part in norm. Consequently, our analysis may not be applicable, for example, to fluid flow problems characterized by a small viscosity coefficient. Some numerical results illustrate our findings.

Keywords. field of values, nonsymmetric saddle-point systems, GMRES convergence, block preconditioners

1 Introduction

The field of values of a matrix is an indispensable tool in linear algebra and its applications. It is defined as follows.

Definition 1.1. Given a matrix $A \in \mathbb{C}^{n \times n}$, the field of values (FOV) of A is defined as

$$W(A) = \left\{ \frac{x^* A x}{x^* x} : \quad x \in \mathbb{C}^n \right\}$$

and the H-field of values of A, given another matrix $H \in \mathbb{C}^{n \times n}$, is defined as

$$W_H(A) = \left\{ \frac{x^* H A x}{x^* H x} : \quad x \in \mathbb{C}^n \right\}.$$

Early work on the topic was published in [11, 16] and in several other papers; see [3] for a recent expository paper that provides an overview of the use of FOV, its history and development, and a comprehensive list of references.

In the context of this work, we are interested in the use of FOV to establish the scalability of Krylov subspace iterative solvers (specifically, GMRES [20]) for large and sparse nonsymmetric saddle-point systems:

$$\begin{bmatrix} F & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} u \\ p \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix},\tag{1}$$

where $F \in \mathbb{R}^{n \times n}$ is nonsymmetric, $B \in \mathbb{R}^{m \times n}$ has full row rank, and $u, f \in \mathbb{R}^n, p, g \in \mathbb{R}^m$.

Significant work has been done on this topic [2, 4, 16, 17, 19], but to the best of our knowledge, the analysis is limited to the situation where 0 is not included in the field of values. Our goal is to perform an FOV analysis for preconditioned saddle-point systems in the case where the origin is included.

Part of our motivation in considering the field of values is that spectral analysis may be limited for this family of linear systems: for nonsymmetric saddle-point systems arising from PDEs, the condition number of the eigenvector matrix of the preconditioned matrix typically increases with the matrix dimensions. In the context of the Navier-Stokes equations, for example, this happens even for a large viscosity coefficient. Thus, an analysis of the eigenvalues of the preconditioned matrix is often insufficient to theoretically prove convergence for nonsymmetric saddle-point matrices.

Throughout this paper, we extensively use the notion of a weighted norm, which we define as follows.

Definition 1.2. Given a Hermitian positive definite matrix $H \in \mathbb{C}^{n \times n}$, the H-norm of a vector $u \in \mathbb{C}^n$ is defined as

$$||u||_H = (u, Hu)^{1/2}.$$

Following the terminology of [7, Eq. (1)], while restricting our attention to discrete linear operators, polynomials, and the *H*-norm, we say that for a closed subset $X \subset \mathbb{C}$ and a matrix A, X is a *K*-spectral set for A if for any polynomial p

$$||p(A)||_H \le K \sup_{z \in X} |p(z)|.$$

Theorem 1.3. [8, Theorem 6] Let A be a matrix of the same dimensions as H. Then, $W_H(A)$ is a $(1 + \sqrt{2})$ -spectral set for A.

In the sequel, we will be using GMRES with respect to *H*-norm, or equivalently the *H*-weighted inner product $\langle \cdot, \cdot \rangle_H$. Applying Theorem 1.3, we obtain the following convergence bound.

Theorem 1.4. [7] Let $r_k = b - Ax_k$ be the residual of the k-th iteration, x_k , of GMRES with respect to the H-norm applied to the linear system Ax = b of the residual, and let \mathbf{P}_j denote all polynomials p of degree $\leq j$ that satisfy p(0) = 1. Then,

$$\frac{\|r_j\|_H}{\|r_0\|_H} \le \min_{p \in \mathbf{P}_j} \|p(A)\|_H \le (1 + \sqrt{2}) \min_{p \in \mathbf{P}_j} \max_{z \in W_H(A)} |p(z)|.$$

...

A challenge is that when $0 \in W_H(A)$, we have $\min_{p \in \mathbf{P}_j} \max_{z \in W_H(A)} |p(z)| = 1$, and Theorem 1.4 fails to provide a useful bound on GMRES convergence in this case. The presence of a zero in the field of values is, in fact, common in saddle-point systems: the (2,2)-block of a saddle-point system preconditioned with a block-diagonal matrix can be 0. Recently, Crouzeix and Greenbaum [7] have defined a convex region with a circular hole and showed that it is a spectral set. This can be used to analyze cases when zero is included in the field of values.

Theorem 1.5. [7] Let A be a matrix of the same dimensions as H. Then, $\Omega_{CG} = W_H(A) \cap \{z \in C : |z| \ge \|A^{-1}\|_H^{-1}\}$ is a $(3 + 2\sqrt{3})$ -spectral set for A.

In [14], the author presents a simple example to illustrate the potential of this result in the context of convergence of GMRES.

The field of values of a matrix is difficult to compute, and in the context of the iterative solution of linear systems, it is necessary to exploit the specific properties of the matrices involved in order to provide concrete conditions for scalability of an iterative solver. Our work extends the family of saddle-point linear systems for which FOV analysis is applicable. In particular, we consider block-diagonal preconditioners and certain block-triangular preconditioners for which no previous FOV analysis is available. On the other hand, our analysis has some limitations compared to the well-studied FOV analysis that excludes the origin. For example, in [17], scaling is effectively used to allow for applying FOV analysis to the discrete Navier-Stokes equations with a small viscosity coefficient when the field of values does not include the origin. In our analysis we are not able to utilize scalings in the same manner, and we require the skew-symmetric part of the linear system to be small norm-wise.

The remainder of this paper is structured as follows. In Section 2 we present an analysis that deals with zero in the field of values. In Section 3 we specialize our results to saddle-point systems. In Section 4 we discuss a few examples of relevant applications and present some numerical results. Finally, we draw some conclusions in Section 5.

2 FOV Analysis that Includes Zero

In this section, we derive sufficient conditions that will serve us in our analysis for saddlepoint systems.

2.1 Preliminaries

Let us present a few known results that we will use in our analysis.

Definition 2.1. For two symmetric positive definite matrices $H_1 \in \mathbb{R}^{n \times n}$ and $H_2 \in \mathbb{R}^{m \times m}$, we define the (H_1, H_2) -norm for a matrix $M \in \mathbb{R}^{m \times n}$ as

$$||M||_{H_1,H_2} = \max_{v \neq 0} \frac{||Mv||_{H_2}}{||v||_{H_1}}.$$

It is immediate from Definition 2.1 that

$$\|H_2^{-1/2}MH_1^{-1/2}\|_2 = \|M\|_{H_1,H_2^{-1}} = \|MH_1^{-1}\|_{H_1^{-1},H_2^{-1}} = \|H_2^{-1}M\|_{H_1,H_2}$$

The following properties from [17] are useful for our analysis.

Lemma 2.2 ([17, Lemma 1]). Let $M \in \mathbb{R}^{m \times n}$ have full rank, and let $H_1 \in \mathbb{R}^{n \times n}$, $H_2 \in \mathbb{R}^{m \times m}$ be two symmetric positive definite matrices. Then

- (i) $||M||_{H_1,H_2^{-1}} = \max_{v \in \mathbb{R}^n \setminus \{0\}} \max_{w \in \mathbb{R}^m \setminus \{0\}} \frac{w^T M v}{||v||_{H_1} ||w||_{H_2}}.$
- (ii) If m = n,

$$\|M^{-1}\|_{H_2^{-1},H_1}^{-1} = \min_{v \in \mathbb{R}^n \setminus \{0\}} \max_{w \in \mathbb{R}^m \setminus \{0\}} \frac{w^T M v}{\|v\|_{H_1} \|w\|_{H_2}}.$$

(iii) If $H_i \in \mathbb{R}^{n_i \times n_i}$, i = 1, 2, 3 are three symmetric and positive definite matrices and $R \in \mathbb{R}^{n_1 \times n_2}$, $Q \in \mathbb{R}^{n_2 \times n_3}$ then

$$||RQ||_{H_3,H_1} \le ||Q||_{H_3,H_2} ||R||_{H_2,H_1}.$$

The following result from [10], adapted to our notation and context, is useful in our analysis.

Theorem 2.3 ([10, Theorem 1]). Let P_n denote the set of polynomials p of degree at most n with p(0) = 1. For a compact set S in the complex plane, with the origin not included in or surrounded by S and no isolated points, define

$$E_n(S) = \min_{p \in P_n} \max_{z \in S} |p(z)|$$

and the corresponding estimated asymptotic convergence factor

$$\rho = \lim_{n \to \infty} (E_n(S))^{1/n}$$

Let g(z) be the Green's function associated with S, defined in the exterior of S, satisfying $\nabla^2 g = 0$ outside of S, $g(z) \to 0$ as $z \to \partial S$, and $g(z) - \log |z| \to C$ as $|z| \to \infty$ for some constants C. Then,

$$\rho = \exp(-g(0)).$$

2.2 Sufficient Conditions

Lemma 2.4. Let $A, H \in \mathbb{R}^{n \times n}$ where A is nonsingular and H is symmetric positive definite. Then, GMRES converges with respect to H-norm in a fixed number of iterations independent of its dimension, n, if the following conditions hold for some constants a, b, and c:

$$||A||_H \le a; \tag{2a}$$

$$\|A^{-1}\|_H \le b; \tag{2b}$$

$$||(HA - A^T H)/2||_{H,H^{-1}} \le c;$$
(2c)

$$bc < 1.$$
 (2d)

Proof. We first derive a bound on the field of values of A. Suppose the conditions hold. Then, for any $z \in W_H(A)$, we have $|z| \leq ||A||_H \leq a$ and

$$\begin{split} |Im(z)| &\leq \max_{x \in \mathbb{C}^n} \left| \left(\frac{x^* HAx}{x^* Hx} - \left(\frac{x^* HAx}{x^* Hx} \right)^* \right) \Big/ 2 \right| \\ &= \max_{x \in \mathbb{C}^n} \left| \left(\frac{x^* (HA - A^T H)x}{2x^* Hx} \right) \right| \\ &\leq \| (HA - A^T H)/2 \|_{H, H^{-1}} \leq c. \end{split}$$

Then,

$$\Omega_{CG} \subseteq \Omega_D := \{ z : \frac{1}{b} \le |z| \le a \} \cap \{ z \in \mathbb{C} : |Im(z)| \le c \}.$$

By Theorem 1.5, we have the GMRES convergence result

$$\frac{\|r_j\|_H}{\|r_0\|_H} \le \min_{p \in \mathbf{P}_j} \|p(A)\|_H \le (3 + 2\sqrt{3}) \min_{p \in \mathbf{P}_j} \max_{z \in \Omega_{CG}} |p(z)|.$$



Figure 1: The shaded region is Ω_D when conditions (2a)–(2d) of Lemma 2.4 hold

Since Condition (2d) holds, the origin is not surrounded by Ω_{CG} , and it follows from Theorem 2.3 that there is always a polynomial of (of some degree) with value 1 at the origin that has a maximum magnitude strictly less than 1 on the closure of this set and hence GMRES converges with an asymptotic rate given by $\exp(-g(0)) < 1$, where g is the Green's function of this set with a pole at ∞ [6, 7].

Remark 2.1. If condition (2d) of Lemma 2.4 does not hold, the iterative solver may still converge but we cannot prove convergence using our technique of proof. Specifically, it is immediate to see that Ω_{CG} is connected and due to the maximum modulus principle, we can only obtain $\min_{p \in \mathbf{P}_j, p(0)=1} \max_{z \in \Omega_{CG}} |p(z)| \ge 1$, which does not indicate convergence; see Figure 2 for a graphical illustration.



Figure 2: The shaded region is Ω_D when $bc \ge 1$ (i.e., when condition (2d) of Lemma 2.4 is violated)

2.3 Scope and Limitations of the Analysis

Recall a widely used definition of FOV equivalence (see, for example, [17]).

Definition 2.5. Given two nonsingular matrices $M, A \in \mathbb{R}^{n \times n}$, M is <u>*H*-field-of-values</u> equivalent to A if there exist positive constants α, β independent of n such that

$$\alpha \le \frac{(MAx, x)_H}{(x, x)_H}, \ \frac{\|MAx\|_H}{\|x\|_H} \le \beta.$$
(3)

If M is H-field-of-values equivalent to A, the FOV of MA is bounded by a well-defined region:

$$W_H(MA) \subseteq \Omega_{\text{FOV}} := \{ z : \alpha \le \text{Re}(z), |z| \le \beta \}.$$

For a geometric illustration of Ω_{FOV} , see Figure 3.



Figure 3: The shaded region is Ω_{FOV} with $\alpha = 0.5$ and $\beta = 1$

The analysis in [17] and elsewhere (see, e.g., [16]) pertaining to the case that 0 is not part of the field of values is based on obtaining convergence independent of the matrix dimensions (or mesh size when discretizations of partial differential equations are concerned) by scaling the preconditioner or the inner product. In that case, Definition 2.5 is a convergence criterion and it allows for making α and β arbitrary (positive) and independent of the matrix dimensions. In contrast, in our case, condition (2d) requires bc to be small. While scaling reduces one of b or c, it increases the other. Therefore, a simple scaling strategy does not work in the case we are considering, which reveals a limitation of our analysis. We note that condition (2b) is rather standard by norm equivalence considerations (see, for example, [17, Lemma 3]). It is condition (2c) that seems to present the difficulty, because it requires the skew-symmetric part of the operator to be smaller than the radius of the inner disk; see Figure 1. Therefore, practically speaking, our analysis is limited to cases where the preconditioned matrix is only mildly nonsymmetric.

However, we note that this lemma can be improved to allow for looser conditions by using a more sophisticated analysis.

Example 2.6. This is a modified example from [14]:

$$A = A_{-1} \oplus A_{+1}$$

where $A_{-1} \in \mathbb{R}^{n \times n}$ and $A_{+1} \in \mathbb{R}^{n \times n}$ are given by

$$A_{-1} = \begin{bmatrix} -1 & 1/4 & & \\ & -1 & 1/4 & & \\ & \ddots & \ddots & \\ & & & -1 & 1/4 \\ & & & & -1 \end{bmatrix}$$

and

$$A_{+1} = \begin{bmatrix} 2 & 1.2 & & \\ & 2 & 1.2 & \\ & & \ddots & \ddots & \\ & & & 2 & 1.2 \\ & & & & & 2 \end{bmatrix}.$$

The field of values of A is a convex hull of two disks centered at -1 with radius 1/4 and at 2 with radius 1.2, independently of the matrix dimensions. The inverse A^{-1} is available analytically, and it can be shown that $||A^{-1}||_2^{-1} \rightarrow \frac{3}{4}$ as $n \rightarrow \infty$; see, for example, [15] for useful relevant results for Toeplitz matrices. For a finite value of n, the norm needs to be computed numerically, and we have experimentally observed that it is bounded between 0.74 and 0.76 for relatively modest values of n.

We provide a graphical illustration in Figure 4. Here c = 1.2 and $b \ge \frac{1}{0.76}$. The condition (2d) is violated, but GMRES would still converge for a linear system with the matrix A because Ω_{CG} does not surround/include the origin. A more careful analysis that tracks the boundary of the FOV (see, e.g., [18]) might result in conditions that are easier to satisfy.

While the limitation we have noted is considerable, our analysis substantially broadens the scope of preconditioners for which FOV analysis can be carried out. In particular, in terms of the quantities of Definition 2.5, our analysis makes it possible to consider

$$\frac{(MAx,x)_H}{(x,x)_H} \le 0. \tag{4}$$

In the upcoming sections, we present specific examples related to discretized fluid flow problems that demonstrate the advantages and the limitations of our analysis.



Figure 4: The shaded region is Ω_{CG} for A.

3 Preconditioned Saddle-Point Matrices with Zero in the Field of Values

Using the results of Lemma 2.4, we now apply our theory to the important case of a non-symmetric saddle-point system.

Consider

$$K = \begin{bmatrix} F & B^T \\ B & 0 \end{bmatrix},\tag{5}$$

where $F \in \mathbb{R}^{n \times n}$ is nonsingular and $B \in \mathbb{R}^{m \times n}$ is full row rank. We assume that F is nonsymmetric and positive real (or positive definite), namely, that $u^T F u > 0$ for all $0 \neq u \in \mathbb{R}^n$.

Let

$$S = BF^{-1}B^T \tag{6}$$

be the Schur complement, and define

$$H = \begin{bmatrix} H_1 & 0\\ 0 & H_2 \end{bmatrix},\tag{7}$$

where $H_1 \in \mathbb{R}^{n \times n}$ and $H_2 \in \mathbb{R}^{m \times m}$ are symmetric positive definite.

To be able to perform our analysis, we need to make some specific assumptions on H_1 and H_2 . We note that these assumptions amount to sufficient conditions, and in practice one may relax them.

Definition 3.1. We set H_1 as the symmetric part of F, and define N as its skew-symmetric part:

$$F = H_1 + N, \qquad H_1 = \frac{F + F^T}{2}, \qquad N = \frac{F - F^T}{2}.$$
 (8)

Note that H_1 is symmetric positive definite by our assumptions on F.

Assumption 3.2. Let α be a constant independent of the matrix dimensions, such that

$$\|N\|_{H_1, H_1^{-1}} \le \alpha. \tag{9}$$

Lemma 3.3. A bound on the weighted norm of F is given by

$$\|F\|_{H_1, H_1^{-1}} \le (1+\alpha).$$

Proof. This is immediate from equation (9).

Lemma 3.4. The inverse of F satisfies

$$||F^{-1}||_{H_1^{-1},H_1} \le 1.$$

Proof. The result can be readily deduced by using the standard property of norms

$$||F^{-1}||_{H_{1}^{-1},H_{1}}^{-1} = \min_{v \in \mathbb{R}^{n} \setminus \{0\}} \max_{w \in \mathbb{R}^{n} \setminus \{0\}} \frac{w^{T} F v}{||v||_{H_{1}} ||w||_{H_{1}}}$$

$$\geq \min_{v \in \mathbb{R}^{n} \setminus \{0\}} \frac{v^{T} F v}{||v||_{H_{1}}^{2}} \geq 1,$$

which is stated in result (ii) of Lemma 2.2 (see also [17, Lemma 1]).

In the problems that we consider, we will assume boundedness of B and a standard inf-sup condition, both of which in fact impose a condition on the choice of H_2 .

Assumption 3.5.

$$||B||_{H_1, H_2^{-1}} \le C_1, \qquad \min_x \frac{||B^T x||_{H_1^{-1}}}{||x||_{H_2}} \ge C_2, \tag{10}$$

where C_1 and C_2 are independent of α and the dimensions of K.

Lemma 3.6. If Assumption 3.5 holds, then

$$||S^{-1}||_{H_2^{-1},H_2} \le (1+\alpha)^2 / C_2^2.$$

Proof. Using (10) and following similar steps to the analysis of [17], we have

$$\begin{split} \|S^{-1}\|_{H_{2}^{-1},H_{2}}^{-1} &= \min_{v \in \mathbb{R}^{m} \setminus \{0\}} \max_{w \in \mathbb{R}^{m} \setminus \{0\}} \frac{w^{T}BF^{-1}B^{T}v}{\|v\|_{H_{2}}} \\ &\geq \min_{v \in \mathbb{R}^{m} \setminus \{0\}} \frac{v^{T}BF^{-1}B^{T}v}{\|v\|_{H_{2}}^{2}} \\ &\geq \min_{v \in \mathbb{R}^{m} \setminus \{0\}} \frac{v^{T}BF^{-1}B^{T}v}{v^{T}BH_{1}^{-1}B^{T}v} \min_{v \in \mathbb{R}^{m} \setminus \{0\}} \frac{\|B^{T}v\|_{H_{1}^{-1}}^{2}}{\|v\|_{H_{2}}^{2}} \\ &\geq C_{2}^{2} \min_{v \in \mathbb{R}^{m} \setminus \{0\}} \frac{v^{T}BF^{-1}B^{T}v}{v^{T}BH_{1}^{-1}B^{T}v}. \end{split}$$

Using [17, Lemma 8] and lemma 3.3, we have

$$\min_{v \in \mathbb{R}^m \setminus \{0\}} \frac{v^T B F^{-1} B^T v}{v^T B H_1^{-1} B^T v} \geq \min_{y \in \mathbb{R}^n \setminus \{0\}} \frac{y^T (I + H_1^{-1/2} N H_1^{-1/2})^{-1} y}{y^T y} \\
= \min_k \operatorname{Re} \left(\frac{1}{\lambda_k (I + H_1^{-1/2} N H_1^{-1/2})} \right) \\
= \min_k \operatorname{Re} \left(\frac{1}{1 + \lambda_k (H_1^{-1/2} N H_1^{-1/2})} \right) \\
= \frac{1}{\max_k \left| \lambda_k (I + H_1^{-1/2} N H_1^{-1/2}) \right|^2} \\
\geq \frac{1}{\|H_1^{-1/2} F H_1^{-1/2}\|_2^2} \\
\geq \frac{1}{(1 + \alpha)^2},$$

which completes the proof.

Finally, we establish notation that will become handy in the following subsections.

Definition 3.7. For a matrix T and a scalar τ , we say that $||T|| \leq \tau$ if $||T|| \leq C\tau$, where C is a constant independent of the dimensions of T.

3.1 Block-Triangular Preconditioners

Let us consider two block-triangular preconditioners:

(i) upper block-triangular preconditioners of the form

$$M_U = \begin{bmatrix} F & B^T \\ 0 & H_2 \end{bmatrix},\tag{11}$$

with left preconditioning under the H-norm;

(ii) lower block-triangular preconditioners of the form

$$M_L = \begin{bmatrix} F & 0\\ B & H_2 \end{bmatrix},\tag{12}$$

with right preconditioning under the H^{-1} -norm.

It is well known that there are some differences in the use of left and right preconditioners. For example, in flexible GMRES it is necessary to use right preconditioning. The correct norm considered in GMRES for finite element discretizations should typically be $\|\cdot\|_{H^{-1}}$ [1].

Consider first the left preconditioner M_U . The preconditioned matrix is given by

$$M_U^{-1}K = \begin{bmatrix} I - F^{-1}B^T H_2^{-1}B & F^{-1}B^T \\ H_2^{-1}B & 0 \end{bmatrix},$$

and its inverse, which is required in order to be able to use Lemma 2.4, is given by

$$(M_U^{-1}K)^{-1} = \begin{bmatrix} I - F^{-1}B^T S^{-1}B & F^{-1}B^T S^{-1}H_2 \\ S^{-1}B & I - S^{-1}H_2 \end{bmatrix}.$$

We now need to establish conditions (2a)–(2d) of Lemma 2.4.

Lemma 3.8 (Proof of condition (2a) for M_U). The *H*-norm of the inverse of the preconditioned matrix associated with the preconditioner M_U satisfies

$$\|(M_U^{-1}K)^{-1}\|_H \lesssim 1.$$

Proof. The proof is obtained by putting together the bounds of Lemmas 3.3, 3.4, and 3.6.

$$\begin{split} \|(M_U^{-1}K)^{-1}\|_H &= \left\| \begin{bmatrix} H_1^{1/2}(I - F^{-1}B^T S^{-1}B)H_1^{-1/2} & H_1^{1/2}(F^{-1}B^T S^{-1}H_2)H_2^{-1/2} \\ H_2^{1/2}(S^{-1}B)H_1^{-1/2} & H_2^{1/2}(I - S^{-1}H_2)H_2^{-1/2} \end{bmatrix} \right\|_2 \\ &\leq \|I - F^{-1}B^T S^{-1}B\|_{H_1} + \|F^{-1}B^T S^{-1}H_2\|_{H_2,H_1} + \|S^{-1}B\|_{H_1,H_2} + \|I - S^{-1}H_2\|_{H_2} \\ &\leq (1 + 2C_1^2(1 + \alpha)^2/C_2^2) + ((1 + \alpha)^2/C_2^2C_1) + 1 + (1 + \alpha)^2/C_2^2 \\ &\lesssim 1. \end{split}$$

Lemma 3.9 (Proof of condition (2b) for M_U). The *H*-norm of the preconditioned matrix associated with the preconditioner M_U satisfies

$$\|M_U^{-1}K\|_H \lesssim 1.$$

Proof. Similarly to the proof of Lemma 3.8,

$$\|M_U^{-1}K\|_H = \left\| \begin{bmatrix} H_1^{1/2}(I - F^{-1}B^T H_2^{-1}B)H_1^{-1/2} & H_1^{1/2}F^{-1}B^T H_2^{-1/2} \\ H_2^{1/2} H_2^{-1}B H_1^{-1/2} & 0 \end{bmatrix} \right\|_2$$

$$\leq \|I - F^{-1}B^T H_2^{-1}B\|_{H_1} + \|F^{-1}B^T\|_{H_2,H_1} + \|B\|_{H_1,H_2^{-1}}$$

$$\leq (1 + (1 + \alpha)^2/C_2^2 C_1^2) + 2C_1$$

$$\lesssim 1.$$

Lemma 3.10 (Proof of condition (2c) for M_U). When $\alpha < \frac{1}{2}$, we have

$$\left\| H(M_U^{-1}K) - (M_U^{-1}K)^T H \right\|_{H,H^{-1}} \lesssim \alpha.$$
 (13)

Proof. We have

$$\|H(M_U^{-1}K) - (M_U^{-1}K)^T H\|_{H,H^{-1}} = \left\| \begin{bmatrix} B_{11} & B_{12} \\ -B_{12}^T & 0 \end{bmatrix} \right\|_2$$

$$\le \|B_{11}\|_2 + 2\|B_{12}\|_2,$$
(14)

where

$$B_{11} = -H_1^{1/2} F^{-1} B^T H_2^{-1} B H_1^{-1/2} + H_1^{-1/2} B^T H_2^{-1} B F^{-T} H_1^{1/2}.$$

$$B_{12} = H_1^{1/2} F^{-1} B^T H_2^{-1/2} - H_1^{-1/2} B^T H_2^{-1/2}$$

$$\|B_{12}\|_2 = \|(H_1 F^{-1} - I) B^T\|_{H_2, H_1^{-1}}$$

$$\leq C_1 \|H_1 F^{-1} - I\|_{H_1^{-1}}$$

$$= C_1 \|H_1^{1/2} (H_1 + N)^{-1} H_1^{1/2} - I\|_2$$

$$= C_1 \|(I + H_1^{-1/2} N H_1^{-1/2})^{-1} - I\|_2.$$

When $\alpha < \frac{1}{2}$ we have

$$||B_{12}||_2 \le C_1 \frac{||N||_{H_1, H_1^{-1}}}{1 - ||N||_{H_1, H_1^{-1}}} \le C_1 \alpha / (1 - \alpha) \le 2C_1 \alpha \lesssim \alpha$$
(15)

and

$$||B_{11}|| = ||H_1F^{-1}B^TH_2^{-1}B - B^TH_2^{-1}BF^{-T}H_1||_{H_1,H_1^{-1}} = ||(F - N)F^{-1}B^TH_2^{-1}B - B^TH_2^{-1}BF^{-T}(F^T - N^T)||_{H_1,H_1^{-1}} = || - NF^{-1}B^TH_2^{-1}B + B^TH_2^{-1}BF^{-T}N^T||_{H_1,H_1^{-1}} \leq ||NF^{-1}B^TH_2^{-1}B||_{H_1,H_1^{-1}} + ||B^TH_2^{-1}BF^{-T}N^T||_{H_1,H_1^{-1}} \leq 2C_1^2\alpha \lesssim \alpha.$$

Substituting the above inequalities into (14), we obtain (13), as required.

The results of Lemmas 3.8–3.10 along with the assumption that α is sufficiently small establish the scalability of the iterations.

Theorem 3.11. Given a saddle-point system with matrix K defined in (5), where $F \in \mathbb{R}^{n \times n}$ is positive real and $B \in \mathbb{R}^{m \times n}$ is full row rank, let H_1 and N be the symmetric and skewsymmetric parts, respectively, of F, as in (8). Let H_2 be a symmetric positive definite matrix, such that the three conditions in (9)–(10) are satisfied. Finally, let H be the block-diagonal matrix defined in (7). Then, for α sufficiently small, GMRES with the left preconditioner M_U under the H-norm will converge in a fixed number of iterations independently of the dimensions of K.

Proof. Lemmas 3.8–3.10 establish conditions (2a)–(2c). Trivially, by Lemmas 3.9 and 3.10, (2d) holds when α is sufficiently small.

We now consider the right preconditioner M_L defined in (12). The analysis is very similar to the left preconditioner case. The details are omitted and we present a theorem analogous to Theorem 3.11. **Theorem 3.12.** Given a saddle-point system with matrix K defined in (5), where $F \in \mathbb{R}^{n \times n}$ is positive real and $B \in \mathbb{R}^{m \times n}$ is full row rank, let H_1 and N be the symmetric and skewsymmetric parts, respectively, of F, as in (8). Let H_2 be a symmetric positive definite matrix, such that the three conditions in (9)–(10) are satisfied. Finally, let H be the block-diagonal matrix defined in (7). Then, for α sufficiently small, GMRES with the right preconditioner M_L under the H^{-1} -norm will converge in a fixed number of iterations independently of the dimensions of K.

Remark 3.1. In practice, H can be replaced with another symmetric positive definite matrix \tilde{H} and results will still hold if H and \tilde{H} are spectrally equivalent: GMRES convergence with H-norm can induce GMRES convergence with \tilde{H} -norm. This is because

$$\|p(A)\|_{H} = \|H^{1/2}(\tilde{H}^{-1/2}\tilde{H}^{1/2})p(A)(\tilde{H}^{-1/2}\tilde{H}^{1/2})H^{-1/2}\|_{2} \le \kappa_{2}(H^{1/2}\tilde{H}^{-1/2})\|p(A)\|_{\tilde{H}}$$

3.2 A Block-Diagonal Preconditioner

The case of a block diagonal preconditioner of the form

$$M_D = \begin{bmatrix} F & 0\\ 0 & H_2 \end{bmatrix} \tag{16}$$

is interesting in the context of this work, because contrary to block-triangular preconditioners, where one might select either an upper block-triangular preconditioner or a lower block-triangular preconditioner along with left or right preconditioning to avoid a situation of having zero in the field of values, here it is immediate that the field of values contains zero regardless of any such choices made. There is no practical difference between left and right preconditioning here, and we proceed with left preconditioning below. The preconditioned matrix is

$$M_D^{-1}K = \begin{bmatrix} I & F^{-1}B^T \\ H_2^{-1}B & 0 \end{bmatrix}$$

and its inverse is

$$(M_D^{-1}K)^{-1} = \begin{bmatrix} I - F^{-1}B^T S^{-1}B & F^{-1}B^T S^{-1}H_2 \\ S^{-1}B & -S^{-1}H_2 \end{bmatrix}$$

The analysis is essentially identical to the analysis in Section 3.1.

Lemma 3.13 (Proof of condition (2a) for M_D). The *H*-norm of the inverse of the preconditioned matrix associated with the preconditioner M_D satisfies

$$||(M_D^{-1}K)^{-1}||_H \lesssim 1.$$

Proof. The proof follows similar steps as for M_U in Lemma 3.8. We need to bound the norm of each block in the inverse, and we apply the bounds obtained in Lemmas 3.3, 3.4, and 3.6:

$$\begin{split} \|I - F^{-1}B^T S^{-1}B\|_{H_1} &\leq 1 + C_1^2 (1+\alpha)^2 / C_2^2 \\ \|F^{-1}B^T S^{-1}H_2\|_{H_2,H_1} &\leq (1+\alpha)^2 / C_2^2 C_1, \\ \|S^{-1}B\|_{H_1,H_2} &\leq C_1, \\ \|S^{-1}H_2\|_{H_2} &\leq (1+\alpha)^2 / C_2^2. \end{split}$$

Combining these, we get the bound for the entire matrix.

Lemma 3.14 (Proof of condition (2b) for M_D). The *H*-norm of the preconditioned matrix associated with the preconditioner M_D satisfies

$$\|M_D^{-1}K\|_H \lesssim 1.$$

Proof. Similar to the analysis for M_U in Lemma 3.9, we bound the norm of each block in the preconditioned matrix:

$$\|I\|_{H_1} = 1,$$

$$\|F^{-1}B^T\|_{H_2,H_1} \le C_1,$$

$$\|H_2^{-1}B\|_{H_1,H_2} \le C_1.$$

Thus, the norm of the entire matrix is bounded by the sum of these norms.

Lemma 3.15 (Proof of condition (2c) for M_D). When $\alpha < \frac{1}{2}$, we have

$$\left\| H(M_2^{-1}K) - (M_2^{-1}K)^T H \right\|_{H,H^{-1}} \lesssim \alpha.$$

Proof. Note that

$$\left\| H(M_D^{-1}K) - (M_D^{-1}K)^T H \right\|_{H,H^{-1}} = \left\| \begin{bmatrix} 0 & B_{12} \\ -B_{12}^T & 0 \end{bmatrix} \right\|_2$$

 $\leq 2 \|B_{12}\|_2,$

where

$$||B_{12}||_2 = ||(H_1F^{-1} - I)B^T||_{H_2, H_1^{-1}}$$

By (15), we complete the proof.

Theorem 3.16. Given a saddle-point system with matrix K defined in (5), where $F \in \mathbb{R}^{n \times n}$ is positive real and $B \in \mathbb{R}^{m \times n}$ is full row rank, let H_1 and N be the symmetric and skewsymmetric parts, respectively, of F, as in (8). Let H_2 be a symmetric positive definite matrix, such that the three conditions in (9)–(10) are satisfied. Finally, let H to be the block-diagonal matrix defined in (7). Then, for α sufficiently small, GMRES with the blockdiagonal preconditioner M_D under the H-norm for left preconditioning or H^{-1} -norm for right preconditioning will converge in a fixed number of iterations independently of the dimensions of K.

3.3 Inexact Preconditioning

To make the iterations practical, one needs to consider computationally inexpensive ways of approximately inverting the preconditioners that we have discussed so far, and using those approximate linear operators as the actual preconditioners. Under mild conditions,

our analysis seems to carry over to such situations. We illustrate this for a block uppertriangular preconditioner that approximates the leading block. Consider

$$\tilde{M}_U = \begin{bmatrix} P_1 & B^T \\ 0 & H_2 \end{bmatrix},$$

where the action of (implicitly) inverting P_1 is computationally practical. Note that

$$\tilde{M}_U^{-1}K = (\tilde{M}_U^{-1}M_U)M_U^{-1}K$$

and

$$\tilde{M}_U^{-1}M_U = \begin{bmatrix} P_1^{-1}F & 0\\ 0 & I \end{bmatrix}$$

Assumption 3.17. We assume $||P_1^{-1}F - I||_{H_1} \le C_3 \alpha$ and $||F^{-1}P_1||_{H_1} \le C_4$.

Based on Assumption 3.17, we have

$$\|\tilde{M}_U^{-1}M_U\|_H \le (1+C_3\alpha) + 1 \lesssim 1$$

and

$$\|(\tilde{M}_U^{-1}M_U)^{-1}\|_H \le \|F^{-1}P_1\|_{H_1} + 1 \lesssim 1.$$

We now examine the sufficient conditions. For condition (2a), we have

$$\|\tilde{M}_U^{-1}K\|_H \le \|\tilde{M}_U^{-1}M_U\|_H \|M_U^{-1}K\|_H \le (1+C_3\alpha)\|M_U^{-1}K\|_H \le 1.$$

For condition (2b), we have

$$\|(\tilde{M}_U^{-1}K)^{-1}\|_H \le \|(\tilde{M}_U^{-1}M_U)^{-1}\|_H \|(M_U^{-1}K)^{-1}\|_H \lesssim 1.$$

For condition (2c), we have

$$\begin{split} \|H(\tilde{M}_{U}^{-1}K) - (\tilde{M}_{U}^{-1}K)^{T}H\|_{H,H^{-1}} &\leq \|H(M_{U}^{-1}K) - (M_{U}^{-1}K)^{T}H\|_{H,H^{-1}} \\ &+ \|H(\tilde{M}_{U}^{-1}M_{U} - I)M_{U}^{-1}K - (M_{U}^{-1}K)^{T}(\tilde{M}_{U}^{-1}M_{U} - I)^{T}H\|_{H,H^{-1}} \\ &\lesssim \alpha + 2\|P_{1}^{-1}F - I\|_{H_{1}}\|M_{U}^{-1}K\|_{H} \\ &\lesssim \alpha. \end{split}$$

Thus, if α is small enough, condition (2d) is satisfied and the iterative solver with \tilde{M}_U as a preconditioner will converge in a fixed number of iterations.

4 Numerical Experiments

We provide a couple of examples of applications from fluid dynamics to validate our analysis.

4.1 Navier-Stokes Equations

Let $\Omega \subset \mathbb{R}^2$ be a bounded domain. The Navier-Stokes equations with pure Dirichlet boundary conditions are [12]

$$-\nu\Delta \mathbf{u} + (\mathbf{u} \cdot \nabla)\mathbf{u} + \nabla p = f \quad \text{in } \Omega,$$
$$\nabla \cdot \mathbf{u} = 0 \quad \text{in } \Omega,$$
$$\mathbf{u} = \mathbf{u}_d \quad \text{on } \partial\Omega,$$

where ν is a viscosity coefficient, **u** is the velocity, *p* is the pressure, and **u** = **u**_d provides the Dirichlet boundary conditions.

Linearizing the equations using the Picard iteration, we obtain

$$-\nu\Delta \mathbf{u} + (\mathbf{b}\cdot\nabla)\mathbf{u} + \nabla p = f \quad \text{in } \Omega,$$
$$\nabla\cdot\mathbf{u} = 0 \quad \text{in } \Omega,$$
$$\mathbf{u} = \mathbf{u}_d \quad \text{on } \partial\Omega,$$

where **b** is the velocity from the previous iteration.

For simplicity, we assume $\mathbf{u}_d = \mathbf{0}$. Define the Sobolev spaces

$$\mathbf{V} = \{ \mathbf{v} \in (H^1(\Omega))^2 : \mathbf{v} = 0 \text{ on } \partial\Omega \}, \quad Q = \{ q \in L^2(\Omega) : \int_{\Omega} q = 0 \}.$$

The weak form involves solving the following system: find $\mathbf{u} \in \mathbf{V}$ and $p \in Q$ such that

$$a(\mathbf{u}, \mathbf{v}) + b(\mathbf{v}, p) = \mathbf{f}(\mathbf{v}) \quad \forall \mathbf{v} \in \mathbf{V},$$
$$b(\mathbf{u}, q) = 0 \quad \forall q \in Q,$$

where the bilinear forms are defined as

$$\begin{split} a(\mathbf{u}, \mathbf{v}) &= \nu \int_{\Omega} \nabla \mathbf{u} \cdot \nabla \mathbf{v} + \int_{\Omega} (\mathbf{b} \cdot \nabla \mathbf{u}) \cdot \mathbf{v}, \\ b(\mathbf{u}, q) &= -\int_{\Omega} (\nabla \cdot \mathbf{u}) q, \end{split}$$

and **f** is the linear functional $\int_{\Omega} \mathbf{f} \cdot \mathbf{v}$.

By using conforming finite element spaces $\mathbf{V}_h \subset \mathbf{V}$ and $Q_h \subset Q$, we discretize these equations and obtain the nonsymmetric saddle-point system

$$\begin{bmatrix} F & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ p \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ 0 \end{bmatrix},\tag{17}$$

where $F = \nu H_1 + N$, $a(\mathbf{u}_h, \mathbf{v}_h) = (F\mathbf{u}_1, \mathbf{v}_1) = \nu(H_1\mathbf{u}_1, \mathbf{v}_1) + (N\mathbf{u}_1, \mathbf{v}_1)$, $(\nabla \mathbf{u}_h, \nabla \mathbf{v}_h) = (H_1\mathbf{u}_1, \mathbf{v}_1)$, and $((\mathbf{b} \cdot \nabla \mathbf{u}_h), \mathbf{v}_h) = (N\mathbf{u}_1, \mathbf{v}_1)$. So far, this is a standard treatment of these equations; see [12].

To make our analysis applicable, we scale the system on the left by $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ and on the right by $\begin{bmatrix} 1 & 0 \\ 0 & \nu \end{bmatrix}$, respectively, and the system becomes

$$\begin{bmatrix} H_1 + \frac{1}{\nu}N & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \frac{1}{\nu}p \end{bmatrix} = \begin{bmatrix} \frac{1}{\nu}\mathbf{f} \\ 0 \end{bmatrix}.$$

Since the problem is given with pure Dirichlet boundary conditions, we have $N^T = -N$, which indicates H_1 is the symmetric part of $H_1 + \frac{1}{\nu}N$. If ν is sufficiently large, then conditions (9) and (10) are satisfied, as we have used conforming elements.

We numerically solve the regularized lid-driven cavity problem using IFISS [13] to illustrate our results. The domain Ω is chosen as $[-1, 1]^2$. Zero boundary conditions are imposed, except we take $u_x = 1 - x^4$ on $\{y = 1, -1 \le x \le 1\}$.

We set $\nu = 1$, because our analysis requires it to be relatively large, and apply the Picard iteration. Since ν is relatively large, the nonlinear iterations converge quickly; we record average iteration counts and examine the performance of the linear solvers. We note that we have observed no significant differences among the linear solver iteration counts throughout the nonlinear iteration. We use the diagonal preconditioner M_D defined in (16) and the upper triangular preconditioner M_U defined in (11). We use left preconditioning for both; the results for right preconditioning with M_L defined in (12) are virtually the same.

Results for a few mesh sizes can be found in Table 1. We observe an excellent level of scalability: the iteration counts are nearly constant for various mesh sizes in all cases. We present our iteration counts in both the ℓ_2 and H norms, and observe that they are nearly identical.

System Size	$\ell_2 \text{ norm}$		<i>H</i> -norm	
	Diagonal	Upper Triangular	Diagonal	Upper Triangular
210	21.0	11.0	21.0	11.3
770	22.5	12.0	23.0	12.0
2,946	23.0	12.5	23.0	12.5
11,522	24.0	13.0	24.0	13.0

Table 1: Average iteration counts for Navier-Stokes

For the diagonal preconditioner, we have computed the parameters of Lamma 2.4 and have observed that $b \approx 2.25$ and $c \approx 0.016$. For the upper-triangular preconditioner, $b \approx 2.06$ and $c \approx 0.035$. In both cases we have bc < 1, as required.

4.2 Stokes-Darcy Equations

Consider the Stokes-Darcy equations on a non-overlapping domain $\Omega = \Omega_s \cup \Omega_d$ with a polygonal interface $\Gamma_I = \partial \Omega_s \cap \partial \Omega_d$:

$$\begin{aligned} -\nabla \cdot (2\nu D(\mathbf{u}) - p\mathbf{I}) &= \mathbf{f}^s & \text{in } \partial\Omega_s, \\ \nabla \cdot \mathbf{u} &= 0 & \text{in } \partial\Omega_s, \\ \mathbf{u} &= \mathbf{g}^s & \text{on } \Gamma_s &= \partial\Omega_s \cap \partial\Omega, \\ -k\Delta\phi &= f^d & \text{in } \Omega_d, \\ \phi &= g^d & \text{on } \Gamma_d, \\ k\nabla\phi \cdot \mathbf{n} &= g^n & \text{on } \Gamma_n, \\ \mathbf{u} \cdot \mathbf{n}_{12} &= -k\nabla\phi \cdot \mathbf{n}_{12} & \text{on } \Gamma_I, \\ (-2\nu D(\mathbf{u}) \cdot \mathbf{n}_{12} + p\mathbf{n}_{12}) \cdot \mathbf{n}_{12} &= \phi & \text{on } \Gamma_I, \\ \mathbf{u} \cdot \boldsymbol{\tau}_{12} &= -2\nu G(D(\mathbf{u})\mathbf{n}_{12}) \cdot \boldsymbol{\tau}_{12} & \text{on } \Gamma_I, \end{aligned}$$

where \mathbf{u} satisfies the incompressibility condition $\nabla \cdot \mathbf{u} = 0$. Here, Ω_S and Ω_d are assumed to be simple domains, e.g., the unit squares in two dimensions, with a polygonal interface. The operator D is defined as $D(\mathbf{u}) = \frac{1}{2}(\nabla \mathbf{u} + \nabla \mathbf{u}^T)$. The physical parameters ν and k denote the viscosity coefficient and hydraulic constant, respectively. The constant G represents an experimentally-determined constant related to the Beavers-Joseph-Saffman interface condition. Finally, \mathbf{n}_{12} and τ_{12} are unit normal and tangential vectors; see [4] for details. We use the finite element discretization described in [4, 9, 5]; some details on the Stokes part are similar to Section 4.1. Full details on the discretization of the entire problem are omitted since this is not the focus of our paper. The discretization yields the following linear system:

$$\mathcal{K}\begin{bmatrix}\mathbf{u}_1\\\phi_1\\p_1\end{bmatrix} = \begin{bmatrix}\nu\mathbf{A}_{\Omega_s} & I_{12}^T & B^T\\-I_{12} & kA_{\Omega_d} & 0\\B & 0 & 0\end{bmatrix}\begin{bmatrix}\mathbf{u}_1\\\phi_1\\p_1\end{bmatrix} = \begin{bmatrix}\mathbf{f}\\f\\0\end{bmatrix},$$
(18)

where \mathbf{u}_1 , ϕ_1 and p_1 represent the vectors of coefficients in the finite element basis expansions, with corresponding continuous finite element solutions denoted by \mathbf{u}_h , p_h and ϕ_h , respectively.

For simplicity of our analysis, we assume $k = \nu$, and consider the following scaled matrix:

$$\begin{bmatrix} \mathbf{A}_{\Omega_s} & \frac{1}{\nu} I_{12}^T & B^T \\ -\frac{1}{\nu} I_{12} & A_{\Omega_d} & 0 \\ B & 0 & 0 \end{bmatrix}.$$

Then, assuming that ν is sufficiently large (which corresponds to requiring α to be sufficiently small in our analysis in Section 3; see (9)), let us define

$$F = \begin{bmatrix} \mathbf{A}_{\Omega_s} & 0\\ 0 & A_{\Omega_d} \end{bmatrix} + \frac{1}{\nu} \begin{bmatrix} 0 & I_{12}^T\\ -I_{12} & 0 \end{bmatrix}.$$

It has been shown in the literature [4, 9] that the inf-sup condition for the matrix $\begin{bmatrix} B & 0 \end{bmatrix}$ is satisfied and that the skew-symmetric operator $\begin{bmatrix} 0 & I_{12}^T \\ -I_{12} & 0 \end{bmatrix}$ is bounded. Therefore, the conditions of Lemma 2.4 are satisfied, and it follows that an iterative solver preconditioned with the block preconditioners discussed in Section 3 will converge independently of the mesh size.

We use the following example from [4]. We choose Ω_s to be $[0,1]^2$ and Ω_d to be $[0,1] \times [1,2]$. Γ_n is $\{x = 0, y \in [0,1]\} \cup \{x = 1, y \in [0,1]\}$. Boundary conditions and right-hand side are computed from the following exact solution:

$$\mathbf{u}(x,y) = [y^2 - 2y + 1 + \nu(2x - 1), x^2 - x - 2\nu(y - 1)]^T;$$

$$p(x,y) = 2\nu(x + y - 1) + \frac{1}{3k} - 4\nu^2;$$

$$\psi(x,y) = \frac{1}{k}(x(1 - x)(y - 1) + \frac{y^3}{3} - y^2 + y) + 2\nu x.$$

We also set $k = \nu = 3$ and G = 1, in order for the parameters to satisfy the conditions of Lemma 2.4.

As we have done for the Navier-Stokes problem in Section 4.1 – here, too, we provide a brief validation of our analysis. We again apply left preconditioning, using the diagonal and the upper-triangular preconditioners, M_D and M_U respectively, defined in (16) and (11).

Our observations are similar to those we made in Section 4.1. The results for a few mesh sizes can be found in Table 2. We again observe an excellent level of scalability, with iteration counts nearly constant for various mesh sizes in all cases. The iteration counts in the ℓ_2 and H norms are nearly identical.

System Size	ℓ_2 -norm		<i>H</i> -norm	
	Diagonal	Upper Triangular	Diagonal	Upper Triangular
633	27	16	29	16
2,545	28	16	30	16
10,209	28	16	30	16
40,897	30	16	28	16

Table 2: Iteration counts for the Stokes-Darcy equations.

For the diagonal preconditioner, we have observed experimentally for the smaller-size problems that the parameters of Lemma 2.4 satisfy $b \approx 9.18$, $c \approx 0.08$, and bc < 1. For the upper triangular preconditioner, $b \approx 8.28$, $c \approx 0.11$ and bc < 1.

5 Concluding Remarks

Our analysis broadens the range of preconditioned saddle-point systems for which FOV analysis may be applied by including cases where zero is included in the field of values. This includes the important family of block-diagonal preconditioners, as well as upper-triangular preconditioners applied with right preconditioning. For these cases, to our knowledge, no FOV analysis was previously available when (4) is true.

When applying Theorem 1.5, a disk must be excluded from the field of values, and the remaining part should not surround the origin, as we have illustrated in Figure 1. To accomplish this, we require the imaginary part of the FOV to be small enough, which means that for the nonsymmetric saddle-point systems we consider, the preconditioned matrix is close to normal, or equivalently, the skew-symmetric part of preconditioned operator needs to be small in norm.

A finer geometric study of the field of values, beyond bounding it just by using the imaginary axis, may allow for loosening the aforementioned restriction. It would be useful to enhance the set of nonsymmetric saddle-point linear systems for which FOV analysis of the type we have offered is applicable.

References

[1] Mario Arioli, Eric Noulard, and Alessandro Russo. Stopping criteria for iterative methods: śapplications to PDE's. Calcolo, 38(2):97–112, 2001.

- [2] Fatemeh Panjeh Ali Beik and Michele Benzi. Preconditioning techniques for the coupled Stokes–Darcy problem: spectral and field-of-values analysis. <u>Numerische Mathematik</u>, 150(2):257–298, 2022.
- [3] Michele Benzi. Some uses of the field of values in numerical analysis. <u>Bollettino</u> dell'Unione Matematica Italiana, 14(1):159–177, 2021.
- [4] Prince Chidyagwai, Scott Ladenheim, and Daniel B Szyld. Constraint preconditioning for the coupled Stokes–Darcy system. <u>SIAM Journal on Scientific Computing</u>, 38(2):A668–A690, 2016.
- [5] Prince Chidyagwai and Beatrice Rivière. Numerical modelling of coupled surface and subsurface flow systems. Advances in Water Resources, 33(1):92–105, 2010.
- [6] Daeshik Choi and Anne Greenbaum. Roots of matrices in the study of GMRES convergence and Crouzeix's conjecture. <u>SIAM Journal on Matrix Analysis and Applications</u>, 36(1):289–301, 2015.
- [7] Michel Crouzeix and Anne Greenbaum. Spectral sets: numerical range and beyond. SIAM Journal on Matrix Analysis and Applications, 40(3):1087–1101, 2019.
- [8] Michel Crouzeix and César Palencia. The numerical range is a $(1+\sqrt{2})$ -spectral set. SIAM Journal on Matrix Analysis and Applications, 38(2):649–655, 2017.
- [9] Marco Discacciati and Alfio Quarteroni. Navier-Stokes/Darcy coupling: modeling, analysis, and numerical approximation. Rev. Mat. Complut, 22(2):315–426, 2009.
- [10] Tobin A Driscoll, Kim-Chuan Toh, and Lloyd N Trefethen. From potential theory to matrix iterations in six steps. SIAM review, 40(3):547–578, 1998.
- [11] Michael Eiermann. Fields of values and iterative methods. Linear algebra and its applications, 180:167–197, 1993.
- [12] Howard Elman, David Silvester, and Andy Wathen. <u>Finite Elements and Fast Iterative</u> <u>Solvers: with Applications in Incompressible Fluid Dynamics</u>. Oxford University Press, 06 2014.
- [13] Howard C. Elman, Alison Ramage, and David J. Silvester. IFISS: A computational laboratory for investigating incompressible flow problems. <u>SIAM Review</u>, 56(2):261– 273, 2014.
- [14] Mark Embree. How descriptive are GMRES convergence bounds? <u>arXiv preprint</u> arXiv:2209.01231, 2022.
- [15] P. C. Hansen. <u>Analysis of Toeplitz Systems</u>, volume 1481 of <u>Lecture Notes in</u> Mathematics. Springer-Verlag, 1991.
- [16] Axel Klawonn and Gerhard Starke. Block triangular preconditioners for nonsymmetric saddle point problems: field-of-values analysis. <u>Numerische Mathematik</u>, 81:577–594, 1999.

- [17] Daniel Loghin and Andrew J Wathen. Analysis of preconditioners for saddle-point problems. SIAM Journal on Scientific Computing, 25(6):2029–2049, 2004.
- [18] Sébastien Loisel and Peter Maxwell. Path-following method to determine the field of values of a matrix with high accuracy. <u>SIAM Journal on Matrix Analysis and Applications</u>, 39(4):1726–1749, 2018.
- [19] Yicong Ma, Kaibo Hu, Xiaozhe Hu, and Jinchao Xu. Robust preconditioners for incompressible MHD models. Journal of Computational Physics, 316:721–746, 2016.
- [20] Youcef Saad and Martin H Schultz. GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. <u>SIAM Journal on scientific and statistical</u> computing, 7(3):856–869, 1986.