Analysis and Preconditioning of Double Saddle-Point Systems

by

Susanne Bradley

B.Sc., Queen's University, 2013

M.Sc., University of British Columbia, 2015

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Computer Science)

The University of British Columbia

(Vancouver)

August 2022

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the thesis entitled:

Analysis and Preconditioning of Double Saddle-Point Systems

submitted by **Susanne Bradley** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy** in **Computer Science**.

Examining Committee:

Chen Greif, Professor, Computer Science, University of British Columbia
Supervisor

Robert Bridson, Adjunct Professor, Computer Science, University of British Columbia
Supervisory Committee Member

Eldad Haber, Professor, Earth Ocean and Atmospheric Sciences, University of British Columbia
University Examiner

Brian Wetton, Professor, Mathematics, University of British Columbia
University Examiner

John W. Pearson, Reader, Mathematics, University of Edinburgh
External Examiner

Additional Supervisory Committee Members:

Michael Friedlander, Professor, Computer Science, University of British Columbia
Supervisory Committee Member

Abstract

This thesis deals with the mathematical analysis and numerical solution of double saddle-point systems.

We derive bounds on the eigenvalues of a generic form of double saddle-point matrices with a positive definite leading block. The bounds are expressed in terms of extremal eigenvalues and singular values of the associated block matrices. Inertia and algebraic multiplicity of eigenvalues are considered as well. The analysis includes bounds for preconditioned matrices based on block diagonal preconditioners using Schur complements, and it is shown that in this case the eigenvalues are clustered within a few intervals bounded away from zero. Analysis for approximations of Schur complements is included. Some numerical observations validate our analytical findings.

We also derive bounds on the eigenvalues of (classical) saddle-point matrices with singular leading blocks. The technique of proof is based on augmentation. Our bounds depend on the principal angles between the ranges or kernels of the matrix blocks. We use these analyses to derive a preconditioner for saddle-point systems with singular leading blocks. Our preconditioning approach is based on augmenting the leading block and using Schur complements of the augmented system. We show that the resulting preconditioned operator has four distinct eigenvalues, and numerical experiments validate the effectiveness of our approach.

We then extend the preconditioner for saddle-point systems with a singular leading block to deal with double saddle-point systems with a singular leading block. The preconditioner is based on augmenting the leading block by a null matrix of one of the off-diagonal blocks, and using the Schur complements of the augmented system.

Lay Summary

This thesis deals with double saddle-point matrices, which are matrices that often arise in multiphysics problems (where one physical process is coupled with another) and constrained optimization (where we want to find the best solution to a problem, subject to certain conditions). These problems are so large that computers may take a long time to solve them or run out of memory unless we exploit the properties of the problem in a clever way. In this thesis we analyze some mathematical properties of these matrices, and use them as a starting point for developing efficient solution methods.

Preface

This thesis describes results in three research articles:

1. S. Bradley and C. Greif. Eigenvalue Bounds for Double Saddle-Point Systems. Submitted in revised form. (27 pages)

2. S. Bradley and C. Greif. Eigenvalue Bounds for Saddle-Point Systems with Singular Leading Blocks. In revision. (17 pages)

3. S. Bradley and C. Greif. Augmentation-Based Preconditioners for Saddle-Point Systems with Singular Leading Blocks. Under review. (17 pages)

All three papers have been submitted to journals. The first paper is described in Chapters 3 and 4. The second is described in Chapter 5. The third paper is described in Sections 6.1-6.3. Section 6.4 was written entirely by me and does not appear in a publication.

All three papers were co-authored with my research supervisor. In all three papers I was responsible for analyses, derived techniques, and numerical experiments, which form the core of this thesis. I received guidance from my research supervisor. All three pieces of work were mostly written by me, with the assistance and editing of my supervisor. These papers appear mostly as they are written for publication, with some changes to organization and notation for consistency throughout the thesis.

This thesis includes several numerical experiments on test problems arising from various applications. I wrote the code for the experiments themselves, but others wrote the code to generate the test matrices. Where code/test matrices are publicly available I have cited accordingly; but I have used code for the interior

point method implementation written by Erin Moulding, code to generate geophysics test problems written by Eldad Haber, and Maxwell test matrices written by Chen Greif and Dominik Schötzau.

Table of Contents

List of Tables

List of Figures

Glossary

CG Conjugate Gradient

GMRES Generalized Minimum Residual

IPM Interior Point Method

KKT Karush-Kuhn-Tucker

LP linear program

MINRES Minimum Residual

PDE partial differential equation

QP quadratic program

Acknowledgments

First and foremost, I would like to thank my supervisor, Chen Greif. Chen, this thesis could not have been written without your unwavering support. Your insight, creativity, and humour have been invaluable (and are pleasant qualities to have in a supervisor if you're going to linger in your Ph.D. program for seven years... *cough, cough*). Thank you for being my supervisor, mentor, and friend.

I would also like to thank my supervisory committee members, Robert Bridson and Michael Friedlander, for their guidance and insight. John Pearson at the University of Edinburgh served as the external examiner on this thesis, and provided thorough and insightful feedback that has improved the quality of the thesis. I'd also like to thank other faculty and staff at UBC who have provided me with valuable support and assistance in completing my studies and working towards career goals: Steve Wolfman, Dinesh Pai, Cinda Heeren, Uri Ascher, Eldad Haber, Brian Wetton, Stephen Gustafson, and the folks at CTLT, to name a few. I'd also like to thank the CS department's administrative staff, help desk, and course coordinators, who have helped me in a variety of ways.

I've had the pleasure to interact with many great students (both undergraduate and graduate) and postdocs at UBC. Thanks to everyone at the SCL, SSL, SPL, and ISW facilitator team for the support and camaraderie.

Financial support for this thesis was provided by a Canada Graduate Scholarship from the Natural Sciences and Engineering Research Council of Canada and a Four-Year Fellowship from the University of British Columbia.

Finally, I'd like to thank my family. My parents have provided a great deal of support, both financial and otherwise, throughout my education, and I couldn't have gotten this far without them. My husband Nick has provided more love and

support than I could have hoped for. And, finally, to my son Virgil: thank you for being a ray of sunshine in my life, and for always reminding me that a Ph.D. is just a Ph.D.

To Virgil and his little sibling

Chapter 1

Introduction

This thesis deals with the analysis and numerical solution of double (block-3×3) saddle-point systems. These systems appear frequently in multiphysics problems, and their numerical solution is of increasing importance and interest. As a result, there has been a recent surge of interest in the iterative solution of multiple saddle-point problems, and our work adds to the increasing body of literature that considers these problems. Recent papers that provide interesting analysis are, for example, [2, 5, 11, 65, 84].

In this chapter, we first provide a mathematical problem statement and discuss the connections between double and classical (block-2×2) saddle-point systems. We then provide a brief overview of iterative solution methods and preconditioning approaches for large, sparse linear systems. We then provide a high-level overview of the techniques we use for eigenvalue analysis and description of our mathematical notation. We conclude with an overview of the structure and contributions of thesis.

1.1 Problem statement

Given positive integer dimensions $n \geq m \geq p$, consider the $(n+m+p) \times (n+m+p)$ double saddle-point system

$$\mathscr{K} u = b, \tag{1.1}$$

where

$$\mathscr{K} = \begin{bmatrix} A & B^T & 0 \\ B & -D & C^T \\ 0 & C & E \end{bmatrix}; \quad u = \begin{bmatrix} x \\ y \\ z \end{bmatrix}; \quad b = \begin{bmatrix} p \\ q \\ r \end{bmatrix}. \tag{1.2}$$

In (1.2), $A \in \mathbb{R}^{n\times n}$ is assumed symmetric positive definite, $D \in \mathbb{R}^{m\times m}$ and $E \in \mathbb{R}^{p\times p}$ are positive semidefinite, and $B \in \mathbb{R}^{m\times n}$, $C \in \mathbb{R}^{p\times m}$.

The matrix $\mathscr{K}$ may be considered a generalization of the block-2×2 or "classical" (to use the terminology of [84]) saddle-point matrix

$$\mathscr{A} = \begin{bmatrix} A & B^T \\ B & -D \end{bmatrix}. \tag{1.3}$$

Classical saddle-point matrices arise as the first-order optimality conditions for equality-constrained quadratic programming problems of the form

$$\min_x \frac{1}{2} x^T A x - f^T x \tag{1.4a}$$

$$\text{subject to } Bx = g. \tag{1.4b}$$

Letting y denote the vector of Lagrange multipliers, an optimal solution of (1.4) is a saddle point for the Lagrangian:

$$\mathscr{L}(x,y) = \frac{1}{2} x^T A x - f^T x + (Bx - g)^T y.$$

This is equivalent to the solution (x,y) of the linear system:

$$\underbrace{\begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix}}_{=:\mathscr{A}_0} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix}, \tag{1.5}$$

where the coefficient matrix $\mathscr{A}_0$ a matrix in classical saddle-point form with $D = 0$. The matrix D in this case is often associated with regularization (or stabilization) [17], and matrices with $D = 0$ are often called unregularized (or unstabilized). Such a stabilization term is commonly used in, for example, the solution of the Stokes

problem [91], which is to find velocity u and pressure p satisfying

$$-\nu\nabla^2 u + \text{grad}p = f \quad \text{in } \Omega, \tag{1.6a}$$

$$\text{div}u = 0 \quad \text{in } \Omega. \tag{1.6b}$$

Block-3×3 matrices in tridiagonal form as in (1.2) can occur when a physical problem with constraints, such as the Stokes problem, is coupled with another physical process, as in the Stokes-Darcy problem [94], which represents fluid flow (represented by the Stokes equations) over a porous medium (where fluid flow within the porous medium is represented by the Darcy equations). The resulting linear system is

$$\begin{bmatrix} A_p & A_\Gamma^T & 0 \\ -A_\Gamma & A_f & B_f^T \\ 0 & B_f & 0 \end{bmatrix} \begin{bmatrix} \phi_h \\ \mathbf{u}_h \\ p_h \end{bmatrix} = \begin{bmatrix} f_{p,h} \\ \mathbf{f}_{f,h} \\ g_h \end{bmatrix}, \tag{1.7}$$

which is equivalent to (1.2) up to a difference in sign. The system $\begin{bmatrix} A_f & B_f^T \\ B_f & 0 \end{bmatrix}$ is the Stokes system, A_p is a symmetric positive Darcy matrix, and A_Γ is a coupling term that represents interface between the fluid and porous flow. For other examples of matrices with higher numbers of blocks resulting from the coupling of physical processes, we refer to, e.g., [27, 46, 50, 69, 72, 74].

Double saddle-point systems also arise in constrained optimization problems that can be expressed in the form (1.4) where the primal variable x itself consists of two components. This occurs in, for example PDE-constrained optimization, where these are the state and control variables. These systems can be viewed as double saddle-point systems (1.2), after some reordering of the variables.

Borrowing from the classical saddle-point terminology, we refer to the special case of double saddle-point matrices with $D = E = 0$ as an *unregularized form* of $\mathscr{K}$, and denote it by $\mathscr{K}_0$:

$$\mathscr{K}_0 = \begin{bmatrix} A & B^T & 0 \\ B & 0 & C^T \\ 0 & C & 0 \end{bmatrix}. \tag{1.8}$$

This simpler form has been analyzed in various ways in [48, 65, 84] and is of much potential interest, as it may be considered a direct generalization of the standard saddle-point form for classical saddle-point matrices:

Classical saddle-point matrices of the form (1.3) have been extensively studied, and their analytical and numerical properties are well understood; see [7, 73] for excellent surveys. Some properties of $\mathscr{K}$ follow from appropriately reordering and partitioning the block matrix and then using known results for block-2×2 saddle-point matrices (as given in, for example, [35, 75, 76, 81]).

Specifically, $\mathscr{K}$ can be reordered and partitioned into a 2×2 block matrix

$$\tilde{\mathscr{K}} = \left[\begin{array}{cc|c} A & 0 & B^T \\ 0 & E & C \\ \hline B & C^T & -D \end{array}\right]. \tag{1.9}$$

While this approach is often effective, we may benefit from considering the block-3×3 formulation $\mathscr{K}$ directly, without resorting to (1.9). When E is rank-deficient or zero (as occurs, for example, in [28, 55]), the leading 2×2 block of $\tilde{\mathscr{K}}$ is singular, even if A is full rank. It is then more challenging to develop preconditioners for $\tilde{\mathscr{K}}$ or to obtain bounds on its eigenvalues, as we are restricted to methods that can handle singular leading blocks. Additionally, by considering the block-3×3 formulation in deriving eigenvalue bounds, we will see in this thesis that we can derive effective bounds using the singular values of B and C (along with the eigenvalues of the diagonal blocks). Analyzing $\tilde{\mathscr{K}}$ using established results for block-2×2 matrices (such as those given by Rusten and Winther [76]) instead requires singular values of the larger off-diagonal block $\begin{bmatrix} B & C^T \end{bmatrix}$, which may be more difficult to obtain than singular values of B and C. (We can estimate the singular values of $\begin{bmatrix} B & C^T \end{bmatrix}$ in terms of those of B and C, but if these estimates are loose it will result in loose eigenvalue bounds.)

The above said, we will observe in this thesis that analysis and solution methods for classical saddle-point systems often provide useful insights in developing techniques for the double saddle-point case. Therefore, in this thesis we do perform some analysis of the classical saddle-point case. We particularly focus on classical saddle-point matrices with a singular leading block; this case is less explored

in the literature, and we shall see that our analyses of these matrices will provide useful tools in developing and analyzing a preconditioner for double saddle-point matrices with a singular leading block.

The distribution of eigenvalues of $\mathscr{K}$ plays a central role in determining the efficiency of iterative solvers. It is therefore useful to gain an understanding of the spectral structure as part of the selection and employment of solvers [86], and this comprises a major portion of this thesis. Effective preconditioners are instrumental in accelerating convergence for sparse and large linear systems; see [6, 63, 78, 92] for general overviews, and [7, 68, 73] for a useful overview of solvers and preconditioners for saddle-point problems. For multiphysics problems it has been demonstrated that exploiting the properties of the underlying discrete differential operators and other characteristics of the problem at hand is beneficial in the development of robust and fast solvers; see, e.g., [21].

1.2 Iterative solution of sparse linear systems

Double saddle-point systems of the form (1.2) are often large and sparse. Direct solvers, such as those based on Gaussian elimination (we refer to [15, 33, 87] for details), may introduce a large amount of fill-in: that is, the matrix decompositions needed for exact inversion may have many more nonzero entries than the original matrix. This can be problematic in the sparse matrix setting, where it is common for an $n \times n$ matrix to have roughly $O(n)$ nonzero entries; depending on the exact sparsity pattern of the matrix, the computed factors may be quite dense ($O(n^2)$ nonzero entries).

Additionally, when dealing with problems arising from the modeling of physical phenomena, a highly accurate solution like we obtain with direct solvers is often not necessary. Whenever we discretize a problem to form a linear system – by discretizing in space [17, 59], in time [4], or discretizing continuous processes like differentiation and integration [85] – we introduce error in the discretization process. As such, even a highly accurate solution of the linear system will still not represent a perfectly accurate solution to the underlying problem; thus there is typically no harm in solving the linear system a bit less accurately, particularly if we can do so at a lower computational cost.

Iterative solution methods for a linear system $Kx = b$ rely on matrix-vector products of the form Kv. Thus, they are well-suited for problems where matrix decompositions are expensive but matrix-vector products are relatively cheap, as is the case for large, sparse matrices. Iterative methods for linear systems fall into two broad categories: stationary iterative methods and Krylov subspace methods.

1.2.1 Stationary iterative methods

Stationary iterative methods solve a linear system with a simpler matrix approximating the original one (often based on some splitting of the original matrix). At each step, the iterate x_{k+1} is updated based on the residual at step k, defined by $r_k = b - Kx_k$. Specifically, if we consider a splitting of our matrix $K = M - N$, a stationary iteration takes the form

$$x_{k+1} = x_k + M^{-1} r_k.$$

Examples of methods of this type include the Richardson method, Jacobi method, Gauss-Seidel, and successive over-relaxation (SOR) method (see [78, Chapter 4]). In practice, these methods converge slowly and are rarely used on their own. They are, however, often used as a preprocessing step to accelerate the convergence of more sophisticated methods, such as multigrid methods (see [42] for an overview) or Krylov subspace methods.

1.2.2 Krylov subspace methods

The Krylov subspace methods select iterates in Krylov subspaces, which take the form:

$$\mathbb{K}_m(K, r_0) = \operatorname{span}\left\{r_0, Kr_0, K^2 r_0, \ldots K^{m-1} r_0\right\},$$

where $r_0 = b - Kx_0$ is the initial residual. The iterate x_m obtained after m iterations of such a method is of the form

$$x_m = x_0 + p_{m-1}(K) r_0,$$

where p_{m-1} is a polynomial of degree $m-1$. There are a variety of methods for forming the basis and selecting the next iterate; we refer to the books of [37, 78] for detailed overviews.

We describe here, at a high level, the Minimum Residual (MINRES) [62] method for symmetric indefinite matrices and the Generalized Minimum Residual (GMRES) [79] method for nonsymmetric matrices. Both GMRES and MINRES have the property of minimizing the norm of the residual at each iteration [3]; that is, the approximate solution x_m at iteration m satisfies:

$$\min_{x_m \in \mathbb{K}_m(K,r_0)} ||b - Kx_m||_2. \tag{1.10}$$

Both the MINRES and GMRES algorithms use an orthonormal basis for the subspace $\mathbb{K}_m(K,r_0)$; we let $Q_m \in \mathbb{R}^{m\times m}$ denote the matrix whose columns form that basis. The orthonormal basis is updated at each step of the algorithm by performing a step of the Arnoldi process (in GMRES) or the Lanczos process (in MINRES).

The first vector in the Krylov subspace is r_0, and thus the first orthonormal basis vector for $\mathbb{K}_m(K,r_0)$ is the unit vector $q_1 = r_0/||r_0||$. At the next step, we take Aq_1 (which is in the same direction as Ar_0) and orthonormalize it against the first basis vector q_1. At subsequent iterations, we construct the basis vector q_{m+1} by taking Aq_m and orthonormalizing against the previous basis vectors.

In GMRES, for any $m \geq 1$, we can write a matrix relation of the form

$$KQ_m = Q_{m+1}H_{m+1,m}, \tag{1.11}$$

where Q_k is a matrix of the first k basis vectors, and $H_{m+1,m} \in \mathbb{R}^{(m+1)\times m}$ is upper Hessenberg, meaning that all entries are zero below the first subdiagonal (stated more mathematically, we have that the matrix entry $H_{m+1,m}(i,j) = 0$ whenever $i > j+1$).

Our aim at iteration m is to find a vector in the Krylov subspace $\mathbb{K}_m(K,r_0)$, which we can write as

$$x_m = x_0 + Q_m z,$$

where $z \in \mathbb{R}^m$ minimizes the residual norm $||b - Kx_m||$. We have

$$||b - Kx_m|| = ||b - Kx_0 - KQ_m z|| = ||r_0 - Q_{m+1} H_{m+1,m}|| = ||Q_{m+1}^T r_0 - H_{m+1,m} z||.$$

Thus, we have now converted the original optimization problem (1.10) to a least-squares problem involving a smaller $(m+1) \times m$ upper Hessenberg matrix $H_{m+1,m}$. We can make the least-squares computations more efficient by maintaining a QR factorization of $H_{m+1,m}$ at each iteration and updating sequentially via Givens rotations; see [85, Section 6.5.3].

We can show that

$$Q_m^T K Q_m = H_{m,m},$$

where $H_{m,m} \in \mathbb{R}^{m \times m}$ is the matrix consisting of the first m rows of $H_{m+1,m}$. Thus, when K is symmetric, the matrix $H_{m,m}$ is symmetric and therefore tridiagonal (because the matrix is Hessenberg). Thus, $H_{m+1,m}$ is also tridiagonal, which allows simplifications to the process described in GMRES. In MINRES, which deals with symmetric indefinite matrices, the Arnoldi process amounts to the Lanczos process and the upper Hessenberg matrix $H_{m+1,m}$ in (1.11) is replaced by the tridiagonal matrix $T_{m+1,m}$:

$$KQ_m = Q_{m+1} T_{m+1,m}, \tag{1.12}$$

The result is that each MINRES iteration is cheaper than a GMRES iteration, and can in fact be implemented via a three-term recurrence relation like in the Conjugate Gradient (CG) method for positive definite matrices; see [43]. Thus, in MINRES we need only store the three most recent basis vectors, while for GMRES we must store all m previously computed basis vectors.

A way in which we may get around this issue for GMRES is by using *restarts*: after some predetermined number of iterations, we discard the computed basis vectors and proceed using the last iterate x_m as a new "initial guess." Restarted GMRES is denoted by GMRES(m), where m is the number of iterations between restarts of the Arnoldi process.

We note that not all iterative solution methods use an orthonormal basis for the Krylov subspace $\mathbb{K}_m(K, r_0)$. Variants such as QMR [25] or Bi-CG and its variants [23, 83, 89] instead use a bi-orthogonalization procedure for nonsymmetric

matrices.

1.3 Preconditioning

A disadvantage of iterative solvers is their potential lack of robustness. This is evident in particular for large, ill-conditioned matrices. For symmetric and diagonalizable matrices, the convergence of methods such as GMRES and MINRES depends on the distribution of eigenvalues of the matrix. (For nonsymmetric matrices which may not be diagonalizable, additional tools to assess the convergence of iterative solvers include the field of values [38] and pseudospectra [88].) For a nonsymmetric diagonalizable matrix $K = X\Lambda X^{-1}$, where $\Lambda = \text{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ is the diagonal matrix of eigenvalues and X the matrix of eigenvectors, the residual norm achieved by the mth step of GMRES satisfies [78, Proposition 6.32]

$$\frac{||r_m||_2}{||r_0||_2} \leq \kappa_2(X) \min_{p \in \mathbb{P}_m, p(0)=1} \max_{i=1,\ldots,n} |p(\lambda_i)|, \tag{1.13}$$

where $\mathbb{P}_m, p(0) = 1$ denotes the set of polynomials of degree less than or equal to m satisfying $p(0) = 1$ and $\kappa_2(X) = ||X||_2||X^{-1}||_2$ denotes the (2-norm) condition number of X. In the symmetric case, X is orthogonal, so the corresponding bound for MINRES reduces to

$$\frac{||r_m||_2}{||r_0||_2} \leq \min_{p \in \mathbb{P}_m, p(0)=1} \max_{i=1,\ldots,n} |p(\lambda_i)|. \tag{1.14}$$

An consequence of (1.13) and (1.14) is that convergence will be rapid if the eigenvalues of Λ are tightly clustered. In particular, if Λ contains s distinct eigenvalues, then MINRES or GMRES will converge in no more than s iterations in exact arithmetic.

Unfortunately, we have no control over the distribution of eigenvalues of our matrix K. This is where the idea of *preconditioning* comes in: we multiply K by another matrix such that the product of these matrices has a more favourable eigenvalue distribution, and we use our iterative solution method on the preconditioned operator.

1.3.1 Overview

A preconditioner is a matrix M that should satisfy a few properties. First, it should be inexpensive to solve the linear system $Mx = b$, because all preconditioned linear solvers will require solving a linear system with M at each step. It should also be nonsingular (for obvious reasons), and should be "close to" K in some way.

The preconditioner can be applied from the left, leading to the preconditioned system

$$M^{-1}Kx = M^{-1}b.$$

Alternatively, the preconditioned can be applied from the right:

$$KM^{-1}u = b, \;\; x := M^{-1}u.$$

It is also common for a preconditioner to be available in factored form

$$M = M_1 M_2,$$

in which we can apply the split preconditioning:

$$M_1^{-1}KM_2^{-1}u = M_1^{-1}b, \;\; x := M_2^{-1}u.$$

We then carry out our iterative solution method using the preconditioned operator. We therefore want the preconditioned operator to have few (or tightly clustered) eigenvalues in order to ensure rapid convergence.

An important constraint is that, if we have a symmetric system K and wish to use MINRES as our solver, we need to maintain symmetry of the preconditioned operator. This means that M must be positive definite: in that case, we can split M into two factors

$$M = M_1 M_1^T, \tag{1.15}$$

and the split preconditioned operator yields the symmetric matrix:

$$M_1^{-1}KM_1^{-T}u = M_1^{-1}b, \;\; x := M_1^{-T}u.$$

We note that a factorization of M of the form (1.15) need not be explicitly available;

see [78, Chapter 9].

Preconditioners can be purely algebraic, based on no prior knowledge of the underlying problem (such as incomplete factorizations or some steps of a stationary iterative method); or they can be made for a specific problem based on knowledge of the properties of the matrix or the problem from which is arises. The latter situation is common in the numerical solution of PDEs: effective preconditioners are often derived by reasoning about the physical properties of the continuous problem, or from known mathematical properties of the continuous or discrete operators. We refer to the survey paper of [92] for more details. It is also possible to use a flexible iterative solver such as FGMRES [77], which allows the preconditioned operator to change at each iteration. A common use for this is to use another iterative method (such as conjugate gradient) set to a low tolerance for some inner solves.

1.3.2 Block preconditioning

Our focus in this thesis is on block matrices; specifically, the double and classical saddle-point matrices $\mathscr{K}$, $\mathscr{K}_0$, $\mathscr{A}$ and $\mathscr{A}_0$, defined in (1.2), (1.8), (1.3), and (1.5), respectively. Monolithic preconditioners, which work on the entire matrix, have been recently shown to be extremely effective. Recent work such as [1] has pushed the envelope towards scalable solvers based on this methodology. Another increasingly important approach is operator preconditioning based on continuous spaces; see [44, 56]. This approach relies on the properties of the underlying continuous differential operators, and uses tools such as Riesz representation and natural norm considerations to derive block diagonal preconditioners. Block preconditioners can also be derived directly by linear algebra considerations accompanied by properties of discretized PDEs; see, for example, [17, 68].

We focus on block preconditioning: that is, preconditioners that are themselves block matrices. These are typically derived in two steps. First, we begin with an *ideal* preconditioner: this is often derived abstractly from a decomposition or inverse formula of the original matrix. With ideal preconditioners, we can derive spectral properties of the preconditioned operator and sometimes derive upper bounds on the number of preconditioned iterations an iterative solver will need to

converge. But there is a downside here, which is that ideal preconditioners tend to involve computationally expensive terms – such as inverses of some of the blocks, Schur complements [60], null spaces [19], or reduced Hessians [67] – that make them too costly to apply in practice. So from the ideal preconditioner, we must develop a *practical* preconditioner, which approximates the expensive terms in the ideal preconditioner. How this is done is typically informed by the problem at hand.

1.3.3 Review of ideal block preconditioners

We begin by reviewing ideal preconditioners for the classical saddle-point matrix $\mathscr{A}_0$. Kuznetsov [53] and Murphy, Golub, and Wathen [60] show that, when A is positive definite and B has full row rank, the preconditioner:

$$\mathscr{P}_{MGW} = \begin{bmatrix} A & 0 \\ 0 & BA^{-1}B^T \end{bmatrix} \tag{1.16}$$

has the property that the preconditioned operator $\mathscr{P}_{MGW}^{-1}\mathscr{A}_0$ has three distinct eigenvalues (equal to 1 and $\frac{1\pm\sqrt{5}}{2}$), meaning that a preconditioned iterative solver (such as MINRES) will converge within 3 iterations, in exact arithmetic.

For the stabilized classical saddle-point matrix $\mathscr{A}$, Ipsen [49] shows that the block triangular preconditioner:

$$\mathscr{P}_I = \begin{bmatrix} A & B^T \\ 0 & -(D+BA^{-1}B^T) \end{bmatrix} \tag{1.17}$$

yields a preconditioned operator with minimal polynomial $(\lambda-1)^2$; therefore, a preconditioned iterative solver (such as GMRES) will converge within 2 iterations, in exact arithmetic. We note that the same result holds for nonsymmetric saddle-point matrices.

The case in which A is singular has been less explored. Greif and Schötzau [39] show that, when A has rank $n-m$, the preconditioner

$$\mathscr{P}_{GS} = \begin{bmatrix} A+B^TWB & 0 \\ 0 & W \end{bmatrix},$$

where $W \in \mathbb{R}^{m\times m}$ is positive definite, yields two distinct eigenvalues (1 and -1) for the preconditioned operator $\mathscr{P}_{GS}^{-1}\mathscr{A}_0$.

For double saddle-point systems with positive definite A, the block diagonal preconditioner

$$\mathscr{M} := \begin{bmatrix} A & 0 & 0 \\ 0 & S_1 & 0 \\ 0 & 0 & S_2 \end{bmatrix}, \tag{1.18}$$

where

$$S_1 = D + BA^{-1}B^T; \quad S_2 = E + CS_1^{-1}C^T, \tag{1.19}$$

is well-defined and positive definite when both S_1 and S_2 are positive definite. The recent papers of Sogn and Zulehner [84] and Cai et al. [11] both analyze the performance of a block-$n \times n$ block diagonal preconditioner analogous to $\mathscr{M}$ defined in (4.1), though the former focus on the spectral properties of the continuous (rather than discretized) preconditioned operator, and Cai et al. focus their analyses of this preconditioner on the case where all diagonal blocks except A are zero. We will provide a detailed analysis of this preconditoner in Chapter 4.

Cai et al. [11] also derive a block triangular preconditioner for double saddle-point systems, analogous to (1.17):

$$\mathscr{M}_T := \begin{bmatrix} A & B^T & 0 \\ 0 & -S_1 & C^T \\ 0 & 0 & S_2 \end{bmatrix}, \tag{1.20}$$

and show that the minimal polynomial of the preconditioned operator is $(\lambda - 1)^3$. As was the case for the 2×2-block triangular preconditioner, this same result holds when $\mathscr{K}$ is nonsymmetric.

To our knowledge, there have been no preconditioners designed for double saddle-point systems $\mathscr{K}$ when A is singular.

1.3.4 From ideal to practical

All the preconditioners described in Section 1.3.3 are too expensive to be applied in practice. To develop a practical solver, it is necessary to develop suitable approx-

imations to the expensive terms (A-inversions, Schur complements) that arise in these preconditioners. For problems arising from the numerical solution of PDEs, a common approach is to replace an expensive differential operator arising in the ideal preconditioner by a cheaper, spectrally equivalent one; see, for example, [9, 39, 47, 52, 93]. Another common approach is to approximate the inversion of some blocks using an inner iterative solver such as a multigrid method; see, for example, [54, 82, 95].

1.4 Eigenvalue analysis

A primary focus of this thesis is generating eigenvalue bounds for block matrices. We employ three primary techniques: *energy estimates*; the *reduced matrix* (or *R-matrix*) method; and the *diagonal matrix* (or *D-matrix*) method. Energy estimates have been widely used to generate eigenvalue bounds for block matrices, most famously by Rusten and Winther [76]. The R-matrix technique is a newer approach. It was used by Sogn and Zulehner [84, Theorem 2.2] to bound the eigenvalues of preconditioned tridiagonal block-$k \times k$ systems (though they do not call it the "R-matrix method"). The D-matrix technique is a variation of the R-matrix technique, also used by Sogn and Zulehner in the same problem setting [84, Theorem 2.1] (though, again, "D-matrix" is our own name). We do not use the D-matrix technique in deriving any of the eigenvalue bounds in this thesis but we include the description here. We will describe here how these methods can be generalized to obtain eigenvalue bounds of any block matrix.

1.4.1 Description of analysis methods

We begin by providing a high-level explanation of all three techniques. For clarity of exposition, we consider deriving an upper bound on positive eigenvalues of a block matrix (lower bounds on negative eigenvalues and internal bounds – i.e., lower bounds on positive eigenvalues and upper bounds on negative eigenvalues – will be discussed later).

Consider a symmetric block-$k \times k$ matrix

$$\mathscr{M} = \begin{bmatrix} A_{11} & A_{21}^T & \cdots & A_{k1}^T \\ A_{21} & A_{22} & & A_{k2}^T \\ \vdots & & \ddots & \vdots \\ A_{k1} & A_{k2} & \cdots & A_{kk} \end{bmatrix},$$

and let $v = \begin{bmatrix} x_1^T & x_2^T & \cdots & x_k^T \end{bmatrix}^T$ be an appropriately partitioned block vector.

Energy estimates With energy estimates, we write out the eigenvalue equations for $\mathscr{M}$:

$$\begin{aligned} A_{11}x_1 + A_{21}^T x_2 + \ldots + A_{k1}^T x_k &= \lambda x_1; \\ &\vdots \\ A_{k1}x_1 + A_{k2}x_2 + \ldots + A_{kk}x_k &= \lambda x_k. \end{aligned}$$

We then perform various manipulations on the eigenvalue equations – often including substituting one variable for another, taking inner products of x_i^T with equation i, using extremal singular values or eigenvalues of blocks to bound terms of the form $x_i^T A_{ij} x_j$, dropping additive terms, and other manipulations – to obtain an inequality involving λ. The inequality is often of the form $p(\lambda) \geq 0$ or $p(\lambda) \leq 0$, where p is a polynomial of degree k or less. This then gives us a bound that is a root of the polynomial p.

R-matrix With the *R*-matrix technique, we seek to bound the quantity $\frac{v^T \mathscr{M} v}{v^T v}$ from above, which provides an upper bound on the eigenvalues of $\mathscr{M}$. We write

$$v^T \mathscr{M} v = \sum_{i=1}^{k} \sum_{j=1}^{k} x_i^T A_{ij} x_j. \tag{1.21}$$

Letting $\lambda_{\max}^{ii}$ denote the maximal eigenvalue of A_{ii} and $\sigma_{\max}^{ij}$ the maximal singular value of A_{ij}, we can bound each term of (1.21) by

$$x_i^T A_{ij} x_j \leq \begin{cases} \lambda_{\max}^{ii} ||x_i||^2 & \text{if } i = j \\ \sigma_{\max}^{ij} ||x_i|| \cdot ||x_j|| & \text{otherwise,} \end{cases} \tag{1.22}$$

with the second case holding as a result of the Cauchy-Schwarz inequality. We then write

$$v^T \mathscr{M} v \leq \begin{bmatrix} ||x_1|| & ||x_2|| & \cdots & ||x_k|| \end{bmatrix} \underbrace{\begin{bmatrix} \lambda_{\max}^{11} & \sigma_{\max}^{12} & \cdots & \sigma_{\max}^{1k} \\ \sigma_{\max}^{12} & \lambda_{\max}^{22} & & \sigma_{\max}^{2k} \\ \vdots & & \ddots & \vdots \\ \sigma_{\max}^{1k} & \sigma_{\max}^{2k} & \cdots & \lambda_{\max}^{kk} \end{bmatrix}}_{=:R} \begin{bmatrix} ||x_1|| \\ ||x_2|| \\ \vdots \\ ||x_k|| \end{bmatrix}. \tag{1.23}$$

An upper bound on $\frac{v^T \mathscr{M} v}{v^T v}$ – and, therefore, an upper bound on the eigenvalues of $\mathscr{M}$ – is given by the maximal eigenvalue of R, which we call the "reduced matrix" because it is $k \times k$ (instead of block-$k \times k$). As before, the resulting bound is the root of a polynomial of degree k or less[1] – specifically, the characteristic polynomial of R.

D-matrix In the D-matrix technique we modify the inequality in (1.22) so that the resulting reduced matrix (shown in (1.23)) is diagonal. We further bound the off-diagonal terms (where $i \neq j$) using the Peter-Paul version of Young's inequality [90, Chapter 6] to obtain:

$$\sigma_{\max}^{ij} ||x_i|| \cdot ||x_j|| \leq \frac{\sigma_{\max}^{ij} \varepsilon_{ij}}{2} ||x_i||^2 + \frac{\sigma_{\max}^{ij}}{2\varepsilon_{ij}} ||x_j||^2,$$

[1] We can easily express the bound as the root of a polynomial of degree strictly less than k if R is singular.

for any $\varepsilon_{ij} > 0$. Then we can write

$$v^T \mathscr{M} v \leq \begin{bmatrix} ||x_1|| & \dots & ||x_k|| \end{bmatrix} R_D \begin{bmatrix} ||x_1|| \\ \vdots \\ ||x_k|| \end{bmatrix},$$

where R_D is a $k \times k$ diagonal matrix whose entries depend on μ_{ii}, σ_{ij}, ε_{ij}. We then select the values of ε_{ij} that yield the tightest upper bound: specifically, we wish to minimize the largest eigenvalue of the reduced matrix R_D, which amounts to selecting ε_{ij} such that all entries of R_D are equal.

1.4.2 Example: re-deriving a bound of Rusten and Winther

We now show how to use energy estimates, the *R*-matrix technique, and the *D*-matrix technique to derive an upper bound on the eigenvalues of the standard classical saddle-point matrix:

$$\mathscr{A}_0 = \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix}.$$

Energy estimates This proof is given in Rusten and Winther [76], but we include it here as a simple example of using energy estimates. The eigenvalue equations for $\mathscr{A}_0$ are

$$Ax + B^T y = \lambda x \tag{1.24a}$$

$$Bx = \lambda y. \tag{1.24b}$$

We can use (1.24b) to write y in terms of x:

$$y = \frac{1}{\lambda} Bx.$$

We substitute this value into (1.24a) to obtain

$$Ax + \frac{1}{\lambda} B^T Bx = \lambda x. \tag{1.25}$$

We then take the inner product of x^T with (1.25), divide each term by $x^T x$, and bound each of the left-hand terms from above to give

$$\mu^A_{\max} + \frac{(\sigma^B_{\max})^2}{\lambda} \geq \lambda,$$

where $\mu^A_{\max}$ denotes the largest eigenvalue of A and $\sigma^B_{\max}$ the largest singular value of B. For $\lambda > 0$, rearranging yields

$$\lambda^2 - \mu^A_{\max}\lambda - (\sigma^B_{\max})^2 \leq 0,$$

from which we conclude that

$$\lambda \leq \frac{\mu^A_{\max} + \sqrt{\mu^A_{\max} + 4(\sigma^B_{\max})^2}}{2}.$$

R-matrix We write

$$\begin{aligned} v^T \mathscr{A}_0 v &= x^T A x + 2x^T B y \\ &\leq \mu^A_{\max} ||x||^2 + 2\sigma^B_{\max} ||x|| \cdot ||y|| \\ &= \begin{bmatrix} ||x|| & ||y|| \end{bmatrix} \underbrace{\begin{bmatrix} \mu^A_{\max} & \sigma^B_{\max} \\ \sigma^B_{\max} & 0 \end{bmatrix}}_{=:R} \begin{bmatrix} ||x|| \\ ||y|| \end{bmatrix}. \end{aligned}$$

An upper bound on λ is given by the larger eigenvalue of R, which is equal to $\frac{\mu^A_{\max} + \sqrt{\mu^A_{\max} + 4(\sigma^B_{\max})^2}}{2}$, which is the same result given by energy estimates.

D-matrix We write

$$\begin{aligned} v^T \mathscr{A}_0 v &= x^T A x + 2x^T B^T y \\ &\leq \mu^A_{\max} ||x||^2 + 2\sigma^B_{\max} ||x|| \cdot ||y|| \\ &\leq \mu^A_{\max} ||x||^2 + \frac{\sigma^B_{\max}}{\varepsilon} ||x||^2 + \varepsilon \sigma^B_{\max} ||y||^2 \qquad \forall \varepsilon > 0 \\ &= \begin{bmatrix} ||x|| & ||y|| \end{bmatrix} \underbrace{\begin{bmatrix} \mu^A_{\max} + \frac{\sigma^B_{\max}}{\varepsilon} & 0 \\ 0 & \varepsilon \sigma^B_{\max} \end{bmatrix}}_{R_D} \begin{bmatrix} ||x|| \\ ||y|| \end{bmatrix}. \end{aligned}$$

The eigenvalues of $\mathscr{A}_0$ are less than or equal to the maximum eigenvalue of R_D. We set

$$\varepsilon = \frac{\mu^A_{\max} + \sqrt{(\mu^A_{\max})^2 + 4(\sigma^B_{\max})^2}}{2\sigma^B_{\max}}.$$

Then

$$R_D = \begin{bmatrix} \frac{\mu^A_{\max} + \sqrt{(\mu^A_{\max})^2 + 4(\sigma^B_{\max})^2}}{2} & 0 \\ 0 & \frac{\mu^A_{\max} + \sqrt{(\mu^A_{\max})^2 + 4(\sigma^B_{\max})^2}}{2} \end{bmatrix},$$

which again yields the same bound.

1.4.3 Extension to other bounds and discussion

The previous example showed how to use energy estimates and the *R*-matrix and *D*-matrix techniques to derive an upper bound on the positive eigenvalues of a block matrix. Lower bounds on negative eigenvalues can be obtained in a similar way. Interior eigenvalue bounds – that is, lower bounds on positive eigenvalues and upper bounds on negative eigenvalues – are less straightforward. Energy estimates often work well for this, though in practice these bounds are harder to obtain than the extremal bounds. The *R*/*D*-matrix methods, on the other hand, are not well-suited for finding internal bounds without significant adaptations. If we are working with a matrix whose block inverse has a fairly simple structure, one option is to form the inverse and find its extremal bounds (using any method of choice) – see [84, proof of Theorem 2.2] for an example of this combined with the *R*-matrix technique to compute interior bounds.

Compared with energy estimates, the benefit of the *R*-matrix method is its sim-

plicity. This is a particular advantage with matrices with many blocks, which can become intractably complex for energy estimates. The drawbacks are that, as mentioned, interior bounds are difficult with the R-matrix technique, and that the R-matrix technique can sometimes yield loose bounds. The R-matrix technique considers each block of the matrix independently, so it may not yield a tight bound in cases where relationships exist between the blocks (as often occurs in, for example, preconditioned matrices). We can sometimes obtain a tighter bound in these casees by using the D-matrix technique instead, though the D-matrix technique is sensitive to the exact configuration of ε_{ij} parameters in the inequality. For instance, in the example of Section 1.4.2, if we instead write that

$$v^T \mathscr{A}_0 v \leq 2\sigma_B^{\max} ||x|| \cdot ||y|| \leq \varepsilon \sigma_B^{\max} ||x||^2 + \frac{\sigma_B^{\max}}{\varepsilon} ||y||^2,$$

we will obtain a loose bound. Like energy estimates, the D-matrix technique can also become very complicated for large block matrices because of the number of ε_{ij} parameters to solve for.

1.5 Notation

Scalars, vectors, matrices, and block matrices In this thesis, scalars will be denoted by lower case Greek or Roman letters. Vectors are denoted by Roman letters (most commonly v, x, y, z). We denote matrices by upper case Roman letters (e.g., A, B), and use calligraphic fonts (e.g., $\mathscr{M}$, $\mathscr{K}$) to refer to block matrices.

Some results in this thesis deal with double saddle-point matrices as defined in (1.2), while others deal with classical saddle-point matrices as in (1.3). To reduce ambiguity, we will denote double saddle-point matrices by the calligraphic letter $\mathscr{K}$ and classical saddle-point matrices by the calligraphic letter $\mathscr{A}$. Similarly, block-3×3 preconditioners for double saddle-point matrices will be demoted by a calligraphic $\mathscr{M}$ and block-2×2 preconditioners for classical saddle-point matrices will be denoted by a calligraphic $\mathscr{P}$.

Eigenvalues and singular values The eigenvalues of the unpreconditioned and preconditioned block matrices will be denoted by λ. We will denote the eigenval-

ues of a matrix $M \in \mathbb{R}^{n\times n}$ by

$$\mu_i(M), \quad i = 1,\ldots,n,$$

and in terms of ordering we will assume that

$$\mu_1(M) \geq \mu_2(M) \geq \cdots \geq \mu_n(M).$$

We follow the same convention for singular values of a rectangular matrix N, but we use σ rather than μ: i.e., the the singular values of $N \in \mathbb{R}^{m\times n}$ are denoted by

$$\sigma_1(N) \geq \sigma_2(N) \geq \cdots \geq \mu_m(N) \geq 0.$$

To increase clarity, we will often refer to the maximal eigenvalues/singular values $\mu_1(M)$ and $\sigma_1(N)$ by $\mu_{\max}(M)$ and $\sigma_{\max}(M)$ respectively. Similarly, we will refer to the minimal values $\mu_n(M)$, $\sigma_m(N)$ by $\mu_{\min}(M)$ and $\sigma_{\min}(M)$. The positive eigenvalues of a matrix will be denoted by an additional "+" subscript – for instance, we denote the smallest nonzero eigenvalue of a semidefinite matrix M by $\mu_{\min,+}(M)$.

For compactness, we will use superscripts rather than parentheses when referring to eigenvalues/singular values of the matrices comprising $\mathscr{K}$; for example, we will write $\mu^A_{\min,+}$ rather than $\mu_{\min,+}(A)$ to refer to the smallest positive eigenvalue of A. Based on this convention, we use the notation of Table 1.1 for eigenvalues and singular values of the matrices comprising $\mathscr{K}$.

matrix	size	type	number	notation	max	min
A	$n\times n$	eigenvalues	n	$\mu^A_i,\ i=1,\ldots,n$	$\mu^A_{\max}$	$\mu^A_{\min}$
B	$m\times n$	singular values	m	$\sigma^B_i,\ i=1,\ldots,m$	$\sigma^B_{\max}$	$\sigma^B_{\min}$
C	$p\times m$	singular values	p	$\sigma^C_i,\ i=1,\ldots,p$	$\sigma^C_{\max}$	$\sigma^C_{\min}$
D	$m\times m$	eigenvalues	m	$\mu^D_i,\ i=1,\ldots,m$	$\mu^D_{\max}$	$\mu^D_{\min}$
E	$p\times p$	eigenvalues	p	$\mu^D_i,\ i=1,\ldots,p$	$\mu^E_{\max}$	$\mu^E_{\min}$

Table 1.1: Summary of notation for eigenvalues and singular values of matrix blocks.

Finally, in some parts of our analyses we denote a symmetric positive semidefinite matrix X by $X \succeq 0$.

1.6 Outline and contributions

This thesis is comprised of seven chapters. In Chapter 2, we provide an overview of applications that lead to double and classical saddle-point matrices. In Chapter 3, we provide a general framework for analysis of the eigenvalues of double saddle-point matrices, with a rather minimal set of assumptions on the matrices involved. We also derive upper and lower bounds on the positive and negative eigenvalues of $\mathscr{K}$. We provide some numerical illustrations to illustrate the tightness of the bounds.

In Chapter 4 we consider preconditioning of double saddle-point matrices when the leading block A is positive definite. We provide eigenvalue bounds on a block diagonal Schur complement-based preconditioner. We also analyze the case in which the preconditioner contain approximations of Schur complements and an approximation of the leading block, and show how our bounds are affected as a result. Some numerical observations indicate that our bounds are tight and effective.

In Chapter 5 we provide eigenvalue bounds for classical saddle-point systems in which A is singular. The challenge in this setting is the lower bound on positive eigenvalues, for which existing eigenvalue bounds reduce to zero. Our general approach involves augmenting the leading block to obtain this bound. The resulting bounds rely on the angles between the kernels of A and B. We then present some numerical results to validate our findings.

In Chapter 6 we consider preconditioning for classical and double saddle-point matrices for which the leading block A is singular. The approach involves a combination of leading block augmentation and Schur complement preconditioning. We will show how "minimally augmenting" the leading block yields a preconditioner with a constant number of eigenvalues in the preconditioned operator in both the classical and double saddle-point case. We also provide a set of numerical experiments for the classical saddle-point preconditioner that illustrate the effectiveness of our approach.

In Chapter 7 we provide concluding remarks and discuss areas for future work.

The main contributions of this thesis are:

1. The derivation of eigenvalue bounds for the unpreconditioned double saddle-point matrix $\mathscr{K}$. Existing analyses of the eigenvalues of unpreconditioned

block-3×3 matrices have often been restricted to specific problems, such as interior-point methods in constrained optimization; see [40, 58].

2. An analysis of the block diagonal multiple saddle-point preconditioner provided in [84], with new analysis in the discretized setting. We also provide an analysis in the case where approximations to the leading block and Schur complements are used.

3. New eigenvalue analyses for classical saddle-point matrices with a singular leading block.

4. A new ideal preconditioner for classical saddle-point matrices with a singular leading block. This preconditioner can be viewed as an extension of [34, 39], with the property that the preconditioned operator has a constant number of eigenvalues regardless of the rank of A.

5. A new ideal preconditioner for double saddle-point matrices with a singular leading block. To our knowledge, this is the first preconditioner that has been designed for this setting.

Chapter 2

Relevant applications

Double saddle-point systems of the form (1.2) arise in several applications. We provide an overview of these applications in Section 2.1. Because Chapter 5 and Chapter 6 of this thesis deal in part with classical saddle-point systems with singular leading blocks, we also discuss these in Section 2.2.

2.1 Double saddle-point systems

2.1.1 Examples arising from optimization

PDE-constrained optimization Consider a discretized linear-quadratic optimization problem of the form:

$$\begin{aligned} &\min_{y,u} \frac{1}{2} y^T C y - y^T w + \frac{\beta}{2} u^T R u \\ &\text{subject to } Ky + Lu = d, \end{aligned} \tag{2.1}$$

where $K \in \mathbb{R}^{n\times n}$ is a stiffness matrix corresponding to a partial differential equation (PDE); $L \in \mathbb{R}^{n\times m}$ is a control matrix; $C \in \mathbb{R}^{n\times n}$ is a positive semidefinite (sometimes positive definite) observation matrix; $R \in \mathbb{R}^{m\times m}$ is a positive definite regularization matrix; and $\beta > 0$ is a regularization parameter (often around 10^{-2} in practice). The vector $y \in \mathbb{R}^n$ denotes the state variables, $u \in \mathbb{R}^m$ the control variables, and $\lambda \in$

$\mathbb{R}^n$ the Lagrange multipliers. The associated Karush-Kuhn-Tucker (KKT) system can be written as a classical saddle-point system:

$$\begin{bmatrix} C & 0 & K^T \\ 0 & \beta R & L^T \\ K & L & 0 \end{bmatrix} \begin{bmatrix} y \\ u \\ \lambda \end{bmatrix} = \begin{bmatrix} w \\ 0 \\ d \end{bmatrix}. \tag{2.2}$$

We can also reorder the unknowns so that the coefficient matrix is in double saddle-point form (1.2):

$$\begin{bmatrix} C & K^T & 0 \\ K & 0 & L \\ 0 & L^T & \beta R \end{bmatrix} \begin{bmatrix} y \\ \lambda \\ u \end{bmatrix} = \begin{bmatrix} w \\ d \\ 0 \end{bmatrix}. \tag{2.3}$$

Preconditioners based on the classical saddle-point formulation (2.2) have been developed in, for example, [51, 66, 71, 80], while preconditioners based on the double saddle-point formulation (2.3) have been developed in [5, 65, 84]. In many cases, the preconditioners are designed for a restricted type of PDE-constrained optimization problem. For instance, it is common to assume that the operator corresponding to K is elliptic [13, 66, 71, 80, 84]. In this case, the matrix K is singular if the underlying PDE has pure Neumann boundary conditions, and positive definite otherwise. Some papers [13, 66, 71] focus primarily on what Kouri et al. [51] refer to as the "idealized case", in which $m = n$ and $C = L = R = M$, where M is a positive definite mass matrix. This case corresponds to distributed control and observations, with identical discretizations for the state and control variables.

Interior point methods Consider the nonlinear optimization problem

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & g(x) \leq 0, \end{aligned} \tag{2.4}$$

where we assume that f and g are convex and twice-differentiable. The interior point algorithm described for this problem in [8] requires solving, at each Newton

iteration, a linear system of the form:

$$\begin{bmatrix} H(x,y) & J(x)^T & 0 \\ J(x) & 0 & I \\ 0 & Z & Y \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta y \\ \Delta z \end{bmatrix} = \begin{bmatrix} -\nabla f(x) - J(x)^T y \\ -g(x) - z \\ \mu e - YZe \end{bmatrix}. \tag{2.5}$$

Here H and J are Hessian and Jacobian matrices, respectively, that change at each iteration, and Y and Z are positive diagonal matrices whose diagonal entries consist of the current iterates of the variables y and z, and here I is the appropriately sized identity matrix. As the iterations proceed, some of the entries approach zero. The matrix in (2.5) can be symmetrized using a diagonal similarity transformation:

$$\begin{aligned} \mathscr{K} &= \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & Z^{1/2} \end{bmatrix}^{-1} \begin{bmatrix} H(x,y) & J(x)^T & 0 \\ J(x) & 0 & I \\ 0 & Z & Y \end{bmatrix} \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & Z^{1/2} \end{bmatrix} \\ &= \begin{bmatrix} H(x,y) & J(x)^T & 0 \\ J(x) & 0 & Z^{1/2} \\ 0 & Z^{1/2} & Y \end{bmatrix}. \end{aligned} \tag{2.6}$$

When (2.4) is a quadratic program (QP) [61, Chapter 16], the matrices H and J are constant. We can write this problem in standard form as:

$$\begin{aligned} &\min_x \; c^T x + \frac{1}{2} x^T H x \quad \text{s.t.} \quad Jx = b, x \geq 0, \\ &\max_{x,y,z} \; b^T y - \frac{1}{2} x^T H x \quad \text{s.t.} \quad J^T y + z - Hx = c, z \geq 0, \end{aligned} \tag{2.7}$$

where inequalities are understood elementwise and y and z are vectors of the Lagrange multipliers. At each iteration of the interior point method [61] for a QP with regularization discussed in [26, 40], the matrix to be solved can be written (after reordering) as

$$\mathscr{K} = \begin{bmatrix} -X & -Z^{1/2} & 0 \\ -Z^{1/2} & H + \rho I & J^T \\ 0 & J & -\delta I \end{bmatrix}, \tag{2.8}$$

where $\delta, \rho \geq 0$ are regularization parameters and X and Z are non-negative diag-

onal matrices (where, again, some of the entries approach zero as the iterations progress). The case in which $H = 0$ corresponds to a linear program (LP) in standard form. The matrix in (2.8) is equivalent to the double saddle-point formulation (1.2) up to a difference in sign.

Constrained weighted least-squares Consider the least-squares problem with linear equality constraints (see [7, sec. 2.2]):

$$\begin{aligned} \min_{y} \ & ||c - Gy||_2 \\ \text{s.t.} \ & Ey = d. \end{aligned}$$

The optimality conditions for this problem are

$$\begin{bmatrix} I & G & 0 \\ G^T & 0 & E^T \\ 0 & E & 0 \end{bmatrix} \begin{bmatrix} r \\ y \\ \lambda \end{bmatrix} = \begin{bmatrix} c \\ 0 \\ d \end{bmatrix}, \tag{2.9}$$

where λ is a vector of Lagrange multipliers.

Geophysical inverse problems Here we consider the example of a geophysical inverse problem described in [41], which involves recovering a model based on observations of a field. The regularized problem is defined by

$$\begin{aligned} \min_{m,u} \ & \frac{1}{2}||Qu - b||^2 + \frac{\beta}{2}||W(m - m_{ref})||^2 \\ \text{s.t.} \ & A(m)u = q, \end{aligned}$$

where β is a regularization parameter, m is a model, m_{ref} is a reference model, W is a weighting matrix, and A is a large, sparse, nonsingular matrix that encodes the model conditions of the field being considered. If Gauss-Newton iterations are

used, the linear system to be solved at each step takes the form

$$\begin{bmatrix} Q^TQ & A^T & 0 \\ A & 0 & G \\ 0 & G^T & \beta W^TW \end{bmatrix} \begin{bmatrix} \delta u \\ \delta\lambda \\ \delta m \end{bmatrix} = - \begin{bmatrix} r_u \\ r_\lambda \\ r_m \end{bmatrix}, \tag{2.10}$$

where G is the Jacobian of A and r_u, r_λ and r_m are the discrete residuals arising from the optimality conditions on the variables u, λ, and m, respectively. In the typical case of sparse observations, G is sparse and Q^TQ has high nullity.

2.1.2 Examples arising from the numerical solution of partial differential equations

Dual-dual finite element formulations Dual-dual finite element methods, as described in [30], solve linear second-order elliptic equations in divergence form by introducing the gradient as an explicit unknown. The matrices associated with the resulting linear systems have the form

$$\begin{bmatrix} A & B_1^T & 0 \\ B_1 & 0 & B^T \\ 0 & B & 0 \end{bmatrix}, \tag{2.11}$$

where A is positive definite, and B_1 and B have full row rank. We refer to [29] for an example in hyperelasticity, and to [28] for an extension of Bramble-Pasciak CG to these systems.

Magma mantle dynamics Rhebergen et al. [72] consider two-phase flow equations for magma/mantle dynamics on a domain in $\Omega \subset \mathbb{R}^d$, where $1 \leq d \leq 3$. This leads to the system of equations:

$$
\begin{aligned}
-\nabla\cdot\left(\eta(\mathbf{D}u-\frac{1}{3}\nabla\cdot u)\right)+\nabla p+\nabla p_c &= \phi\mathbf{e}_3\\
-\nabla\cdot u+\nabla\cdot k\nabla p &= \nabla\cdot k\mathbf{e}_3,\\
-\nabla\cdot u-\zeta^{-1}p_c &= 0,
\end{aligned}
$$

where: u is the matrix velocity; $\mathbf{D}u$ is the total strain rate; p is the dynamic pressure; $p_c=-\eta\nabla\cdot u$ is an auxiliary variable; $\mathbf{e}_3$ is the unit vector in the direction aligned with gravity; ϕ is the porosity; η is the shear viscosity; ζ is the bulk viscosity; k is the permeability; and $\alpha=\zeta/\eta-1/3$. Decomposing the boundary of the domain by $\Gamma_D\cup\Gamma_N=\nabla\Omega$ where $\Gamma_D\cap\Gamma_N=\emptyset$, the boundary conditions are given by

$$
\begin{aligned}
u &= g \ \text{ on } \Gamma_D,\\
\eta\mathbf{D}u\cdot n-\left(\frac{1}{3}\eta\nabla\cdot u+p+p_c\right)\mathbf{n} &= g_N \ \text{ on } \Gamma_N,\\
-k(\nabla p-\mathbf{e}_3)\cdot\mathbf{n} &= 0 \ \text{ on } \Delta\Omega,
\end{aligned}
$$

where g and g_N are given boundary data and $\mathbf{n}$ is an outward unit normal vector.

The discrete problem then reads

$$
\begin{bmatrix} \eta K & G^T & G^T\\ G & -kC & 0\\ G & 0 & -\zeta^{-1}Q\end{bmatrix}\begin{bmatrix} u\\ p\\ p_c\end{bmatrix}=\begin{bmatrix} f\\ g\\ 0\end{bmatrix},
$$

which we can reorder as

$$
\begin{bmatrix} -kC & G & 0\\ G^T & \eta K & G^T\\ 0 & G & -\zeta^{-1}Q\end{bmatrix}\begin{bmatrix} p\\ u\\ p_c\end{bmatrix}=\begin{bmatrix} g\\ f\\ 0\end{bmatrix}, \tag{2.12}
$$

Here K and Q are positive definite, C is positive semidefinite, and G is assumed to be rank-deficient by 1. Note that the matrix in (2.12) is equivalent to the one in (1.2), up to a difference in sign.

Boundary element tearing and interconnecting methods The inexact data-sparse version of the boundary element tearing and interconnecting (BETI) method, as presented by Langer et al. [55], results in a linear system of the form

$$\begin{bmatrix} V & -K & 0 \\ -K^T & -D & B^T \\ 0 & B & 0 \end{bmatrix} \begin{bmatrix} t \\ u \\ \lambda \end{bmatrix} = \begin{bmatrix} g \\ f \\ 0 \end{bmatrix}. \tag{2.13}$$

Here V, D, and K are all block diagonal matrices whose blocks correspond to discretized boundary integral operators for each element. The matrix V is positive definite, while D generally has high nullity, as many of the blocks along the diagonal of D are singular. The null space of each singular block of D is spanned by the vector $(1,1,\ldots,1)^T$. The matrix B enforces the continuity of the local potential vectors across subdomain boundaries. In preconditioning (2.13), Langer et al. use a regularization technique to make D nonsingular.

Darcy-Stokes Following the formulation of [10], the linear system associated with the Darcy-Stokes equations can be written in block form as:

$$\underbrace{\begin{bmatrix} A_p & A_\Gamma^T & 0 \\ -A_\Gamma & A_f & B_f^T \\ 0 & B_f & 0 \end{bmatrix}}_{=:\mathscr{K}_{DS}} \begin{bmatrix} \phi_h \\ \mathbf{u}_h \\ p_h \end{bmatrix} = \begin{bmatrix} f_{p,h} \\ \mathbf{f}_{f,h} \\ g_h \end{bmatrix}, \tag{2.14}$$

where ϕ_h is the piezometric head (essentially a scaled and shifted version of the Darcy pressure variable), and $\mathbf{u}_h$ and p_h are the Stokes velocity and pressure, respectively. In the block matrix, A_p is the (SPD) Darcy matrix, A_f is the vector Laplacian (also SPD), A_Γ is the coupling matrix encoding the interface conditions, and B_f is a discrete divergence operator.

In its original formulation this matrix is nonsymmetric and has a positive (rather than negative) definite $(2,2)$-block. However, we can symmetrize the system (and make it conform to our assumptions on positive vs. negative diagonal blocks) by

writing

$$\underbrace{\begin{bmatrix} A_p & -A_\Gamma^T & 0 \\ -A_\Gamma & -A_f & -B_f^T \\ 0 & -B_f & 0 \end{bmatrix}}_{=:\tilde{\mathscr{K}}_{DS}} \begin{bmatrix} \phi_h \\ -\mathbf{u}_h \\ -p_h \end{bmatrix} = \begin{bmatrix} f_{p,h} \\ \mathbf{f}_{f,h} \\ g_h \end{bmatrix}.$$

We refer also to [12] for further details on preconditioning.

Poromechanics Ferronato et al. [21] present a formulation for coupled poromechanical equations that leads to a (reordered) Jacobian matrix of the form

$$\mathscr{K}_P = \begin{bmatrix} K & -Q & 0 \\ Q^T & P & \gamma B^T \\ 0 & -B & A \end{bmatrix},$$

where: K is an SPD elastic stiffness matrix; P is a diagonal SPD matrix representing the fluid flow capacity; A is an SPD mass matrix for Darcy's velocity in mixed form; Q and B are rectangular blocks coupling displacements and Darcy's velocities to the pressure unknowns; and γ is a positive parameter equal to either Δt or $0.5\Delta t$, where Δt is the integration timestep.

As in the Darcy-Stokes case, we can transform the matrix $\mathscr{K}_P$ into a symmetric matrix with alternating positive/negative definiteness of the diagonal blocks by applying a matrix transformation to obtain a symmetric matrix $\tilde{\mathscr{K}}_P$:

$$\tilde{\mathscr{K}}_P = \mathscr{K}_P \underbrace{\begin{bmatrix} I & 0 & 0 \\ 0 & -I & 0 \\ 0 & 0 & \frac{1}{\gamma} I \end{bmatrix}}_{=:\Lambda} = \begin{bmatrix} K & Q & 0 \\ Q^T & -P & B^T \\ 0 & B & \frac{1}{\gamma} A \end{bmatrix}.$$

2.2 Saddle-point systems with a singular leading block

Parts of this thesis deal with classical saddle-point systems with singular leading blocks. In this section, we describe some examples of applications that generate such systems. Some of these examples are, in fact, double saddle-point systems

described in Section 2.1, where the systems are re-ordered and partitioned into block-2×2 systems.

Time-harmonic Maxwell equations In the time-harmonic Maxwell equations in lossless media with perfectly conducting boundaries and constant coefficients, the problem is to find the vector field u and multiplier p such that

$$\begin{aligned} \nabla \times \nabla \times u + \nabla p &= f \ \text{ in } \ \Omega, \\ \nabla \cdot u &= 0 \ \text{ in } \ \Omega, \\ u \times n &= 0 \ \text{ on } \ \partial\Omega, \\ p &= 0 \ \text{ on } \ \partial\Omega, \end{aligned}$$

where the domain Ω is a subset of $\mathbb{R}^2$ or $\mathbb{R}^3$. Discretizing with Nédélec finite elements for u and nodal elements for p [39, Section 2.2] yields a linear system of the form

$$\begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} u \\ p \end{bmatrix} = \begin{bmatrix} g \\ 0 \end{bmatrix}, \tag{2.15}$$

where A is a discrete curl-curl operator, B is a negative discrete divergence operator, and M is the finite element mass matrix. The matrix A satisfies $\text{nullity}(A) = m$ (a situation we refer to later in this thesis as A being *lowest-rank*), and its null space is given by the space of gradient functions.

Interior point methods While interior point methods were discussed in the previous section as an example of a double saddle-point system, we focus here particularly on the solution of a QP without regularization. Each step of a primal-dual interior point method requires solving a linear system of the form [24]:

$$\begin{bmatrix} H & J^T & -I \\ J & 0 & 0 \\ -Z & 0 & -X \end{bmatrix} \begin{bmatrix} \Delta x \\ -\Delta y \\ \Delta z \end{bmatrix} = \begin{bmatrix} -c - Hx + J^T y + z \\ b - Jx \\ XZe - \tau e \end{bmatrix}, \tag{2.16}$$

where X and Z are diagonal matrices consisting of the current x and z iterates, respectively, $\tau > 0$ is the barrier parameter, and e is a vector of ones. Because X is

diagonal, we can perform a step of block Gaussian elimination to reduce (2.16) to block-2×2 form:

$$\begin{bmatrix} H + X^{-1}Z & J^T \\ J & 0 \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} = \begin{bmatrix} -c - HxJ^T y + \tau X^{-1} e \\ b - Jx \end{bmatrix}. \tag{2.17}$$

Some entries of both X and Z approach zero as the IPM iterations proceed, so the leading block becomes increasingly ill-conditioned, with the largest magnitude entries occurring along the diagonal. Thus the leading block may become numerically singular, particularly if H is singular.

Double saddle-point systems with an all-zero (3, 3)-block We note that all double saddle-point systems with $E = 0$ can be written as a classical saddle-point system with a singular leading block. Specifically, given a linear system

$$\begin{bmatrix} A & B^T & 0 \\ B & -D & C^T \\ 0 & C & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} p \\ q \\ r \end{bmatrix},$$

we obtain a classical saddle-point system with a rank-deficient leading block by reordering the second and third block unknowns:

$$\left[\begin{array}{cc|c} A & 0 & B^T \\ 0 & 0 & C \\ \hline B & C^T & -D \end{array}\right] \begin{bmatrix} x \\ z \\ y \end{bmatrix} = \begin{bmatrix} p \\ r \\ q \end{bmatrix}.$$

Thus, of our examples from Section 2.1 the following can also be considered examples of classical saddle-point systems with a singular leading block: constrained weighted least-squares (2.9); geophysical inverse problems (2.10); dual-dual finite element formulations (2.11); boundary element tearing and interconnecting methods (2.13); and the Darcy-Stokes equations (2.14), along with some finite element formulations of the Stokes equations [18, 31].

Chapter 3

Eigenvalue bounds when A is positive definite

3.1 Inertia and solvability conditions

We first discuss the inertia and conditions for nonsingularity of $\mathscr{K}$ defined in Equation 1.2. Recall that the inertia of a matrix is the triplet denoting the number of its positive, negative, and zero eigenvalues [45, Definition 4.5.6].

Proposition 3.1. *The following conditions are necessary for $\mathscr{K}$ to be invertible:*

(i) $\ker(A) \cap \ker(B) = \{0\}$*;*

(ii) $\ker(B^T) \cap \ker(D) \cap \ker(C) = \{0\}$*;*

(iii) $\ker(C^T) \cap \ker(E) = \{0\}$.

A sufficient condition for $\mathscr{K}$ to be invertible is that A, S_1, and S_2 are invertible.

Proof. We begin with the proof of statement (i) by assuming to the contrary that the intersection of the kernels is not empty – namely, there exists a nonzero vector x such that $Ax = Bx = 0$. This would mean that the block vector $\begin{bmatrix} x^T & 0 & 0 \end{bmatrix}^T$ was a null vector of $\mathscr{K}$, which would imply that $\mathscr{K}$ was singular. Similar reasoning proves (ii) and (iii).

For the sufficient condition, we observe that when A, S_1, and S_2 are invertible we can write a block-LDL^T factorization of $\mathscr{K}$:

$$\begin{bmatrix} A & B^T & 0 \\ B & -D & C^T \\ 0 & C & E \end{bmatrix} = \begin{bmatrix} I & 0 & 0 \\ BA^{-1} & I & 0 \\ 0 & -CS_1^{-1} & I \end{bmatrix} \underbrace{\begin{bmatrix} A & 0 & 0 \\ 0 & -S_1 & 0 \\ 0 & 0 & S_2 \end{bmatrix}}_{=:\mathscr{D}} \begin{bmatrix} I & A^{-1}B^T & 0 \\ 0 & I & -S_1^{-1}C^T \\ 0 & 0 & I \end{bmatrix}. \tag{3.1}$$

When A and S_1 are invertible, S_2 is well-defined and $\mathscr{K}$ is invertible if and only if $\mathscr{D}$ is invertible. The stated result follows. □

In this chapter and in Chapter 4, we assume that the sufficient condition holds: namely, that A, S_1, S_2 are invertible. Given our assumptions that D and E are semidefinite, this is equivalent to A, S_1, and S_2 being positive definite. This allows us to obtain the following result on the inertia of $\mathscr{K}$, which will be useful in deriving our bounds.

Lemma 3.2 (Inertia of a double saddle-point matrix)**.** *If A, S_1, and S_2 are positive definite, the matrix $\mathscr{K}$ has $n+p$ positive eigenvalues and m negative eigenvalues.*

Proof. When A, S_1, and S_2 are symmetric positive definite, Sylvester's Law of Inertia tells us the inertia of $\mathscr{K}$ is the same as that of $\mathscr{D}$ defined in (3.1); the stated result follows. □

Corollary 3.3. *Let a,b,c,d, and e be scalars with $a > 0$, $d,e \geq 0$, $s_1 := d + \frac{b^2}{a} > 0$ and $s_2 := e + \frac{c^2}{s_1} > 0$. Any cubic polynomial of the form*

$$p(\lambda) = \lambda^3 + (d-a-e)\lambda^2 + (ae - ad - de - b^2 - c^2)\lambda + (ade + ac^2 + b^2 e)$$

has two positive real roots and one negative real root.

Proof. Consider the 3×3 matrix

$$P = \begin{bmatrix} a & b & 0 \\ b & -d & c \\ 0 & c & e \end{bmatrix}.$$

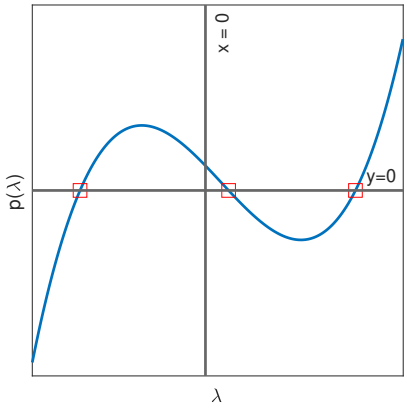

Figure 3.1: Plot of a cubic polynomial $p(\lambda)$ of the form described in Corollary 3.3, with two positive roots and one negative root.

Using the characteristic polynomial of P, it is straightforward to confirm that

$$\begin{aligned}\det(\lambda I - P) &= \lambda^3 - \mathrm{Tr}(P)\lambda^2 - \frac{1}{2}\left(\mathrm{Tr}(P^2) - \mathrm{Tr}^2(P)\right)\lambda - \det(P) \\ &= p(\lambda).\end{aligned}$$

Because P is symmetric its eigenvalues are real, and because P is a double saddle-point matrix with $n = m = p = 1$, the two positive and one negative root follow by Lemma 3.2. See Figure 3.1 for a graphical illustration. □

3.2 Derivation of bounds

Let us define three cubic polynomials, as follows:

$$p(\lambda) = \lambda^3 + (\mu^D_{\max} - \mu^A_{\min})\lambda^2 - \left(\mu^A_{\min}\mu^D_{\max} + (\sigma^B_{\max})^2 + (\sigma^C_{\min})^2\right)\lambda + \mu^A_{\min}(\sigma^C_{\min})^2; \tag{3.2a}$$

$$
\begin{aligned}
q(\lambda) = \lambda^3 &+ (\mu^D_{\min} - \mu^A_{\max} - \mu^E_{\max})\lambda^2 \\
&+ \left(\mu^A_{\max}\mu^E_{\max} - \mu^A_{\max}\mu^D_{\min} - \mu^D_{\min}\mu^E_{\max} - (\sigma^B_{\max})^2 - (\sigma^C_{\max})^2\right)\lambda \\
&+ \left(\mu^A_{\max}\mu^D_{\min}\mu^E_{\max} + \mu^A_{\max}(\sigma^C_{\max})^2 + (\sigma^B_{\max})^2\mu^E_{\max}\right);
\end{aligned} \tag{3.2b}
$$

$$
\begin{aligned}
r(\lambda) = \lambda^3 &+ (\mu^D_{\max} - \mu^A_{\min} - \mu^E_{\min})\lambda^2 \\
&+ \left(\mu^A_{\min}\mu^E_{\min} - \mu^A_{\min}\mu^D_{\max} - \mu^D_{\max}\mu^E_{\min} - (\sigma^B_{\max})^2 - (\sigma^C_{\max})^2\right)\lambda \\
&+ (\mu^A_{\min}\mu^D_{\max}\mu^E_{\min} + \mu^A_{\min}(\sigma^C_{\max})^2 + (\sigma^B_{\max})^2\mu^E_{\min}).
\end{aligned} \tag{3.2c}
$$

All three of these polynomials are of the form described in Corollary 3.3. Thus, all roots are real and each polynomial has two positive roots and one negative root. For notational convenience, we will denote the negative root of $p(\lambda)$, for example, by p^-, and use subscripts *max* and *min* to distinguish between the two positive roots. For example, $p^+_{\max}$ will denote the largest positive root and $p^+_{\min}$ will denote the smallest positive root. The same notational rules apply to $q(\lambda)$ and $r(\lambda)$.

Theorem 3.4 (Eigenvalue bounds, matrix $\mathscr{K}$)**.** *Suppose A is symmetric positive definite. Using the notation established in Equation 3.2, the eigenvalues of $\mathscr{K}$ are bounded within the intervals*

$$
\left[r^-, \frac{\mu^A_{\max} - \sqrt{(\mu^A_{\max})^2 + 4(\sigma^B_{\min})^2}}{2}\right] \bigcup \left[p^+_{\min}, q^+_{\max}\right]. \tag{3.3}
$$

Proof. *Upper bound on positive eigenvalues.* We let $v = \begin{bmatrix} x^T & y^T & z^T \end{bmatrix}^T$ be a vector with $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$, and $z \in \mathbb{R}^p$. Because $\mathscr{K}$ is symmetric, we can derive an upper bound on the eigenvalues of $\mathscr{K}$ by bounding the value of $\frac{v^T \mathscr{K} v}{v^T v}$. We can write

$$
v^T \mathscr{K} v = x^T A x - y^T D y + z^T E z + 2x^T B^T y + 2y^T C^T z. \tag{3.4}
$$

We use Cauchy-Schwarz to bound the mixed bilinear forms – such as $x^T B^T y$ – by, for example,

$$
-||x|| \cdot ||B^T y|| \leq x^T B^T y \leq ||x|| \cdot ||B^T y||.
$$

Using this and the eigenvalues/singular values of the block of $\mathscr{K}$, we can bound

(3.4) from above by

$$v^T \mathscr{K} v \leq \mu^A_{\max}||x||^2 - \mu^D_{\min}||y||^2 + \mu^E_{\max}||z||^2 + 2\sigma^B_{\max}||x|| \cdot ||y|| + 2\sigma^C_{\max}||y|| \cdot ||z||$$
$$= \begin{bmatrix} ||x|| & ||y|| & ||z|| \end{bmatrix} \underbrace{\begin{bmatrix} \mu^A_{\max} & \sigma^B_{\max} & 0 \\ \sigma^B_{\max} & -\mu^D_{\min} & \sigma^C_{\max} \\ 0 & \sigma^C_{\max} & \mu^E_{\max} \end{bmatrix}}_{=:R} \begin{bmatrix} ||x|| \\ ||y|| \\ ||z|| \end{bmatrix}.$$

An upper bound on $\frac{v^T \mathscr{K} v}{v^T v}$ is therefore given by the maximal eigenvalue of R. The largest positive eigenvalue of $\mathscr{K}$ is therefore less than or equal to the largest root of the characteristic polynomial $\det(\lambda I - R)$, which yields the desired result.

Lower bound on negative eigenvalues. The proof is similar to that for the upper bound on the positive eigenvalues. Using Cauchy-Schwarz and the eigenvalues/singular values of the blocks, we bound (3.4) from below by writing

$$v^T \mathscr{K} v \geq \mu^A_{\min}||x||^2 - 2\sigma^B_{\max}||x|| \cdot ||y|| - \mu^D_{\max}||y||^2 - 2\sigma^C_{\max}||y|| \cdot ||z|| + \mu^E_{\min}||z||^2$$
$$= \begin{bmatrix} ||x|| & ||y|| & ||z|| \end{bmatrix} \underbrace{\begin{bmatrix} \mu^A_{\min} & -\sigma^B_{\max} & 0 \\ -\sigma^B_{\max} & -\mu^D_{\max} & -\sigma^C_{\max} \\ 0 & -\sigma^C_{\max} & \mu^E_{\min} \end{bmatrix}}_{=:R} \begin{bmatrix} ||x|| \\ ||y|| \\ ||z|| \end{bmatrix}.$$

A lower bound on $\frac{v^T \mathscr{K} v}{v^T v}$ is therefore given by the smallest eigenvalue of R. Taking the characteristic polynomial of R yields the stated result.

Upper bound on negative eigenvalues. We derive an upper bound on the negative eigenvalues of $\mathscr{K}$ by finding a lower bound on the negative eigenvalues of $\mathscr{K}^{-1}$. We begin by partitioning $\mathscr{K}$ as

$$\mathscr{K} = \left[\begin{array}{cc|c} A & B^T & 0 \\ B & -D & C^T \\ \hline 0 & C & E \end{array}\right] = \begin{bmatrix} \mathscr{K}_2 & \bar{C}^T \\ \bar{C} & E \end{bmatrix}, \tag{3.5}$$

where $\mathscr{K}_2 = \begin{bmatrix} A & B^T \\ B & -D \end{bmatrix}$ and $\bar{C} = \begin{bmatrix} 0 & C \end{bmatrix}$. By [7, Equation (3.4)],

$$\mathscr{K}^{-1} = \begin{bmatrix} \mathscr{K}_2^{-1} + \mathscr{K}_2^{-1}\bar{C}^T S_2^{-1}\bar{C}\mathscr{K}_2^{-1} & -\mathscr{K}_2^{-1}\bar{C}S_2^{-1} \\ -S_2^{-1}\bar{C}\mathscr{K}_2^{-1} & S_2^{-1} \end{bmatrix},$$

where $S_2 = E - \bar{C}\mathscr{K}_2^{-1}\bar{C}^T = E + CS_1^{-1}C^T$. Notice that

$$\mathscr{K}^{-1} = \begin{bmatrix} \mathscr{K}_2^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} \mathscr{K}_2^{-1}\bar{C}^T \\ -I \end{bmatrix} S_2^{-1} \begin{bmatrix} \bar{C}\mathscr{K}_2^{-1} & -I \end{bmatrix}.$$

Because the second term is positive semidefinite, we conclude that the eigenvalues of $\mathscr{K}^{-1}$ are greater than or equal to the eigenvalues of $\begin{bmatrix} \mathscr{K}_2^{-1} & 0 \\ 0 & 0 \end{bmatrix}$. Thus, a lower bound on the negative eigenvalues of $\mathscr{K}_2^{-1}$ is also a lower bound on the negative eigenvalues of $\mathscr{K}^{-1}$. This means that an upper bound on the negative eigenvalues of $\mathscr{K}_2$ is also an upper bound on the negative eigenvalues of $\mathscr{K}$. The desired result now follows from Silvester and Wathen [81, Lemma 2.2].

Lower bound on positive eigenvalues. We begin by noting that, because E is positive semidefinite, the eigenvalues of $\mathscr{K}$ are greater than or equal to those of

$$\mathscr{K}_{E=0} = \begin{bmatrix} A & B^T & 0 \\ B & -D & C^T \\ 0 & C & 0 \end{bmatrix}.$$

Moreover, $\mathscr{K}$ and $\mathscr{K}_{E=0}$ have the same inertia, by Lemma 3.2. Thus, the smallest positive eigenvalue of $\mathscr{K}$ is greater than or equal to the smallest positive eigenvalue of $\mathscr{K}_{E=0}$. We therefore obtain a lower bound on the positive eigenvalues of $\mathscr{K}$ by using energy estimates with $\mathscr{K}_{E=0}$. The eigenvalue problem associated with $\mathscr{K}_{E=0}$ is

$$\begin{bmatrix} A & B^T & 0 \\ B & -D & C^T \\ 0 & C & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \lambda \begin{bmatrix} x \\ y \\ z \end{bmatrix}. \tag{3.6}$$

The third block row of (3.6) gives $z = \frac{1}{\lambda} Cy$. The first block row gives

$$B^T y = (\lambda I - A)x.$$

We now consider two cases based on the value of λ.

Case I: $\lambda < \mu^A_{\min}$*:* If $\lambda < \mu^A_{\min}$, then $(\lambda I - A)$ is negative definite, so we can write

$$x = (\lambda I - A)^{-1} B^T y.$$

Substituting this into the second block row of (3.6), pre-multiplying by y^T and rearranging gives

$$\lambda y^T y = y^T B(\lambda I - A)^{-1} B^T y - y^T D y + \frac{1}{\lambda} y^T C^T C y \tag{3.7a}$$

$$\geq \frac{(\sigma^B_{\max})^2}{\lambda - \mu^A_{\min}} y^T y - \mu^D_{\max} y^T y + \frac{(\sigma^C_{\min})^2}{\lambda} y^T y. \tag{3.7b}$$

Dividing by $y^T y$, using the fact that $\lambda > 0$ and $\lambda - \mu^A_{\min} < 0$, and rearranging yields

$$\underbrace{\lambda^3 + (\mu^D_{\max} - \mu^A_{\min})\lambda^2 - \left(\mu^A_{\min}\mu^D_{\max} + (\sigma^B_{\max})^2 + (\sigma^C_{\min})^2\right)\lambda + \mu^A_{\min}(\sigma^C_{\min})^2}_{=p(\lambda)} \leq 0.$$

By applying Corollary 3.3 with $a = \mu^A_{\min}$, $b = \sigma^B_{\max}$, $c = \sigma^C_{\min}$, and $d = \mu^D_{\max}$, we know that this polynomial has two positive roots. Moreover, $p(\lambda)$ is negative between these two roots: this follows from the fact that $p(0) > 0$ and $\lim_{\lambda \to \infty} p(\lambda) = \infty$. Therefore, we conclude that in this case λ is greater than or equal to the smaller positive root of $p(\lambda)$, namely $p^+_{\min}$.

Case II: $\lambda \geq \mu^A_{\min}$*:* If $\lambda \geq \mu^A_{\min}$, then $(\lambda I - A)$ may be indefinite (and possibly singular). However, we can still obtain a lower bound for λ by observing that $\mu^A_{\min}$ is greater than $p^+_{\min}$. As stated earlier, $p(\lambda)$ has two positive roots and the value of $p(\lambda)$ is negative between those roots. Moreover, these are the only positive values

of λ for which $p(\lambda) < 0$. We observe, after simplification, that

$$p(\mu^A_{\min}) = -(\sigma^B_{\max})^2 \mu^A_{\min} < 0.$$

Therefore, the bound $\lambda \geq p^+_{\min}$ also holds in this case, which completes the proof. □

Remark 3.5. From Theorem 3.4 we see that when B and C are rank deficient, the internal bounds are zero. Similarly, when A is rank-deficient the lower positive bound is zero. Under mild conditions on the ranks and kernels of D and E, the statement of the theorem and its proof may be revised to obtain nonzero internal bounds in this case. However, doing so is rather technical, and since the case of full rank B and C is common, further details on this end case are omitted.

The matrix $\mathscr{K}_0$ is a special case of $\mathscr{K}$, and bounds on its eigenvalues can be obtained as a direct consequence of Theorem 3.4.

Corollary 3.6 (Eigenvalue bounds, matrix $\mathscr{K}_0$)**.** *Define the following three cubic polynomials as special cases of p, q and r defined in* (3.2) *with $D = E = 0$:*

$$\hat{p}(\lambda) = \lambda^3 - \mu^A_{\min}\lambda^2 - \left((\sigma^B_{\max})^2 + (\sigma^C_{\min})^2\right)\lambda + \mu^A_{\min}(\sigma^C_{\min})^2; \tag{3.8a}$$

$$\hat{q}(\lambda) = \lambda^3 - \mu^A_{\max}\lambda^2 - \left((\sigma^B_{\max})^2 + (\sigma^C_{\max})^2\right)\lambda + \mu^{\max}_A(\sigma^C_{\max})^2; \tag{3.8b}$$

$$\hat{r}(\lambda) = \lambda^3 - \mu^A_{\min}\lambda^2 - \left((\sigma^B_{\max})^2 + (\sigma^C_{\max})^2\right)\lambda + \mu^A_{\min}(\sigma^C_{\max})^2. \tag{3.8c}$$

Using (3.8)*, the eigenvalues of $\mathscr{K}_0$ are bounded within the intervals*

$$\left[\hat{r}^-, \frac{\mu^A_{\max} - \sqrt{(\mu^A_{\max})^2 + 4(\sigma^B_{\min})^2}}{2}\right] \bigcup \left[\hat{p}^+_{\min}, \hat{q}^+_{\max}\right]. \tag{3.9}$$

The proof of Corollary 3.6 is omitted; it is similar to the proof of Theorem 3.4, with simplifications arising from setting $D = E = 0$.

3.3 Tightness of the bounds

While the tightness of the bounds we have obtained depends on the problem at hand, all bounds presented in this section are attainable, as we will demonstrate

with small examples. For the extremal bounds (upper positive and lower negative), we note that both bounds hold for all double saddle-point matrices with $n = m = p = 1$, as the characteristic polynomial of the 3×3 matrix is the same as the polynomial given in the upper positive and lower negative bounds of Theorem 3.4.

For the upper bound on negative eigenvalues, we consider the example (with $n = m = 2$, and $p = 1$):

$$\mathscr{K} = \left[\begin{array}{cc|cc|c} \mu^A_{\max} & 0 & \sigma^B_{\min} & 0 & 0 \\ 0 & \mu^A_{\max} & 0 & \sigma^B_{\min} & 0 \\ \hline \sigma^B_{\min} & 0 & 0 & 0 & 0 \\ 0 & \sigma^B_{\min} & 0 & -\mu^D & \sigma^C \\ \hline 0 & 0 & 0 & \sigma^C & \mu^E \end{array}\right].$$

We can permute the rows and columns of $\mathscr{K}$ to obtain a block diagonal matrix:

$$\mathscr{K}_D = \left[\begin{array}{cc|ccc} \mu^A_{\max} & \sigma^B_{\min} & 0 & 0 & 0 \\ \sigma^B_{\min} & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & \mu^A_{\max} & \sigma^B_{\min} & 0 \\ 0 & 0 & \sigma^B_{\min} & -\mu^D & \sigma^C \\ 0 & 0 & 0 & \sigma^C & \mu^E \end{array}\right].$$

The upper left block of $\mathscr{K}_D$ has as an eigenvalue $\frac{\mu^A_{\max} - \sqrt{(\mu^A_{\max})^2 + 4(\sigma^B_{\min})^2}}{2}$, which is the bound given by Theorem 3.4.

Finally, for the lower positive bound, we consider the matrix (with $n = m = p = 2$)

$$\mathscr{K} = \left[\begin{array}{cc|cc|cc} \mu^A_{\min} & 0 & \sigma^B_{\max} & 0 & 0 & 0 \\ 0 & \mu^A_{\min} & 0 & \sigma^B_{\max} & 0 & 0 \\ \hline \sigma^B_{\max} & 0 & -\mu^D_{\max} & 0 & \sigma^C_{\min} & 0 \\ 0 & \sigma^B_{\max} & 0 & -\mu^D_{\max} & 0 & \sigma^C_{\min} \\ \hline 0 & 0 & \sigma^C_{\min} & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma^C_{\min} & 0 & \mu^E \end{array}\right].$$

As in the previous example, we can permute the rows and columns to obtain a

block diagonal matrix:

$$\mathscr{K}_D = \left[\begin{array}{ccc|ccc} \mu^A_{\min} & \sigma^B_{\max} & 0 & 0 & 0 & 0 \\ \sigma^B_{\max} & -\mu^D_{\max} & \sigma^C_{\min} & 0 & 0 & 0 \\ 0 & \sigma^C_{\min} & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & \mu^A_{\min} & \sigma^B_{\max} & 0 \\ 0 & 0 & 0 & \sigma^B_{\max} & -\mu^D_{\max} & \sigma^C_{\min} \\ 0 & 0 & 0 & 0 & \sigma^C_{\min} & \mu^E \end{array}\right].$$

The characteristic polynomial of the upper left block is precisely $p(\lambda)$ defined in (3.2a); thus, the bound of Theorem 3.4 is obtained.

3.4 Numerical experiments

In this section we consider two slightly different variants of a Poisson control problem as in [71]. First, we consider a distributed control problem with Dirichlet boundary conditions:

$$\min_{u,f} \frac{1}{2}||u-\hat{u}||^2_{L_2(\Omega)} + \frac{\beta}{2}||f||^2_{L_2(\Omega)} \tag{3.10a}$$

$$\text{s.t.} \quad -\nabla^2 u = f \ \text{ in } \Omega, \tag{3.10b}$$

$$u = g \ \text{ on } \partial\Omega, \tag{3.10c}$$

where u is the state, $\hat{u}$ is the desired state, $0 < \beta \ll 1$ is a regularization parameter, f is the control, g is a source function, and Ω is the domain with boundary $\partial\Omega$. After discretization, we can write the Lagrangian as:

$$\mathscr{L} = \frac{1}{2}u_h^T M u_h - u_h^T b + ||\hat{u}||_2^2 + \beta f_h^T M f_h + \lambda^T (Ku - Mf_h),$$

where: M is the discrete mass matrix; K is the discrete Laplacian stiffness matrix; u_h and f_h are the finite element discretizations of u and f; b is a vector of inner products of $\hat{u}$ with the finite element basis functions; d contains the terms coming from the boundary values of u_h; and λ is the Lagrange multiplier corresponding to the constraint. Using the stationarity conditions of $\mathscr{L}$, defined by setting $\nabla\mathscr{L} = 0$,

we obtain the system

$$\underbrace{\begin{bmatrix} M & K & 0 \\ K & 0 & -M \\ 0 & -M & \beta M \end{bmatrix}}_{=:\mathscr{K}} \begin{bmatrix} u_h \\ \lambda \\ f_h \end{bmatrix} = \begin{bmatrix} b \\ d \\ 0 \end{bmatrix}, \tag{3.11}$$

where M is a symmetric positive definite mass matrix and K is a symmetric positive definite discrete Laplacian. All blocks of $\mathscr{K}$ are square (i.e., $n = m = p$).

As a second experiment we consider a boundary control problem:

$$\min_{u,f} \frac{1}{2}||u - \hat{u}||^2_{L_2(\Omega)} + \frac{\beta}{2}||g||^2_{L_2(\partial\Omega)} \tag{3.12a}$$

$$\text{s.t.} \quad -\nabla^2 u = 0 \ \text{ in } \Omega, \tag{3.12b}$$

$$\frac{\partial u}{\partial n} = g \ \text{ on } \partial\Omega, \tag{3.12c}$$

which after discretization yields the linear system

$$\underbrace{\begin{bmatrix} M & K & 0 \\ K & 0 & -E \\ 0 & -E^T & \beta M_b \end{bmatrix}}_{=:\mathscr{K}_\partial} \begin{bmatrix} u_h \\ \lambda \\ g_h \end{bmatrix} = \begin{bmatrix} b_\partial \\ d_\partial \\ 0 \end{bmatrix}, \tag{3.13}$$

where $M_b \in \mathbb{R}^{n_b \times n_b}$ (with $n_b < n$) is a boundary mass matrix. Thus, in the distributed control problem the mass matrix in the (2,3)/(3,2)-block is square and in the boundary control problem we consider a rectangular version of it.

In all experiments that follow we set Ω to be the unit square. We use uniform **Q1** finite elements and set $\beta = 10^{-3}$. The MATLAB code of Rees [70] was used to generate the linear systems.

Here we compare the eigenvalues of $\mathscr{K}$ (3.11) and $\mathscr{K}_\partial$ (3.13) to the eigenvalue bounds predicted by Theorem 3.4. We use MATLAB's `eigs`/`svds` functions to compute the minimum and maximum eigenvalues/singular values of the matrix blocks M, K, M_b, and E.

We note that for the distributed control matrix $\mathscr{K}$, all blocks are square and

the (3,3)-block is positive definite; therefore, we can also use our results to obtain bounds on the re-ordered matrix

$$\mathscr{K}_{\text{flip}} = \begin{bmatrix} \beta M & -M & 0 \\ -M & 0 & K \\ 0 & K & M \end{bmatrix}. \tag{3.14}$$

Both orderings generate the same extremal bounds but different interior bounds.

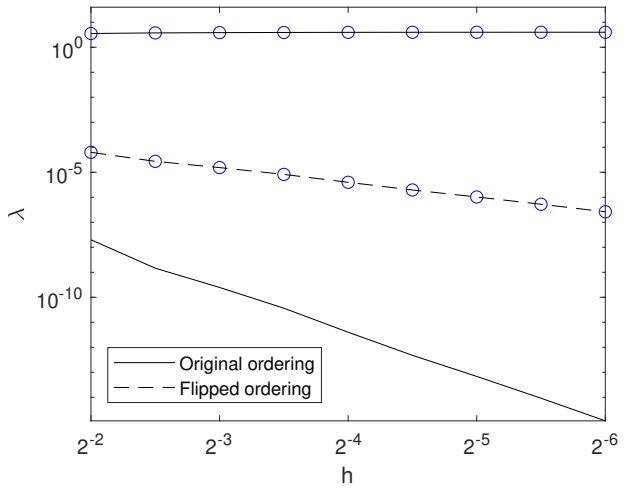

Figure 3.2: Largest and smallest positive eigenvalues of $\mathscr{K}$. Blue circles indicate the eigenvalues, and black lines the bounds given by Theorem 3.4. The dashed lines indicate the bounds obtained by applying Theorem 3.4 to the reordered matrix $\mathscr{K}_{\text{flip}}$.

Comparisons of the predicted eigenvalue bounds to the actual eigenvalues are shown in Figures 3.2-3.3 (distributed control) and 3.4-3.5 (boundary control). In the distributed control case, we show the bounds obtained for both the original matrix $\mathscr{K}$ and those for the reordered matrix $\mathscr{K}_{\text{flip}}$. In all cases, the bounds for the extremal eigenvalues are quite tight: this is because the K block has the largest eigenvalues ($O(1)$ compared to $O(h^2)$ for the others – see [17, Proposition 1.29 and Theorem 1.32]). Therefore, the largest positive and smallest negative eigenvalue of both $\mathscr{K}$ and $\mathscr{K}_\partial$ tend towards $\mu^K_{\max}$ and $-\mu^K_{\max}$, respectively. Referring to the

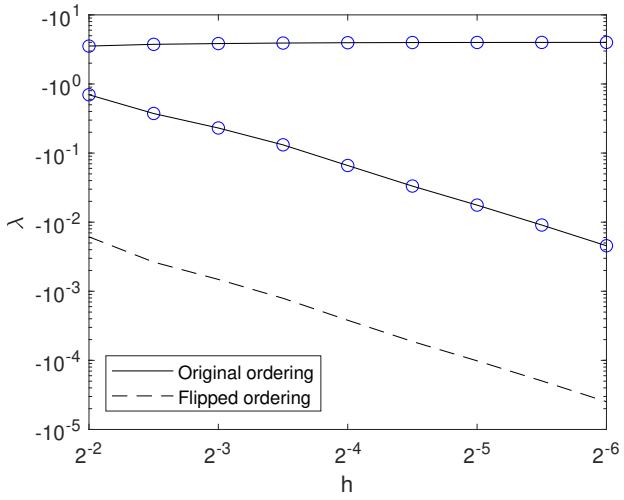

Figure 3.3: Largest and smallest negative eigenvalues of $\mathscr{K}$. Blue circles indicate the eigenvalues, and black lines the bounds given by Theorem 3.4. The dashed lines indicate the bounds obtained by applying Theorem 3.4 to the reordered matrix $\mathscr{K}_{\text{flip}}$.

proofs of the extremal bounds in Theorem 3.4, the 3×3 reduced matrices R will contain a $\mu^K_{\max}$ (or $-\mu^K_{\max}$) term, with other lower-order terms. This means that the extremal eigenvalues of R (and therefore the predicted eigenvalue bounds) will also be close to $\pm\mu^K_{\max}$.

It is more difficult to capture the interior bounds. In the distributed control case, each ordering (either the original ordering with M in the leading block or the flipped ordering with βM in the leading block) gives one bound that is quite tight and one that is loose. In the boundary control case, both interior bounds are loose.

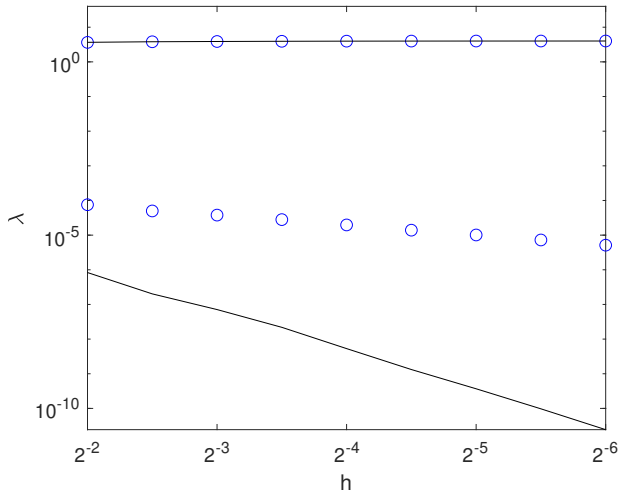

Figure 3.4: Largest and smallest positive eigenvalues of $\mathscr{K}_\partial$. Blue circles indicate the eigenvalues, and black lines the bounds given by Theorem 3.4.

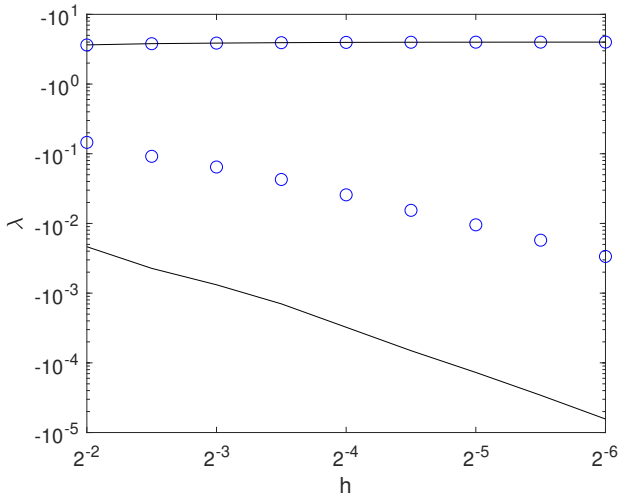

Figure 3.5: Largest and smallest positive eigenvalues of $\mathscr{K}_\partial$. Blue circles indicate the eigenvalues, and black lines the bounds given by Theorem 3.4.

Chapter 4

Preconditioning when A is positive definite

We consider the block diagonal preconditioner:

$$\mathscr{M} := \begin{bmatrix} A & 0 & 0 \\ 0 & S_1 & 0 \\ 0 & 0 & S_2 \end{bmatrix}, \tag{4.1}$$

where

$$S_1 = D + BA^{-1}B^T; \quad S_2 = E + CS_1^{-1}C^T. \tag{4.2}$$

We assume that S_1 and S_2 are both positive definite.

The preconditioner $\mathscr{M}$ is based on Schur complements. It has been considered in, for example, [11, 48, 84], and is a natural extension of [49, 60] for block-2×2 matrices of the form (1.3). In practice the Schur complements S_1 and S_2 defined in (4.2) are too expensive to form and invert exactly. It is therefore useful to consider approximations to those matrices when a practical preconditioner is to be developed, and we include an analysis of that scenario.

The properties of the preconditioning approach (4.2) have been the focus of several recent articles. The paper [84] analyzes the performance of a block-$n \times n$ block diagonal preconditioner analogous to $\mathscr{M}$ defined in (4.1), concentrating on the spectral properties of the continuous preconditioned operator. The paper [65]

provides additional analytical results and bounds, and in [64] the concentration of eigenvalues near zero is discussed and an alternative preconditioning approach is offered for multiple saddle-point systems with a larger number of blocks. The papers [11, 48] focus their analyses on the case where all diagonal blocks of $\mathcal{K}$ except A are zero. See also Remark 4.6.

Our analysis extends the work in the literature in a few useful ways. We use energy estimates and other analytical tools to prove our results in a variety of cases. Our proof techniques enable a refined analysis of the eigenvalues of the unpreconditioned matrix when only $D = 0$, which is an interesting case because it arises commonly in applications such as PDE-constrained optimization [71]. Here we obtain bounds that cannot be easily derived from the analysis of the case in which both D and E are (potentially) nonzero. Finally and most significantly, we provide eigenvalue bounds for the preconditioned system when approximations of the leading block and Schur complement inversions are used in the challenging case where $D, E \neq 0$. Our assumptions are minimal and the analysis is broader than the analysis in [48] for the unregularized matrix $\mathcal{K}_0$.

4.1 Eigenvalue bounds for block diagonal preconditioning

We now derive bounds for the preconditioned matrix $\mathcal{M}^{-1}\mathcal{K}$, with $\mathcal{M}$ defined in (4.1).

4.1.1 Inertia and eigenvalue multiplicity

We begin with some observations on inertia and eigenvalue multiplicity. To simplify the presentation and proof of this result, we restrict ourselves to the case that B and C have full row rank. In some later results, we will lift this restriction to allow for rank-deficient B and C.

Theorem 4.1 (Inertia and algebraic multiplicity, matrix $\mathcal{M}^{-1}\mathcal{K}$)**.** *Let $\mathcal{K}$ be defined as in* (1.2) *and $\mathcal{M}$ as in* (4.1)*, and suppose that B and C have full row rank. The preconditioned matrix $\mathcal{M}^{-1}\mathcal{K}$ has:*

(i) m negative eigenvalues;

(ii) p eigenvalues in $(0,1)$;

(iii) $n-m$ eigenvalues equal to 1; and

(iv) m eigenvalues greater than 1.

Proof. Because $\mathscr{M}$ is symmetric positive definite, $\mathscr{M}^{1/2}$ exists and is invertible, and the inertias of $\mathscr{M}^{-1/2}\mathscr{K}\mathscr{M}^{-1/2}$ and $\mathscr{M}^{-1}\mathscr{K}$ are equal to the inertia of $\mathscr{K}$. Thus, Lemma 3.2 establishes that $\mathscr{M}^{-1}\mathscr{K}$ has $n+p$ positive and m negative eigenvalues, which proves (i).

For (ii)-(iv) we split the $n+p$ positive eigenvalues into eigenvalues less than, equal to, or greater than 1. For this we compute the inertia of the shifted, split-preconditioned matrix

$$\mathscr{M}^{-1/2}\mathscr{K}\mathscr{M}^{-1/2} - I = \begin{bmatrix} 0 & \tilde{B}^T & 0 \\ \tilde{B} & -\tilde{D} - I & \tilde{C}^T \\ 0 & \tilde{C} & \tilde{E} - I \end{bmatrix},$$

where $\tilde{B} = S_1^{-1/2} B A^{-1/2}$, $\tilde{C} = S_2^{-1/2} C S_1^{-1/2}$, $\tilde{D} = S_1^{-1/2} D S_1^{-1/2}$, $\tilde{E} = S_2^{-1/2} E S_2^{-1/2}$, and I is an identity matrix of appropriate dimension (with slight abuse of notation, we use I to denote an identity matrix of any size). The number of positive, negative, and zero eigenvalues of $\mathscr{M}^{-1/2}\mathscr{K}\mathscr{M}^{-1/2} - I$ will be equal to the number of eigenvalues of $\mathscr{M}^{-1}\mathscr{K}$ greater than, less than, or equal to 1, respectively.

Noting that $\tilde{E} + \tilde{C}\tilde{C}^T = I$, we write

$$\mathscr{M}^{-1/2}\mathscr{K}\mathscr{M}^{-1/2} - I = \begin{bmatrix} 0 & \mathscr{B} \\ \mathscr{B}^T & \mathscr{T} \end{bmatrix},$$

where

$$\mathscr{T} := \begin{bmatrix} -\tilde{D} - I & \tilde{C}^T \\ \tilde{C} & -\tilde{C}\tilde{C}^T \end{bmatrix} \quad \text{and} \quad \mathscr{B} = \begin{bmatrix} \tilde{B}^T & 0 \end{bmatrix}.$$

The matrix $\mathscr{B}$ is in $\mathbb{R}^{n\times(m+p)}$ and has rank m. We define a matrix

$$N := \begin{bmatrix} \mathbf{0}_{m\times p} \\ I_p \end{bmatrix},$$

where $\mathbf{0}_{m\times p}$ is the $m \times p$ zero matrix and I_p is the $p \times p$ identity matrix. The columns of N form a basis for $\ker(\mathscr{B})$. Denote the inertia of M by $\mathrm{In}(M) = (n_+, n_-, n_0)$. It is well known (see, e.g., [36, Lemma 3.4]) that

$$\mathrm{In}(\mathscr{M}^{-1/2}\mathscr{K}\mathscr{M}^{-1/2} - I) = \mathrm{In}(N^T\mathscr{T}N) + (m, m, n-m).$$

Because $N^T\mathscr{T}N = -\tilde{C}\tilde{C}^T$ is negative definite, this gives $\mathrm{In}(\mathscr{M}^{-1/2}\mathscr{K}\mathscr{M}^{-1/2} - I) = (m, m+p, n-m)$, which yields (ii)-(iv). □

4.1.2 Derivation of bounds

It is possible to obtain eigenvalue bounds on the preconditioned system $\mathscr{M}^{-1}\mathscr{K}$ by using the results of Theorem 3.4 on the (symmetric) preconditioned system $\mathscr{M}^{-1/2}\mathscr{K}\mathscr{M}^{-1/2}$; however, some of the resulting bounds will be loose. The reason for this is that Theorem 3.4 uses the *R*-matrix technique, which assumes no relationships between the blocks of $\mathscr{K}$ and considers each block individually. In this section, we will derive tight eigenvalue bounds using energy estimates, which allow us to fully exploit the relationships between the blocks of the preconditioned system.

We begin by recalling a result that follows (after minor notational adjustments) from Horn and Johnson [45, Theorem 7.7.3], which will be useful throughout the analysis that follows.

Lemma 4.2. *Let M and N be symmetric positive semidefinite matrices such that $M+N$ is positive definite. Then all eigenvalues of $(M+N)^{-1}M$ are in $[0,1]$.*

The following result follows directly from [84, Lemma A.1] and simplifies the presentation of some of the subsequent results in this section.

Lemma 4.3. *The roots of the cubic polynomial $\lambda^3 - \lambda^2 - 2\lambda + 1$ are given by*

$$\lambda = \left\{ 2\cos\left(\frac{\pi}{7}\right),\ 2\cos\left(\frac{3\pi}{7}\right),\ 2\cos\left(\frac{5\pi}{7}\right) \right\},$$

which are approximately equal to 1.8019, 0.4450, and -1.2470, respectively.

The case $D = E = 0$

When D and E are both zero, $\mathscr{M}^{-1}\mathscr{K}$ has six distinct eigenvalues given by: 1, $\frac{1\pm\sqrt{5}}{2}$, $2\cos\left(\frac{\pi}{7}\right), 2\cos\left(\frac{3\pi}{7}\right)$, and $2\cos\left(\frac{5\pi}{7}\right)$. The last three are the roots of the cubic polynomial $\lambda^3 - \lambda^2 - 2\lambda + 1$, per Lemma 4.3. The proof follows directly from [84, Theorem 2.3]. See also [11, 48] for later proofs.

The case $D = 0$ and $E \succeq 0$

When $D = 0$, it is necessary that B have full row rank in order for S_1 to be invertible; however, C may be rank-deficient. Suppose that C^T has nullity k. The following result holds.

Theorem 4.4 (Eigenvalue bounds, matrix $\mathscr{M}^{-1}\mathscr{K}$, $D = 0$, $E \neq 0$)**.** *When $D = 0$ and $E \neq 0$, the eigenvalues of $\mathscr{M}^{-1}\mathscr{K}$ are given by: $\lambda = 1$ with multiplicity $n - m + k$; $\lambda = \frac{1\pm\sqrt{5}}{2}$, each with multiplicity $m - p + k$; and $p - k$ eigenvalues located in each of the three intervals:*

- $I_1 = \left[2\cos\left(\frac{5\pi}{7}\right), \frac{1-\sqrt{5}}{2}\right) \approx [-1.618, -0.618)$
- $I_2 = \left[2\cos\left(\frac{3\pi}{7}\right), 1\right) \approx [0.4450, 1)$
- $I_3 = \left(\frac{1+\sqrt{5}}{2}, 2\cos\left(\frac{\pi}{7}\right)\right] \approx (1.618, 1.8019]$

Proof. We write out the (left-)preconditioned operator

$$\mathscr{M}^{-1}\mathscr{K} = \begin{bmatrix} I & A^{-1}B^T & 0 \\ S_1^{-1}B & 0 & S_1^{-1}C^T \\ 0 & S_2^{-1}C & S_2^{-1}E \end{bmatrix},$$

with corresponding eigenvalue equations

$$x + A^{-1}B^T y = \lambda x; \tag{4.3a}$$

$$S_1^{-1}Bx + S_1^{-1}C^T z = \lambda y; \tag{4.3b}$$

$$S_2^{-1}Cy + S_2^{-1}Ez = \lambda z. \tag{4.3c}$$

We obtain $n - m$ eigenvectors for $\lambda = 1$ by choosing $x \in \ker(B)$ and $y, z = 0$.

By considering $y \in \ker(C)$ and $z = 0$, we obtain eigenvalues $\lambda = \frac{1\pm\sqrt{5}}{2}$, each with geometric multiplicity $m - p + k$.

For the remaining eigenvalues, we assume that $z \neq 0$ and $\lambda \notin \{1, \frac{1\pm\sqrt{5}}{2}\}$. From (4.3a) we obtain

$$x = \frac{1}{\lambda - 1} A^{-1} B^T y,$$

which we substitute into (4.3b) and rearrange to get

$$y = \frac{\lambda - 1}{\lambda^2 - \lambda - 1} S_1^{-1} C^T z.$$

Substituting this into (4.3c) gives

$$\frac{\lambda - 1}{\lambda^2 - \lambda - 1} S_2^{-1} C S_1^{-1} C^T z + S_2^{-1} E z = \lambda z. \tag{4.4}$$

Because $S_2 = E + C S_1^{-1} C^T$, we can write $S_2^{-1} E = I - S_2^{-1} C S_1^{-1} C^T$. We substitute this into (4.4) and rearrange to obtain

$$\Big(-\lambda^2 + 2\lambda \Big) S_2^{-1} C S_1^{-1} C^T z = \Big(\lambda^3 - 2\lambda^2 + 1 \Big) z.$$

Let $\{\delta_j, v_j\}$, for $1 \leq j \leq p$, denote an eigenpair of $S_2^{-1} C S_1^{-1} C^T$. For k of these eigenpairs corresponding to $v_j \in \ker(C^T)$, we note that

$$S_2^{-1} E v_j = (I - S_2^{-1} C S_1^{-1} C^T) v_j = v_j,$$

and therefore $\begin{bmatrix} 0 & 0 & v_j^T \end{bmatrix}^T$ is an eigenvector of $\mathscr{M}^{-1}\mathscr{K}$ with $\lambda = 1$. For the $p - k$ remaining eigenpairs, we have $0 < \delta_j \leq 1$ (by Lemma 4.2). We can then write

$$\Big(-\lambda_j^2 + 2\lambda_j \Big) \delta_j z_j = \Big(\lambda_j^3 - 2\lambda_j^2 + 1 \Big) z_j,$$

where λ_j corresponds to an eigenpair $\{\delta_j, z_j\}$ of $S_2^{-1} C S_1^{-1} C^T$. Because $z_j \neq 0$ this implies that

$$\lambda_j^3 - (2 - \delta_j)\lambda_j^2 - 2\delta_j \lambda_j + 1 = 0. \tag{4.5}$$

Thus, each of the $p - k$ positive eigenvalues δ_j of $S_2^{-1} C S_1^{-1} C^T$ yields three distinct

corresponding eigenvalues $\lambda_j^{(1)}, \lambda_j^{(2)}$, and $\lambda_j^{(3)}$ of $\mathscr{M}^{-1}\mathscr{K}$, corresponding to the roots of the cubic polynomial (4.5). These $3(p-k)$ eigenvalues, combined with the eigenvalues described earlier, account for all eigenvalues of $\mathscr{M}^{-1}\mathscr{K}$. Substituting $\delta_j = 0$ and $\delta_j = 1$ into (4.5) gives us the three intervals for these eigenvalues stated in the theorem. □

The case $D \succeq 0$ and $E = 0$

When $D \succeq 0$ and $E = 0$, the bounds are the same as when D and E are both nonzero, which are given next.

The case $D, E \succeq 0$

Theorem 4.5 (Eigenvalue bounds, matrix $\mathscr{M}^{-1}\mathscr{K}$, $D, E \succeq 0$)**.** *The eigenvalues of $\mathscr{M}^{-1}\mathscr{K}$ are bounded within the intervals*

$$\left[-\frac{1+\sqrt{5}}{2}, \frac{1-\sqrt{5}}{2}\right] \cup \left[2\cos\left(\frac{3\pi}{7}\right), 2\cos\left(\frac{\pi}{7}\right)\right],$$

which are approximately $[-1.618, -0.618] \cup [0.4450, 1.8019]$.

Remark 4.6. A proof for the upper bound on the positive eigenvalues is presented in [84, Theorem 2.1] and one for the lower bound on the positive eigenvalues is given in [84, Theorem 2.2]. A different proof for the four bounds appears in [64, Theorem 5.3]. We provide below an alternative technique of proof based on energy estimates.

Proof. *Upper bound on positive eigenvalues.* We know from Theorem 4.1 that the upper bound on positive eigenvalues is greater than 1, so we assume here that $\lambda > 1$. We begin by writing the eigenvalue equations as

$$Ax + B^T y = \lambda Ax; \tag{4.6a}$$

$$Bx - Dy + C^T z = \lambda S_1 y; \tag{4.6b}$$

$$Cy + Ez = \lambda S_2 z. \tag{4.6c}$$

From (4.6a) we get

$$x = \frac{1}{\lambda - 1} A^{-1} B^T y, \tag{4.7}$$

and from (4.6c) we get $z = (\lambda S_2 - E)^{-1} C y$ (because $\lambda > 1$, we are guaranteed that $\lambda S_2 - E = (\lambda - 1)E + \lambda C S_1^{-1} C^T$ is positive definite). Substituting these values back into (4.6b) and pre-multiplying by y^T gives

$$\left(\frac{1}{\lambda - 1}\right) y^T B A^{-1} B^T y + y^T (D + \lambda S_1) y - y^T C^T (\lambda S_2 - E)^{-1} C y = 0.$$

Recalling that $S_1 = D + B A^{-1} B^T$ we can rewrite this as

$$\left(\frac{1}{1-\lambda} + \lambda\right) y^T S_1 y + \left(1 - \frac{1}{1-\lambda}\right) y^T D y - y^T C^T (\lambda S_2 - E)^{-1} C y = 0. \tag{4.8}$$

Because $\lambda > 1$, we have

$$\begin{aligned} y^T C^T (\lambda S_2 - E)^{-1} C y &\leq y^T C^T \Big(\lambda (S_2 - E)\Big)^{-1} C y \\ &= \frac{1}{\lambda} y^T C^T \left(C S_1^{-1} C^T\right)^{-1} C y. \end{aligned}$$

Therefore, (4.8) gives

$$\left(\frac{1}{1-\lambda} + \lambda\right) y^T S_1 y + \left(1 - \frac{1}{1-\lambda}\right) y^T D y - \frac{1}{\lambda} y^T C^T (C S_1^{-1} C^T)^{-1} C y \leq 0. \tag{4.9}$$

Next, we let $\tilde{y} = S_1^{1/2} y$ and rewrite the first and third terms in (4.9) in terms of $\tilde{y}$:

$$\left(\frac{1}{1-\lambda} + \lambda\right) \tilde{y}^T \tilde{y} + \left(1 - \frac{1}{1-\lambda}\right) y^T D y - \frac{1}{\lambda} \tilde{y}^T \underbrace{S_1^{-1/2} C^T (C S_1^{-1} C^T)^{-1} C S_1^{-1/2}}_{:=P} \tilde{y} \leq 0. \tag{4.10}$$

Because P defined in (4.10) is an orthogonal projector, we have $\tilde{y}^T P \tilde{y} \leq \tilde{y}^T \tilde{y}$. And because $\lambda > 1$, we have $1 - \frac{1}{1-\lambda} > 0$, which means that the second term $\left(1 - \frac{1}{1-\lambda}\right) y^T D y$ is non-negative and can be dropped from the inequality. The inequality (4.10) therefore becomes

$$\left(\frac{1}{1-\lambda} + \lambda\right) \tilde{y}^T \tilde{y} - \frac{1}{\lambda} \tilde{y}^T y \leq 0,$$

which we can divide by $\tilde{y}^T\tilde{y}$ rearrange to give

$$\lambda^3 - \lambda^2 - 2\lambda + 1 \leq 0.$$

This, combined with the assumption that $\lambda > 1$, gives us the stated result that λ is less than or equal to the largest root of $\lambda^3 - \lambda^2 - 2\lambda + 1$. The polynomial $\lambda^3 - \lambda^2 - 2\lambda + 1$ is of the form given in Corollary 3.3 with $a = b = c = 1$, $d = e = 0$, so its value is negative between the two positive roots.

Lower bound on negative eigenvalues. We begin from (4.8) and note that when $\lambda < 0$, by similar reasoning as was shown for the upper bound on positive eigenvalues,

$$\left(\frac{1}{1-\lambda} + \lambda\right) y^T S_1 y + \left(1 - \frac{1}{1-\lambda}\right) y^T D y - \frac{1}{\lambda} y^T C^T (C S_1^{-1} C^T)^{-1} C y \geq 0.$$

Rewriting the inequality in terms of $\tilde{y} = S_1^{1/2} y$ gives

$$\left(\frac{1}{1-\lambda} + \lambda\right) \tilde{y}^T\tilde{y} + \left(1 - \frac{1}{1-\lambda}\right) \tilde{y}^T S_1^{-1/2} D S_1^{-1/2} \tilde{y} - \frac{1}{\lambda} \tilde{y}^T P \tilde{y} \geq 0,$$

where P is the orthogonal projector defined in (4.10). For the second term, note that $S_1^{-1/2} D S_1^{-1/2}$ is similar to $S_1^{-1} D$ which has all eigenvalues between 0 and 1 (by Lemma 4.2). Thus, both $\tilde{y}^T S_1^{-1/2} D S_1^{-1/2} \tilde{y}$ and $\tilde{y}^T P \tilde{y}$ are less than or equal to $\tilde{y}^T\tilde{y}$ and we can therefore write

$$\left(\frac{1}{1-\lambda} + \lambda\right) \tilde{y}^T\tilde{y} + \left(1 - \frac{1}{1-\lambda}\right) \tilde{y}^T\tilde{y} - \frac{1}{\lambda} \tilde{y}^T\tilde{y} \geq 0,$$

which, after dividing by $\tilde{y}^T\tilde{y}$ and simplifying, yields

$$\lambda^2 + \lambda - 1 \leq 0.$$

This along with the assumption that $\lambda < 0$ gives the desired bound of $\lambda \geq -\frac{1+\sqrt{5}}{2}$.

Lower bound on positive eigenvalues. Assume that $0 < \lambda < 1$ (we know from Theorem 4.1 that the lower bound is in this interval). Substituting (4.7) into (4.6b)

and solving for y gives

$$y = \underbrace{\left(\frac{1}{1-\lambda}BA^{-1}B^T + D + \lambda S_1\right)^{-1}}_{=:Q} C^T z, \tag{4.11}$$

When $0 < \lambda < 1$, the value $\frac{1}{1-\lambda}$ is positive, so we are guaranteed that Q in (4.11) is positive definite. If $z \in \ker(C^T)$, (4.11) gives $y = 0$ and (4.7) gives $x = 0$, and we can see from the eigenvalue equations (4.6a)-(4.6c) that this eigenvector corresponds to $\lambda = 1$ (because $Ez = S_2 z$ for $z \in \ker(C^T)$). This contradicts our assumption that $\lambda < 1$; thus, we assume $z \notin \ker(C^T)$.

We can then write (4.6c) as

$$C\left(\frac{1}{1-\lambda}BA^{-1}B^T + D + \lambda S_1\right)^{-1} C^T z + (1-\lambda)Ez - \lambda C S_1^{-1} C^T z = 0. \tag{4.12}$$

When $0 < \lambda < 1$, we have $\frac{1}{1-\lambda} > 1$. Therefore, if we take the inner product of z^T with (4.12), replace D by $\frac{1}{1-\lambda}D$, and drop the non-negative term $(1-\lambda)z^T Ez$, we obtain the inequality

$$z^T C\left(\left(\frac{1}{1-\lambda} + \lambda\right) S_1\right)^{-1} C^T z - \lambda z^T C S_1^{-1} C^T z \leq 0.$$

After simplifying and dividing by $z^T C S_1^{-1} C^T z$, we obtain

$$\lambda^3 - \lambda^2 - 2\lambda + 1 \leq 0. \tag{4.13}$$

This, combined with the assumption that $0 < \lambda < 1$, gives us that λ must be greater than or equal to the smaller positive root of (4.13), as required.

Upper bound on negative eigenvalues. Assume that $\lambda < 0$. We begin from Equation (4.8) and note that $\lambda S_2 - E$ is negative definite. Therefore,

$$\left(\frac{1}{1-\lambda} + \lambda\right) y^T S_1 y + \left(1 - \frac{1}{1-\lambda}\right) y^T D y \leq 0.$$

Since $1 - \frac{1}{1-\lambda} > 0$, the $y^T D y$ term is non-negative and can be dropped while keep-

ing the inequality, giving us

$$\left(\frac{1}{1-\lambda}+\lambda\right) y^T S_1 y \leq 0.$$

We thus require

$$\frac{1}{1-\lambda}+\lambda \leq 0 \text{ with } \lambda < 0,$$

which leads to the desired bound $\lambda \leq \frac{1-\sqrt{5}}{2}$. □

Remark 4.7. The bounds in Theorem 4.5 are different than the bounds of Theorem 4.4: the lower bound on the negative eigenvalues is looser, and the inclusion set for positive eigenvalues is a single interval that strictly contains and is larger than the union of the two positive intervals in Theorem 4.4, $I_2 \cup I_3$. While the lower bound on negative eigenvalues of Theorem 4.4 may be obtained in the proof of Theorem 4.5 by assuming $D = 0$, the two positive intervals in Theorem 4.4 are obtained thanks to the specific technique of proof we use in that proof, and we believe that they cannot easily be obtained from the proof of Theorem 4.5 or by other means.

4.2 Bounds for block diagonal preconditioners with approximations of Schur complements

In practice, it is too expensive to invert A, S_1, and S_2 exactly. In this section we examine eigenvalue bounds on matrices of the form $\tilde{\mathscr{M}}^{-1}\mathscr{K}$, where $\tilde{\mathscr{M}}$ uses symmetric positive definite and ideally spectrally equivalent approximations for A, S_1, and S_2. Specifically, we consider the approximate block diagonal preconditioner

$$\tilde{\mathscr{M}} = \begin{bmatrix} \tilde{A} & 0 & 0 \\ 0 & \tilde{S}_1 & 0 \\ 0 & 0 & \tilde{S}_2 \end{bmatrix}, \tag{4.14}$$

with $\tilde{A}$, $\tilde{S}_1$ and $\tilde{S}_2$ satisfying the following:

Assumption 4.8. *Let $\Lambda(\cdot)$ denote the spectrum of a matrix. The diagonal blocks of the approximate preconditioner $\tilde{\mathscr{M}}$, given by $\tilde{A}$, $\tilde{S}_1$, and $\tilde{S}_2$, satisfy:*

i. $\Lambda(\tilde{A}^{-1}A) \in [\alpha_0, \beta_0]$;

ii. $\Lambda(\tilde{S}_1^{-1}S_1) \in [\alpha_1, \beta_1]$;

iii. $\Lambda(\tilde{S}_2^{-1}S_2) \in [\alpha_2, \beta_2]$,

where $0 < \alpha_i \leq 1 \leq \beta_i$.

We note that to obtain spectral equivalence we seek approximations that yield values of α_i independent of the mesh size and bounded uniformly away from zero. It is also worth noting that the values of α_i and β_i are typically not explicitly available. We briefly address this at the end of this section, following our derivation of the bounds.

To simplify our analyses, we define:

$$\tilde{A}^{-1}A =: Q_0; \tag{4.15a}$$

$$\tilde{S}_1^{-1}S_1 =: Q_1; \tag{4.15b}$$

$$\tilde{S}_2^{-1}S_2 =: Q_2. \tag{4.15c}$$

To derive the eigenvalue bounds, we consider the split preconditioned matrix

$$\tilde{\mathscr{M}}^{-1/2}\mathscr{K}\tilde{\mathscr{M}}^{-1/2} = \begin{bmatrix} \tilde{Q}_0 & \tilde{B}^T & 0 \\ \tilde{B} & -\tilde{D} & \tilde{C}^T \\ 0 & \tilde{C} & \tilde{E} \end{bmatrix}, \tag{4.16}$$

where $\tilde{Q}_0 = \tilde{A}^{-1/2}A\tilde{A}^{-1/2}$, $\tilde{B} = \tilde{S}_1^{-1/2}B\tilde{A}^{-1/2}$, $\tilde{C} = \tilde{S}_2^{-1/2}C\tilde{S}_1^{-1/2}$, $\tilde{D} = \tilde{S}_1^{-1/2}D\tilde{S}_1^{-1/2}$, and $\tilde{E} = \tilde{S}_2^{-1/2}E\tilde{S}_2^{-1/2}$. We proceed by bounding the eigenvalues and singular values of the blocks of $\tilde{\mathscr{M}}^{-1/2}\mathscr{K}\tilde{\mathscr{M}}^{-1/2}$ and then applying the results for general double saddle-point matrices presented in Section 3.2. In order to avoid providing internal eigenvalue bounds equal to zero (see Remark 3.5) we assume here that B and C have full row rank.

Lemma 4.9. *When B and C have full row rank, bounds on the eigenvalues/singular values of the blocks of the matrix* $\tilde{\mathscr{M}}^{-1/2}\mathscr{K}\tilde{\mathscr{M}}^{-1/2}$ *are as follows:*

- $\tilde{Q}_0$*: eigenvalues are in* $[\alpha_0, \beta_0]$.
- $\tilde{B}$*: singular values are in* $\left[\sqrt{\frac{\alpha_0\alpha_1}{1+\eta_D}}, \sqrt{\beta_0\beta_1}\right]$*, where* η_D *is the maximal eigenvalue of* $(BA^{-1}B^T)^{-1}D$.

- *$\tilde{C}$: singular values are in* $\left[\sqrt{\frac{\alpha_1\alpha_2}{1+\eta_E}}, \sqrt{\beta_1\beta_2}\right]$*, where* η_E *is the maximal eigenvalue of* $(CS_1^{-1}C^T)^{-1}E$.
- *$\tilde{D}$: eigenvalues are in* $[0, \beta_1]$.
- *$\tilde{E}$: eigenvalues are in* $[0, \beta_2]$.

Proof. The eigenvalue bounds on $\tilde{Q}_0$ follow from the fact that $\tilde{Q}_0$ is similar to Q_0. For $\tilde{D}$ and $\tilde{E}$, the lower bounds follow from the fact that D and E are semidefinite (if D and/or E are definite, this bound will be loose, but we use the zero bound to simplify some results that we present in this section). For the upper bound on $\tilde{D}$, we note that $\tilde{D} = \tilde{S}_1^{-1/2} D \tilde{S}_1^{-1/2}$, which is similar to $\tilde{S}_1^{-1} D = Q_1 S_1^{-1} D$. The eigenvalues of $S_1^{-1}D$ are less than or equal to 1 by Lemma 4.2, meaning that those of $Q_1 S_1^{-1} D$ are less than or equal to β_1. Analogous reasoning gives the upper bound for $\tilde{E}$.

We now present the results for $\tilde{B}$. For the upper bound, note that the matrix $\tilde{B}\tilde{B}^T = \tilde{S}_1^{-1/2} B \tilde{A}^{-1} B^T \tilde{S}_1^{-1/2}$ is similar to $\tilde{S}_1^{-1} B \tilde{A}^{-1} B^T = Q_1 S_1^{-1} B \tilde{A}^{-1} B^T$. Because the eigenvalues of Q_1 are in $[\alpha_1, \beta_1]$, we need only bound the eigenvalues of $S_1^{-1} B \tilde{A}^{-1} B^T$. These are the same as the nonzero eigenvalues of

$$\tilde{A}^{-1} B^T S_1^{-1} B = Q_0 A^{-1} B^T S_1^{-1} B.$$

The nonzero eigenvalues of $A^{-1}B^T S_1^{-1} B$ are the same as those of $S_1^{-1} B A^{-1} B^T$, which are all less than or equal to 1 by Lemma 4.2. Thus, the eigenvalues of $Q_0 A^{-1} B^T S_1^{-1} B$ are less than or equal to β_0, from which we conclude that the eigenvalues of $\tilde{B}\tilde{B}^T$ are less than or equal to $\beta_0\beta_1$, giving an upper singular value bound of $\sqrt{\beta_0\beta_1}$. Similarly, a lower bound on the eigenvalues of $\tilde{B}\tilde{B}^T$ is given by $\alpha_0\alpha_1$ times a lower bound on the eigenvalues of $S_1^{-1} B A^{-1} B^T$. Because we have assumed that B is full rank, $BA^{-1}B^T$ is invertible, implying that

$$S_1^{-1} B A^{-1} B^T = \left((BA^{-1}B^T)^{-1} S_1\right)^{-1} = \left(I + (BA^{-1}B^T)^{-1} D\right)^{-1}.$$

The stated result then follows because the eigenvalues of $S_1^{-1} B A^{-1} B^T$ are greater than or equal to $\frac{1}{1+\eta_D}$, where η_D is the maximal eigenvalue of $(BA^{-1}B^T)^{-1}D$. Thus, a lower bound on the singular values of $\tilde{B}$ is given by the square root of this value.

The bounds for $\tilde{C}$ are obtained in the same way as those for $\tilde{B}$. □

We can now present bounds on the eigenvalues of $\tilde{\mathscr{M}}^{-1/2}\mathscr{K}\tilde{\mathscr{M}}^{-1/2}$. We define three cubic polynomials:

$$u(\lambda) = \lambda^3 + (\beta_1 - \alpha_0)\lambda^2 - \left(\alpha_0\beta_1 + \frac{\alpha_1\alpha_2}{1+\eta_E} + \beta_0\beta_1\right)\lambda + \frac{\alpha_0\alpha_1\alpha_2}{1+\eta_E}; \quad (4.17a)$$

$$v(\lambda) = \lambda^3 - (\beta_0 + \beta_2)\lambda^2 + (\beta_0\beta_2 - \beta_0\beta_1 - \beta_1\beta_2)\lambda + 2\beta_0\beta_1\beta_2; \quad (4.17b)$$

$$w(\lambda) = \lambda^3 + (\beta_1 - \alpha_0)\lambda^2 - (\alpha_0\beta_1 + \beta_0\beta_1 + \beta_1\beta_2)\lambda + \alpha_0\beta_1\beta_2. \quad (4.17c)$$

These polynomials all have two positive roots and one negative root, by Corollary 3.3. As in Chapter 3, we let u^- denote the (single) negative root of a polynomial u, and let $u^+_{\min}$ and $u^+_{\max}$ respectively denote the smallest and largest positive roots.

Theorem 4.10 (Eigenvalue bounds, matrix $\tilde{\mathscr{M}}^{-1/2}\mathscr{K}\tilde{\mathscr{M}}^{-1/2}$)**.** *When B and C have full row rank, the eigenvalues of $\tilde{\mathscr{M}}^{-1/2}\mathscr{K}\tilde{\mathscr{M}}^{-1/2}$ are bounded within the intervals*

$$\left[w^-, \frac{\beta_0 - \sqrt{\beta_0^2 + \frac{4\alpha_0\alpha_1}{1+\eta_D}}}{2}\right] \bigcup \left[u^+_{\min}, v^+_{\max}\right]. \quad (4.18)$$

Proof. The stated bounds follow from Lemma 4.9 and Theorem 3.4. □

Remark 4.11. Obtaining η_D and η_E requires computation of eigenvalues of two matrices related to the Schur complements S_1 and S_2: $(S_1 - D)^{-1}D = (BA^{-1}B^T)^{-1}D$ and $(S_2 - E)^{-1}E = (CS_1^{-1}C^T)^{-1}E$, respectively. As we have previously mentioned, in practice when solving (1.1) the matrices S_1 and S_2 would typically be approximated by sparse and easier-to-invert matrices, rather than formed and computed explicitly. Therefore, we cannot expect to compute η_D and η_E exactly. Spectral equivalence relations may be helpful in providing reasonable approximations here. For example, in common formulations of the Stokes-Darcy problem [10] the matrix A is a discrete negative Laplacian and B and C are discrete divergence operators (or scaled variations thereof). In such a case, for certain finite element discretizations $BA^{-1}B^T = S_1 - D$ is spectrally equivalent to the mass matrix in the pressure space [17, Section 5.5]. Denote it by Q. Then, estimating the maximal eigenvalue of η_D amounts to computing an approximation to the maximal eigenvalue of $Q^{-1}D$, which is computationally straightforward given the favorable spectral properties of

Q. If the above-mentioned Schur complement is approximated by the mass matrix, then it can be shown that $CS_1^{-1}C^T = S_2 - E$ is strongly related to the scalar Laplacian, and therefore maximal eigenvalue of η_E would be relatively easy to approximate as well.

When $D = 0$, η_D and the maximal eigenvalue of $\tilde{D}$ are zero. Similarly, when $E = 0$, η_E and the maximal eigenvalue of $\tilde{E}$ are zero. In these cases, we can simplify some of the bounds of Theorem 4.10. Some of the cubic polynomials that define the bounds will change in these cases. We will use a bar (e.g., $\bar{u}$) to denote cubic polynomials where $D = 0$, a hat (e.g., $\hat{u}$) to denote the polynomials where $E = 0$, and both (e.g., $\hat{\bar{u}}$) to denote both D and E being zero. We define the following cubic polynomials:

$$\bar{u}(\lambda) = \lambda^3 - \alpha_0\lambda^2 - \left(\frac{\alpha_1\alpha_2}{1+\eta_E} + \beta_0\beta_1\right)\lambda + \frac{\alpha_0\alpha_1\alpha_2}{1+\eta_E}; \tag{4.19a}$$

$$\hat{u}(\lambda) = \lambda^3 + (\beta_1 - \alpha_0)\lambda^2 - (\alpha_0\beta_1 + \alpha_1\alpha_2 + \beta_0\beta_1)\lambda + \alpha_0\alpha_1\alpha_2; \tag{4.19b}$$

$$\hat{\bar{u}}(\lambda) = \lambda^3 - \alpha_0\lambda^2 - (\alpha_1\alpha_2 + \beta_0\beta_1)\lambda + \alpha_0\alpha_1\alpha_2; \tag{4.19c}$$

$$\hat{v}(\lambda) = \lambda^3 - \beta_0\lambda^2 - (\beta_0\beta_1 + \beta_1\beta_2)\lambda + \beta_0\beta_1\beta_2; \tag{4.19d}$$

$$\bar{w}(\lambda) = \lambda^3 - \alpha_0\lambda^2 - (\beta_0\beta_1 + \beta_1\beta_2)\lambda + \alpha_0\beta_1\beta_2. \tag{4.19e}$$

The following results then follow directly from Theorem 4.10; the proof is omitted.

Corollary 4.12. ***In the case*** $D = 0,\ E \neq 0$***:*** *the eigenvalues of* $\mathscr{M}^{-1/2}\mathscr{K}\mathscr{M}^{-1/2}$ *are bounded within the intervals*

$$\left[\bar{w}^-, \frac{\beta_0 - \sqrt{\beta_0^2 + 4\alpha_0\alpha_1}}{2}\right] \bigcup \left[\bar{u}^+_{\min}, v^+_{\max}\right]. \tag{4.20}$$

In the case $D \neq 0,\ E = 0$***:*** *the eigenvalues are bounded in*

$$\left[w^-, \frac{\beta_0 - \sqrt{\beta_0^2 + \frac{4\alpha_0\alpha_1}{1+\eta_D}}}{2}\right] \bigcup \left[\hat{u}^+_{\min}, \hat{v}^+_{\max}\right]. \tag{4.21}$$

In the case $D = 0,\ E = 0$***:*** *the eigenvalues are bounded in*

$$\left[\bar{w}^{-}, \frac{\beta_0 - \sqrt{\beta_0^2 + 4\alpha_0\alpha_1}}{2}\right] \bigcup \left[\hat{\hat{u}}^{+}_{\min}, \hat{v}^{+}_{\max}\right]. \tag{4.22}$$

Remark 4.13. There is some looseness in the bounds of Theorem 4.10 when applied to the exact preconditioning case (i.e., $\alpha_i = \beta_i = 1$). This is a consequence of the fact that Theorem 4.10 is based on the bounds for unpreconditioned matrices, which consider each matrix block individually with the *R*-matrix method, as opposed to the energy estimates approach in Section 4.1, which fully exploits the relationships between the blocks of the preconditioned matrix. For the extremal (lower negative and upper positive) bounds, the looseness is minor: Theorem 4.10 gives a bound of 2 for the positive eigenvalues and approximately -1.9 on the negative eigenvalues, while we know from Theorem 4.5 that tight bounds are approximately 1.8 and -1.6, respectively. For the interior bounds, however, we note that the bounds may be quite loose if η_D or η_E is very large. Fortunately, this is often not a concern in practical settings as having a large η_D or η_E generally implies that the spectral norm of D or E is large relative to that of $BA^{-1}B^T$ or $CS_1^{-1}C^T$, respectively. Often D and/or E are regularization terms, which tend to have a fairly small norm. Nonetheless, it should be acknowledged that the interior bounds may be pessimistic for some problems.

Remark 4.14. The values of α_i and β_i are rarely available in practical applications; they are typically coercivity constants or related quantities, which are proven to be independent of the mesh size but are not known explicitly. The values of β_i should not typically generate a difficulty, and approximating them by 1 or a value close to 1 should provide a reasonable approximation for the bounds.

For the α_i, let us provide some (partial) observations. For the solutions of the quadratic equations in Corollary 4.12, we can use the Taylor approximation $\sqrt{1+x} \lessapprox 1 + \frac{x}{2}$ for $0 < x \ll 1$ to conclude that

$$\frac{\beta_0 - \sqrt{\beta_0^2 + 4\alpha_0\alpha_1}}{2} \gtrapprox -\frac{\alpha_0\alpha_1}{\beta_0}.$$

This means that if α_0 and α_1 are small, then the above displayed expression would be a valid (albeit slightly less tight) upper negative bound, and if the α_i are uniformly bounded away from zero, then so is the bound.

4.3 Numerical experiments

We now consider preconditioning strategies for PDE-constrained optimization matrices defined in (3.13)-(3.14). We note that, in this somewhat simplified problem, optimal preconditioners for the classical saddle-point formulation have been proposed in, e.g., [66, 71], and are shown to be highly effective. We examine eigenvalue bounds with both exact and approximate Schur complements.

For the distributed control problem, we work with the reordered matrix $\mathscr{K}_{\text{flip}}$ (3.14). The Schur complement preconditioner for $\mathscr{K}_{\text{flip}}$ is

$$\mathscr{M} = \begin{bmatrix} \beta M & 0 & 0 \\ 0 & \frac{1}{\beta} M & 0 \\ 0 & 0 & M + \beta K M^{-1} K \end{bmatrix}. \tag{4.23}$$

The first and second blocks are mass matrices, which are cheap to invert, so we leave these terms as they are. For the second Schur complement $S_2 = M + \beta K M^{-1} K$, we use the approximation $\tilde{S}_2 = \left(M + \sqrt{\beta} K\right) M^{-1} \left(M + \sqrt{\beta} K\right)$ proposed by Pearson and Wathen [66] to obtain the preconditioner:

$$\tilde{\mathscr{M}} = \begin{bmatrix} \beta M & 0 & 0 \\ 0 & \frac{1}{\beta} M & 0 \\ 0 & 0 & \left(M + \sqrt{\beta} K\right) M^{-1} \left(M + \sqrt{\beta} K\right) \end{bmatrix}. \tag{4.24}$$

Per [66, Theorem 4], the eigenvalues of $\tilde{S}_2^{-1} S_2$ satisfy $\Lambda(\tilde{S}_2^{-1} S_2) \in \left[\frac{1}{2}, 1\right]$. Thus $\tilde{\mathscr{M}}$ satisfies Theorem 4.8 with $\alpha_0 = \beta_0 = \alpha_1 = \beta_1 = 1$, $\alpha_2 = \frac{1}{2}$, and $\beta_2 = 1$.

Plots of the preconditioned eigenvalues for $h = 2^{-4}$ are shown in Figure 4.1-4.2. The preconditioned matrix eigenvalues are denoted by the blue dots, and eigenvalue bounds are shown by lines. Recall that, for the exact preconditioner shown in Figure 4.1, Theorem 4.4 predicts that the positive eigenvalues will be contained in two intervals. The bounds on the two sub-intervals are denoted by

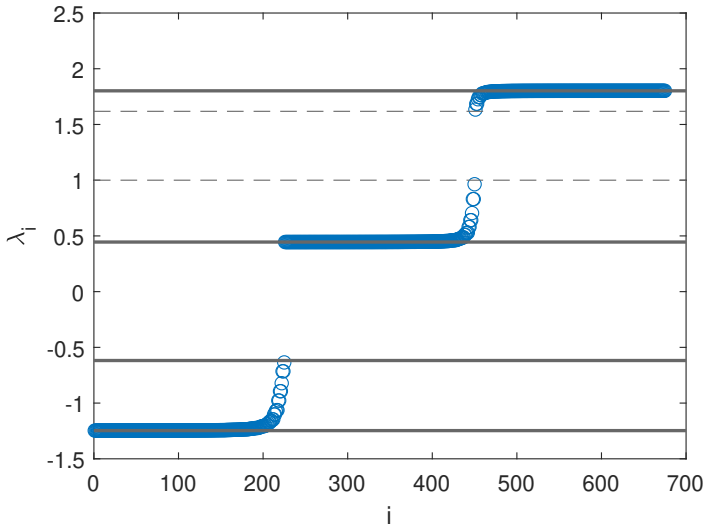

Figure 4.1: Eigenvalue plots for $\mathscr{K}_{\text{flip}}$ with exact preconditioner $\mathscr{M}$. Eigenvalues are shown by the blue circles; eigenvalue intervals predicted by Theorem 4.4 are shown by lines.

dashed lines.

The value η_E, defined as the maximal eigenvalue of

$$(CS_1^{-1}C^T)^{-1}E = \beta(KM^{-1}K)^{-1}M,$$

is approximately 2.6×10^{-7}. We note that for 2D problems with uniform **Q1** finite element discretizations the value η_E is $O(\beta h^4)$, and will thus be small in general.

Comparing Figures 4.1 and 4.2, we notice that the bounds on the negative eigenvalue bounds do not change when we use the approximate Schur complement, but the lower positive bound becomes smaller (from 0.4450 with the exact Schur complement to 0.2929 for the approximate Schur complement) and the upper positive bound becomes larger (1.8019 in the exact case and 2 in the approximate case). We note that the eigenvalues appear to be very close to the predicted bounds except for the upper positive eigenvalues of $\tilde{\mathscr{M}}^{-1}\mathscr{K}$. As discussed in Remark 3, this kind of minor looseness in the upper bound can happen when we have highly accurate Schur complement approximations (as in this case, where we are exactly

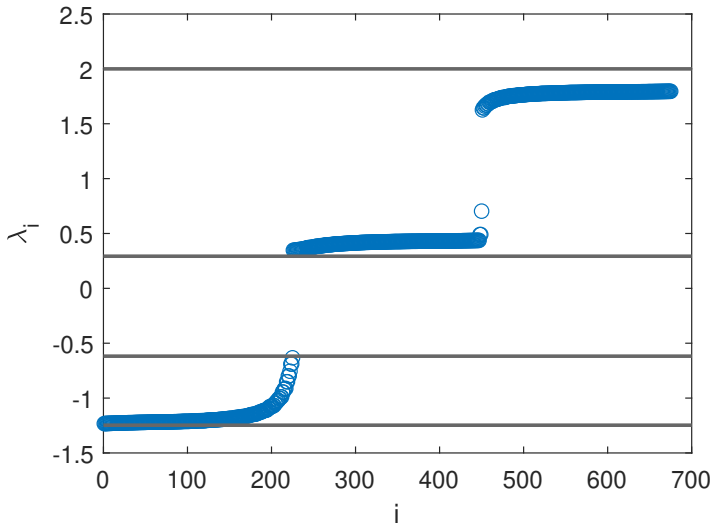

Figure 4.2: Eigenvalue plots for $\mathscr{K}_{\text{flip}}$ with approximate preconditioner $\tilde{\mathscr{M}}$. Eigenvalues are shown by the blue circles; eigenvalue bounds from Corollary 4.12 are shown by lines.

inverting the two M blocks).

For the matrix $\mathscr{K}_\partial$ arising from the boundary control problem, the Schur complement preconditioner is

$$\mathscr{M}_\partial = \begin{bmatrix} M & 0 & 0 \\ 0 & KM^{-1}K & 0 \\ 0 & 0 & \beta M_b + E^T \left(KM^{-1}K\right)^{-1} E \end{bmatrix}. \tag{4.25}$$

In practice, the first Schur complement $KM^{-1}K$ can be inverted approximately with, for example, a multigrid method. For the second Schur complement, we note that for the 2D Poisson control problem on a uniform **Q1** grid, the eigenvalues of K are between $O(h^2)$ and $O(1)$, while those of the mass matrices are all $O(h^2)$. Therefore, the term βM_b will dominate $E^T \left(KM^{-1}K\right)^{-1} E$ for all but very small values of β (which are not commonly used in practice). Therefore, an approximate

preconditioner for $\mathscr{K}_\partial$ is given by

$$\tilde{\mathscr{M}}_\partial = \begin{bmatrix} M & 0 & 0 \\ 0 & KM^{-1}K & 0 \\ 0 & 0 & \beta M_b \end{bmatrix}. \tag{4.26}$$

We note that this preconditioner is presented in [71], though it is derived there from the block-2×2 formulation of $\mathscr{K}$.

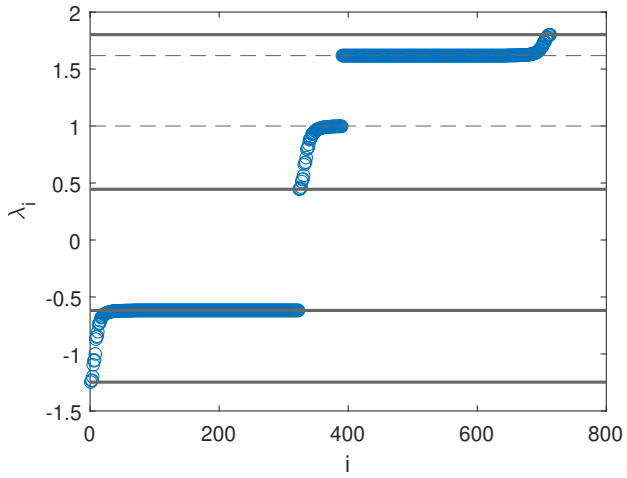

Figure 4.3: Eigenvalue plots for the boundary control problem $\mathscr{K}_\partial$, preconditioned by exact preconditioner $\mathscr{M}$. Eigenvalue intervals predicted by Theorem 4.4) are shown by horizontal lines.

Plots of the preconditioned eigenvalues for $h = 2^{-4}$ are shown in Figures 4.3-4.4. For $\tilde{\mathscr{M}}_\partial^{-1}\mathscr{K}_\partial$, we note that there is a single large-magnitude positive eigenvalue and a large-magnitude negative eigenvalue, whose absolute values are nearly 200, and they make it difficult to see how the other eigenvalues compare to to those of $\mathscr{M}_\partial^{-1}\mathscr{K}_\partial$. Thus in Figure 4.4, we omit the largest and smallest eigenvalues of $\tilde{\mathscr{M}}_\partial^{-1}\mathscr{K}_\partial$ and overlay the others on the corresponding eigenvalues of $\mathscr{M}_\partial^{-1}\mathscr{K}_\partial$. We notice that most other eigenvalues of $\tilde{\mathscr{M}}_\partial^{-1}\mathscr{K}_\partial$ remain close to those of $\mathscr{M}_\partial^{-1}\mathscr{K}_\partial$. We also note that the performance of this preconditioner depends on β: in particular, the performance of the Schur complement approximation deteriorates for small β. We refer to [66, 71] for further discussion of this.

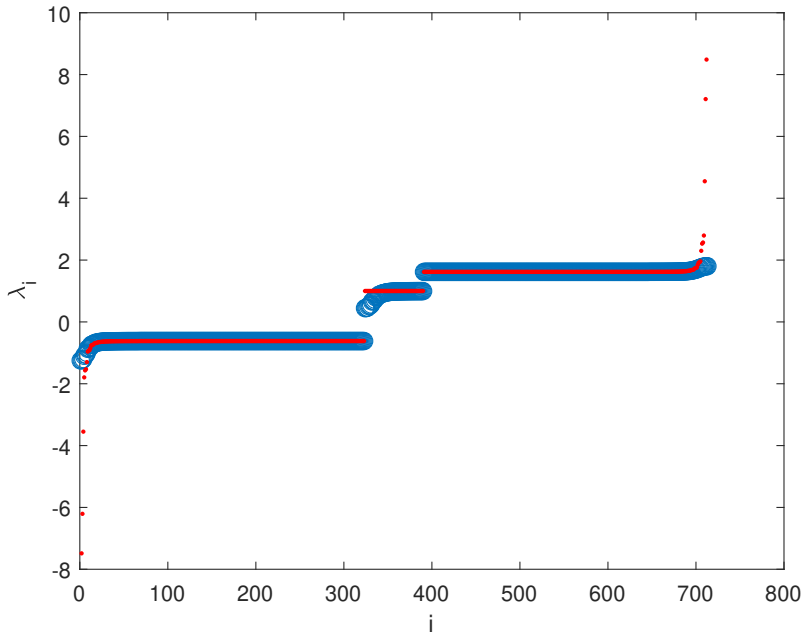

Figure 4.4: All but the single smallest and largest eigenvalues of $\mathscr{M}_\partial^{-1}\mathscr{K}_\partial$ (blue) and $\tilde{\mathscr{M}}_\partial^{-1}\mathscr{K}_\partial$ (red).

It is evident from Figure 4.3 that our bounds are tight and effective. Unlike in the boundary control example, the Schur complement approximation $\tilde{S}_2$ used here does not have β- and h-independent constants α_2, β_2 such that $\Lambda(\tilde{S}_2^{-1}S_2) \in [\alpha_2, \beta_2]$, so our analyses on Schur complement approximations in Section 4.2 are difficult to apply. Nonetheless, we observe from Figure 4.4 that most of the eigenvalues of $\tilde{\mathscr{M}}_\partial^{-1}\mathscr{K}_\partial$ remain very close to those of $\mathscr{M}_\partial^{-1}\mathscr{K}_\partial$. Thus, we see that the eigenvalue bounds for the "ideal" Schur complement preconditioner may still be of use, provided that we have an effective approximation of the Schur complement.

Chapter 5

Eigenvalue bounds when A is singular

In this chapter we consider the unregularized classical saddle-point system

$$\begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix}, \tag{5.1}$$

where $A \in \mathbb{R}^{n\times n}$ is symmetric positive semidefinite and $B \in \mathbb{R}^{m\times n}$ has full row rank, with $m < n$. We denote the coefficient matrix by

$$\mathscr{A}_0 = \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix}. \tag{5.2}$$

We assume throughout that $\mathscr{A}_0$ is invertible. Our goal in this chapter is to derive eigenvalue bounds for $\mathscr{A}_0$ under the assumption that A is singular.

To illustrate the challenge posed by the problem in hand, recall the following result of Rusten and Winther [76, Lemma 2.1]. We note that in their analyses it is assumed that A is positive definite (as opposed to semidefinite); however, the proof of this lemma does not rely on this, so the result still holds when A is semidefinite.

Lemma 5.1. *The eigenvalues of $\mathscr{A}_0$ are bounded in the union of intervals*

$$I^- \cup I^+,$$

where

$$I^- = \left[\frac{1}{2}(\mu_{\min} - \sqrt{\mu_{\min}^2 + 4\sigma_{\max}^2}), \frac{1}{2}(\mu_{\max} - \sqrt{\mu_{\max}^2 + 4\sigma_{\min}^2})\right]$$

and

$$I^+ = \left[\mu_{\min}, \frac{1}{2}(\mu_{\max} + \sqrt{\mu_{\max}^2 + 4\sigma_{\max}^2})\right].$$

When A is singular, the upper bounds on both positive and negative values of $\mathscr{A}_0$ are unchanged, and the lower negative bound reduces to $-\sigma_{\max}$. However, the lower bound on positive eigenvalues reduces to zero, which is not a useful bound. In particular, when the null spaces of A and B are well separated, the matrix $\mathscr{A}_0$ may be well-conditioned. In this work, we derive a nonzero bound on the positive eigenvalues of $\mathscr{A}_0$ by considering the principal angles between the ranges/kernels of A and B. Ruiz et al. [75] also developed a lower positive eigenvalue bound using principal angles, though their analyses assume a positive definite A.

In section 5.1 we discuss our general approach of augmenting the leading block of a saddle-point matrix to obtain a lower bound on the positive eigenvalues. In section 5.2 we provide our bounds, which rely on the angles between the kernel of A and B. We then present numerical observations in 5.3.

5.1 Lower positive eigenvalue bounds using leading block augmentation

Example. As a motivating example that illustrates the range of possibilities, consider the coefficient matrix

$$\mathscr{A}_0 = \begin{bmatrix} 1 & 0 & b_1 \\ 0 & 0 & b_2 \\ b_1 & b_2 & 0 \end{bmatrix} \text{ where } A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \text{ and } B = \begin{bmatrix} b_1 & b_2 \end{bmatrix}, \tag{5.3}$$

with $b_1^2 + b_2^2 = 1$ and $b_1, b_2 > 0$. The eigenvalues of A and singular value of B are the same for all such b_1, b_2, but the lowest positive eigenvalue of $\mathscr{A}_0$ varies depending on b_1 and b_2. The eigenvalues λ of $\mathscr{A}_0$ are the roots of the cubic polynomial $p(\lambda) = \lambda^3 - \lambda^2 - \lambda + b_2^2$. This polynomial has two positive roots and and one

negative root, by Corollary 3.3; the smaller positive root approaches zero as b_2 goes to zero (i.e., when A and B have overlapping null spaces), but as b_2 goes to 1 (i.e., when A and B have orthogonal null spaces) the smaller positive root approaches 1.

We now present a general approach for deriving nonzero bounds for the lower positive eigenvalues of $\mathscr{A}_0$ when A is singular. We recall the following result [22, 32]:

Lemma 5.2. *Let*

$$\mathscr{A}_0(W) = \begin{bmatrix} A + B^T W B & B^T \\ B & 0 \end{bmatrix}, \tag{5.4}$$

where $W \in \mathbb{R}^{m\times m}$. If $\mathscr{A}_0$ and $\mathscr{A}_0(W)$ are both nonsingular, then

$$\mathscr{A}_0^{-1} = (\mathscr{A}_0(W))^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & W \end{bmatrix}. \tag{5.5}$$

We will assume that W is positive semidefinite and the leading block $A_W := A + B^T W B$ of $\mathscr{A}_0(W)$ is positive definite. We can use this along with (5.5) to derive a nonzero bound on the lower positive eignvalues of $\mathscr{A}_0$, using a free matrix parameter W.

Theorem 5.3. *Let $W \in \mathbb{R}^{m\times m}$ be a symmetric positive semidefinite matrix. Then the positive eigenvalues of $\mathscr{A}_0$ are greater than or equal to*

$$\min\left\{\mu_{\min}(A_W), \frac{1}{\mu_{\max}(W)}\right\}.$$

Proof. We derive a lower bound on the positive eigenvalues of $\mathscr{A}_0$ by considering an upper bound on the eigenvalues of $\mathscr{A}_0^{-1}$. By combining [7, Equation (3.4)] and (5.5), we obtain

$$\mathscr{A}_0^{-1} = \begin{bmatrix} A_W^{-1} - A_W^{-1} B^T S_W^{-1} B A_W^{-1} & A_W^{-1} B^T S_W^{-1} \\ S_W^{-1} B A_W^{-1} & -S_W^{-1} + W \end{bmatrix}, \tag{5.6}$$

where $S_W = BA_W^{-1}B^T$. Notice that we can write

$$\mathscr{A}_0^{-1} = \begin{bmatrix} A_W^{-1} & 0 \\ 0 & W \end{bmatrix} - \begin{bmatrix} A_W^{-1}B^T \\ -I \end{bmatrix} S_W^{-1} \begin{bmatrix} BA_W^{-1} & -I \end{bmatrix}.$$

Because the subtracted term is positive semidefinite, we conclude that the eigenvalues of $\mathscr{A}_0^{-1}$ are less than or equal to the eigenvalues of

$$\begin{bmatrix} A_W^{-1} & 0 \\ 0 & W \end{bmatrix}.$$

The stated result follows. □

5.2 Augmentation-based bounds when $W = \gamma I$

As in section 5.1, we consider the augmented matrix $\mathscr{A}_0(W)$, but in this case we restrict ourselves to the case where

$$W = \gamma I,$$

as is done in [20, 32]. For simplicity we write $A_\gamma = A + \gamma B^T B$ and $\mathscr{A}_\gamma = \mathscr{A}_0(\gamma I)$. In this case, the lower bound on positive eigenvalues presented in Theorem 5.3 reduces to $\min\left\{\mu_{\min}(A_\gamma), \frac{1}{\gamma}\right\}$.

We first consider the special case where $\operatorname{rank}(A) = n - m$. We say here that A is *lowest-rank* because if its rank were any lower then $\mathscr{A}_0$ would necessarily be singular. It was shown in [19, 20] that A_γ and $\mathscr{A}_\gamma$ have unique properties, which we will use here to refine the bound on lower positive eigenvalues given in Theorem 5.3. We return in Section 5.2.2 to the general case, where A is assumed to be rank-deficient but not lowest-rank.

5.2.1 Bounds when A is lowest-rank

Theorem 5.4. *When* $\operatorname{rank}(A) = n - m$*, we have*

$$\mu_{\min}(A_\gamma) \geq \rho \cdot \min\left\{\mu_{\min}^+(A), \gamma\sigma_{\min}^2(B)\right\}, \tag{5.7}$$

where $\rho \leq 1$ is a constant that does not depend on γ.

Proof. We begin by writing a decomposition of A_γ as was done in [20]. Let

$$A = U\Lambda U^T, \;\; B = QSV^T$$

be the economy-size singular value decompositions of A and B.

The matrices Λ and U comprise of the eigenpairs of A that correspond to its nonzero eigenvalues, and the columns of V are the set of eigenvectors of B^TB that correspond to its nonzero eigenvalues. We can then write

$$A_\gamma = P\Sigma P^T, \tag{5.8}$$

where

$$P = \begin{bmatrix} U & V \end{bmatrix}, \;\; \Sigma = \begin{bmatrix} \Lambda & 0 \\ 0 & \gamma S^2 \end{bmatrix}.$$

The decomposition in (5.8) resembles an eigenvalue decomposition, but is not an eigenvalue decomposition in general because the columns of V will not be orthogonal to those of U.

We then derive a lower bound on the eigenvalues of A_γ by obtaining an upper bound on the eigenvalues of A_γ^{-1}. We can write

$$\mu_{\max}(A_\gamma^{-1}) = ||A_\gamma^{-1}|| = ||P^{-T}\Sigma^{-1}P^{-1}|| \leq ||\Sigma^{-1}|| \cdot ||P^{-1}||^2.$$

The largest eigenvalue of Σ^{-1} is equal to $\max\left\{\frac{1}{\mu_{\min}^+}, \frac{1}{\gamma\sigma_{\min}^2}\right\}$. The stated result follows by setting $\rho = ||P^{-1}||^{-2}$. We note that $\rho \leq 1$, with equality when U and V are mutually orthogonal (that is, when the range of A is orthogonal to the range of B^T). To show this is the case, consider $x \in \ker(A)$. We then have

$$P^Tx = \begin{bmatrix} U^Tx \\ V^Tx \end{bmatrix} = \begin{bmatrix} 0 \\ V^Tx \end{bmatrix}.$$

Since V is orthogonal, this gives $||P^Tx|| \leq ||x||$. Defining $q = P^Tx$, this implies that

$$||q|| \leq ||P^{-T}q||,$$

meaning that $||P^{-T}||$ (and therefore $||P^{-1}||$) is greater than or equal to 1. Thus, $\rho \leq 1$.

□

We now provide a value for $\rho = ||P^{-1}||^{-2}$ in terms of the principal angles between range(A) and range(B^T). Let

$$\theta_i, i = 1, \ldots, m$$

denote these angles. The cosines $\cos(\theta_i)$ of these angles are given by the singular values of $U^T V$ (or $V^T U$).

Lemma 5.5. *Let* $\theta_{\min}$ *denote the minimum principal angle between* range(A) *and* range(B^T). *Then*

$$||P^{-1}|| = \frac{1}{\sqrt{1 - \cos(\theta_{\min})}},$$

which implies that ρ *defined in* (5.7) *is given by*

$$\rho = 1 - \cos(\theta_{\min}).$$

Proof. We proceed by analyzing the eigenvalues of $P^T P$, using the fact that

$$||P^{-1}|| = \frac{1}{\sqrt{\mu_{\min}(P^T P)}}.$$

We write $P^T P$ in block form:

$$P^T P = \begin{bmatrix} U^T \\ V^T \end{bmatrix} \begin{bmatrix} U & V \end{bmatrix} = \begin{bmatrix} I & U^T V \\ V^T U & I \end{bmatrix}.$$

The (1,1)-block of $P^T P$ is size $(n-m) \times (n-m)$ and the (2,2)-block is size $m \times m$. We now assume without loss of generality that $n - m \geq m$. (If $n - m < m$, we can reorder the blocks of $P^T P$ such that the (1,1)-block is larger, and use the same analysis as below.)

Letting $v = \begin{bmatrix} x^T & y^T \end{bmatrix}^T$ be an appropriately partitioned eigenvector, we write

the eigenvalue equations for P^TP:

$$x + U^TVy = \lambda x; \tag{5.9a}$$

$$V^TUx + y = \lambda y. \tag{5.9b}$$

There is an eigenvalue $\lambda = 1$ with multiplicity $n-2m$, which we observe by choosing $x \in \ker(V^TU)$ and $y = 0$. For the remaining $2m$ eigenvalues, we assume $\lambda \neq 1$. From (5.9a) we have $x = \frac{1}{\lambda-1}U^TVy$, which we substitute into (5.9b) to obtain

$$y = \frac{1}{(\lambda-1)^2}V^TUU^TVy. \tag{5.10}$$

The eigenvalues of V^TUU^TV are given by $\cos^2(\theta_i)$, where θ_i are the principal angles between range(A) and range(B^T). Thus, for each θ_i we can write (5.10) as

$$y = \frac{\cos^2(\theta_i)}{(\lambda_i-1)^2}y,$$

implying that

$$\lambda_i = 1 \pm \cos(\theta_i).$$

Thus each θ_i yields two distinct eigenvalues. Together with the $n-2m$ eigenvalues with $\lambda = 1$, this accounts for all n eigenvalues of P^TP. Therefore, the smallest eigenvalue of P^TP is given by $1-\cos(\theta_{\min})$; the stated result follows. □

We can use the results we have established for matrices with lowest-rank A to derive a lower bound on the positive eigenvalues of $\mathscr{A}_0$ that does not require us to know the eigenvalues of A_γ. We saw in Theorem 5.3 that for $W = \gamma I$, the bound is given by $\min\left\{\mu_{\min}(A_\gamma), \frac{1}{\gamma}\right\}$. As γ decreases, the value of $\mu_{\min}(A_\gamma)$ approaches zero (because A_γ approaches A); thus, we achieve the best possible lower bound when

$$\frac{1}{\gamma} = \mu_{\min}(A_\gamma).$$

Since we do not generally know the value of $\mu_{\min}(A_\gamma)$, we can instead select $\frac{1}{\gamma}$ to be equal to the reciprocal of the lower bound on $\mu_{\min}(A_\gamma)$ given by Theorem 5.4

and Lemma 5.5. That is, we find a γ that satisfies

$$\frac{1}{\gamma} = (1-\cos(\theta_{\min}))\min\left\{\mu^{+}_{\min}, \gamma\sigma^{2}_{\min}\right\}.$$

Depending on which of the arguments to the min function is smaller, we either have

$$\frac{1}{\gamma} = \mu^{+}_{\min}(1-\cos(\theta_{\min}))$$

or we have $\frac{1}{\gamma} = (1-\cos(\theta_{\min}))\cdot\gamma\sigma^{2}_{\min}$, which implies that

$$\frac{1}{\gamma} = \sigma_{\min}\sqrt{1-\cos(\theta_{\min})}.$$

Therefore, if we select

$$\frac{1}{\gamma} = \min\left\{\mu^{+}_{min}(1-\cos(\theta_{\min})), \sigma_{\min}\sqrt{1-\cos(\theta_{\min})}\right\},$$

we know that $\mu_{\min}(A_\gamma)$ will be greater than or equal to this value of $\frac{1}{\gamma}$. This gives the following result:

Theorem 5.6. *When* $\operatorname{rank}(A) = n-m$*, the positive eigenvalues of* $\mathscr{A}_0$ *are greater than or equal to*

$$\min\left\{\mu^{+}_{\min}(1-\cos(\theta_{\min})), \sigma_{\min}\sqrt{1-\cos(\theta_{\min})}\right\}.$$

In some problems such as the Maxwell equation, knowledge of the null spaces of A and B may be more readily available than the ranges of A and B^T [39]. For these settings, it is convenient to re-frame the result of Theorem 5.6 to rely on the angle between kernels rather than the angle between ranges. Because $\ker(A)$ and $\ker(B)$ are respectively orthogonal to $\operatorname{range}(A)$ and $\operatorname{range}(B^T)$, the principal angles are the same between both pairs of subspaces. The following result then holds.

Corollary 5.7. *Let* $\operatorname{rank}(A) = n-m$ *and let* $\psi_{\min}$ *denote the minimum principal angle between* $\ker(A)$ *and* $\ker(B)$*. The positive eigenvalues of* $\mathscr{A}_0$ *are greater than*

or equal to

$$\min\left\{\mu^{+}_{\min}(1-\cos(\psi_{\min})), \sigma_{\min}\sqrt{1-\cos(\psi_{\min})}\right\}.$$

5.2.2 Bounds when A is not lowest-rank

We now return to the case in which A is rank-deficient but not lowest-rank, and discuss how the results of the previous section can be extended to this case. Let us denote the eigenvalue decomposition of A by:

$$A = U\Lambda U^T.$$

Let $\Lambda^{\max}_{n-m}$ be a diagonal matrix of the $n-m$ largest eigenvalues of Λ and $\Lambda^{\min}_{m}$ be a diagonal matrix of the m smallest. Similarly, let $U^{\max}_{n-m}$ denote the eigenvectors corresponding to the $n-m$ largest eigenvalues and $U^{\min}_{m}$ the eigenvectors corresponding to the m smallest eigenvalues. We then have

$$A = \begin{bmatrix} U^{\max}_{n-m} & U^{\min}_{m} \end{bmatrix} \begin{bmatrix} \Lambda^{\max}_{n-m} & 0 \\ 0 & \Lambda^{\min}_{m} \end{bmatrix} \begin{bmatrix} \left(U^{\max}_{n-m}\right)^T \\ \left(U^{\min}_{m}\right)^T \end{bmatrix}. \tag{5.11}$$

As before, if we consider a weight matrix $W = \gamma I$, a lower bound on the positive eigenvalues of $\mathscr{A}_0$ is given by

$$\min\left\{\frac{1}{\gamma}, \mu_{\min}(A_\gamma)\right\},$$

as this bound does not depend on the nullity of A. When A is not lowest-rank, the bound of Theorem 5.4 is not immediately applicable. However, we note from (5.11) that

$$A = A^{\max}_{n-m} + A^{\min}_{m}, \tag{5.12}$$

where $A^{\max}_{n-m} = U^{\max}_{n-m}\Lambda^{\max}_{n-m}\left(U^{\max}_{n-m}\right)^T$ is a semidefinite matrix with rank $n-m$ and $A^{\min}_{m} = U^{\min}_{m}\Lambda^{\min}_{m}\left(U^{\min}_{m}\right)^T$ is a semidefinite matrix with rank less than or equal to m. Thus, the eigenvalues of A_γ are all greater than or equal to those of

$$A^{\max}_{n-m} + \gamma B^T B =: A^{\max}_{\gamma}.$$

The eigenvalue μ_{n-m} is the smallest eigenvalue in $\Lambda_{n-m}^{\max}$ and therefore the smallest positive eigenvalue of $A_{n-m}^{\max}$. Let $\tilde{\theta}_{\min}$ denote the minimum principal angle between $\text{range}(A_{n-m}^{\max})$ and $\text{range}(B^T)$. By Theorem 5.4 and Lemma 5.5, we have

$$\mu_{\min}(A_\gamma) \geq \mu_{\min}(A_\gamma^{\max}) \geq \left(1-\cos(\tilde{\theta}_{\min})\right) \cdot \min\left\{\tilde{\mu}_{\min}, \gamma\sigma_{\min}^2\right\}.$$

As we did before, we can select $\frac{1}{\gamma}$ to be equal to the smaller of these two values to obtain a lower bound on the positive eigenvalues of $\mathscr{A}_0$ that does not require forming an augmented matrix. The proof of the following theorem is similar to that of Theorem 5.6 and is omitted.

Theorem 5.8. *Let A be semidefinite with $n-m \leq \text{rank}(A) \leq n$. The positive eigenvalues of $\mathscr{A}_0$ are greater than or equal to*

$$\min\left\{\mu_{n-m}(1-\cos(\tilde{\theta}_{\min})), \sigma_{\min}\sqrt{1-\cos(\tilde{\theta}_{\min})}\right\},$$

where μ_{n-m} denotes the $(n-m)$-th largest eigenvalue of A and $\tilde{\theta}_{\min}$ the smallest principal angle between $\text{range}(B^T)$ and the subspace spanned by the eigenvectors corresponding to the $n-m$ largest eigenvalues of A. (Or, equivalently, $\tilde{\theta}_{\min}$ is the smallest principal angle between $\ker(B)$ and the subspace spanned by the eigenvectors corresponding to the m smallest eigenvalues of A – see Corollary 5.7.)

Remark 5.9. Our approach in deriving the previous result was to convert a non-lowest-rank A into a lowest-rank $\tilde{A}$ by removing the part of the spectrum corresponding to the m smallest eigenvalues. However, removing this part of the spectrum of A is not always a good choice, in that it may lead to an overly pessimistic bound. For example, consider the matrix (with $n=3$ and $m=2$):

$$\mathscr{A}_0 = \left[\begin{array}{ccc|cc} 1 & 0 & 0 & 0 & 1 \\ 0 & \alpha & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{array}\right] =: \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix},$$

where $0 < \alpha < 1$. The positive eigenvalues of $\mathscr{A}_0$ are $\alpha, 1$, and $\frac{1+\sqrt{5}}{2}$. The "non-

removed" eigenvector $U_{n-m}^{\max}$, which is in this case the eigenvector corresponding to $\lambda = 1$, is:

$$U_{n-m}^{\max} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}.$$

Because this eigenvector is in the range of B^T, the value $\tilde{\theta}_{\min}$ is 0, meaning that Theorem 5.8 gives a bound of 0. We would obtain a better bound if, instead of keeping the part of the spectrum of A that corresponds to the eigenvalue $\lambda = 1$, we kept the portion of the spectrum corresponding to $\lambda = \alpha$ (this would in fact give a tight bound of α). However, the issue of optimizing what subspace of range(A) to use in order to obtain a bound is beyond the scope of this thesis.

5.3 Numerical experiments

Here we test our eigenvalue bounds on two real problems. The first is a Maxwell problem described in Equation 2.15. In this problem, A is lowest-rank. Figure 5.1 shows the predicted bound (as a solid line), the actual smallest positive eigenvalue (dashed line) for various values of γ for a Maxwell matrix with $n = 6080$ and $m = 1985$.

The second is a matrix arising from an Interior Point Method (IPM) solution to a QP problem, described in Section 2.1.1. In Figure 5.2 we show the results of our bounds on the first IPM iteration on TOMLAB[1] Problem 17 for which the saddle-point matrix $\mathscr{A}_0$ is numerically singular. This problem has $n = 293$ and $m = 286$. For the particular matrix shown in the experiment below (which arises in the 12th iteration of the IPM), there are 115 "numerically zero" eigenvalues of the leading block (which we define as those less than machine epsilon times the largest eigenvalue of that block).

In both cases the actual smallest positive eigenvalue $\mu_{\min}(\mathscr{K})$ occurs precisely where $\frac{1}{\gamma} = \mu_{\min}(A_\gamma)$. The bounds for the Maxwell matrix are rather tight, in the sense that they are of the same order of magnitude as the eigenvalue (we also see this with Maxwell matrices of other sizes): the predicted eigenvalue bound is 0.0453 while the actual smallest positive eigenvalue is 0.0611.

[1] Test matrices available at https://tomopt.com/tomlab/.

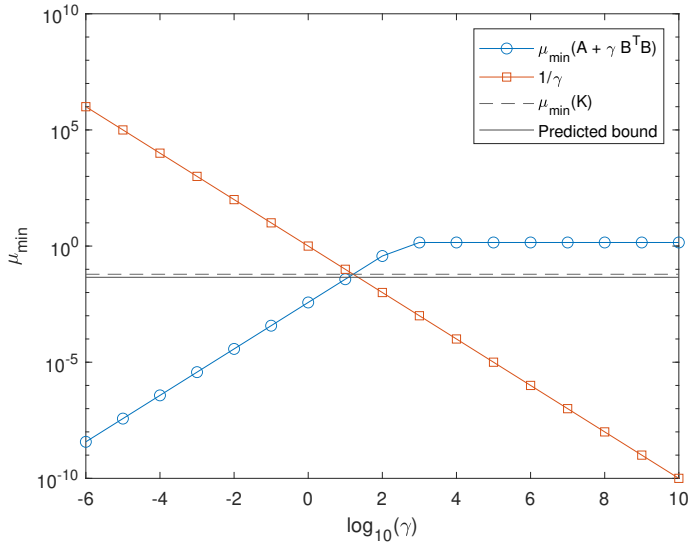

Figure 5.1: Comparison of predicted and actual smallest positive eigenvalue bounds at various values of γ for the Maxwell matrix (lowest rank)

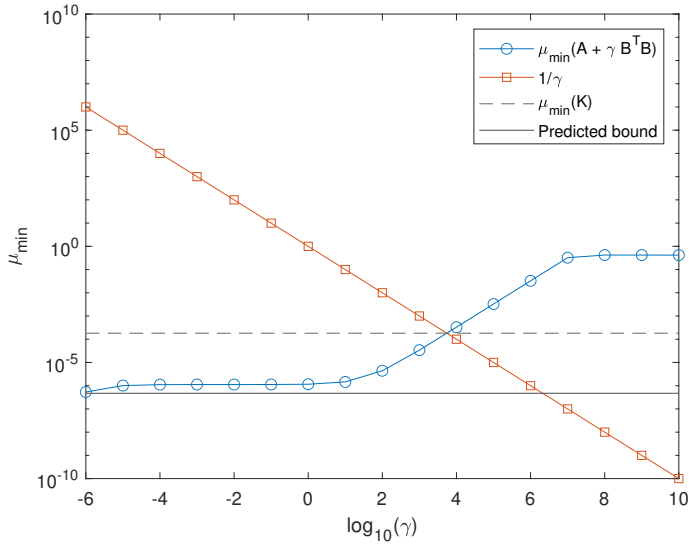

Figure 5.2: Comparison of predicted and actual smallest positive eigenvalue bounds at various values of γ for the IPM matrix for TOMLAB QP 17

The bound for the TOMLAB problem is looser: the predicted bound is 4.716×10^{-7} while the actual smallest positive eigenvalue is 1.817×10^{-4}. Recall that our approach for deriving the bound for a matrix with A that does not have the lowest rank consisted of two steps: (1) implicitly convert the matrix to one with a lowest-rank leading block by "dropping" part of the spectrum of A corresponding to the smallest positive eigenvalues; and (2) estimate the lower bound for the matrix with the lowest-rank leading block using the results of Section 5.2.1, using the fact that this will also be a lower bound for the original matrix. Because our bound in the non-lowest-rank case relies on "dropping" part of the spectrum of A, as discussed in Section 5.2.2, we might in general expect that to lead to some looseness in the bound.

However, the dropping is not the cause of the looseness in this case of the TOMLAB problem, as the saddle-point matrix we obtain by simply replacing A with its dropped portion $A_{n-m}^{\max}$ (defined in (5.12)) has almost the same smallest positive eigenvalue as the original matrix (1.810×10^{-4}, compared with 1.817×10^{-4}). Thus, the looseness in this bound does not come from the dropping part of the spectrum of A to create a lowest-rank matrix, but rather in the estimation of the lower positive eigenvalue bound of the modified matrix.

Chapter 6

Preconditioning when A is singular

In this chapter we consider preconditioning strategies for the unregularized classical saddle-point matrix

$$\mathscr{A}_0 = \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} \tag{6.1}$$

and double saddle-point matrix

$$\mathscr{K}_0 = \begin{bmatrix} A & B^T & 0 \\ B & 0 & C^T \\ 0 & C & 0 \end{bmatrix}, \tag{6.2}$$

under the assumption that A is singular. Our strategy in both cases involves leading block augmentation (as we saw in Chapter 5) combined with taking Schur complements of the augmented systems. In both the classical and double saddle-point settings, we obtain positive definite preconditioners that yield preconditioned operators with a constant number of eigenvalues; thus, a preconditioned iterative solver such as MINRES will converge in a constant number of iterations in the absence of round-off error. Of course, such preconditioners are expensive to apply and some terms must be approximated, and we consider strategies for this as well.

Sections 6.1-6.3 deal with the classical saddle-point system (6.1). In Sec-

tion 6.1 we review some known results when A is lowest-rank, including the existing block diagonal preconditioner of [39]. Then in Section 6.2 we use the eigenvalue analyses in Chapter 5 to extend these to the case where A is not lowest-rank, and show how we can still maintain some of the desirable properties of the preconditioner in the lowest-rank case through careful choice of the augmentation parameter. We then include numerical observations in Section 6.3. We then extend this preconditioning technique to the double saddle-point setting in Section 6.4.

6.1 Preconditioning when A is lowest-rank

In this section we consider the classical saddle-point matrix $\mathscr{A}_0$ (6.1). Recall from Chapter 5 that A is considered to be lowest-rank if $\text{rank}(A) = n - m$, because if its rank were any lower than $\mathscr{A}_0$ would necessarily be singular. Recall also that, given a semidefinite weight matrix $W \in \mathbb{R}^{m\times m}$, we define the augmented leading block by $A_w := A + B^T W B$, and let

$$\mathscr{A}_0(W) := \begin{bmatrix} A_W & B^T \\ B & 0 \end{bmatrix}.$$

We assume that A_W is invertible; when A is lowest-rank, this means that W must be positive definite.

When A is lowest-rank, the blocks of $\mathscr{A}_0$ and those of the augmented matrix $\mathscr{A}_0(W)$ interact in unique ways, which provide useful tools in the design and analysis of preconditioners. Estrin and Greif [19, Theorem 3.5] provide the following result on the Schur complement of $\mathscr{A}_0(W)$:

Proposition 6.1. *Suppose* $\text{nullity}(A) = m$ *and let* $W \in \mathbb{R}^{m\times m}$ *be an invertible matrix. Then*

$$B(A + B^T W B)^{-1} B^T = W^{-1}.$$

We also recall the following result [20, Corollary 2.1] applying to more general matrices, which we will use repeatedly in our analyses:

Lemma 6.2. *For matrices* $M, N \in \mathbb{R}^{n\times n}$ *with* $\text{rank}(M) = r$, $\text{rank}(n) = n - r$ *and*

$M+N$ nonsingular, the matrix $(M+N)^{-1}M$ is a projector with rank r. Moreover,

$$M(M+N)^{-1}N = 0.$$

We consider the block diagonal preconditioner [39]

$$\mathscr{P}_W = \begin{bmatrix} A_W & 0 \\ 0 & W^{-1} \end{bmatrix}, \tag{6.3}$$

where W is positive definite and $A_W = A + B^T W B$. Let us denote the blocks of the split preconditioned operator $\mathscr{P}_W^{-1/2}\mathscr{A}_0\mathscr{P}_W^{-1/2}$ as follows:

$$\mathscr{P}_W^{-1/2}\mathscr{A}_0\mathscr{P}_W^{-1/2} = \begin{bmatrix} A_W^{-1/2}AA_W^{-1/2} & A_W^{-1/2}B^T W^{1/2} \\ W^{1/2}BA_W^{-1/2} & 0 \end{bmatrix} =: \begin{bmatrix} \tilde{A} & \tilde{B}^T \\ \tilde{B} & 0 \end{bmatrix}. \tag{6.4}$$

Lemma 6.3. *When* $\operatorname{rank}(A) = n - m$*, the blocks of* $\mathscr{P}_W^{-1/2}\mathscr{A}_0\mathscr{P}_W^{-1/2}$ *satisfy the following:*

(i) All nonzero eigenvalues of $\tilde{A}$ *are equal to 1;*

(ii) All singular values of $\tilde{B}$ *are equal to 1;*

(iii) The subspaces $\operatorname{range}(\tilde{A})$ *and* $\operatorname{range}(\tilde{B}^T)$ *are orthogonal.*

Proof. To prove (i), we note that $\tilde{A}$ is similar to $A_W^{-1}A$, which is a projector by Lemma 6.2. Lemma 6.1 gives us that $BA_W^{-1}B^T = W^{-1}$, and therefore

$$\tilde{B}\tilde{B}^T = W^{1/2}BA_W^{-1}B^T W^{1/2} = I,$$

which proves (ii). We prove (iii) by showing that $\operatorname{range}(\tilde{B}^T) \in \ker(\tilde{A})$. We write

$$\tilde{A}\tilde{B}^T = A_W^{-1/2}AA_W^{-1}B^T W^{-1/2} = 0,$$

where the second equality follows from the result of [19, Proposition 2.6], which shows that $A_W^{-1}B^T$ is a null-space matrix of A. □

We now consider what the results of Lemma 6.3 tell us about the eigenvalues of $\mathscr{P}_W^{-1}\mathscr{A}_0$ when $\operatorname{rank}(A) = n - m$. The orthogonality of $\operatorname{range}(\tilde{A})$ and $\operatorname{range}(\tilde{B}^T)$

means that the value of $\cos(\theta_{\min})$ in Theorem 5.6 is 1, and thus that the positive eigenvalues are greater than or equal to the minimum of the smallest positive eigenvalue of $\tilde{A}$ and the smallest singular value of $\tilde{B}$. These are both equal to 1, by parts (i)-(ii) of Lemma 6.3. Because the maximal eigenvalues of $\tilde{A}$ and singular values of $\tilde{B}$ are also equal to 1, all negative eigenvalues are equal to -1 and all positive eigenvalues are less than or equal to 1 (as a consequence of Lemma 2.1 of Rusten and Winther [76]). This yields the following result, which is also shown via a different proof method in [39, Theorem 4.1]; we refer to their proof for derivation of the multiplicities of the eigenvalues.

Proposition 6.4. *When* $\operatorname{rank}(A) = n - m$*, the matrix* $\mathscr{P}_W^{-1}\mathscr{A}_0$ *has two distinct eigenvalues given by* 1 *and* -1 *with algebraic multiplicities n and m, respectively.*

Proposition 6.4 tells us that, when A has maximal nullity there is a block diagonal preconditioner that yields a preconditioned operator with two distinct eigenvalues. This is similar to the block diagonal preconditioner of [60], which yields a preconditioner with three distinct eigenvalues in the case that A is positive definite. What has not yet been developed is a preconditioner that gives a constant number of distinct eigenvalues for the "in-between" case where A is rank-deficient, but not maximally so. This is the focus of the next section.

6.2 Preconditioning when A is not lowest-rank

6.2.1 Preconditioner derivation

Let us now consider the case in which A has nullity k, with $k < m$. We will now consider how we can devise a preconditioner to preserve (perhaps approximately) the properties listed in Lemma 6.3 in the case where A is not lowest-rank.

Let us consider a general block diagonal preconditioner of the form

$$\mathscr{P} = \begin{bmatrix} A+G & 0 \\ 0 & C \end{bmatrix},$$

where C is positive definite and G is a semidefinite matrix such that $A+G$ is posi-

tive definite. As before, let us define the split preconditioned system:

$$\begin{aligned}\mathscr{P}^{-1/2}\mathscr{A}_0\mathscr{P}^{-1/2} &= \begin{bmatrix} (A+G)^{-1/2}A(A+G)^{-1/2} & (A+G)^{-1/2}B^TC^{-1/2} \\ C^{-1/2}B(A+G)^{-1/2} & 0 \end{bmatrix} \\ &=: \begin{bmatrix} \tilde{A} & \tilde{B}^T \\ \tilde{B} & 0 \end{bmatrix}.\end{aligned}$$

Property (i) of Lemma 6.3 holds whenever $\text{rank}(G) = k$; see Lemma 6.2. It is also straightforward to verify, using a similar process as in the proof of Lemma 6.3, that Property (ii) holds if and only if

$$C = B(A+G)^{-1}B^T.$$

Property (iii) of Lemma 6.3 holds because, in that Lemma's setting,

$$A(A+G)^{-1}B^T = 0.$$

We can write this as

$$\begin{aligned}A(A+G)^{-1}B^T &= (A+G-G)(A+G)^{-1}B^T \\ &= B^T - G(A+G)^{-1}B^T.\end{aligned} \tag{6.5}$$

Suppose that G has rank k, as we have already established will ensure Property (i). Then, as a consequence of Lemma 6.2, $G(A+G)^{-1}$ is a projector onto the range of G. From (6.5) we see that Property (iii) will hold if $G(A+G)^{-1}$ is a projector onto the range of B^T; however, this is clearly not possible if $\text{rank}(G) = k < m$. But we note that if we set

$$G = B^T W_k B,$$

where W_k is a symmetric positive semidefinite matrix of rank k, this matrix will be a projector onto a rank-k subspace of $\text{range}(B^T)$. While Property (iii) will not hold in this case because we will not have $\tilde{A}\tilde{B}^T = 0$, we instead have that $\text{nullity}(\tilde{A}\tilde{B}^T) = k$ (which is the highest nullity we can achieve, as from (6.5) we have a rank-k term being subtracted from B).

Thus, we consider the preconditioner:

$$\mathscr{P}_k = \begin{bmatrix} A_k & 0 \\ 0 & S_k \end{bmatrix}, \tag{6.6}$$

where $A_k = A + B^T W_k B$ and $S_k = B A_k^{-1} B^T$, with $\text{rank}(W_k) = \text{nullity}(A) = k$. This is the same preconditioner analyzed in [34], but with the additional assumption that $\text{rank}(W_k) = k$.

Remark 6.5. We note that, when A is maximally rank-deficient, the preconditioner $\mathscr{P}_k$ reduces to that of Greif and Schötzau defined in eq. (6.3). When A is positive definite, then $\mathscr{P}_k$ is equivalent to the preconditioner of Murphy, Golub, and Wathen shown in eq. (1.16).

6.2.2 Preconditioner analysis

We begin by presenting a few lemmas that will be necessary for our analysis.

Lemma 6.6. *When* $\text{rank}(W_k) = \text{nullity}(A) = k$*,*

$$(B A_k^{-1} B^T)^{-1} = W_k + (BB^T)^{-1} B(A - AVA) B^T (BB^T)^{-1},$$

where $V = Z(Z^T AZ)^{-1} Z^T$ *with* $Z \in \mathbb{R}^{n \times (n-m)}$ *being a null-space matrix of* B*.*

Proof. The proof follows by considering the block inverses of $\mathscr{A}_0$ and

$$\mathscr{A}_0(W_k) := \begin{bmatrix} A_k & B^T \\ B & 0 \end{bmatrix}.$$

Let $Z \in \mathbb{R}^{n \times (n-m)}$ denote a matrix whose columns form a basis for $\ker(B)$. The inverse of $\mathscr{A}_0$ is (see [7, Eq. (3.8)]):

$$\mathscr{A}_0^{-1} = \begin{bmatrix} V & (I - VA) B^T (BB^T)^{-1} \\ (BB^T)^{-1} B(I - AV) & -(BB^T)^{-1} B(A - AVA) B^T (BB^T)^{-1} \end{bmatrix},$$

where $V = Z(Z^T AZ)^{-1} Z^T$; we note that $Z^T AZ$ must be nonsingular for any nonsingular $\mathscr{A}_0$ (see [7]). The result then follows from Lemma 5.2 and the fact that the (2,2)-block of $(\mathscr{A}_0(W_k))^{-1}$ is equal to $-(B A_k^{-1} B^T)^{-1}$ (see [7, Eq. (3.4)]). □

Lemma 6.7. *The matrix VA is a projector. Moreover, when* $\operatorname{rank}(W_k) = \operatorname{nullity}(A) = k$*, the following results hold:*

(i) The matrix $A_k^{-1}A$ is a projector;

(ii) The matrices VA and $A_k^{-1}A$ commute.

Proof. By writing $VA = Z(Z^TAZ)^{-1}Z^TA$, it is clear that VA is a projector onto $\ker(B)$. Item (i) holds because of Lemma 6.2.

To verify (ii), we first note that

$$VAA_k^{-1}A = VA,$$

because AA_k^{-1} is a projector (this follows from the fact that $A_k^{-1}A = (AA_k^{-1})^T$ is a projector) onto the range of A. Because $A_k^{-1}A = I - A_k^{-1}B^TW_kB$, we can write

$$A_k^{-1}AZ = Z - A_k^{-1}B^TWBZ = Z.$$

Therefore,

$$\begin{aligned} A_k^{-1}AVA &= A_k^{-1}AZ(Z^TAZ)^{-1}Z^TA \\ &= Z(Z^TAZ)^{-1}Z^TA \\ &= VA \\ &= VAA_k^{-1}A. \end{aligned}$$

□

Theorem 6.8. *Let $\mathscr{A}_0$ be nonsingular with A having nullity k, and let $W_k \in \mathbb{R}^{m\times m}$ be a rank-k matrix such that $A + B^TW_kB$ is positive definite. The preconditioned operator $\mathscr{P}_k^{-1}\mathscr{A}_0$ has four distinct eigenvalues:*

- $\lambda = -1$ *with multiplicity k;*
- $\lambda = 1$ *with multiplicity $n - m + k$;*
- $\lambda = \frac{1\pm\sqrt{5}}{2}$*, each with multiplicity $m - k$.*

Proof. We consider the eigenvalue equations for the preconditioned system:

$$Ax + B^T y = \lambda A_k x; \tag{6.7a}$$

$$Bx = \lambda S_k y. \tag{6.7b}$$

From (6.7b) we obtain $y = \frac{1}{\lambda} S_k^{-1} Bx$. Substituting this into (6.7a) and re-arranging yields

$$A_k^{-1} Ax + \frac{1}{\lambda} A_k^{-1} B^T S_k^{-1} Bx - \lambda x = 0. \tag{6.8}$$

By Lemma 6.6, we can write

$$\begin{aligned} A_k^{-1} B^T S_k^{-1} B = {} & A_k^{-1} B^T W_k B \\ & + A_k^{-1} B^T (BB^T)^{-1} B(A - AVA) B^T (BB^T)^{-1} B. \end{aligned} \tag{6.9}$$

As was discussed in the proof of Lemma 6.7, VA is a projector onto $\ker(B)$, meaning that $I - VA$ is a projector onto $\text{range}(B)$. Because $B^T (BB^T)^{-1} B$ is an orthogonal projector onto this subspace, we have

$$(I - VA) B^T (BB^T)^{-1} B = I - VA.$$

Similarly, $B^T (BB^T)^{-1} B(I - AV) = I - AV$. Thus, we can further simplify (6.9), using relations we developed in Lemma 6.7:

$$\begin{aligned} A_k^{-1} B^T S_k^{-1} B &= A_k^{-1} B^T W_k B + A_k^{-1} (A - AVA) \\ &= I - A_k^{-1} AVA \\ &= I - VA. \end{aligned}$$

We can thus rewrite (6.8) as

$$A_k^{-1} Ax - \frac{1}{\lambda} VAx + \left(\frac{1}{\lambda} - \lambda \right) x = 0. \tag{6.10}$$

By Lemma 6.7, $A_k^{-1} A$ and VA are commuting projectors; thus, they have the same

eigenvectors. Because VA has rank $n-m$ and $A_k^{-1}A$ has rank $n-k$, we have

$$\text{range}(VA) \subseteq \text{range}(A_k^{-1}A) \text{ and } \ker(A_k^{-1}A) \subseteq \ker(VA).$$

We now consider x in the ranges/kernels of these projectors.

Case I: When $x \in \ker(A)$, (6.10) becomes

$$\left(\frac{1}{\lambda} - \lambda\right) x = 0. \tag{6.11}$$

We note that x cannot be zero, as (4.6a) would necessarily imply $y = 0$. Thus, (6.11) gives k eigenvectors corresponding to each of the eigenvalues $\lambda = \pm 1$.

Case II: When $x \in \text{range}(VA)$ (and therefore also in $\text{range}(A_k^{-1}A)$), (6.10) becomes

$$(1-\lambda)\, x = 0,$$

which gives $n-m$ additional eigenvectors corresponding to the eigenvalue $\lambda = 1$.

Case III: if $x \in \ker(VA)$ and $\text{range}(A_k^{-1}A)$ (we know there are $m-k$ such vectors because the projectors commute), (6.10) becomes

$$\left(1 + \frac{1}{\lambda} - \lambda\right) x = 0,$$

which gives the eigenvalues $\lambda = \frac{1 \pm \sqrt{5}}{2}$, each with geometric multiplicity $m-k$.

Cases I-III account for all $n+m$ eigenvectors of $\mathscr{P}_k^{-1}\mathscr{A}_0$. □

6.2.3 Schur complement approximations

In practice, the blocks A_k and S_k of the ideal preconditioner $\mathscr{P}_k$ defined in (6.6) are too expensive to invert exactly. While developing suitable approximation strategies for these terms often requires some knowledge of the problem at hand, we provide here two strategies for approximately inverting the Schur complement S_k.

First, recall from Lemma 6.1 that when A is lowest-rank we have $S_k^{-1} = W_k$. Thus, when A is not lowest-rank but is close (i.e., has high nullity), it is reasonable

to use an approximation of the form

$$S_k^{-1} \approx W_k + \alpha I, \tag{6.12}$$

where α is a small positive value. We add the αI term because if A is not maximally rank-deficient then W_k will be singular. We refer to this strategy as the "WkI Schur complement approximation."

For our second strategy, recall that Lemma 6.6 tells us that

$$\begin{aligned} S_k^{-1} &= W_k + (BB^T)^{-1}B(A - AVA)B^T(BB^T)^{-1} \\ &= W_k + (BB^T)^{-1}BA\underbrace{(I - VA)}_{=:P}B^T(BB^T)^{-1}. \end{aligned}$$

Since VA is a projector whose range is $\ker(B)$ and whose kernel is $\ker(Z^TA)$, the matrix $P = (I - VA)$ has range given by $\ker(Z^TA)$ and kernel given by $\ker(B)$. Thus, we consider replacing the projector $(I - VA)$ by the orthogonal projector onto range(B), defined by $P_B = B^T(BB^T)^{-1}B$. This matrix has the same kernel as P but a different range, and has the advantage of yielding a considerably simpler second term, as we can write:

$$\begin{aligned} (BB^T)^{-1}BAP_BB^T(BB^T)^{-1} &= (BB^T)^{-1}BAB^T(BB^T)^{-1}BB^T(BB^T)^{-1} \\ &= (BB^T)^{-1}BAB^T(BB^T)^{-1}. \end{aligned}$$

Thus, we can consider the Schur complement approximation:

$$S_k^{-1} \approx W_k + (BB^T)^{-1}BAB^T(BB^T)^{-1}. \tag{6.13}$$

We note that this modified second term is similar to the BFBT preconditioner proposed by Elman [16] for the Navier-Stokes equations; thus, we refer to this as the "BFBT Schur complement approximation."

6.3 Numerical experiments

In this section we consider implementations of the block diagonal preconditioner described in Section 6.2. All experiments are run in MATLAB R2021a on a com-

modity desktop PC. We report computation times for all experiments. The code is not optimized for efficiency and the measurements do not represent what would be possible with an optimized, state-of-the-art code base; they are included as a way to compare the computational costs of different approaches and validate our analytical observations.

6.3.1 Selection of weight matrix

Here we detail our general approach for choosing W_k. For simplicity, all our matrices W_k are diagonal matrices with either 1 or 0 on the diagonal; thus, the augmented matrix A_k is equal to A in addition to k terms of the form $b_i b_i^T$, where b_i denotes the ith column of B^T. Hence, our task of selecting W_k becomes the task of selecting which columns of B^T to use in to augment A.

We begin by forming a matrix A_{drop} formed by eliminating very small elements of A (for our purposes, we eliminate those matrix entries whose absolute values are less than machine epsilon times the largest magnitude entry in A). We then select columns of B^T that increase the structural rank of A_{drop} until the matrix $A_{drop} + \sum_i b_i b_i^T$ has full structural rank. These selected columns of B^T do not guarantee that the augmented matrix $A + \sum_i b_i b_i^T$ has full numerical rank or is sufficiently well-conditioned to avoid convergence problems, so in some cases we add additional columns of B^T; in this case, we greedily select the sparsest columns of B^T to reduce fill-in of A_k.

We note that, in general, this approach of selecting W_k does *not* guarantee a "minimal-rank" augmentation; that is, the rank of W_k may be greater than the nullity of A. Finding a W_k with rank exactly equal to the nullity of A such that the augmented matrix A_k is sufficiently well-conditioned to avoid numerical difficulty requires knowledge of the null-space of A and of which vectors in B^T will span that null space. That said, in many practical applications, for example in problems arising from discretizations of PDEs, some information on the discrete differential operators and their null space is often available and comes handy.

6.3.2 Constrained optimization problems

Problem statement

Here we consider preconditioning matrices that arise in the IPM solution of a linear or quadratic program. Recall from Section 2.2 that each step of a primal-dual interior-point method (IPM) to solve (2.7) requires solving a linear system of the form [61]:

$$\begin{bmatrix} H + X^{-1}Z & J^T \\ J & 0 \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} = \begin{bmatrix} -c - HxJ^T y + \tau X^{-1} e \\ b - Jx \end{bmatrix}. \tag{6.14}$$

See [61] for full details. Some entries of the diagonal matrices X and Z approach zero as the IPM iterations proceed, so the leading block of the saddle-point matrix becomes increasingly ill-conditioned, with the largest magnitude entries occurring along the diagonal. Thus the leading block may become nearly singular or numerically singular, particularly if H is singular.

Description of test problems

We use an implementation of the predictor-corrector algorithm of Mehrotra [57]. The matrices for linear programming problems were obtained from the Sparse Suite matrix collection [14], and the quadratic programming problems are from TOMLAB[1]. A summary of the test suite of LP problems used in our experiments is given in Table 6.1.

Comparison of different augmentation and approximation strategies

In this experiment we consider preconditioners of the form

$$\mathscr{P} = \begin{bmatrix} \tilde{A}_{aug} & 0 \\ 0 & B\hat{A}_{aug}^{-1}B^T, \end{bmatrix} \tag{6.15}$$

where $\tilde{A}_{aug}$ and $\hat{A}_{aug}$ are approximations (potentially the same approximation) of an augmented leading block A. Our experiments are on matrices that arise while

[1] Test matrices available at https://tomopt.com/tomlab/.

Problem ID	m	n	nnz($\mathscr{A}_0$)
lp_80bau3b	2,262	12,061	35,325
lp_bandm	305	472	2,966
lp_capri	271	482	2,378
lp_finnis	497	1,064	3,824
lp_fit1p	627	1,677	11,545
lp_ganges	1,309	1,706	8,643
lp_lofti	153	366	1,502
lp_maros_r7	3,136	9,408	154,256
lp_osa_14	2,337	5,497	371,894
lp_osa_30	4,350	104,374	708,862
lp_pilot87	2,030	6,680	81,629
lp_scfxm1	330	600	3,332
lp_scsd8	397	2,750	11,334
lp_stair	356	614	4,617
lp_standmps	467	1,274	5,152
lp_stocfor2	2,157	3,045	12,402
lp_truss	1,000	8,806	36,642
lp_vtp_base	198	346	1,397

Table 6.1: Summary of linear programming (LP) problems used in numerical experiments. The value $nnz(\mathscr{A}_0)$ gives the number of nonzeros arising in the saddle-point system at each interior point method (IPM) iteration.

applying an interior-point method on an LPs, so the leading block A is diagonal. We consider three augmentation strategies:

1. Partial augmentation: we take $A_{aug} = A + B^T W_k B$, where we form W_k by selecting just enough rows of B such that $A_{drop} + B^T W_k B$ has full structural rank, where A_{drop} is the matrix obtained by setting to zero all elements of A with absolute value less than or equal to machine-epsilon times the largest absolute magnitude value of A.

2. Full augmentation: we take $A_{aug} = A + B^T B$.

3. Identity augmentation: we take $A_{aug} = A + \rho I$, for some positive ρ.

For A_{aug} arising from partial and full augmentation, we consider three approximations for $\tilde{A}_{aug}$ and $\hat{A}_{aug}$ in (6.15):

1. Ideal approximation (ID): $\tilde{A}_{aug} = \hat{A}_{aug} = A_{aug}$. (This is too expensive to use in practice but we include it here for comparison purposes.)

2. Diagonal approximation (D): $\tilde{A}_{aug} = \hat{A}_{aug} = \text{diag}(A_{aug})$.

3. Incomplete Cholesky approximation (IC) : $\tilde{A}_{aug} = \text{IC}(A_{aug})$ and $\hat{A}_{aug} = \text{diag}(A_{aug})$. We use ICT with drop tolerance of 0.01 (meaning that non-diagonal elements of the factorization with absolute value less than 0.01 times the norm of that row of A_{aug} are dropped).

For the identity-based augmentation, the matrix A_{aug} is diagonal, so we solve it exactly (that is, $\tilde{A}_{aug} = \hat{A}_{aug} = A_{aug}$).

Problem ID	Partial			Full			Identity
	ID	D	IC	ID	D	IC	ID
80bau3b	5 (0.03)	22 (0.03)	230 (0.02)	18 (2.0)	122 (0.02)	254 (0.01)	43 (0.02)
maros_r7	22 (3.7)	22 (0.2)	56 (0.1)	2 (2.2)	19 (0.1)	26 (0.1)	11 (0.1)

Table 6.2: MINRES iteration counts for partial, full and identity-augmentation preconditioners for the lp_80bau3b and lp_maros_r7 problems, using various block approximation strategies (ID=ideal, D=diagonal, IC=incomplete Cholesky). Time per iteration (in seconds) is given in parentheses.

Problem ID	Partial augmentation			Full augmentation		
	Rank(W)	nnz(A_W)	nnz(IC(A_W))	Rank(W)	nnz(A_W)	nnz(IC(A_W))
80bau3b	2	12,249	12,101	2,262	456,943	14,183
maros_r7	2,511	1,101,752	31,343	3,136	1,230,928	10,761

Table 6.3: Comparison of memory usage for partial and full augmentation for the lp_80bau3b and lp_maros_r7 problem.

We use matrices that arise from IPMs on the test problems lp_80bau3b and lp_maros_r7. Iteration counts and time per iteration are given in Tables 6.2 and 6.3.

We observe that for lp_80bau_3b, the partial augmentation preconditioner outperforms the full augmentation preconditioner in terms of both iteration count and memory usage. This is because the leading block of this matrix is only mildly rank-deficient, so we only need a low-rank augmentation to make it nonsingular

(which leads to a much sparser augmented matrix than the full augmentation); additionally, when we fully augment this matrix we are far away from the "ideal" amount of augmentation (i.e., the rank of augmentation that would yield a constant number of eigenvalues in an ideally-preconditioned iterative solver) because the leading block is nowhere near lowest-rank.

In contrast, the leading block for `lp_maros_r7` is highly rank-deficient, as even the minimal amount of augmentation to obtain a structurally nonsingular leading block requires using most of the rows of B (2,511, when m for this problem is 3,136). And we observe that, in cases like these where the nullity of the leading block is high, we are close enough to the lowest-rank case that full augmentation performs well. In this case, it actually performs better than the partial augmentation in terms of iteration counts and computation time because the fully augmented leading block is more well-conditioned than the partially augmented leading block. Recall that our procedure for choosing W_k only looks at structural rank, and does not guarantee that the augmented matrix is actually nonsingular (so we may still encounter numerical difficulties without further augmentation).

Finally, we note that the incomplete Cholesky approximation strategy is less effective than the diagonal approximation strategy. One reason for this is that by the time IPM matrices are singular, the largest magnitude entries tend to occur along the diagonal; thus, a diagonal leading block approximation is generally effective (as we will see in the next set of experiments). The other is that when we used the incomplete Cholesky in the leading block, we avoided using the inverse of the incomplete Cholesky factors in the Schur complement approximation to avoid introducing too much computational expense. Thus, the Schur complement approximation is not equal to $B\tilde{A}_{aug}^{-1}B^T$ (where $\tilde{A}_{aug}$ is the selected leading block approximation); and as we saw in Section 6.2, this has an impact on the theoretical properties of the preconditioned operator.

Running partial augmentation preconditioners on LP test suites

Here we consider preconditioning the complete set of problems described in Table 6.4. The matrices reported below are the first matrices for which the IPM generates a matrix with a numerically singular leading block. We consider the partial

augmentation preconditioner of the form (6.15) with the diagonal leading block approximation strategy: that is, we define a preconditioner $\mathscr{P}_D$ of the form (6.15) using $\tilde{A}_{aug} = \hat{A}_{aug} = \text{diag}(A_{aug})$. In all cases, we select W_k by augmenting A until the matrix $A_{drop} + B^T W_k B$ is structurally nonsingular. MINRES solver tolerance is set to a relative residual norm of 10^{-8}.

Problem ID	rank(W_k)	nnz(A_k)	$\mathscr{P}_D$	
			Iters	Time per iter
80bau3b	1	12,117	20	0.02
bandm	5	1,444	40	0.003
capri	13	2,230	67	0.003
finnis	29	11,184	77	0.006
fit1p	5	2,545	28	0.06
ganges	88	2,690	41	0.01
lofti	13	966	194	0.001
maros_r7	64	73,102	26	0.2
osa_14	34	98,459,317	171	0.06
osa_30	4	354,880,632	80	0.1
pilot87	5	133,798	37	0.2
scfxm1	1	840	32	0.003
scsd8	36	16,826	6	0.003
stair	33	9,994	11	0.006
standmps	2	557,906	65	0.004
stocfor2	61	3,411	9	0.1
truss	15	18,468	34	0.005
vtp_base	10	3,126	125	0.002

Table 6.4: MINRES iteration counts and time per iteration (in seconds) of the partial augmentation preconditioners with diagonal approximations of A_k.

Eigenvalues of the preconditioned operator $\mathscr{P}_D^{-1}\mathscr{A}_0$ are shown in Figure 6.1 for lp_fit1p problem. Notice that there is strong clustering around the eigenvalues $1, \frac{1\pm\sqrt{5}}{2}$.

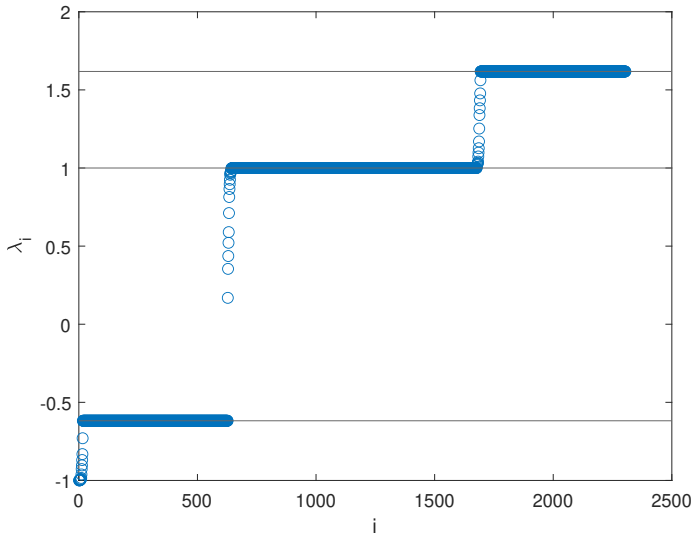

Figure 6.1: Eigenvalues of preconditioned operator $\mathscr{P}_D^{-1}\mathscr{A}_0$ for matrix arising in the IPM solution of the `lp_fit1p` problem. Horizontal lines are shown at $y = \pm 1, \frac{1\pm\sqrt{5}}{2}$.

Using preconditioned MINRES iterations in an IPM

Here we consider using preconditioned inner solves in an IPM solver. For our test problems, we use the LP `lp_stocfor2` and the TOMLAB QP problem 37 (which has $m = 490$; $n = 1275$; 3,288 nonzeros in the Jacobian matrix; and 290 in the Hessian). Our preconditioning approach at each iteration is as follows:

- If the leading block A is nonsingular, we use the preconditioner

$$\mathscr{P}_{LP} = \begin{bmatrix} A & 0 \\ 0 & BA^{-1}B^T \end{bmatrix}$$

for the LP (recall that in this context A is diagonal), and

$$\mathscr{P}_{QP} = \begin{bmatrix} \mathrm{IC}(A) & 0 \\ 0 & B(\mathrm{diag}(A))^{-1}B^T \end{bmatrix}$$

for the QP, with an ICT drop tolerance of 0.01.

- If the leading block A is singular, we select the lowest-rank W_k to make $A_{drop} + B^T W_k B$ nonsingular, and use the preconditioner

$$\mathscr{P} = \begin{bmatrix} \text{diag}(A_k) & 0 \\ 0 & B(\text{diag}(A_k))^{-1} B^T \end{bmatrix}.$$

We solve the IPM to a duality gap tolerance of 10^{-6} and use an inner tolerance of 10^{-7} for the MINRES solves.

We see that for both problems, using inexact solves results in modestly more IPM iterations, as we would expect. For the LP, the leading block was nonsingular for the first 21 iterations and numerically singular for the final 10. For the QP, the leading block was nonsingular for the first 22 iterations and singular for the last 16. Notice that the average MINRES iteration counts are correspondingly higher for the QP. This is because, at the LP steps with a nonsingular leading block, we were able to use an ideal preconditioner because the leading block is diagonal, and convergence was always achieved in roughly three iterations. Additionally, the nonzero Hessian in the QP has some additional terms in the leading block that are dropped in the diagonal leading block approximation once the leading block becomes singular.

Problem		Direct inner solve	MINRES inner solve		
ID	Type	IPM iterations	IPM iterations	Inner iters (average)	
				Predictor	Corrector
`stocfor2`	LP	27	31	4.1	4.1
`TOMLAB37`	QP	31	38	35.1	36.6

Table 6.5: Comparison of IPM iterations using a direct vs. preconditioned MINRES solver for the inner linear system solves. Average number of inner MINRES iterations are reported for both the predictor and corrector steps.

Testing different block approximation strategies

Here we test the WkI Schur complement approximation strategy (see Eq. (6.12)). We use a matrix that arises at the 20th iteration of the IPM solution for the LP

maros_r7 and use $\beta = 0.5$. As we have seen in our earlier LP experiments, by the time the IPM iterations have advanced enough to create a numerically singular leading block, the diagonal has enough large entries that the augmented matrix A_k is mostly diagonally dominant. Thus, using $\text{diag}(A_k)$ is often effective in approximating A_k. We include comparisons between the preconditioners in which:

- A_k approximated by $\text{diag}(A_k)$ and S_k^{-1} is approximated by $B\,\text{diag}(A_k)^{-1}B^T$ (the preconditioner P_D explored in the previous set of experiments);
- A_k is approximated by $\text{diag}(A_k)$ and S_k^{-1} is approximated by $W_k + \beta I$ ("Diagonal+WkI" or "D+WkI").

For this experiment, our weight matrix W_k has rank 2,911 (the minimum required to achieve structural nonsingularity of $A_{drop} + B^T W_k B$).

A convergence plot is shown in Figure 6.2. The P_D preconditioner converges in 11 iterations and 1.4 seconds (0.1 seconds per iteration), and the Diagonal+WkI preconditioner in 102 iterations and 0.18 seconds (0.0018 seconds per iteration). While this is a significantly higher iteration count, we notice that this preconditioner is extremely cheap (in that it is fully diagonal) and thus results in faster computational time overall. We note that a basic Jacobi iteration on the original system (or Jacobi on the leading block combined with the WkI approximation of the Schur complement) does not lead to convergence. Thus, the leading block augmentation has utility in arriving at this surprisingly simple-looking preconditioner.

6.3.3 A geophysical inverse problem

Problem statement

Here we consider the solution of a geophysical inverse problem, as described in Section 2.2. If Gauss-Newton iterations are used, the linear system to be solved at each step takes the form

$$\begin{bmatrix} Q^T Q & 0 & A^T \\ 0 & \beta W^T W & G^T \\ A & G & 0 \end{bmatrix} \begin{bmatrix} \delta u \\ \delta m \\ \delta \lambda \end{bmatrix} = - \begin{bmatrix} r_u \\ r_m \\ r_\lambda \end{bmatrix}, \tag{6.16}$$

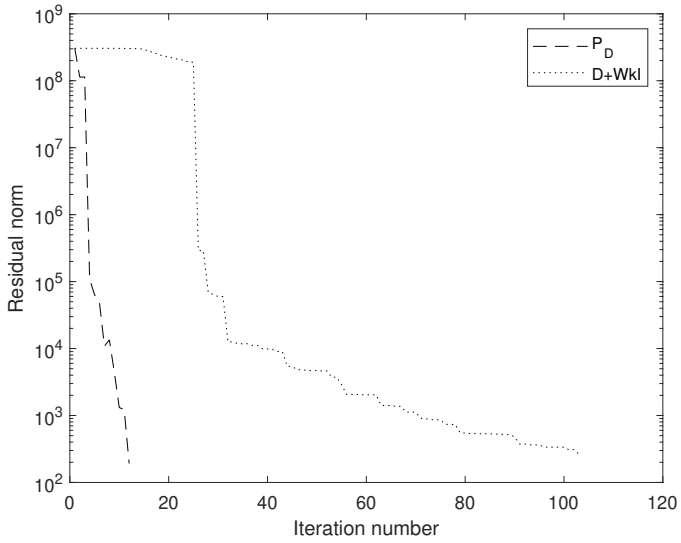

Figure 6.2: Comparison of block approximation strategies (diagonal leading block + $B(\text{diag}(A_k))^{-1}B^T$ Schur complement; Diagonal leading block+WkI Schur complement) for a matrix arising from an IPM on the `lp_maros_r7` problem.

with G being the Jacobian of A. In the typical case of sparse observations, Q^TQ has high nullity.

Testing different block approximation strategies

In this experiment we test the BFBT Schur complement approximation strategy (Eq. (6.13)). We set the regularization parameter $\beta = 10^{-3}$. The leading block is highly singular, so we augment A by all of B to avoid numerical difficulties (as simply augmenting by enough rows of B to make the augmented matrix structurally nonsingular still leads to a matrix that is highly ill-conditioned).

Recall that the BFBT Schur complement approximation requires two solves for BB^T. Fortunately, for the geophysics problem, this term is sparse and banded. Thus, in computing this approximation, we will solve exactly for the BB^T terms.

We note that the augmented matrix $A + B^TB$ has an interesting structure, as we can see in Figure 6.3: if we partition the matrix four blocks with the (1,1)-

block of size m and the (2,2)-block of size $n-m$, we observed that the (1,1)- and (2,2)-blocks are banded (e.g., for a problem with $m = 9,261$ and $n = 17,261$, the bandwidths are 848 and 421, respectively), and can therefore be solved less expensively than the entire matrix $A + B^T B$. Thus, we can use block Jacobi to approximately solve A_k. Because stationary methods are often not especially effective as preconditioners, we will instead use block Jacobi as a preconditioner for an inner preconditioned conjugate gradient (PCG) solver for A_k.

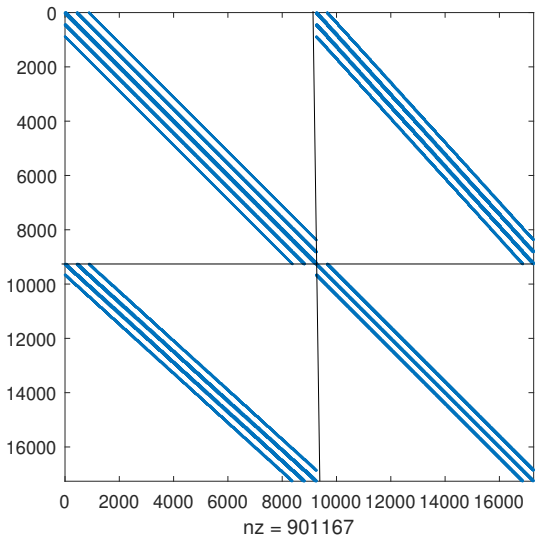

Figure 6.3: Sparsity pattern of $A_k = A + B^T B$ for a geophysics problem with $m = 9,261$ and $n = 17,261$.

Thus, in these experiments, we compare the preconditioners in which:

- A_k is inverted exactly (which is generally not practical for large problems but is included here for validation and comparison), and S_k^{-1} is approximated with the BFBT approximation. We denote this by "Akinv+BFBT."
- A_k is inverted approximately using CG to an inner tolerance of 0.1, with block Jacobi as a preconditioner, and S_k^{-1} is approximated by the BFBT approximation. We denote this by "CG+BFBT."

We use MINRES for the Akinv+BFBT preconditioner. Because the CG+BFBT

preconditioner involves a nonstationary iteration in the outer iteration, we must use a different outer solver. Accordingly, for the CG+BFBT preconditioner we will use FGMRES with restarts every 30 iterations, denoted by FGMRES(30).

m	n	Akinv+BFBT		CG+BFBT	
		Iters	Time per iter	Iters	Time per iter
2,197	3,195	6	0.21	9	0.20
4,913	9,009	6	1.07	10	0.76
9,261	17,261	8	2.87	10	2.26

Table 6.6: Results (solver iteration counts and time per iteration) geophysics problems of varying size. Akinv+BFBT = exact solve for A_k, BFBT approximation for S_k; CG+BFBT = block Jacobi preconditioned CG for A_k, BFBT for S_k.

Results are shown in Table 6.6. The Akinv+BFBT preconditioner performs well in terms of iteration count, but includes a very expensive term in the A_k solve. We note, however, that the number of preconditioned iterations is very close to what we would expect of the ideal preconditioner (with exact solves for both A_k and S_k), which highlights the effectiveness of the BFBT Schur complement approximation for this problem. The CG+BFBT preconditioner achieves similar convergence to the Akinv+BFBT – in particular, the number of iterations appears to be independent of problem size – and is modestly less expensive per iteration in terms of compute time (we avoid the direct solve for A_k, but have some added expense from the inner CG solves and additional orthogonalization for FGMRES). On average, the inner PCG solves required 28.7 iterations for the first test problem (with $m = 2,197$ and $n = 3,195$), 35.1 iterations for the second problem (with $m = 4,913$ and $n = 9,009$), and 35.8 iterations for the third (with $m = 9,261$ and $n = 17,261$). For larger problems, we speculate that CG+BFBT will outperform Akinv+BFBT by larger margins.

6.4 Preconditioning of double saddle-point systems

In this section, we extend the preconditioning approach of the previous section to the unregularized double saddle-point matrix $\mathscr{K}_0$ (6.2). We note that the unregularized case is comparatively less common in the double saddle-point context than

in the classical saddle-point context, though examples arise in, for example, dual-dual mixed finite element formulations [28], constrained weighted least squares problems, and some finite element formulations of the Stokes equation [18, 31]. Additionally, the approach we describe here has not been deployed as part of a practical, preconditioned solver; thus, the work presented here should be viewed as a theoretical starting point for the development of a preconditioned solver.

We begin by extending the definition of "lowest-rank" to the double saddle-point context. The following proposition gives some invertibility condition on $\mathscr{K}_0$.

Proposition 6.9. *The following conditions are necessary for $\mathscr{K}_0$ to be invertible:*

(1) $p \leq m \leq n+p$;

(2) $\text{rank}(B) \geq m-p$;

(3) $\text{rank}(A) \geq n-(m-p)$.

Proof. We will prove the statements of the proposition in the order (2), (3), (1).

For (2), we begin by noting that if C is not full rank then $\mathscr{K}_0$ is clearly singular. We also must have that $\ker(B^T) \cap \ker(C) = \emptyset$, as if we have a nonzero vector y in $\ker(B^T) \cap \ker(C) = \emptyset$ then $\begin{bmatrix} 0 & y^T & 0 \end{bmatrix}^T$ will be a null vector of $\mathscr{K}_0$. Statement (2) follows from this, along with the fact that B^T and C have m columns and C has rank p.

To prove statement (3), we partition $\mathscr{K}$ into a 2×2 block matrix as follows:

$$\mathscr{K} = \left[\begin{array}{cc|c} A & B^T & 0 \\ B & 0 & C^T \\ \hline 0 & C & 0 \end{array}\right].$$

Because $\mathscr{K}_0$ is invertible, the (2,1)-block $\begin{bmatrix} 0 & C \end{bmatrix}$ must have full row rank (because C must have full row rank). We can define a null space of this block using

$$\ker(\begin{bmatrix} 0 & C \end{bmatrix}) = \text{span of } \begin{bmatrix} I_n & 0 \\ 0 & Z_C \end{bmatrix}, \tag{6.17}$$

where I_n is the $n \times n$ identity matrix and $Z_C \in \mathbb{R}^{m \times (m-p)}$ is a matrix whose columns form a basis for $\ker(C)$. From [7, eq. (3.8)], we can see that $\mathscr{K}$ is invertible if and

only if the reduced Hessian, defined by

$$\begin{bmatrix} I_n & 0 \\ 0 & Z_C \end{bmatrix} \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} I_n & 0 \\ 0 & Z_C^T \end{bmatrix} = \begin{bmatrix} A & B^T Z_C \\ Z_C^T B & 0 \end{bmatrix}, \tag{6.18}$$

is invertible. A necessary condition for this is for $\ker(A) \cap \ker(Z_C^T B) = \emptyset$. Because $Z_C^T B \in \mathbb{R}^{(m-p)\times n}$ (and can easily be shown to have full rank because of the requirement that $\ker(B^T) \cap \ker(C) = \emptyset$), this proves the requirement that $\operatorname{rank}(A) \geq n - (m-p)$.

The requirement that the reduced Hessian be invertible also implies that $m - p \leq n$, which establishes the second inequality of statement (1). The first inequality follows from the requirement that C have full row rank. □

Thus, we say that the leading block A of a double saddle-point matrix is *lowest-rank* if $\operatorname{rank}(A) = n - (m-p)$. As a consequence of statement (3) of Proposition 6.9, $\mathcal{K}_0$ will necessarily be singular if the rank of A is any lower.

6.4.1 Leading block augmentation

As in the classical saddle-point case, our approach is to augment the singular leading block A so it becomes positive definite, which then enables us to design preconditioners based on the Schur complements of the augmented system. Recall from Lemma 5.2 that augmenting A in the classical saddle-point case only changed one block of the inverse of the matrix. The following result gives us a way to accomplish this in the double saddle-point case.

Proposition 6.10. *Let $F \in \mathbb{R}^{m\times m}$ be a matrix satisfying*

$$CF = 0.$$

Define $A_F := A + B^T F B$ and let

$$\mathcal{K}_0(F) = \begin{bmatrix} A_F & B^T & 0 \\ B & 0 & C^T \\ 0 & C & 0 \end{bmatrix}. \tag{6.19}$$

Then

$$(\mathscr{K}_0(F))^{-1} = \mathscr{K}_0^{-1} - \begin{bmatrix} 0 & 0 & 0 \\ 0 & F & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Proof. We will approach this problem by re-ordering and partitioning $\mathscr{K}_0$ into a (classical) saddle-point matrix:

$$\mathscr{K}_0 = \left[\begin{array}{cc|c} A & 0 & B^T \\ 0 & 0 & C \\ \hline B & C^T & 0 \end{array}\right].$$

We know from Lemma 5.2 that, for any $m \times m$ matrix W,

$$\left[\begin{array}{cc|c} A+B^TWB & B^TWC & B^T \\ C^TWB & C^TWC & C \\ \hline B & C^T & 0 \end{array}\right]^{-1} = \left[\begin{array}{cc|c} A & 0 & B^T \\ 0 & 0 & C \\ \hline B & C^T & 0 \end{array}\right]^{-1} - \begin{bmatrix} 0 & 0 & 0 \\ 0 & W & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

We seek an augmentation strategy that only changes one block of the matrix inverse (as this does), but we would rather not fill in any of the zero blocks of $\mathscr{K}_0$. Thus, we select a specific weight matrix F satisfying $CF = 0$; this yields the desired result. □

Since a goal of leading block augmentation is to convert a semidefinite leading block to a positive definite one, we present the following result to prove that this is still achievable even with the additional constraint that the weight matrix be a null matrix of C.

Proposition 6.11. *Let $\mathscr{K}_0$ is invertible, and let F be an $m \times m$ matrix satisfying $CF = 0$. Then $A + B^TFB$ is positive definite for any F with rank $m - p$.*

Proof. Assume to the contrary that $\mathscr{K}_0$ is nonsingular but $A + B^TFB$ is singular; then, there exists a nonzero vector $x \in \mathbb{R}^n$ such that $(A + B^TFB)x = 0$. Because both A and B^TFB are semidefinite, it must be the case that

$$Ax = 0 \text{ and } B^TFBx = 0.$$

We cannot have a nonzero $x \in \ker(A) \cap \ker(B)$, because if that were the case then the block vector $\begin{bmatrix} x^T & 0^T & 0^T \end{bmatrix}^T$ would be a null vector of $\mathscr{K}_0$. Thus, we must have

$$x \in \ker(A) \text{ and } Bx \in \ker(F). \tag{6.20}$$

Because F has rank $m-p$, its kernel is given by range(C^T). Thus, (6.20) is equivalent to

$$x \in \ker(A) \text{ and } Bx = C^T z,$$

for some $z \in \mathbb{R}^p$. However, this implies that $\begin{bmatrix} x^T & 0^T & -z^T \end{bmatrix}^T$ is a null vector $\mathscr{K}_0$. Thus, we have shown by contradiction that there can be no nonzero null vector of $A + B^T F B$, and therefore $A + B^T F B$ is positive definite. □

We note that the condition that F have rank $m-p$ is a sufficient condition for $A + B^T F B$ to be invertible, but it is not necessary unless A has nullity $m-p$. In general, if A has nullity k, then it is clear that having rank$(F) \geq k$ is a necessary (though not sufficient) condition for invertibility of $A + B^T F B$.

6.4.2 Preconditioner derivation and analysis

Our approach is identical to the approach we used in deriving our classical saddle-point preconditioner: namely, we augment A with as low-rank a weight matrix as possible, and then use the block diagonal Schur complement of the augmented system. Thus, our preconditioner is defined by:

$$\mathscr{M}_F = \begin{bmatrix} A_F & 0 & 0 \\ 0 & S_{F,1} & 0 \\ 0 & 0 & S_{F,2} \end{bmatrix},$$

where $A_F = A + B^T F B$, $S_{F,1} = B A_F^{-1} B^T$ and $S_{F,2} = C S_{F,1}^{-1} C^T$, with F being a matrix of rank nullity(A) that satisfies $CF = 0$. Note that, in order for all these Schur complements to be defined, we require B to be invertible, even though this is not a necessary condition for invertibility of $\mathscr{K}_0$ (as we saw in Proposition 6.9). Thus, we assume for the remainder of this section that B is invertible.

This preconditioner is similar to standard Schur complement approaches, but

in practice it also requires some knowledge of the null-space of the block C in order to form the weight matrix F. In this sense, it is conceptually similar to null-space-based preconditioners for classical saddle-point systems, such as, e.g., [19, 67].

Theorem 6.12. *Let A have nullity k (where $0 \leq k \leq m-p$), B have full row rank, and let $F \in \mathbb{R}^{m\times m}$ be a matrix of rank k that satisfies $CF = 0$. Define the preconditioner*

$$\mathscr{M}_F = \begin{bmatrix} A_F & 0 & 0 \\ 0 & S_{F,1} & 0 \\ 0 & 0 & S_{F,2} \end{bmatrix},$$

where $A_F = A + B^T FB$, $S_{F,1} = BA_F^{-1}B^T$ and $S_{F,2} = CS_{F,1}^{-1}C^T$. The preconditioned operator $\mathscr{M}_F^{-1}\mathscr{K}_0$ has seven distinct eigenvalues given by:

- $\lambda = 1$*, with multiplicity $n-m+k$;*
- $\lambda = -1$*, with multiplicity k;*
- $\lambda = \frac{1\pm\sqrt{5}}{2}$*, each with multiplicity $(m-p)-k$;*
- *The roots of the cubic polynomial $\lambda^3 - \lambda^2 + 2\lambda + 1$ (approximately $-1.2470, 0.4450$, and 1.8019), each with multiplicity p.*

Proof. The eigenvalue equations for the preconditioned system are:

$$Ax + B^T y = \lambda A_F x \tag{6.21a}$$
$$Bx + C^T z = \lambda S_{F,1} y \tag{6.21b}$$
$$Cy = \lambda S_{F,2} z. \tag{6.21c}$$

We begin by searching for eigenvectors with $z = 0$. Eq. (6.21c) tells us that $y \in \ker(C)$. Then, subject to the this constraint, the eigenvalue equations reduce to

$$Ax + B^T y = \lambda A_F x \tag{6.22a}$$
$$Bx = \lambda S_1(F) y. \tag{6.22b}$$

These are precisely the eigenvalue equations (6.7a)-(6.7b) that arise from the anal-

ysis of the block-2×2 case in Theorem 6.8; by obtaining

$$y = \frac{1}{\lambda} S_{F,1}^{-1} Bx \tag{6.23}$$

from (6.22b), substituting into (6.22a), and using identical reasoning as was used in Theorem 6.8, we obtain the equality:

$$A_F^{-1} Ax - \frac{1}{\lambda} VAx + \left(\frac{1}{\lambda} - \lambda\right) x = 0, \tag{6.24}$$

where $V = Z(Z^T AZ)^{-1} Z^T$ with $Z \in \mathbb{R}^{n\times m}$ being a null-space matrix of B. At this point, our analysis cannot proceed in exactly the same manner as in Theorem 6.8 (by simply considering x in the ranges/kernels of the projectors $A_F^{-1}A$ and VA) because we now have the added constraint due to (6.23) that

$$x \in \ker(CS_{F,1}^{-1} B).$$

However, we note that this is clearly true of any $x \in \ker(B)$ (and therefore any $x \in \text{range}(VA)$), as in that case $S_{F,1}^{-1} Bx = 0$. It is in fact also true of any $x \in \ker(A)$. Recall from Lemma 6.6 that

$$S_{F,1}^{-1} = F + (BB^T)^{-1} B(A - AVA) B^T (BB^T)^{-1}. \tag{6.25}$$

Therefore,

$$\begin{aligned} CS_{F,1}^{-1} B &= CFB + C(BB^T)^{-1} B(A - AVA) B^T (BB^T)^{-1} B \\ &= C(BB^T)^{-1} B(A - AVA) B^T (BB^T)^{-1} B. \end{aligned} \tag{6.26}$$

Recalling that $(I - VA)$ is a projector onto $\text{range}(B)$ and $B^T (BB^T)^{-1} B$, we see that

$$C(BB^T)^{-1} B(A - AVA) B^T (BB^T)^{-1} B = C(BB^T)^{-1} B(I - AV)A,$$

and therefore $CS_{F,1}^{-1} Bx = 0$ for any $x \in \ker(A)$. So, by identical reasoning as in Theorem 6.8, we immediately obtain:

- k eigenvectors corresponding to each of the eigenvalues $\lambda = \pm 1$ by selecting

$x \in \ker(A)$.

- $n-m$ additional eigenvectors corresponding to $\lambda = 1$ by setting $x \in \text{range}(VA)$.

We can then obtain the $m-k-p$ eigenvectors corresponding to the eigenvalues $\lambda = \frac{1\pm\sqrt{5}}{2}$ by taking

$$x \in \ker(VA) \cap \text{range}(A_F^{-1}A) \cap \ker(CS_{F,1}^{-1}B).$$

We now consider the case in which $z \neq 0$. We assume that $\lambda \notin \{\pm 1, \frac{1\pm\sqrt{5}}{2}\}$, and we also assume that $x \in \text{range}(A_F^{-1}A)$ (we will confirm the validity of this assumption at the end of the proof). Because $A_F^{-1}A$ is a projector, this means $A_F^{-1}Ax = x$, and (6.21a) gives

$$x = \frac{1}{\lambda - 1} A_F^{-1} B^T y. \tag{6.27}$$

Substituting this into (6.21b) and solving for y yields

$$y = \frac{\lambda - 1}{\lambda^2 - \lambda - 1} S_{F,1}^{-1} C^T z. \tag{6.28}$$

Substituting this into (6.21c) and rearranging yields

$$(\lambda^3 - \lambda^2 - 2\lambda + 1) S_{F,2} z = 0.$$

This, combined with the positive definiteness of $S_{F,2}$ yields the remaining eigenvalues stated in the theorem and accounts for all $n+m+p$ eigenvalues of $\mathscr{M}_F^{-1}\mathscr{K}_0$.

It remains for us to verify our earlier assumption that $x \in \text{range}(A_F^{-1}A)$; we will show here that this holds for any z. From Eqs. (6.27) and (6.28) we can write:

$$x = \frac{1}{\lambda^2 - \lambda - 1} A_F^{-1} B^T S_{F,1}^{-1} C^T z. \tag{6.29}$$

Using (6.25), recalling that $B^T(BB^T)^{-1}B(I - AV) = (I - AV)$, and simplifying (6.29) yields

$$x = \frac{1}{\lambda^2 - \lambda - 1} A_F^{-1} A (I - VA) B^T (BB^T)^{-1} C^T z,$$

which is clearly in $\text{range}(A_F^{-1}A)$, as assumed earlier. □

As in the block-2×2 case, the multiplicity of eigenvalues in the preconditioned operator is smaller in the case when A has minimal rank. The following result follows directly from Theorem 6.12.

Corollary 6.13. *When A is lowest-rank (i.e., has nullity $m-p$), the preconditioned operator $\mathscr{M}_F^{-1}\mathscr{K}_0$ has five distinct eigenvalues given by:* $1,-1$*, and the roots of the cubic polynomial* $\lambda^3-\lambda^2+2\lambda+1$ *(approximately* $-1.2470, 0.4450$*, and* 1.8019*).*

Remark Analogously to the block-2×2 case, we observe that the preconditioner $\mathscr{M}_F$ reduces to the standard Schur complement preconditioner (4.1) when A is positive definite.

Chapter 7

Conclusions

7.1 Summary

In this thesis we provided eigenvalue analysis and considered preconditioning techniques for double saddle-point systems with full-rank and singular leading blocks, as well as for classical saddle-point systems with singular leading blocks. While our focus has been more on theoretical analysis than on practical implementations, a solid theoretical understanding of matrices and preconditioners is an essential first step in the development of solvers.

For double saddle-point systems with a positive definite leading block, we have provided effective eigenvalue bounds based on the eigenvalues and singular values of the matrix blocks. The increasing importance of double saddle-point systems requires attention to spectral properties of the matrices involved. We have shown that energy estimates are an effective tool for obtaining eigenvalue bounds in this case. We then used these results to analyze the performance of a block diagonal Schur complement preconditioner, including in the more realistic scenario where the leading block and Schur complements are replaced by approximations.

We then considered matrices in which the leading block is singular. We have made some contributions to existing work on classical saddle-point systems by providing a nonzero lower bound for the positive eigenvalues of these matrices based on the angles between the ranges of the blocks, and we then used these analyses to construct a preconditioner that yields a constant number of eigenvalues

of the preconditioned operator, regardless of the rank of the leading block. We then extended this to the double saddle-point context and created a preconditioner that again yields a constant number of eigenvalues of the preconditioned operator.

7.2 Future work

1. In terms of eigenvalue analysis, following Remark 4.13, specific assumptions on the magnitudes of the norms of the matrices D and E may yield additional results and insights on the bounds. The rank structure of the blocks may also have a significant effect, and it may be useful to eliminate the positive definiteness requirement of A and/or consider rank-deficient B and C.

2. It would be useful to consider and further develop preconditioners for nonsymmetric double saddle-point systems, This will likely involve considering block triangular, rather than block diagonal, preconditioners (some theoretical work in this area has already been done in [11]). We note that when symmetry is lost, eigenvalue bounds may not be a sufficient tool for predicting convergence rates of Krylov subspace solvers, and we must instead resort to such tools as fields of values [38] or pseudospectra [88]. It may nonetheless be possible to derive effective practical solvers for certain problems, even in the absence of some of the theoretical guarantees we can achieve in the symmetric case. Both theoretical analysis and practical experimentation could lead to useful developments.

3. For classical saddle-point systems with singular leading blocks, future work in this area includes improving the bound in the non-lowest-rank case (for instance, by intelligently selecting the portion of the spectrum that is "dropped" from A).

4. In terms of preconditioning for classical saddle-point systems with a singular leading block, a limitation of our approach is the need to construct a suitable weight matrix W_k. We describe a method for doing this in the numerical experiments that looks only at the structural rank of a modified augmented matrix, but saw in a few cases that this does not guarantee that the augmented matrix will be numerically nonsingular. Another potential area for improve-

ment is that in our experiments we restricted ourselves to diagonal weight matrices with all ones and zeros along the diagonal; it is worth exploring whether other matrix structures or scalings could be more effective.

5. There are several steps required before our preconditioner for double saddle-point systems with a singular leading block can be incorporated into a practical solver. One is, again, the choice of a suitable weight matrix, which is more difficult in this case because the ideal weight matrix is a null matrix of C. There is also a need to develop some strategies (which will likely be problem-dependent) for approximating the expensive terms in the preconditioner, as we did in the classical saddle-point case. Because the unregularized case is less common for double saddle-point systems than classical saddle-point systems, it would also be useful to extend this preconditioner to the case where $D, E \neq 0$.

Bibliography

[1] J. H. Adler, T. R. Benson, E. C. Cyr, P. E. Farrell, S. P. MacLachlan, and R. S. Tuminaro. Monolithic multigrid methods for magnetohydrodynamics. *SIAM Journal on Scientific Computing*, 43(5):S70–S91, 2021. doi:10.1137/20M1348364. URL https://doi.org/10.1137/20M1348364. → page 11

[2] F. Ali Beik and M. Benzi. Iterative methods for double saddle point systems. *SIAM Journal on Matrix Analysis and Applications*, 39(2):902–921, 2018. doi:10.1137/17M1121226. URL https://doi.org/10.1137/17M1121226. → page 1

[3] U. M. Ascher and C. Greif. *A First Course in Numerical Methods*. Society for Industrial and Applied Mathematics, USA, 2011. ISBN 0898719976. → page 7

[4] K. Atkinson, W. Han, and D. Stewart. *Numerical Solution of Ordinary Differential Equations*. 01 2009. ISBN 9780470042946. doi:10.1002/9781118164495. → page 5

[5] A. Beigl, J. Sogn, and W. Zulehner. Robust preconditioners for multiple saddle point problems and applications to optimal control problems. *SIAM Journal on Matrix Analysis and Applications*, 41(4):1590–1615, 2020. doi:10.1137/19M1308426. URL https://doi.org/10.1137/19M1308426. → pages 1, 25

[6] M. Benzi. Preconditioning techniques for large linear systems: a survey. *J. Comput. Phys.*, 182(2):418–477, 2002. ISSN 0021-9991. doi:10.1006/jcph.2002.7176. URL https://doi.org/10.1006/jcph.2002.7176. → page 5

[7] M. Benzi, G. H. Golub, and J. Liesen. Numerical solution of saddle point problems. *Acta Numerica*, 14:1–137, 2005. → pages 4, 5, 27, 39, 71, 87, 104

[8] L. Bergamaschi, J. Gondzio, and G. Zilli. Preconditioning indefinite systems in interior point methods for optimization. *Computational Optimization and Applications*, 28(2):149–171, July 2004. doi:10.1023/B:COAP.0000026882.34332.1b. URL https://doi.org/10.1023/B:COAP.0000026882.34332.1b. → page 25

[9] N. Bootland, A. Bentley, C. Kees, and A. Wathen. Preconditioners for two-phase incompressible Navier–Stokes flow. *SIAM Journal on Scientific Computing*, 41(4):B843–B869, 2019. doi:10.1137/17M1153674. URL https://doi.org/10.1137/17M1153674. → page 14

[10] M. Cai, M. Mu, and J. Xu. Preconditioning techniques for a mixed Stokes/Darcy model in porous media applications. *Journal of Computational and Applied Mathematics*, 233(2):346 – 355, 2009. ISSN 0377-0427. doi:https://doi.org/10.1016/j.cam.2009.07.029. URL http://www.sciencedirect.com/science/article/pii/S0377042709004269. → pages 30, 61

[11] M. Cai, G. Ju, and J. Li. Schur complement based preconditioners for twofold and block tridiagonal saddle point problems, 2021. → pages 1, 13, 48, 49, 52, 113

[12] P. Chidyagwai, S. Ladenheim, and D. B. Szyld. Constraint preconditioning for the coupled Stokes–Darcy system. *SIAM Journal on Scientific Computing*, 38(2):A668–A690, 2016. doi:10.1137/15M1032156. URL https://doi.org/10.1137/15M1032156. → page 31

[13] Y. Choi, C. Farhat, W. Murray, and M. Saunders. A practical factorization of a schur complement for PDE-constrained distributed optimal control. *Journal of Scientific Computing*, 65:576–597, 11 2015. doi:10.48550/ARXIV.1312.5653. → page 25

[14] T. A. Davis and Y. Hu. The University of Florida sparse matrix collection. *ACM Trans. Math. Softw.*, 38(1), Dec 2011. ISSN 0098-3500. doi:10.1145/2049662.2049663. URL https://doi.org/10.1145/2049662.2049663. → page 93

[15] J. W. Demmel. *Applied Numerical Linear Algebra*. SIAM, Jan. 1997. doi:10.1137/1.9781611971446. → page 5

[16] H. C. Elman. Preconditioning for the steady-state Navier–Stokes equations with low viscosity. *SIAM Journal on Scientific Computing*, 20(4):

1299–1316, 1999. doi:10.1137/S1064827596312547. URL https://doi.org/10.1137/S1064827596312547. → page 91

[17] H. C. Elman, D. J. Silvester, and A. J. Wathen. *Finite elements and fast iterative solvers: with applications in incompressible fluid dynamics*. Numerical Mathematics and Scientific Computation. Oxford University Press, New York, 2005. ISBN 978-0-19-852868-5; 0-19-852868-X. → pages 2, 5, 11, 45, 61

[18] V. J. Ervin, J. S. Howell, and I. Stanculescu. A dual-mixed approximation method for a three-field model of a nonlinear generalized Stokes problem. *Computer Methods in Applied Mechanics and Engineering*, 197(33):2886 – 2900, 2008. ISSN 0045-7825. doi:https://doi.org/10.1016/j.cma.2008.01.022. URL http://www.sciencedirect.com/science/article/pii/S004578250800039X. → pages 33, 104

[19] R. Estrin and C. Greif. On nonsingular saddle-point systems with a maximally rank deficient leading block. *SIAM Journal on Matrix Analysis and Applications*, 36(2):367–384, 2015. → pages 12, 72, 83, 84, 108

[20] R. Estrin and C. Greif. Towards an optimal condition number of certain augmented Lagrangian-type saddle-point matrices. *Numerical Linear Algebra with Applications*, 23(4):693–705, 2016. → pages 72, 73, 83

[21] M. Ferronato, A. Franceschini, C. Janna, N. Castelletto, and H. A. Tchelepi. A general preconditioning framework for coupled multiphysics problems with application to contact- and poro-mechanics. *Journal of Computational Physics*, 398:108887, 2019. ISSN 0021-9991. doi:https://doi.org/10.1016/j.jcp.2019.108887. URL https://www.sciencedirect.com/science/article/pii/S0021999119305856. → pages 5, 31

[22] R. Fletcher. An Ideal Penalty Function for Constrained Optimization. *IMA Journal of Applied Mathematics*, 15(3):319–342, 06 1975. ISSN 0272-4960. doi:10.1093/imamat/15.3.319. URL https://doi.org/10.1093/imamat/15.3.319. → page 71

[23] R. Fletcher. Conjugate gradient methods for indefinite systems. In G. A. Watson, editor, *Numerical Analysis*, pages 73–89, Berlin, Heidelberg, 1976. Springer Berlin Heidelberg. ISBN 978-3-540-38129-7. → page 8

[24] A. Forsgren. Inertia-controlling factorizations for optimization algorithms. *Applied Numerical Mathematics*, 43(1):91–107, 2002. ISSN 0168-9274. doi:https://doi.org/10.1016/S0168-9274(02)00119-8. URL https://www.sciencedirect.com/science/article/pii/S0168927402001198. 19th Dundee Biennial Conference on Numerical Analysis. → page 32

[25] R. W. Freund and N. M. Nachtigal. Qmr: a quasi-minimal residual method for non-Hermitian linear systems, 1991. → page 8

[26] M. P. Friedlander and D. Orban. A primal–dual regularized interior-point method for convex quadratic programs. *Mathematical Programming Computation*, 4(1):71–107, Mar 2012. ISSN 1867-2957. doi:10.1007/s12532-012-0035-2. URL https://doi.org/10.1007/s12532-012-0035-2. → page 26

[27] E. Gartland, Jr and A. Ramage. Local stability and a renormalized Newton method for equilibrium liquid crystal director modeling. Working Paper 9, University of Strathclyde, 2012. → page 3

[28] G. Gatica and N. Heuer. A dual-dual formulation for the coupling of mixed-FEM and BEM in hyperelasticity. *SIAM Journal on Numerical Analysis*, 38:380–400, July 2000. doi:10.1137/S0036142999363486. → pages 4, 28, 104

[29] G. Gatica and N. Heuer. A dual-dual formulation for the coupling of mixed-FEM and BEM in hyperelasticity. *SIAM Journal on Numerical Analysis*, 38:380–400, July 2000. doi:10.1137/S0036142999363486. → page 28

[30] G. N. Gatica and N. Heuer. An expanded mixed finite element approach via a dual-dual formulation and the minimum residual method. *Journal of Computational and Applied Mathematics*, 132(2):371 – 385, 2001. ISSN 0377-0427. doi:https://doi.org/10.1016/S0377-0427(00)00440-4. URL http://www.sciencedirect.com/science/article/pii/S0377042700004404. → page 28

[31] G. N. Gatica, M. González, and S. Meddahi. A low-order mixed finite element method for a class of quasi-Newtonian Stokes flows. Part I: a priori error analysis. *Computer Methods in Applied Mechanics and Engineering*, 193(9):881 – 892, 2004. ISSN 0045-7825. doi:https://doi.org/10.1016/j.cma.2003.11.007. URL http://www.sciencedirect.com/science/article/pii/S0045782503005929. → pages 33, 104

[32] G. H. Golub and C. Greif. On solving block-structured indefinite linear systems. *SIAM J. Sci. Comput.*, 24(6):2076–2092, 2003. → pages 71, 72

[33] G. H. Golub and C. F. Van Loan. *Matrix Computations (3rd Ed.)*. Johns Hopkins University Press, Baltimore, MD, USA, 1996. ISBN 0-8018-5414-8. → page 5

[34] G. H. Golub, C. Greif, and J. M. Varah. An algebraic analysis of a block diagonal preconditioner for saddle point systems. *SIAM Journal on Matrix Analysis and Applications*, 27(3):779–792, 2005. doi:10.1137/04060679X. URL https://doi.org/10.1137/04060679X. → pages 23, 87

[35] N. Gould and V. Simoncini. Spectral analysis of saddle point matrices with indefinite leading blocks. *SIAM J. Matrix Analysis Applications*, 31: 1152–1171, 01 2009. doi:10.1137/080733413. → page 4

[36] N. I. Gould. On practical conditions for the existence and uniqueness of solutions to the general equality quadratic programming problem. *Mathematical Programming*, 32:90–99, 1985. → page 51

[37] A. Greenbaum. *Iterative Methods for Solving Linear Systems*. Society for Industrial and Applied Mathematics, 1997. doi:10.1137/1.9781611970937. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611970937. → page 7

[38] A. Greenbaum. Generalizations of the field of values useful in the study of polynomial functions of a matrix. *Linear Algebra and its Applications*, 347 (1):233–249, 2002. ISSN 0024-3795. doi:https://doi.org/10.1016/S0024-3795(01)00555-9. URL https://www.sciencedirect.com/science/article/pii/S0024379501005559. → pages 9, 113

[39] C. Greif and D. Schötzau. Preconditioners for the discretized time-harmonic Maxwell equations in mixed form. *Numer. Linear Algebra Appl.*, 14(4): 281–297, 2007. → pages 12, 14, 23, 32, 76, 83, 84, 85

[40] C. Greif, E. Moulding, and D. Orban. Bounds on eigenvalues of matrices arising from interior-point methods. *SIAM Journal on Optimization*, 24(1): 49–83, 2014. → pages 23, 26

[41] E. Haber, U. M. Ascher, and D. Oldenburg. On optimization techniques for solving nonlinear inverse problems. *Inverse Problems*, 16(5):1263, 2000. URL http://stacks.iop.org/0266-5611/16/i=5/a=309. → page 27

[42] W. Hackbusch. *Multi-Grid Methods and Applications*, volume 4. 01 1985. ISBN 3-540-12761-5. doi:10.1007/978-3-662-02427-0. → page 6

[43] M. R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *Journal of research of the National Bureau of Standards*, 49: 409–436, 1952. → page 8

[44] R. Hiptmair. Operator preconditioning. *Comput. Math. Appl.*, 52(5): 699–706, 2006. ISSN 0898-1221. doi:10.1016/j.camwa.2006.10.008. URL https://doi.org/10.1016/j.camwa.2006.10.008. → page 11

[45] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1 edition, 1985. doi:10.1017/9781139020411. → pages 34, 51

[46] V. E. Howle and R. C. Kirby. Block preconditioners for finite element discretization of incompressible flow with thermal convection. *Numerical Linear Algebra with Applications*, 19(2):427–440, 2012. doi:10.1002/nla.1814. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/nla.1814. → page 3

[47] V. E. Howle, R. C. Kirby, and G. Dillon. Block preconditioners for coupled physics problems. *SIAM Journal on Scientific Computing*, 35(5): S368–S385, 2013. doi:10.1137/120883086. URL https://doi.org/10.1137/120883086. → page 14

[48] N. Huang and C.-F. Ma. Spectral analysis of the preconditioned system for the 3 × 3 block saddle point problem. *Numer. Algorithms*, 81(2):421–444, jun 2019. ISSN 1017-1398. doi:10.1007/s11075-018-0555-6. URL https://doi.org/10.1007/s11075-018-0555-6. → pages 4, 48, 49, 52

[49] I. C. F. Ipsen. A note on preconditioning nonsymmetric matrices. *SIAM J. Sci. Comput.*, 23(3):1050–1051, 2001. ISSN 1064-8275. doi:10.1137/S1064827500377435. URL https://doi.org/10.1137/S1064827500377435. → pages 12, 48

[50] G. Ke, E. Aulisa, G. Bornia, and V. Howle. Block triangular preconditioners for linearization schemes of the Rayleigh–Bénard convection problem. *Numerical Linear Algebra with Applications*, 24(5):e2096, 2017. doi:10.1002/nla.2096. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/nla.2096. e2096 nla.2096. → page 3

[51] D. P. Kouri, D. Ridzal, and R. Tuminaro. KKT preconditioners for PDE-constrained optimization with the Helmholtz equation. *SIAM Journal on Scientific Computing*, 43(5):S225–S248, 2021. doi:10.1137/20M1349199. URL https://doi.org/10.1137/20M1349199. → page 25

[52] Y. Kuznetsov. Spectrally equivalent preconditioners for mixed hybrid discretizations of diffusion equations on distorted meshes. *Journal of Numerical Mathematics - J NUMER MATH*, 11:61–74, 03 2003. doi:10.1515/156939503322004891. → page 14

[53] Y. A. Kuznetsov. Efficient iterative solvers for elliptic finite element problems on nonmatching grids. 10(3):187–212, 1995. doi:doi:10.1515/rnam.1995.10.3.187. URL https://doi.org/10.1515/rnam.1995.10.3.187. → page 12

[54] F. Laakmann, P. E. Farrell, and L. Mitchell. An augmented Lagrangian preconditioner for the magnetohydrodynamics equations at high Reynolds and coupling numbers, 2021. URL https://arxiv.org/abs/2104.14855. → page 14

[55] U. Langer, G. Of, O. Steinbach, and W. Zulehner. Inexact data-sparse boundary element tearing and interconnecting methods. *SIAM Journal on Scientific Computing*, 29(1):290–314, 2007. doi:10.1137/050636243. URL https://doi.org/10.1137/050636243. → pages 4, 30

[56] K.-A. Mardal and R. Winther. Preconditioning discretizations of systems of partial differential equations. *Numer. Linear Algebra Appl.*, 18(1):1–40, 2011. ISSN 1070-5325. doi:10.1002/nla.716. URL https://doi.org/10.1002/nla.716. → page 11

[57] S. Mehrotra. On the implementation of a primal-dual interior point method. *SIAM Journal on Optimization*, 2(4):575–601, 1992. doi:10.1137/0802028. URL https://doi.org/10.1137/0802028. → page 93

[58] B. Morini, V. Simoncini, and M. Tani. Spectral estimates for unreduced symmetric KKT systems arising from interior point methods. *Numer. Linear Algebra Appl.*, 23:776–800, 2016. → page 23

[59] K. W. Morton. *Numerical solution of partial differential equations / K.W. Morton and David Mayers.* Cambridge University Press, Cambridge ;, 2nd ed. edition, 2005. ISBN 0521607930. → page 5

[60] M. F. Murphy, G. H. Golub, and A. J. Wathen. A note on preconditioning for indefinite linear systems. *SIAM J. Sci. Comput.*, 21(6):1969–1972, 2000. ISSN 1064-8275. doi:10.1137/S1064827599355153. URL https://doi.org/10.1137/S1064827599355153. → pages 12, 48, 85

[61] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, New York, NY, USA, 2e edition, 2006. → pages 26, 93

[62] C. C. Paige and M. A. Saunders. Solutions of sparse indefinite systems of linear equations. *SIAM J. Numer. Anal.*, 12(4):617–629, 1975. ISSN 0036-1429. doi:10.1137/0712047. URL https://doi.org/10.1137/0712047. → page 7

[63] J. W. Pearson and J. Pestana. Preconditioners for Krylov subspace methods: an overview. *GAMM-Mitt.*, 43(4):e202000015, 35, 2020. ISSN 0936-7195. doi:10.1002/gamm.202000015. URL https://doi.org/10.1002/gamm.202000015. → page 5

[64] J. W. Pearson and A. Potschka. A preconditioned inexact active-set method for large-scale nonlinear optimal control problems. 2021. https://arxiv.org/abs/2112.05020. → pages 49, 54

[65] J. W. Pearson and A. Potschka. A note on symmetric positive definite preconditioners for multiple saddle-point systems, 2022. → pages 1, 4, 25, 48

[66] J. W. Pearson and A. J. Wathen. A new approximation of the Schur complement in preconditioners for PDE-constrained optimization. *Numerical Linear Algebra with Applications*, 19(5):816–829, 2012. doi:https://doi.org/10.1002/nla.814. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/nla.814. → pages 25, 64, 67

[67] J. Pestana and T. Rees. Null-space preconditioners for saddle point systems. *SIAM Journal on Matrix Analysis and Applications*, 37(3):1103–1128, 2016. doi:10.1137/15M1021349. URL https://doi.org/10.1137/15M1021349. → pages 12, 108

[68] J. Pestana and A. J. Wathen. Natural preconditioning and iterative methods for saddle point systems. *SIAM Review*, 57(1):71–91, 2015. ISSN 0036-1445. doi:10.1137/130934921. URL https://doi.org/10.1137/130934921. → pages 5, 11

[69] E. G. Phillips, H. C. Elman, E. C. Cyr, J. N. Shadid, and R. P. Pawlowski. A block preconditioner for an exact penalty formulation for stationary MHD. *SIAM Journal on Scientific Computing*, 36(6):B930–B951, 2014. doi:10.1137/140955082. URL https://doi.org/10.1137/140955082. → page 3

[70] T. Rees. Github - tyronerees/poisson-control. https://github.com/tyronerees/poisson-control, 2010. Accessed: 2021-10-22. → page 44

[71] T. Rees, H. Dollar, and A. Wathen. Optimal solvers for PDE-constrained optimization. *SIAM Journal on Scientific Computing*, 32(1):271–298, 2010. doi:10.1137/080727154. URL https://doi.org/10.1137/080727154. → pages 25, 43, 49, 64, 67

[72] S. Rhebergen, G. Wells, A. Wathen, and R. Katz. Three-field block preconditioners for models of coupled magma/mantle dynamics. *SIAM Journal on Scientific Computing*, 37(5):A2270–A2294, 2015. doi:10.1137/14099718X. → pages 3, 28

[73] M. Rozložník. *Saddle-point problems and their iterative solution*. Nečas Center Series. Birkhäuser/Springer, Cham, 2018. ISBN 978-3-030-01430-8; 978-3-030-01431-5. doi:10.1007/978-3-030-01431-5. URL https://doi.org/10.1007/978-3-030-01431-5. → pages 4, 5

[74] M. Ruggeri. *Coupling and numerical integration of the Landau-Lifshitz-Gilbert equation*. PhD thesis, Technische Universität Wien, 2016. → page 3

[75] D. Ruiz, A. Sartenaer, and C. Tannier. Refining the lower bound on the positive eigenvalues of saddle point matrices with insights on the interactions between the blocks. *SIAM Journal on Matrix Analysis and Applications*, 39(2):712–736, 2018. doi:10.1137/16M108152X. URL https://doi.org/10.1137/16M108152X. → pages 4, 70

[76] T. Rusten and R. Winther. A preconditioned iterative method for saddlepoint problems. *SIAM J. Matrix Anal. Appl.*, 13:887–904, 1992. → pages 4, 14, 17, 69, 85

[77] Y. Saad. A flexible inner-outer preconditioned gmres algorithm. *SIAM Journal on Scientific Computing*, 14(2):461–469, 1993. doi:10.1137/0914028. URL https://doi.org/10.1137/0914028. → page 11

[78] Y. Saad. *Iterative methods for sparse linear systems*. Society for Industrial and Applied Mathematics, Philadelphia, PA, second edition, 2003. ISBN 0-89871-534-2. doi:10.1137/1.9780898718003. URL https://doi.org/10.1137/1.9780898718003. → pages 5, 6, 7, 9, 11

[79] Y. Saad and M. H. Schultz. Gmres: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM Journal on Scientific and Statistical Computing*, 7(3):856–869, 1986. doi:10.1137/0907058. URL https://doi.org/10.1137/0907058. → page 7

[80] J. Schöberl and W. Zulehner. Symmetric indefinite preconditioners for saddle point problems with applications to PDE-constrained optimization problems. *SIAM Journal on Matrix Analysis and Applications*, 29(3): 752–773, 2007. doi:10.1137/060660977. URL https://doi.org/10.1137/060660977. → page 25

[81] D. Silvester and A. Wathen. Fast iterative solution of stabilised Stokes systems part II: Using general block preconditioners. *SIAM Journal on Numerical Analysis*, 31(5):1352–1367, 1994. doi:10.1137/0731070. URL https://doi.org/10.1137/0731070. → pages 4, 39

[82] D. Silvester, H. Elman, D. Kay, and A. Wathen. Efficient preconditioning of the linearized Navier–Stokes equations for incompressible flow. *Journal of Computational and Applied Mathematics*, 128(1):261–279, 2001. ISSN 0377-0427. doi:https://doi.org/10.1016/S0377-0427(00)00515-X. URL https://www.sciencedirect.com/science/article/pii/S037704270000515X. Numerical Analysis 2000. Vol. VII: Partial Differential Equations. → page 14

[83] T. Sogabe, M. Sugihara, and S.-L. Zhang. An extension of the conjugate residual method to nonsymmetric linear systems. *Journal of Computational and Applied Mathematics*, 226(1):103–113, 2009. ISSN 0377-0427. doi:https://doi.org/10.1016/j.cam.2008.05.018. URL https://www.sciencedirect.com/science/article/pii/S0377042708002264. Special Issue: The First International Conference on Numerical Algebra and Scientific Computing (NASC06). → page 8

[84] J. Sogn and W. Zulehner. Schur complement preconditioners for multiple saddle point problems of block tridiagonal form with application to optimization problems. *IMA Journal of Numerical Analysis*, 39(3): 1328–1359, 05 2018. ISSN 0272-4979. doi:10.1093/imanum/dry027. URL

https://doi.org/10.1093/imanum/dry027. → pages 1, 2, 4, 13, 14, 19, 23, 25, 48, 51, 52, 54

[85] J. C. Strikwerda. *Finite Difference Schemes and Partial Differential Equations, Second Edition*. Society for Industrial and Applied Mathematics, 2004. doi:10.1137/1.9780898717938. URL https://epubs.siam.org/doi/abs/10.1137/1.9780898717938. → pages 5, 8

[86] C. Tannier. *Study of block diagonal preconditioners using partial spectral information to solve linear systems arising in constrained optimization problems*. PhD thesis, Namur Institute for Complex Systems, 2016. → page 5

[87] L. N. Trefethen and D. Bau. *Numerical Linear Algebra*. SIAM, 1997. ISBN 0898713617. → page 5

[88] L. N. Trefethen and M. Embree. *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators*. Princeton University Press, 2005. ISBN 9780691119465. URL http://www.jstor.org/stable/j.ctvzxx9kj. → pages 9, 113

[89] H. A. van der Vorst. Bi-CGSTAB: A fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems. *SIAM Journal on Scientific and Statistical Computing*, 13(2):631–644, 1992. doi:10.1137/0913035. URL https://doi.org/10.1137/0913035. → page 8

[90] F. Warner. *Foundations of differentiable manifolds and Lie groups*. Springer, New York, 1983. → page 16

[91] A. Wathen and D. Silvester. Fast iterative solution of stabilised stokes systems part i: Using simple diagonal preconditioners. *SIAM Journal on Numerical Analysis*, 30(3):630–649, 1993. ISSN 00361429. URL http://www.jstor.org/stable/2158202. → page 3

[92] A. J. Wathen. Preconditioning. *Acta Numer.*, 24:329–376, 2015. ISSN 0962-4929. doi:10.1017/S0962492915000021. URL https://doi.org/10.1017/S0962492915000021. → pages 5, 11

[93] M. Wathen, C. Greif, and D. Schötzau. Preconditioners for mixed finite element discretizations of incompressible MHD equations. *SIAM Journal on Scientific Computing*, 39(6):A2993–A3013, 2017. doi:10.1137/16M1098991. URL https://doi.org/10.1137/16M1098991. → page 14

[94] U. Wilbrandt. *Stokes–Darcy Equations: Analytic and Numerical Analysis*, pages 109–151. 01 2019. ISBN 978-3-030-02903-6. doi:10.1007/978-3-030-02904-3_6. → page 3

[95] J. Xia, P. E. Farrell, and F. Wechsung. Augmented Lagrangian preconditioners for the Oseen-Frank model of nematic and cholesteric liquid crystals, 2020. URL https://arxiv.org/abs/2004.07329. → page 14