PRECONDITIONERS FOR GENERALIZED SADDLE-POINT PROBLEMS

BY

CHRISTOPHER MARTIN SIEFERT

B.S., College of William and Mary, 2000

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2006

Urbana, Illinois

Abstract

Generalized saddle point problems arise in a number of applications, ranging from optimization and metal deformation to fluid flow and PDE-governed optimal control. We focus our discussion on the most general case, making no assumption of symmetry or definiteness in the matrix or its blocks. As these problems are often large and sparse, preconditioners play a critical role in speeding the convergence of Krylov methods for these problems. We first examine two types of preconditioners for these problems, one block-diagonal and one indefinite, and present analyses of the eigenvalue distributions of the preconditioned matrices. We also investigate the use of approximations for the Schur complement matrix in these preconditioners and develop eigenvalue analysis accordingly.

Second, we examine new developments in probing methods, inspired by graph coloring methods for sparse Jacobians, for building approximations to Schur complement matrices. We then present an analysis of these techniques and their accuracy. In addition, we provide a mathematical justification for their use in approximating Schur complements and suggest the use of approximate factorization techniques to decrease the computational cost of applying the inverse of the probed matrix.

Finally, we consider the effect of our preconditioners on four applications. Two of these applications come from the realm of fluid flow, one using a finite element discretization and the other using a spectral discretization. The third application involves the stress relaxation of aluminum strips at low stress levels. The final application involves mesh parameterization and flattening.

For these applications, we present results illustrating the eigenvalue bounds on our preconditioners and demonstrating the theoretical justification of these methods. We also present convergence and timing results, showing the effectiveness of our methods in practice. Specifically the use of probing methods for approximating the Schur compliment matrices in our preconditioners is empirically justified. We also investigate the h-dependence of our preconditioners one model fluid problem, and demonstrate empirically that our methods do not suffer from a deterioration in convergence as the problem size increases.

Acknowledgments

There are many people deserving of thanks for their assistance in making this thesis a reality, but none more so than my advisor, Eric de Sturler. His sage advice, thoughtful insights and all-around good company played no small role in making this work happen. I am deeply grateful to him for everything he has done!

I would also like to thank Michael Heath for his role as the Numerical Analysis group "patriarch." As the unofficial advisor to all of the graduate students, Mike's wisdom (and critique, when warranted) has helped me to hone my presentation skills and helped me to succinctly explain my work to my professional colleagues.

The other members of my committee, Robert Skeel and Armand Beaudoin, are also deserving of special thanks, Bob for coming all the way back from Purdue to serve on my committee and Armand for providing our metal deformation application with the assistance of his former student Lihua Zhu. I am also grateful to the NSF for their funding of my first three years of graduate school through their graduate fellowship, the Center for Simulation of Advanced Rockets and their funding of my research for two more years, and Eric de Sturler and David Ceperley allowing me to finish up my thesis and begin looking at $O(n)$ methods for particle simulation.

When asked once what I thought about graduate school, I replied that I really enjoyed the collaborative environment at UIUC. The graduate students in our NA group have provided not only wonderful fellowship, but have helped me to hone and refine many of the ideas I have had in my time at graduate school. To David Alber, Naomi Caldwell, Bill Cochran, Zhen

Cheng, Eric Cyr, Rebecca Hartman-Baker, Vanessa Lopez, Greg Mackey, Mike Parks, Ali Pinar, Ryan Syzpowski, Hanna VanderZee, and Shun Wang, I offer special thanks.

I am also deeply indebted to the priests, sisters and staff of my "home away from home" at St. John's Catholic Newman Center. Their example, prayers and encouragement have been an inspiration to me in my own work. I am especially grateful to Fr. Fortunatus Bijura, Sr. Anna Flannigan, Fr. Brian Herlocker, Sr. Janice Keenan, Dr. Kenneth Howell, Shawn Reeves and Msgr. Stuart Swetland.

My friends have also played a special role in encouraging me throughout my time here. Nathan Buxhoeveden, Joe Carnahan, Rami Dass, Dwight Divine, Meghan Faulkner, Brandon Field, Aaron Foege, Kathryn Guy, Chris Hamelly, Nathan Hearn, Tony Kaap, Phil Lee, Rosie Ray, Nick Romero and Cathy Roodhouse all deserve a special mention. Nick also gave me the great idea of choosing patron saints for my thesis. I am greatful to Sts. Francis de Sales and Thomas Aquinas for their prayers on my behalf.

Penultimately, I need to thank my parents, Daniel and Frances Siefert for their love, dedication, support, encouragement and downright silly gifts throughout the years. They have been wonderful to me and I am forever indebted to them. Where I would be without them, I do not know, and do not care to speculate.

Finally, I offer thanks to God for giving me all of these wonderful blessings! For without Him, none of this would have come into being. *Vivat Jesus!*

Christopher Martin Siefert

Urbana, IL.

Feast of St. Elizabeth Ann Seaton, January 4th, 2006.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AINV	Approximate Inverse Preconditioner
AMG	Algebraic Multigrid
GMRES	Generalized Minimum Residual Method
ILU	Incomplete LU Factorization
ILU(0)	Incomplete LU Factorization with Zero Fill
ILUT	Incomplete LU Factorization with a Drop Tolerance
KKT	Karush-Kuhn-Tucker
LFO	Largest-First Ordering
PDE	Partial Differential Equation
SP	Structured Probing
SP-Exact	Structured Probing with an Exact Factorization
SP-ILU(0)	Structured Probing with an ILU(0) Factorization
SPAI	Sparse Approximate Inverse Preconditioner
SVD	Singular Value Decomposition

List of Symbols

$\\|\cdot\\|$	2-norm
$\\|\cdot\\|_1$	1-norm
$\\|\cdot\\|_\infty$	∞-norm
$\Delta(G)$	Maximum vertex degree in graph G
$\delta(G)$	Minimum vertex degree in graph G
$\bar{\delta}_2(G)$	Average vertex distance-2 degree in graph G
$\Lambda(A)$	Set of eigenvalues of matrix A.
$\rho(A)$	Spectral radius of matrix A.
$\sigma(A)$	Set of singular values of matrix A.

1 Introduction

In the context of constrained optimization, the Karush-Kuhn-Tucker (KKT) conditions [46, p. 185] state that a constrained minimizer must be a *saddle-point* of an augmented problem. Specifically, it must be a minimizer in the primal variables and a maximizer on the space of Lagrange multipliers. Linearizing these conditions with a Newton iteration yields linear system of the form

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

where A is symmetric positive definite. These linear systems are referred to as *saddle-point problems.* But as Benzi, Golub and Liesen [5] point out, many other problems have similar structural and eigenvalue characteristics to saddle-point problems, but do not fit the form (1.1). These problems are referred to as *generalized saddle-point problems*, and are are of the form

$$\mathcal{A} \begin{bmatrix} x \\ y \end{bmatrix} \equiv \begin{bmatrix} A & B^T \\ C & D \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix}, \tag{1.2}$$

where $A \in \mathbb{R}^{n\times n}$, $D \in \mathbb{R}^{m\times m}$, and $n > m$. Benzi, Golub and Liesen [5] describe generalized saddle-point problems as satisfying one or more of the following criteria:

C1. A is symmetric.

C2. The symmetric part of A, $H = \frac{1}{2}(A + A^T)$ is positive semidefinite.

C3. $C = B$.

C4. D is symmetric and negative semidefinite.

C5. $D = 0$.

This description is adequate enough to describe many problems of interest. While saddle-point problems satisfy all of these criteria, generalized saddle-point problems only satisfy some. For instance, a stabilized finite element discretization of the Navier-Stokes equations, might only satisfy C2, C3 and C4. A spectral discretization of the same problem may only satisfy C4 or C5.

However, some problems which have similar structural and eigenvalue properties to saddle-point problems may not satisfy any of the above criteria. In that light, we wish to introduce a sixth criterion, namely:

C6. $\|D\|$ is sufficiently small that the system retains the qualities of a saddle-point problem.

All the systems we consider meet criterion C6 and many will also satisfy some of C1–C5 as well. Systems that satisfy some or all of these criteria may arise from:

- Constrained optimization [24, 32],
- Domain decomposition [27, 28],
- Finite element discretizations for fluid dynamics [25, 55, 56, 60],
- Finite element discretizations for magnetostatics [47],
- Mesh parameterization in computer graphics [40, 52],
- Metal deformation [63],
- Mixed formulations of elliptic PDEs[1, 2, 12],
- Optimal control [3, 4],
- Spectral discretizations for fluid dynamics [8].

Our goal is to develop effective and efficient preconditioners for this broader class of generalized saddle-point problems. Specifically, we consider problems that are large and sparse, although our methods are useful for problems with dense blocks, so long as matrix-vector products with those blocks can be performed efficiently.

In Chapter 2, we briefly survey major preconditioning strategies in the literature for generalized saddle-point problems. Here we largely follow the survey by Benzi, Golub and Liesen [5]. We also go into more detail about methods closely related to the ones we present.

Chapter 3 focuses on the development of four preconditioners for generalized saddle-point problems. The first two preconditioners will involve exact Schur complements of a matrix related to the saddle-point problem considered. As these Schur complement matrices can be expensive to form and solve with, variants of these preconditioners that allow for computationally inexpensive approximations are desirable. The other two preconditioners allow for such approximations. A detailed analysis of the eigenvalues of the preconditioned systems derived from all four preconditioners is included. This chapter, which represents original research, forms the theoretical core of this work. It will largely follow a recent paper by Siefert and de Sturler [53].

As we have developed preconditioners that allow for the use of approximate Schur complements, Chapter 4 will focus on general methods for generating approximations to Schur complements. Our discussion will focus on techniques based on *probing*, which was originally presented by Chan and Mathew [14] and other related techniques from the optimization community [19]. We present a more powerful version of the probing technique, which utilizes recent advances from the optimization community [31], which we will refer to as *structured probing* [54]. We present a theoretical justification of the application of these methods to Schur complements with a decay property and present some analysis of the quality of the approximations generated by this technique. Sections 4.1 and 4.2 largely follow a paper by Siefert and de Sturler [54] and like the rest of the chapter, represent original research.

Chapter 5 details several applications to which we apply our preconditioners. These applications range from fluid dynamics, to metal deformation, to mesh parameterization. Experimental results for these applications will be discussed in Chapter 6. This includes some results from both the papers by Siefert and de Sturler [53, 54] as well as new material. Finally, we will summarize our results in Chapter 7.

2 Background

We choose Krylov methods as our solution technique for generalized saddle-point problems. These methods are iterative techniques that solve the problem $\mathcal{A}x = b$, by solving an iterative sequence of minimization problems. For instance, GMRES [50], solves

$$\min_{y_i \in \mathcal{K}_i(\mathcal{A}, r_0)} \|b - \mathcal{A}y_i\|, \tag{2.1}$$

where

$$\mathcal{K}_i(\mathcal{A}, r_0) = \operatorname{span}\left\{r_0, \mathcal{A}r_0, \ldots, \mathcal{A}^{i-1}r_0\right\},$$

at each step. Here, $\mathcal{K}_i(\mathcal{A}, r_0)$ is called the Krylov space [58, p. 267], and $r_0 = b - \mathcal{A}x_0$ represents the initial residual. The convergence of these methods is heavily dependent on the location eigenvalues the matrix $\mathcal{A}$. If r_i represents the residual at step i, then

$$\frac{\|r_i\|}{\|b\|} \leq \kappa(V) \inf_{p_i \in P_i} \|p_i\|_{\Lambda(\mathcal{A})}, \tag{2.2}$$

where P_i represents the set of all polynomials of degree i or less with $p_i(0) = 1$, V represents the eigenvector matrix of $\mathcal{A}$ and $\Lambda(\mathcal{A})$ represents the set of the eigenvalues of $\mathcal{A}$ [58, p. 271], assuming that $\mathcal{A}$ is diagonalizible.

Note that in (2.2), the convergence rate of the method is influenced by the minimization of a polynomial over the set of eigenvalues. The closer together these eigenvalues are, the smaller this minimum value will be. Thus, to guarantee rapid convergence of our Krylov methods, we want to modify our linear system so that the eigenvalues are close together. We do this by either pre- or post-multiplying the system by another matrix chosen to cluster the eigenvalues. This technique is known as *preconditioning.* The preconditioner matrices are almost never applied directly to $\mathcal{A}$, instead they are applied in the form of matrix-vector products inside the Krylov method.

One other property possessed by some Krylov methods is that of *finite termination.* This means that if an $n \times x$ matrix has k distinct eigenvalues with $k < n$, a Krylov method with a finite termination property will converge to the exact solution in at most k iterations. We can see this clearly from (2.2). If a matrix has k distinct eigenvalues, then a polynomial of degree k can be found that will be zero at each of the eigenvalues. Thus, after k steps, we have found the exact solution.

With respect to generalized saddle-point problems, preconditioners fall into several categories. An extensive survey on solution methods for generalized saddle-point problems, including preconditioning techniques is given by Benzi, Golub and Liesen [5]. We briefly summarize preconditioning techniques here.

All of the below preconditioners have one characteristic in common — a linear system must be solved using a Schur complement matrix, or some approximation thereof. While an exact Schur complement is preferable from a theoretical perspective, sometimes the expense of forming and solving with the exact Schur complement can make the preconditioner too expensive to use. In these cases, preconditioners that allow for approximate Schur complements are an attractive alternative.

The first category of preconditioners consider is that of block-diagonal preconditioners. The basic block-diagonal preconditioner for (1.2) is

$$\mathcal{P}_1 \quad = \quad \begin{bmatrix} A & 0 \\ 0 & S_1 \end{bmatrix}, \tag{2.3}$$

where S_1 is the Schur complement for the matrix in (1.2). This preconditioner was proposed by several different authors. For the $D = 0$ case, it was proposed by Murphy, Golub and Wathen [44]. For the $D \neq 0$ case, it was proposed by Ipsen [38]. Discussion of these preconditioners will follow in Section 2.1. A variant of this preconditioner, which allows for an approximation to (splitting of) the (1,1) block A, was presented by de Sturler and Liesen [20], and is detailed in Section 2.2.

Fisher, Ramage, Silvester and Wathen [29] present a variant of (2.3) for the case where A is symmetric positive definite, $C = B$ and $D = 0$, namely

$$\mathcal{P}_2 = \begin{bmatrix} \frac{1}{\eta}A & 0 \\ 0 & \pm S_2 \end{bmatrix}, \tag{2.4}$$

where S_2 is an approximation to the Schur complement S_1. They provide an analysis of the eigenvalues of this preconditioned system based on the singular values of the preconditioned (1,2) block.

Silvester and Wathen [56, 60] discuss block-diagonal preconditioners specifically for stabilized Stokes problems, which have $B = C$, A positive definite and D negative semidefinite. They use positive definite approximations for both A and the Schur complement and show eigenvalue bounds for the preconditioned system.

The second category of preconditioners we consider is that of block-triangular preconditioners. These preconditioners are of the form

$$\mathcal{P}_3 = \begin{bmatrix} F & 0 \\ C & S_1 \end{bmatrix}, \tag{2.5}$$

where $F \approx A$ and S_1 is the Schur complement matrix. They were introduced by Bramble and Pasciak [12]. The systems they consider have $B = C$, A positive definite and D negative semidefinite. The Uzawa method [59], can also be viewed as a reformulated block-triangular preconditioner. Block-triangular preconditioners for several variants of (1.2) are discussed in more detail in Section 2.1.

The third category of preconditioners we consider is that of constraint or indefinite preconditioners. These preconditioners have the a similar block-structure to the original matrix. These constraint preconditioners take the form

$$\mathcal{P}_4 \;=\; \begin{bmatrix} F & B^T \\ C & D \end{bmatrix}, \tag{2.6}$$

where $F \approx A$. Applying the inverse of these preconditioners requires the application of the inverse of the Schur complement. We can see this clearly from the factorization

$$\mathcal{P}_4 \;=\; \begin{bmatrix} I & 0 \\ CF^{-1} & I \end{bmatrix} \begin{bmatrix} F & 0 \\ 0 & S_1 \end{bmatrix} \begin{bmatrix} I & F^{-1}B^T \\ 0 & I \end{bmatrix}.$$

Eigenvalue analysis for the preconditioned systems in the case where $A = A^T$, $C = B$ and $D = 0$ is provided by Keller, Gould and Wathen [39]. For the case where A and D are symmetric, and $C = B$, eigenvalue analysis is provided by Dollar [23].

In the case where $D = 0$, de Sturler and Liesen [20] propose and analyze a variant of a constraint preconditioner that is described in more detail in Section 2.2. For a particular problem in the A s.p.d., $C = B$ and $D \neq 0$ case, Perugia and Simoncini [47] also present a constraint preconditioner. Variants of $\mathcal{P}_4$ using approximate Schur complements have also been developed, notably by Perugia and Simoncini [47]. Their work is discussed in more detail in Section 2.3.

2.1 The Preconditioners of Murphy, Golub, Wathen and of Ipsen

Murphy, Golub and Wathen [44] present two preconditioners for problem (1.2) in the special case where $D = 0$. The first such preconditioner, which was later extended by de Sturler and Liesen [20], is

$$\mathcal{P}_5 = \begin{bmatrix} A & 0 \\ 0 & CA^{-1}B^T \end{bmatrix}, \tag{2.7}$$

assuming that both A, and the Schur complement $CA^{-1}B^T$ are invertible (otherwise (1.2) is singular). The preconditioned matrix has the form

$$\mathcal{P}_5^{-1}\mathcal{A} = \begin{bmatrix} I & A^{-1}B^T \\ (CA^{-1}B^T)^{-1}C & 0 \end{bmatrix}, \tag{2.8}$$

has at most three distinct non-zero eigenvalues, namely 1, $\frac{1\pm\sqrt{5}}{2}$ [44, Remark 1]. This also holds for both right and symmetric preconditioning. A natural consequence of this property of the eigenvalues is that Krylov methods with a finite termination property will converge in at most three iterations [44, Remark 3]. However, this is more expensive than a direct solution of the problem [20].

Murphy, Golub and Wathen also propose a second, symmetric indefinite preconditioner, namely

$$\mathcal{P}_6 = \begin{bmatrix} A & B^T \\ 0 & CA^{-1}B^T \end{bmatrix}. \tag{2.9}$$

The preconditioned matrix, $\mathcal{P}_2^{-1}\mathcal{A}$, has exactly two eigenvalues, namely ± 1 [44, Remark 4]. It is noted that one can multiply the (2,2) block of (2.9) by -1 to yield a preconditioned system where the only eigenvalue is one, but such a system is no longer diagonalizable.

Ipsen [38] proposes three preconditioners as extensions of Murphy, Golub and Wathen's work [44] to systems of form (1.2) with a non-zero (2,2) block ($D \neq 0$). The first, a direct analogue of (2.7), serves as an inspiration for our work in Chapter 3, along with the work of de Sturler and Liesen [20]. Ipsen suggests the preconditioner

$$\mathcal{P}_7 = \begin{bmatrix} A & 0 \\ 0 & -(D - CA^{-1}B^T) \end{bmatrix}, \tag{2.10}$$

and notes that the preconditioned system has at most three distinct non-zero eigenvalues, which, although she does not enumerate them, are clearly 1 and $\frac{1 \pm \sqrt{5}}{2}$ [38, Remark 1].

Ipsen also suggests an analogue for (2.9), namely,

$$\mathcal{P}_8 = \begin{bmatrix} A & B^T \\ 0 & -(D - CA^{-1}B^T) \end{bmatrix}. \tag{2.11}$$

Again, the preconditioned system has eigenvalues ±1, and the (2,2) block can be multiplied by -1 to yield a preconditioned system where the only eigenvalue is one [38, Propositions 1 and 2].

2.2 The Preconditioners of de Sturler and Liesen

De Sturler and Liesen [20] developed extensions of the preconditioners of Murphy, Golub and Wathen [44] (but not Ipsen [38]) to allow for the use of approximations to the inverse of the (1,1) block, A. They propose splitting the (1,1) block into

$$A = F - E, \tag{2.12}$$

in a fashion such that F is invertible. The resulting block-diagonal preconditioner is

$$\mathcal{P}_9 \quad = \quad \begin{bmatrix} F & 0 \\ 0 & CF^{-1}B^T \end{bmatrix}, \tag{2.13}$$

with the assumption that $CF^{-1}B^T$ is also invertible. Note that $CF^{-1}B^T$ is the Schur complement of the matrix

$$\begin{bmatrix} F & B^T \\ C & 0 \end{bmatrix}. \tag{2.14}$$

Note that this matrix is a constraint preconditioner. The preconditioned matrix $\mathcal{P}_9^{-1}\mathcal{A}$ is then,

$$\mathcal{P}_9^{-1}\mathcal{A} \quad = \quad \begin{bmatrix} I - F^{-1}E & F^{-1}B^T \\ (CF^{-1}B^T)^{-1}C & 0 \end{bmatrix}. \tag{2.15}$$

The preconditioned matrix, regardless of whether left, right or symmetric preconditioning is used, is of the form

$$\mathcal{B}(F) \quad = \quad \begin{bmatrix} I - S & N \\ M & 0 \end{bmatrix}, \tag{2.16}$$

where S, N and M are defined based on the type of preconditioning used. The authors also define the matrix

$$\mathcal{B}(0) \quad = \quad \begin{bmatrix} I & N \\ M & 0 \end{bmatrix}, \tag{2.17}$$

which plays a crucial role in the eigenvalue perturbation analysis of the preconditioned matrix, as well as in the derivation of their second preconditioned system. The second preconditioned system offered by de Sturler and

Liesen [20], requires first the inverse of $\mathcal{B}(0)$, namely

$$\mathcal{B}(0)^{-1} = \begin{bmatrix} I - NM & N \\ M & -I \end{bmatrix}. \tag{2.18}$$

When we multiply the block-diagonally preconditioned system (2.16) by $\mathcal{B}(0)^{-1}$, what is referred to as the *related system* results, namely

$$\begin{bmatrix} I - (I - NM)S & 0 \\ -MS & I \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \hat{f} \\ \hat{g} \end{bmatrix}. \tag{2.19}$$

This system possesses better eigenvalue clustering and iterative methods converge faster than those run on systems preconditioned with (2.15) [20, 40].

A key component of the work of de Sturler and Liesen [20] is the analysis of the eigenvalue perturbation of the preconditioned matrices (2.15) and (2.19). As the discussion in Chapter 3 expands on this analysis, we repeat some of their major results. First, they present the eigenvalue decomposition of the matrix $\mathcal{B}(0)$, (2.17), namely,

$$\mathcal{B}(0)\mathcal{Y} = \mathcal{Y} \begin{bmatrix} I & 0 & 0 \\ 0 & \Lambda^{+} & 0 \\ 0 & 0 & \Lambda^{-} \end{bmatrix}, \tag{2.20}$$

$$\text{where } \mathcal{Y} = \begin{bmatrix} U_1 & U_2 & U_2 \\ 0 & \Lambda^{+^{-1}} M U_2 & \Lambda^{-^{-1}} M U_2 \end{bmatrix}, \tag{2.21}$$

$U_1 \in \mathrm{I\!R}^{n \times (n-m)}$ such that $MU_1 = 0$, $U_2 \in \mathrm{I\!R}^{n \times m}$ forms a basis for range (NM), and $\Lambda^{\pm} = \frac{1 \pm \sqrt{5}}{2} I_m$ [20, Theorem 3.3]. It is important to note that both U_1 and U_2 are orthogonal matrices.

From here, de Sturler and Liesen employ a well-known result in matrix perturbation theory [57, Theorem IV.1.12] to yield the eigenvalue perturbation bound:

$$|\lambda_{\mathcal{B}} - \lambda| \quad \leq c_S \left\| [U_1, U_2]^{-1} S [U_1, U_2] \right\|, \tag{2.22}$$

where $c_S = \sqrt{2 + \frac{1-\sqrt{5}}{5}}$, $\lambda_{\mathcal{B}}$ is an eigenvalue of $\mathcal{B}(F)$ and $\lambda \in \left\{1, \frac{1\pm\sqrt{5}}{2}\right\}$ [20, Theorem 3.5]. This bound is restated in its final form [20, Corollary 3.7] in terms of ω_1, the maximum singular value of the matrix $U_1^T U_2$, namely,

$$|\lambda_{\mathcal{B}} - \lambda| \quad \leq c_S \left(\frac{1+\omega_1}{1-\omega_1}\right)^{1/2} \|S\|. \tag{2.23}$$

For the related system matrix (2.19), they present the bound,

$$|1 - \lambda_R| \quad \leq (1 - \omega_1^2)^{-1/2} \|S\|, \tag{2.24}$$

where λ_R is an eigenvalue of the matrix (2.19) [20, Theorem 4.4].

Besides eigenvalue results, de Sturler and Liesen present a result on the satisfaction of the second set of equations (constraints) of (1.2) when $D = 0$. This result applies to Krylov methods applied to the related system (2.19). Specifically, if the starting Krylov iterate, x_0, satisfies $Cx_0 = 0$, then every successive Krylov iterate, x_k, also satisfies $Cx_k = 0$ [20, Theorem 4.2]. If no better guess is available, the preconditioned right-hand side, $\hat{f}$, satisfies $C\hat{f} = 0$, and can be used as a starting guess [20, Theorem 4.1].

2.3 Preconditioners of Perugia and Simoncini

Perugia and Simoncini consider the problem (1.2), with $C = B^T$, A positive definite, D negative semidefinite and the Schur complement matrix $-(D - BB^T)$ positive definite [47]. The specific application they consider is a mixed finite-element formulation of a magnetostatics problem. In this particular application both A and the Schur complement matrix $-(D - BB^T)$ are spectrally equivalent to the identity [47, Lemma 1], which influences their choice of preconditioners.

Perugia and Simoncini present four preconditioners, two block-diagonal and two indefinite. All of these are closely related to the preconditioners described in Sections 2.1 and 2.2 and those which will be described in Chapter 3. Their block-diagonal preconditioner with an exact Schur complement is

$$\mathcal{P}_{10} = \begin{bmatrix} I & 0 \\ 0 & -(D - BB^T) \end{bmatrix}. \tag{2.25}$$

When the problem is preconditioned symmetrically by $\mathcal{P}_{10}$, then the preconditioned eigenvalues are either -1 or cluster in two intervals which are functions of the spectral equivalence constants for the (1,1) block A [47, Proposition 2].

Recognizing that forming the Schur complement matrix $S_1 = -(D - BB^T)$ could be expensive, a second, "quasi-optimal," block-diagonal preconditioner is also offered. In the notation which will be described in Chapter 3, let S_2 be an approximation to the Schur complement matrix $S_1 = -(D - BB^T)$. Then, their "quasi-optimal" block-diagonal preconditioner is

$$\mathcal{P}_{11} = \begin{bmatrix} I & 0 \\ 0 & S_2 \end{bmatrix}. \tag{2.26}$$

The preconditioners (2.25) and (2.26) are a special cases of (3.1) and (3.47) respectively with a fixed "splitting" of $F = I$. The eigenvalues of the system preconditioned by $\mathcal{P}_{11}$ are in two intervals, which are the intervals for the system preconditioned by $\mathcal{P}_{10}$, rescaled by the spectral equivalence constants between $\mathcal{P}_{10}$ and $\mathcal{P}_{11}$ [47, Proposition 3].

The third preconditioner they present is the indefinite preconditioner

$$\mathcal{P}_{12} = \begin{bmatrix} I & B^T \\ B & D \end{bmatrix}, \tag{2.27}$$

The eigenvalues of this system are either 1 or in an interval defined by the spectral equivalence constants for A. This preconditioner can be factored as

$$\mathcal{P}_{12} = \begin{bmatrix} I & 0 \\ B & I \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & (D - BB^T) \end{bmatrix} \begin{bmatrix} I & B^T \\ B & I \end{bmatrix}, \tag{2.28}$$

and this factorization used to derive the “quasi-optimal” version of their indefinite preconditioner. Again, let S_2 be an approximation to the (Schur complement) matrix $S_1 = -(D - BB^T)$. The “quasi-optimal” indefinite preconditioner is then

$$\mathcal{P}_{13} = \begin{bmatrix} I & B^T \\ B & BB^T - S_2 \end{bmatrix}, \tag{2.29}$$

$$= \begin{bmatrix} I & 0 \\ B & I \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & -S_2 \end{bmatrix} \begin{bmatrix} I & B^T \\ 0 & I \end{bmatrix}. \tag{2.30}$$

Concrete bounds on the location of the eigenvalues of the problem as preconditioned by $\mathcal{P}_{13}$ are not presented, though qualitative statements about their clustering are made and computed eigenvalues are shown for their target application.

3 Preconditioners: Theory and Analysis

The discussion in this chapters largely follows the recent work by Siefert and de Sturler [53]. Specifically, we first discuss preconditioners with exact Schur complements in Sections 3.1 and 3.2. This largely follows the discussions of Sections 2 and 3 of Siefert and de Sturler [53], respectively. After this, we discuss similar these preconditioners where we replace the exact Schur complement matrix with an approximation. This discussion, in Section 3.3 and 3.4, largely follows the discussion of Section 4 of Siefert and de Sturler [53].

3.1 Block-Diagonal Preconditioner with Exact Schur Complements

The block-diagonal preconditioner of de Sturler and Liesen [20], (2.15), is only applicable to problems of the form (1.2) where $D = 0$. We propose

$$\mathcal{P}(F) = \begin{bmatrix} F & 0 \\ 0 & -(D - CF^{-1}B^T) \end{bmatrix}, \tag{3.1}$$

as a generalization of their preconditioner to the $D \neq 0$ case [53]. We note that preconditioning either from the left or the right yields a system of the form

$$\mathcal{B}(F)\begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix} = \begin{bmatrix} I - S & N \\ M & Q \end{bmatrix}\begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix} = \begin{bmatrix} \tilde{f} \\ \tilde{g} \end{bmatrix}, \tag{3.2}$$

where $\mathcal{B}(F)$ is either $\mathcal{P}^{-1}\mathcal{A}$ or $\mathcal{A}\mathcal{P}^{-1}$. More explicitly,

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

Both two-sided (using an LU factorization) preconditioning and symmetric preconditioning (if $F = F^T$, $D = D^T$ and $C = B$) are also possible.

While this preconditioned system shares many properties with the case described by de Sturler and Liesen, we note that several key properties change. In the $D = 0$ case, $MN = I$ [20]. Here we have

$$\begin{aligned}
MN &= -(D - CF^{-1}B^T)^{-1}CF^{-1}B^T, \\
&= -(D - CF^{-1}B^T)^{-1}(-D + CF^{-1}B^T + D), \\
&= I + Q. \qquad (3.3)
\end{aligned}$$

This is true for the left-preconditioned, right-preconditioned, two-sided and symmetric cases. In the $D = 0$ case NM is a projector [20]. In our case, it is not, as $(NM)^2 = NM + NQM$ [53].

3.1.1 Eigenvalue Analysis of $\mathcal{B}_0$

The location of the eigenvalues of the block-diagonally preconditioned system play an important role in determining the convergence of Krylov methods. Consequently, we begin our analysis of this preconditioner by considering the eigenvalues of the matrix

$$\mathcal{B}_0 = \begin{bmatrix} I & N \\ M & Q \end{bmatrix}, \qquad (3.4)$$

and bounding the eigenvalues of the preconditioned matrix $\mathcal{B}(F)$ using perturbation theory. We make the assumption that $I+Q$ (and therefore B^T and C) have full rank and that Q is diagonalizable. We discuss the rank-deficient case in Section 3.1.2. Our goal is to find λ, u and v such that

$$u + Nv = \lambda u \tag{3.5}$$

$$Mu + Qv = \lambda v. \tag{3.6}$$

We begin by considering all eigenpairs such that $\lambda = 1$. From (3.3), we have $Q = MN - I$. We substitute $\lambda = 1$ into (3.5) and $Q = MN - I$ into (3.6) to yield

$$Nv = 0 \quad \text{and} \quad Mu = 2v. \tag{3.7}$$

Since B^T has full (column) rank (by assumption), $Nv = 0$ if and only if $v = 0$. Thus, $\mathcal{B}_0$ has eigenpairs of the form

$$\left(1, \begin{bmatrix} u \\ 0 \end{bmatrix}\right), \qquad \text{where } u \in \operatorname{null}(M). \tag{3.8}$$

Since we have assumed that C (and thus M) has full (row) rank, then $\mathcal{B}_0$ has $n - m$ eigenpairs of this type.

Next, we consider $\lambda \neq 1$. We solve (3.5) for u and substitute the result into (3.6), yielding

$$\lambda Qv = (\lambda^2 - \lambda - 1)v. \tag{3.9}$$

Therefore, v must be an eigenvector of Q. Since Q is diagonalizable (by assumption), we have $Qv_j = \delta_j v_j$, for $j = 1, \ldots, m$. Solving (3.9) for λ

yields

$$\lambda_j^{\pm} \quad = \quad \frac{(1+\delta_j) \pm \sqrt{4+(1+\delta_j)^2}}{2}. \tag{3.10}$$

Substituting $\delta_j v_j$ for Qv_j in (3.6) gives u. Finally, we rescale the eigenvector by $(\lambda_j^{\pm} - 1)$ to yield

$$\left(\lambda_j^{\pm}, \begin{bmatrix} Nv_j \\ (\lambda_j^{\pm} - 1)v_j \end{bmatrix} \right). \tag{3.11}$$

Note that $\lambda_j^- \neq 1$ for any δ_j and $\lambda_j^+ = 1$ only if $\delta_j = -1$. By assumption, $I + Q$ has full rank and this is precluded. Therefore, $\mathcal{B}_0$ has $2m$ independent eigenvectors corresponding to $\lambda \neq 1$ and is diagonalizable.

Let $\Lambda^+ = \operatorname{diag}(\lambda_j^+)$ and $\Lambda^- = \operatorname{diag}(\lambda_j^-)$, (where $\operatorname{diag}(\cdot)$ denotes a diagonal matrix with the arguments given). Let U_1 be an orthonormal basis for $\operatorname{null}(M)$, cf. (3.8), and let U_2 be the matrix with normalized columns such that $u_j = Nv_j$, where $Qv_j = \delta_j v_j$, cf. (3.11). Then an eigenvector matrix of $\mathcal{B}_0$ is

$$\mathcal{Y} \equiv \left[\begin{array}{c|c} Y_{11} & Y_{12} \\ \hline Y_{21} & Y_{22} \end{array} \right] \quad = \quad \left[\begin{array}{cc|c} U_1 & U_2 & U_2 \\ \hline 0 & V(\Lambda^+ - I) & V(\Lambda^- - I) \end{array} \right]. \tag{3.12}$$

We can write the block-wise inverse [37, Section 0.7.3] of $\mathcal{Y}$,

$$\mathcal{Z} = \mathcal{Y}^{-1} \quad = \quad \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix}. \tag{3.13}$$

Let $\Upsilon^+ = \operatorname{diag}((\lambda_j^- - 1)/(\lambda_j^- - \lambda_j^+))$ and $\Upsilon^- = \operatorname{diag}((\lambda_j^+ - 1)/(\lambda_j^- - \lambda_j^+))$.

Then,

$$\begin{align*}
Z_{11} &= \left(Y_{11} - Y_{12}Y_{22}^{-1}Y_{21}\right)^{-1}, \\
&= \begin{bmatrix} I_{n-m} & 0 \\ 0 & \Upsilon^{+} \end{bmatrix} Y_{11}^{-1} \;=\; \hat{I}_n Y_{11}^{-1}, \tag{3.14} \\
Z_{21} &= -Y_{22}^{-1}Y_{21}Z_{11}, \\
&= -\begin{bmatrix} 0 & (\Lambda^{-} - I)^{-1}(\Lambda^{+} - I)\Upsilon^{+} \end{bmatrix} Y_{11}^{-1}, \\
&= -\begin{bmatrix} 0 & \Upsilon^{-} \end{bmatrix} Y_{11}^{-1}. \tag{3.15}
\end{align*}$$

Using $[U_1\ U_2]^{-1}NV = [0\ I]^T$ we also have

$$\begin{align*}
Z_{22} &= (Y_{22} - Y_{21}Y_{11}^{-1}Y_{12})^{-1}, \\
&= \left(V(\Lambda^{-} - I) - \begin{bmatrix} 0 & V(\Lambda^{+} - I) \end{bmatrix} \begin{bmatrix} U_1 & U_2 \end{bmatrix}^{-1} NV\right)^{-1}, \\
&= \left(V(\Lambda^{-} - I) - \begin{bmatrix} 0 & V(\Lambda^{+} - I) \end{bmatrix} \begin{bmatrix} 0 \\ I \end{bmatrix}\right)^{-1}, \\
&= \left(V(\Lambda^{-} - \Lambda^{+})\right)^{-1}, \tag{3.16} \\
Z_{12} &= Y_{11}^{-1}Y_{22}Z_{22} \;=\; -\begin{bmatrix} U_1 & U_2 \end{bmatrix}^{-1} U_2 Z_{22}, \\
&= -\begin{bmatrix} 0 \\ (\Lambda^{-} - \Lambda^{+})^{-1}V^{-1} \end{bmatrix}. \tag{3.17}
\end{align*}$$

For the $D = 0$ case, this reduces to the decomposition given by de Sturler and Liesen [20]. Lemma 3.1.1, which is unique to this work, simplifies the proofs which follow.

Lemma 3.1.1. *Let $\mathcal{Y}$ be defined as in (3.12). Let*

$$L_1 = \begin{bmatrix} I & 0 & 0 \\ 0 & \Upsilon^+ & 0 \\ 0 & 0 & \Upsilon^- \end{bmatrix}, \tag{3.18}$$

$$L_2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & -(\Lambda^- - \Lambda^+)^{-1} & 0 \\ 0 & 0 & (\Lambda^- - \Lambda^+)^{-1} \end{bmatrix}, \tag{3.19}$$

$$R_1 = \begin{bmatrix} U_1 & 0 & 0 \\ 0 & U_2 & 0 \\ 0 & 0 & U_2 \end{bmatrix}, \tag{3.20}$$

$$R_2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & (\Lambda^+ - I) & 0 \\ 0 & 0 & (\Lambda^- - I) \end{bmatrix}. \tag{3.21}$$

Then,

$$\mathcal{Y}^{-1} \begin{bmatrix} R & 0 \\ 0 & 0 \end{bmatrix} \mathcal{Y} = L_1 \begin{bmatrix} Y_{11}^{-1} R \\ -[0\ I] Y_{11}^{-1} R \end{bmatrix} \begin{bmatrix} I & I & I \end{bmatrix} R_1, \tag{3.22}$$

$$\mathcal{Y}^{-1} \begin{bmatrix} 0 & T \\ 0 & 0 \end{bmatrix} \mathcal{Y} = L_1 \begin{bmatrix} Y_{11}^{-1} TV \\ -[0\ I] Y_{11}^{-1} TV \end{bmatrix} \begin{bmatrix} 0 & I & I \end{bmatrix} R_2, \tag{3.23}$$

$$\mathcal{Y}^{-1} \begin{bmatrix} 0 & 0 \\ X & 0 \end{bmatrix} \mathcal{Y} = L_2 \begin{bmatrix} 0 \\ V^{-1} X \\ V^{-1} X \end{bmatrix} \begin{bmatrix} I & I & I \end{bmatrix} R_1, \tag{3.24}$$

$$\mathcal{Y}^{-1} \begin{bmatrix} 0 & 0 \\ 0 & W \end{bmatrix} \mathcal{Y} = L_2 \begin{bmatrix} 0 \\ V^{-1} WV \\ V^{-1} WV \end{bmatrix} \begin{bmatrix} 0 & I & I \end{bmatrix} R_2. \tag{3.25}$$

Proof: For the sake of notation, denote the matrices being multiplied by $\mathcal{Y}^{-1}$ and $\mathcal{Y}$ as $\hat{R}, \hat{T}, \hat{X}$ and $\hat{W}$, respectively. For the non-zero (1,1) block

(3.22), we have

$$\begin{aligned}
\mathcal{Y}^{-1}\hat{R}\mathcal{Y} &= \begin{bmatrix} \mathcal{Z}_{11}R\mathcal{Y}_{11} & \mathcal{Z}_{11}R\mathcal{Y}_{12} \\ \mathcal{Z}_{21}R\mathcal{Y}_{11} & \mathcal{Z}_{21}R\mathcal{Y}_{12} \end{bmatrix}, \\
&= \begin{bmatrix} \hat{I}_n Y_{11}^{-1}RU_1 & \hat{I}_n Y_{11}^{-1}RU_2 & \hat{I}_n Y_{11}^{-1}RU_2 \\ -[0\ \Upsilon^-]Y_{11}^{-1}RU_1 & -[0\ \Upsilon^-]Y_{11}^{-1}RU_2 & -[0\ \Upsilon^-]Y_{11}^{-1}RU_2 \end{bmatrix}.
\end{aligned}$$

For the non-zero (1,2) block (3.23), we have

$$\begin{aligned}
\mathcal{Y}^{-1}\hat{T}\mathcal{Y} &= \begin{bmatrix} \mathcal{Z}_{11}T\mathcal{Y}_{21} & \mathcal{Z}_{11}T\mathcal{Y}_{22} \\ \mathcal{Z}_{21}T\mathcal{Y}_{21} & \mathcal{Z}_{21}T\mathcal{Y}_{22} \end{bmatrix}, \\
&= \begin{bmatrix} \hat{I}_n[0\ Y_{11}^{-1}TV(\Lambda^+ - I)] & \hat{I}_n Y_{11}^{-1}TV(\Lambda^- - I) \\ -[0\ \Upsilon^-][0\ Y_{11}^{-1}TV(\Lambda^+ - I)] & -[0\ \Upsilon^-]Y_{11}^{-1}TV(\Lambda^- - I) \end{bmatrix}.
\end{aligned}$$

For the non-zero (2,1) block (3.24), we have

$$\begin{aligned}
\mathcal{Y}^{-1}\hat{X}\mathcal{Y} &= \begin{bmatrix} \mathcal{Z}_{12}X\mathcal{Y}_{11} & \mathcal{Z}_{12}X\mathcal{Y}_{12} \\ \mathcal{Z}_{22}X\mathcal{Y}_{11} & \mathcal{Z}_{22}X\mathcal{Y}_{12} \end{bmatrix}, \\
&= L_2 \begin{bmatrix} 0 & 0 & 0 \\ V^{-1}XU_1 & V^{-1}XU_2 & V^{-1}XU_2 \\ V^{-1}XU_1 & V^{-1}XU_2 & V^{-1}XU_2 \end{bmatrix}.
\end{aligned}$$

For the non-zero (2,2) block (3.25), we have

$$\begin{aligned}
\mathcal{Y}^{-1}\hat{W}\mathcal{Y} &= \begin{bmatrix} \mathcal{Z}_{12}W\mathcal{Y}_{21} & \mathcal{Z}_{12}W\mathcal{Y}_{22} \\ \mathcal{Z}_{22}W\mathcal{Y}_{21} & \mathcal{Z}_{22}W\mathcal{Y}_{22} \end{bmatrix}, \\
&= L_2 \begin{bmatrix} 0 & 0 & 0 \\ 0 & V^{-1}WV(\Lambda^+ - I) & V^{-1}WV(\Lambda^- - I) \\ 0 & V^{-1}WV(\Lambda^+ - I) & V^{-1}WV(\Lambda^- - I) \end{bmatrix}.
\end{aligned}$$

□

Our next theorem, a modified version of [53, Theorem 2.1], allows us to bound the eigenvalues of the preconditioned system. This theorem applies regardless of whether the the block-diagonal preconditioner is applied from the left, from the right or from both sides.

Theorem 3.1.2. *Consider matrices $\mathcal{B}(F)$ of the form (3.2). Let $\mathcal{Y}$ be an eigenvector matrix of $\mathcal{B}_0$ given in (3.12). Then, for each eigenvalue $\lambda_{\mathcal{B}}$ of $\mathcal{B}(F)$, there exists an eigenvalue λ of $\mathcal{B}_0$, such that*

$$\begin{aligned} |\lambda_{\mathcal{B}} - \lambda| &\leq \left\| \mathcal{Y}^{-1} \begin{bmatrix} S & 0 \\ 0 & 0 \end{bmatrix} \mathcal{Y} \right\| && (3.26) \\ &\leq 2 \max\left(1, \|\Upsilon^+\|, \|\Upsilon^-\|\right) \|Y_{11}^{-1} S Y_{11}\|. && (3.27) \end{aligned}$$

Proof: Since $\mathcal{B}_0$ is diagonalizable, (3.26) follows from a classic result in perturbation theory [57, Theorem IV.1.12]. Using Lemma 3.1.1 with $R = S$ gives

$$\begin{aligned} |\lambda_{\mathcal{B}} - \lambda| &\leq \left\| L_1 \begin{bmatrix} Y_{11}^{-1} S \\ -[0\ I] Y_{11}^{-1} S \end{bmatrix} \begin{bmatrix} I & I & I \end{bmatrix} R_1 \right\|, \\ &\leq \max(1, \|\Upsilon^+\|, \|\Upsilon^-\|) \left\| \begin{bmatrix} Y_{11}^{-1} S \\ -[0\ I] Y_{11}^{-1} S \end{bmatrix} \begin{bmatrix} U_1 & U_2 & U_2 \end{bmatrix} \right\|. \end{aligned}$$

Using the consistency of the 2-norm we can simplify this to (see also [20])

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

□

The $\Upsilon^{\pm}$ terms can only be large if $\delta_j \approx -1 \pm 2i$. For the problems discussed in Chapter 5, the δ_j's are well-separated from this value. This is because $\|D\|$ is small and the problems and preconditioners are relatively well-conditioned. Lemma 3.1.3 [53, Lemma 2.2] provides bounds on the $\|\Upsilon^{\pm}\|$. The lemma considers the case where the δ_j's are real (and thus bounded away from $-1 \pm 2i$) separately. This is condition holds in the important case that D is symmetric and the Schur complement is definite. We first define

$$p(z) = 4 + (1+z)^2, \tag{3.28}$$

and now restate [53, Lemma 2.2].

Lemma 3.1.3. *Let Υ^+ and Υ^- be defined as above.*

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

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

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

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

Proof: Substituting $\lambda_j^{\pm}$ from (3.10) in $\Upsilon^+ = \operatorname{diag}(\lambda_j^- - 1)/(\lambda_j^- - \lambda_j^+)$ and $\Upsilon^- = \operatorname{diag}(\lambda_j^+ - 1)/(\lambda_j^- - \lambda_j^+)$ gives

$$\begin{aligned} \Upsilon^{\pm} &= \operatorname{diag}\left(\frac{1-\delta_j}{2\sqrt{4+(1+\delta_j)^2}} \pm \frac{1}{2}\right), \\ &= \operatorname{diag}\left(\frac{1-\delta_j}{2\sqrt{p(\delta_j)}} \pm \frac{1}{2}\right). \end{aligned} \tag{3.29}$$

For the real case, consider the diagonal entries of the Υ's as functions of $\delta_j \in \mathbb{R}$. These functions have a critical point at $\delta_j = -3$. The maximal absolute value of the diagonal entries of Υ^+, $(1+\sqrt{2})/2$, is found at this critical point.

The supremum of the diagonal entries of Υ^- occurs when δ_j is large, and that value is 1. In the case of $\delta_j \geq -1$, the absolute value of Υ^+ at $\delta_j = -1$ is maximal, namely 1. Therefore, in this case $\max(1, \|\Upsilon^+\|, \|\Upsilon^-\|) = 1$.

For the complex case, note that $p(\delta) = (\delta + 1 + 2i)(\delta + 1 - 2i)$. Since the distance between $-1-2i$ and $-1+2i$ is 4, 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 $|p(\delta)|$ completes the proof. □

In practice, the bound for the complex case is relatively 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.

Our derivation of a bound on $\left\|Y_{11}^{-1}SY_{11}\right\|$ follows the approach of de Sturler and Liesen [20]. Recall from (3.12) 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$\,(NM)$. This allows us to restate [53, Lemma 2.3].

Lemma 3.1.4. *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,*

$$\left\|Y_{11}^{-1}SY_{11}\right\| \leq \kappa(\Theta)\left(\frac{1+\omega_1}{1-\omega_1}\right)^{1/2}\|S\|. \tag{3.30}$$

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

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

As V_2 is orthogonal, we have $\sigma(U_2) = \sigma(\Theta)$, where $\sigma(\cdot)$ represents the set of singular values its argument. As U_2 has normalized columns, $\|U_2\| \geq 1$ and

$\|U_2^{-1}\| \geq 1$. Thus, $\|\Theta\| \geq 1$, and $\|\Theta^{-1}\| \geq 1$. So, our bound simplifies to

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

where the second inequality follows from the bound on $\kappa([U_1 \ \ V_2])$ from Lemma 3.6 in [20]. □

Finally we can bring the above results together [53, Corollary 2.4].

Corollary 3.1.5. *Let Θ and ω_1 be defined as above.*

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

$$|\lambda_{\mathcal{B}} - \lambda| \quad \leq \quad (1+\sqrt{2})\kappa(\Theta) \left(\frac{1+\omega_1}{1-\omega_1}\right)^{1/2} \|S\|. \tag{3.31}$$

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

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

Proof: Use Lemmas 3.1.3 and 3.1.4 in Theorem 3.1.2. □

We see that the clustering of the eigenvalues depends mainly on $\|S\|$ and the size of the δ_j, unless $\omega_1 \approx 1$, or $\kappa(\Theta)$ large. This implies that the block-diagonally preconditioned system can have as many as $2m+1$ eigenvalue clusters, one for $\lambda = 1$ and one for each $\lambda_j^{\pm}$. Even in the case where $\|S\|$ is small, the convergence of Krylov methods may still be poor for the block-diagonally preconditioned system. Examples in Chapter 6 will illustrate this. However, the block-diagonal preconditioner is a useful step to a better preconditioner, described in Section 3.2.

3.1.2 Rank Deficiency in $I + Q$

In Section 3.1, we assumed that $I + Q$ is full rank. Now we relax that assumption. There are three potential sources of rank-deficiency in $I + Q$.

The first two are rank-deficiency in C and B^T. The third is when there exist vectors v such that $Nv \neq 0$ and $Nv \in \text{null}\,(M)$. This implies that $MNv = (I+Q)v = 0$, where v is an eigenvector of Q. This occurs when F^{-1} (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 (3.10) 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 and splitting are invertible, M must also be rank deficient by l_c. Thus, the number of eigenpairs of the form (3.8) 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.

By (3.7), 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 3.1 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 [35]) for $\lambda = 1$. These vectors satisfy the equations

$$Nv = \tilde{u} \quad \text{and} \quad Mu = 2v, \tag{3.32}$$

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 (3.7) or (3.32).

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 $k = l_b = l_c$, $k - l_c = 0$, 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$\,(U_1) =$ null$\,(M)$. Let $\widetilde{V}$ be such that $\widetilde{V}^T \widetilde{V} = I_{l_c}$ and range$(\widetilde{V}) =$ 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

$$\mathcal{Y} \quad = \quad \left[\begin{array}{cc|cc} U_1^{(n-m+l_c)} & U_2^{(m-l_c)} & N\tilde{V}^{(l_c)} & U_2^{(m-l_c)} \\ \hline 0 & \widehat{V}(\widehat{\Lambda}^+ - I) & -2\widetilde{V} & \widehat{V}(\widehat{\Lambda}^- - I) \end{array} \right], \qquad (3.33)$$

where superscripts in the top row indicate the number of columns. The corresponding eigenvalues are those from (3.8) and (3.10). We can then use the eigenvector matrix of $\mathcal{B}_0$ given in (3.33) 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 3.1.5 . An important example of this case is the stabilized Navier-Stokes (Oseen) problem [26], where $C = B$ and F is positive definite.

3.2 Related System with Exact Schur Complements

Following de Sturler and Liesen [20], we outline an alternative solution method. In the $D = 0$, this approach leads to so-called constraint preconditioners, cf. [10, 11, 33, 47]. We begin our derivation of a similar system for the $D \neq 0$ case by splitting (3.1) as follows,

$$\begin{aligned} \mathcal{B}(F) \begin{bmatrix} x \\ y \end{bmatrix} &= \begin{bmatrix} I - S & N \\ M & Q \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}, \\ &= \left(\mathcal{B}_0 - \begin{bmatrix} S & 0 \\ 0 & 0 \end{bmatrix} \right) \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \tilde{f} \\ \tilde{g} \end{bmatrix}. \end{aligned} \tag{3.34}$$

Note that

$$\mathcal{B}_0^{-1} = \begin{bmatrix} I - NM & N \\ M & -I \end{bmatrix}. \tag{3.35}$$

Left-multiplying (3.34) by $\mathcal{B}_0^{-1}$ and splitting yields the fixed point iteration,

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} (I - NM)S & 0 \\ MS & 0 \end{bmatrix} \begin{bmatrix} x_k \\ y_k \end{bmatrix} + \begin{bmatrix} \hat{f} \\ \hat{g} \end{bmatrix}. \tag{3.36}$$

This iteration is essentially the same as the $D = 0$ case described in [11, 20]. As x_{k+1} and y_{k+1} depend only on x_k, we need to iterate only on the x_k variables; cf. [9, pp. 214–215] and [20]. The x-component of the fixed point of (3.36) satisfies the so-called *related system* for the fixed-point iteration [35],

$$(I - (I - NM)S)x = \hat{f}. \tag{3.37}$$

The full-size related system (including the y component) and $D \neq 0$ has been examined elsewhere for special cases. In [47], A is symmetric positive definite and spectrally equivalent to the identity, and so a fixed splitting $F = I$ is used. In [33], F is symmetric positive definite. In both of these cases $B = C$.

3.2.1 Eigenvalue Analysis

Let U_1 and U_2 be defined as in (3.12), $\Delta = \mathrm{diag}(\delta_j)$ and let $U_2 = V_2\Theta$, with $V_2^T V_2 = I$. Then, $NMU_1 = 0$, $NMU_2 = NMNV = NV(I + \Delta)$, and therefore

$$(I - NM) \;=\; \begin{bmatrix} U_1 & V_2 \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & -\Theta\Delta\Theta^{-1} \end{bmatrix} \begin{bmatrix} U_1 & V_2 \end{bmatrix}^{-1}. \tag{3.38}$$

In the rank-deficient case, we can use (3.33). Thus, 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),

$$U_1^T V_2 = \Phi\Omega\Psi^T, \text{ where } 1 > \omega_1 \geq \omega_2 \geq \ldots \geq \omega_m. \tag{3.39}$$

Following [20], 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,

$$V_2 = U_1\Phi\Omega\Psi^T + W\Sigma\Psi^T. \tag{3.40}$$

This allows us to restate [53, Theorem 3.1]

Theorem 3.2.1. *Let U_1,V_2 and ω_1 be defined as above. Let λ_R be an eigenvalue of the related system matrix in (3.37). Then,*

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

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

Proof: The proof of this theorem largely follows [20]. 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,

$$\begin{aligned}
\|I-NM\| &= \left\| [U_1\, V_2] \begin{bmatrix} I & 0 \\ 0 & Z \end{bmatrix} [U_1\, V_2]^{-1} \right\|, && (3.41) \\
&\leq \left\| [U_1\, V_2] \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} [U_1\, V_2]^{-1} \right\| \\
&\quad + \left\| [U_1\, V_2] \begin{bmatrix} 0 & 0 \\ 0 & Z \end{bmatrix} [U_1\, V_2]^{-1} \right\|, && (3.42) \\
&\leq (1-\omega_1^2)^{-1/2}(1+\|Z\|). && (3.43)
\end{aligned}$$

Note that the first term in (3.42) is the norm of an oblique projection. Given the SVD in (3.39), this norm equals $(1-\omega_1^2)^{-1/2}$ [43, Section 5.15]. We establish the bound for the second term as follows

$$\left\| [U_1\, V_2] \begin{bmatrix} 0 & 0 \\ 0 & Z \end{bmatrix} [U_1\, V_2]^{-1} \right\| = \max_{U_1a+V_2b\neq 0} \frac{\|V_2Zb\|}{\|U_1a+V_2b\|}.$$

Without loss of generality we may assume $\|b\|=1$, so that $\|V_2Zb\| \leq \|Z\|$. From (3.40) we see that $\|U_1a+V_2b\| = \|U_1a+U_1\Phi\Omega\Psi^Tb+W\Sigma\Psi^Tb\|$, which for any given b is minimized by $a = -\Phi\Omega\Psi^Tb$. This gives $\|U_1a+V_2b\| = \|W\Sigma\Psi^Tb\|$, which in turn is minimized for $b=\psi_1$. Hence, we have

$$\begin{aligned}
\left\| [U_1\, V_2] \begin{bmatrix} 0 & 0 \\ 0 & Z \end{bmatrix} [U_1\, V_2]^{-1} \right\| &= \max_{U_1a+V_2b\neq 0} \frac{\|V_2Zb\|}{\|U_1a+V_2b\|}, \\
&\leq (1-\omega_1^2)^{-1/2}\|Z\|. && (3.44)
\end{aligned}$$

So, using (3.41)–(3.44) we have $\rho((I-NM)S) \leq (1-\omega_1^2)^{-1/2}(1+\|\Theta\Delta\Theta\|)\|S\|$, which concludes our proof. □

The following corollary [53, Corollary 3.2] shows that the influence of $\kappa(\Theta)$ need not be large if the δ_j's are well-clustered.

Corollary 3.2.2. *Let* $\hat{\delta} = \arg\min_{z\in\mathbb{C}} \max_j |z - \delta_j|$ *and* $\tilde{\delta}_j = \delta_j - \hat{\delta}$*, then*

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

Proof: Note that $\Delta = \hat{\delta}I + \mathrm{diag}(\tilde{\delta}_j)$, 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 3.2.2 is no larger than the the corresponding term for the block-diagonally preconditioned system (Corollary 3.1.5). 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.

3.2.2 Satisfying 'Constraints'

In the $D = 0$ case [20], the second block of equations in (1.2) often represents a set of constraints. For the $D \neq 0$ case, this may or may not true. So-called constraint preconditioners (in the $D = 0$ case) have the advantage that each iterate of a Krylov subspace method for the preconditioned system satisfies the constraints, if the initial guess is chosen appropriately. Fixed point methods such as (3.36) often satisfy the constraints after a single step. This is the case for the fixed-point method proposed in [20] for $D = 0$. We can show an analogous property for the $D \neq 0$ case, by restating [53, Lemma 3.3].

Lemma 3.2.3. *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 (3.36) satisfy $Mx_k + Qy_k = \tilde{g}$ in (3.1) and $Cx_k + Dy_k = g$ in (1.2).*

Proof: From (3.34)–(3.36) and the equality $MN = I + Q$ we have

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

Thus, the fixed-point method satisfies the second block of equations of (3.36) exactly after one step. Because the block diagonal preconditioner (3.1) is invertible, the second block of equations of (1.2) are also satisfied after one step. □

Following trivially from this, we restate [53, Corollary 3.4].

Corollary 3.2.4. *After the first iteration of (3.36), all fixed-point updates are in the null space of $[M\ Q]$.*

We can also show that the iterates of a Krylov subspace method will satisfy the constraints if the initial guess satisfies the constraints (cf. [20]). We first restate a general result, namely [53, Theorem 3.5], 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 (1.2).

Theorem 3.2.5. *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

$$x^{(m)}=x^{(0)}+\sum_{i=0}^{m-1}\alpha_i(I-A)^i r^{(0)}=x^{(0)}+\gamma_0 r^{(0)}+A\sum_{i=1}^{m-1}\gamma_i A^{i-1}r^{(0)}. \quad (3.45)$$

Finally, we multiply (3.45) by C, and note that $Cx^{(0)}=d$, $Cr^{(0)}=0$ and $CA=0$. Therefore,

$$Cx^{(m)}=Cx^{(0)}+\gamma_0 Cr^{(0)}+CA\sum_{i=1}^{m-1}\gamma_i A^{i-1}r^{(0)}=d. \quad (3.46)$$

□

From here we can restate [53, Corollary 3.6].

Corollary 3.2.6. *The iterates, $[x^{(m)^T}, y^{(m)^T}]^T$, of any Krylov method applied to the full $n+m$ related system for (3.36) satisfy $Mx^{(m)}+Qy^{(m)}=\tilde{g}$ and $Cx^{(m)}+Dy^{(m)}=g$ if the initial guess is the result of at least one step of fixed point iteration (3.36).*

Proof: Use Theorem 3.2.5, with A as fixed-point iteration matrix in (3.36), $b=[\hat{f}^T\ \hat{g}^T]^T$, $C=[M\ Q]$ and $d=\hat{g}$. □

3.3 Block-Diagonal Preconditioner with Approximate Schur Complements

As the Schur complement matrix $S_1=-D(-CF^{-1}B^T)$ may be expensive to compute or factor, we would like to be able to use inexpensive approximations instead. We now consider the effect of such an approximation on the eigenvalue clustering of the preconditioned matrices. Let

$S_2 \approx S_1$ denote our approximation to the Schur complement. For the left-preconditioned case, let $S_2^{-1}S_1 = I + \mathcal{E}$. For the right-preconditioned case, let $S_1 S_2^{-1} = I + \mathcal{E}$. Two-sided preconditioning can be handled similarly. Our new block-diagonal preconditioner is then

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

We refer to the resulting preconditioned matrix as $\mathcal{B}(F, S_2)$. The left-preconditioned system of equations is

$$\begin{aligned} \begin{bmatrix} \tilde{f} \\ \tilde{g} \end{bmatrix} &= \begin{bmatrix} I - S & N \\ M_2 & Q_2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}, \\ &= \left(\begin{bmatrix} I & N \\ M & Q \end{bmatrix} - \begin{bmatrix} S & 0 \\ -\mathcal{E}M & -\mathcal{E}Q \end{bmatrix} \right) \begin{bmatrix} x \\ y \end{bmatrix}, \end{aligned} \qquad (3.47)$$

where M, N and Q are defined as in Section 3.1, $M_2 = S_2^{-1}C$ 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 $Q_2 = (I + \mathcal{E})Q$. The right-preconditioned system is

$$\begin{aligned} \begin{bmatrix} \tilde{f} \\ \tilde{g} \end{bmatrix} &= \begin{bmatrix} I - S & N_2 \\ M & Q_2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}, \\ &= \left(\begin{bmatrix} I & N \\ M & Q \end{bmatrix} - \begin{bmatrix} S & -\mathcal{E}N \\ 0 & -\mathcal{E}Q \end{bmatrix} \right) \begin{bmatrix} x \\ y \end{bmatrix}, \end{aligned} \qquad (3.48)$$

where M, N and Q are defined as in Section 3.1, $N_2 = B^T S_2^{-1}$ and $Q_2 = DS_2^{-1}$. Note also that $N_2 = B^T S_1^{-1} S_1 S_2^{-1} = N(I + \mathcal{E})$ and analogously $Q_2 = Q(I + \mathcal{E})$.

Using (3.47), we can bound the eigenvalues of $\mathcal{B}(F, S_2)$ by considering the perturbation of the eigenvalues of $\mathcal{B}_0$ analogously to our bounds in Section 3.1. We consider first the left-preconditioned version, restating [53, Theorem 4.1].

Theorem 3.3.1. *Let $\lambda_{S,\mathcal{E}}$ be an eigenvalue of the (left-preconditioned) matrix $\mathcal{B}(F, S_2)$, λ be an eigenvalue of $\mathcal{B}_0$ and $Qv_j = \delta_j v_j$.*

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

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

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

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

3. If $D = 0$, then

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

Proof: In Section 3.1.1 we have already derived the eigendecomposition of $\mathcal{B}_0$. From this decomposition we get the following perturbation bound (see [57, Theorem IV.1.12]),

$$\begin{aligned} |\lambda_{\mathcal{B}} - \lambda| &\leq \left\|\mathcal{Y}^{-1}\begin{bmatrix} S & 0 \\ -\mathcal{E}M & -\mathcal{E}Q \end{bmatrix}\mathcal{Y}\right\|, \\ &\leq \left\|\mathcal{Y}^{-1}\begin{bmatrix} S & 0 \\ 0 & 0 \end{bmatrix}\mathcal{Y}\right\| + \left\|\mathcal{Y}^{-1}\begin{bmatrix} 0 & 0 \\ \mathcal{E}M & \mathcal{E}Q \end{bmatrix}\mathcal{Y}\right\|. \end{aligned} \tag{3.49}$$

Corollary 3.1.5 gives bounds for the first term in (3.49). 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

$$\begin{bmatrix} 0 & 0 \\ \mathcal{E}M & \mathcal{E}Q \end{bmatrix}\mathcal{Y} \;=\; \left[\begin{array}{cc|c} 0 & 0 & 0 \\ \hline 0 & -\mathcal{E}V(I+\Delta\Lambda^+) & -\mathcal{E}V(I+\Delta\Lambda^-) \end{array}\right] \tag{3.50}$$

We then multiply (3.50) from the left by $\mathcal{Y}^{-1}$, see (3.13)–(3.17), and refactor to yield

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

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

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

The remainder of the proof concerns the bounds on the right hand side of (3.51) for each particular case.

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

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

Clearly, $|1/(\lambda_j^- - \lambda_j^+)|$ obtains its maximum at $\delta_j = -1$. This yields $|1/(\lambda_j^- - \lambda_j^+)| \leq 1/2$. We can use this in (3.51) 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 3.1.3 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

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

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 (3.51) yields

$$\|\mathcal{X}\| \leq \frac{2+(1+\sqrt{5})\alpha + 2\alpha^2}{\sqrt{2\left(\sqrt{5}-\alpha\right)}}\kappa(V)\|\mathcal{E}\|. \tag{3.51}$$

We can then substitute this result into (3.49) 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 (3.49) using Theorem 3.1.2, Lemma 3.1.3 for $\delta \geq -1$ and Lemma 3.1.4 where $\kappa(\Theta) = 1$. This follows from the fact that U_2 can be chosen to be orthogonal (see [20]).

For the second term in (3.49), since $Q = 0$, $\delta_j = 0$, so $\lambda_j^- - \lambda_j^+ = -\sqrt{5}$, and we can choose $V = I$. We then substitute this into (3.51). □

As a side note, in the complex case the term involving α 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.

For right preconditioning, introduce two lemmas (not found in [53]), which bound the norms of two key quantities.

Lemma 3.3.2. *Let $\Lambda^\pm$ be defined as above. Then,*

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

$$\|\Lambda^\pm - I\| \quad \leq \quad \frac{1 + \alpha + \sqrt{4 + (1+\alpha)^2}}{2}.$$

2. If $D = 0$, then

$$\|\Lambda^\pm - I\| \quad = \quad \frac{1 + \sqrt{5}}{2}.$$

The proof of this lemma is trivial.

Lemma 3.3.3. *Let $\Lambda^\pm$ and Δ be defined as above. Then,*

1. If $\delta_j \in \mathrm{I\!R}$, for $j = 1, \ldots, m$, then

$$\|(\Lambda^- - \Lambda^+)^{-1}\Delta\| \quad \leq \quad \frac{\sqrt{5}}{2}.$$

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

$$\|(\Lambda^- - \Lambda^+)^{-1}\Delta\| \quad \leq \quad \frac{\alpha}{\sqrt{2(\sqrt{5} - \alpha)}}.$$

3. If $D = 0$, then

$$\|(\Lambda^- - \Lambda^+)^{-1}\Delta\| \quad = \quad 0.$$

Proof: For the real case, note that the the supremum of $|(\lambda_j^- - \lambda_j^+)^{-1}\delta_j|$ as a function of δ_j occurs at $\delta_j = -5$. For the complex case, use the lower bound on $|p(\delta_j)|$ developed in the proof of Lemma 3.1.3 [53, Lemma 2.2]. The $D = 0$ case is trivial. □

These lemmas allow us to state bounds for the right-preconditioned system which are not included in [53]. Recall that for the right-preconditioned case, $Q_2 = Q(I + \mathcal{E})$ and $N_2 = N(I + \mathcal{E})$.

Theorem 3.3.4. *Let $\lambda_{S,\mathcal{E}}$ be an eigenvalue of the (right-preconditioned) matrix $\mathcal{B}(F, S_2)$, λ be an eigenvalue of $\mathcal{B}_0$ and $Qv_j = \delta_j v_j$.*

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

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

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

$$\begin{aligned} |\lambda_{S,\mathcal{E}} - \lambda| \quad \leq \quad & 2\max\left(1, \frac{1}{2} + \frac{1+\alpha}{2\sqrt{2\left(\sqrt{5}-\alpha\right)}}\right)\kappa(\Theta)\left(\frac{1+\omega_1}{1-\omega_1}\right)^{1/2}\|S\| \\ & + \max\left(1 + \frac{\alpha}{\sqrt{2(\sqrt{5}-\alpha)}}, \frac{1}{2} + \frac{1+3\alpha}{2\sqrt{2(\sqrt{5}-\alpha)}}\right) \\ & \cdot\left(1+\alpha+\sqrt{4+(1+\alpha)^2}\right)\kappa(V)\|\mathcal{E}\|. \end{aligned}$$

3. If $D = 0$, then

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

Proof: Using Lemma 3.1.1, we have $R = S$, $T = N\mathcal{E}$ and $W = Q\mathcal{E}$. We handle the (1,1) block using Corollary 3.1.5 [53, Corollary 2.4].

For the (2,1) block (3.24), the $\|L_1\|$ term is bounded by Lemma 3.1.3 [53, Lemma 2.2] and Lemma 3.3.2 bound the $\|R_2\|$ term in the complex and $D = 0$ cases. For the middle term, we note that $N = U_2V^{-1}$, so $\|Y_{11}^{-1}U_2\| = \|[U_1\ U_2]^{-1}U_2\| = 1$. So we have

$$\left\| \begin{bmatrix} Y_{11}^{-1}N\mathcal{E} \\ -[0\ I]Y_{11}^{-1}N\mathcal{E} \end{bmatrix} \begin{bmatrix} 0 & I & I \end{bmatrix} \right\| \leq 2\kappa(V)\|\mathcal{E}\|.$$

For the (2,2) block (3.25), we note that $QV = V\Delta$, so $V^{-1}Q = \Delta V^{-1}$. Using this on the middle block, allows us to employ Lemma 3.3.3. We treat the $|\|R_2\|$ term as before. □

3.4 Related System with Approximate Schur Complements

Following the approach of Section 3.2 to generate the related system for this problem, we would generate precisely the related system derived from (3.36), with S_1^{-1} instead of S_2^{-1} [20]. Instead, we use an alternative splitting of $\mathcal{B}(F, S_2)$. Though shown for the left-preconditioned system, the splitting is the same for right-preconditioning.

$$\mathcal{B}(F, S_2) = \begin{bmatrix} I & N \\ M_2 & Q_2 + \mathcal{E} \end{bmatrix} - \begin{bmatrix} S & 0 \\ 0 & \mathcal{E} \end{bmatrix},$$

We then 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, for the left-preconditioned system, we have

$$\begin{bmatrix} I-(I-NM_2)S & -N\mathcal{E} \\ -M_2S & I+\mathcal{E} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \hat{f} \\ \hat{g} \end{bmatrix}. \tag{3.52}$$

For the right-preconditioned system, we have

$$\begin{bmatrix} I-(I-N_2M)S & -N_2\mathcal{E} \\ -MS & I+\mathcal{E} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \hat{f} \\ \hat{g} \end{bmatrix}. \tag{3.53}$$

For a problem in magnetostatics, a linear system similar to (3.52) was derived by Perugia and Simoncini [47]. Their preconditioned system (2.30) is described in Section 2.3. If we employ the same choices for the splitting and approximations, we obtain basically the same system to be solved. However, Perugia and Simoncini [47] only outline the qualitative behavior of the eigenvalues in the case that $\mathcal{E}$ is sufficiently small.

Now we present a bound on the eigenvalues of the left-preconditioned system (the bound for the right-preconditioned system is almost identical) [53, Theorem 4.2].

Theorem 3.4.1. *For any eigenvalue, λ_R, of the related system matrix (3.52),*

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

Proof: Note that the matrix in (3.52) 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 [57]) yields

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

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

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

□

The terms $\|N\|$ and $\|M_2\|$ in the bound from Theorem 3.4.1 are generally benign. They are bounded by the norms of the off-diagonal blocks of the unpreconditioned matrix (1.2) 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 (1.2), unless (1.2) 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.

4 Probing Methods for Approximating Schur Complements

Replacing the exact Schur complement in the preconditioners discussed in Chapter 3 with an approximation can significantly improve their efficiency. Although in some cases a natural approximation is cheaply available (e.g. the pressure mass matrix for Navier-Stokes), in general this is not the case. While it is generally inexpensive to multiply with the Schur complement matrix without forming it (i.e. multiplying with each of the component pieces) forming it can be prohibitively expensive. Thus, we restrict our attention to matrix approximation techniques that can be performed using only matrix-vector multiplication. Such techniques have been around since the work of Curtis, Powell and Reid [19] and they were first applied to Schur complement problems by Chan and Mathew [14]. However, the extensions to the work of Curtis, Powell and Reid [19] by the optimization community [15, 16, 41] have only recently been used for Schur complement approximation. We outline this work by Siefert and de Sturler [54] in Sections 4.1 and 4.2. The remainder of the chapter is original.

4.1 Background and Algorithmic Statement

Curtis, Powell and Reid [19] introduced a technique for the estimation of a sparse Jacobian matrix by finite differencing. They note that if the sparsity pattern of the Jacobian matrix is known — specifically, if it is banded with bandwidth b — then the matrix can be exactly reconstructed using b matrix-vector products. The idea behind this method is to partition the columns of the matrix into sets such for any row, there is at most one column in the set that has a non-zero entry in that row. For each set, one then multiplies the matrix with a single vector, which exactly captures the entries in the columns belonging to that set. For banded matrices, computing this set of columns can be done with modular arithmetic. This procedure is detailed in Algorithm 1. It was first applied to the approximation of Schur complement matrices by Chan and Mathew [14] where it referred to as *probing.* We will refer to it as *banded probing* to differentiate it from more advanced techniques. As a convention we will assume that vectors are numbered starting at 1.

```
Algorithm 1   $\widetilde{K} = \mathsf{Banded\ Probing}(K \in \mathbb{R}^{n\times n}, b \in \mathbb{Z}^n)$
1: Generate $b$ vectors, $x_1, \ldots, x_b$, such that
       $$x_i(j) = \begin{cases} 1 & \text{if } j \bmod b == i \bmod b \\ 0 & \text{otherwise} \end{cases}$$
2: Compute $w_i = Kx_i$, for $i = 1, \ldots, p$.
3: Build $\widetilde{K}$ such that
       $$\tilde{K}_{i,j} = w_i\left((j-1) \bmod b + 1\right)$$
Note: If $b$ is the bandwidth of $K$, then $\widetilde{K} = K$ and probing is exact.
```

Figure 4.1 shows a how steps 1 and 2 of the banded probing algorithm work for a small tridiagonal matrix. We take b is taken to be the bandwidth of the matrix, so that all of the entries are exactly captured. One drawback of this approach is that it does not necessarily generate symmetric approximations $\widetilde{K}$, even if K is symmetric [14]. To remedy this, Chan and Mathew propose an alternative algorithm (which is not linear in K), which is described in Algorithm 2.

Although Curtis, Powell and Reid saw the potential to extend probing methods to non-banded matrices, they offered no suggestion on how to accomplish this. Nearly a decade later, McCormick [41] and Coleman and Moré [15] developed such a technique in the context of their work on sparse Jacobians and Hessians. These authors independently noted that

$$
\begin{bmatrix}
\mathbf{a_1} & b_2 & & & \\
\mathbf{c_1} & a_2 & \mathrm{b}_3 & & \\
 & c_2 & \mathrm{a}_3 & \mathbf{b_4} & \\
 & & \mathrm{c}_3 & \mathbf{a_4} & b_5 \\
 & & & \mathbf{c_4} & a_5
\end{bmatrix}
\begin{bmatrix}
\mathbf{1} & 0 & 0 \\
0 & \mathit{1} & 0 \\
0 & 0 & 1 \\
\mathbf{1} & 0 & 0 \\
0 & \mathit{1} & 0
\end{bmatrix}
=
\begin{bmatrix}
\mathbf{a_1} & b_2 & 0 \\
\mathbf{c_1} & a_2 & \mathrm{b}_3 \\
\mathbf{b_4} & c_2 & \mathrm{a}_3 \\
\mathbf{a_4} & b_5 & \mathrm{c}_3 \\
\mathbf{c_4} & a_5 & 0
\end{bmatrix}
$$

Figure 4.1. Banded probing on a tridiagonal matrix, using the vectors $e_1 + e_4$, $e_2 + e_5$ and e_3.

Algorithm 2 $\widetilde{K}$ = Symmetric Banded Probing($K \in \mathbb{R}^{n \times n}, b \in \mathbb{Z}^+$)

```
1: $M$ =Banded Probing($K, b$).
2: Let
       $\widetilde{K}_{i,j} = \begin{cases} M_{i,j} & \text{if } |M_{i,j}| < |M_{j,i}| \\ M_{j,i} & \text{otherwise} \end{cases}$
```

Note: $\widetilde{K} = \widetilde{K}^T$. If b is the bandwidth of K, then $\widetilde{K} = K$ and probing is exact.

the problem of partitioning the columns of a matrix into sets such that for any row, there is at most one column in the set that has a non-zero entry in that row, can be viewed as a graph coloring problem. Given a graph $G = (V, E)$ with vertices V and edges E, a graph coloring assigns colors to the vertices such that no two neighboring vertices have the same color. This means that if $(v_i, v_j) \in E$, v_i and v_j must have different colors. From this coloring, a set of vectors can be chosen to perform the probing process. This general approach, which is referred to by Cullum and Tůma as partial matrix estimation [18], will be called *structured probing* and is described on Algorithm 3.

Like the banded case, we must have*a priori* knowledge of the sparsity pattern of the matrix K. Here we use a sparsity matrix $H \in \{0, 1\}^{n \times n}$ to represent that pattern, as opposed to choosing a bandwidth b as in the banded case. For an exact reconstruction, we would choose H to have the same sparsity pattern as K. For an approximation, we would choose H so that the large entries of K are captured. We then use that matrix H to define a graph G (step 1) on which we will perform the graph coloring (step 2). The graph coloring problem is NP-complete [17, Chapter 36], so a variety of different heuristics are available. The question of which graph and which coloring heuristic to choose will be dealt with in Section 4.2

```
Algorithm 3   $\widetilde{K} = \mathsf{Structured\ Probing}(K \in \mathbb{R}^{n\times n}, H \in \{0,1\}^{n\times n})$
1: Compute a graph $G$ derived from $H$.
2: Perform a graph coloring on $G$ to generate a mapping $\phi : \{1,\ldots,n\} \to \{1,\ldots,p\}^n$, where $p$ is the number of colors. The color for vertex $i$ is given by $\phi(i)$.
3: Generate the matrix of probing vectors $X \in \{0,1\}^{n\times p}$ such that
       $X_{i,j} = \begin{cases} 1 & \text{if } \phi(i) = j, \\ 0 & \text{otherwise.} \end{cases}$
4: Compute $W = KX$.
5: Build $\widetilde{K}$ using $W$ and the sparsity pattern of $H$.
Note: If $K_{i,j} \neq 0 \to H_{i,j} \neq 0$, then $\widetilde{K} = K$ and structured probing is exact.
```

After choosing H, the graph G and the coloring algorithm to generate the mapping ϕ, computing the matrix of probing vectors, X, (step 3) is straightforward. We then multiply by K (step 4) and then perform the construction of $\widetilde{K}$ (step 5). This final step is not difficult, For each i, j s.t. $H_{i,j} \neq 0$, we set $\widetilde{K}_{i,j} = W_{i,\phi(j)}$.

Questions with respect to the choice of graph, G, and coloring algorithm will be discussed in more detail in Section 4.2. We discuss the analytic properties of structured probing in Section 4.3 and the applicability of structured probing to Schur complement matrices in Section 4.4. Next, we analyze of the quality of approximations to Schur complements generated by structured probing in Section 4.5. Finally, in Section 4.6 we discuss of the use of inexact factorizations and inverses in the context of structured probing.

4.2 Graph Coloring and Structured Probing

A undirected graph, $G = (V, E)$, is defined by a set of vertices V and a set of unordered pairs of vertices, E, called edges. Vertices u and v are adjacent if $\{u, v\} \in E$. A path of length k is a sequence of vertices $v_1, \ldots, v_{k+1}$ such that v_i is adjacent to v_{i+1}, for $i = 1, \ldots, k$. Two vertices are distance-k neighbors if the shortest path connecting them has length at most k. Let $N_k(v)$ represent the set of distance-k neighbors of v. Directed graphs are defined similarly, except that the edge pairs are ordered, meaning that u is adjacent to v if and only if $(u, v) \in E$. The existence of $(v, u) \in E$, is irrelevant to u being adjacent to v. Distance metrics, path lengths and neighborhoods on a directed graph are defined accordingly.

A distance-k coloring of a graph $G = (V, E)$ is a mapping $\phi : V \to \{1, \ldots, p\}$, such that $\phi(u) \neq \phi(v)$ if u and v are distance-k neighbors. This mapping ϕ assigns a color to each vertex in the graph. The standard, "textbook" graph coloring (coloring G such that no two distance-1 neighbors have the same color) is a distance-1 coloring. Distance-2 coloring, then, requires that distance-1 neighbors, and distance-1 neighbors of distance-1 neighbors, have different colors. Let $\Delta(G)$ and $\delta(G)$ be the maximum and minimum vertex degree of G, respectively. Let $\bar{\delta_2}(G)$ be the average number of distance-2 neighbors per vertex in G.

Consider a sparse symmetric matrix $H \in \{0, 1\}^{n \times n}$. Define the vertex set $V = \{v_1, \ldots, v_n\}$, that is one vertex corresponding to each column of H. For the purpose of structured probing, each color and its associated set of columns corresponds to a single probing vector. Thus we need to choose a graph and coloring such that

$$\forall i, j, k \in \{1, \ldots, n\} \text{ s.t. } H_{i,j}, H_{i,k} \neq 0, \qquad \phi(v_j) \neq \phi(v_k). \tag{4.1}$$

This property ensures that the coloring partitions the matrix into sets of columns with non-overlapping sparsity patterns.

The *column intersection graph*, $G_1(H) = (V, E_1)$, is defined such that

$$(v_i, v_j) \in E_1, \qquad \text{if } \exists k \text{ s.t. } H_{k,i} \neq 0 \text{ and } H_{k,j} \neq 0.$$

This means that two vertices are adjacent if their columns have sparsity patterns that intersect. Performing a distance-1 coloring on this graph satisfies (4.1). The use of this particular graph was proposed by Coleman and Moré [15].

McCormick [41] proposed the use of the *adjacency graph*, $G_2(H) = (V,E_2)$, where

$$(v_i, v_j) \in E_2, \qquad \text{if } \exists H_{i,j} \neq 0.$$

For this graph, a distance-1 coloring is not sufficient to satisfy (4.1). Instead, a distance-2 coloring is required. As Gebremedhin, Manne and Pothen [31] show experimentally, this approach is significantly more efficient in terms of space complexity and execution time, although the methods have the same time complexity. In light of this, we will use distance-2 coloring on G_2 for all of the non-symmetric matrices we consider. Algorithm 4, as presented by Gebremedhin, et al. [31], shows the simplest greedy distance-2 coloring algorithm on an arbitrary graph G. This algorithm works by coloring each vertex with the lowest numbered legal color. It determines this by recording all of the illegal colors in the array forbiddenColors and choosing the lowest-numbered remaining color. Running this algorithm takes $O(n\bar{\delta_2})$ time [31, Lemma 3.9].

Algorithm 4 ϕ = Greedy Distance-2 Coloring($G = (V, E)$)

```
Initialize forbiddenColors with some $a \notin V$.
For $i = 1, \dots, n$
    For each colored vertex $w \in N_2(v_i)$
        forbiddenColors[$\phi(w)$] = $v_i$.
    Set $\phi(v_i)$ = minimum $c$ s.t. forbiddenColors[$c$] $\neq v_i$.
```

Cullum and Tůma [18] suggest balanced coloring as an alternative approach. This approach attempts to maximize the number of vertices colored by the least commonly used color at each step. Thus, when a number of colors are legal for a given vertex, we choose the color that has been used the least. Algorithm 5 states a heuristic algorithm for balanced coloring.

Algorithm 5 $\phi = \mathsf{Balanced\ Distance\text{-}2\ Coloring}(G = (V, E))$

```
Set $p = 1 + \max\deg(V)$, the initial number of colors to consider.
Set $c_j = 0$, for $j = 1, \ldots, p$, be the number of times each color has been used.
Initialize $\mathsf{forbiddenColors}$ with some $a \notin V$.
For $i = 1, \ldots, n$
    For each colored vertex $w \in N_2(v_i)$
        $\mathsf{forbiddenColors}[\phi(w)] = v_i$.
    Set $colored$ = false and $nuses = \infty$.
    For $j = 1, \ldots, p$
        If $\mathsf{forbiddenColors}[j] \neq v_i$ and $c_j < nuses$ then
            Set $nuses = c_j$, $colored$ = true and $\phi(v_i) = j$.
    If $colored$ = false then
        Set $\phi(v_i) = p + 1$.
        $p = p + 1$.
    Set $c_{\phi(v_i)} = c_{\phi(v_i)} + 1$.
```

The idea behind this heuristic is that by balancing the number of nodes colored by each color, the algorithm should use fewer colors. For generating approximations with structured probing, this heuristic has a potential advantage in reducing the error $\widetilde{K}$. Probing vectors corresponding to heavily used colors will often have more error than vectors corresponding to less frequently used colors. This error comes from the lumping in of small entries in K that are not captured by H. Running this algorithm takes $O(n(\Delta(G)^2 + p))$ time.

If we approximate a symmetric matrix, we can use symmetry to reduce the number of colors used. In this case, we do not need to accurately capture the (i, j) entry if we can reconstruct the (j, i) entry accurately. This modified technique is known as distance-$\frac{3}{2}$ coloring. Algorithm 6, developed by Powell and Toint [48], shows a greedy distance-$\frac{3}{2}$ coloring algorithm. This particular statement of the algorithm is taken from Gebremedhin, et. al [31].

For the purpose of comparison with banded probing techniques, we introduce one final coloring method that we call *prime divisor* coloring. This algorithm computes a distance-2 coloring that is implicitly performed on

Algorithm 6 ϕ = Greedy Distance-$\frac{3}{2}$ Coloring($G = (V, E)$)

```
Initialize forbiddenColors with some $a \notin V$.
For $i = 1, \ldots, n$
   For each $w \in N_1(v_i)$
      If $w$ is colored then
         forbiddenColors[$\phi(w)$] = $v_i$.
      For each colored vertex $x \in N_1(w)$
         If $w$ is not colored then
            forbiddenColors[$\phi(x)$] = $v_i$.
         else
            If $phi(x) < \phi(w)$ then
               forbiddenColors[$\phi(x)$] = $v_i$.
   Set $\phi(v_i)$ = minimum $c$ s.t. forbiddenColors[$c$] $\neq v_i$.
```

the adjacency graph of the matrix. In this approach, we choose probing vectors that have regular patterns, as the vectors banded probing do. However, we choose the number of vectors such that we can exactly reconstruct the desired sparsity pattern. We do this by choosing the number of vectors to be relatively prime to the differences between column indices of nonzero coefficients in a row, for all rows. More precisely, choose the number of colors p, such that p is relatively prime to all elements of the set $\{\ i - j \mid \text{for some } k,\ H_{k,i} = 1 \text{ and } H_{k,j} = 1\}$. This algorithm implicitly performs a distance-2 coloring on the adjacency graph of H, although we do not use the graph explicitly. Instead, we operate directly on H. Algorithm 7 presents this approach. If H comes from a fixed stencil on a regular grid,

Algorithm 7 ϕ = Prime Divisor Distance-2 Coloring($H \in \{0, 1\}^{n,n}$)

```
Initialize illegalP with the empty set.
For $i = 1, \ldots, n$
   For $j$ s.t. $H_{i,j} \neq 0$
      For $k$ s.t. $H_{i,k} \neq 0$ and $k > j$
         illegalP $\leftarrow (k - j)$.
For $i = 2, \ldots, n$
   If $\forall j \in$ illegalP, $j \bmod i \neq 0$ then
      $p = i$.
      break
For $i = 1, \ldots, n$
   Set $\phi(v_i) = i \bmod p$.
```

we need only consider a single "representative" row of the matrix (i.e., a row for a point away from the boundaries), and use that row to choose the number of colors p. For such problems, this algorithm takes $O(p + \Delta(G)^2)$ work, where $\Delta(G)$ is the highest degree vertex on the mesh. However, this algorithm takes $O(n(p+\Delta(G)^2))$ work for an unstructured meshes. It can be accelerated somewhat by only using prime numbers in Step 5 of Algorithm 7, but that does not improve the asymptotic efficiency.

4.3 Analytic Properties of Probing Methods

In their discussion of probing, Chan and Mathew outline several analytic properties of approximation generated by banded probing. The algorithm is linear in K and preserves (strict) diagonal dominance [14]. We will prove similar properties for structured probing. In the notation of Algorithm 3, (4.1) guarantees that X, the matrix of probing vectors, has exactly 1 non-zero entry per-row. Let $W = KX$, where K is the matrix we are approximating. Then,

$$W_{i,j} = \sum_{\phi(k)=j} K_{i,k}, \qquad \text{for } i = 1, \ldots, n \text{ and } j = 1, \ldots, p. \tag{4.2}$$

We can use this to derive several theoretical results.

Theorem 4.3.1. *Let H, K be valid input matrices for Algorithm 3 and $\widetilde{K}$ be its output. Then for any i,*

$$\sum_{j=1}^{n} |\widetilde{K}_{i,j}| \leq \sum_{j=1}^{n} |K_{i,j}|.$$

Proof: Let ϕ be the coloring generated in Algorithm 3. In building the approximation $\widetilde{K}$ in Step 5 of Algorithm 3,

$$\forall H_{i,j} \neq 0 \ , \ \widetilde{K}_{i,j} = W_{i,\phi(j)}. \tag{4.3}$$

Thus,

$$\forall i, \quad \sum_{\substack{j \\ H_{i,j}\neq 0}} |\widetilde{K}_{i,j}| \;=\; \sum_{\substack{j \\ H_{i,j}\neq 0}} |W_{i,\phi(j)}|.$$

The property (4.1) guarantees that this sum will never include the same entry of W twice. Thus, we have

$$\begin{aligned}
\forall i, \quad \sum_{\substack{j \\ H_{i,j}\neq 0}} |W_{i,\phi(j)}| &\leq \sum_{k=1}^{p} |W_{i,k}|, \\
&= \sum_{k=1}^{p}\sum_{j=1}^{n} |K_{i,j}X_{j,k}|, \\
&= \sum_{j=n}^{p} \left(|K_{i,j}| \sum_{k=1}^{p} |X_{j,k}| \right), \\
&= \sum_{j=1}^{n} |K_{i,j}|.
\end{aligned}$$

□

Corollary 4.3.2. *Let H, K be valid input matrices for Algorithm 3 and $\widetilde{K}$ be its output. Then, $\|\widetilde{K}\|_\infty \leq \|K\|_\infty$.*

Like banded probing, structured probing also preserves (strict) diagonal dominance, if a certain assumption is met.

Theorem 4.3.3. *Let H and K be valid input matrices for Algorithm 3 where $H_{i,i} = 1$ for $i = 1, \ldots, n$ and K is (strictly) diagonally dominant. Then $\widetilde{K}$, the output of Algorithm 3, is also (strictly) diagonally dominant.*

Proof: Diagonal dominance of K implies that for each row $i = 1, \ldots, n$,

$$|K_{i,i}| \;\geq\; \sum_{j\neq i} |K_{i,j}|.$$

Therefore,

$$|K_{i,i}| - \sum_{\substack{j\neq i \\ \phi(j)=\phi(i)}} |K_{i,j}| \quad \geq \quad \sum_{\phi(j)\neq\phi(i)} |K_{i,j}|. \tag{4.4}$$

But by (4.2) and (4.3),

$$\widetilde{K}_{i,i} = K_{i,i} + \sum_{\substack{j\neq i \\ \phi(j)=\phi(i)}} K_{i,j},$$

and by the triangle inequality,

$$|\widetilde{K}_{i,i}| \quad \geq \quad |K_{i,i}| - \sum_{\substack{j\neq i \\ \phi(j)=\phi(i)}} |K_{i,j}|. \tag{4.5}$$

By (4.4) and (4.5),

$$|\widetilde{K}_{i,i}| \quad \geq \quad \sum_{\phi(j)\neq\phi(i)} |K_{i,j}|. \tag{4.6}$$

Also note that by (4.2) and (4.3),

$$\sum_{\substack{j\neq i \\ H_{i,j}\neq 0}} |\widetilde{K}_{i,j}| \quad = \quad \sum_{\substack{j\neq i \\ H_{i,j}\neq 0}} \left| \sum_{\phi(k)=\phi(j)} K_{i,k} \right|. \tag{4.7}$$

Since $H_{i,i} \neq 0$, by (4.1) we know that $\phi(k) \neq \phi(i)$. Likewise, we know that no $|K_{i,k}|$ term is included twice. Thus,

$$\begin{aligned} \sum_{\substack{j\neq i \\ H_{i,j}\neq 0}} |\widetilde{K}_{i,j}| \quad &\leq \quad \sum_{\substack{j\neq i \\ H_{i,j}\neq 0}} \sum_{\phi(k)=\phi(j)} |K_{i,k}|, \\ &\leq \quad \sum_{\phi(j)\neq\phi(i)} |K_{i,j}|. \end{aligned} \tag{4.8}$$

Combining (4.6) and (4.8) as follows

$$|\widetilde{K}_{i,i}| \geq \sum_{\phi(j)\neq\phi(i)} |K_{i,j}| \geq \sum_{\substack{j \text{ s.t. } j\neq i \\ H_{i,j}\neq 0}} |\widetilde{K}_{i,j}| \geq \sum_{j} |\widetilde{K}_{i,j}|, \tag{4.9}$$

completes the proof. Note that the above inequalities can be replaced with strict inequalities if K is strictly diagonally dominant. □

4.4 Structured Probing for Schur Complements

The effectiveness of structured probing to approximate the Schur complement is based on the assumption that a reasonably accurate sparse approximation of the Schur complement exists. The sparsity pattern of this approximation, or some pattern similar to it, serves as the inspiration for the choice of sparsity pattern, H. Therefore, we consider the issue of the type of problems for which such an approximation exists.

Demko, Moss and Smith [21] show several results demonstrating bounds on the decay of the entries of inverses of banded matrices with distance from the diagonal. While their bounds are generally applicable, they do not necessarily demonstrate a decay that can be exploited to generate a sparse approximation to a matrix. For example, the decay constant may be bounded by a number greater than one, implying that the magnitude of entries can increase with distance from the diagonal. They show similar results for sparse (non-banded) positive definite matrices. In Section 4.4.1, we reformulate their results on sparse matrices in terms of distance on the adjacency graph of the matrix and extend them to indefinite matrices. We remark on the relevance of these results to Schur complements in Section 4.5.1.

4.4.1 Extending the Results of Demko, Moss and Smith

In their paper, Demko, Moss and Smith [21] present bounds on the entries of the inverses of banded matrices, both finite and infinite. They note first that for any positive definite operator A and real polynomial p,

$$\left\|A^{-1} - p(A)\right\| = \max_{x \in \lambda(A)} \left|\frac{1}{x} - p(x)\right|, \tag{4.10}$$

where $\lambda(A)$ represents the spectrum of A. They also recall a result of Meinardus [42, p. 33] on the accuracy of polynomial approximation of $1/x$. This result, repeated in Lemma 4.4.1, plays a key role in their analysis.

Lemma 4.4.1. *Let $f(x) = 1/x$, $a, b \in \mathbb{R}$ s.t. $0 < a < b$, $r = b/a$ and $q(r) = (\sqrt{r} - 1)/(\sqrt{r} + 1)$. Let π_n be the set of polynomials of degree at most n. Then,*

$$\inf_{p \in \pi_n} \|f - p\|_\infty = \frac{(1 + \sqrt{r})^2 q^{n+1}}{2ar}.$$

This key property leads Demko, et al. to derive bounds on the entries of the inverse of a banded matrix. Their results for positive definite banded matrices [21, Proposition 2.1, Theorem 2.4] are presented here in Theorem 4.4.2,

Theorem 4.4.2. *Let $A \in \mathbb{R}^{n \times n}$ be a p.d., banded matrix with bandwidth m. Let $[a, b]$ be the smallest interval containing $\lambda(A)$. Then,*

$$|A^{-1}_{i,j}| \leq \|A^{-1}\| \max\left(1, \frac{\left(1 + \sqrt{\kappa(A)}\right)^2}{2\kappa(A)}\right) \left(\frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1}\right)^{\frac{2|i-j|}{m}}.$$

Demko, et. al provide a similar result for indefinite banded matrices [21, Proposition 2.3, Theorem 2.4], which we repeat in Theorem 4.4.3

Theorem 4.4.3. *Let $A \in \mathbb{R}^{n \times n}$ be a banded matrix with bandwidth m.*

Let $[a, b]$ be the smallest interval containing $\lambda(AA^*)$. Then,

$$|A^{-1}_{i,j}| \leq (m+1)\|A^{-1}\| \max\left(1, \left(\frac{1+\kappa(A)}{\sqrt{2}\kappa(A)}\right)^2\right) \lambda_1^{|i-j|-m},$$

where

$$\lambda_1 = \left(\frac{\kappa(A)-1}{\kappa(A)+1}\right)^{\frac{1}{m}}.$$

These results show that the inverses of banded matrices have entries that decay with distance from the diagonal. Such decay is useful for the creation of sparse approximations if the matrix is well-conditioned. This result provides an intuitive foundation for the applicability of banded probing to certain Schur complement problems, like the 2-D non-overlapping domain decomposition problem studied by Chan and Mathew [14]. If the matrix F is banded, D is sparse and C and B^T are "local" (they do not combine columns/rows of F^{-1} that are far apart), then the Schur complement matrix $-(D - CF^{-1}B^T)$ will share the decay characteristics of the inverse of a banded matrix.

Demko, et. al go on to state a similar result that applies to the inverse of (non-banded) sparse matrices [21, Proposition 5.l]. They describe this property in terms of support and decay sets, but we have developed an alternative formulation in terms of graph distance.

Theorem 4.4.4. *Let $G = (V, E)$ be the adjacency graph of a positive definite sparse matrix $A \in \mathbb{R}^{n\times n}$ and let $\mathsf{dist}(\cdot,\cdot)$ be the shortest-path distance metric defined on G. Then,*

$$|A^{-1}_{i,j}| \leq \frac{\left(1+\sqrt{\kappa(A)}\right)^2}{2\|A\|} \left(\frac{\sqrt{\kappa(A)}-1}{\sqrt{\kappa(A)}+1}\right)^{\mathsf{dist}(v_i,v_j)+1}.$$

Proof: Demko, et. al [21, Proposition 5.1] define

$$\begin{aligned}
S_n(A) &= \bigcup_{k=0}^{n} \Big\{ (i,j) \text{ s.t. } A^k(i,j) \neq 0 \Big\}, \\
D_n(A) &= (\{1,\ldots,n\} \times \{1,\ldots,n\}) \setminus S_n(A),
\end{aligned}$$

and prove that

$$\sup \Big\{ |A^{-1}_{i,j}| \ : \ (i,j) \in D_n(A) \Big\} \leq \frac{\left(1+\sqrt{\kappa(A)}\right)^2}{2\|A\|} \left(\frac{\sqrt{\kappa(A)}-1}{\sqrt{\kappa(A)}+1} \right)^{n+1}. \tag{4.11}$$

Let $G_n(V, E_n)$ be the adjacency graph of A^n. Then, $E_n \subset \{(v_i, N_n(v_i))\}$, the set of all paths between distance-n neighbors in G. Therefore,

$$D_n(A) = \{(i,j) \notin E_n\}.$$

Thus,

$$\{(i,j) \ : \ \mathsf{dist}(v_i, v_j) > n\} \subset D_n(A). \tag{4.12}$$

Combining (4.11) with (4.12) and noting that the bound can be applied element-wise completes the proof. □

We have also developed an extension of Theorem 4.4.4 to the indefinite case.

Theorem 4.4.5. *Let $A \in \mathbb{R}^{n\times n}$ be a non-singular matrix with at most b non-zero entries per column. Let $G = (V, E)$ be the adjacency graph of matrix AA^T. and let $\mathsf{dist}(\cdot,\cdot)$ be the shortest-path distance metric defined on G. Then,*

$$|A^{-1}_{i,j}| \leq \frac{b\|A\|_1 \left(1+\kappa(A)\right)^2}{2\|A\|} \left(\frac{\kappa(A)-1}{\kappa(A)+1} \right)^{\mathsf{dist}(v_i,v_j)}.$$

Proof: Take $i, j \in 1, \ldots, n$. Note that $A^{-1} = A^T(AA^T)^{-1}$. Thus,

$$\begin{aligned}
|A^{-1}_{i,j}| &= \left| \sum_{A^T_{i,k} \neq 0} A^T_{i,k} (AA^T)^{-1}_{k,j} \right|, \\
&\leq \sum_{A^T_{i,k} \neq 0} |A^T_{i,k}| \left| (AA^T)^{-1}_{k,j} \right|, \\
&\leq \|A\|_1 \sum_{A^T_{i,k} \neq 0} \left| (AA^T)^{-1}_{k,j} \right|. \qquad (4.13)
\end{aligned}$$

As (AA^T) is s.p.d., we can apply Theorem 4.4.4 to the right hand side of (4.13), yielding

$$|A^{-1}_{i,j}| \leq \frac{\|A\|_1 (1 + \kappa(A))^2}{2\|A\|} \sum_{A_{k,i} \neq 0} \left(\frac{\kappa(A) - 1}{\kappa(A) + 1} \right)^{\mathsf{dist}(v_i, v_k) + 1}. \qquad (4.14)$$

Since the summands are all less than one, any v_k with the smallest value of $\mathsf{dist}(v_i, v_k)$ will serve as an upper bound for all the summands. As v_k is a neighbor of v_j, $\mathsf{dist}(v_i, v_j) - 1 \leq \mathsf{dist}(v_i, v_k)$. Since A has at most b non-zeros per column, there are at most b entries in the sum. Thus,

$$|A^{-1}_{i,j}| \leq \frac{b\|A\|_1 (1 + \kappa(A))^2}{2\|A\|} \left(\frac{\kappa(A) - 1}{\kappa(A) + 1} \right)^{\mathsf{dist}(v_i, v_j)}.$$

□

4.5 Accuracy of Approximations by Structured Probing

Algorithm 3 for structured probing generates an approximation $\widetilde{K}$, given a matrix K and a sparsity pattern H. Assume that H is chosen such that $H_{i,j} = 0$ if $K_{i,j} = 0$. We split K such that

$$K = \bar{K} + \widehat{K}, \qquad (4.15)$$

where $H_{i,j} \neq 0 \Leftrightarrow \bar{K}_{i,j} = K_{i,j}$ and $\widehat{K}_{i,j} = 0$. This means that H and $\bar{K}$ share sparsity patterns and that H and $\widehat{K}$ have sparsity patterns that are disjoint. The matrix $\bar{K}$ is what the structured probing algorithm would output if $\widehat{K} = 0$.

We also assume that H is chosen such that $\widehat{K}$ has no large entries. Specifically, we assume that there exists some $\epsilon_1 > 0$ such that

$$|\widehat{K}_{i,j}| \leq \epsilon_1. \tag{4.16}$$

Define the error introduced by structured probing to be

$$R = K - \widetilde{K}. \tag{4.17}$$

Define the error in the approximation of $\bar{K}$ by $\widetilde{K}$ to be

$$L = \widetilde{K} - \bar{K}. \tag{4.18}$$

As $\widetilde{K}$ and $\bar{K}$ have sparsity patterns that are disjoint with that of $\widehat{K}$, so does L. Finally, we define

$$c_k = \sum_{i=1}^{n} X_{i,k}, \tag{4.19}$$

which is the number of nodes colored by color k. This allows us to state a bound on the entries of R.

Theorem 4.5.1. *Let H and K be valid input matrices for Algorithm 3. Let R be defined as in (4.17), ϵ_1 defined as in (4.16) Then,*

$$|R_{i,j}| \leq \begin{cases} \epsilon_1, & \text{if } H_{i,j} = 0, \\ (c_{\phi(j)} - 1)\epsilon_1, & \text{if } H_{i,j} \neq 0. \end{cases}$$

Proof: Fix $i, j \in \{1, \ldots, n\}$. Note that by (4.15)–(4.18), $R = \widehat{K} - L$ and that $\widehat{K}$ and L have disjoint sparsity patterns. Thus, if $H_{i,j} = 0$, then

$R_{i,j} = \widehat{K}_{i,j}$ and $|R_{i,j}| \leq \epsilon_1$.

For the $H_{i,j} \neq 0$ case, we have by (4.2), (4.3) and (4.15), that

$$\widetilde{K}_{i,j} = K_{i,j} + \sum_{\substack{k \neq j \\ \phi(k)=\phi(j)}} \widehat{K}_{i,k}. \tag{4.20}$$

By (4.18), (4.19), and (4.20), we have

$$\begin{aligned} L_{i,j} &= \sum_{\substack{k \neq j \\ \phi(k)=\phi(j)}} \widehat{K}_{i,k}, \\ |L_{i,j}| &\leq \sum_{\substack{k \neq j \\ \phi(k)=\phi(j)}} |\widehat{K}_{i,k}|, \\ &\leq \epsilon_1 \sum_{k \neq j} |X_{\phi(j),k}|, \\ &\leq \epsilon_1 (c_{\phi(j)} - 1). \end{aligned}$$

□

Theorem 4.5.1 then allows us to state bounds on particular norms of R.

Lemma 4.5.2. *Let H and K be valid input matrices for Algorithm 3. Let R be defined as in (4.17), ϵ_1 defined as in (4.16) and b_H by the minimum number of non-zeros per row in H and b_K be the maximum number of non-zeros per row in K. Then,*

$$\begin{aligned} \|R\|_1 &\leq n \max_k (1, c_k - 1)\epsilon_1, && (4.21) \\ \|R\|_\infty &\leq 2(b_K - b_H)\epsilon_1. && (4.22) \end{aligned}$$

Proof: The bound on $\|R\|_1$ follows directly from Theorem 4.5.1. For the bound on $\|R\|_\infty$, recall that

$$|R_{i,j}| = \begin{cases} |K_{i,j}|, & \text{if } H_{i,j} = 0, \\ \sum\limits_{\substack{k \neq j \\ \phi(k)=\phi(j)}} \widehat{K}_{i,k}, & \text{if } H_{i,j} \neq 0. \end{cases} \tag{4.23}$$

Then,

$$\sum_{j=1}^{n} |R_{i,j}| = \sum_{\substack{j \\ H_{i,j}\neq 0}} \left(\sum_{\substack{k\neq j \\ \phi(k)=\phi(j)}} |K_{i,k}| \right) + \sum_{\substack{j \\ H_{i,j}=0}} |K_{i,k}|.$$

By (4.1) the first sum will never include the same term more than once. Let H have b_i non-zero entries in row i, then we have

$$\begin{aligned} \sum_{j=1}^{n} |R_{i,j}| &\leq \; = 2 \sum_{\substack{j \\ H_{i,j}\neq 0}} |K_{i,j}|, \\ &\leq \; 2(b_K - b_i)\epsilon_1 \end{aligned}$$

Taking the maximum such row sum completes the lemma. □

Note that this bound can be improved if there exists A, such that $K = A^{-1}$, and that Theorem 4.4.4 or Theorem 4.4.5 gives useful decay. While Lemma 4.5.3 does not require $q < 1$, the entries of R will only be small if that is the case.

Lemma 4.5.3. *Let H and K be valid input matrices for Algorithm 3. Let $G = (V, E)$ be the adjacency graph of K and let the adjacency graph of H be a subgraph of G. Assume that K satisfies $|K_{i,j}| \leq cq^{\mathsf{dist}(v_i,v_j)}$, for some $c, q > 0$ and where $\mathsf{dist}(\cdot,\cdot)$ is the shortest-path distance metric on G. Let R be defined as in (4.17). Then,*

$$|R_{i,j}| \leq \begin{cases} cq^{\mathsf{dist}(v_i,v_j)}, & \textit{if } H_{i,j} = 0, \\ c\sum_{\substack{k\neq j \\ \phi(k)=\phi(j)}} q^{\mathsf{dist}(v_i,v_k)}, & \textit{if } H_{i,j} \neq 0. \end{cases}$$

Proof: Substitute $|K_{i,j}| \leq cq^{\mathsf{dist}(v_i,v_j)}$ into (4.23). □

We can also restate Lemma 4.5.3 in terms of the maximum vertex degree of G, denoted as $\Delta(G)$.

Lemma 4.5.4. *Let H and K be valid input matrices for Algorithm 3. Let $G = (V, E)$ be the adjacency graph of K, with maximal degree $\Delta(G)$ and* *diameter* $\mathsf{diam}(G)$*, and let the adjacency graph of* H *be a subgraph of* G*. Assume that* K *satisfies* $|K_{i,j}| \leq cq^{\mathsf{dist}(v_i,v_j)}$*, for some* $c, q > 0$ *and where* $\mathsf{dist}(\cdot,\cdot)$ *is the shortest-path distance metric defined on* G*. Let* R *be defined as in (4.17). Then, if* $\Delta(G)q \neq 1$*,*

$$|R_{i,j}| \leq \begin{cases} cq^{\mathsf{dist}(v_i,v_j)}, & \text{if } H_{i,j} = 0, \\ c\frac{(\Delta(G)q)^{\mathsf{diam}(G)} - (\Delta(G)q)^3}{\Delta(G)q - 1}, & \text{if } H_{i,j} \neq 0. \end{cases}$$

Proof: Since v_i has at most $\Delta(G)$ neighbors, Lemma 4.5.3 tells us that

$$|R_{i,j}| \leq c \sum_{k=3}^{\mathsf{diam}(G)} (\Delta(G)q)^k,$$

because no distance-1 or distance-2 neighbors of v_i will have contributions to $\widehat{K}_{i,j}$. Summing the geometric series completes the proof. □

If $\Delta(G)q < 1$, then the entries of R corresponding to non-zero entries of H are also small. This particular bound is not dependent on the particular graph coloring used and is only applicable to problems with rapid decay.

4.5.1 Analysis for Schur Complements

In order for the Schur complement matrix $S_1 = -(D - CF^{-1}B^T)$ to be well-approximated by structured probing, we must first have a useful decay in the entries of F^{-1}. This means that entries of F^{-1} corresponding to points which are far away from each other on the adjacency graph of F must be small. Condition 4.5.5 expresses this rigorously.

Condition 4.5.5. *Let* $F \in \mathbb{R}^{n \times n}$*. Let* $G = (V, E)$ *be the adjacency graph of* F *and let* $\mathsf{dist}(\cdot,\cdot)$ *be the shortest path distance metric on* G*. Assume that there exists* $\epsilon > 0$ *and* $d_f \in \mathbb{Z}^+$ *such that*

$$\forall k, l \in \{1, \ldots, n\}, \text{ s.t. } \mathsf{dist}(v_k, v_l) \geq d_f, \qquad |F^{-1}_{k,l}| \leq \epsilon.$$

If the entries of F^{-1} decay with distance on the adjacency graph of F, the exact Schur complement $S_1 = -(D - CF^{-1}B^T)$ should also have a decay property, assuming B^T and C are "local" (e.g. they only combine rows/columns of F^{-1} corresponding to nearby points on the graph of F) and D is also sparse. This idea of locality is expressed rigorously in Condition 4.5.6.

Condition 4.5.6. *Let $F \in \mathbb{R}^{n\times n}$, and $B, C \in \mathbb{R}^{n\times m}$. Let $G = (V, E)$ be the adjacency graph of F and let $\mathsf{dist}(\cdot,\cdot)$ be the shortest path distance metric on G. For a given $i, j \in \{1,\dots,m\}$, assume that $\exists k_0, l_0 \in \{1,\dots,n\}$ such that $(B^T)_{l_0,j} \neq 0$ and $C_{i,k_0} \neq 0$. Assume also that $\exists d_b, d_c \in \mathbb{Z}^+$ such that*

$$\begin{aligned} \forall l \in \{1,\dots,n\} \text{ s.t. } (B^T)_{l,j} \neq 0, \qquad & \mathsf{dist}(v_l, v_{l_0}) < d_b, \\ \forall k \in \{1,\dots,n\} \text{ s.t. } C_{i,k} \neq 0, \qquad & \mathsf{dist}(v_{k_0}, v_k) < d_c. \end{aligned}$$

Matrices that satisfy these two conditions lead to Schur complements that have useful decay. Theorem 4.5.7 expresses this more rigorously.

Theorem 4.5.7. *Let $F \in \mathbb{R}^{n\times n}$ satisfy Condition 4.5.5 and $B, C \in \mathbb{R}^{n\times m}$ satisfy Condition 4.5.6. Find d_f as per Condition 4.5.5. Given $i, j \in \{1,\dots,m\}$, find k_0, l_0, d_b, d_c as per Condition 4.5.6. If $\mathsf{dist}(v_{k_0}, v_{l_0}) \geq d_f + d_b + d_c$, then*

$$|(CF^{-1}B^T)_{i,j}| \quad \leq \quad \epsilon \|(B^T)_{\cdot,j}\|_1 \|C_{i,\cdot}\|_1,$$

where $\|(B^T)_{\cdot,j}\|_1$ and $\|C_{i,\cdot}\|_1$ represent the 1-norm of the jth column and ith row of B^T and C respectively.

Proof: Given $i, j \in \{1,\dots,m\}$, assume $\exists k_0, l_0 \in \{1,\dots,n\}$ and $\exists d_f, d_b, d_c \in$

$\mathbb{Z}^+$, such that $\mathsf{dist}(v_{k_0}, v_{l_0}) \geq d_f + d_b + d_c$. Then,

$$\begin{aligned} (CF^{-1}B^T)_{i,j} &= \sum_{k=1}^{n} C_{i,k} \sum_{l=1}^{n} F^{-1}_{k,l} (B^T)_{l,j}, \\ &= \sum_{\substack{k \\ \mathsf{dist}(v_{k_0}, v_k) < d_c}} C_{i,k} \sum_{\substack{l \\ \mathsf{dist}(v_l, v_{l_0}) < d_b}} F^{-1}_{k,l} (B^T)_{l,j}. \end{aligned} \tag{4.24}$$

By the triangle inequality,

$$\mathsf{dist}(v_k, v_l) \geq \mathsf{dist}(v_{k_0}, v_{l_0}) - \mathsf{dist}(v_{k_0}, v_k) - \mathsf{dist}(v_l, v_{l_0}). \tag{4.25}$$

But $\mathsf{dist}(v_{k_0}, v_{l_0}) \geq d_f + d_b + d_c$, $\mathsf{dist}(v_{k_0}, v_k) < d_c$ and $\mathsf{dist}(v_l, v_{l_0}) < d_b$, thus $\mathsf{dist}(v_k, v_l) \geq d_f$. So, for each term in the sum in (4.24), $|F^{-1}_{k,l}| \leq \epsilon$. So,

$$\begin{aligned} |(CF^{-1}B^T)_{i,j}| &\leq \epsilon \sum_{\substack{k \\ \mathsf{dist}(v_{k_0}, v_k) < d_c}} |C_{i,k}| \sum_{\substack{l \\ \mathsf{dist}(v_l, v_{l_0}) < d_b}} |(B^T)_{l,j}|, \\ &= \epsilon \|B^T_{\cdot,j}\|_1 |C_{i,\cdot}\|_1. \end{aligned}$$

□

Several other observations follow immediately from Theorem 4.5.7.

Lemma 4.5.8. *Let $F \in \mathbb{R}^{n \times n}$ satisfy Condition 4.5.5 and $B, C \in \mathbb{R}^{n \times m}$ satisfy Condition 4.5.6. Find d_f as per Condition 4.5.5. Given $i, j \in \{1, \ldots, m\}$, find k_0, l_0, d_b, d_c as per Condition 4.5.6. If $\mathsf{dist}(v_{k_0}, v_{l_0}) \geq d_f + d_b + d_c$, then*

1. $| - (D - CF^{-1}B^T)_{i,j}| \leq |D_{i,j}| + \epsilon \|(B^T)_{\cdot,j}\|_1 |C_{i,\cdot}\|_1$,
2. $|(CF^{-1}B^T)_{i,j}| \leq \epsilon \|B^T\|_1 \|C\|_\infty$,
3. $| - (D - CF^{-1}B^T)_{i,j}| \leq |D_{i,j}| + \epsilon \|B^T\|_1 \|C\|_\infty$.

Proof: All three properties follow immediately from Theorem 4.5.7 and the definition of the 1 and ∞ norms. □

We can also restate these theorems such that they do not explicitly invoke the locality of B^T and C. This result is somewhat less elegant, but can make the above bounds tighter. To do this we choose an optimal k_0 and l_0 in Condition 4.5.6 rather than choosing an arbitrary one.

Theorem 4.5.9. *Let $F \in \mathbb{R}^{n\times n}$ satisfy Condition 4.5.5. Find d_f as per Condition 4.5.5. Let $B, C \in \mathbb{R}^{n\times m}$. For a fixed $i, j \in \{1, \ldots, m\}$, choose $k_0, l_0 \in \{1, \ldots, n\}$ such that $B^T_{l_0,j} \neq 0$ and $C_{i,k_0} \neq 0$ and that $\mathsf{dist}(v_{k_0}, v_{l_0})$ is minimized. If $\mathsf{dist}(v_{k_0}, v_{l_0}) \geq d_f$, then*

$$|(CF^{-1}B^T)_{i,j}| \quad \leq \quad \epsilon\|(B^T)_{\cdot,j}\|_1\|C_{i,\cdot}\|_1,$$

where $\|(B^T)_{\cdot,j}\|_1$ and $\|C_{i,\cdot}\|_1$ represent the 1-norm of the jth column and ith row of B^T and C respectively.

Proof: Given $i, j \in \{1, \ldots, m\}$, choose $\exists k_0, l_0 \in \{1, \ldots, n\}$ such that $\mathsf{dist}(v_{k_0}, v_{l_0})$ is minimized. If $\mathsf{dist}(v_{k_0}, v_{l_0}) \geq d_f$,

$$(CF^{-1}B^T)_{i,j} \quad = \quad \sum_{k=1}^{n} C_{i,k} \sum_{l=1}^{n} F^{-1}_{k,l}(B^T)_{l,j}. \tag{4.26}$$

Since $\mathsf{dist}(v_{k_0}, v_{l_0})$ is a minimum for all (k_0, l_0) such that $(B^T)_{l_0,j} \neq 0$ and $C_{i,k_0} \neq 0$, we have $\mathsf{dist}(v_k, v_l) \geq \mathsf{dist}(v_{k_0}, v_{l_0})$, for all k and l in the sum in (4.26). Therefore, $|F^{-1}_{k,l}| \leq \epsilon$ for each term in the sum in (4.26). So,

$$\begin{aligned} |(CF^{-1}B^T)_{i,j}| \quad &\leq \quad \epsilon \sum_{k=1}^{n} |C_{i,k}| \sum_{l=1}^{n} |(B^T)_{l,j}|, \\ &= \quad \epsilon\|(B^T)_{\cdot,j}\|_1|C_{i,\cdot}\|_1. \end{aligned}$$

□

We can also show similar results to Lemma 4.5.8.

Lemma 4.5.10. *Let $F \in \mathbb{R}^{n\times n}$ satisfy Condition 4.5.5. Find d_f as per Condition 4.5.5. Let $B, C \in \mathbb{R}^{n\times m}$ and $D \in \mathbb{R}^{m\times m}$. For a fixed $i, j \in$*

$\{1, \ldots, m\}$, *choose* $k_0, l_0 \in \{1, \ldots, n\}$ *such that* $(B^T)_{l_0,j} \neq 0$ *and* $C_{i,k_0} \neq 0$ *and that* $\mathsf{dist}(v_{k_0}, v_{l_0})$ *is minimized. If* $\mathsf{dist}(v_{k_0}, v_{l_0}) \geq d_f$*, then*

1. $| - (D - CF^{-1}B^T)_{i,j}| \leq |D_{i,j}| + \epsilon \|(B^T)_{\cdot,j}\|_1 |C_{i,\cdot}\|_1,$

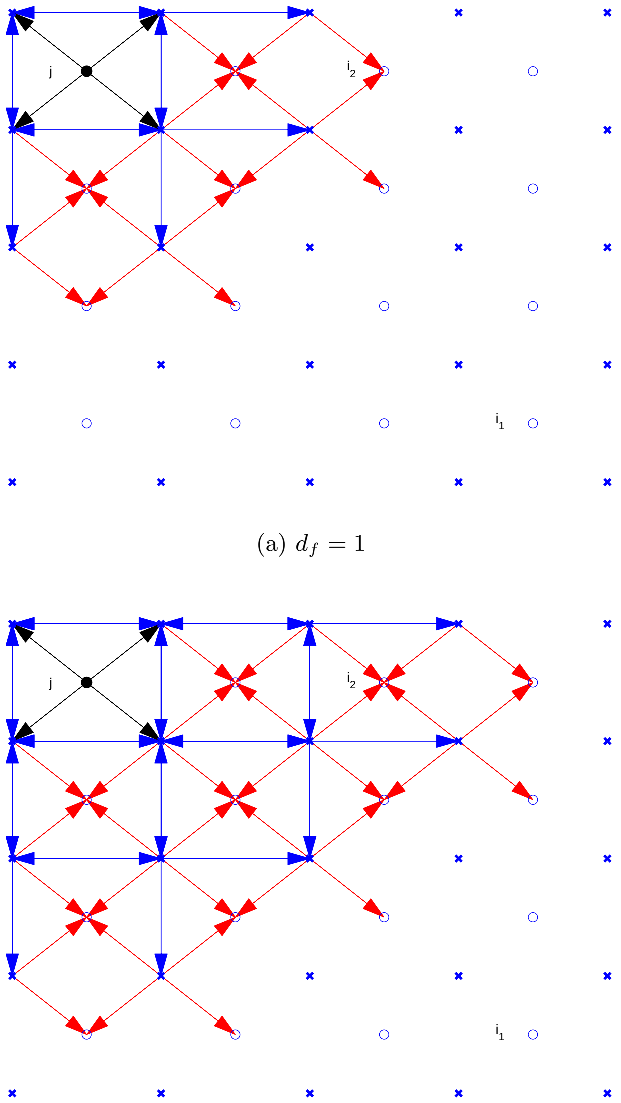
2. $|(CF^{-1}B^T)_{i,j}| \leq \epsilon \|B^T\|_1 \|C\|_\infty,$

3. $| - (D - CF^{-1}B^T)_{i,j}| \leq |D_{i,j}| + \epsilon \|B^T\|_1 \|C\|_\infty.$

Proof: All three properties follow immediately from Theorem 4.5.9 and the definition of the 1 and ∞ norms. □

Figures 4.2(a) and 4.2(b) show a graphical interpretation of Theorem 4.5.9 for a problem using a staggered finite difference mesh. To simplify visualization, we assume that $D = 0$.

The x-variables in (1.2) are represented by blue x's and the y-variables are represented by blue circles. We consider the y-variable corresponding to node j, and a column of $CF^{-1}B^T$, and represent is as a black node in Figure 4.2. Black arrows point to the x-variables, l, such that $(B^T)_{l,j} \neq 0$. These are the x-variables "affected" by the y-variable j. Blue arrows show the effect of the large entries of F^{-1}, pointing to the x-variables k, such that $|(F^{-1}B^T)_{k,j}|$ is large. Here we assume that all entries that are further than one (Figure 4.2(a)) or two 4.2(b) links away on the adjacency graph of F are small. Red arrows show the effect of C, pointing the the y-variables, i, such that $|(CF^{-1}B^T)_{i,j}|$ is large. Note that point i_1 is "far" from j on the graph of CFB^T , and thus $|(CF^{-1}B^T)_{i_1,j}|$ is small. Point i_2 is "close" to j and therefore $|(CF^{-1}B^T)_{i_2,j}|$ may be large. In a situation like this, we can derive appropriate stencils for the Schur complement over the finite element or finite difference mesh as the sparsity pattern for H. This is the approach we take in Section 6.



(a) $d_f = 1$

(b) $d_f = 2$

Figure 4.2. Entries of the Schur complement affected by node j through B^T (black arrows), $F^{-1}B^T$, assuming that the entries of F^{-1} decay after one edge ($d_f = 1$) or two edges ($d_f = 2$) on the adjacency graph (blue arrows) and $CF^{-1}B^T$ (red arrows), on a staggered finite difference grid. Note that $|(CF^{-1}B^T)_{i_1,j}|$ will be small and $|(CF^{-1}B^T)_{i_2,j}|$ may be large.

4.6 Structured Probing, Approximate Factorizations and Inverses

When probing is used to approximate narrowly banded matrices, factoring the resulting approximation is generally inexpensive as little fill is created in the factors. Structured probing, on the other hand, allows for much more complicated sparsity patterns, which may result in a sizable amount of fill-in during the factorization process. However, since the result of the probing process is an explicitly stored sparse matrix, many of the standard sparse linear algebra techniques are available to solve this problem.

One approach is to use an ILU or ILUT factorization [49] as an alternative to an exact factorization of the matrix resulting from structured probing. Results in Sections 6.1.4 and 6.3 will illustrate the effectiveness of this approach. Sparse approximate inverse techniques, like AINV [6, 7] or SPAI [34] are also promising candidates for decreasing the computational cost of applying the inverse of the matrix resulting from structured probing.

5 Applications

Generalized saddle-point problems arise from a multitude of different applications; therefore we consider applications from several sources. This chapter describes the applications for which experimental data is presented in Chapter 6. Our first application describes fluid flow in a lid-driven cavity. This problem, described in Section 5.1, is modeled by the incompressible Navier-Stokes equations in two dimensions and is discretized with finite elements. Our second application, described in Section 5.2, also describes 2-D fluid flow, this time modeled by the incompressible Stokes equations, but is discretized using a spectral collocation approach. The third application involves stress relaxation of thin strips of metal. This application, which uses a modified Hart's model [30] and is described in Section 5.3, is a three-dimensional problem discretized using finite elements. Our final application arises from an optimization problem. Here we consider the flattening of a three-dimensional surface mesh into a two-dimensional planar mesh. This application is described in Section 5.4.

5.1 Navier-Stokes: Lid-Driven Cavity

The Navier-Stokes equations for incompressible flow are given by

$$\begin{aligned} \frac{\partial \mathbf{u}}{\partial t} + \mathbf{u} \cdot \nabla \mathbf{u} - \nu \Delta \mathbf{u} + \nabla p &= 0, \\ \nabla \cdot \mathbf{u} = 0, & \end{aligned} \tag{5.1}$$

on some region Ω. Here $\mathbf{u}$ is the velocity field, p is the pressure and $\nu > 0$ is the viscosity constant (the inverse of the Reynolds number in a nondimensional setting).

The description of our particular application, a 2-D leaky lid-driven cavity discretized with finite elements, follows Elman, Silvester and Wathen [26]. The implementation is from the associated MATLAB software. Specifically, we consider the steady-state Oseen equations, [45, Section 21.13],

$$\begin{aligned} \mathbf{w} \cdot \nabla \mathbf{u} - \nu \Delta \mathbf{u} + \nabla p &= 0, \\ \nabla \cdot \mathbf{u} = 0, & \end{aligned} \tag{5.2}$$

on $\Omega = [0,\ 1]^2$, with $\mathbf{u}$, p, and ν defined as above. For the wind function, $\mathbf{w}$, we choose the so-called “divergence-free vortex,”

$$\mathbf{w} = \begin{bmatrix} 2y(1-x^2) \\ -2x(1-y^2) \end{bmatrix}. \tag{5.3}$$

We discretize the space using mixed $Q1 - P0$ finite elements, that is, a piecewise linear approximation for the velocity and a piecewise constant approximation for pressure. The linear system resulting from the finite element discretization of (5.2) is of the form

$$\begin{bmatrix} A & B^T \\ B & D \end{bmatrix} \begin{bmatrix} u \\ p \end{bmatrix} = \begin{bmatrix} f \\ 0 \end{bmatrix}, \tag{5.4}$$

with $D = -\beta\widehat{D}$, where $\beta > 0$ is the stabilization parameter, and $\widehat{D}$ is a positive semidefinite stabilization matrix. This application satisfies conditions C3, C4 and C6 from Chapter 1. We set the y-component of the velocity to zero on all four boundaries and the x-component of the velocity to zero on all boundaries except where $y = 1$, where it is set to one. Figure 5.1 shows the boundary conditions for our experiments. Figure 5.2 shows the finite element mesh for the case of the 16×16 grid. Velocity unknowns are shown in blue, as are connections between the velocity unknowns. Pressure unknowns and pressure-pressure connectivity (due to stabilization) are shown in red. Velocity-pressure connectivity is shown in green.

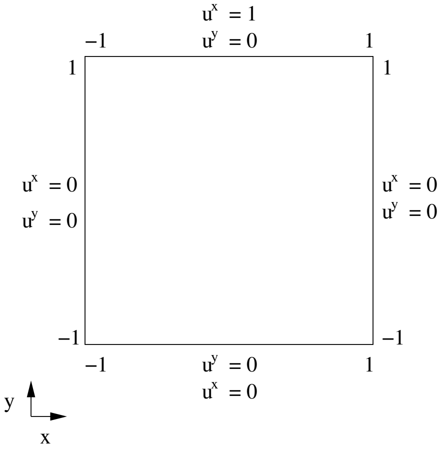


Figure 5.1. Boundary conditions for the 2-D lid-driven cavity problem, modeled using the steady-state Navier-Stokes equations (5.1).

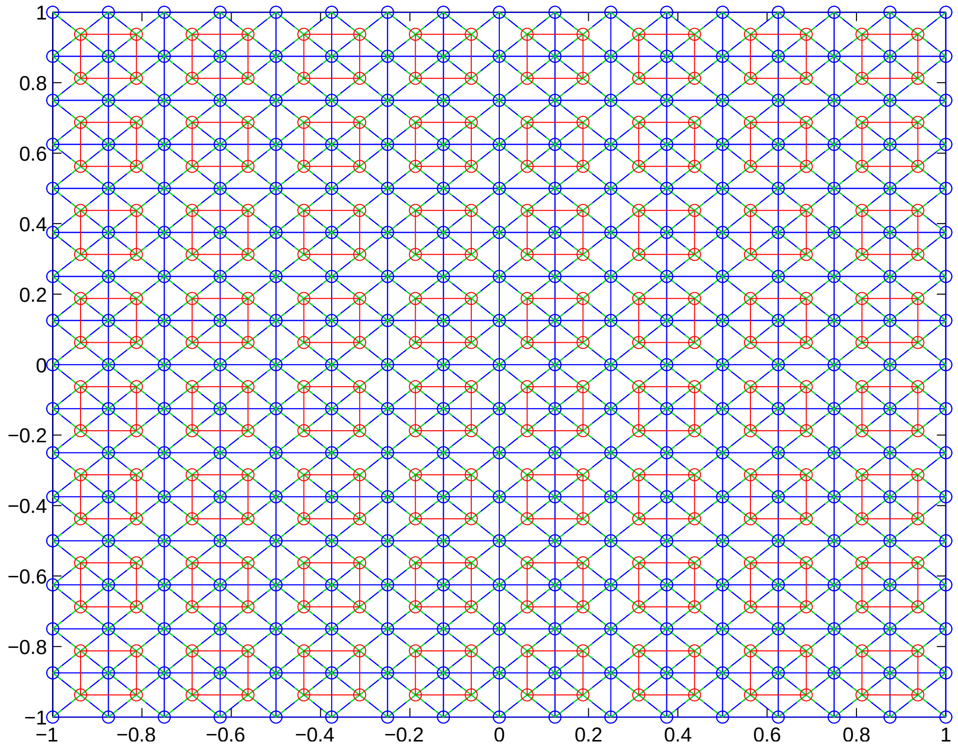
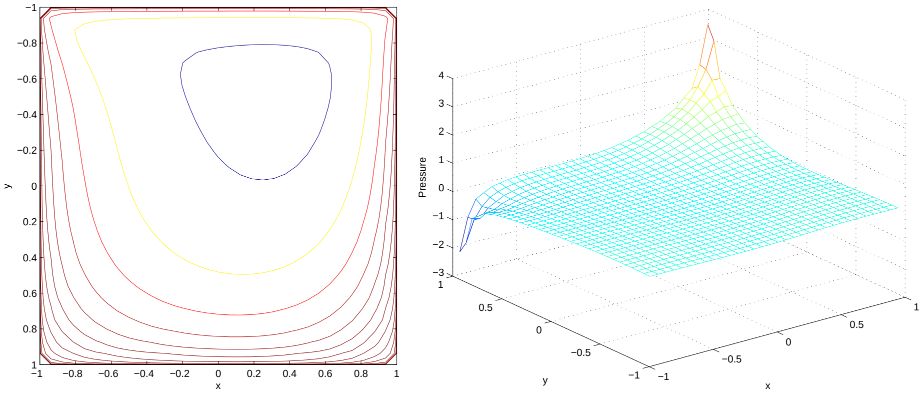


Figure 5.2. Finite element mesh showing velocity-velocity (blue), pressure-pressure (red) and velocity-pressure (green) connectivity, for the 2-D lid-driven cavity problem, modeled using the steady-state Navier-Stokes equations (5.1).

Figures 5.3(a) and 5.3(b) show some of the streamlines and the pressure field, respectively, of the solution of (5.4) on a 32×32 grid with viscosity parameter $\nu = .1$ and stabilization parameter $\beta = .25$. These choices represent the typical values of ν and β that we employ in Section 6.1.



(a) Velocity Streamlines

(b) Pressure Values

Figure 5.3. Solution of the 2-D lid-driven cavity problem, modeled using the steady-state Navier-Stokes equations, with $\nu = .1$ and $\beta = .25$.

5.2 Incompressible Stokes: Spectral Collocation

The incompressible Stokes equations are given by

$$\begin{aligned} -\nu\Delta\mathbf{u} + \nabla p &= \mathbf{f}, \\ \nabla \cdot \mathbf{u} = 0, & \end{aligned} \tag{5.5}$$

on some region Ω. We consider the use of the spectral collocation approach described by Bernardi, Canuto and Maday [8] on $\Omega = (-1,1)^2$. This involves using the Chebyshev nodes associated with Gauss-Lobatto quadrature as our collocation sites and assuming homogeneous Dirchlet boundary conditions in velocity space. This generates linear systems of the form

$$\begin{bmatrix} A & B^T \\ C & 0 \end{bmatrix} \begin{bmatrix} u \\ p \end{bmatrix} = \begin{bmatrix} f \\ 0 \end{bmatrix}. \tag{5.6}$$

In the nomenclature of Chapter 1, this problem satisfies C5 (and thus C4 and C6). Notably, the spectral collocation method creates a structurally asymmetric system where $B \neq C$. This can be seen from the weak form of the system,

$$\begin{aligned} a(\mathbf{u},\mathbf{v}) + b_1(\mathbf{v},p) &= <\mathbf{f},\mathbf{v}>, & (5.7)\\ b_2(\mathbf{u},q) &= 0, & (5.8) \end{aligned}$$

where

$$\begin{aligned} a(\mathbf{u},\mathbf{v}) &= \nu \int_\Omega \nabla\mathbf{u}\cdot\nabla(\mathbf{v}\omega)\, d\Omega, & (5.9)\\ b_1(\mathbf{v},p) &= -\int_\Omega (\nabla\cdot(\mathbf{v}\omega))\, p\, d\Omega, & (5.10)\\ b_2(\mathbf{u},q) &= -\int_\Omega (\nabla\cdot\mathbf{u})\, q\omega\, d\Omega, & (5.11) \end{aligned}$$

and ω is the Chebyshev weight function. The location of this weighting function in (5.9)–(5.11) produces the asymmetry in A and the asymmetry between C and B^T.

Our particular MATLAB implementation of the 2-D incompressible Stokes problem follows the algorithm presented by Bernardi, et al. [8, Section 5.1]. Results for this problem are described in Section 6.2.

5.3 Metal Deformation: Stress Relaxation in Thin Strips

We consider the simulation of in-service stress relaxation of formed sheet metal components. Stress relaxation is the time-dependent deformation of a material under constant strain. When this relaxation is "in-service," this deformation occurs while the particular formed material is in use. The motivation for this particular application comes from the stress relaxation of formed sheet metal parts, such as beverage can ends. In the case of the can end, in-service relaxation leads to a decrease in can end buckle pressure. As buckle strength is a critical property for can ends, accurate modeling (and simulation) is important.

These simulations complement the numerical work of Zhu, Beaudoin and MacEwen [64], and the description of this particular application follows the work of Zhu [62]. Our simulation code, which uses a finite element formulation, was implemented in FORTRAN 90 by Lihua Zhu, the first author of the aforementioned paper.

Following Zhu et al. [64], we consider a simple bent beam relaxation test. This produces deformations similar to that of can end formation. Specifically, we consider a 88.9 × 12.7 × 0.236 mm beam of AA5812-H19, an aluminum-magnesium alloy. The simulated test includes three phases:

1. Loading: The strip is bent around a pipe.
2. Holding: The strip is clamped around the pipe.
3. Unloading: The strip is removed from the pipe, undergoing springback.

A modified Hart's model [30] (derived from Hart's model [36]), provides the governing equations for the system. Figure 5.4 shows a pictorial representation of the constitutive equations of modified Hart's model in the form of a rheological diagram. The upper branch of Figure 5.4 represents the macro-scale effects: elastic and anelastic deformation, which represented by the spring, as well as plasticity, which represented by the dotted line. In this regime, we consider anelastic (a) and plastic (ε_p) strain. The lower branch of Figure 5.4 represents the micro-scale effects, namely, micro-anelasticity and micro-plasticity. Micro-anelastic strain (a_t) is represented by the spring, and micro-plastic strain (ε_{pt}) is represented by the plastic dashpot. These branches give us both micro (σ_t) and macro (σ_a) components of stress and when combined, yield the total stress (σ) and inelastic strain (σ_i).

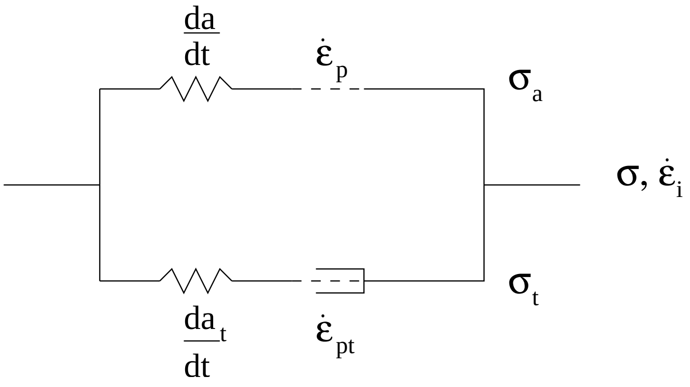


Figure 5.4. Rheological diagram of the modified Hart's model[30]

The finite element discretization is described in detail in [62, Chapter 5]. Each element generates element stiffness matrices of the form

$$\begin{bmatrix} (K_E^{-1} + K_P^{-1})^{-1} & B^T \\ B & D \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix}, \tag{5.12}$$

where the (1,1) block is positive definite and has both plastic (K_P) and elastic (K_E) components, and the (2,2) block results from a slight compressibility. The y-variables correspond to nodes in the center of each element, allowing us to use probing techniques on the element-element connectivity graph. In the language of Chapter 1, this problem satisfies conditions C1 (and thus C2), C3 and C6. Results for this problem are described in Section 6.3.

5.4 Optimization: Mesh Flattening

Computing optimal planar triangulations of 3-D surface meshes, also known as mesh flattening, has been a topic of recent interest in the computer graphics and meshing communities. These algorithms are often used for applications such as surface parameterization and remeshing. Sheffer and de Sturler [52] formulate this problem in terms of a constrained optimization problem, with the goal of minimizing the relative angular deformation. Their algorithm minimizes the relative distortion in the angles on the 2-D mesh from their 3-D counterparts, after normalization for non-planarity, subject to constraints that enforce the validity of the 2-D mesh. The relevant constants and variables of the problem are defined as follows.

- Let α_i^j represent the jth angle in the ith triangle on the 2-D mesh.
- Let β_i^j represent the jth angle in the ith triangle on the 3-D mesh.
- Let w_j^i represent the weight associated with the jth angle in the ith triangle.
- Define the function $j(k)$, such that $\alpha_i^{j(k)}$ and $\beta_i^{j(k)}$ represent the angle at node N_k in the ith triangle in the 2-D and 3-D meshes, respectively.
- Define

$$\phi_i^{j(k)} = \begin{cases} 2\pi\beta_i^{j(k)}\frac{1}{\sum_m \beta_m^{j(k)}}, & \text{If } N_k \text{ is an interior node,} \\ \beta_i^{j(k)}, & \text{If } N_k \text{ is on the boundary.} \end{cases} \tag{5.13}$$

Then, the underlying optimization problem can be stated as

$$\min_{\alpha} \sum_i \sum_{j=1}^{3} (\alpha_i^j - \phi_i^j)^2 w_i^j, \tag{5.14}$$

subject to

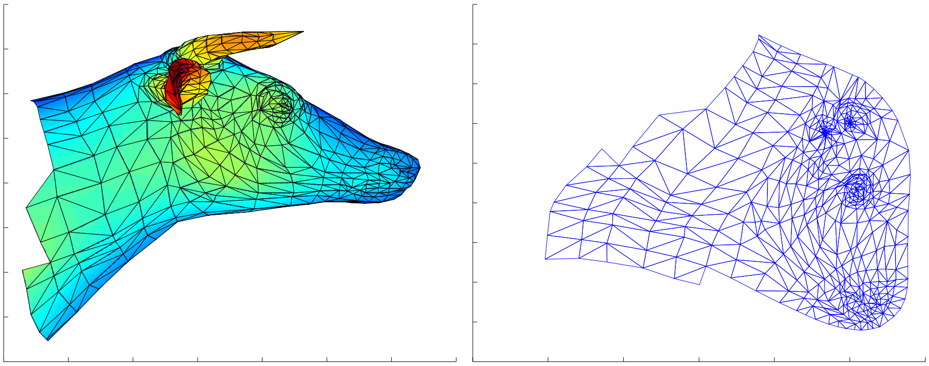
$$\begin{aligned} &\alpha_i^j \geq \epsilon_2 > 0, && \text{for all } i \text{ and } j = 1,2,3, \\ &\alpha_i^1 + \alpha_i^2 + \alpha_i^3 - \pi = 0, && \text{for all } i, \\ &\sum_i \alpha_i^{j(k)} = 2\pi, && \forall k \text{ s.t. } N_k \text{ is an interior node,} \\ &\frac{\prod_i \sin\left(\alpha_i^{j(k)+1 \mod 3+1}\right)}{\prod_i \sin\left(\alpha_i^{j(k)-1 \mod 3+1}\right)} = 1, && \forall k \text{ s.t. } N_k \text{ is an interior node.} \end{aligned}$$

Problem (5.14) is then solved with Newton's method, which yields linear systems of the form

$$\left[\begin{array}{cc|c} A + A_k & B^T & C_k^T \\ B & 0 & 0 \\ \hline C_k & 0 & 0 \end{array}\right] \begin{bmatrix} \alpha \\ y \\ z \end{bmatrix} = \begin{bmatrix} f \\ g \\ h \end{bmatrix}, \tag{5.15}$$

at each step, where the subscripted matrices change from iteration to iteration. Liesen, et al. [40] propose partitioning the matrix in the fashion indicated in (5.15) and preconditioning it in the fashion of de Sturler and Liesen [20]. In the nomenclature of Chapter 1, this problem satisfies C1–C6.

A solution of this optimization problem is given in Figures 5.5(a) and 5.5(b). They illustrate the original mesh and the corresponding flattened mesh from half of the head of a cow, respectively. This particular mesh contains 567 nodes and 1071 triangles. In addition, Zayer, Rössl and Seidel [61] propose several variants on the original formulation of Sheffer and de Sturler [52]. We choose to employ their modified wheel condition [61, Section 6], which replaces the multiplicative constraint in (5.14) with a summation of



(a) Mesh in 3-D

(b) Flattened mesh

Figure 5.5. Mesh flattening on a mesh representing half of the head of a cow (cow_halfh).

logarithms, namely,

$$\sum_i \log\sin\left(\alpha_i^{j(k)+1 \mod 3+1}\right) - \sum_i \log\sin\left(\alpha_i^{j(k)-1 \mod 3+1}\right) = 0,$$

for all k such that N_k is an interior node. This makes A_k diagonal. Our original C++ code was written by Alla Sheffer, but we have heavily modified it to increase efficiency and have included the modified wheel condition of Zayer, et al. [61].

6 Experimental Results

We discuss the application of our preconditioners to a variety of problems. These experiments serve to illustrate the effectiveness of the preconditioners themselves as well as the bounds described in Chapter 3. We also consider the application of structured probing techniques for approximate Schur complements and how they perform in the context of our preconditioners. We will use GMRES [50] as our Krylov method of choice and will run the algorithm without restarting for our example problems.

First, we consider the lid-driven cavity application described in Section 5.1. We use this application as a test-bed to illustrate both the preconditioner performance and the predictive power of the bounds described in Chapter 3. We illustrate how the accuracy of the splitting, F, and approximate Schur complement, S_2, affect the convergence of the preconditioned system. We also explore experimentally how the convergence of the preconditioned system depends on h. Finally, we use structured probing with our preconditioners and analyze the performance. These results are described in Section 6.1 and largely follow the presentation of Siefert and de Sturler [53, 54].

Second, in Section 6.2, we consider the application of our preconditioners to the incompressible Stokes problem described in Section 5.2. This problem allows us to consider the performance of our preconditioners for an application where $B \neq C$, in the notation of (1.2). We focus on convergence behavior and eigenvalue perturbation. We also briefly explore the scalability of the preconditioner with respect to the maximum polynomial degree used in the spectral collocation method. These results are an expanded version of what is reported by Siefert and de Sturler [53].

Next, in Section 6.3, we consider the effectiveness of our preconditioners applied to the metal deformation problem described in Section 5.3. Here, we focus on convergence and timing results for our preconditioned system. We pay special attention to the effectiveness of structured probing for approximating the Schur complement matrices in our preconditioners. We also explore how the convergence of the preconditioned system depends on h. This presentation expands significantly on the results presented by Siefert and de Sturler [54].

Finally, in Section 6.4, we consider the mesh flattening problem described in Section 5.4. Here, we focus on the application of approximate Schur complements, including structured probing. We present both convergence and timing results for this problem, focusing on the total linear solver cost for the entire non-linear iteration. These results have not been presented elsewhere.

6.1 Navier-Stokes: Lid-Driven Cavity

Unless otherwise specified, we perform our experiments on a 16×16 grid, with viscosity parameter $\nu = .1$ and stabilization parameter $\beta = .25$. After removing the constant pressure mode, this system has 705 unknowns. For splitting the (1,1) block we take two different approaches. The first involves employing a geometric multigrid technique [13]. Since multigrid cycles are actually matrix splittings, we use a number of multigrid V-cycles to define the splitting of the (1,1) block. For each V-cycle we use three SOR-Jacobi pre- and post-smoothing steps with relaxation parameter $\omega = .25$. As a purely algebraic alternative, we employ an ILUT factorization of the (1,1) block and vary the drop tolerance to change the accuracy of our splitting [49].

We begin by considering preconditioners with exact Schur complements in Section 6.1.1. From there we explore preconditioners with approximate Schur complements in Section 6.1.2. We also consider the issue of h-dependence in Section 6.1.3 and finally consider the use of probing-based approximations to Schur complements in Section 6.1.4.

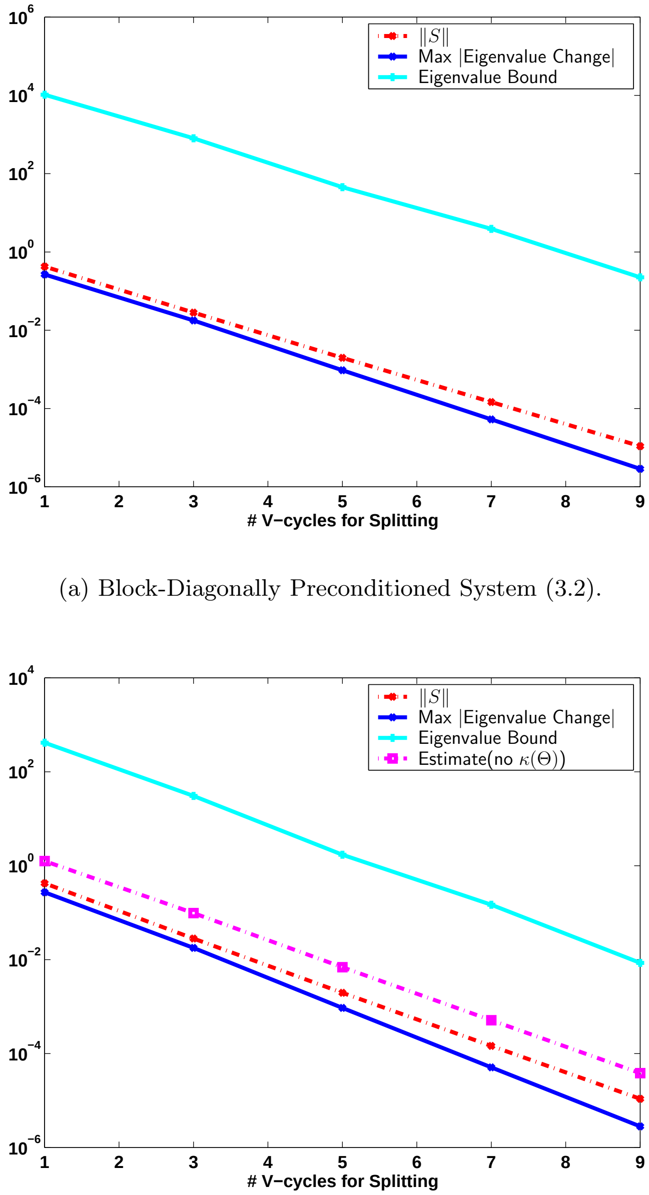
6.1.1 Exact Schur Complement Case

Eigenvalue Clustering and Bounds

Our first objective is to illustrate the clustering of the eigenvalues of our preconditioned system as well as the values of the bounds described in Chapter 3. We begin with the exact Schur complement case and vary the number of multigrid V-cycles for the splitting of the (1,1) block between one and nine. Figure 6.1(a) shows the maximum absolute eigenvalue perturbation from $\lambda \in \{1, \lambda_j^{\pm}\}$ for the block-diagonally preconditioned system (3.2) and Figure 6.1(b) shows the maximum absolute eigenvalue perturbation from 1 for the related system (3.37).

As we use a better splitting for A (more V-cycles), we see that the eigenvalue bound decreases with approximately the same rate as the corresponding eigenvalue perturbations, although the bound is pessimistic. This pessimism is mostly due to the $\kappa(\Theta)$ term, which is introduced in Lemma 3.1.4. Figure 6.1(b) includes an “estimate” of the perturbation for the related system, which consists of the bound in Corollary 3.2.2 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 description of the behavior of the eigenvalue perturbation as the splitting improves.

In comparing the block-diagonally preconditioned system (3.2) with the related system (3.37) 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) factors



(a) Block-Diagonally Preconditioned System (3.2).

(b) Related System (3.37).

Figure 6.1. 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 for the lid-driven cavity modeled with Navier-Stokes.

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, suggests that their convergence behavior will differ.

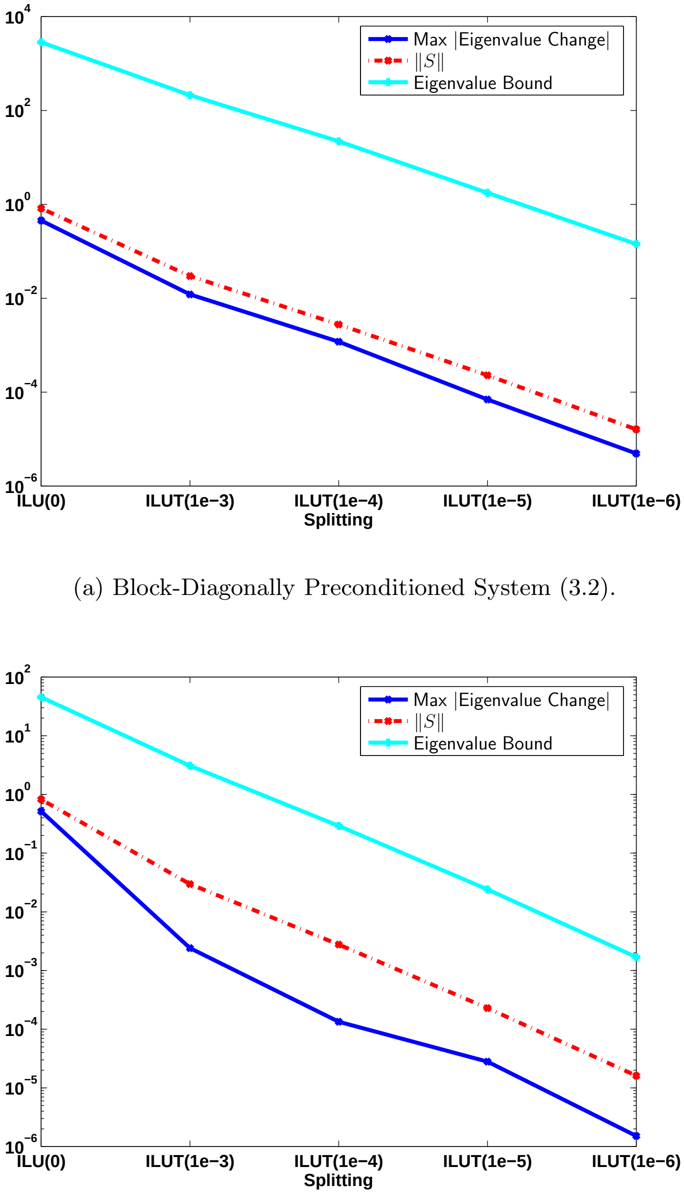
We see similar results for the case using ILU(0) or ILUT splittings. Figures 6.2(a) and 6.2(b) shows the results for the block-diagonally preconditioned system (3.2) and the related system (3.37), respectively. In this case, the eigenvalue bound is about two orders of magnitude smaller for the related system, even though the actual eigenvalue perturbation is about the same in magnitude. Again, due to the number of clusters in the block-diagonally preconditioned case, we should see slower convergence.

GMRES Convergence

Figures 6.3(a) and 6.3(b) show the convergence history for preconditioned GMRES for the block-diagonally preconditioned system (3.2) and the related system (3.37), respectively. Note that GMRES on the related system converges in significantly fewer iterations, for any choice of the number of V-cycles. This demonstrates the relative performance difference between the block-diagonally preconditioned system and the related system, as suggested by the aforementioned eigenvalue clustering and bounds. We see similar results in Figures 6.4(a) and 6.4(b) when we use ILUT for a splitting, instead of multigrid.

6.1.2 Approximate Schur Complement Case

We use an ILUT decomposition [49] to approximate the Schur complement. While this may not be a practical choice, it serves our purposes, 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 $1\mathrm{e}-3$ to $5\mathrm{e}-8$.



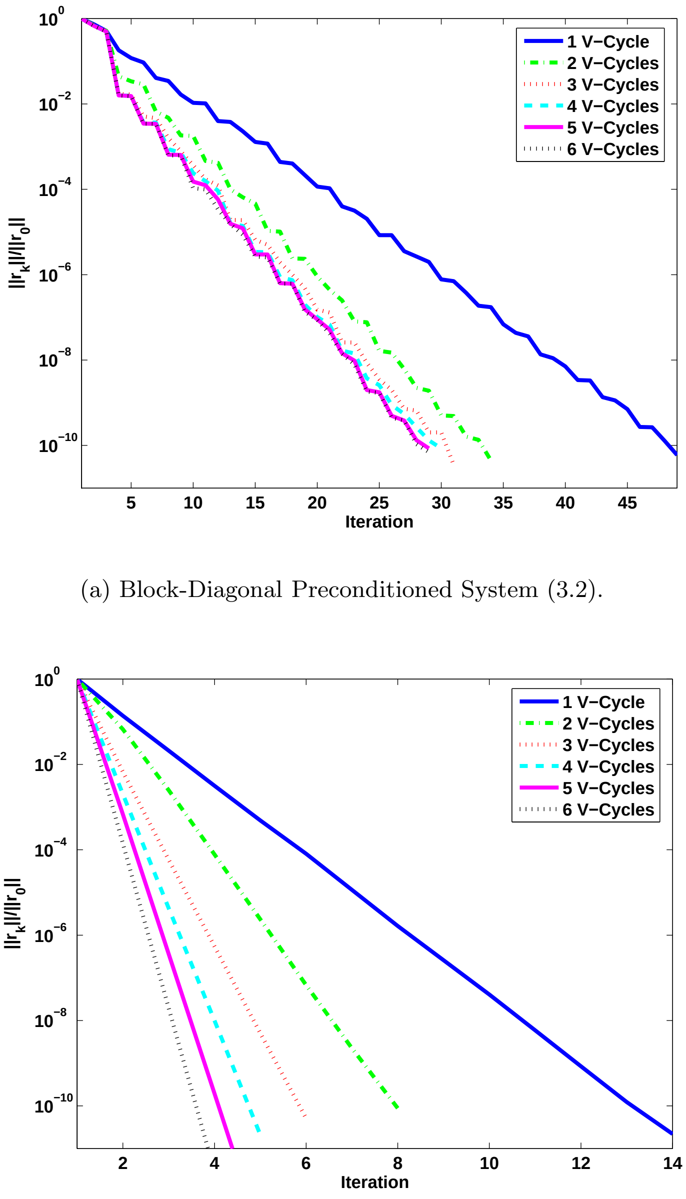
(a) Block-Diagonally Preconditioned System (3.2).

(b) Related System (3.37).

Figure 6.2. Maximum absolute eigenvalue perturbation and perturbation bounds, for both types of preconditioners, using the exact Schur complement and varying the ILUT drop tolerance for the lid-driven cavity modeled with Navier-Stokes.

Eigenvalue Clustering and Bounds

We begin by varying the drop tolerance for the inexact Schur complement while using a fixed splitting. For a fixed multigrid splitting, we set the

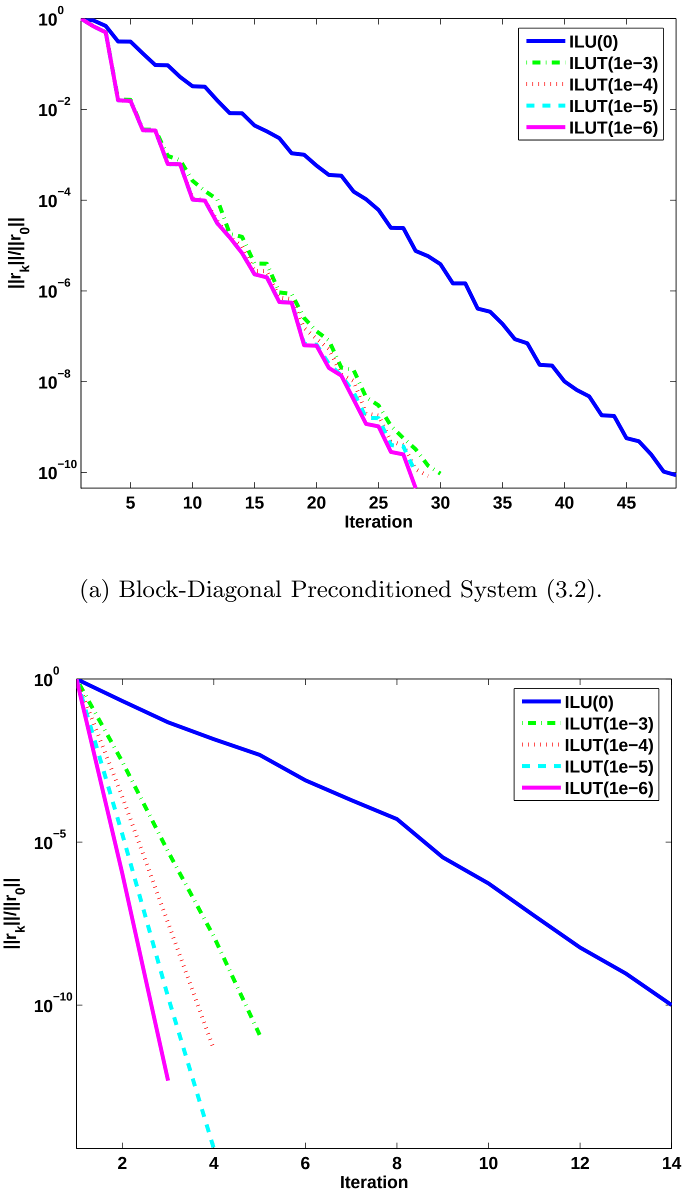


(a) Block-Diagonal Preconditioned System (3.2).

(b) Related System (3.37).

Figure 6.3. Convergence of GMRES for both types of preconditioners, using the exact Schur complement and varying the number of V-cycles for the splitting on the lid-driven cavity modeled with Navier-Stokes.

number of V-cycles to seven. For a fixed ILUT splitting, we set the drop tolerance to 1e−5. After this we present results where we vary the splittings for the (1,1) block, some using V-cycles and some using ILUT, while fixing the drop tolerance for the ILUT decomposition of the Schur complement to



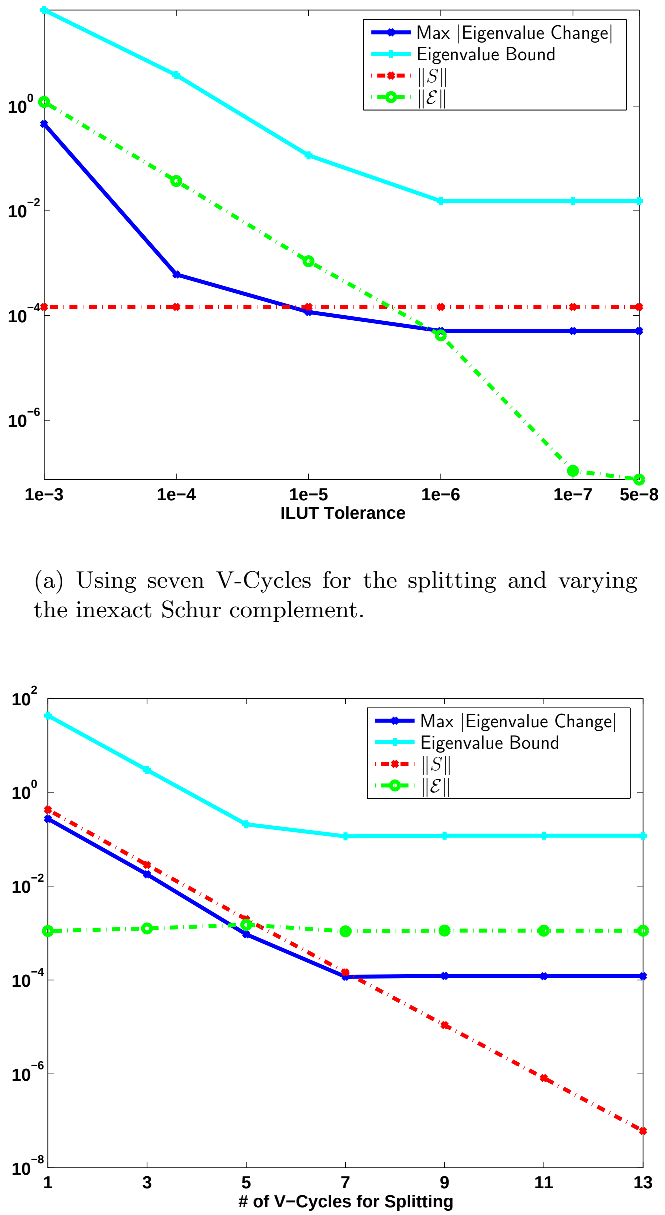
(a) Block-Diagonal Preconditioned System (3.2).

(b) Related System (3.37).

Figure 6.4. Convergence of GMRES for both types of preconditioners, using the exact Schur complement and varying the ILUT tolerance for the splitting on the lid-driven cavity modeled with Navier-Stokes.

$1e-5$. These results are shown in Figures 6.5(a), 6.6(a), 6.5(b) and 6.6(b), respectively.

Note that in both plots, shortly after $\|S\|$ is less than $\|\mathcal{E}\|$, or vice versa, the eigenvalue perturbation (and bound) cease to decrease. This suggests

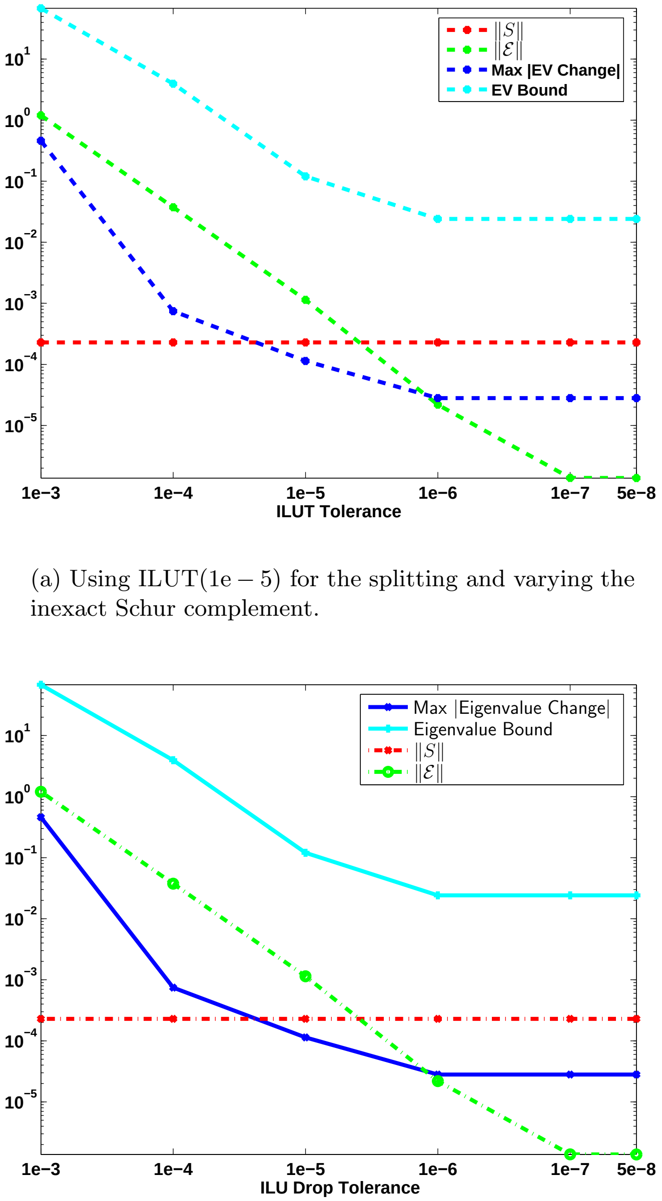


(a) Using seven V-Cycles for the splitting and varying the inexact Schur complement.

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

Figure 6.5. The effects of $\|S\|$ and $\|\mathcal{E}\|$ on related system (3.52) using the inexact Schur complement and a multigrid splitting for the lid-driven cavity modeled with Navier-Stokes.

that the behavior of the eigenvalue bound (Theorem 3.4.1) is indicative of the actual eigenvalue perturbation, and that using a significantly more accurate splitting than Schur complement approximation, or vice versa, yields little



(a) Using ILUT($1e-5$) for the splitting and varying the inexact Schur complement.

(b) Using the inexact Schur complement with ILUT($1e-5$) and varying the ILUT tolerance for the splitting.

Figure 6.6. The effects of $\|S\|$ and $\|\mathcal{E}\|$ on related system (3.52) using the inexact Schur complement and an ILUT splitting for the lid-driven cavity modeled with Navier-Stokes.

benefit. Note also that for reasonable choices of splitting and approximation to the Schur complement, the bounds are less than 1, indicating that the eigenvalues are clustered away from the origin. This should lead to very rapid convergence for Krylov methods.

GMRES Convergence

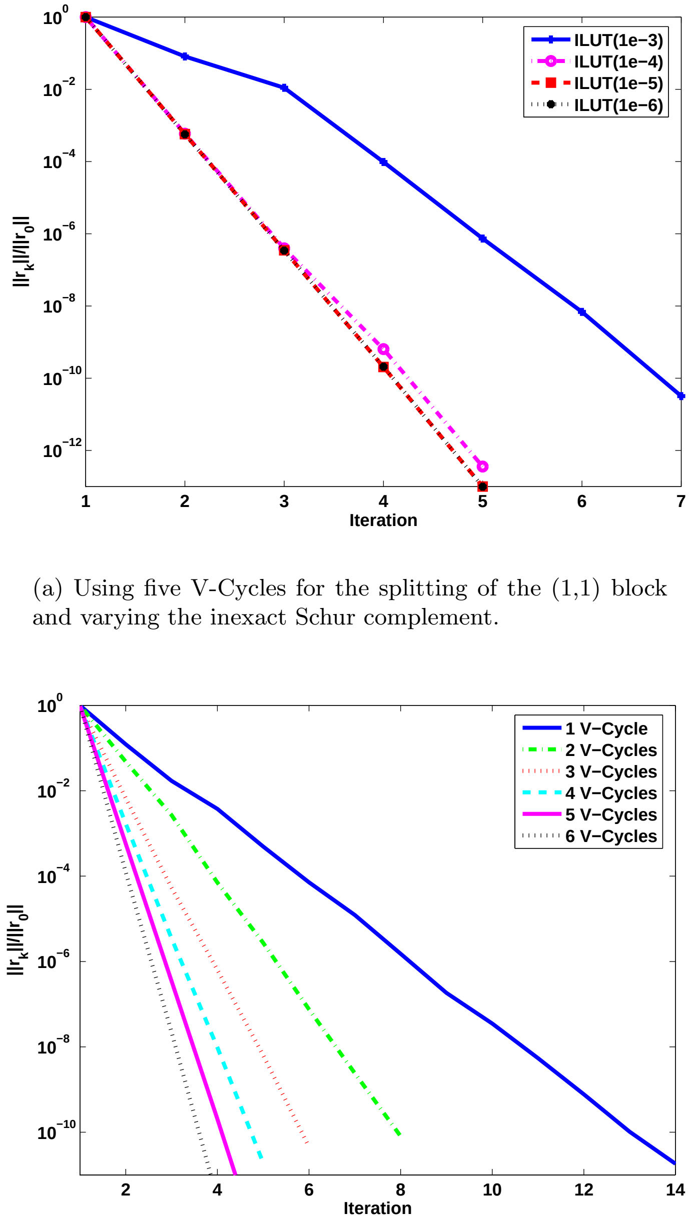
Figures 6.7(a) and 6.7(b) show the convergence of GMRES for the related system (3.52) using multigrid as a splitting. Figures 6.8(a) and 6.8(b) show similar results using ILUT as a splitting. This allows us to see the effects of improving the splitting (both multigrid and ILUT) and the approximation to the Schur complement.

The convergence results are quite good, regardless of the choice of splitting. However, like the results in Figures 6.3(a) and 6.3(b), the convergence rates in Figures 6.7(a), 6.7(b), 6.8(a) and 6.8(b) hit 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. As illustrated above, this is due to the eigenvalue clustering of the preconditioned system.

6.1.3 Examining h-dependence

Varying the number of grid points per dimension, $N = 1/h$, gives us some insight into the h-dependence of the related system (3.52). Table 6.1 summarizes these results when we use multigrid to split the (1,1) block. We do not expect similar results for the using ILUT to split the (1,1) block, since that preconditioner does not generate h-independent approximations for the convection-diffusion operator. Thus, it would be unreasonable to expect an ILUT splitting of the (1,1) block to lead to an h-independent preconditioner for the overall problem.

With respect to Table 6.1, note that the modulus of δ_j decreases with h, regardless of which splitting we use. This means that the terms that are functions of the δ_j's in the theorems of Sections 3.1, 3.2, 3.3 and 3.4, will be very small. Also note that the convergence of GMRES for the related system (3.52) depends only mildly on h. A good splitting and a sufficiently accurate approximate Schur complement appear to lead to h-independent

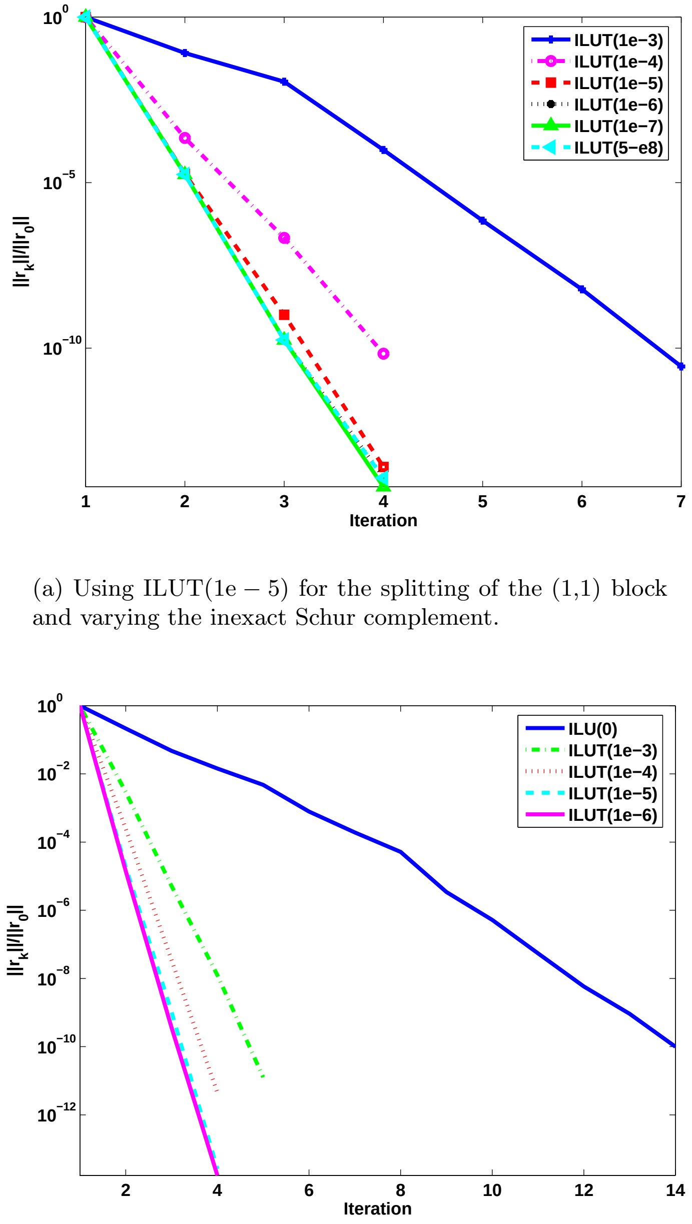


(a) Using five V-Cycles for the splitting of the (1,1) block and varying the inexact Schur complement.

(b) Using the inexact Schur complement with ILUT(1e − 5) and varying the number of V-cycles for the splitting of the (1,1) block.

Figure 6.7. Convergence results for the related system (3.52) using an inexact Schur complement and a multigrid splitting for the lid-driven cavity modeled with Navier-Stokes.

convergence.



(a) Using ILUT($1e-5$) for the splitting of the (1,1) block and varying the inexact Schur complement.

(b) Using the inexact Schur complement with ILUT($1e-5$) and varying the ILUT tolerance for the splitting of the (1,1) block.

Figure 6.8. Convergence results for the related system (3.52) using an inexact Schur complement and an ILUT splitting for the lid-driven cavity modeled with Navier-Stokes.

6.1.4 Structured Probing

In addition to demonstrating the effectiveness of our preconditioners with approximate Schur complements generated by structured probing, we use

		Approximate Schur Complement			
N	max $\lvert\delta_j\rvert$	ILUT(1e-3)	ILUT(1e-4)	ILUT(1e-5)	ILUT(1e-6)
Splitting: 1 V-Cycle					
4	1.85e+00	10	10	10	10
8	5.61e-01	13	13	13	13
16	1.50e-01	15	14	14	14
32	3.84e-02	18	14	13	13
Splitting: 3 V-Cycles					
4	1.72e+00	6	6	6	6
8	5.92e-01	6	6	6	6
16	1.60e-01	8	6	6	6
32	4.07e-02	13	7	6	6
Splitting: 5 V-Cycles					
4	1.72e+00	5	5	5	5
8	5.92e-01	5	4	4	4
16	1.60e-01	7	5	5	5
32	4.07e-02	13	6	5	5
Splitting: 7 V-Cycles					
4	1.72e+00	4	4	4	4
8	5.92e-01	5	4	4	4
16	1.60e-01	7	4	4	4
32	4.07e-02	13	6	4	4

Table 6.1. Effect of varying the number of grid points per dimension (N) on the maximum modulus δ and the number of GMRES iterations for the related system (3.52) for various numbers of V-Cycles for the splitting and the ILUT tolerance for the approximate Schur complement on the lid-driven cavity modeled with Navier-Stokes

the Navier-Stokes problem to illustrate the superiority of structured probing to banded probing. Specifically, we focus on the role of the sparsity pattern chosen for the approximate Schur complement.

We use the prime divisor method (Algorithm 7) for graph coloring to isolate the role of this chosen sparsity pattern. This allows us to use the same probing vectors for banded probing and for structured probing. Thus, the only difference between the two methods is the sparsity pattern, H, used for the construction of the approximation. Hence, we are not trying to get the most out of structured probing, but rather demonstrate that even using the same vectors as banded probing, reconstruction based on a better sparsity pattern leads to much better eigenvalue clustering and convergence for our preconditioned system.

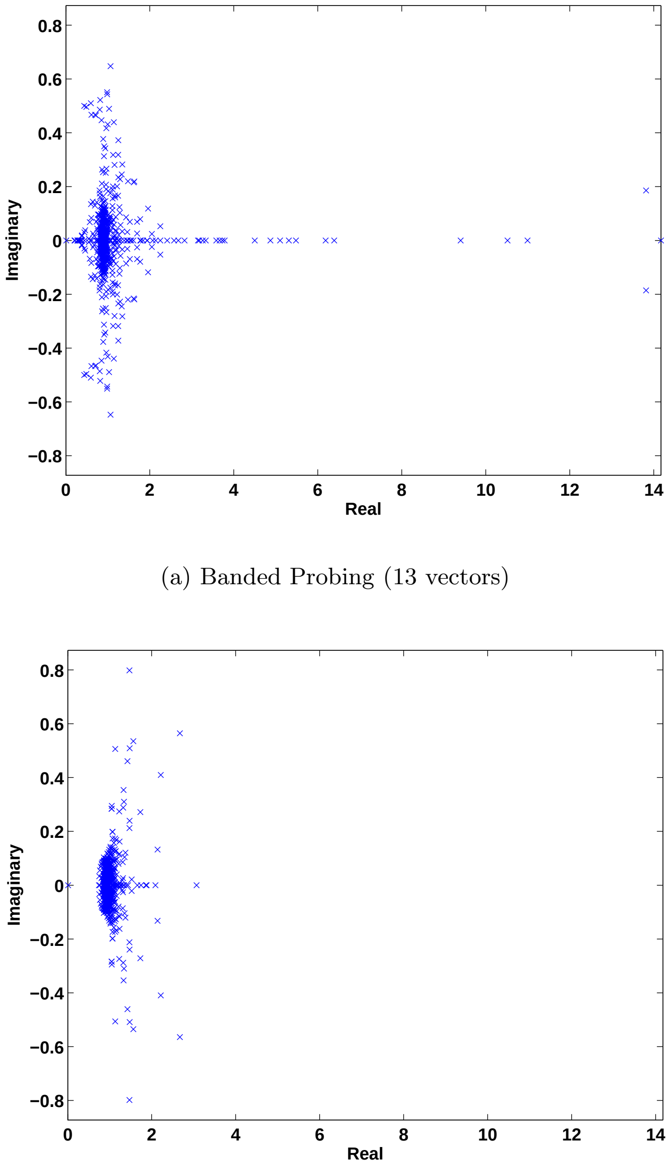
Eigenvalue Clustering

Figures 6.9(a) and 6.9(b) show the eigenvalue distributions for the related system (3.52) with banded probing and structured probing using 13 probing vectors. For scaling purposes, we exclude two negative eigenvalues at approximately $(-73, 0)$ and $(-113, 0)$ for the banded probing case. We use a nine-point stencil on the pressure connectivity graph to define the sparsity pattern H for structured probing. As the prime divisor coloring (Algorithm 7) requires 13 vectors for this stencil, we also perform banded probing using the same 13 vectors.

Structured probing yields much better clustering than banded probing, especially near the origin. The system resulting from structured probing has only one small eigenvalue (about 0.01); the others are well separated from zero. The system resulting from banded probing has many eigenvalues clustering near the origin. This should lead to poor convergence behavior.

We see similar results for the eigenvalues of the block-diagonally preconditioned system (3.47) for banded and structured probing; see Figures 6.10(a) and 6.10(b). Both probing and structured probing yield one small eigenvalue (about 0.01), but structured probing clusters eigenvalues much further away from the origin.

Figures 6.11(a) and 6.11(b) show the eigenvalues for the block-diagonally preconditioned system (3.47) and related system related system for structured probing with both five-point (seven vector) and nine-point (thirteen vector) stencils. Note that barring a few outliers, the eigenvalue clustering is significantly better for the nine-point stencil with both kinds of preconditioners, especially near the origin. Krylov-subspace methods tend to find and "remove" outlying eigenvalues quickly. Therefore, these eigenvalues do not affect the convergence rate after some number of initial iterations. Thus,

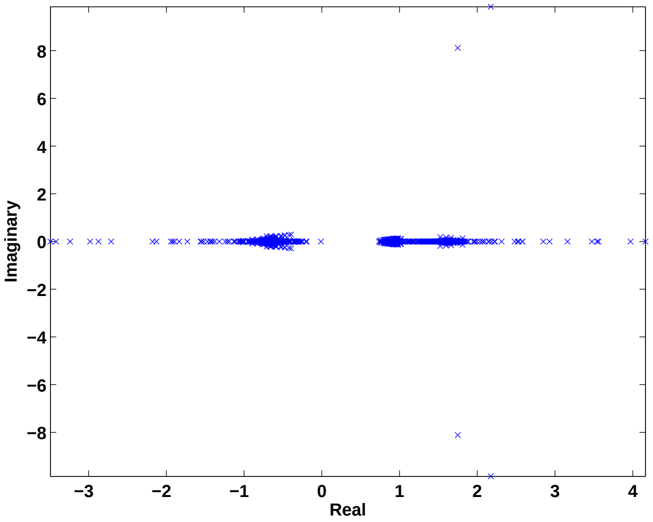


(a) Banded Probing (13 vectors)

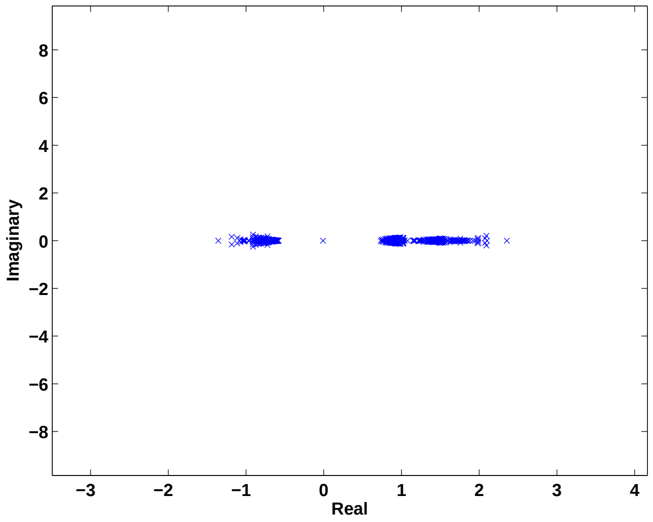
(b) Structured Probing (9pt. stencil / 13 vectors)

Figure 6.9. Eigenvalues for the related system (3.52) with one V-cycle as splitting of the (1,1) block and approximate Schur complements using banded and structured probing with exact factorizations on the lid-driven cavity modeled with Navier-Stokes.

the significantly better eigenvalue clustering obtained using the nine-point stencil should lead to a significantly improved convergence rate for GMRES.



(a) Banded Probing (13 vectors)

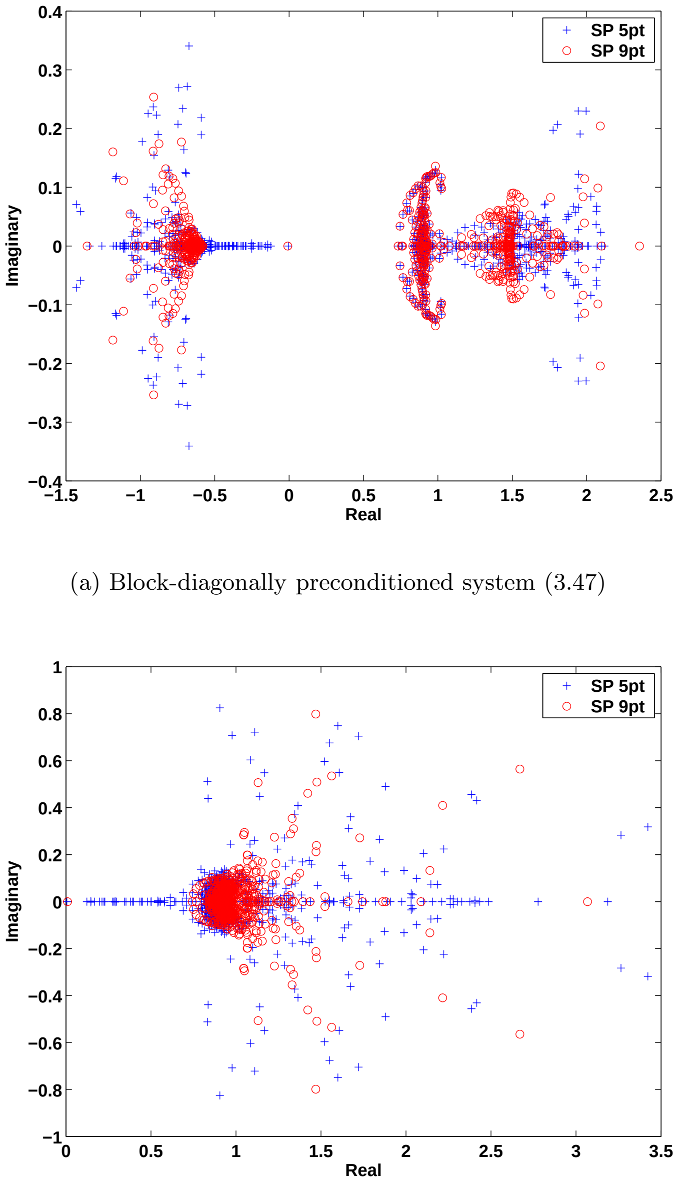


(b) Structured Probing (9pt. stencil / 13 vectors)

Figure 6.10. Eigenvalues for the block-diagonally preconditioned system (3.47) with one V-cycle as splitting of the (1,1) block and approximate Schur complements using banded and structured probing with exact factorizations on the lid-driven cavity modeled with Navier-Stokes.

GMRES Convergence

Figures 6.12(a) and 6.12(b) show the convergence of GMRES for the preconditioned systems. The difference between banded and structured probing is



(a) Block-diagonally preconditioned system (3.47)

(b) Related system (3.52)

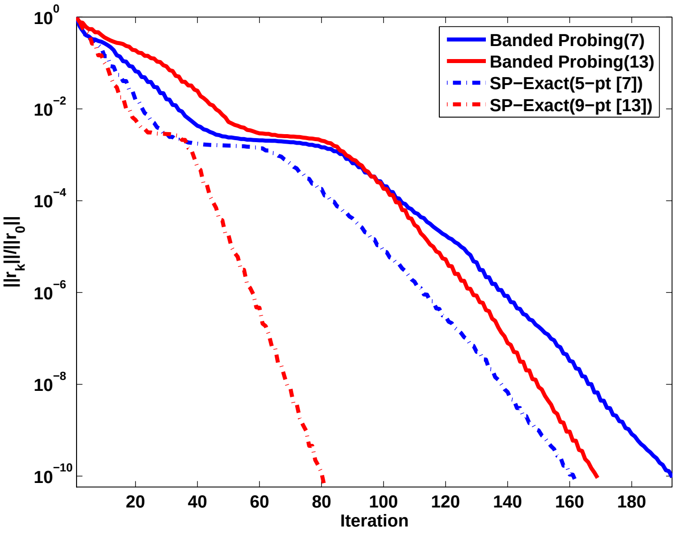
Figure 6.11. Eigenvalues for the block-diagonally preconditioned system (3.47) and the related system (3.52) with one V-cycle as splitting of the (1,1) block and approximate Schur complements using structured probing with exact factorizations on the lid-driven cavity modeled with Navier-Stokes.

quite pronounced. For both systems, structured probing with a five-point stencil using seven vectors has a lower iteration count than banded probing with thirteen vectors. This is due to tighter eigenvalue clustering. We also note that using the related system (3.52) leads to significantly faster convergence than using the block-diagonally preconditioned system (3.47) for all probing variants. Finally, we note that the contrast in eigenvalue clustering between the 5-point and 9-point stencils for structured probing, shown in Figure 6.11, has significant effects on the convergence of both the block-diagonally preconditioned and related systems. In both cases, it approximately halves the number of GMRES iterations required to converge.

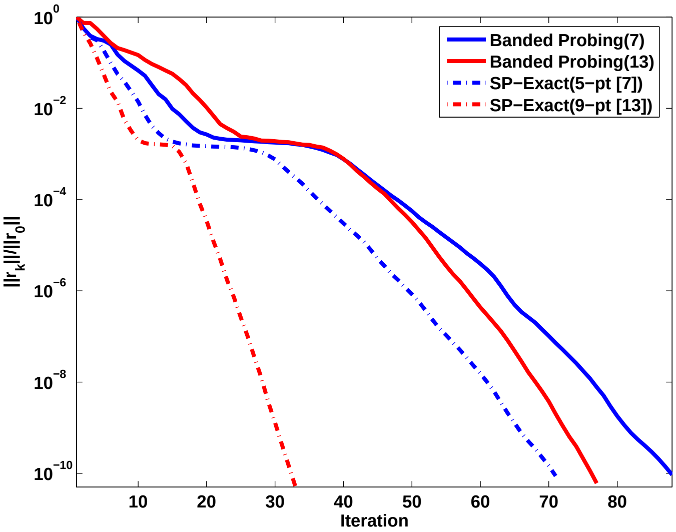
Inexact Factorizations

As discussed in Section 4.6, we also examine the use of incomplete factorizations for the approximate Schur complement matrices generated by structured probing as a means of further reducing the cost of the preconditioners (3.47) and (3.52). In practice, this leads to a negligible deterioration in convergence while reducing the overhead of applying structured probing significantly. We use an ILU(0) factorization for this problem. For symmetric problems an IC(0) factorization should be used. Since ILU(0) and IC(0) have linear cost in the number of unknowns, the overall cost remains $O(m)$.

Figure 6.13 shows the eigenvalue distributions for both preconditioned systems with structured probing using a nine point stencil (13 vectors) for both the exact and ILU(0) factorizations of the approximate Schur complement. Figure 6.14 shows the convergence results for both preconditioned systems, using structured probing with 9 and with 13 vectors. Using ILU(0) instead of an exact factorization changes the eigenvalue distribution slightly, but leaves the clustering essentially unchanged. The impact of such a change on the convergence behavior is negligible. Given the significant difference in cost between exact and inexact factorizations, using ILU(0) is more cost-effective than an exact factorization.



(a) Block-diagonally preconditioned system (3.47)

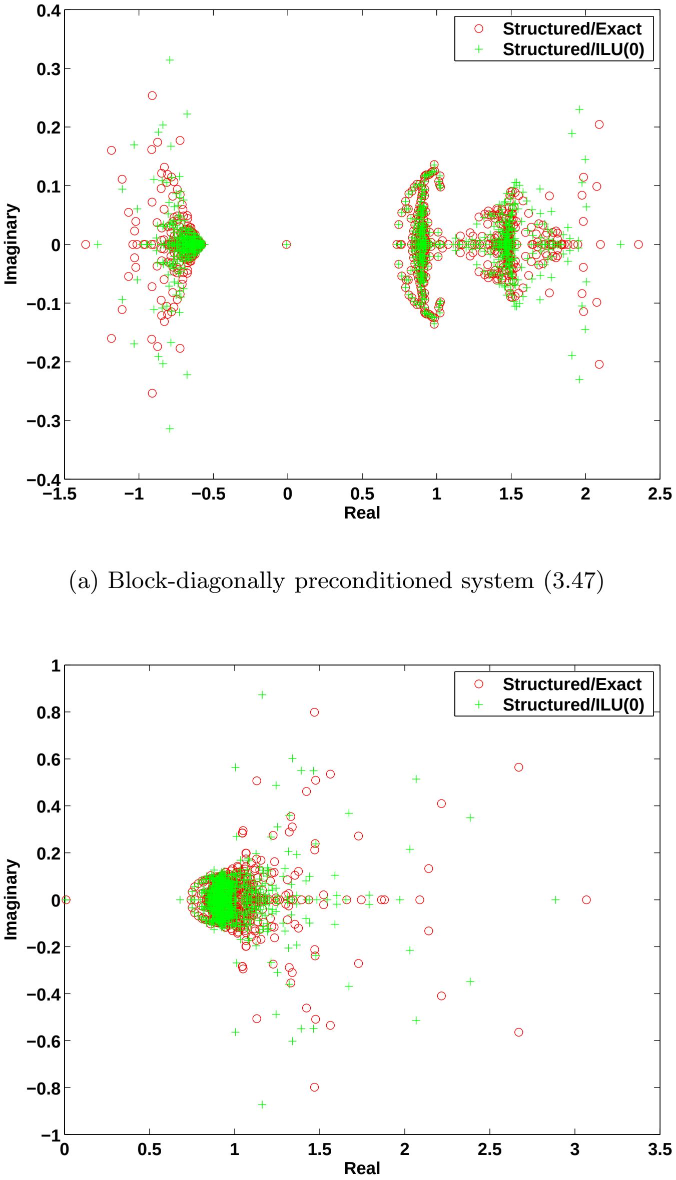


(b) Related system (3.52)

Figure 6.12. GMRES convergence with one V-cycle as splitting of the (1,1) block and approximate Schur complements using banded and structured probing with exact factorizations on the lid-driven cavity modeled with Navier-Stokes.

6.1.5 Examining h-dependence with Structured Probing

As demonstrated in Section 6.1.3, the related system (3.52) yields practically h-independent convergence when a sufficiently accurate ILUT factorization

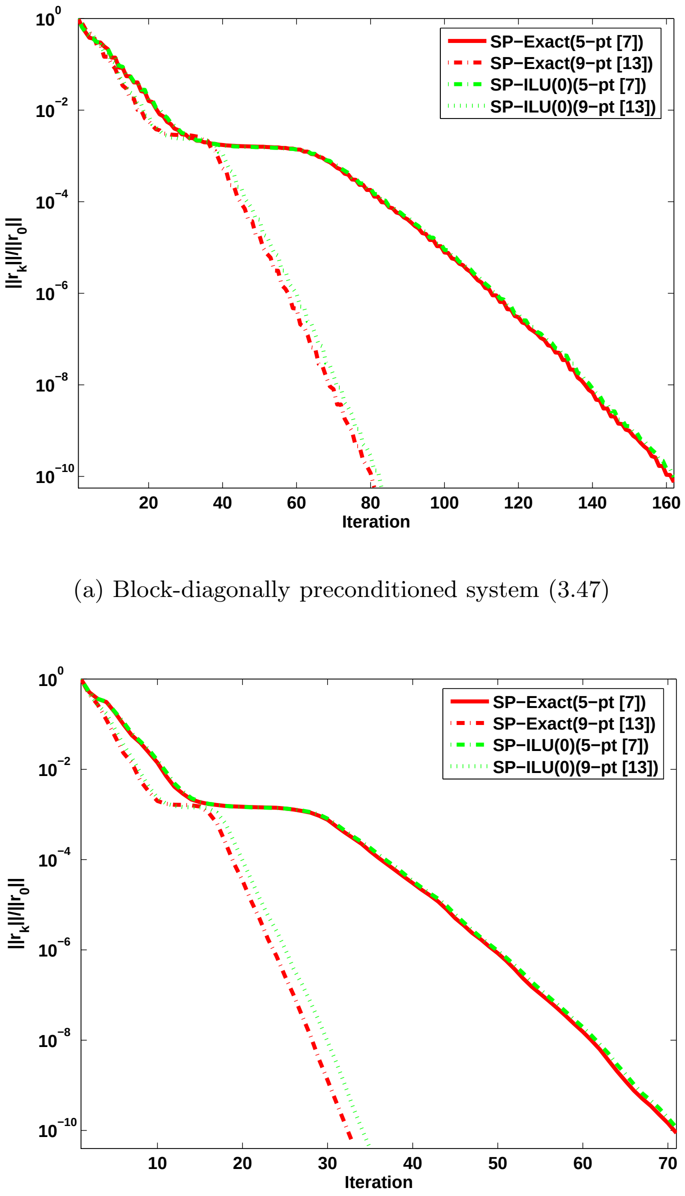


(a) Block-diagonally preconditioned system (3.47)

(b) Related system (3.52)

Figure 6.13. Eigenvalues for the block-diagonally preconditioned system (3.47) the related system (3.52) with one V-cycle as splitting of the (1,1) block and approximate Schur complements using structured probing (13 vectors) with exact and ILU(0) factorizations on the lid-driven cavity modeled with Navier-Stokes.

is used to approximate the Schur complement. We also consider the h-dependence of the related system using banded probing as well as structured probing with an ILU(0) factorization to approximate the Schur complement.



(a) Block-diagonally preconditioned system (3.47)

(b) Related system (3.52)

Figure 6.14. GMRES convergence with one V-cycle as the splitting of the (1,1) block and an inexact Schur complement computing using structured probing, using both exact and ILU(0) factorizations on the lid-driven cavity modeled with Navier-Stokes.

We vary the number of V-cycles of multigrid for the splitting, F, between one and four. Table 6.2 shows the estimated condition number, as computed by MATLAB's condest routine, for the related system (3.52) in each case. Since banded probing sometimes yields very poorly conditioned systems, we have

	Structured Probing H Stencil			Banded Probing w/nvecs equal to		
N	5-pt	9-pt	13-pt	5-pt	9-pt	13-pt
Splitting: 1 V-Cycle						
16	1.37e+01	6.73e+01	2.47e+02	8.06e+04	6.16e+03	8.06e+04
32	1.60e+01	2.44e+02	2.21e+03	1.49e+05	3.58e+04	5.41e+05
64	1.13e+01	5.64e+02	6.90e+03	4.02e+06	1.28e+09	4.02e+06
128	1.54e+01	8.56e+02	3.19e+04	2.41e+14	1.91e+09	2.56e+08
Splitting: 2 V-Cycles						
16	1.61e+01	7.07e+01	4.13e+02	7.31e+04	4.58e+03	7.31e+04
32	1.48e+01	2.69e+02	5.86e+03	2.64e+05	2.63e+04	2.14e+06
64	1.14e+01	5.21e+02	1.73e+04	3.37e+10	7.64e+07	3.37e+10
128	1.42e+01	6.95e+02	1.36e+05	7.99e+10	2.34e+09	2.91e+07
Splitting: 3 V-Cycles						
16	1.69e+01	6.86e+01	5.35e+02	6.47e+04	4.51e+03	6.47e+04
32	1.47e+01	2.95e+02	6.78e+03	1.78e+05	2.50e+04	3.21e+06
64	1.15e+01	5.22e+02	5.86e+04	6.40e+07	4.29e+07	6.40e+07
128	1.40e+01	7.22e+02	5.70e+04	2.21e+11	5.19e+08	2.71e+08
Splitting: 4 V-Cycles						
16	1.71e+01	6.86e+01	5.49e+02	7.52e+04	4.55e+03	7.52e+04
32	1.46e+01	3.06e+02	4.36e+03	1.60e+05	2.50e+04	3.59e+06
64	1.15e+01	5.22e+02	3.96e+03	4.89e+07	4.38e+07	4.89e+07
128	1.40e+01	7.35e+02	1.52e+05	2.81e+11	4.37e+08	9.08e+07

Table 6.2. Estimated condition number of the related system (3.52) using multigrid V-Cycles as a splitting of the (1,1) block and using structured probing with an ILU(0) factorization or banded probing to approximate the Schur complement. We use various levels of h-refinement in lid-driven cavity problem modelled with Navier-Stokes.

chosen to use unpreconditioned residual (rather than the preconditioned residual) as our convergence criterion, stopping when the residual is below $1\mathrm{e}-10$. Table 6.3 shows the corresponding GMRES convergence results.

We see that the related system with structured probing and an ILU(0) factorization shows relatively mild h-dependence. This scaling is significantly better than that of banded probing, which is not surprising, as the Schur complement matrix in this problem does not resemble a banded matrix.

	Structured Probing H Stencil			Banded Probing w/nvecs equal to		
N	5-pt	9-pt	13-pt	5-pt	9-pt	13-pt
Splitting: 1 V-Cycle						
16	77	37	33	127	74	127
32	102	57	52	161	119	318
64	119	82	74	643	1500($*$)	643
128	132	100	93	1500($*$)	1500($*$)	1500($*$)
Splitting: 2 V-Cycles						
16	75	34	27	125	76	125
32	103	52	42	162	122	313
64	119	76	63	1500($*$)	1500($*$)	1500($*$)
128	127	95	81	1500($*$)	1500($*$)	377
Splitting: 3 V-Cycles						
16	75	34	27	123	76	123
32	102	51	42	162	122	312
64	119	75	62	638	267	638
128	126	94	81	1500($*$)	1500($*$)	1500($*$)
Splitting: 4 V-Cycles						
16	75	34	27	122	76	122
32	102	51	42	162	122	312
64	119	73	63	634	1500($*$)	634
128	126	91	81	1500($*$)	1500($*$)	1500($*$)

Table 6.3. Number of GMRES iterations for the related system (3.52) using multigrid V-Cycles as a splitting of the (1,1) block and using structured probing with an ILU(0) factorization or banded probing to approximate the Schur complement. We use various levels of h-refinement in lid-driven cavity problem modelled with Navier-Stokes. Asterisks ($*$) indicate that the method did not converge in the listed number of iterations.

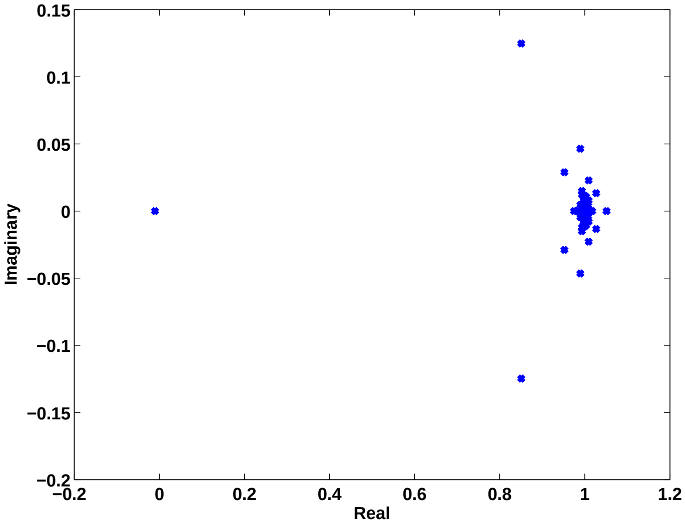
6.2 Incompressible Stokes: Spectral Collocation

For this problem, we build approximations with polynomials of degree up to 22. Our largest system will be of size 1241. We employ an odd-even ordering scheme on our velocity unknowns to exploit the orthogonality properties of Chebyshev polynomials and block-diagonalize the (1,1) block of our system. We use ILUT for both the splitting and inexact Schur complements of our spectral collocation matrix. For the splitting, we choose a drop tolerance of $1\mathrm{e}-4$. For the inexact Schur complement, we vary the drop tolerance between $1\mathrm{e}-3$ and $1\mathrm{e}-5$. Figure 6.15(a) shows the eigenvalues of the related system for the largest problem we considered with polynomials of degree up to 22. Except for a single small eigenvalue (of order $1\mathrm{e}-2$), the other eigenvalues are nicely clustered around one. As expected, this leads to rapid convergence.

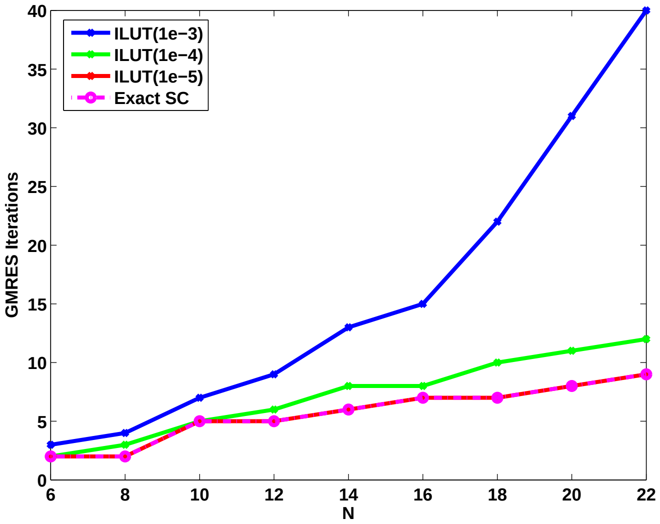
Second, we note that for the ILUT($1\mathrm{e}-4$) splitting, most choices of inexact Schur complement yield very rapid convergence, as shown in Figure 6.15(b). Moreover, the convergence of GMRES on the related system (except with the ILUT($1\mathrm{e}-3$) inexact Schur complement) depends only weakly on the maximum polynomial degree N. Thus, even for fully non-symmetric problems, our preconditioners are effective, and they show the potential of scaling well to larger problems.

6.3 Metal Deformation: Stress Relaxation in Thin Strips

For our first experiments, we consider a problem on a $101 \times 13 \times 2$ grid of points, which yields 5222 x-unknowns and 1200 y-unknowns, once boundary conditions are specified. For this problem, we employ three splitting strategies for the (1,1) block, A. First we use a diagonal splitting. Second, we use a banded splitting with a semi-bandwidth of four, or a bandwidth



(a) Eigenvalues of the related system (3.52) for $N = 22$ problem using ILUT($1e - 4$) inexact Schur complement.



(b) GMRES iterations of the related system (3.52) varying inexact Schur complement and maximum polynomial degree (N).

Figure 6.15. Related system (3.52) using ILUT($1e - 4$) splitting and an inexact Schur complement for spectral collocation method on the incompressible Stokes equations.

of nine. Third, we use an ILU(0) splitting. For our approximate Schur complement, we employ both structured probing and banded probing. For structured probing, we use a distance-2 balanced coloring on the element-

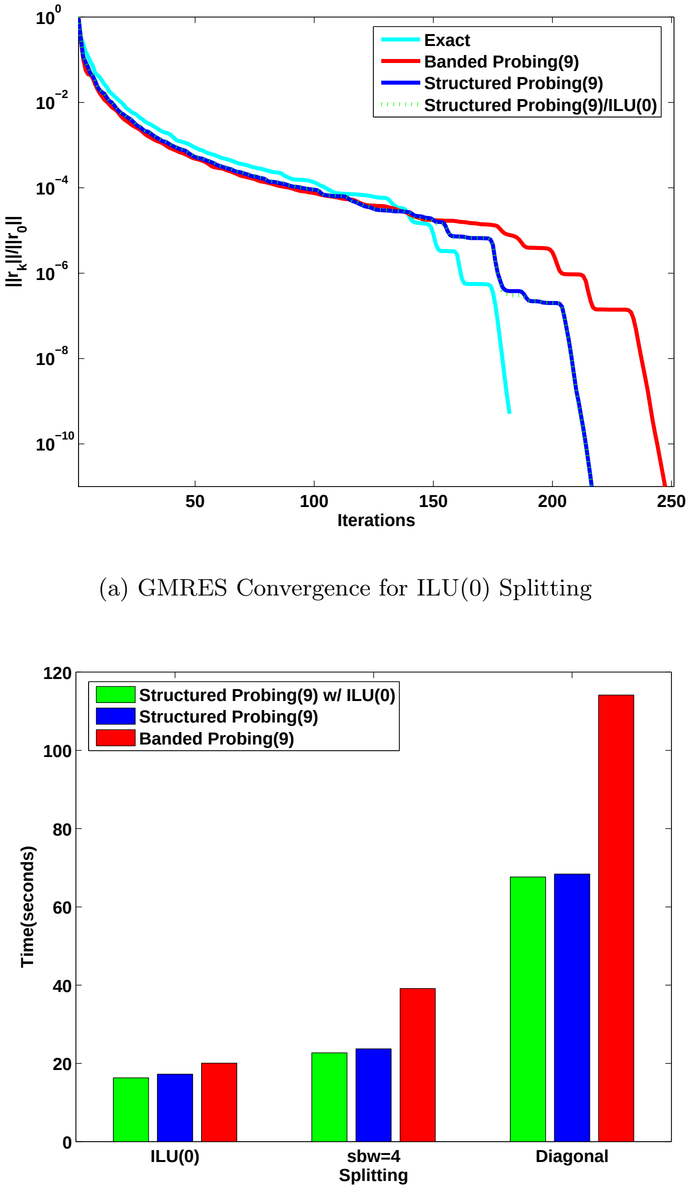
element connectivity graph. This coloring allows us to build our approximation using only nine probing vectors. We also perform banded probing using nine vectors. For comparison, we also include results using the exact Schur complement.

6.3.1 Preconditioner Performance

Figure 6.16(a) shows convergence results for the related system (3.52) using an ILU(0) splitting for the (1,1) block and banded probing, structured probing (using both exact and ILU(0) factorizations) and the exact Schur complement for a single non-linear iteration. With respect to the choice of approximate Schur complement we see that structured probing leads to faster convergence than banded probing.

It should be noted that this problem models a long, thin, piece of metal. So, although the problem is three dimensional, it is similar in nature to a one-dimensional problem (the elements are also ordered appropriately). Therefore, banded probing for the Schur complement does very well, as this problem shows the same type of 1-D decay as the 2-D domain decomposition problems for which banded probing was designed [14]. This is a relatively easy 3-D problem for banded probing and the improvements due to structured probing should be viewed in that light.

Figure 6.16(b) shows the corresponding wall clock solution time for a single linear system in the overall nonlinear iteration. These results are from a single run on a Sun V880 machine with 8 GB of ram running Solaris 8. Here we see that structured probing leads to a savings in time of 15% to 40% over banded probing, depending on the choice of splitting. In addition, using an inexact factorization on the probing matrix saves an additional 5% or so of execution time. These results show not only the efficacy of structured probing, but also the potential benefit of the use of inexact factorizations on the probed matrices.



(a) GMRES Convergence for ILU(0) Splitting

(b) Wall Clock Time

Figure 6.16. GMRES convergence and wall clock time for various probing-based inexact Schur complements in the related system (3.52) for three different splittings of the (1,1) block (ILU(0), banded matrix with semi-bandwidth four, and diagonal) for the metal deformation problem.

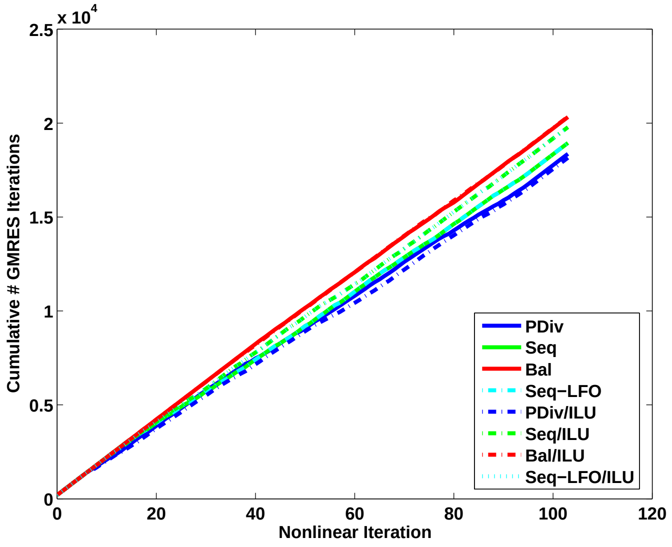
6.3.2 Inexact Factorizations

However, the benefit from using approximate factorizations is not always as pronounced as it is in Figure 6.16. Figures 6.17(a) and 6.17(b) show the cumulative number of GMRES iterations and wall clock time for the metal deformation problem across all 104 non-linear iterations. These timing results represent an average over ten runs on a Sun V880 machine with 8 GB of ram running Solaris 8. Four coloring schemes are used for structured probing — prime divisor, balanced, sequential and sequential with a largest-first ordering. The resulting matrix from structured probing is then factored with either an exact or an ILU(0) factorization. Both of the sequential colorings as well as the balanced coloring use 9 vectors, while the less efficient prime divisor coloring uses 17. The use of more vectors reduces the error in the approximate Schur complement somewhat, which explains why the prime divisor coloring leads to faster GMRES convergence and execution time. Choosing a larger sparsity pattern, H, for the approximate Schur complement and performing the coloring with a more efficient algorithm would be better use of computational effort.

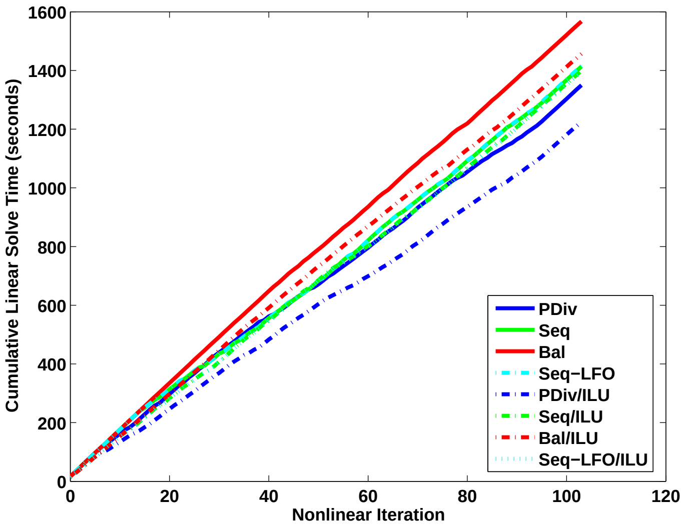
Since all of the methods are similar in performance, it is also informative to look at the extra cost in GMRES iterations and the time savings offered by using the ILU(0) factorization. Figures 6.18(a) and 6.18(b) show that for this problem, the using an ILU(0) with sequential coloring generally takes more GMRES iterations than using an exact factorization, using balanced coloring usually takes about the same number using prime divisor coloring often costs a little less. While all three algorithms save wall clock time by using an ILU(0) factorization, that gain is minimal for the sequential coloring.

6.3.3 Examining h-dependence

Varying the number of grid points per dimension, gives us some insight into the h-dependence of the related system (3.52) on our metal deformation problem. Here we multiply the number of points per dimension by $2, 3$ and 4, yielding systems of size 39,722, 121,466 and 273,258 respectively. For the



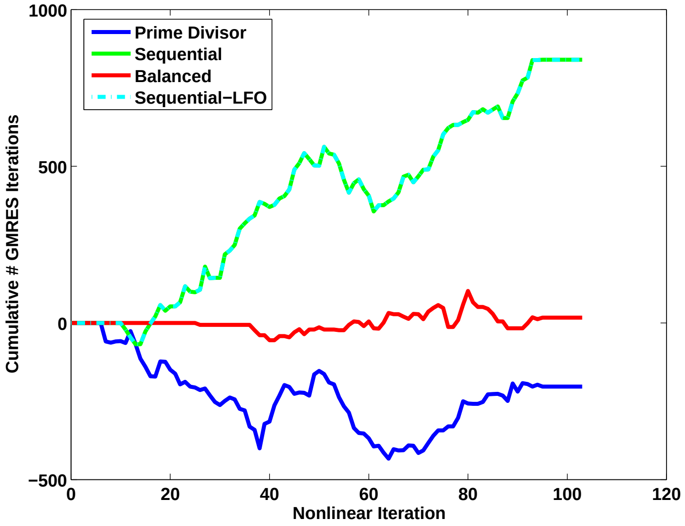
(a) Cumulative Number of GMRES Iterations



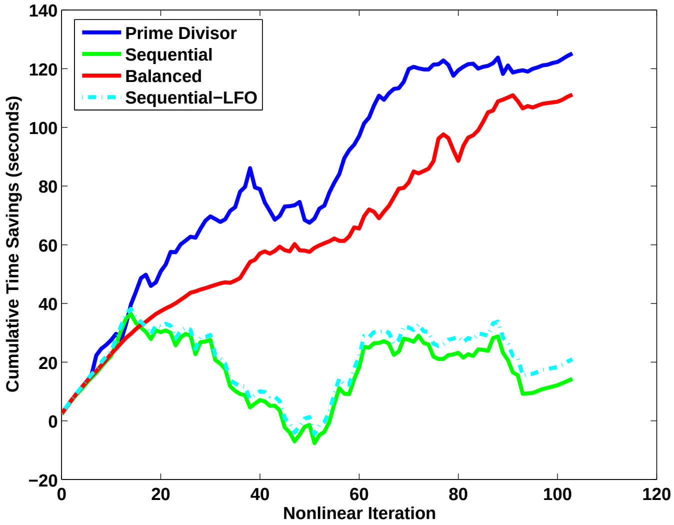
(b) Cumulative Wall Clock Time

Figure 6.17. Cumulative number of GMRES iterations and wall clock time for the related system (3.52) using structured probing (prime divisor, greedy sequential with a natural ordering, greedy sequential with an LFO ordering and balanced coloring), both with an exact an ILU(0) factorization, using an ILU(0) splitting of the (1,1) block of the metal deformation problem.

splitting of the (1,1) block, we use both ILU(0) and ML [51], a smoothed-aggregation algebraic multigrid (AMG). For ML, we choose a single step of SOR-Jacobi a pre- and post-smoother, with weighting parameter $\omega =$



(a) Cumulative Number of Additional GMRES Iterations



(b) Cumulative Wall Clock Time Saved

Figure 6.18. Cumulative cost in GMRES iterations and savings in time for the related system (3.52) using ILU(0) to factor the matrix resulting from structured probing (prime divisor, greedy sequential with a natural ordering, greedy sequential with an LFO ordering and balanced coloring) over the use of an exact factorization, using an ILU(0) splitting for the (1,1) block of the metal deformation problem.

0.5. We also set the aggregation threshold to 0.1, meaning that entries are dropped in the coarsening phase if $|A_{i,j}| \leq 0.1\sqrt{|A_{i,i}A_{j,j}|}$. All other parameters, including the uncoupled coarsening technique are set to the ML defaults. We then apply structured probing using either a balanced or a prime divisor coloring using both exact and ILU(0) factorizations of the resulting approximate Schur complement.

Table 6.4 shows the number of GMRES iterations necessary to solve the related system (3.52) to a tolerance of $1\mathrm{e}-10$. These results represent a minimum over five runs on a machine with a 2.6 GHz Intel Xeon CPU and 2 GB of ram running version 2.6.8 of the Linux kernel. Note that while the system does not demonstrate h-independence, the scaling with h is not particularly severe. Better scaling, in terms of iterations, could be achieved using a more accurate splitting or Schur complement, but that additional effort would take more wall-clock time. Results for the largest system are not complete, as certain combinations of splittings and approximate Schur complements require more memory than the machine has. Table 6.5 shows the wall-clock times corresponding to Table 6.4. While clearly h-dependent, these times scale relatively well with the increase in the problem size.

6.4 Optimization: Mesh Flattening

Table 6.6 provides information about the different meshes on which we run the flattening algorithm. The linear systems generated by the flattening algorithm range from fairly small (cat_head with 1,276 unknowns) to moderately large (bethf and beth1 with about 22,000 unknowns).

As exact Schur complement approaches were explored for this problem by de Sturler and Liesen [20], we consider only the use of approximate Schur complements generated with structured probing. We use a fixed splitting that does not change between iterations, namely, the matrix A in (5.15). We use a greedy distance-2 coloring on the connectivity graph of the internal nodes of the mesh in order to perform structured probing.

Figure 6.7 shows the number of GMRES iterations needed at each nonlinear iteration in order to converge to a tolerance of $1\mathrm{e}-10$. The use of

	Structured Coloring			Banded w/nvecs equal to	
# Unknowns	Prime Div.	Balanced	Greedy	Prime Div.	Greedy
Splitting: Exact					
6422	18	32	25	31	25
39722	23	38	31	51	50
121466	25	34	35	—	—
273258	—	—	—	—	—
Splitting: ILU(0)					
6422	219	230	225	252	252
39722	432	477	478	533	533
121466	644	700	730	794	794
273258	—	—	—	—	—
Splitting: AMG					
6422	153	159	159	164	164
39722	215	227	227	260	240
121466	223	241	249	289	314
273258	266	285	307	—	—

Table 6.4. Number of GMRES iterations for the related system (3.52) using an ILU(0) or AMG splitting of the (1,1) block and banded or structured probing to approximate the Schur complement, for various levels of h-refinement on a single non-linear iteration in the metal deformation problem. Dashes indicate insufficient memory to run that particular combination.

	Structured Coloring			Banded w/nvecs equal to	
# Unknowns	Prime Div.	Balanced	Greedy	Prime Div.	Greedy
Splitting: Exact					
6422	7.79e-01	1.07e+00	8.91e-01	1.11e+00	8.91e-01
39722	1.32e+01	1.68e+01	1.46e+01	2.26e+01	1.95e+01
121466	9.91e+01	1.03e+02	1.04e+02	—	—
273258	—	—	—	—	—
Splitting: ILU(0)					
6422	4.09e+00	4.31e+00	4.22e+00	4.60e+00	4.67e+00
39722	1.17e+02	1.40e+02	1.38e+02	1.66e+02	1.65e+02
121466	1.44e+03	9.89e+02	1.08e+03	1.19e+03	1.27e+03
273258	—	—	—	—	—
Splitting: AMG					
6422	8.87e+00	9.10e+00	9.06e+00	9.28e+00	9.28e+00
39722	9.80e+01	1.03e+02	1.03e+02	1.18e+02	1.04e+02
121466	3.80e+02	4.09e+02	4.25e+02	4.55e+02	4.85e+02
273258	8.20e+02	9.12e+02	1.01e+03	—	—

Table 6.5. Wall-clock time (seconds) for the related system (3.52) using an ILU(0) or AMG splitting of the (1,1) block and banded or structured probing to approximate the Schur complement, for various levels of h-refinement on a single non-linear iteration in the metal deformation problem. Dashes indicate insufficient memory to run that particular combination.

Mesh	Nodes	Elements	Angles	System Size
cat_head	135	257	771	1276
balls	547	1032	3096	5102
isis	458	879	2637	4362
cow_halfh	567	1071	3213	5296
fu	1104	2126	6378	10552
bethf	2267	4432	13296	22074
beth1	2258	4429	13287	22062

Table 6.6. Number of nodes, elements, angles, and unknowns for each mesh used in the flattening problem.

		Nonlinear Iteration								
Problem	Method	1	2	3	4	5	6	7	8	9
isis	Exact SC	1	10	9	9					
	Structured Prob.	42	40	37	35					
cat_head	Exact SC	1	9	8	8					
	Structured Prob.	18	16	15	14					
balls	Exact SC	1	8	9	10					
	Structured Prob.	20	21	20	21					
cow_halfh	Exact SC	1	13	11	10	11	12			
	Structured Prob.	27	37	30	28	28	28			
fu	Exact SC	1	11	12	14					
	Structured Prob.	31	30	28	28					
beethf	Exact SC	1	19	14	13	12	9	8	8	10
	Structured Prob.	42	46	44	41	35	31	29	27	30
beeth1	Exact SC	1	19	14	13	12	9	8	8	10
	Structured Prob.	42	46	44	41	35	31	29	27	30

Table 6.7. Number of GMRES iterations needed for each nonlinear iteration in the mesh flattening problem.

structured probing requires more iterations than the use of the exact Schur complement, but the iteration count is still modest. If forming the exact Schur complement is expensive, structured probing still should yield a time savings.

Figure 6.19 shows the total linear solver time, across all nonlinear iterations, for three solvers — the related system with an exact Schur complement (3.37), the related system with a structured probing Schur complement (3.52) and SuperLU, a direct solver [22]. These timing results represent an average over ten runs on a Sun V880 machine with 8 GB of RAM running

Solaris 8. Compared with using an exact Schur complement, the structured probing approximation saves between 60%, on the smallest mesh, to 92% of the total linear solution time. On the small meshes, SuperLU is faster than the iterative solver, which is to be expected. However, on the largest two meshes using the related system with structured probing saves about 11% over using SuperLU.

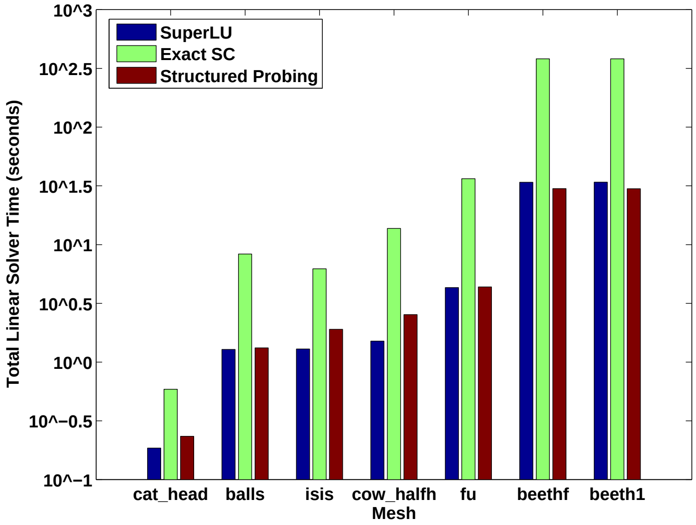


Figure 6.19. Total linear solver time for complete nonlinear solve, using SuperLU [22], related system with exact Schur complement (3.37) and related system with a structured probing Schur complement (3.52) on a variety of mesh flattening problems

.

7 Conclusions

We have discussed preconditioners for non-symmetric generalized saddle-point problems (1.2), with a focus on the general, non-symmetric case. In Chapter 3, we generalized the block-diagonal preconditioner and the related system from de Sturler and Liesen [20] to allow for a non-zero (2,2) block (problems that do not satisfy condition C5). We presented an analysis of the location of the eigenvalues both of these preconditioned systems. As the Schur complement matrices involved in these preconditioners can be expensive to form and factor, we also extended these preconditioners to allow for the use of approximations to the Schur complement. The eigenvalue analysis of Sections 3.1 and 3.2 was extended accordingly.

In Chapter 4, we presented structured probing, a graph coloring technique intended for generating approximations to Schur complement matrices. In addition to proving some analytic properties about the probing process, we extended the results of Demko, et al. [21] to show graph-based decay for sparse matrices and used these results to derive bounds on the accuracy of matrix approximations generated by structured probing. This allows us to argue for the applicability of probing techniques to problems not derived from a system of PDEs. Finally, we specialized these bounds to approximating Schur complement matrices.

Chapter 5 presented the four applications for which we presented results in Chapter 6. Our experiments illustrated the predictive power of the eigenvalue analysis of Chapter 3 as well as the performance of our preconditioners in practice. We also demonstrated empirical h-independent convergence for the lid-driven cavity problem.

Our results also demonstrated the effectiveness of structured probing as Schur complement approximation technique, especially when compared to classic probing. This reduced the wall-clock solution time significantly from using an exact Schur complement and led to more rapid solutions than a direct solver on moderate-sized problems. We also proposed the use of inexact factorizations of the matrices resulting from structured probing, and demonstrated the effectiveness of this idea in reducing the total solution time.

References

[1] M. Arioli, J. Maryška, M. Rozložník, and M. Tůma, *Dual variable methods for mixed-hybrid finite element approximation of the potential fluid flow problem in porous media*, Tech. Rep. RAL-TR-2000-023, Rutherford Appleton Laboratory, December 2004.

[2] I. Babuška, *The finite element method with Lagrange multipliers*, Numer. Math., 20 (1973), pp. 179–193.

[3] A. Battermann and E. Sachs, *An indefinite preconditioner for KKT systems arising in optimal control problems*, tech. rep., Universität Trier, 2000.

[4] ——, *Block preconditioners for KKT system in PDE-governed optimal control problems*, in Workshop on Fast Solution of Discretized Optimization Problems, Weierstrass Institude for Applied Analysis and Stochastics, Berlin, 2001, pp. 1–18.

[5] M. Benzi, G. Golub, and J. Liesen, *Numerical solution of saddle point problems*, Acta Numer., 14 (2005), pp. 1–137.

[6] M. Benzi, C. Meyer, and M. Tůma, *A sparse approximate inverse preconditioner for the conjugate gradient method*, SIAM J. Sci. Comput., 17 (1996), pp. 1135–1149.

[7] M. Benzi and M. Tůma, *A sparse approximate inverse preconditioner for nonsymmetric linear systems*, SIAM J. Sci. Comput., 19 (1998), pp. 968–994.

[8] 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 (1988), pp. 1237–1271.

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

[10] D. Braess, P. Deuflhard, and K. Lipnikov, *A subspace cascadic multigrid method for mortar elements*, Computing, 69 (2002), pp. 205–225.

[11] D. Braess and R. Sarazin, *An efficient smoother for the Stokes problem*, Applied Numerical Mathematics, 23 (1997), pp. 3–19.

[12] J. Bramble and J. Pasciak, *A preconditioning technique for indefinite systems resulting from mixed approximations of elliptic problems*, Math. Comp., 50 (1988), pp. 1–17.

[13] W. Briggs, V. Henson, and S. McCormick, *A Multigrid Tutorial*, SIAM, 2nd ed., 2000.

[14] T. Chan and T. Mathew, *The interface probing technique in domain decomposition*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 212–238.

[15] T. Coleman and J. Moré, *Estimation of sparse Jacobian matrices and graph coloring problems*, SIAM J. Numer. Anal., 20 (1983), pp. 287–309.

[16] ——, *Estimation of sparse Hessian matrices and graph coloring problems*, Math. Programming, 28 (1984), pp. 243–270.

[17] T. Cormen, C. Leiserson, and R. Rivest, *Introduction to Algorithms*, MIT Press, 1990.

[18] J. Cullum and M. Tůma, *Matrix-free preconditioning using partial matrix estimation*, Tech. Rep. 898, Institute of Computer Science, Academy of Sciences of the Czech Republic, April 2004.

[19] A. Curtis, M. Powell, and J. Reid, *On the estimation of sparse Jacobian matrices*, J. Inst. Math. Appl., 13 (1974), pp. 117–119.

[20] E. de Sturler and J. Liesen, *Block-diagonal and constraint preconditioners for nonsymmetric indefinite linear systems. Part I: Theory*, SIAM J. Sci. Comput., 26 (2005), pp. 1598–1619.

[21] S. Demko, W. Moss, and P. Smith, *Decay rates for inverses of band matrices*, Math. Comp., 43 (1984), pp. 491–499.

[22] J. Demmel, S. Eisenstat, J. Gilbert, X. Li, and J. Liu, *A supernodal approach to sparse partial pivoting*, SIAM J. Mat. Anal. Appl., 20 (1999), pp. 720–755.

[23] H. Dollar, *Extending constrainr preconditioners for saddle point problems*, Tech. Rep. NA-05/02, Numerical Analysis Group, Oxford University, January 2005.

[24] N. Dyn and W. Ferguson, Jr., *The numerical solution of equality-constrained quadratic programming problems*, Math. Comp., 41 (1983), pp. 165–170.

[25] H. Elman and D. Silvester, *Fast nonsymmetric iterations and preconditioning for Navier-Stokes equations*, SIAM J. Sci. Comput., 17 (1996), pp. 33–46.

[26] H. Elman, D. Silvester, and A. 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.

[27] C. Farhat, K. Pierson, and M. Lesoinne, *The second generation of FETI methods and their application to the parallel solution of large-scale linear and geometrically nonlinear structural analysis problems*, Computer Methods in Applied Mechanics and Engineering, 184 (2000), pp. 333–374.

[28] C. Farhat and F.-X. Roux, *Implicit parallel processing in structural mechanics*, in Computational Mechanics Advances, J. T. Oden, ed., vol. 2 (1), North-Holland, 1994, pp. 1–124.

[29] B. Fischer, A. Ramage, D. Silverter, and A. Wathen, *Minimum residual methods for augmented systems*, BIT, 38 (1998), pp. 527–543.

[30] H. Garmestani, M. Vaghar, and E. Hart, *A unified model for inelastic deformation of polycrystalline materials — application to transient behavior in cyclic loading and relaxation*, International Journal of Plasticity, (2001), pp. 1367–1391.

[31] A. Gebremedhin, F. Manne, and A. Pothen, *What color is your Jacobian? Graph coloring for computing derivatives*, SIAM Review, 47 (2005), pp. 629–705.

[32] P. Gill, W. Murray, D. Ponceleón, and M. Saunders, *Preconditioners for indefinite systems arising in optimization*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 292–311.

[33] G. Golub and A. Wathen, *An iteration for indefinite systems and its application to the Navier-Stokes equations*, SIAM J. Sci. Comput., 19 (1998), pp. 530–539.

[34] M. Grote and T. Huckle, *Parallel preconditioning with sparse approximate inverses*, SIAM J. Sci. Comput., 18 (1997), pp. 838–853.

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

[36] E. Hart, *Consitutive relations for the noneleastic deformation of metals*, J. Eng. Mater. Technol., 98 (1976), pp. 193–202.

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

[38] I. Ipsen, *A note on preconditioning nonsymmetric matrices*, SIAM J. Sci. Comput., 23 (2001), pp. 1050–1051.

[39] C. Keller, N. Gould, and A. Wathen, *Constraint preconditioning for indefinite linear systems*, SIAM Journal on Matrix Analysis and Applications, 21 (2000), pp. 1300–1317.

[40] J. Liesen, E. de Sturler, A. Sheffer, Y. Aydin, and C. Siefert, *Efficient computation of planar triangulations*, in Proceedings of the 10th International Meshing Roundtable, 2001.

[41] S. McCormick, *Optimal approximation of sparse Hessians and its equivalence to a graph coloring problem*, Math. Programming, 26 (1983), pp. 153–171.

[42] G. Meinardus, *Approximation of Functions: Theory and Numerical Methods*, vol. 13 of Springer Tracts in Natural Philosophy, Springer-Verlag, 1967. Translated by L.L. Schumaker.

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

[44] M. Murphy, G. Golub, and A. Wathen, *A note on preconditioning for indefinite linear systems*, SIAM J. Sci. Comput., 21 (2000), pp. 2969–2972.

[45] R. L. Panton, *Incompressible Flow*, John Wiley and Sons, 2nd ed., 1996.

[46] A. Peressini, F. Sullivan, and J. Uhl, Jr., *The Mathematics of Nonlinear Programming*, Springer-Verlag, 1988.

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

[48] M. Powell and P. Toint, *On the estimation of sparse Hessian matrices*, SIAM J. Numer. Anal., 16 (1979), pp. 1060–1074.

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

[50] Y. Saad and M. Schultz, *GMRES: A generalized minimum residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.

[51] M. Sala, M. Gee, J. Hu, and R. Tuminaro, *ML4.0 Smoothed Aggregation User's Guide*, Tech. Rep. SAND2005-4819, Sandia National Laboratories, Albuquerque, NM USA, 2005.

[52] A. Sheffer and E. de Sturler, *Parameterization of faceted surfaces for meshing using angle-based flattening*, Engineering with Computers, 17 (2001), pp. 326–337. Special issue: 10 best papers from the 9th International Meshing Round Table Conference.

[53] C. Siefert and E. de Sturler, *Preconditioners for generalized saddle-point problems*, Tech. Rep. UIUCDCS-R-2004-2448, Department of Computer Science, University of Illinois at Urbana-Champaign, June 2004. Accepted for publication in SIAM J. Numer. Anal.

[54] ——, *Probing methods for generalized saddle-point problems*, Tech. Rep. UIUCDCS-R-2005-2540, Department of Computer Science, University of Illinois at Urbana-Champaign, March 2005. Submitted to Electronic Transaction in Numerical Analysis(ETNA).

[55] 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 (2001), pp. 261–279. Numerical analysis 2000, Vol. VII, Partial differential equations.

[56] D. Silvester and A. Wathen, *Fast iterative solution of stabilised Stokes systems Part II: Using general block preconditioners*, SIAM J. Numer. Anal, 31 (1994), pp. 1352–1367.

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

[58] L. Trefethen and D. Bau III, *Numerical Linear Algebra*, SIAM, Philadelphia, 1997.

[59] H. Uzawa, *Iterative methods for concave programming*, in Studies in Linear and Nonlinear Programming, K. Arrow, L. Hurwicz, and H. Uzawa, eds., Stanford University Press, Stanford, CA, 1958, pp. 154–165.

[60] A. Wathen and D. Silvester, *Fast iterative solution of stabilised Stokes systems Part I: Using simple diagonal preconditioners*, SIAM J. Numer. Anal, 30 (1993), pp. 630–649.

[61] R. Zayer, C. Rössl, and H.-P. Seidel, *Variations on angle based flattening*, in Proceedings of Multiresolution in Geometric Modelling 2003, 2003, pp. 285–296.

[62] L. Zhu, *An Assessment of In-Service Stress Relaxation of a Work-Hardened Al-Mg Alloy*, PhD thesis, University of Illinois at Urbana-Champaign, 2003.

[63] L. Zhu, A. Beaudoin, and S. MacEwan, *A study of kinetics in stress relaxation of AA 5182*, in Proceedings of TMS Fall 2001: Microstructural Modeling and Prediction During Thermomechanical Processing, 2001, pp. 189–199.

[64] ——, *An assessment of in-service stress relaxation of a work-hardened aluminum magnesium alloy*, J. Eng. Mater. Technol., 126 (2004), pp. 157–163.

Author's Biography

Christopher Martin Siefert was born in Queens, NY on the 28^{th} of October, 1978. He received his **Bachelor of Science** degree in Computer Science and Mathematics from the College of William and Mary, Williamsburg, VA, with highest honors in Computer Science. He joined the graduate program in Computer Science at the University of Illinois at Urbana-Champaign in the fall of 2000.