

**CONSTRAINT PRECONDITIONING OF SADDLE POINT
PROBLEMS**

A Dissertation
Submitted to
the Temple University Graduate Board

in Partial Fulfillment
of the Requirements for the Degree of
DOCTOR OF PHILOSOPHY

by
Scott Ladenheim
May, 2015

Examining Committee Members:

Daniel B. Szyld, Advisory Chair, Mathematics
Benjamin Seibold, Mathematics
Isaac Klapper, Mathematics
Prince Chidyagwai, University of Loyola-Maryland, Mathematics

©

by

Scott Ladenheim

May, 2015

All Rights Reserved

ABSTRACT

CONSTRAINT PRECONDITIONING OF SADDLE POINT PROBLEMS

Scott Ladenheim

DOCTOR OF PHILOSOPHY

Temple University, May, 2015

Professor Daniel B. Szyld, Chair

This thesis is concerned with the fast iterative solution of linear systems of equations of saddle point form. Saddle point problems are a ubiquitous class of matrices that arise in a host of computational science and engineering applications. The focus here is on improving the convergence of iterative methods for these problems by preconditioning. Preconditioning is a way to transform a given linear system into a different problem for which iterative methods converge faster.

Saddle point matrices have a very specific block structure and many preconditioning strategies for these problems exploit this structure. The preconditioners considered in this thesis are constraint preconditioners. This class of preconditioner mimics the structure of the original saddle point problem.

In this thesis, we prove norm- and field-of-values-equivalence for constraint preconditioners associated to saddle point matrices with a particular structure. As a result of these equivalences, the number of iterations needed for convergence of a constraint preconditioned minimal residual Krylov subspace method is bounded, independent of the size of the matrix. In particular, for saddle point systems that arise from the finite element discretization of partial differential equations (p.d.e.s), the number of iterations it takes for GMRES to converge for these constraint preconditioned systems is bounded (asymptotically), independent of the size of the mesh width. Moreover, we extend these results when appropriate inexact versions of the constraint preconditioner are

used.

We illustrate this theory by presenting numerical experiments on saddle point matrices that arise from the finite element solution of coupled Stokes-Darcy flow. This is a system of p.d.e.s that models the coupling of a free flow to a porous media flow by conditions across the interface of the two flow regions. We present experiments in both two and three dimensions, using different types of elements (triangular, quadrilateral), different finite element schemes (continuous, discontinuous Galerkin methods), and different geometries. In all cases, the effectiveness of the constraint preconditioner is demonstrated.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my family. Without the continual support and love from both of my parents and sisters I would certainly not be where I am today. My passion for life and learning is directly attributable to them and for that I am forever grateful.

I would also like to acknowledge several professors who have played a prominent role in my academic pursuits. First, both Professors Uday Banerjee and Dan Zacharia at Syracuse University, for giving me a strong mathematical foundation and starting me on this journey. I would also like to thank Dr. Carla Martin and soon to be Dr. Emily Miller for my first mathematical research experience; working with both of you at James Madison University made me want to become a research mathematician. I would also like to thank my collaborator Prince Chidyagwai for working with me on this interesting problem. I look forward to our continued research together. I would also like to thank Drs. Panayot Vassilevski and Umberto Villa as well as Lawrence Livermore National Laboratory for supporting me during two summer internships and for truly introducing me to the world of high performance computing. Finally, I would like to thank Professor Valeria Simoncini and the Department of Mathematics at the University of Bologna for hosting me through three of the most fantastic months of my life during the Fall 2013 semester.

I would also like to thank the entire Department of Mathematics at Temple University for giving me a chance. From the moment I set foot in the department I have felt welcome and at home. I will always carry fond memories of attending classes, the Applied Mathematics and Scientific Computing seminar, and the daily interactions with my fellow students and professors.

Lastly, I cannot thank my advisor Professor Daniel Szyld enough for all that he has done these past six years. I remember meeting Daniel on my first visit to Temple. He had returned that day from a conference and though visibly tired, he made time to welcome me to Temple and encouraged me to participate in the Applied Mathematics and Scientific Computing seminar. It

is through his knowledge, assistance, and patience, that I have developed into a proficient mathematician. Thank you for all of the wonderful opportunities and sharing your passion for learning, life, food and most importantly, good coffee with me. I am glad that our paths crossed and that we have shared this time together. I look forward to our future scientific collaborations and a lifelong friendship.

I dedicate this thesis to my sisters Valerie and Katie, the two brightest stars
in my sky.

TABLE OF CONTENTS

ABSTRACT	iii
ACKNOWLEDGEMENT	v
DEDICATION	vii
LIST OF FIGURES	x
LIST OF TABLES	xii
1 INTRODUCTION	1
2 SADDLE POINT PROBLEMS: PROPERTIES AND SOLUTION	6
2.1 Basic properties	7
2.2 Overview of solution methods	8
2.3 The finite element method	16
2.4 Multigrid methods	19
3 CONSTRAINT PRECONDITIONERS FOR SADDLE POINT PROBLEMS	23
3.1 Norm- and field-of-values-equivalences	24
3.2 Inexact constraint preconditioning	30
4 CONSTRAINT PRECONDITIONING OF THE COUPLED STOKES-DARCY SYSTEM	33
4.1 The coupled Stokes-Darcy system	34
4.2 Finite element solution of the coupled Stokes-Darcy system	36
4.3 Preconditioning the coupled Stokes-Darcy system	40
4.4 Numerical Results	44
4.4.1 2D Test Problem: trigonometric solution	44

4.4.2	2D Test Problem: robustness with respect to physical parameters	47
4.4.3	3D: Inexact Preconditioning	54
4.4.4	3D Test Problem: cube domain	57
4.4.5	3D Test Problem: rectangular prism domain	59
4.4.6	3D Test Problem: discontinuous permeability field	61
5	CONCLUSIONS	68
	REFERENCES	70

LIST OF FIGURES

4.1	The domain $\Omega = \Omega_1 \cup \Omega_2$. The Stokes flow region is Ω_1 . The Darcy region is Ω_2 . The Stokes boundary (excluding the interface) Γ_1 is colored blue. The interface Γ_{12} is colored red. The Darcy boundary is composed of Dirichlet (Γ_{2D}) and Neumann (Γ_{2N}) parts.	35
4.2	Plots of the preconditioned H -field-of-values, $\mathcal{W}_H(\mathcal{AP}_{con}^{-1})$ for two mesh widths $h = 2^{-3}, 2^{-4}$. The second figure is magnified near the origin to show that the field-of-values is contained in \mathbb{C}^+ . The dimension of the coupled Stokes-Darcy matrices here are $n = 521$ and $n = 2065$	42
4.3	Spectra of $\Lambda(\mathcal{AP}_{con_i}^{-1})$ for $i = D, T$. The blue circles are the eigenvalues of \mathcal{A} , the red stars are the eigenvalues of $\mathcal{AP}_{con_D}^{-1}$, and the black squares are the eigenvalues of $\mathcal{AP}_{con_T}^{-1}$. The dimensions of the Darcy-conG and Darcy-DG matrices are respectively, $n = 521$ and 2065 , and $n = 1217$ and 4865	43
4.4	Exact and approximate solution computed with \mathcal{P}_{con_D} for Problem (4.14) with Darcy-DG and $h = 2^{-4}$	45
4.5	Residual convergence of the preconditioned GMRES algorithm for Problem (4.14) with a mesh discretization of $h = 2^{-8}$ and $h = 2^{-7}$, respectively.	46
4.6	Residual convergence of the preconditioned GMRES algorithm for Problem (4.15) with $\kappa = 1$ and $\nu = 1$. The corresponding mesh widths are $h = 2^{-8}$ and $h = 2^{-7}$, respectively.	48
4.7	The cube computational domain. The interface between the Stokes and Darcy regions Γ_{12} is colored red. The boundary of the Stokes domain Γ_1 is colored blue. Dirichlet boundary conditions are enforced on all five sides of Γ_1 . On the four lateral sides of the Darcy boundary denoted Γ_{2N} and colored green we enforce Neumann boundary conditions. Finally, the bottom of the Darcy boundary colored in grey and denoted Γ_{2D} is the Dirichlet portion of the boundary.	57

4.8	Inexact preconditioned GMRES solution obtained using \mathcal{P}_{con_D} . The mesh width is $h = 2^{-4}$ and the system size is $n = 576213$.	59
4.9	The rectangular prism computational domain. The interface between the Stokes and Darcy regions Γ_{12} is colored red. The boundary of the Stokes domain is comprised of two portions. The first piece $\Gamma_{1,0}$ consisting of the four lateral sides of the Stokes boundary, is colored blue and denotes where the velocity is zero. The second piece $\Gamma_{1,D}$ at the top of the Stokes boundary is colored yellow and denotes the inflow boundary condition. The Darcy boundary consists of Neumann boundary conditions on the four lateral sides colored green and denoted by Γ_{2N} . The Dirichlet portion of the Darcy boundary, denoted Γ_{2D} is the bottom of the rectangular prism and is colored grey.	61
4.10	3D Prism solutions. The plots show slices of the coupled Stokes-Darcy velocity magnitude and the Darcy pressure. The size of the system matrix for this problem is $n = 265277$ corresponding to a mesh width of $h = 2^{-3}$	63
4.11	Log-log plot of the drop in pressure Δp in the Darcy domain against the decrease of κ	64
4.12	The cube computational domain with a discontinuous permeability field in the Darcy region. The interface between the Stokes and Darcy regions Γ_{12} is colored red. The boundary of the Stokes domain Γ_1 is colored blue. Dirichlet boundary conditions are enforced on all five sides of Γ_1 . On the four lateral sides of the Darcy domain denoted Γ_{2N} and colored in green we enforce Neumann boundary conditions. Finally, on the bottom portion of the Darcy domain denoted Γ_{2D} and colored grey we enforce Dirichlet boundary conditions.	65
4.13	3D discontinuous permeability solutions. The plots show the slices of the coupled Stokes-Darcy velocity magnitude and the Darcy pressure. The size of the system matrix for this problem is $n = 76653$ corresponding to a mesh width of $h = 2^{-3}$	66

LIST OF TABLES

2.1	Characterization of the solution vectors, $\mathbf{x}_m = \mathbf{x}_0 + V_m \mathbf{y}$, for FOM and GMRES.	12
4.1	Number of iterations and CPU times for convergence of Problem (4.14) with $\kappa = 1$, $\nu = 1$	46
4.2	Number of iterations and CPU times for convergence of Problem (4.15) with $\kappa = 1$, $\nu = 1$	48
4.3	Number of iterations and CPU times for convergence of Problem (4.15) with Darcy-conG, varying κ , and fixed $\nu = 1$	50
4.4	Number of iterations and CPU times for convergence of Problem (4.15) with Darcy-DG, varying κ , and fixed $\nu = 1$	51
4.5	Number of iterations and CPU times for convergence of Problem (4.15) with Darcy-conG, varying ν , and fixed $\kappa = 1$	52
4.6	Number of iterations and CPU times for convergence of Problem (4.15) with Darcy-DG, varying ν , and fixed $\kappa = 1$	53
4.7	Table of FGMRES iterations and CPU time in seconds for Problem (4.17) with inexact preconditioners. Varying κ with fixed $\nu = 1$. A * indicates the maximum run time of 24 hours was reached.	58
4.8	Table of FGMRES iterations and time in seconds for flow in the rectangular prism domain with the considered inexact preconditioners.	62
4.9	Table of FGMRES iterations and CPU time in seconds for the discontinuous permeability field problem with inexact preconditioners. A * indicates the maximum run time of 24 hours was reached.	67

CHAPTER 1

INTRODUCTION

The focus of this thesis is on the iterative solution of large-scale linear systems of equations of the form

$$\mathcal{A}\mathbf{x} = \begin{bmatrix} A & B^T \\ C & -D \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \mathbf{b}, \quad (1.1)$$

where $A \in \mathbb{R}^{n \times n}$, $B, C \in \mathbb{R}^{m \times n}$ and $D \in \mathbb{R}^{m \times m}$. This type of linear system is called a saddle point system, and the coefficient matrix is called a saddle point matrix.

These types of problem are ubiquitous in a host of science and engineering applications as they generally arise from energy minimization problems subject to some type of linear constraint; see, e.g., [55]. We briefly list various various application areas where these kinds of matrices arise, for instance in optimization problems, where they are referred to as Karush-Kuhn-Tucker (KKT) systems (see, e.g., [34, 42]), constrained least squares problems (see, e.g., [28]), mixed finite element methods (see, e.g., [11]), as well as from the finite element discretization of (coupled) fluid flow problems; see, e.g., [15, 23]. The saddle point matrices that arise in the above application areas are typically large, i.e., $n + m \approx 10^6$ or larger, and also sparse, meaning a majority of the entries of the matrix \mathcal{A} are zero.

The fast numerical solution for these types of problem is an important task, and a very active area of research; see for instance [6] and the references

therein. Here, the focus is on improving the convergence of preconditioned Krylov subspace methods, specifically the generalized minimal residual (GMRES) method [49], for the solution of (1.1). We consider a class of preconditioners, called constraint preconditioners (see, e.g., [19, 36, 40, 44]), that are indefinite and mimic the structure of the original saddle point problem.

In this thesis we prove norm-equivalence and field-of-values- (f.o.v.-) equivalence between constraint preconditioners and saddle point matrices. As a result of these equivalences, constraint preconditioned minimal residual Krylov subspace methods have convergence properties that are independent of the size of the matrix. Thus, for a sequence of saddle point matrices (of increasing size) corresponding to increasingly refined finite element discretizations of a given partial differential equation (p.d.e.), the number of iterations it takes for a constraint preconditioned Krylov subspace method to converge is bounded, regardless of the mesh discretization size.

This thesis is structured as follows, in Chapter 2, several fundamental properties of saddle point matrices are reviewed, an overview of existing methods for the solution of such problems is given, and some important concepts of the finite element method and multigrid methods are presented. This chapter introduces key conceptual ideas that are used throughout the thesis. The solution overview includes direct methods, basic fixed point iterations, and preconditioned Krylov subspace methods.

In Chapter 3, the theory concerning the convergence properties of constraint preconditioned GMRES is presented. The important consequence of this theory is that a minimal residual Krylov subspace method applied to constraint preconditioned saddle point systems converges in a number of iterations that is bounded independently of the size of the matrix. In particular, for saddle point matrices that arise from the finite element discretization of certain p.d.e.s, the convergence of the constraint preconditioned operator is optimal with respect to the mesh width.

In Chapter 4, the proposed preconditioning approach is used to solve saddle point matrices arising from the finite element discretization of coupled Stokes-

Darcy flow. We consider experiments in both two and three dimensions that illustrate the theoretical results for a variety of geometries and two types of discretization schemes. Concluding remarks are naturally given in the final chapter.

Before proceeding, the following notation is introduced. Boldface is used to denote vectors and calligraphic upper case letters denote block matrices with standard upper case letters for the sub-blocks. The conjugate transpose of a complex-valued matrix A is A^* , where the element of the matrix $A_{ij}^* = \bar{A}_{ji}$. For real-valued matrices A , a superscript T denotes the transpose A^T , and in this case $A_{ij}^T = A_{ji}$. A square matrix is Hermitian when $A^* = A$, and is symmetric when $A^T = A$. A square matrix is skew-Hermitian when $A^* = -A$ and skew-symmetric when $A^T = -A$. A square matrix is positive definite provided $\mathbf{x}^T A \mathbf{x} > 0$ for all non-zero \mathbf{x} , and positive semi-definite if $\mathbf{x}^T A \mathbf{x} \geq 0$ for all non-zero \mathbf{x} .

The inner product between vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ is denoted by (\mathbf{x}, \mathbf{y}) and the induced vector norm by $\|\mathbf{x}\| = (\mathbf{x}, \mathbf{x})^{1/2}$. The standard Euclidean inner product is given by

$$\mathbf{x}^* \mathbf{y} = \sum_{i=1}^n \bar{x}_i y_i$$

and induces the 2-norm (Euclidean-norm) $\|\cdot\|_2 = (\cdot, \cdot)^{1/2}$.

For a symmetric, positive definite matrix H , one can define an H -inner product $(\mathbf{x}, \mathbf{y})_H = \mathbf{x}^* H \mathbf{y}$ and corresponding norm $\|\cdot\|_H = (\cdot, \cdot)_H^{1/2}$. This vector norm also induces the corresponding matrix norm

$$\|M\|_H = \max_{\mathbf{v} \in \mathbb{R}^n \setminus \{0\}} \frac{\|M \mathbf{v}\|_H}{\|\mathbf{v}\|_H}. \quad (1.2)$$

The spectrum of a matrix A is the set of all its eigenvalues and is denoted by

$$\Lambda(A) = \{\lambda : A \mathbf{v} = \lambda \mathbf{v}, \mathbf{v} \in \mathbb{R}^n \setminus \{0\}\}.$$

The spectral radius, $\rho(A) = \max_{\lambda \in \Lambda(A)} |\lambda|$, is the maximum modulus $|\lambda|$ over all eigenvalues in the spectrum $\Lambda(A)$. The H -field of values, for a symmetric

positive definite matrix H , is the following set in the complex plane,

$$\mathcal{W}_H(A) = \left\{ z \in \mathbb{C} : z = \frac{\mathbf{x}^* H A \mathbf{x}}{\mathbf{x}^* H \mathbf{x}}, \forall \mathbf{x} \neq 0 \right\}.$$

The standard field of values, or numerical range [57], is recovered by taking $H = I$, where I is the identity matrix.

The condition number of a square matrix is denoted $\kappa(\mathcal{A}) = \|\mathcal{A}\| \|\mathcal{A}^{-1}\|$. For Hermitian (symmetric) positive definite matrices, $\|\mathcal{A}\| = \lambda_{max}(\mathcal{A})$ and $\|\mathcal{A}^{-1}\| = 1/\lambda_{min}(\mathcal{A})$, where $\lambda_{max}(\mathcal{A})$ and $\lambda_{min}(\mathcal{A})$ denote the maximum and minimum eigenvalues of \mathcal{A} , respectively. Thus, for a Hermitian positive definite matrix \mathcal{A} , the condition number is $\kappa(\mathcal{A}) = \lambda_{max}(\mathcal{A})/\lambda_{min}(\mathcal{A})$.

Let $\mathbf{x} = (x_1, \dots, x_d)^T \in \mathbb{R}^d$. Scalar valued functions are denoted with lower case letters, i.e., $p(\mathbf{x})$, where $p : \mathbb{R}^d \rightarrow \mathbb{R}$. Vector-valued functions are denoted with boldface letters, i.e., $\mathbf{u}(\mathbf{x}) = (u_1(\mathbf{x}), \dots, u_n(\mathbf{x}))$, where $\mathbf{u} : \mathbb{R}^d \rightarrow \mathbb{R}^n$. In this thesis we will only consider the cases when $d = 2, 3$ and $n = 2, 3$.

Define the divergence operator applied to a vector-valued function \mathbf{u} by

$$\nabla \cdot \mathbf{u} = \begin{bmatrix} \sum_{j=1}^d \frac{\partial}{\partial x_j} u_1 \\ \vdots \\ \sum_{j=1}^d \frac{\partial}{\partial x_j} u_n \end{bmatrix}, \quad (1.3)$$

where $\frac{\partial}{\partial x_j}$ denotes differentiation with respect to the variable x_j .

The gradient operator of a vector-valued function is an $n \times d$ matrix whose entries are defined by

$$(\nabla \mathbf{u})_{ij} = \frac{\partial}{\partial \mathbf{x}_j} u_i. \quad (1.4)$$

In addition, we introduce the following set of function spaces and norms; see, e.g., [23, 25]. The $L^p(\Omega)$ function space for scalar functions u is defined as

$$L^p(\Omega) = \left\{ u : \Omega \rightarrow \mathbb{R} : \int_{\Omega} u^p < \infty \right\}, \quad (1.5)$$

which is equipped with the following norm

$$\|u\|_{L^p(\Omega)} = \left(\int_{\Omega} u^p \right)^{1/p}. \quad (1.6)$$

In this thesis, the focus is only for the case $p = 2$, i.e., for $L^2(\Omega)$. The following notation for the Sobolev space is defined as

$$H^1(\Omega) = \{u : \Omega \rightarrow \mathbb{R} : \|u\|_2 < \infty, \|\nabla u\|_2 < \infty\}, \quad (1.7)$$

with corresponding norm

$$\|u\|_{H^1(\Omega)} = \left(\|u\|_{L^2(\Omega)}^2 + \|\nabla u\|_{L^2(\Omega)}^2 \right)^{1/2}.$$

Lastly, the following fundamental definitions are introduced. These definitions will be crucial in establishing bounds that are independent of the size of the matrix on the norm of the GMRES residual for constraint preconditioned saddle point systems. As a result of these bounds, the number of iterations for GMRES to converge is also bounded independent of the size of the constraint preconditioned system.

Definition 1.1. *Two nonsingular matrices $M, N \in \mathbb{R}^{n \times n}$ are H -norm equivalent, $M \sim_H N$, if there exist α_0, β_0 independent of n such that the following holds for all $\mathbf{x} \in \mathbb{R}^n \setminus \{0\}$*

$$\alpha_0 \leq \frac{\|M\mathbf{x}\|_H}{\|N\mathbf{x}\|_H} \leq \beta_0.$$

Definition 1.2. *Two nonsingular matrices $M, N \in \mathbb{R}^{n \times n}$ are H -field-of-values equivalent, $M \approx_H N$, if there exist α_0, β_0 independent of n such that the following holds for all $\mathbf{x} \in \mathbb{R}^n \setminus \{0\}$*

$$\alpha_0 \leq \frac{(MN^{-1}\mathbf{x}, \mathbf{x})_H}{(\mathbf{x}, \mathbf{x})_H} \quad \text{and} \quad \|MN^{-1}\|_H \leq \beta_0.$$

CHAPTER 2

SADDLE POINT PROBLEMS: PROPERTIES AND SOLUTION

This chapter serves as an introduction to saddle point matrices. Here, we introduce fundamental factorizations and properties that are used throughout the thesis. In addition, we give a brief overview of existing solution methods for such problems. The chapter ends by reviewing the finite element method and multigrid methods. The final two sections demonstrate how linear systems arise from the discretization of a p.d.e. and introduces an important class of solution methods for such problems.

2.1 Basic properties

Let the (1,1)-block $A \in \mathbb{R}^{n \times n}$, of the saddle point coefficient matrix (1.1) be nonsingular, then \mathcal{A} admits the following factorizations

$$\begin{bmatrix} A & B^T \\ C & -D \end{bmatrix} = \begin{bmatrix} I & 0 \\ CA^{-1} & I \end{bmatrix} \begin{bmatrix} A & 0 \\ 0 & S \end{bmatrix} \begin{bmatrix} I & A^{-1}B^T \\ 0 & I \end{bmatrix} \quad (2.1a)$$

$$= \begin{bmatrix} I & 0 \\ CA^{-1} & I \end{bmatrix} \begin{bmatrix} A & B^T \\ 0 & S \end{bmatrix} \quad (2.1b)$$

$$= \begin{bmatrix} A & 0 \\ C & S \end{bmatrix} \begin{bmatrix} I & A^{-1}B^T \\ 0 & I \end{bmatrix}, \quad (2.1c)$$

where $S = -(D + CA^{-1}B^T)$ is the Schur complement of A in the block matrix \mathcal{A} ; see, e.g., [6, 28, 59]. Based on any of the above factorizations, it is easy to see that the matrix \mathcal{A} is nonsingular if and only if the Schur complement is invertible. For an overview of the necessary and sufficient conditions for S^{-1} to exist, see [6]. For the situations considered in this thesis S^{-1} exists and we have the following expression for the inverse

$$\mathcal{A}^{-1} = \begin{bmatrix} A^{-1} + A^{-1}B^T S^{-1}CA^{-1} & -A^{-1}B^T S^{-1} \\ -S^{-1}CA^{-1} & S^{-1} \end{bmatrix}. \quad (2.2)$$

An important subset of the class of saddle point matrices is formed when A is Hermitian positive definite, $B = C$ is full rank, and $D = 0$ so that

$$\mathcal{A} = \begin{bmatrix} A & C^T \\ C & 0 \end{bmatrix}.$$

In this case, the Schur complement $S = -CA^{-1}C^T$ is nonsingular and negative definite. Note that by the factorization (2.1a), we have the following congruence relation

$$\begin{bmatrix} A & 0 \\ 0 & S \end{bmatrix} = \begin{bmatrix} I & 0 \\ -CA^{-1} & I \end{bmatrix} \begin{bmatrix} A & C^T \\ C & 0 \end{bmatrix} \begin{bmatrix} I & -A^{-1}C^T \\ 0 & I \end{bmatrix}. \quad (2.3)$$

By Sylvester's law of inertia (see, e.g., [35]) it follows that the matrix \mathcal{A} is indefinite, admitting n positive eigenvalues corresponding to A and m negative

eigenvalues corresponding to S . For further results on the spectra of saddle point matrices see [6] and the references therein.

2.2 Overview of solution methods

There are two options for solving (1.1), namely, direct and iterative methods. Direct methods work by first computing a factorization $\mathcal{A} = \mathcal{M}\mathcal{N}$ and the solution $\mathbf{x} = \mathcal{N}^{-1}\mathcal{M}^{-1}\mathbf{b}$ is computed by (easier) solves with the factors. Two commonly used factorizations are the LU and QR factorizations [48, 57]. The factors of the LU decomposition are a lower and upper triangular matrix for which the solution is then computed by forward- and then back-substitution. The QR factorization produces an orthogonal matrix Q , i.e., $Q^*Q = I$, and an upper triangular matrix R . The solution can then be computed via back-substitution and multiplication by Q^* .

Direct methods are valued for their ability to produce exact solutions, i.e., solutions that would be exact in the absence of round-off error, in a fixed number of steps. In fact, there are very fast and advanced direct methods that are able to exploit the sparsity of a given matrix, such as sparse direct methods [20]. However, for large-scale linear systems, especially those arising from the finite element discretization of p.d.e.s in three dimensions, fill-in of the factorized matrices and the associated growth in memory storage make these methods impractical.

A more economical alternative are iterative methods, which mitigate some of the unavoidable complexity costs associated with direct methods. Starting from an initial vector \mathbf{x}_0 , iterative methods work by producing a sequence of approximations $\{\mathbf{x}_m\}$ that converge to the true solution. The method is stopped and is said to have converged when an appropriate measure of the error, i.e., $\mathbf{e} = \mathbf{x}_m - \mathbf{x}$, or the residual, $\mathbf{r}_m = \mathbf{b} - \mathcal{A}\mathbf{x}_m$ is within a prescribed tolerance of the solution.

The early development of modern iterative methods for general linear systems of equation were fixed point iterations based on splittings of the coefficient

matrix, i.e.,

$$\mathcal{A} = \mathcal{M} - \mathcal{N}. \quad (2.4)$$

From this splitting, a stationary iterative method can be defined by the recurrence

$$\mathbf{x}_{k+1} = \mathcal{M}^{-1}\mathcal{N}\mathbf{x}_k + \mathcal{M}^{-1}\mathbf{b} \quad (2.5a)$$

$$= (I - \mathcal{M}^{-1}\mathcal{A})\mathbf{x}_k + \mathcal{M}^{-1}\mathbf{b}. \quad (2.5b)$$

The second equality follows directly from $\mathcal{N} = \mathcal{M} - \mathcal{A}$.

Two classical stationary iterations are the Jacobi method and Gauss-Seidel method. The Jacobi method results from setting $\mathcal{M} = \mathcal{D}$, where \mathcal{D} is the diagonal of the matrix \mathcal{A} . The Gauss-Seidel method results from taking $\mathcal{M} = \mathcal{L}$ where \mathcal{L} is the lower triangular part of the matrix \mathcal{A} .

The error, $\mathbf{e}_{k+1} = \mathbf{x}_{k+1} - \mathbf{x}$, for this type of iterative method satisfies

$$\mathbf{e}_{k+1} = \mathcal{M}^{-1}\mathcal{N}\mathbf{e}_k = (\mathcal{M}^{-1}\mathcal{N})^k \mathbf{e}_0. \quad (2.6)$$

The necessary and sufficient conditions for such an iteration to converge are that the spectral radius, $\rho(\mathcal{M}^{-1}\mathcal{N}) < 1$. The main advantages of these types of iterative methods are the ease of implementation, as well as the fixed storage cost at each iteration. However, the convergence of these types of methods is linear in the convergence factor $\rho(\mathcal{M}^{-1}\mathcal{N})$, often making the speed of convergence unsatisfactory.

Evolving out of these simple iterative methods, are the more modern Krylov subspace methods, which offer the potential for super-linear convergence [14, 53, 54, 58]. Krylov subspace methods are some of the most popular and widely-used iterative methods for the solution of large, sparse linear systems, including those of the form (1.1). Let

$$\mathcal{K}_m(\mathcal{A}, \mathbf{v}) = \text{span}\{\mathbf{v}, \mathcal{A}\mathbf{v}, \mathcal{A}^2\mathbf{v}, \dots, \mathcal{A}^{m-1}\mathbf{v}\}, \quad (2.7)$$

denote a Krylov subspace of dimension m . These subspaces are nested, i.e., $\mathcal{K}_m \subset \mathcal{K}_{m+1}$. Krylov subspace methods work by finding an approximate solution $\mathbf{x}_m \in \mathbf{x}_0 + \mathcal{K}_m(\mathcal{A}, \mathbf{r}_0)$ for which the corresponding residual $\mathbf{r}_m = \mathbf{b} - \mathcal{A}\mathbf{x}_m$

satisfies the Petrov-Galerkin condition, that is,

$$\mathbf{r}_m \perp \mathcal{L}_m, \quad (2.8)$$

where \mathcal{L}_m belongs to another set of nested subspaces and is commonly referred to as the constraint space [48]. Note that any element of a Krylov subspace can be identified with a polynomial of degree no greater than $m - 1$, i.e., $\mathbf{x}_m = \mathbf{x}_0 + p_{m-1}(\mathcal{A})\mathbf{r}_0$, where $p_{m-1}(\lambda) \in \mathbb{P}_{m-1} := \{p(\lambda) : p(\lambda) = \sum_{i=0}^{m-1} \alpha_i \lambda^i\}$ is the set of polynomials of degree no larger than $m - 1$. Consequently,

$$\mathbf{r}_m = \mathbf{b} - \mathcal{A}(\mathbf{x}_0 + p_{m-1}(\mathcal{A})\mathbf{r}_0) \quad (2.9a)$$

$$= (I - \mathcal{A}p_{m-1}(\mathcal{A}))\mathbf{r}_0 \quad (2.9b)$$

$$= q_m(\mathcal{A})\mathbf{r}_0, \quad (2.9c)$$

where $q_m(\lambda) \in \mathbb{P}_m$ and $q(0) = 1$. This viewpoint is critical when establishing bounds on the residual for various Krylov subspace methods.

Different Krylov subspace methods are then determined by the choice of the nested subspaces \mathcal{L}_m . For instance, when $\mathcal{L}_m = \mathcal{K}_m$, one obtains the Full Orthogonalization Method (FOM) [48]. Additionally, if \mathcal{A} is symmetric positive definite (s.p.d.) the FOM algorithm simplifies to the Conjugate Gradient (CG) method of Hestenes and Steiffel [32]. The CG algorithm also possesses the highly desirable short-term recurrence property. Specifically, the basis vectors of the Krylov subspace satisfy a three-term recurrence, implying that the storage costs are fixed for each iteration of the algorithm. Additionally, the convergence of CG is completely determined by the eigenvalues [28, 48] and this property can be exploited for developing extremely effective preconditioning strategies.

Another customary choice is $\mathcal{L}_m = \mathcal{AK}_m$, which is equivalent to the minimal residual condition

$$\|\mathbf{r}_m\|_2 = \min_{\mathbf{x} \in \mathbf{x}_0 + \mathcal{AK}_m} \|\mathbf{b} - \mathcal{A}\mathbf{x}\|_2. \quad (2.10)$$

This choice of constraint space leads to the class of minimal residual methods for which the classical example is the Generalized Minimal Residual (GMRES)

method of Saad and Schultz [49]. Again, simplifications can be made in the case when the matrix is symmetric, though possibly indefinite, resulting in another short-term recurrence method and this is the MINRES algorithm [43].

A necessary component of any Krylov subspace method is an algorithm that constructs an orthogonal basis for $\mathcal{K}_m(A, \mathbf{r}_0)$. This is accomplished by the Arnoldi algorithm, given in Algorithm 1. At step j in the algorithm, the vector $A\mathbf{v}_j$ is orthogonalized against all previous vectors \mathbf{v}_i in a Gram-Schmidt type procedure.

Algorithm 1 Arnoldi

- 1: Given initial residual \mathbf{r}_0 , set $\beta = \|\mathbf{r}_0\|_2$ and $\mathbf{v}_1 = \mathbf{r}_0/\beta$.
 - 2: **for** $k = 1, \dots, m$ **do**
 - 3: $\mathbf{w} = A\mathbf{v}_k$
 - 4: **for** $j = 1, \dots, k$ **do**
 - 5: $h_{j,k} = \mathbf{v}_j^T \mathbf{w}$
 - 6: $\mathbf{w} = \mathbf{w} - \mathbf{v}_j h_{j,k}$
 - 7: **end for**
 - 8: $h_{k+1,k} = \|\mathbf{w}\|_2$
 - 9: $\mathbf{v}_{k+1} = \mathbf{w}/h_{k+1,k}$
 - 10: **end for**
-

At the end of Algorithm 1, the basis vectors \mathbf{v}_j can be gathered into the columns of the matrix V_m and the constants $h_{j,k}$ into the matrix \bar{H}_m . This gives the classical Arnoldi relation

$$AV_m = V_{m+1}\bar{H}_m = V_m H_m + h_{m+1,m} v_{m+1} e_m^T, \quad (2.11)$$

where $V_m \in \mathbb{R}^{n \times m}$ and $\bar{H}_m \in \mathbb{R}^{(m+1) \times m}$ is an upper Hessenberg matrix. Note that due to the orthogonality of V_m we have that $V_m^T AV_m = H_m$. For the case when A is symmetric, the upper Hessenberg matrix H_m becomes a tridiagonal matrix and is usually denoted T_m .

The Arnoldi relation also forms the basis for constructing approximate solutions for Krylov subspace methods. Recall that at the m^{th} step of a Krylov

subspace method an approximate solution of the form $\mathbf{x}_m = \mathbf{x}_0 + V_m \mathbf{y}$ is produced. Using (2.11), we have the equivalent formulation of the residual,

$$\mathbf{r}_m = \mathbf{b} - A\mathbf{x}_m = \mathbf{r}_0 - AV_m \mathbf{y} = \mathbf{r}_0 - V_{m+1} \bar{H}_m \mathbf{y}. \quad (2.12)$$

Therefore, using (2.11) and (2.12), we summarize in Table 2.1 the following characterizations of the vector \mathbf{y} depending on the choice of the constraint space \mathcal{L}_m in (2.8).

$\mathcal{L}_m = \mathcal{K}_m$	$\mathcal{L}_m = A\mathcal{K}_m$
$\mathbf{r}_m \perp \mathcal{K}_m(A, \mathbf{r}_0)$	$\ \mathbf{r}_m\ _2 = \min_{\mathbf{x} \in \mathbf{x}_0 + A\mathcal{K}_m} \ \mathbf{b} - A\mathbf{x}_m\ _2$
$\iff V_{m+1}^T \mathbf{r}_m = 0$	$\iff \ \mathbf{r}_m\ _2 = \min_{\mathbf{y}} \ \mathbf{r}_0 - V_{m+1} \bar{H}_m \mathbf{y}\ _2$
$\iff \beta \mathbf{e}_1 - H_m \mathbf{y} = 0$	$\iff \min_{\mathbf{y}} \ V_{m+1}(\beta \mathbf{e}_1 - \bar{H}_m \mathbf{y})\ _2$
$\iff \mathbf{y} = H_m^{-1}(\beta \mathbf{e}_1)$	$\mathbf{y} = \min \ \beta \mathbf{e}_1 - \bar{H}_m \mathbf{y}\ _2$

Table 2.1: Characterization of the solution vectors, $\mathbf{x}_m = \mathbf{x}_0 + V_m \mathbf{y}$, for FOM and GMRES.

These characterizations then allow us to state the FOM and GMRES methods in Algorithm 2.

We now state two standard bounds on the convergence of the GMRES algorithm. These bounds will be used in Chapter 3 to establish mesh-independent convergence properties for the MINRES and GMRES algorithm when applied to constraint preconditioned saddle point matrices.

The first bound has the form

$$\frac{\|\mathbf{r}_m\|_2}{\|\mathbf{r}_0\|_2} \leq \kappa(X) \min_{q_m \in \mathbb{P}_m} \max_{\lambda_i \in \Lambda(A)} |q(\lambda_i)| \quad (2.13)$$

This bound can be derived by assuming that the system matrix is diagonalizable, i.e., $A = X\Lambda X^{-1}$. Then, using the polynomial representation of the GMRES residual, recall (2.9), we have that

Algorithm 2 FOM/GMRES

- 1: Given initial residual \mathbf{r}_0 , set $\beta = \|\mathbf{r}_0\|_2$ and $\mathbf{v}_1 = \mathbf{r}_0/\beta$.
 - 2: **for** $k = 1, \dots$, convergence **do**
 - 3: $\mathbf{w} = A\mathbf{v}_1$
 - 4: **for** $j = 1, \dots, k$ **do**
 - 5: $h_{j,k} = \mathbf{v}_j^T \mathbf{w}$
 - 6: $\mathbf{w} = \mathbf{w} - \mathbf{v}_j h_{j,k}$
 - 7: **end for**
 - 8: $h_{k+1,k} = \|\mathbf{w}\|_2$
 - 9: $\mathbf{v}_{k+1} = \mathbf{w}/h_{k+1,k}$
 - 10: **if** GMRES **then**
 - 11: Compute $\mathbf{y}_k^{GMRES} = \min \|\beta \mathbf{e}_1 - \bar{H}\mathbf{y}\|_2$
 - 12: **else if** FOM **then**
 - 13: Compute $\mathbf{y}^{FOM} = H_m^{-1}(\beta \mathbf{e}_1)$
 - 14: **end if**
 - 15: $\mathbf{x}_k = \mathbf{x}_0 + V_k \mathbf{y}_k$
 - 16: **end for**
-

$$\begin{aligned}
\|\mathbf{r}_m\|_2 &\leq \min_{\substack{q_m \in \mathbb{P}_m, \\ q_m(0)=1}} \|q_m(\mathcal{A})\|_2 \|\mathbf{r}_0\|_2 \\
&= \min_{\substack{q_m \in \mathbb{P}_m, \\ q_m(0)=1}} \|X q_m(\Lambda) X^{-1}\|_2 \|\mathbf{r}_0\|_2 \\
&= \kappa(X) \min_{\substack{q_m \in \mathbb{P}_m, \\ q_m(0)=1}} \max_{\lambda_i \in \Lambda(\mathcal{A})} |q_m(\lambda_i)| \|\mathbf{r}_0\|_2,
\end{aligned}$$

which is precisely the bound given in (2.13).

When $\kappa(X) = 1$ the convergence is determined completely by the spectrum. This is true, in particular, for symmetric matrices. We remark that even if the matrix is not diagonalizable, similar bounds can still be derived. In this situation, the bounds are based on the Jordan decomposition $A = XJX^{-1}$, see for instance [24, 54].

The GMRES residual can also be bounded using the field of values. This bound assumes that $0 \notin \mathcal{W}(A)$, hence, the minimal distance from the field of values to the origin, denoted $\nu(\mathcal{W}(\mathcal{A})) = \min_{z \in \mathcal{W}(\mathcal{A})} |z|$, is positive. The field of values bound then takes the form

$$\|\mathbf{r}_m\|_2 \leq (1 - \nu(\mathcal{W}(\mathcal{A}))\nu(\mathcal{W}(\mathcal{A}^{-1})))^{m/2} \|\mathbf{r}_0\|_2, \quad (2.14)$$

see, e.g., [21, 38].

Obtaining sufficiently rapid convergence of a given Krylov subspace method usually requires an additional crucial component, namely, preconditioning. Preconditioning is a way to transform the original linear system into a different system for which Krylov subspace method convergence is faster. Specifically, a preconditioner is an invertible operator \mathcal{P}^{-1} that multiplies the original matrix either on the right or left.

For normal problems, the bound in (2.13) implies that a good preconditioner should cluster the eigenvalues so that the value $|q(\lambda_i)|$ is small. In fact, many preconditioning strategies attempt to cluster the eigenvalues. However, for non-normal problems, the eigenvalues alone do not determine the convergence of the problem.

In the case of right preconditioning, this reformulates the original linear system (1.1) into the following form

$$A\mathcal{P}^{-1}\mathbf{u} = \mathbf{b} \quad \mathbf{u} = \mathcal{P}\mathbf{x}. \quad (2.15)$$

Thus, a right-preconditioned Krylov subspace method finds at the m^{th} step, $\mathbf{u}_m \in \mathbf{u}_0 + \mathcal{K}_m(A\mathcal{P}^{-1}, \mathbf{r}_0)$, such that

$$\|\mathbf{r}_m\|_2 = \|\mathbf{b} - A\mathcal{P}^{-1}\mathbf{u}_m\|_2 = \min_{\mathbf{u} \in \mathbf{u}_0 + \mathcal{K}_m(A\mathcal{P}^{-1}, \mathbf{r}_0)} \|\mathbf{b} - A\mathcal{P}^{-1}\mathbf{u}\|_2. \quad (2.16)$$

By definition, we have that the residual satisfies

$$\mathbf{r}_m = \mathbf{b} - A\mathcal{P}^{-1}\mathbf{u}_m = \mathbf{b} - A\mathbf{x}_m. \quad (2.17)$$

Hence, right-preconditioned minimal residual methods minimize the same residual as standard, non-preconditioned methods.

Using the definition of the vector \mathbf{u}_m in (2.15) and the fact that

$$\mathcal{P}^{-1}\mathcal{K}_m(A\mathcal{P}^{-1}, \mathbf{r}_0) = \mathcal{K}_m(\mathcal{P}^{-1}A, \mathcal{P}^{-1}\mathbf{r}_0),$$

we have the following characterization of right-preconditioned minimal residual Krylov subspace methods for the vector \mathbf{x}_m . Find $\mathbf{x}_m \in \mathbf{x}_0 + \mathcal{K}_m(\mathcal{P}^{-1}A, \mathcal{P}^{-1}\mathbf{r}_0)$ such that the norm of the residual is minimal over this affine space. We conclude by noting that left-preconditioned minimal residual methods extract approximate solutions from the same space as right-preconditioned ones, but minimize the norm of the preconditioned residual, $\mathcal{P}^{-1}\mathbf{r}_m$; see, e.g., [48].

There are two fundamental block preconditioning techniques for the solution of (1.1), namely block diagonal and block lower triangular preconditioners of the form

$$\mathcal{P}_{\text{bd}} = \begin{bmatrix} P_1 & 0 \\ 0 & P_2 \end{bmatrix}, \quad (2.18a)$$

$$\mathcal{P}_{\text{lt}} = \begin{bmatrix} P_1 & 0 \\ C & P_2 \end{bmatrix}. \quad (2.18b)$$

Typically, P_1 is chosen to approximate the (1,1) block of \mathcal{A} and P_2 is chosen to approximate the Schur complement, i.e. $P_1 \approx A$ and $P_2 \approx -S$, respectively. These choices are motivated by the following two fundamental theorems proved in [41].

Theorem 2.1. *Let \mathcal{A} be of the form (1.1) with $D = 0$, and the blocks of \mathcal{P}_{bd} such that $P_1 = A$ and $P_2 = -S = CA^{-1}B^T$, then the matrix $\mathcal{T}_{bd} = \mathcal{P}_{bd}^{-1}\mathcal{A}$ satisfies $\mathcal{T}_{bd}(\mathcal{T}_{bd} - I)(\mathcal{T}_{bd}^2 - \mathcal{T}_{bd} - I) = 0$.*

Theorem 2.2. *Let \mathcal{A} be of the form (1.1) with $D = 0$, and the blocks of \mathcal{P}_{lt} such that $P_1 = A$ and $P_2 = -S = CA^{-1}B^T$, then the matrix $\mathcal{T}_{lt} = \mathcal{P}_{lt}^{-1}\mathcal{A}$ satisfies $(\mathcal{T}_{lt} - I)^2 = 0$.*

The significant consequence of these two theorems is that in exact arithmetic, a minimal residual Krylov subspace method applied to the nonsingular preconditioned systems \mathcal{T}_{bd} , \mathcal{T}_{lt} , converges in three and two iterations, respectively.

In practice though, computations are not done in exact arithmetic. Using exact representations of both the A -block and Schur complement is typically impractical and this motivates the use of more economical, inexact versions of these blocks. Ideally, these cheaper, inexact versions of the preconditioner still maintains the favorable convergence properties of the exact versions of the preconditioner. We mention the following references for further theoretical developments on block diagonal preconditioners [17, 52] and block triangular preconditioners [37, 60] for the solution of (1.1).

2.3 The finite element method

Here, we briefly review the finite element method. The finite element method is one of the most popular and widely used methods for numerically solving a given p.d.e. For classical theoretical treatments of the method, see for instance [9, 16]. For more modern texts, see for instance [27] where implementation details are discussed, and [23, 59] for topics on the fast iterative

solution and block preconditioning for matrices that arise from the finite element discretization of p.d.e.s, respectively.

To simplify the exposition and introduce notation for subsequent sections, we consider the solution of a Poisson problem with homogeneous Dirichlet boundary conditions on a domain Ω with boundary $\partial\Omega$, that is,

$$-\nabla \cdot (\nabla u) = f, \quad \text{in } \Omega, \quad (2.19a)$$

$$u = 0, \quad \text{on } \partial\Omega. \quad (2.19b)$$

The idea of the finite element method is to first partition or mesh the domain Ω into simple smaller pieces, typically tetra- or hexahedrons. The characteristic size of the elements in the discretized domain is called the mesh width and is denoted by h . The next step is to construct an appropriate finite dimensional solution space, corresponding to the discretized domain, i.e., $V^h \subset V$, where V is the infinite-dimensional solution space. The solution space V is precisely defined shortly. The approximate finite element solution is constructed from this finite dimensional space and the idea is that as the mesh width $h \rightarrow 0$, the approximate finite element solution converges to the true solution.

The finite element space V^h is typically formed by associating to each vertex, or node \mathbf{x}_i of the mesh, a locally supported basis function $\phi_i(\mathbf{x})$. This type of nodal basis function can be defined as follows

$$\phi_i(\mathbf{x}) = \begin{cases} 1 & \mathbf{x} = \mathbf{x}_i, \\ 0 & \mathbf{x} = \mathbf{x}_j. \end{cases}$$

Therefore, any element of this finite element space V^h can be expressed by a linear combination of the basis functions, i.e.,

$$u_h = \sum_{i=1}^n u_i \phi_i(\mathbf{x}). \quad (2.20)$$

In order to find the finite element solution, the original p.d.e. is transformed into its weak or variational formulation. This is accomplished by first multiplying the p.d.e. by test functions v which also satisfy the homogeneous

boundary conditions, i.e., $v|_{\partial\Omega} = 0$, and then integrating over the entire domain, yielding

$$-\int_{\Omega} (\nabla \cdot (\nabla u))v \, d\mathbf{x} = \int_{\Omega} f v \, d\mathbf{x}. \quad (2.21)$$

Using the Green's formula [25, p. 628], we have that

$$-\int_{\Omega} (\nabla \cdot (\nabla u))v \, d\mathbf{x} = \int_{\Omega} \nabla u \cdot \nabla v \, d\mathbf{x} - \int_{\partial\Omega} v(\nabla u \cdot \mathbf{n}) \, dS \quad (2.22)$$

$$= \int_{\Omega} \nabla u \cdot \nabla v \, d\mathbf{x}, \quad (2.23)$$

where the second equality follows from $v|_{\partial\Omega} = 0$.

Introducing the following notation for the corresponding bilinear form and linear functional defined above,

$$a(u, v) := \int_{\Omega} \nabla u \cdot \nabla v \, d\mathbf{x}, \quad (2.24)$$

$$(f, v) := \int_{\Omega} f v \, d\mathbf{x}. \quad (2.25)$$

The continuous weak problem is the following, find $u \in V$ such that

$$a(u, v) = (f, v) \quad \forall v \in V. \quad (2.26)$$

Thus, solutions of the above weak problem only require the existence of first order derivatives, rather than the existence of second order derivatives required for the original p.d.e. The natural function space to search for solutions is then $V := H_0^1(\Omega) = \{v \in H^1(\Omega) : v = 0 \text{ on } \partial\Omega\}$, the Sobolev space introduced in (1.7) with the condition that the functions have homogeneous boundary conditions.

The discrete solution is then found by inserting the approximate solution, u_h into the above bilinear form and testing against all of the basis functions in the finite element space V^h , i.e.,

$$a(u_h, \phi_i) = (f, \phi_i) \quad \text{for } i = 1, \dots, n. \quad (2.27)$$

The above equations are equivalent to the following linear system of equations for the unknowns, also referred to as degrees of freedom (d.o.f.s) u_i ,

$$\mathbf{A}\mathbf{u} = \mathbf{f}. \quad (2.28)$$

The entries of the matrix A and the vectors \mathbf{u} , \mathbf{f} , are

$$A_{ij} = \int_{\Omega} \phi_i \phi_j \, d\mathbf{x}, \quad \mathbf{u}_i = u_i, \quad \text{and} \quad \mathbf{f}_i = \int_{\Omega} \mathbf{f} \phi_i \, d\mathbf{x}. \quad (2.29)$$

The finite element solution u_h is then obtained by solving this linear system for the above vector \mathbf{u} . Note that the dimension of the system matrix A is equal to the number of d.o.f.s and increases as the mesh is refined. Moreover, due to the basis functions being locally supported on the mesh, the resulting linear system is sparse. In Chapter 4, the finite element solution for the Stokes-Darcy system, a more complex set of coupled p.d.e.s is introduced.

2.4 Multigrid methods

As described in the previous section, a crucial step in any finite element method is the solution of a linear system of equations for the variables of interest. Here, we introduce a very important and powerful class of methods called multigrid methods. These methods were originally developed for solving the linear systems arising from the finite difference (or finite element) discretization of p.d.e.s based on a hierarchy of meshes. For references on the classical geometric multigrid method (MG) based on this hierarchy of grids, see, for instance, [12, 30]. For more modern references on algebraic multigrid methods (AMG), which are solvers based only on the structure of the matrix rather than a hierarchy of refined meshes, see, for instance, [8, 26, 46]. Though originally developed for solving the linear systems from discretized p.d.e.s, these methods are also extremely efficient and powerful preconditioners. In this thesis, the focus is on the latter case, where multigrid methods are used as preconditioners for the solution of linear systems by Krylov subspace methods.

The finite element method produces a sequence of linear systems of the form of

$$\mathcal{A}^h \mathbf{x} = \mathbf{b}^h, \quad (2.30)$$

where each system corresponds to a particular mesh in a hierarchy of increasingly refined meshes. The solution of such matrices is often difficult for

iterative solvers due to the dependence of the condition number $\kappa(\mathcal{A}^h)$ on the mesh width, i.e., decreasing the mesh width induces a corresponding increase in the conditioner number. Thus, a goal in the design of many preconditioners \mathcal{P} is that the condition number of the preconditioned system $\kappa(\mathcal{A}^h\mathcal{P}^{-1})$ does not depend on the mesh width. Preconditioners with this property are said to be optimal with respect to the mesh

Here, we describe a basic multigrid method, known as the V-cycle, that converges independently of the mesh for elliptic problems of the form (2.19); see, e.g., [7, 23, 59]. As a result, it can be shown that the V-cycle is an optimal preconditioner with respect to the mesh width for the solution of (2.30). This result is established using the concept of spectral equivalence, a special case of f.o.v.-equivalence that is obtained by setting $H = I$ in Definition 1.2. The precise definition is given below.

Definition 2.1. *Two nonsingular, symmetric matrices $M, N \in \mathbb{R}^{n \times n}$ are spectrally equivalent if there exist positive constants α_0, β_0 , independent of the mesh width such that the following holds for all $\mathbf{x} \in \mathbb{R}^n \setminus \{0\}$*

$$\alpha_0 \leq \frac{(MN^{-1}\mathbf{x}, \mathbf{x})}{(\mathbf{x}, \mathbf{x})} \text{ and } \|MN^{-1}\| \leq \beta_0 \quad (2.31)$$

For symmetric problems in general, this definition establishes the important result that an effective preconditioner is one which captures the spectra of the original operator. In Chapter 3.2, spectral equivalence is used to establish the optimality of inexact versions of constraint preconditioners for saddle point problems.

To introduce the idea of a multigrid V-cycle, let $\mathbf{x} = (\mathcal{A}^h)^{-1}\mathbf{b}^h$ denote the true solution and $\tilde{\mathbf{x}}$ an approximation. Note that the error, $\mathbf{e} = \mathbf{x} - \tilde{\mathbf{x}}$ satisfies the following relation

$$\mathcal{A}^h\mathbf{e} = \mathcal{A}^h(\mathbf{x} - \tilde{\mathbf{x}}) = (\mathbf{b}^h - \mathcal{A}^h\tilde{\mathbf{x}}) = \tilde{\mathbf{r}}, \quad (2.32)$$

where $\tilde{\mathbf{r}}$ is corresponding residual. This says that the error satisfies an equation of the form (2.30) but with the residual $\tilde{\mathbf{r}}$ as the right hand side.

The idea of multigrid methods is to incorporate information from all of the grids to solve (2.30). First, consider the simple situation where the hierarchy of grids consists of two grids only, a fine grid and a coarse grid. Let $R : V^h \rightarrow V^H$ denote the restriction operator from the fine grid space to the coarse grid space, and $P : V^H \rightarrow V^h$ denote the prolongation operator from the coarse grid space to the fine grid space. The matrices on the fine grid and coarse grid are denoted by \mathcal{A}^h and $\mathcal{A}^H := R\mathcal{A}^hP$, respectively.

Multigrid methods work by first applying a stationary iterative method, often a (weighted) Jacobi or Gauss-Seidel iteration, which is defined by the choice of the matrix \mathcal{M} in the splitting, recall equation (2.4). In this context, the application of these iterative methods, see equation (2.5b), is called a smoothing process. The nomenclature arises because such stationary iterations are effective at reducing, or smoothing the highly oscillatory error components. However, the smooth error components still remain. Thus, in order to improve the smoothed fine-grid solution, some approximation of the error is added to it. The next key ingredient of these methods is obtaining an approximation to the error. The idea is to use the coarse grid to obtain this approximation, where smooth error becomes more oscillatory.

The coarse grid error can be obtained by restricting the newly formed residual $\tilde{\mathbf{r}}$ to the coarse grid and then solving (2.32) with the restricted residual and coarse grid matrix \mathcal{A}^H , i.e.,

$$\mathcal{A}^H \mathbf{e}^H = (R\mathcal{A}^hP)\mathbf{e}^H = R\tilde{\mathbf{r}}. \quad (2.33)$$

The coarse grid error, $\mathbf{e}^H = (\mathcal{A}^H)^{-1}R\tilde{\mathbf{r}}$ is then prolonged to the fine grid $P\mathbf{e}^H$ and added to the previously smoothed solution, $\tilde{\mathbf{x}} = \tilde{\mathbf{x}} + P\mathbf{e}^H$. The subsequent iterate, \mathbf{x}_1 is then obtained by applying a smoothing iteration to $\tilde{\mathbf{x}}$. This smoothing iteration is defined by the matrix \mathcal{M}^T . The reason for using \mathcal{M}^T is to ensure that the two-grid error propagation matrix defined by this algorithm is symmetric with respect to the \mathcal{A} inner product; see [23]. The above outlined procedure is summarized in Algorithm 3.

The full multigrid V-cycle is obtained from the two-grid method by recur-

Algorithm 3 Two-Grid Algorithm

- 1: Given \mathcal{A}^h , \mathbf{b}^h , \mathbf{x}_0 , and operators R, P .
 - 2: **for** $k=0, \dots$, until convergence **do**
 - 3: Pre-smooth: $\tilde{\mathbf{x}} = (I - \mathcal{M}^{-1}\mathcal{A}^h)\mathbf{x}_k + \mathcal{M}^{-1}\mathbf{b}^h$.
 - 4: Restrict the residual: $\mathbf{r}^H = R\tilde{\mathbf{r}} = R(\mathbf{b}^h - \mathcal{A}^h\tilde{\mathbf{x}})$.
 - 5: Solve for the coarse grid correction: $\mathbf{e}^H = (\mathcal{A}^H)^{-1}\mathbf{r}^H$.
 - 6: Prolongate the coarse grid error: $\mathbf{e}^h = P\mathbf{e}^H$.
 - 7: Correct the smoothed approximation: $\tilde{\mathbf{x}} = \tilde{\mathbf{x}} + \mathbf{e}^h$.
 - 8: Post-smooth: $\mathbf{x}_{k+1} = (I - \mathcal{M}^{-T}\mathcal{A}^h)\tilde{\mathbf{x}} + \mathcal{M}^{-T}\mathbf{b}^h$.
 - 9: Check convergence.
 - 10: **end for**
-

sively calling the two-grid method to solve for the coarse grid correction in line 5 of Algorithm 3. The power of multigrid methods is that they are scalable, i.e., they offer the potential of solving a linear system of size n in $\mathcal{O}(n)$ operations, and optimal with respect to the mesh width for elliptic problems, i.e., they converge independently of the mesh width. For proofs of these result, which use the notion of spectral equivalence, see for instance [23, 59].

CHAPTER 3

CONSTRAINT

PRECONDITIONERS FOR

SADDLE POINT PROBLEMS

In this chapter we consider the solution of an important subclass of saddle point systems (1.1). For this case, the blocks of the coefficient matrices are such that $A \in \mathbb{R}^{n \times n}$ is nonsingular, $B = C \in \mathbb{R}^{m \times n}$ is of full rank, and $D = 0$, i.e.,

$$\mathcal{A}\mathbf{x} = \begin{bmatrix} A & C^T \\ C & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \mathbf{b}. \quad (3.1)$$

Here, we establish results on the norm- and field-of-values-equivalence between saddle point matrices (3.1) to constraint preconditioners of the form

$$\mathcal{P}_{\text{con}} = \begin{bmatrix} P & C^T \\ C & 0 \end{bmatrix}. \quad (3.2)$$

The focus here is on sequences of such matrices arising from the finite element discretization of p.d.e.s, where the dimension of these large, sparse matrices grows as the mesh width is decreased. It is assumed that each element

of this sequence of matrices satisfies the following stability conditions

$$\max_{\mathbf{w} \in \mathbb{R}^n \setminus \{0\}} \max_{\mathbf{v} \in \mathbb{R}^n \setminus \{0\}} \frac{\mathbf{w}^T \mathcal{A} \mathbf{v}}{\|\mathbf{w}\|_H \|\mathbf{v}\|_H} \leq c_1, \quad (3.3a)$$

$$\min_{\mathbf{w} \in \mathbb{R}^n \setminus \{0\}} \max_{\mathbf{v} \in \mathbb{R}^n \setminus \{0\}} \frac{\mathbf{w}^T \mathcal{A} \mathbf{v}}{\|\mathbf{w}\|_H \|\mathbf{v}\|_H} \geq c_2, \quad (3.3b)$$

where c_1, c_2 are positive constants that are independent of the size of the matrix (correspondingly, mesh width). These conditions, known as the Babuška-Brezzi conditions [3, 10, 11], are necessary for the existence of a unique finite element solution. Moreover, the above stability bounds are crucial for establishing results on the norm- and field-of-values-equivalence of constraint preconditioners (3.2) to saddle point matrices (3.1).

3.1 Norm- and field-of-values-equivalences

Recall the fundamental Definitions 1.1 and 1.2 of norm- and field-of-values-equivalence between two matrices M and N . A direct consequence of norm-equivalence, i.e., $M \sim_H N$ are the following bounds

$$\|MN^{-1}\|_H \leq \beta_0, \quad (3.4a)$$

$$\|NM^{-1}\|_H \leq \alpha_0^{-1}. \quad (3.4b)$$

Moreover, H -norm-equivalence is an equivalence relation since it satisfies the following three properties:

- reflexive , $M \sim_H M$,
- symmetric, $M \sim_H N \Rightarrow N \sim_H M$,
- transitive, $M \sim_H N, N \sim_H L \Rightarrow M \sim_H L$.

It is worth noting that f.o.v.-equivalence implies that the H -field of values $\mathcal{W}_H(MN^{-1}) \subset \mathbb{C}^+$. In addition, f.o.v.-equivalence is a stronger condition, as $M \approx_H N \Rightarrow M \sim_H N$. However, f.o.v.-equivalence is not an equivalence relation since it is not reflexive [39].

The utility of these equivalences comes from the following two theorems; see, e.g., [22, 29, 48].

Theorem 3.1. *If $M \sim_H N$ and MN^{-1} is symmetric and indefinite with respect to $(\cdot, \cdot)_H$, the MINRES algorithm applied to M with right preconditioner N converges in a number of iterations independent of the dimension and the residuals satisfy*

$$\frac{\|\mathbf{r}_k\|_H}{\|\mathbf{r}_0\|_H} \leq 2 \left(\frac{\beta_0 - \alpha_0}{\beta_0 + \alpha_0} \right)^{k/2}.$$

Theorem 3.2. *If $M \approx_H N$, the GMRES algorithm converges with respect to $(\cdot, \cdot)_H$ in a number of iteration independent of the dimension and the residuals satisfy*

$$\frac{\|\mathbf{r}_k\|_H}{\|\mathbf{r}_0\|_H} \leq \left(1 - \frac{\alpha_0^2}{\beta_0^2} \right)^{k/2}.$$

In [39] the authors establish conditions for norm- and field-of-values-equivalence between saddle point matrices of the form (3.1) to block diagonal and block lower triangular preconditioners of the form

$$\mathcal{P}_{\text{bd}} = \begin{bmatrix} P_1 & 0 \\ 0 & P_2 \end{bmatrix}, \quad \mathcal{P}_{\text{lt}}(\rho) = \begin{bmatrix} P_1 & 0 \\ C & \rho P_2 \end{bmatrix}. \quad (3.5)$$

The factor ρ is a scaling parameter for the (2,2)-block of the block lower triangular preconditioner and is required in establishing the f.o.v.-equivalence of \mathcal{P}_{lt} [39].

Here, we extend these results, and establish the optimality of constraint preconditioners. We do this by first proving the norm-equivalence, $\mathcal{P}_{\text{con}} \sim_H \mathcal{A}$ and then the stronger, f.o.v.-equivalence, $\mathcal{P}_{\text{con}} \approx_H \mathcal{A}$.

With the aid of the following definitions and lemmas (proved in [39]) we show that the constraint preconditioner (3.2) is H -norm equivalent (and subsequently H -f.o.v equivalent) to the operator \mathcal{A} in (3.1), where

$$H = \begin{bmatrix} H_1 & 0 \\ 0 & H_2 \end{bmatrix} \quad (3.6)$$

and H_1, H_2 are symmetric positive definite (s.p.d.) matrices.

Recall the following general definition of the matrix norm.

Definition 3.1. Let $M \in \mathbb{R}^{m \times n}$ and $H_1 \in \mathbb{R}^{n \times n}$, $H_2 \in \mathbb{R}^{m \times m}$ be two s.p.d. matrices, then

$$\|M\|_{H_1, H_2} = \max_{\mathbf{v} \in \mathbb{R}^n \setminus \{0\}} \frac{\|M\mathbf{v}\|_{H_2}}{\|\mathbf{v}\|_{H_1}}.$$

Note that if $H_1 = H_2 = H$ we recover the matrix norm defined in (1.2). Moreover, we have the following useful set of equalities

$$\|H_2^{-1/2} M H_1^{-1/2}\|_2 = \|M\|_{H_1, H_2^{-1}} = \|M H_1^{-1}\|_{H_1^{-1}, H_2^{-1}} = \|H_2^{-1} M\|_{H_1, H_2}. \quad (3.7)$$

Additionally, given a s.p.d. matrix H_3 , then the following matrix norm inequality holds

$$\|MN\|_{H_3, H_1} \leq \|N\|_{H_3, H_2} \|M\|_{H_2, H_1}. \quad (3.8)$$

The following lemmas are stated without proof, for complete details, see [39].

Lemma 3.1. Let (3.3) hold, then $H \sim_{H^{-1}} \mathcal{A}$ and $H^{-1} \sim_H \mathcal{A}^{-1}$, and in particular

$$\|H^{-1} \mathcal{A}\|_H = \|\mathcal{A} H^{-1}\|_{H^{-1}} \leq c_1, \quad (3.9a)$$

$$\|\mathcal{A}^{-1} H\|_H = \|H \mathcal{A}^{-1}\|_{H^{-1}} \leq c_2^{-1}. \quad (3.9b)$$

Lemma 3.2. Let (3.3) hold and assume $P \sim_{H^{-1}} H$, then

$$P \sim_{H^{-1}} \mathcal{A} \quad \text{and} \quad P^{-1} \sim_H \mathcal{A}^{-1}.$$

Lemma 3.3. Let (3.3) hold then $\|A\|_{H_1, H_1^{-1}} \leq c_1$, $\|C\|_{H_1, H_2^{-1}} \leq c_1$.

Lemma 3.4. Let (3.3) hold. If there exists $c_3 > 0$ independent of n such that

$$\min_{\mathbf{w} \in \mathbb{R}^n \setminus \{0\}} \max_{\mathbf{v} \in \mathbb{R}^n \setminus \{0\}} \frac{\mathbf{w}^T A \mathbf{v}}{\|\mathbf{w}\|_{H_1} \|\mathbf{v}\|_{H_1}} \geq c_3,$$

then $S = CA^{-1}C^T$, the negative Schur complement, satisfies $S \sim_{H_2^{-1}} H_2$ and $H_2^{-1} \sim_{H_2} S^{-1}$. Hence there exists positive c_4 independent of n such that $\|S^{-1}\|_{H_2^{-1}, H_2} \leq c_4$.

It is worth noting that if A is replaced by an approximation, \tilde{A} , then the previous lemma still holds for the corresponding approximate Schur complement, $\tilde{S} = C\tilde{A}^{-1}C^T$.

Lemma 3.5.

$$\|M\|_{H_1, H_2^{-1}} = \|M^T\|_{H_2, H_1^{-1}}.$$

The theorem and proof of the norm-equivalence of constraint preconditioners (3.2) to saddle point matrices (3.1) is given below.

Theorem 3.3. *Let \mathcal{P}_{con} be defined as in (3.2), H be as in (3.6) and let (3.3) and the hypotheses of Lemma 3.4 hold. If $P \sim_{H_1^{-1}} H_1$, then $\mathcal{P}_{con} \sim_{H^{-1}} \mathcal{A}$ and $\mathcal{P}_{con}^{-1} \sim_H \mathcal{A}^{-1}$.*

Proof. We prove that $\mathcal{P}_{con} \sim_{H^{-1}} \mathcal{A}$ since $\mathcal{P}_{con}^{-1} \sim_H \mathcal{A}^{-1}$ follows similarly. By Lemma 3.2 we need only show that $\mathcal{P}_{con} \sim_{H^{-1}} H$ because then by transitivity the result is proven. To prove the above equivalence we bound both $\|H^{-1/2}\mathcal{P}_{con}H^{-1/2}\|_2$ and $\|H^{1/2}\mathcal{P}_{con}^{-1}H^{1/2}\|_2$. By the assumption that $P \sim_{H_1^{-1}} H_1$ we know there exists β_1 such that $\|PH_1^{-1}\|_{H_1^{-1}} \leq \beta_1$. Now consider

$$H^{-1/2}\mathcal{P}_{con}H^{-1/2} = \begin{bmatrix} H_1^{-1/2}PH_1^{-1/2} & H_1^{-1/2}C^T H_2^{-1/2} \\ H_2^{-1/2}CH_1^{-1/2} & 0 \end{bmatrix}. \quad (3.10)$$

We can bound the 2-norm of the above matrix as follows

$$\begin{aligned} \|H^{-1/2}\mathcal{P}_{con}H^{-1/2}\|_2 &\leq \|H_1^{-1/2}PH_1^{-1/2}\|_2 + \|H_2^{-1/2}CH_1^{-1/2}\|_2 + \|H_1^{-1/2}C^T H_2^{-1/2}\|_2 \\ &= \|H_1^{-1}P\|_{H_1} + \|C\|_{H_1, H_2^{-1}} + \|C^T\|_{H_2, H_1^{-1}} \\ &\leq \beta_1 + c_1 + c_1. \end{aligned}$$

The first inequality is a result of expressing the block matrix as a sum of three matrices. The equality in the second line is obtained using (3.7). Lastly, the final bound comes from (3.3) and the fact that $\|H_1^{-1}P\|_{H_1} = \|PH_1^{-1}\|_{H_1^{-1}} \leq \beta_1$.

Using (2.2), the inverse of \mathcal{P}_{con} is

$$\mathcal{P}_{con}^{-1} = \begin{bmatrix} P^{-1} + P^{-1}C^T S^{-1}CP^{-1} & -P^{-1}C^T S^{-1} \\ -S^{-1}CP^{-1} & S^{-1} \end{bmatrix},$$

and therefore

$$H^{1/2}\mathcal{P}_{con}^{-1}H^{1/2} = \begin{bmatrix} H_1^{1/2}(P^{-1} + P^{-1}C^T S^{-1}CP^{-1})H_1^{1/2} & -H_1^{1/2}(P^{-1}C^T S^{-1})H_2^{1/2} \\ -H_2^{1/2}(S^{-1}CP^{-1})H_1^{1/2} & H_2^{1/2}S^{-1}H_2^{1/2} \end{bmatrix}.$$

Hence, $\|H^{1/2}\mathcal{P}_{con}^{-1}H^{1/2}\|_2$ can be bounded by bounding the following five terms

$$\begin{aligned} (I) &= \|H_1^{1/2}P^{-1}H_1^{1/2}\|_2, \\ (II) &= \|H_1^{1/2}P^{-1}C^T S^{-1}CP^{-1}H_1^{1/2}\|_2, \\ (III) &= \|H_1^{1/2}P^{-1}C^T S^{-1}H_2^{1/2}\|_2, \\ (IV) &= \|H_2^{1/2}S^{-1}CP^{-1}H_1^{1/2}\|_2, \\ (V) &= \|H_2^{1/2}S^{-1}H_2^{1/2}\|_2. \end{aligned}$$

Note that

$$(I) = \|P^{-1}H_1\|_{H_1} \leq \alpha_1^{-1}$$

and

$$(V) = \|S^{-1}\|_{H_2^{-1}, H_2} \leq c_4.$$

Further note that

$$\begin{aligned} (II) &= \|H_1^{1/2}P^{-1}H_1^{1/2}H_1^{-1/2}C^T H_2^{-1/2}H_2^{1/2}S^{-1}H_2^{1/2}H_2^{-1/2}CH_1^{-1/2}H_1^{1/2}P^{-1}H_1^{1/2}\|_2 \\ &\leq \|H_1^{1/2}P^{-1}H_1^{1/2}\|_2 \|H_1^{-1/2}C^T H_2^{-1/2}\|_2 \|H_2^{1/2}S^{-1}H_2^{1/2}\|_2 \|H_2^{-1/2}CH_1^{-1/2}\|_2 \|H_1^{1/2}P^{-1}H_1^{1/2}\|_2 \\ &\leq \alpha_1^{-1}c_1c_4c_1\alpha_1^{-1}. \end{aligned}$$

The terms (III) and (IV) are then bounded in a similar manner. \square

With the aid of the previous theorem we can show the following result on the H -f.o.v.-equivalence of the constraint preconditioner \mathcal{P}_{con} to \mathcal{A} .

Theorem 3.4. *Let $\mathcal{A}, \mathcal{P}_{con}$ be defined as in (3.1), (3.2), respectively. Furthermore let (3.3) and Lemma 3.4 hold. Let*

$$H = \begin{bmatrix} \rho H_1 & 0 \\ 0 & H_2 \end{bmatrix}.$$

If $A \approx_{H_1^{-1}} P$, then there exists $\rho_0 > 0$ such that $\mathcal{A} \approx_{H^{-1}} \mathcal{P}_{con}$ for all $\rho \geq \rho_0$ provided $\|AP^{-1} - I\|_{H_1^{-1}} \leq \rho_0^{-1}$.

Proof. Since $A \approx_{H_1^{-1}} P$ there exists α_0, β_0 such that for all $\mathbf{x} \neq 0$,

$$\alpha_0 \leq \frac{\langle AP^{-1}\mathbf{x}, \mathbf{x} \rangle_{H_1^{-1}}}{\langle \mathbf{x}, \mathbf{x} \rangle_{H_1^{-1}}} \quad \text{and} \quad \|AP^{-1}\|_{H_1^{-1}} \leq \beta_0.$$

Moreover, f.o.v.-equivalence implies norm-equivalence so we have that $A \sim_{H_1^{-1}} P$ and due to (3.3), $A \sim_{H_1^{-1}} H_1$. Thus, by transitivity of norm-equivalence, $P \sim_{H_1^{-1}} H_1$. By Theorem 3.3, $\|\mathcal{AP}_{con}^{-1}\|_{H^{-1}}$ is bounded from above. The only remaining piece is to show the existence of a lower bound for all $\mathbf{x} \neq 0$

$$\alpha \mathbf{x}^T H^{-1} \mathbf{x} \leq \mathbf{x}^T H^{-1} \mathcal{AP}_{con}^{-1} \mathbf{x}. \quad (3.11)$$

We have that

$$H^{-1} \mathcal{AP}_{con}^{-1} = \begin{bmatrix} \rho H_1^{-1}(AP^{-1} - AP^{-1}C^T S^{-1}CP^{-1} + C^T S^{-1}CP^{-1}) & \rho H_1^{-1}(AP^{-1} - I)C^T S^{-1} \\ 0 & H_2^{-1} \end{bmatrix}.$$

Therefore, when $\mathbf{x} = \begin{bmatrix} x_1^T, x_2^T \end{bmatrix}^T$, the product $\mathbf{x}^T H^{-1} \mathcal{AP}_{con}^{-1} \mathbf{x}$ is

$$\begin{aligned} & \rho(x_1^T H_1^{-1} AP^{-1} x_1 + x_1^T H_1^{-1} (I - AP^{-1}) C^T S^{-1} CP^{-1} x_1) \\ & + \rho x_1^T H_1^{-1} (AP^{-1} - I) C^T S^{-1} x_2 + x_2^T H_2^{-1} x_2. \end{aligned}$$

As $x_2^T H_2^{-1} x_2 = \|x_2\|_{H_2^{-1}}^2$, we establish the desired bound (3.11) by bounding the following three pieces from below

$$\begin{aligned} (I) &= x_1^T H_1^{-1} AP^{-1} x_1, \\ (II) &= x_1^T H_1^{-1} (I - AP^{-1}) C^T S^{-1} CP^{-1} x_1, \\ (III) &= x_1^T H_1^{-1} (AP^{-1} - I) C^T S^{-1} x_2. \end{aligned}$$

By the f.o.v.-equivalence $A \approx_{H^{-1}} P$, $(I) \geq \alpha_0 \|x_1\|_{H_1^{-1}}^2$. In order to bound piece (II) from below, first consider the bound on the following absolute value

$$|x_1^T H^{-1} (I - AP^{-1}) C^T S^{-1} CP^{-1} x_1| \leq \|I - AP^{-1}\|_{H_1^{-1}} \|C^T S^{-1} CP^{-1}\|_{H_1^{-1}} \|x_1\|_{H_1^{-1}}^2.$$

From (3.7) we have

$$\begin{aligned} & \|C^T S^{-1} CP^{-1}\|_{H_1^{-1}} \\ &= \|H_1^{-1/2} C^T S^{-1} CP^{-1} H_1^{1/2}\|_2 \\ &\leq \|H_1^{-1/2} C^T H_2^{-1/2}\|_2 \|H_2^{1/2} S^{-1} H_2^{1/2}\|_2 \|H_2^{-1/2} C H_1^{-1/2}\|_2 \|H_1^{1/2} P^{-1} H_1^{1/2}\|_2 \\ &\leq c_1^2 c_4 \alpha_1^{-1}. \end{aligned}$$

Choosing $\|I - AP^{-1}\|_{H_1^{-1}} \leq \rho^{-1}$, we obtain the desired mesh-independent lower bound for (II) by negating the absolute value. Lastly, for piece (III) we have that

$$\begin{aligned}
& |x_1^T H_1^{-1} (AP^{-1}) C^T S^{-1} x_2| \\
& \leq \|(AP^{-1} - I) C^T S^{-1}\|_{H_2^{-1}, H_1^{-1}} \|x_1\|_{H_1^{-1}} \|x_2\|_{H_2^{-1}} \\
& \leq \|C^T S^{-1}\|_{H_2^{-1}, H_1} \|AP^{-1} - I\|_{H_1^{-1}, H_1^{-1}} \|x_1\|_{H_1^{-1}} \|x_2\|_{H_2^{-1}} \\
& \leq \|S^{-1}\|_{H_2^{-1}, H_2} \|C^T\|_{H_2, H_1^{-1}} \|AP^{-1} - I\|_{H_1^{-1}} \|x_1\|_{H_1^{-1}} \|x_2\|_{H_2^{-1}} \\
& \leq c_1 c_4 \|AP^{-1} - I\|_{H_1^{-1}}.
\end{aligned}$$

And since $\|AP^{-1} - I\|_{H_1^{-1}} \leq \rho^{-1}$ we obtain a mesh-independent bound for (III). Combining these bounds together, we obtain a lower bound of the following form

$$(\rho a_1 + a_2) \|x_1\|_{H_1^{-1}}^2 - a_3 \|x_1\|_{H_1^{-1}} \|x_2\|_{H_2^{-1}} + \|x_2\|_{H_2^{-1}}^2, \quad (3.12)$$

where $a_1 = \alpha_0$, $a_2 = c_1^2 c_4 \alpha_1^{-1}$ and $a_3 = c_1 c_4$. By selecting the constant $\rho_0 = (1 + a_3^2 - 2a_2)/2a_1$, (3.12) simplifies to

$$\begin{aligned}
& \frac{1 + a_3^2}{2} \|x_1\|_{H_1^{-1}}^2 - a_3 \|x_1\|_{H_1^{-1}} \|x_2\|_{H_2^{-1}} + \frac{1}{2} \|x_2\|_{H_2^{-1}}^2 \\
& = \frac{1}{2} \left(\|x_1\|_{H_1^{-1}}^2 + \|x_2\|_{H_2^{-1}}^2 \right) + \frac{a_3^2}{2} \|x_1\|_{H_1^{-1}}^2 - a_3 \|x_1\|_{H_1^{-1}} \|x_2\|_{H_2^{-1}} + \frac{1}{2} \|x_2\|_{H_2^{-1}}^2, \\
& = \frac{1}{2} \left(\|x_1\|_{H_1^{-1}}^2 + \|x_2\|_{H_2^{-1}}^2 \right) + \left(\frac{a_3}{2} \|x_1\|_{H_1^{-1}} - \|x_2\|_{H_2^{-1}} \right)^2, \\
& \geq \frac{1}{2} \left(\|x_1\|_{H_1^{-1}}^2 + \|x_2\|_{H_2^{-1}}^2 \right).
\end{aligned}$$

This shows we have the desired lower bound of the form $\alpha \|\mathbf{x}\|_{H^{-1}}^2$ with $\alpha = 1/2$.

□

3.2 Inexact constraint preconditioning

The theory of the previous section established mesh-independent GMRES convergence for constraint preconditioned saddle point systems using direct methods, i.e., exact solves, for the block solves. For a sufficiently resolved

numerical solution, small mesh widths are required, and the corresponding linear systems grow in size. For these large-scale computations, particularly problems arising from the discretization of p.d.e.s in three-dimensions, exact solves become prohibitively expensive and are therefore impractical. This requires replacing exact solves with faster, more economical inexact methods that ideally maintain the previously established mesh independent convergence of GMRES. In order to ensure that this property is maintained, we replace the exact block solves with P by iterative methods preconditioned with spectrally equivalent multigrid operators; see Definition 2.1 in Chapter 2.4.

Thus, the situation under consideration is replacing P by some \tilde{P} , giving an inexact constraint preconditioner of the form

$$\tilde{\mathcal{P}}_{\text{con}} = \begin{bmatrix} \tilde{P} & C^T \\ C & 0 \end{bmatrix}. \quad (3.13)$$

Here, we consider an operator \tilde{P} that corresponds to the solution of A by a Krylov subspace method preconditioned with a spectrally equivalent multigrid preconditioner.

Note that by Definition 2.1, if \tilde{P} is spectrally equivalent to A , then there exist mesh-independent constants, α, β such that for all $\mathbf{x} \neq 0$

$$\alpha \leq \frac{(A\tilde{P}^{-1}\mathbf{x}, \mathbf{x})}{(\mathbf{x}, \mathbf{x})} \quad \text{and} \quad \|A\tilde{P}^{-1}\| \leq \beta. \quad (3.14)$$

The following theorem, the inexact analogue of Theorem 3.4, is shown for the inexact constraint preconditioner defined in (3.13).

Theorem 3.5. *Let \mathcal{A} , $\tilde{\mathcal{P}}_{\text{con}}$ be defined as in (3.1), (3.13), respectively. Additionally, let (3.3) and the hypotheses of Lemma 3.4 hold but with A replaced by \tilde{P} . Let*

$$H = \begin{bmatrix} \rho H_1 & 0 \\ 0 & H_2 \end{bmatrix}.$$

If \tilde{P} satisfies the bounds (3.14), then the results of Theorem 3.4 remain true with $P = \tilde{P}$.

Proof. Recall that in the proof of Theorem 3.4 we first established an upper bound on $\|\mathcal{A}\tilde{\mathcal{P}}_{con}^{-1}\|_{H^{-1}}$ by appealing to Theorem 3.3. This theorem can still be applied in the inexact case due to the upper bound, $\|A\tilde{P}\|_{H_1^{-1}} \leq \beta$, from (3.14).

The second part of the proof established the necessary lower bound for field-of-values-equivalence. This required bounding three different pieces from below. The line of argument in this proof is essentially the same with only a few differences. First, the existence of lower bounds for $x_1^T H_1^{-1} A \tilde{P}^{-1} x_1$ comes from (3.14). The next important difference concerns the bounds on the approximate Schur complement, $\tilde{S} = -C\tilde{P}^{-1}C^T$, which are taken care of by the approximate version of Lemma 3.4 that holds. Therefore, for the inexact solves we still have mesh-independent f.o.v.-equivalence. \square

The important consequences of this theory is that exact representations of the block inverses are not required in order to maintain the desired mesh-independent convergence properties of constraint preconditioned GMRES. In fact, all that is necessary is that a spectrally equivalent method for the matrix \tilde{P} be used in the preconditioner. This allows a great deal of flexibility in the types of methods that can be used as preconditioners while still maintaining mesh-independent convergence. For instance, \tilde{P} could be a multigrid method itself, or a call to an iterative method with an optimal multigrid preconditioner. In Chapter 4, inexact versions of the considered preconditioners for the solution of the coupled Stokes-Darcy system are presented.

CHAPTER 4

CONSTRAINT PRECONDITIONING OF THE COUPLED STOKES-DARCY SYSTEM

In this chapter, we present numerical results that illustrate the theory of Chapter 3. We consider the solution by preconditioned GMRES of the saddle point system that arises from the finite element discretization of the coupled Stokes-Darcy system. This is a set of p.d.e.s that models the coupling of a freely flowing fluid, governed by the Stokes equations, to a porous media flow, governed by the Darcy equations, via conditions across the interface of the two flow regions.

We present numerical experiments in both two and three dimensions. The considered 2D experiments are for tetrahedral elements and we consider two types of discretization schemes in the Darcy region, namely, standard continuous finite elements (conG), where the finite element basis functions are continuous, and discontinuous Galerkin (DG) methods, where the finite element basis functions are discontinuous across elements. We compare exact implementations of two versions of the constraint preconditioner against stan-

dard block diagonal and block lower triangular preconditioners.

The numerical experiments in 3D are implemented using the `deal.II` finite element library with hexahedral elements [4]. The numerical experiments have been performed using the high performance computing cluster, Owl's Nest, at Temple University [1]. The computations were done in serial and on highmem nodes where the processor is a 2x Intel Xeon X5677. We consider flow in several domains with differing geometries and varying the permeabilities in the Darcy region. Due to the size of the matrices considered, it is not practical to use exact versions of the considered preconditioners, thus, inexact versions of the preconditioner are introduced.

4.1 The coupled Stokes-Darcy system

Consider fluid flowing in a domain $\Omega = \Omega_1 \cup \Omega_2$; see Figure 4.1. For simplicity, in this thesis we consider the case where Ω is a rectangle in 2D and a rectangular prism in 3D, but most of what we say equally applies to other geometries. The flow in Ω_1 is modeled by the Stokes equations

$$-\nabla \cdot (2\nu D(\mathbf{u}_1) - p_1 \mathbf{I}) = \mathbf{f}_1, \quad \text{in } \Omega_1, \quad (4.1a)$$

$$\nabla \cdot \mathbf{u}_1 = 0, \quad \text{in } \Omega_1, \quad (4.1b)$$

$$\mathbf{u}_1 = 0, \quad \text{on } \Gamma_1 := \partial\Omega_1 \setminus \Gamma_{12}. \quad (4.1c)$$

The velocity and pressure in Ω_1 are denoted by \mathbf{u}_1 , p_1 , respectively. The coefficient $\nu > 0$ is the kinematic viscosity, the function \mathbf{f}_1 is an external force acting on the fluid, \mathbf{I} is the identity matrix, and $D(\mathbf{u}_1) = \frac{1}{2}(\nabla \mathbf{u}_1 + \nabla \mathbf{u}_1^T)$ is the rate of strain tensor which describes how the fluid deforms in response to stress. Lastly, Γ_1 is the boundary of the domain Ω_1 excluding the interface Γ_{12} .

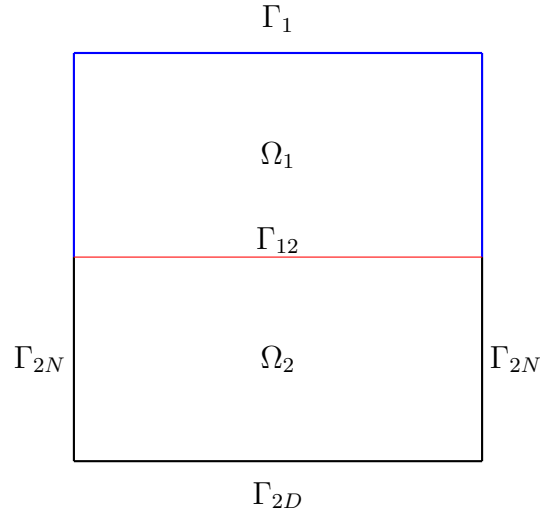


Figure 4.1: The domain $\Omega = \Omega_1 \cup \Omega_2$. The Stokes flow region is Ω_1 . The Darcy region is Ω_2 . The Stokes boundary (excluding the interface) Γ_1 is colored blue. The interface Γ_{12} is colored red. The Darcy boundary is composed of Dirichlet (Γ_{2D}) and Neumann (Γ_{2N}) parts.

The flow in Ω_2 is modeled by Darcy's Law

$$-\nabla \cdot \mathbf{K} \nabla p_2 = f_2, \quad \text{in } \Omega_2, \quad (4.2a)$$

$$-\mathbf{K} \nabla p_2 = \mathbf{u}_2, \quad \text{in } \Omega_2, \quad (4.2b)$$

$$p_2 = g_D, \quad \text{on } \Gamma_{2D}, \quad (4.2c)$$

$$\mathbf{K} \nabla p_2 \cdot \mathbf{n}_2 = g_N, \quad \text{on } \Gamma_{2N}. \quad (4.2d)$$

The velocity and pressure in Ω_2 are denoted \mathbf{u}_2 , p_2 , respectively and the function f_2 is an external force acting on the fluid. The functions g_D and g_N are prescribed on the portions of the boundary corresponding to the Dirichlet (Γ_{2D}) and Neumann (Γ_{2N}) boundary conditions, respectively, so that the Darcy boundary is $\Gamma_2 := \partial\Omega_2 \setminus \Gamma_{12} = \Gamma_{2D} \cup \Gamma_{2N}$. The vector \mathbf{n}_2 denotes the outward unit normal vector to Γ_{2N} . The symmetric positive definite matrix \mathbf{K} represents the hydraulic conductivity of the fluid and describes how well flow moves through a particular point of the porous media. For isotropic flow we have that the hydraulic conductivity matrix is a scaled identity matrix with

scaling factor κ , i.e., $\mathbf{K} = \kappa \mathbf{I}$.

Let $\mathbf{n}_{12}, \boldsymbol{\tau}_{12}$, denote the unit normal vector directed from Ω_1 to Ω_2 and unit tangential vector to the interface, respectively. The model is completed by specifying the following coupling (interface) conditions between the two domains

$$\mathbf{u}_1 \cdot \mathbf{n}_{12} = -\mathbf{K} \nabla p_2 \cdot \mathbf{n}_{12}, \quad (4.3a)$$

$$(-2\nu D(\mathbf{u}_1) \mathbf{n}_{12} + p_1 \mathbf{n}_{12}) \cdot \mathbf{n}_{12} = p_2, \quad (4.3b)$$

$$\mathbf{u}_1 \cdot \boldsymbol{\tau}_{12} = -2\nu G(D(\mathbf{u}_1) \mathbf{n}_{12}) \cdot \boldsymbol{\tau}_{12}, \quad (4.3c)$$

where (4.3a) ensures mass conservation across the interface, (4.3b) ensures the balance of normal forces across the interface, and (4.3c) is the Beavers-Joseph-Saffman (BJS) law. The BJS law, which was determined experimentally, states that the tangential velocity of the fluid is proportional to the shear (tangential) stress. The proportionality constant G , called the BJS constant, is experimentally determined and depends on the material properties of the porous medium; see [5, 50]. The third condition is essential to complete the model. This is due to the Stokes equations having second order derivatives for the velocity variable and first order derivatives for the pressure variable, while the opposite is true in the Darcy domain, that is, the velocity contains first order derivatives and the pressure contains second order derivatives; see [18].

4.2 Finite element solution of the coupled Stokes-Darcy system

We consider the solution of the fully coupled set of p.d.e.s by the finite element method; recall Chapter 2.3. Thus, the original set of p.d.e.s given by (4.1) and (4.2) must be transformed into its weak, or variational form. This is done by multiplying the p.d.e.s by appropriate test functions, integrating over the domain, using Green's formulas [25, p. 628] (integration by parts), and incorporating the boundary and interface conditions, which then defines

the corresponding bilinear forms for the weak problem. In this case, the choice of the function spaces where the velocity and pressures are found is now very important.

To introduce the function spaces where the weak solutions of the coupled Stokes-Darcy problem are to be found, let

$$X_1 = \{\mathbf{v}_1 \in (H^1(\Omega_1))^2 : \mathbf{v}_1 = 0 \text{ on } \Gamma_1\}, \quad Q_1 = L^2(\Omega_1),$$

be the Stokes velocity and pressure spaces, respectively, and let

$$Q_2 = \{q_2 \in H^1(\Omega_2) : q_2 = 0 \text{ on } \Gamma_{2D}\},$$

be the Darcy pressure space. Let $X = X_1 \times Q_2$ and $Q = Q_1$. Associated to these spaces are the corresponding norms (see, e.g., [13, 18])

$$\|(\mathbf{u}_1, p_2)\|_X = \left(\kappa \|p_2\|_{H^1(\Omega_2)}^2 + 2\nu \|D(\mathbf{u}_1)\|_{L^2(\Omega_1)}^2 \right)^{1/2}, \quad (4.4a)$$

$$\|p_1\|_Q = \|p_1\|_{L^2(\Omega_1)}. \quad (4.4b)$$

The weak formulation of the coupled Stokes-Darcy problem (4.1), (4.2), (4.3), is to find $\mathbf{u}_1 \in X_1$, $p_1 \in Q_1$, and $p_2 \in Q_2$ such that

$$a((\mathbf{u}_1, p_2), (\mathbf{v}_1, q_2)) + b^*(p_1, (\mathbf{v}_1, q_2)) = \mathbf{f}(\mathbf{v}_1, q_2) \quad \forall \mathbf{v}_1 \in X_1, \forall q_2 \in Q_2, \quad (4.5a)$$

$$b((\mathbf{u}_1, p_2), q_1) = 0 \quad \forall q_1 \in Q_1, \quad (4.5b)$$

where

$$a((\mathbf{u}_1, p_2), (\mathbf{v}_1, q_2)) = a_{\Omega_1}(\mathbf{u}_1, \mathbf{v}_1) + a_{\Omega_2}(p_2, q_2) + a_{\Gamma_{12}}((\mathbf{u}_1, p_2), (\mathbf{v}_1, q_2)), \quad (4.6)$$

$$b((\mathbf{u}_1, p_2), q_1) = - \int_{\Omega_1} (\nabla \cdot \mathbf{u}_1) q_1 \, dx, \quad (4.7)$$

and

$$\begin{aligned} a_{\Omega_1}(\mathbf{u}_1, \mathbf{v}_1) &= 2\nu \int_{\Omega_1} D(\mathbf{u}_1) : D(\mathbf{v}_1) + \frac{1}{G} \int_{\Gamma_{12}} (\mathbf{u}_1 \cdot \boldsymbol{\tau}_{12})(\mathbf{v}_1 \cdot \boldsymbol{\tau}_{12}), \\ a_{\Omega_2}(p_2, q_2) &= \int_{\Omega_2} \mathbf{K} \nabla p_2 \cdot \nabla q_2, \\ a_{\Gamma_{12}}((\mathbf{u}_1, p_2), (\mathbf{v}_1, q_2)) &= \int_{\Gamma_{12}} (p_2 \mathbf{v}_1 - q_2 \mathbf{u}_1) \cdot \mathbf{n}_{12}. \end{aligned}$$

Lastly,

$$\mathbf{f}(\mathbf{v}_1, q_2) = \int_{\Omega_1} \mathbf{f}_1 \cdot \mathbf{v}_1 + \int_{\Omega_2} f_2 q_2 + \int_{\Gamma_{2N}} g_N q_2. \quad (4.9)$$

The following lemma, proved in [13, 18], is used to establish the well-posedness of the weak problem (4.5) for continuous finite elements.

Lemma 4.1. *Let $a(\cdot, \cdot)$ be defined as in (4.6) and $b(\cdot, \cdot)$ as in (4.7) with function spaces X and Q , then the following hold.*

1. $a(\cdot, \cdot)$ is continuous on $X \times X$, that is, there exists a positive constant c_1 such that

$$|a((\mathbf{u}_1, p_2), (\mathbf{v}_1, q_2))| \leq c_1 \|(\mathbf{u}_1, p_2)\|_X \|(\mathbf{v}_1, q_2)\|_X,$$

2. $a(\cdot, \cdot)$ is coercive on X , that is, there exists a positive constant c_2 such that

$$a((\mathbf{u}_1, p_2), (\mathbf{u}_1, p_2)) \geq c_2 \|(\mathbf{u}_1, p_2)\|_X^2,$$

3. $b(\cdot, \cdot)$ is continuous on $X \times Q$ and satisfies the inf-sup condition, that is, there exists a positive constant β such that $\forall q_1 \in Q, \exists (\mathbf{u}_1, q_2) \in X, (\mathbf{u}_1, q_2) \neq 0$, such that

$$b((\mathbf{u}_1, q_2), q_1) \geq \beta \|(\mathbf{u}_1, q_2)\|_X \|q_1\|_Q.$$

The existence and uniqueness of the weak problem (4.5) then follows from Lemma 4.1 and the theory of Brezzi-Fortin [10, 11].

For proofs that the bilinear forms and linear functionals in (4.5) satisfy the necessary and sufficient conditions to apply the Brezzi-Fortin theory when the Darcy-domain is discretized with a DG method; see [45].

To find the discrete solution, the next step is to select appropriate finite element spaces, that is, choose $X_1^h \subset X_1, Q_1^h \subset Q_1$ that satisfy discrete versions of the above continuity, coercivity, and inf-sup conditions for the Stokes velocity and pressure in addition to selecting $Q_2^h \subset Q_2$ for the finite Darcy pressure space.

For example one could choose MINI finite element spaces [2, 15] or Taylor-Hood elements [23, 56] for X_1^h, Q_1^h . The discrete pressure space for the Darcy domain Q_2^h consists of piecewise continuous polynomials.

The derivation of this weak form of the coupled Stokes-Darcy problem has been done here only for continuous finite element spaces. Should the Darcy domain be discretized using a DG method, then the Darcy pressure space changes, as well as the bilinear form a_{Ω_2} ; see [15] for further details.

The linear system is derived as follows. The discrete velocity and pressure solutions are assumed to be a linear combination of the basis functions for the finite element spaces, X_1^h, Q_1^h , and Q_2^h , i.e.,

$$\mathbf{u}_{\Omega_1}^h(\mathbf{x}) = \sum_{i=1}^{n_u} u_i \phi_i(\mathbf{x}), \quad (4.10)$$

$$p_{\Omega_1}^h(\mathbf{x}) = \sum_{i=1}^{n_{p_1}} p_{1,i} \psi_{1,i}(\mathbf{x}), \quad (4.11)$$

$$p_{\Omega_2}^h(\mathbf{x}) = \sum_{i=1}^{n_{p_2}} p_{2,i} \psi_{2,i}(\mathbf{x}), \quad (4.12)$$

where n_u, n_{p_1} , and n_{p_2} , denote the number of Stokes velocity d.o.f.s, Stokes pressure d.o.f.s, and Darcy pressure d.o.f.s, respectively.

The linear system for the above velocity and pressure coefficients is obtained by substituting the above functions into the corresponding bilinear forms of problem (4.5) and testing against each of the basis functions. If we enumerate the Darcy pressure degrees of freedom first, the Stokes velocity degrees of freedom second, and the Stokes pressure degrees of freedom third, then the fully coupled, discrete version of (4.5) is the following linear system of equations

$$\mathcal{A}\mathbf{x} = \begin{bmatrix} A_{\Omega_2} & A_{\Gamma_{12}}^T & 0 \\ -A_{\Gamma_{12}} & A_{\Omega_1} & B^T \\ 0 & B & 0 \end{bmatrix} \begin{bmatrix} p_{\Omega_2} \\ \mathbf{u}_{\Omega_1} \\ p_{\Omega_1} \end{bmatrix} = \begin{bmatrix} f_{2,h} \\ \mathbf{f}_{1,h} \\ g_h \end{bmatrix} = b. \quad (4.13)$$

Here, A_{Ω_2} is the discrete version of $a_{\Omega_2}(\cdot, \cdot)$, A_{Ω_1} the discrete version of $a_{\Omega_1}(\cdot, \cdot)$, B is the discrete version of $b(\cdot, \cdot)$, and $A_{\Gamma_{12}}$ is the discrete version of $a_{\Gamma_{12}}(\cdot, \cdot)$.

Setting

$$A = \begin{bmatrix} A_{\Omega_2} & A_{\Gamma_{12}}^T \\ -A_{\Gamma_{12}} & A_{\Omega_1} \end{bmatrix}, \quad C = \begin{bmatrix} 0 & B \end{bmatrix},$$

and recalling (3.1), it is easy to see that the system of equations (4.13) is of saddle point form

$$\mathcal{A} = \begin{bmatrix} A & C^T \\ C & 0 \end{bmatrix}.$$

We remark that if the continuous Galerkin method is used to solve for the flow in both the Stokes and Darcy regions, then the matrices A_{Ω_1} , A_{Ω_2} are both symmetric and positive definite. However, due to the interface block $A_{\Gamma_{12}}$, the (1,1) block A of the saddle point matrix \mathcal{A} is non-symmetric. It is possible though to scale the second row of the block matrix \mathcal{A} in (4.13) by -1 to obtain a symmetric but still indefinite matrix \mathcal{A} . Additionally, if one models the flow in the Darcy region with a non-symmetric discontinuous Galerkin method, as proposed in [15], the matrix A_{Ω_2} is not symmetric and the fully coupled system can not be made symmetric.

4.3 Preconditioning the coupled Stokes-Darcy system

To determine the discrete Stokes velocity, Stokes pressure, and Darcy pressure we solve the large, sparse, and non-symmetric saddle point matrix \mathcal{A} with preconditioned GMRES [49]. We consider the following preconditioners

$$\begin{aligned} \mathcal{P}_+ &= \begin{bmatrix} A_{\Omega_2} & 0 & 0 \\ 0 & A_{\Omega_1} & 0 \\ 0 & 0 & M_p \end{bmatrix}, & \mathcal{P}_{T_1}(\rho) &= \begin{bmatrix} A_{\Omega_2} & 0 & 0 \\ 0 & A_{\Omega_1} & 0 \\ 0 & B & -\rho M_p \end{bmatrix}, \\ \mathcal{P}_{T_2}(\rho) &= \begin{bmatrix} A_{\Omega_2} & 0 & 0 \\ -A_{\Gamma_{12}} & A_{\Omega_1} & 0 \\ 0 & B & -\rho M_p \end{bmatrix}, & \mathcal{P}_C(\rho) &= \begin{bmatrix} A_{\Omega_2} & A_{\Gamma_{12}}^T & 0 \\ -A_{\Gamma_{12}} & A_{\Omega_1} & 0 \\ 0 & B & -\rho M_p \end{bmatrix}. \end{aligned}$$

These preconditioners were considered in [13] and are standard block diagonal and block triangular preconditioners. The matrix M_p in the $(3, 3)$ -block of each of the above preconditioners is the pressure mass matrix. This corresponds to a mass matrix for the Stokes pressure space, i.e., a discretization of the bilinear form $(\psi_{1,j}, \psi_{1,i})_{\Omega_1}$. The reason for this matrix to be included in each of the preconditioners is that the pressure mass matrix is spectrally equivalent to the Schur complement $S = BA_{\Omega_1}^{-1}B^T$ of the Stokes sub-matrix; see, e.g., [13, 23]. Moreover, utilizing the theory established in [39], Cai, Mu, and Xu [13], showed norm-equivalence between the above preconditioners and the system matrix \mathcal{A} . Thus, the preconditioned operator $\mathcal{P}^{-1}\mathcal{A}$, has a spectrum that is bounded independently of the mesh width of the finite element discretization.

However, constraint (indefinite) preconditioners of the form (3.2) (see, e.g., [36, 44, 51]) were not addressed in [13]. The following two indefinite preconditioners are proposed for the solution of the coupled Stokes-Darcy system.

$$\mathcal{P}_{con_D} = \begin{bmatrix} A_{\Omega_2} & 0 & 0 \\ 0 & A_{\Omega_1} & B^T \\ 0 & B & 0 \end{bmatrix}, \mathcal{P}_{con_T} = \begin{bmatrix} A_{\Omega_2} & 0 & 0 \\ -A_{\Gamma_{12}} & A_{\Omega_1} & B^T \\ 0 & B & 0 \end{bmatrix}.$$

Appealing to Theorems 3.3 and 3.4, which respectively establish norm- and f.o.v.-equivalence for constraint preconditioners, the above two constraint preconditioners are shown to be f.o.v.-equivalent to the system matrix \mathcal{A} .

By construction, the fully coupled operator \mathcal{A} satisfies the the Babuška-Brezzi conditions (3.3a)-(3.3b). Moreover, for the exact versions of the constraint preconditioner, the matrix P in the $(1, 1)$ -block is both norm- and f.o.v.-equivalent to the matrix A . In addition, the operator A satisfies the conditions of Lemma 3.4. Therefore, we can apply Theorem 3.4 and conclude that the constraint preconditioner \mathcal{P}_{con} is both norm- and f.o.v.-equivalent to the operator \mathcal{A} .

In order to visually illustrate Theorem 3.4 we computed the field-of-values for two small-scale matrices, corresponding to mesh widths $h = 2^{-3}, 2^{-4}$, using

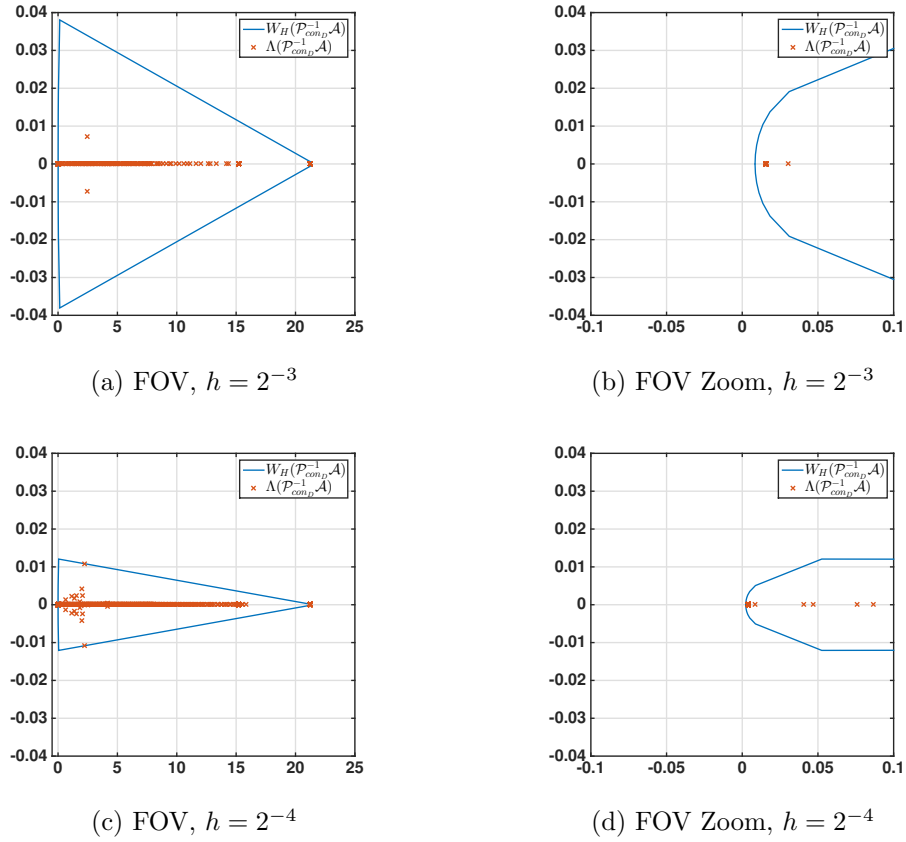


Figure 4.2: Plots of the preconditioned H -field-of-values, $\mathcal{W}_H(\mathcal{AP}_{con}^{-1})$ for two mesh widths $h = 2^{-3}, 2^{-4}$. The second figure is magnified near the origin to show that the field-of-values is contained in \mathbb{C}^+ . The dimension of the coupled Stokes-Darcy matrices here are $n = 521$ and $n = 2065$.

the `fv` function in the matrix computation toolbox package [33]. The matrix used for this computation was the discrete coupled Stokes-Darcy operator from problem (4.15). The results are shown in Figure 4.2. Observe that the H -field-of-values, where H is the discretized norm corresponding to (4.4), are contained in the right hand side of the complex plane, i.e., $\mathcal{W}_H(\mathcal{AP}_{con}^{-1}) \subset \mathbb{C}^+$. Moreover, as the mesh is refined, the f.o.v. stays bounded.

In Figure 4.3, the computed spectra of $\Lambda(\mathcal{A})$, $\Lambda(\mathcal{AP}_{con_D}^{-1})$, and $\Lambda(\mathcal{AP}_{con_T}^{-1})$ are displayed (as blue circles, red stars, and black squares respectively).

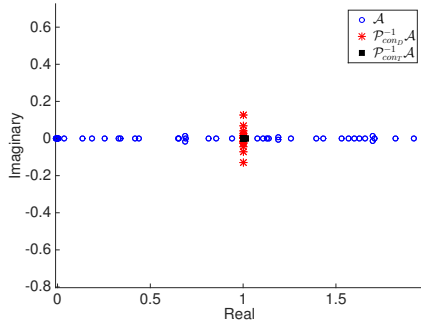
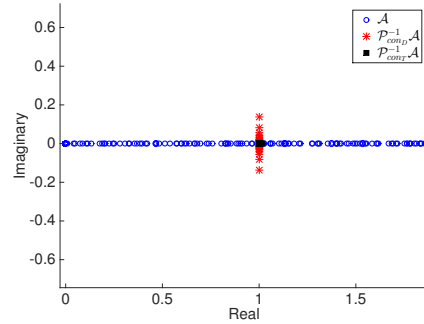
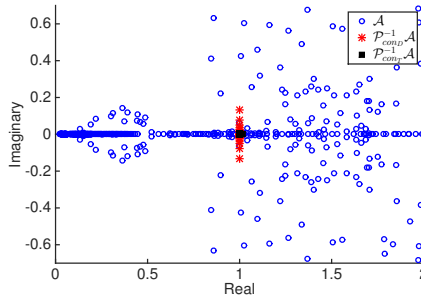
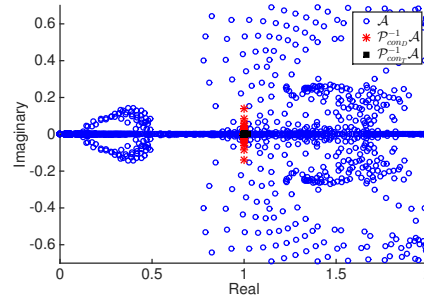
(a) Darcy-conG Spectrum, $h = 2^{-3}$ (b) Darcy-conG Spectrum, $h = 2^{-4}$ (c) Darcy-DG Spectrum, $h = 2^{-3}$ (d) Darcy-DG Spectrum, $h = 2^{-4}$

Figure 4.3: Spectra of $\Lambda(\mathcal{AP}_{con_i}^{-1})$ for $i = D, T$. The blue circles are the eigenvalues of \mathcal{A} , the red stars are the eigenvalues of $\mathcal{AP}_{con_D}^{-1}$, and the black squares are the eigenvalues of $\mathcal{AP}_{con_T}^{-1}$. The dimensions of the Darcy-conG and Darcy-DG matrices are respectively, $n = 521$ and 2065 , and $n = 1217$ and 4865 .

Figures 4.3a and 4.3b correspond to the case when the Darcy domain is discretized with continuous finite element spaces. Figures 4.3c and 4.3d display the same information but correspond to the case when the Darcy domain is discretized using a non-symmetric DG method. The underlying problem these discrete operators represent is from Problem (4.15). The mesh widths considered for both the Darcy-conG and Darcy-DG problems are $h = 2^{-3}, 2^{-4}$. Observe that in both cases the spectra are bounded as the mesh is refined.

4.4 Numerical Results

4.4.1 2D Test Problem: trigonometric solution

Here, we consider a simple test case for a coupled Stokes-Darcy system with the following velocity and pressure solution

$$\begin{cases} \mathbf{u}_1(x, y) &= \left[-\cos\left(\frac{\pi}{2}y\right)\sin\left(\frac{\pi}{2}x\right) + 1.0, \sin\left(\frac{\pi}{2}y\right)\cos\left(\frac{\pi}{2}x\right) - 1.0 + x \right]^T, \\ p_1(x, y) &= 1 - x, \\ p_2(x, y) &= \frac{2}{\pi}\cos\left(\frac{\pi}{2}x\right)\cos\left(\frac{\pi}{2}y\right) - y(x - 1). \end{cases} \quad (4.14)$$

The boundary conditions (see Figure 4.1) are chosen to be compatible with the above solution. This problem is solved on a triangular mesh. For the Stokes velocity and pressure basis functions we use MINI finite elements and piecewise linear elements, respectively [2]. For the Darcy pressure basis functions we consider piecewise linear elements for the continuous Galerkin case and discontinuous polynomials of degree two for the discontinuous Galerkin case.

We solve the discrete linear system using preconditioned GMRES with the considered preconditioners. The stopping criterion for the algorithm is when the relative residual, $\|\mathbf{r}_k\|_2/\|\mathbf{r}_0\|_2 < 10^{-8}$. A plot of the exact solution and the computed solution where the Darcy domain is discretized by a second order DG method and using the preconditioner \mathcal{P}_{con_D} is given in Figure 4.4.

Figure 4.5 contains a semilog plot of the decrease in the norm of the relative residual against the number of iterations. The iteration counts and CPU time in seconds for increasingly refined meshes are given in Table 4.1, where the preconditioners \mathcal{P}_- and \mathcal{P}_{T_1} are left out since the number of iterations for convergence and CPU timings are nearly identical to \mathcal{P}_+ and \mathcal{P}_{T_2} , respectively. Observe that the number of iterations to converge for both versions of the constraint preconditioner does not increase as the mesh is refined, as expected from Theorem 3.4. Moreover, it takes under ten iterations, regardless of the size of the system matrix, for GMRES to converge for both versions of

the constraint preconditioner. Additionally, the CPU times of both constraint preconditioners are more competitive than the other considered preconditioners.

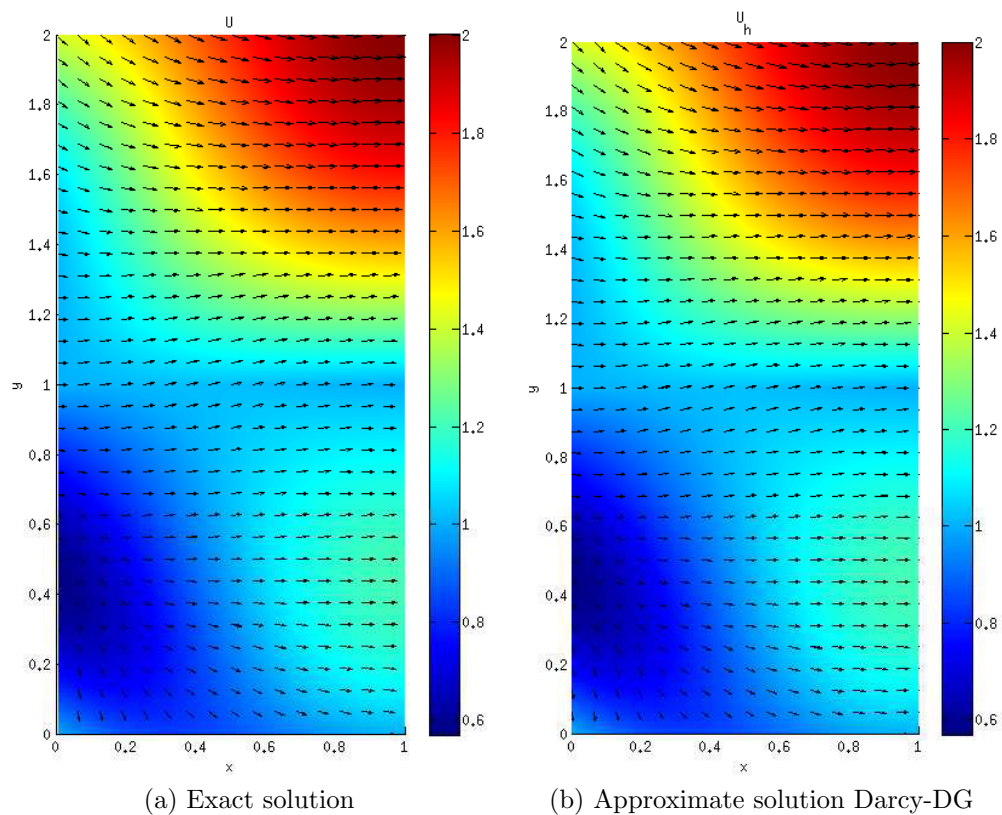


Figure 4.4: Exact and approximate solution computed with \mathcal{P}_{con_D} for Problem (4.14) with Darcy-DG and $h = 2^{-4}$.

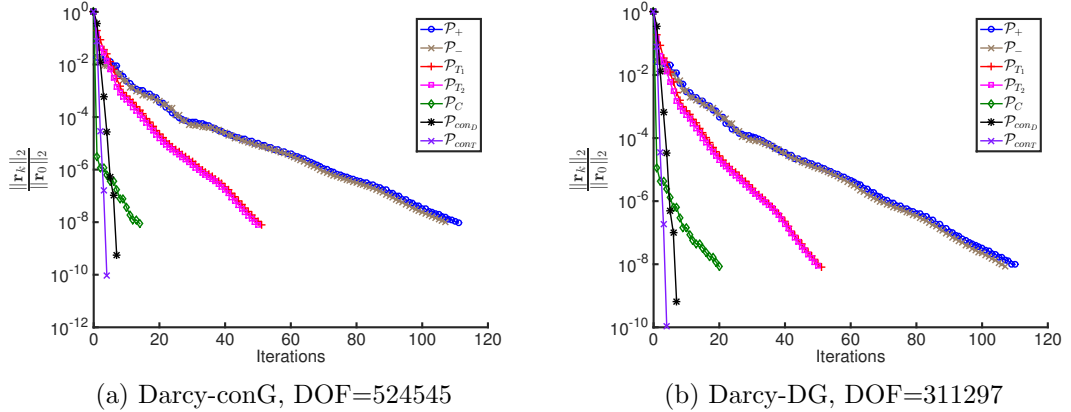


Figure 4.5: Residual convergence of the preconditioned GMRES algorithm for Problem (4.14) with a mesh discretization of $h = 2^{-8}$ and $h = 2^{-7}$, respectively.

Table 4.1: Number of iterations and CPU times for convergence of Problem (4.14) with $\kappa = 1$, $\nu = 1$.

(a) Darcy -conG						
h	DOF	\mathcal{P}_+	$\mathcal{P}_{T_2}(0.6)$	$\mathcal{P}_C(0.6)$	\mathcal{P}_{con_D}	\mathcal{P}_{con_T}
2^{-3}	521	74 (0.12)	41 (0.04)	33 (0.03)	7 (0.01)	4 (0.01)
2^{-4}	2065	87 (0.29)	45 (0.12)	33 (0.09)	7 (0.03)	4 (0.02)
2^{-5}	8225	95 (1.24)	45 (0.51)	32 (0.36)	7 (0.14)	4 (0.16)
2^{-6}	32833	105 (6.62)	48 (2.98)	26 (1.69)	7 (0.77)	4 (1.01)
2^{-7}	131201	110 (41.4)	50 (18.3)	20 (7.80)	7 (4.25)	4 (6.36)
2^{-8}	524545	111 (210)	50 (89.6)	14 (25.9)	7 (22.5)	4 (42.6)

(b) Darcy-DG						
h	DOF	\mathcal{P}_+	$\mathcal{P}_{T_2}(0.6)$	$\mathcal{P}_C(0.6)$	\mathcal{P}_{con_D}	\mathcal{P}_{con_T}
2^{-3}	1217	74 (0.21)	41 (0.09)	33 (0.07)	7 (0.02)	4 (0.01)
2^{-4}	4865	88 (1.03)	45 (0.44)	34 (0.33)	7 (0.08)	4 (0.05)
2^{-5}	19457	95 (5.94)	45 (2.79)	31 (1.98)	7 (0.51)	4 (0.33)
2^{-6}	77825	105 (35.3)	48 (14.4)	24 (7.57)	7 (2.72)	4 (1.65)
2^{-7}	311297	110 (188)	50 (79.0)	20 (33.2)	7 (14.0)	4 (7.94)

4.4.2 2D Test Problem: robustness with respect to physical parameters

Here, we consider a coupled Stokes-Darcy problem with the following polynomial velocity and pressure solution

$$\begin{cases} \mathbf{u}_1(x, y) &= [y^2 - 2y + 1 + \nu(2x - 1), x^2 - x - (y - 1)2\nu]^T, \\ p_1(x, y) &= 2\nu(x + y - 1) + \frac{1}{3\kappa} - 4\nu^2, \\ p_2(x, y) &= \frac{1}{\kappa}(x(1 - x)(y - 1) + \frac{y^3}{3} - y^2 + y) + 2\nu x. \end{cases} \quad (4.15)$$

In this case, $\mathbf{K} = \kappa\mathbf{I}$. The boundary conditions (see Figure 4.1) are again chosen to be compatible with the above solution. We solve this problem on a triangular mesh using MINI finite elements and piecewise linear elements for the Stokes velocity and pressure basis functions, respectively. Again, two types of Darcy pressure basis functions are considered, piecewise linear elements for the conG case and discontinuous polynomials of degree two for the DG case. Here, we independently vary both the permeability of the porous media κ and the viscosity ν over a range of progressively lower values. This is to illustrate the robustness of the two constraint preconditioners with respect to these physical parameters.

$\kappa = 1$ and $\nu = 1$

We first consider the case when $\kappa = 1$ and $\nu = 1$. Figure 4.6 is a semilog plot of the reduction of the relative residual against the number of iterations. Table 4.2 contains the iteration counts and CPU times for this problem. Observe that regardless of the type of discretization scheme used, the constraint preconditioners outperform the other proposed standard block diagonal and block lower triangular preconditioners with respect to both iteration count and CPU timings. Additionally, the iteration counts for the constraint preconditioners are bounded as the mesh is refined, further illustrating Theorem 3.4.

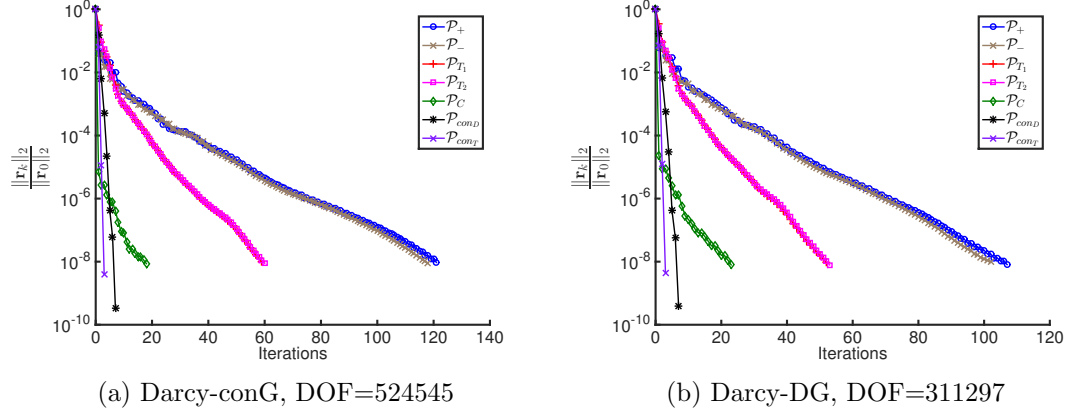


Figure 4.6: Residual convergence of the preconditioned GMRES algorithm for Problem (4.15) with $\kappa = 1$ and $\nu = 1$. The corresponding mesh widths are $h = 2^{-8}$ and $h = 2^{-7}$, respectively.

h	DOF	\mathcal{P}_+	$\mathcal{P}_{T_2}(0.6)$	$\mathcal{P}_C(0.6)$	\mathcal{P}_{con_D}	\mathcal{P}_{con_T}
2^{-3}	521	80 (0.12)	46 (0.04)	37 (0.03)	7 (0.01)	4 (0.01)
2^{-4}	2065	89 (0.27)	53 (0.13)	39 (0.09)	7 (0.02)	3 (0.02)
2^{-5}	8225	104 (1.26)	54 (0.64)	36 (0.42)	7 (0.16)	3 (0.12)
2^{-6}	32833	113 (7.81)	57 (3.70)	31 (2.02)	7 (0.93)	3 (0.71)
2^{-7}	131201	119 (40.3)	61 (19.9)	26 (8.70)	7 (4.63)	3 (4.04)
2^{-8}	524545	121 (288)	60 (126)	18 (36.2)	7 (28.2)	3 (26.0)

(a) Darcy -conG

h	DOF	\mathcal{P}_+	$\mathcal{P}_{T_2}(0.6)$	$\mathcal{P}_C(0.6)$	\mathcal{P}_{con_D}	\mathcal{P}_{con_T}
2^{-3}	1217	72 (0.21)	41 (0.09)	33 (0.07)	7 (0.02)	4 (0.01)
2^{-4}	4865	86 (1.08)	46 (0.51)	35 (0.39)	7 (0.09)	3 (0.04)
2^{-5}	19457	93 (5.96)	47 (2.82)	32 (2.00)	7 (0.49)	3 (0.25)
2^{-6}	77825	101 (35.2)	50 (19.4)	28 (11.2)	7 (3.48)	3 (1.66)
2^{-7}	311297	107 (221)	53 (99.5)	23 (43.7)	7 (16.3)	3 (7.41)

(b) Darcy-DG

Table 4.2: Number of iterations and CPU times for convergence of Problem (4.15) with $\kappa = 1$, $\nu = 1$.

Varying the permeability

Consider the solution by preconditioned GMRES of the matrix arising from the finite element discretization of the Stokes-Darcy system with the solution given by Problem (4.15). We again compare the number of iterations and CPU timings for each of the proposed preconditioners as the permeability κ is decreased. Table 4.3 contains the convergence information for the Darcy-conG case and Table 4.4 contains the convergence information for the Darcy-DG case. Observe that the number of iterations for each preconditioner increases as κ is decreased. Thus, decreasing κ can be viewed in some sense as making the problem more difficult for the considered preconditioners. Further observe that in terms of iteration counts the two considered constraint preconditioners are always superior. The constraint preconditioners are also superior in terms of CPU times for the Darcy-DG case. In the Darcy-conG case, the fully coupled block lower triangular preconditioner \mathcal{P}_C provides the best times for low values of κ , however, the constraint preconditioners are still close competitors.

Varying the viscosity

Again, we consider the solution of Problem (4.15) but in this section we compare the effect that varying the viscosity ν has on the number of iterations and CPU timings for each of the proposed preconditioners. Table 4.5 contains the convergence information for the Darcy-conG case and Table 4.6 contains the convergence information for the Darcy-DG case.

Observe again that as ν is decreased, the number of iterations to converge for each of the considered preconditioners increases. Decreasing ν also corresponds to making the problem more challenging for the iterative solvers. The two constraint preconditioners are again superior in terms of iteration counts. The CPU timings are also superior for the two constraint preconditioners as ν is decreased for the Darcy-DG problem. The CPU timings for the constraint preconditioners Darcy-conG case are superior for $\nu = 10^{-1}, 10^{-2}$. For the smaller values $\nu = 10^{-3}, 10^{-4}$ the CPU timings for \mathcal{P}_C are better.

Table 4.3: Number of iterations and CPU times for convergence of Problem (4.15) with Darcy-conG, varying κ , and fixed $\nu = 1$.

(a) $\kappa = 10^{-1}$

h	DOF	\mathcal{P}_+	$\mathcal{P}_{T_2}(0.6)$	$\mathcal{P}_C(0.6)$	\mathcal{P}_{con_D}	\mathcal{P}_{con_T}
2^{-3}	521	82 (0.12)	46 (0.04)	37 (0.03)	9 (0.01)	5 (0.01)
2^{-4}	2065	102 (0.33)	54 (0.13)	38 (0.09)	9 (0.03)	5 (0.02)
2^{-5}	8225	112 (1.50)	56 (0.65)	35 (0.40)	9 (0.18)	5 (0.16)
2^{-6}	32833	122 (8.32)	59 (3.70)	28 (1.73)	9 (1.08)	5 (0.98)
2^{-7}	131201	127 (45.5)	63 (21.9)	23 (8.07)	9 (5.99)	5 (6.35)
2^{-8}	524545	128 (358)	63 (131)	15 (30.4)	9 (35.1)	5 (37.4)

(b) $\kappa = 10^{-2}$

h	DOF	\mathcal{P}_+	$\mathcal{P}_{T_2}(0.6)$	$\mathcal{P}_C(0.6)$	\mathcal{P}_{con_D}	\mathcal{P}_{con_T}
2^{-3}	521	91 (0.15)	48 (0.04)	41 (0.04)	13 (0.01)	7 (0.01)
2^{-4}	2065	121 (0.41)	60 (0.15)	46 (0.11)	14 (0.05)	7 (0.03)
2^{-5}	8225	134 (1.89)	62 (0.76)	43 (0.50)	15 (0.29)	7 (0.21)
2^{-6}	32833	145 (10.5)	65 (3.92)	38 (2.19)	15 (1.64)	7 (1.24)
2^{-7}	131201	157 (55.2)	69 (21.0)	31 (9.26)	15 (8.71)	7 (7.96)
2^{-8}	524545	142 (306)	63 (119)	23 (41.73)	15 (45.4)	7 (65.0)

(c) $\kappa = 10^{-3}$

h	DOF	\mathcal{P}_+	$\mathcal{P}_{T_2}(0.6)$	$\mathcal{P}_C(0.6)$	\mathcal{P}_{con_D}	\mathcal{P}_{con_T}
2^{-3}	521	94 (0.15)	49 (0.04)	42 (0.04)	15 (0.01)	8 (0.01)
2^{-4}	2065	126 (0.43)	62 (0.16)	46 (0.11)	20 (0.07)	11 (0.05)
2^{-5}	8225	152 (2.29)	68 (0.82)	41 (0.46)	22 (0.45)	11 (0.33)
2^{-6}	32833	169 (12.1)	72 (4.68)	35 (2.16)	23 (2.60)	12 (2.13)
2^{-7}	131201	185 (74.0)	77 (25.3)	30 (9.29)	23 (13.7)	12 (13.1)
2^{-8}	524545	195 (526)	80 (174)	25 (48.0)	23 (78.4)	12 (73.8)

(d) $\kappa = 10^{-4}$

h	DOF	\mathcal{P}_+	$\mathcal{P}_{T_2}(0.6)$	$\mathcal{P}_C(0.6)$	\mathcal{P}_{con_D}	\mathcal{P}_{con_T}
2^{-3}	521	95 (0.15)	50 (0.04)	42 (0.04)	15 (0.01)	8 (0.01)
2^{-4}	2065	131 (0.45)	65 (0.17)	46 (0.11)	29 (0.10)	15 (0.06)
2^{-5}	8225	163 (2.36)	77 (0.93)	41 (0.45)	35 (0.67)	18 (0.50)
2^{-6}	32833	202 (17.6)	86 (5.83)	34 (2.19)	37 (4.64)	19 (3.71)
2^{-7}	131201	231 (109)	95 (31.8)	27 (8.45)	38 (23.5)	20 (22.4)
2^{-8}	524545	251 (864)	101 (189)	21 (34.6)	38 (123)	19 (114)

Table 4.4: Number of iterations and CPU times for convergence of Problem (4.15) with Darcy-DG, varying κ , and fixed $\nu = 1$.

(a) $\kappa = 10^{-1}$						
h	DOF	\mathcal{P}_+	$\mathcal{P}_{T_2}(0.6)$	$\mathcal{P}_C(0.6)$	\mathcal{P}_{con_D}	\mathcal{P}_{con_T}
2^{-3}	1217	76 (0.22)	42 (0.09)	33 (0.07)	9 (0.02)	5 (0.01)
2^{-4}	4865	92 (1.09)	48 (0.50)	34 (0.36)	9 (0.10)	5 (0.06)
2^{-5}	19457	100 (7.00)	49 (3.28)	31 (2.16)	9 (0.70)	5 (0.42)
2^{-6}	77825	110 (42.0)	52 (17.7)	26 (9.25)	9 (3.84)	5 (2.25)
2^{-7}	311297	113 (237)	55 (104)	20 (37.0)	9 (18.6)	5 (10.1)

(b) $\kappa = 10^{-2}$						
h	DOF	\mathcal{P}_+	$\mathcal{P}_{T_2}(0.6)$	$\mathcal{P}_C(0.6)$	\mathcal{P}_{con_D}	\mathcal{P}_{con_T}
2^{-3}	1217	83 (0.25)	44 (0.10)	37 (0.09)	13 (0.03)	7 (0.2)
2^{-4}	4865	110 (1.38)	53 (0.58)	39 (0.43)	14 (0.16)	8 (0.10)
2^{-5}	19457	123 (8.14)	55 (3.83)	38 (2.63)	15 (1.12)	7 (0.56)
2^{-6}	77825	132 (46.2)	57 (18.6)	33 (10.8)	15 (5.75)	7 (2.73)
2^{-7}	311297	140 (268)	61 (104)	28 (47.9)	15 (30.0)	7 (13.3)

(c) $\kappa = 10^{-3}$						
h	DOF	\mathcal{P}_+	$\mathcal{P}_{T_2}(0.6)$	$\mathcal{P}_C(0.6)$	\mathcal{P}_{con_D}	\mathcal{P}_{con_T}
2^{-3}	1217	86 (0.26)	45 (0.10)	38 (0.09)	15 (0.04)	8 (0.02)
2^{-4}	4865	115 (1.51)	56 (0.66)	40 (0.47)	21 (0.24)	12 (0.14)
2^{-5}	19457	143 (10.8)	62 (4.22)	37 (2.53)	23 (1.67)	12 (0.92)
2^{-6}	77825	159 (64.2)	66 (21.9)	33 (10.7)	23 (8.47)	12 (4.37)
2^{-7}	311297	167 (329)	69 (118)	29 (49.4)	23 (44.4)	12 (22.3)

(d) $\kappa = 10^{-4}$						
h	DOF	\mathcal{P}_+	$\mathcal{P}_{T_2}(0.6)$	$\mathcal{P}_C(0.6)$	\mathcal{P}_{con_D}	\mathcal{P}_{con_T}
2^{-3}	1217	85 (0.26)	46 (0.11)	39 (0.09)	15 (0.04)	8 (0.02)
2^{-4}	4865	117 (1.55)	59 (0.69)	41 (0.48)	31 (0.40)	16 (0.20)
2^{-5}	19457	149 (10.3)	71 (4.54)	38 (2.48)	37 (2.56)	20 (1.41)
2^{-6}	77825	195 (81.2)	82 (31.0)	33 (12.0)	38 (16.6)	20 (8.57)
2^{-7}	311297	218 (486)	88 (141)	27 (43.3)	38 (69.0)	20 (33.6)

Table 4.5: Number of iterations and CPU times for convergence of Problem (4.15) with Darcy-conG, varying ν , and fixed $\kappa = 1$.

(a) $\nu = 10^{-1}$

h	DOF	\mathcal{P}_+	$\mathcal{P}_{T_2}(0.6)$	$\mathcal{P}_C(0.6)$	\mathcal{P}_{con_D}	\mathcal{P}_{con_T}
2^{-3}	521	89 (0.15)	39 (0.04)	37 (0.03)	9 (0.01)	5 (0.01)
2^{-4}	2065	105 (0.38)	47 (0.13)	41 (0.11)	9 (0.04)	5 (0.03)
2^{-5}	8225	111 (1.51)	49 (0.59)	39 (0.46)	9 (0.18)	5 (0.17)
2^{-6}	32833	121 (8.15)	54 (3.55)	35 (2.28)	9 (0.96)	5 (1.24)
2^{-7}	131201	125 (43.8)	55 (18.4)	31 (10.7)	9 (5.19)	5 (7.59)
2^{-8}	524545	129 (303)	57 (104)	24 (44.1)	9 (28.8)	4 (43.3)

(b) $\nu = 10^{-2}$

h	DOF	\mathcal{P}_+	$\mathcal{P}_{T_2}(0.6)$	$\mathcal{P}_C(0.6)$	\mathcal{P}_{con_D}	\mathcal{P}_{con_T}
2^{-3}	521	113 (0.19)	50 (0.05)	37 (0.03)	12 (0.02)	7 (0.01)
2^{-4}	2065	150 (0.57)	66 (0.19)	43 (0.11)	13 (0.05)	7 (0.04)
2^{-5}	8225	162 (2.49)	69 (0.90)	40 (0.49)	14 (0.27)	7 (0.27)
2^{-6}	32833	162 (12.1)	70 (4.73)	37 (2.41)	14 (1.53)	7 (1.65)
2^{-7}	131201	165 (61.8)	71 (23.8)	32 (10.8)	14 (8.12)	7 (10.1)
2^{-8}	524545	169 (368)	73 (159)	28 (56.2)	14 (49.6)	7 (73.4)

(c) $\nu = 10^{-3}$

h	DOF	\mathcal{P}_+	$\mathcal{P}_{T_2}(0.6)$	$\mathcal{P}_C(0.6)$	\mathcal{P}_{con_D}	\mathcal{P}_{con_T}
2^{-3}	521	137 (0.24)	73 (0.08)	45 (0.04)	15 (0.01)	8 (0.01)
2^{-4}	2065	237 (1.15)	117 (0.42)	50 (0.14)	20 (0.07)	11 (0.06)
2^{-5}	8225	270 (5.50)	134 (2.14)	46 (0.58)	22 (0.42)	12 (0.40)
2^{-6}	32833	282 (27.1)	138 (11.6)	42 (3.11)	23 (2.75)	12 (3.00)
2^{-7}	131201	266 (114)	137 (50.2)	37 (12.4)	23 (13.5)	12 (18.5)
2^{-8}	524545	261 (732)	137 (282)	32 (54.6)	23 (65.3)	12 (106)

(d) $\nu = 10^{-4}$

h	DOF	\mathcal{P}_+	$\mathcal{P}_{T_2}(0.6)$	$\mathcal{P}_C(0.6)$	\mathcal{P}_{con_D}	\mathcal{P}_{con_T}
2^{-3}	521	146 (0.28)	78 (0.09)	50 (0.05)	15 (0.01)	8 (0.01)
2^{-4}	2065	341 (2.11)	170 (0.72)	59 (0.18)	28 (0.10)	15 (0.08)
2^{-5}	8225	459 (13.3)	231 (4.71)	58 (0.78)	34 (0.71)	18 (0.72)
2^{-6}	32833	487 (59.0)	246 (22.7)	52 (3.44)	37 (4.03)	20 (4.74)
2^{-7}	131201	485 (273)	235 (125)	46 (18.2)	38 (24.2)	20 (32.7)
2^{-8}	524545	479 (1649)	234 (537)	42 (70.7)	3 (111)	20 (193)

Table 4.6: Number of iterations and CPU times for convergence of Problem (4.15) with Darcy-DG, varying ν , and fixed $\kappa = 1$.

(a) $\nu = 10^{-1}$						
h	DOF	\mathcal{P}_+	$\mathcal{P}_{T_2}(0.6)$	$\mathcal{P}_C(0.6)$	\mathcal{P}_{con_D}	\mathcal{P}_{con_T}
2^{-3}	1217	89 (0.27)	39 (0.08)	37 (0.08)	9 (0.02)	5 (0.01)
2^{-4}	4865	105 (1.27)	46 (0.50)	41 (0.44)	9 (0.11)	5 (0.06)
2^{-5}	19457	112 (7.04)	49 (2.98)	39 (2.38)	9 (0.67)	5 (0.37)
2^{-6}	77825	121 (40.1)	54 (16.3)	34 (10.4)	9 (3.39)	5 (1.95)
2^{-7}	311297	125 (251)	55 (94.8)	31 (54.8)	9 (18.9)	5 (10.2)

(b) $\nu = 10^{-2}$						
h	DOF	\mathcal{P}_+	$\mathcal{P}_{T_2}(0.6)$	$\mathcal{P}_C(0.6)$	\mathcal{P}_{con_D}	\mathcal{P}_{con_T}
2^{-3}	1217	116 (0.36)	52 (0.12)	37 (0.08)	13 (0.03)	7 (0.02)
2^{-4}	4865	152 (2.01)	67 (0.74)	43 (0.46)	14 (0.16)	8 (0.09)
2^{-5}	19457	164 (10.9)	69 (4.25)	40 (2.44)	14 (0.97)	7 (0.55)
2^{-6}	77825	163 (56.0)	70 (21.7)	37 (11.5)	14 (5.51)	7 (2.62)
2^{-7}	311297	165 (319)	71 (118)	33 (58.4)	14 (27.1)	7 (13.0)

(c) $\nu = 10^{-3}$						
h	DOF	\mathcal{P}_+	$\mathcal{P}_{T_2}(0.6)$	$\mathcal{P}_C(0.6)$	\mathcal{P}_{con_D}	\mathcal{P}_{con_T}
2^{-3}	1217	135 (0.43)	72 (0.18)	45 (0.10)	15 (0.03)	8 (0.02)
2^{-4}	4865	249 (4.05)	124 (1.61)	50 (0.55)	22 (0.26)	12 (0.14)
2^{-5}	19457	284 (23.6)	137 (9.95)	46 (3.09)	23 (1.61)	12 (0.97)
2^{-6}	77825	285 (121)	138 (52.7)	42 (15.5)	23 (10.2)	12 (5.06)
2^{-7}	311297	273 (669)	137 (233)	38 (60.7)	23 (43.7)	12 (21.1)

(d) $\nu = 10^{-4}$						
h	DOF	\mathcal{P}_+	$\mathcal{P}_{T_2}(0.6)$	$\mathcal{P}_C(0.6)$	\mathcal{P}_{con_D}	\mathcal{P}_{con_T}
2^{-3}	1217	142 (0.47)	77 (0.20)	51 (0.12)	15 (0.03)	8 (0.02)
2^{-4}	4865	350 (6.51)	181 (2.61)	59 (0.67)	31 (0.36)	16 (0.19)
2^{-5}	19457	481 (45.7)	231 (17.4)	58 (3.55)	38 (2.54)	20 (1.37)
2^{-6}	77825	511 (218)	246 (85.4)	52 (15.4)	38 (13.6)	21 (7.21)
2^{-7}	311297	485 (1285)	246 (481)	46 (76.8)	38 (72.1)	21 (40.5)

4.4.3 3D: Inexact Preconditioning

In the following sections, we consider several 3D experiments that have been computed on hexadral meshes with continuous finite element basis functions in both the Darcy and Stokes domains. The numerical implementation uses the `deal.II` finite element library [4].

In order to consider the solution of large matrices that arise from the finite element discretization of such 3D problems, inexact versions of the preconditioners must be used. This means that inner block solves are now computed by another iterative method that is preconditioned with a spectrally equivalent multigrid method. In this case, the action of the preconditioner is no longer fixed at each iteration and so a flexible Krylov subspace method, in particular flexible GMRES (FGMRES) [47], which allows for the preconditioner to change at each inner iteration is used as the Krylov solver. Here, we describe the inexact implementations of the following three preconditioners, $\tilde{\mathcal{P}}_+$, $\tilde{\mathcal{P}}_{T_1}(\rho)$, and $\tilde{\mathcal{P}}_{con_D}$, which we consider in the solution of the 3D numerical experiments.

Recall that for the continuous finite element basis functions, the matrices A_{Ω_i} are s.p.d. Thus, for the inexact implementation of $\tilde{\mathcal{P}}_+$, we consider a preconditioner of the form

$$\tilde{\mathcal{P}}_+^{-1} = \begin{bmatrix} \tilde{A}_{\Omega_2}^{-1} & 0 & 0 \\ 0 & \tilde{A}_{\Omega_1}^{-1} & 0 \\ 0 & 0 & M_p^{-1} \end{bmatrix},$$

where the action of the matrix $\tilde{A}_{\Omega_i}^{-1}$ is computed by a call to the conjugate gradient method [32] preconditioned with a spectrally equivalent multigrid V-cycle. The inverse of the mass matrix M_p^{-1} is trivial to compute as mass matrices corresponding to first order continuous finite element basis functions are diagonal matrices.

The inexact version of the block lower triangular preconditioned is derived

by applying the factorization (2.1b) to $\tilde{\mathcal{P}}_{T_1}(\rho)$, giving

$$\tilde{\mathcal{P}}_{T_1}(\rho) = \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & B\tilde{A}_{\Omega_1}^{-1} & I \end{bmatrix} \begin{bmatrix} \tilde{A}_{\Omega_2} & 0 & 0 \\ 0 & \tilde{A}_{\Omega_1} & 0 \\ 0 & 0 & -\rho M_p \end{bmatrix}.$$

Thus, the inverse is

$$\tilde{\mathcal{P}}_{T_1}(\rho)^{-1} = \begin{bmatrix} \tilde{A}_{\Omega_2}^{-1} & 0 & 0 \\ 0 & \tilde{A}_{\Omega_1}^{-1} & 0 \\ 0 & 0 & -\rho M_p^{-1} \end{bmatrix} \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & -B\tilde{A}_{\Omega_1}^{-1} & I \end{bmatrix}.$$

The action of this matrix on a vector is given by

$$\begin{aligned} \tilde{\mathcal{P}}_{T_1}(\rho)\mathbf{x} &= \begin{bmatrix} \tilde{A}_{\Omega_2}^{-1} & 0 & 0 \\ 0 & \tilde{A}_{\Omega_1}^{-1} & 0 \\ 0 & 0 & -\rho M_p^{-1} \end{bmatrix} \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & -B\tilde{A}_{\Omega_1}^{-1} & I \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \\ &= \begin{bmatrix} \tilde{A}_{\Omega_2}^{-1}x_1 \\ \tilde{A}_{\Omega_1}^{-1}x_2 \\ -\rho M_p^{-1}(x_3 - B\tilde{A}_{\Omega_1}^{-1}x_2) \end{bmatrix}. \end{aligned}$$

For the inexact version of this block lower triangular preconditioner, the actions of $\tilde{A}_{\Omega_i}^{-1}$ are again replaced by a calls to the conjugate gradient algorithm with spectrally equivalent multigrid V-cycle preconditioners. Compared to $\tilde{\mathcal{P}}_+$, the cost for applying $\tilde{\mathcal{P}}_{T_1}$ is an additional matrix-vector product with B .

Lastly, in order to apply the constraint preconditioner inexactly and in a manner consistent with Theorem 3.5, we utilize the block structure of the preconditioner, recall that

$$\mathcal{P}_{con_D}^{-1} = \begin{bmatrix} A_{\Omega_2}^{-1} & 0 \\ 0 & \begin{bmatrix} A_{\Omega_1} & B^T \\ B & 0 \end{bmatrix}^{-1} \end{bmatrix}. \quad (4.16)$$

Thus, the action of $\mathcal{P}_{con_D}^{-1}$ amounts to two independent block solves, one with the Darcy operator and the other with the Stokes operator. The inexact

version of this preconditioner ignores the coupling between the two problems. The motivation of this inexact version of the constraint preconditioner is that we can use existing fast solvers for the two sub-problems.

The action of the Darcy and Stokes operators are represented by multigrid preconditioned short-term recurrence Krylov subspace methods. Specifically, the action of the inverse of the Darcy operator $A_{\Omega_2}^{-1}$, is replaced by calls to the preconditioned conjugate gradient method with a spectrally equivalent multigrid V-cycle.

The Stokes operator

$$\begin{bmatrix} A_{\Omega_1} & B^T \\ B & 0 \end{bmatrix}^{-1},$$

is symmetric and indefinite. Thus, the action of this matrix is implemented with a call to the preconditioned MINRES method [43]. The preconditioner in this case is a block diagonal preconditioner of the form

$$P^{-1} = \begin{bmatrix} \tilde{A}_{\Omega_1}^{-1} & \\ & \tilde{M}_p^{-1} \end{bmatrix},$$

where \tilde{A}_{Ω_1} (essentially the representation of a vector Laplacian operator) is a multigrid V-cycle, and M_p is the pressure-mass matrix.

Each of the considered preconditioners requires a stopping criterion for the inner iterative method. In each of the following numerical experiments the inner method is stopped when the relative residual has dropped below a tolerance of 10^{-2} . It is possible to have inner solves with increased accuracy but this just increases the total solve time of the method. Moreover, we found this lax tolerance to be sufficient for the following numerical experiments.

4.4.4 3D Test Problem: cube domain

Here, we consider coupled Stokes-Darcy flow in a three dimensional cube, $\Omega = [0, 2]^3$, see Figure 4.7, with the following prescribed solution

$$\begin{cases} \mathbf{u}_1(x, y, z) &= [0, 0, -1]^T, \\ p_1(x, y, z) &= z - 1, \\ p_2(x, y, z) &= z - 1. \end{cases} \quad (4.17)$$

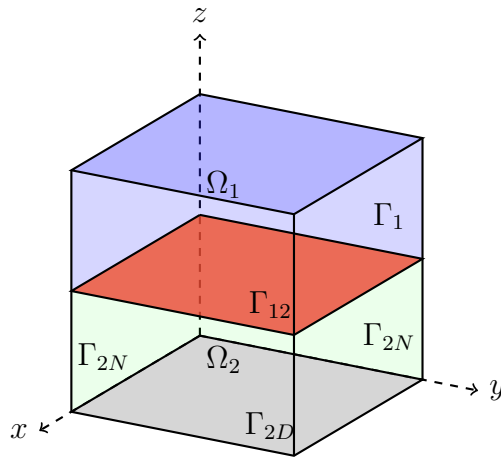


Figure 4.7: The cube computational domain. The interface between the Stokes and Darcy regions Γ_{12} is colored red. The boundary of the Stokes domain Γ_1 is colored blue. Dirichlet boundary conditions are enforced on all five sides of Γ_1 . On the four lateral sides of the Darcy boundary denoted Γ_{2N} and colored green we enforce Neumann boundary conditions. Finally, the bottom of the Darcy boundary colored in grey and denoted Γ_{2D} is the Dirichlet portion of the boundary.

In this case, the system parameters are set to be $\kappa = \nu = 1$. On the five faces of the Stokes boundary, Dirichlet boundary conditions are prescribed to match the velocity solution. On the four lateral faces of the the Darcy boundary, Neumann boundary conditions are enforced. Note that for this solution, the Neumann boundary conditions are homogeneous as $\nabla p_2 \cdot \mathbf{n}_2 = 0$.

We solve this problem numerically on a hexadral mesh using the software package `deal.II` [4]. For the Stokes velocity and pressure basis functions we consider biquadratic and bilinear finite elements, respectively. This is the familiar Taylor-Hood pair and is inf-sup stable [56]. For the Darcy pressure basis functions we consider continuous biquadratic elements.

We solve the discrete linear system using the inexact versions of the following three preconditioners, \mathcal{P}_+ , $\mathcal{P}_{T_1}(0.6)$ and \mathcal{P}_{con_D} . The inexact versions of these preconditioner are not fixed at each iteration and for this reason we use FGMRES [47]. The stopping criterion is when the relative residual, $\|\mathbf{r}_k\|_2/\|\mathbf{r}_0\|_2 < 10^{-6}$. The results of this experiment are reported in Table 4.7.

h	elements	DOF	\mathcal{P}_+	$\mathcal{P}_{T_1}(0.6)$	\mathcal{P}_{con_D}
2^{-1}	64	1695	39 (13.7)	39 (13.7)	6 (5.07)
2^{-2}	512	10809	44 (112)	43 (111)	6 (77.1)
2^{-3}	4096	76653	43 (1294)	41 (1262)	6 (858)
2^{-4}	32768	576213	42 (16449)	41(16256)	5 (18979)

(a) $\kappa = 1$

h	elements	DOF	\mathcal{P}_+	$\mathcal{P}_{T_1}(0.6)$	\mathcal{P}_{con_D}
2^{-1}	64	1695	86 (29.5)	107 (35.7)	10 (8.49)
2^{-2}	512	10809	87 (227)	115 (294)	10 (111)
2^{-3}	4096	76653	84 (2524)	112 (3420)	9 (1260)
2^{-4}	32768	576213	84 (33095)	104 (41619)	9 (32026)

(b) $\kappa = 10^{-1}$

h	elements	DOF	\mathcal{P}_+	$\mathcal{P}_{T_1}(0.6)$	\mathcal{P}_{con_D}
2^{-1}	64	1695	970 (317)	1470 (495)	23 (17.2)
2^{-2}	512	10809	1294 (3237)	2057 (5348)	22 (255)
2^{-3}	4096	76653	1500 (43235)	2053 (62647)	20 (2822)
2^{-4}	32768	576213	*	*	19 (71862)

(c) $\kappa = 10^{-2}$

Table 4.7: Table of FGMRES iterations and CPU time in seconds for Problem (4.17) with inexact preconditioners. Varying κ with fixed $\nu = 1$. A * indicates the maximum run time of 24 hours was reached.

Observe that as the mesh is uniformly refined the number of FGMRES

iterations to converge with the constraint preconditioner remains bounded. In addition, the constraint preconditioner is competitive in its CPU timings. We also remark that sparse direct solvers were competitive until the third level of refinement and ran out of memory for the matrix of size 576213.

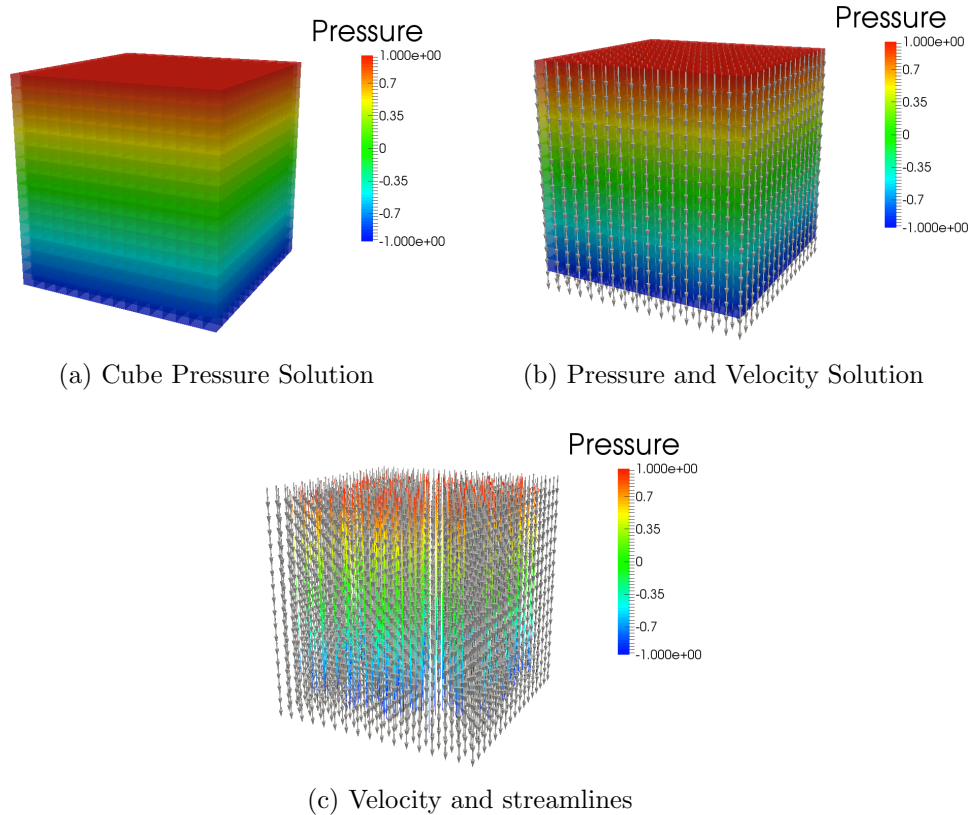


Figure 4.8: Inexact preconditioned GMRES solution obtained using $\mathcal{P}_{con,D}$. The mesh width is $h = 2^{-4}$ and the system size is $n = 576213$.

4.4.5 3D Test Problem: rectangular prism domain

In this set of numerical experiments, we consider coupled Stokes-Darcy flow in a rectangular box, $[0, 0.05] \times [0, 0.05] \times [0, 0.25]$, see Figure 4.9. The plane $z = 0.10$, colored red, is the interface Γ_{12} between the Stokes and Darcy domains. The domain considered here has a more elongated free flow region

compared to that of the porous flow region. At the top of the Stokes boundary, colored yellow and denoted $\Gamma_{1,D}$ we prescribe a constant inflow velocity along the z -direction, i.e., $\mathbf{u}_1 = (0, 0, -0.1)^T$. On the four lateral walls of the Stokes boundary, colored blue, we prescribed no slip or zero-velocity boundary conditions. In the Darcy domain, the velocity is fixed to be zero at the four lateral walls of the boundary, which are colored green and denoted by Γ_{2N} . At the bottom of the Darcy boundary denoted by Γ_{2D} and colored grey, the value of the Darcy pressure is fixed at zero.

We consider the solution of the discrete Stokes-Darcy operator and compare the inexact versions of the three preconditioners, \mathcal{P}_+ , $\mathcal{P}_{T_1}(0.6)$ and \mathcal{P}_{con_D} . Here, we vary the permeability of the porous medium over the following range of values $\kappa = \{10^{-2}, 10^{-3}, 10^{-4}\}$. We again compare iteration counts and CPU timings. These experiments were inspired by those in [31] and also illustrate how the proposed constraint preconditioned solver is able to reproduce the expected linear increase in the pressure drop, the difference between the pressure at the top and bottom of the Darcy domain, as the permeability κ is decreased.

The convergence information is contained in Table 4.8. Observe that in terms of iteration counts, the inexact constraint preconditioner is superior. However, for this problem, the timings for the inexact block diagonal preconditioner is superior. For this problem, the size of the Stokes region, and consequently, the Stokes sub-matrix is far larger in size compared to the Darcy operator. Thus, more time is spent in the MINRES solution of the larger Stokes problem, increasing the time for the considered constraint preconditioner.

Various plots from the inexact constraint preconditioned solutions for the considered values of κ are given in Figure 4.10. The plots are slices of the velocity magnitude and the Darcy pressure. The velocity magnitude slices serve to illustrate the parabolic flow profile in the Stokes flow region. The Darcy pressure solution illustrates the gradual drop in pressure across the Darcy region. The pressure in the Stokes domain is colored blue as we only plot the Darcy pressure in this case. Lastly, the plot that shows the expected

linear drop in pressure across the Darcy domain as κ is decreases is given in Figure 4.11.

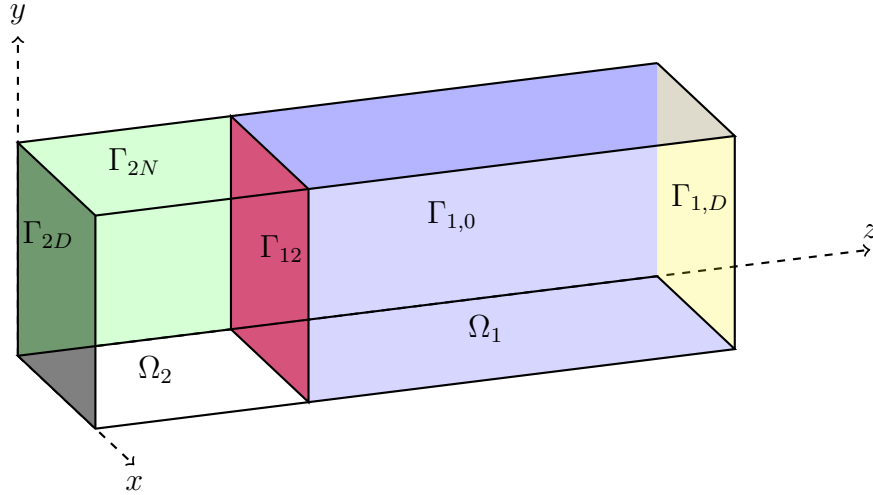


Figure 4.9: The rectangular prism computational domain. The interface between the Stokes and Darcy regions Γ_{12} is colored red. The boundary of the Stokes domain is comprised of two portions. The first piece $\Gamma_{1,0}$ consisting of the four lateral sides of the Stokes boundary, is colored blue and denotes where the velocity is zero. The second piece $\Gamma_{1,D}$ at the top of the Stokes boundary is colored yellow and denotes the inflow boundary condition. The Darcy boundary consists of Neumann boundary conditions on the four lateral sides colored green and denoted by Γ_{2N} . The Dirichlet portion of the Darcy boundary, denoted Γ_{2D} is the bottom of the rectangular prism and is colored grey.

4.4.6 3D Test Problem: discontinuous permeability field

In this section, we again consider numerical experiments for the coupled Stokes-Darcy system in a cube domain; see Figure 4.12. However, in this situation the permeability of the porous medium consists of two values, κ_1 and κ_2 . The value κ_2 is the value inside of the sphere $B_r(\mathbf{x}_c)$ (colored black) that is centered at the point $\mathbf{x}_c = (1, 1, 0.5)^T$ of radius $r = 0.25$ and outside of this

h	elements	DOF	\mathcal{P}_+	$\mathcal{P}_{T_1}(0.6)$	\mathcal{P}_{con_D}
2^{-1}	200	5879	38 (44.8)	32 (39.3)	5 (57.1)
2^{-2}	1600	37489	40 (441)	36 (408)	5 (1190)
2^{-3}	12800	265277	44 (5978)	38 (5307)	5 (14146)

(a) $\kappa = 10^{-2}$

h	elements	DOF	\mathcal{P}_+	$\mathcal{P}_{T_1}(0.6)$	\mathcal{P}_{con_D}
2^{-1}	200	5879	42 (51)	38 (48.2)	6 (69.4)
2^{-2}	1600	37489	46 (515)	43 (494)	6 (706)
2^{-3}	12800	265277	48 (6636)	46 (6448)	6 (10717)

(b) $\kappa = 10^{-3}$

h	elements	DOF	\mathcal{P}_+	$\mathcal{P}_{T_1}(0.6)$	\mathcal{P}_{con_D}
2^{-1}	200	5879	53 (65.5)	49 (63.9)	9 (96.2)
2^{-2}	1600	37489	64 (737)	65 (782)	9 (953)
2^{-3}	12800	265277	66 (9385)	72 (10371)	9 (15249)

(c) $\kappa = 10^{-4}$

Table 4.8: Table of FGMRES iterations and time in seconds for flow in the rectangular prism domain with the considered inexact preconditioners.

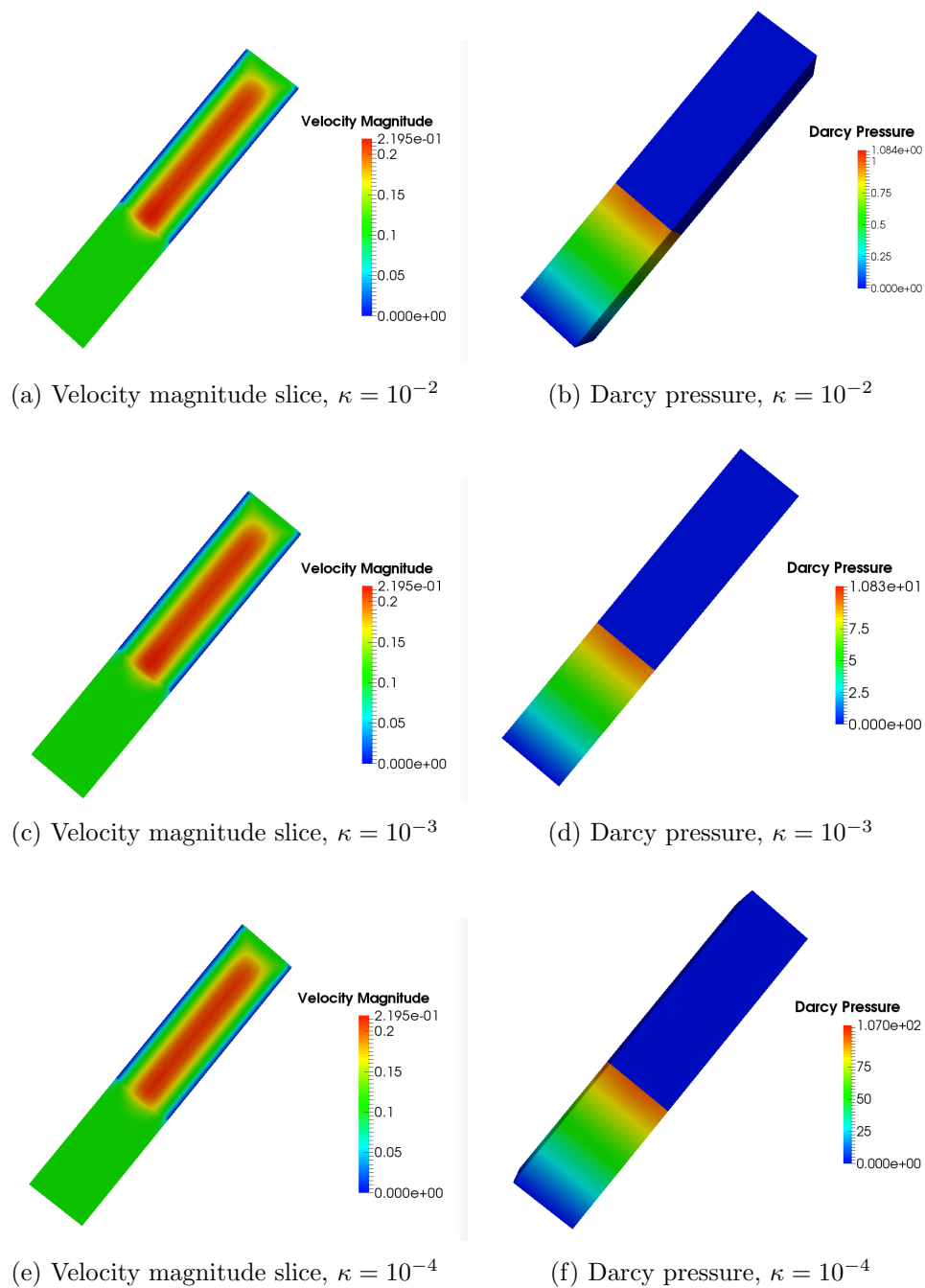


Figure 4.10: 3D Prism solutions. The plots show slices of the coupled Stokes-Darcy velocity magnitude and the Darcy pressure. The size of the system matrix for this problem is $n = 265277$ corresponding to a mesh width of $h = 2^{-3}$.

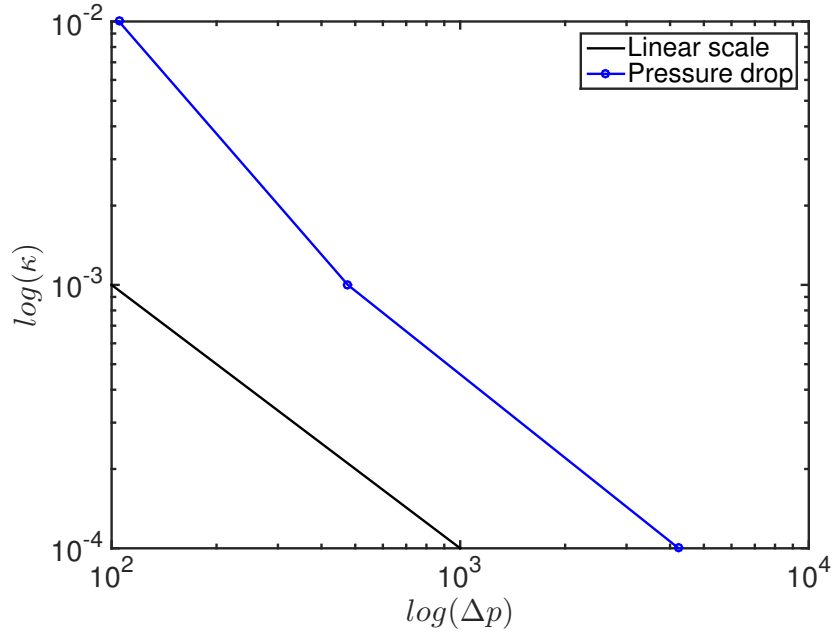


Figure 4.11: Log-log plot of the drop in pressure Δp in the Darcy domain against the decrease of κ .

sphere, the value is κ_1 . The boundary conditions are the same as in Problem (4.17). Here, we consider the following set of pairs for the discontinuous permeability values $(\kappa_0, \kappa_1) = \{(10^{-2}, 10^{-5}), 10^{-3}, 10^{-6}), (10^{-4}, 10^{-7})\}$.

The number of iterations for the relative residual to converge to the tolerance $\|\mathbf{r}_k\|_2 / \|\mathbf{r}_0\|_2 < 10^{-6}$ and CPU timings are given in Table 4.9. Observe that as the permeabilities decrease the inexact constraint preconditioner becomes more competitive. In fact, for the low permeability range, the number of outer iterations drastically increases for both the block diagonal and block lower triangular preconditioners, whereas the number of iterations remains bounded for the constraint preconditioner. These experiments also highlight the inherent difficulty each preconditioner has for solving these more challenging problems as the time for each solver increases for smaller values of κ . In order to solve these more challenging problems for smaller mesh widths, parallel calculations are needed. Plots of the Darcy velocity and pressure are

given in Figure 4.13. Observe how the fluid bends around the impermeable enclosure as well as the gradual pressure drop from the top to the bottom of the Darcy domain.

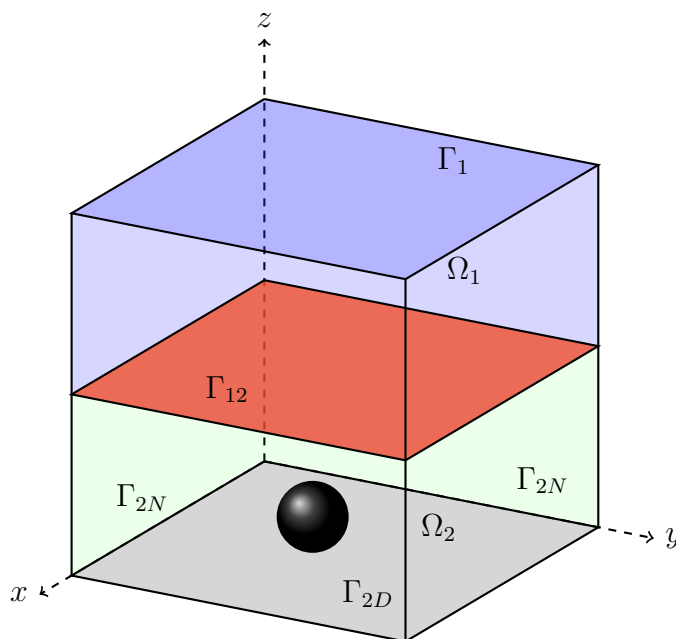
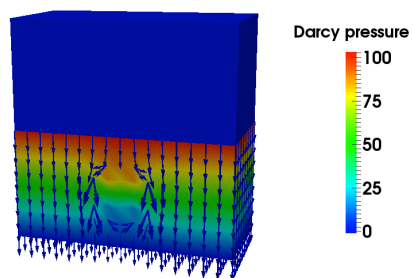
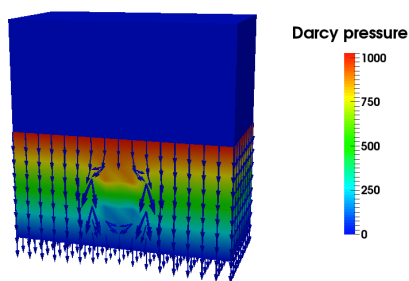


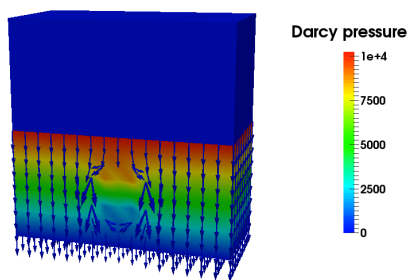
Figure 4.12: The cube computational domain with a discontinuous permeability field in the Darcy region. The interface between the Stokes and Darcy regions Γ_{12} is colored red. The boundary of the Stokes domain Γ_1 is colored blue. Dirichlet boundary conditions are enforced on all five sides of Γ_1 . On the four lateral sides of the Darcy domain denoted Γ_{2N} and colored in green we enforce Neumann boundary conditions. Finally, on the bottom portion of the Darcy domain denoted Γ_{2D} and colored grey we enforce Dirichlet boundary conditions.



(a) Darcy velocity and pressure $(\kappa_1, \kappa_2) = (10^{-2}, 10^{-5})$



(b) Darcy velocity and pressure $(\kappa_1, \kappa_2) = (10^{-3}, 10^{-6})$



(c) Darcy velocity and pressure $(\kappa_1, \kappa_2) = (10^{-4}, 10^{-7})$

Figure 4.13: 3D discontinuous permeability solutions. The plots show the slices of the coupled Stokes-Darcy velocity magnitude and the Darcy pressure. The size of the system matrix for this problem is $n = 76653$ corresponding to a mesh width of $h = 2^{-3}$.

h	elements	DOF	\mathcal{P}_+	$\mathcal{P}_{T_1}(0.6)$	\mathcal{P}_{con_D}
2^{-2}	512	10809	111 (330)	127 (381)	22 (250)
2^{-3}	4096	76653	114 (3725)	136 (4537)	21 (3051)

(a) $(\kappa_1, \kappa_2) = (10^{-2}, 10^{-5})$

h	elements	DOF	\mathcal{P}_+	$\mathcal{P}_{T_1}(0.6)$	\mathcal{P}_{con_D}
2^{-2}	512	10809	221 (636)	249 (745)	55 (566)
2^{-3}	4096	76653	250 (7725)	293 (9179)	53 (7132)

(b) $(\kappa_1, \kappa_2) = (10^{-3}, 10^{-6})$

h	elements	DOF	\mathcal{P}_+	$\mathcal{P}_{T_1}(0.6)$	\mathcal{P}_{con_D}
2^{-2}	512	10809	366 (1105)	362 (1108)	122 (1189)
2^{-3}	4096	76653	513 (15770)	570 (17803)	135 (16665)

(c) $(\kappa_1, \kappa_2) = (10^{-4}, 10^{-7})$

Table 4.9: Table of FGMRES iterations and CPU time in seconds for the discontinuous permeability field problem with inexact preconditioners. A * indicates the maximum run time of 24 hours was reached.

CHAPTER 5

CONCLUSIONS

In this thesis, we have considered the preconditioned iterative solution of large, sparse, saddle point problems. The major theoretical contribution of this thesis is the result on f.o.v.-equivalence between exact and inexact versions of constraint preconditioners to saddle point matrices of the form (3.1). The theory is relevant for saddle point matrices that arise in the finite element discretization of certain p.d.e.s. The important consequence being that, under certain conditions, the GMRES algorithm applied to constraint preconditioned saddle point problems converges in a number of iterations that is independent of the mesh width. Thus, this type of preconditioner is optimal with respect to the mesh width.

In addition to the theoretical results on constraint preconditioners, we performed a variety of numerical experiments on the saddle point system that arises from the finite element discretization of the coupled Stokes-Darcy problem. These experiments were performed for a variety of geometries, in both two and three dimensions, for two discretization schemes (continuous and discontinuous Galerkin methods), as well as a test problem with a discontinuous permeability field. In each of the considered cases, and as expected from the proven theoretical results, the number of iterations for constraint preconditioned GMRES (FGMRES for inexact versions of the preconditioners) to converge was bounded. Moreover, the CPU timings of the constraint precon-

ditioned GMRES algorithm in the considered numerical experiments were as competitive, or better than standard block diagonal and block lower triangular preconditioners. In fact, for more difficult problems, in particular for the large-scale 3D problems with small permeabilities in the Darcy region, the constraint preconditioner was more effective than the other considered preconditioners.

The proposed inexact version of the constraint preconditioner also introduces an interesting paradigm for the solution of coupled problems. This inexact version of the constraint preconditioner essentially ignores the coupling, i.e., $A_{\Gamma_{12}}$, and in some sense can be viewed as a decoupled Stokes-Darcy operator. The effectiveness of this preconditioner then amounts to being able to accurately solve the individual Darcy and Stokes problem quickly with iterative methods. With fast inner iterative solves for these problems, this decoupled constraint preconditioner offers a method that converges in only a few outer iterations, even for extremely large-scale problems in 3D. It would also be natural to apply this decoupled preconditioning approach for other interesting coupled multi-physics problems that exhibit a similar structure to the systems considered here. Another natural extension for this research would be for coupled Navier-Stokes-Darcy flow where the non-linear convection term now plays an important role in the governing equations and solution.

It has been demonstrated, both in theory and in the considered numerical experiments, that constraint preconditioners are a viable and powerful preconditioning method for saddle point problems. At first, it may seem unnatural to precondition a saddle point problem with another saddle point problem. However, these indefinite, constraint preconditioners are extremely effective and can be more economical, both in terms of iteration count and CPU time than other standard block diagonal and block lower triangular preconditioners, especially for challenging large-scale 3D problems of interest.

REFERENCES

- [1] Owl's Nest High Performance Computing, Temple University. www.hpc.temple.edu/owlsnest/OwlsnestUserGuide.html.
- [2] Douglas N. Arnold, Franco Brezzi, and Michel Fortin. A stable finite element for the Stokes equations. *Calcolo*, 21:337–344, 1984.
- [3] Ivo Babuška. Error-bounds for finite element method. *Numerische Mathematik*, 16:322–333, 1971.
- [4] Wolfgang Bangerth, Timo Heister, Luca Heltai, Guido Kanschat, Martin Kronbichler, Matthias Maier, Bruno Turcksin, and Toby D. Young. The `deal.ii` library, version 8.2. *Archive of Numerical Software*, 3, 2015.
- [5] Gordon S. Beavers and Daniel D. Joseph. Boundary conditions at a naturally permeable wall. *Journal of Fluid Mechanics*, 30:197–207, 1967.
- [6] Michele Benzi, Gene H. Golub, and Jörg Liesen. Numerical solution of saddle point problems. *Acta Numerica*, 14:1–137, 2005.
- [7] Dietrich Braess and Wolfgang Hackbusch. A new convergence proof for the multigrid method including the V-cycle. *SIAM Journal on Numerical Analysis*, 20:967–975, 1983.
- [8] Achi Brandt. Algebraic multigrid theory: The symmetric case. *Applied Mathematics and Computation*, 19:23–56, 1986.

- [9] Susanne C. Brenner and Ridgway Scott. *The Mathematical Theory of Finite Element Methods*. Springer, New York, 3rd edition, 2008.
- [10] Franco Brezzi. On the existence, uniqueness and approximation of saddle-point problems arising from lagrangian multipliers. *ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique*, 8:129–151, 1974.
- [11] Franco Brezzi and Michel Fortin. *Mixed and Hybrid Finite Element Methods*. Springer New York, 1991.
- [12] William L. Briggs, Van Emden Henson, and Steve F. McCormick. *A Multigrid Tutorial*. SIAM, Philadelphia, 2nd edition, 2000.
- [13] Mingchao Cai, Mo Mu, and Jinchao Xu. Preconditioning techniques for a mixed Stokes/Darcy model in porous media applications. *Journal of Computational and Applied Mathematics*, 233:346–355, 2009.
- [14] Stephen L. Campbell, Ilse C.F. Ipsen, C. Tim Kelley, and Carl D. Meyer. GMRES and the minimal polynomial. *BIT Numerical Mathematics*, 36:664–675, 1996.
- [15] Prince Chidyagwai and Beatrice Rivière. Numerical modelling of coupled surface and subsurface flow systems. *Advances in Water Resources*, 33:92–105, 2010.
- [16] Philippe G. Ciarlet. *The Finite Element Method for Elliptic Problems*. North-Holland, Amsterdam, 1978.
- [17] Eric de Sturler and Jörg Liesen. Block-diagonal and constraint preconditioners for nonsymmetric indefinite linear systems. Part I: Theory. *SIAM Journal on Scientific Computing*, 26:1598–1619, 2005.
- [18] Marco Discacciati and Alfio Quarteroni. Navier-Stokes/Darcy coupling: modeling, analysis, and numerical approximation. *Revista Matemática Complutense*, 22:315–426, 2009.

- [19] H. Sue Dollar. Constraint-style preconditioners for regularized saddle point problems. *SIAM Journal on Matrix Analysis and Applications*, 29:672–684, 2007.
- [20] Iain S. Duff, Albert Maurice Erisman, and Jon Ker Reid. *Direct Methods for Sparse Matrices*. Oxford University Press, Oxford, 1986.
- [21] Michael Eiermann and Oliver G. Ernst. Geometric aspects of the theory of krylov subspace methods. *Acta Numerica 2001*, 10:251–312, 2001.
- [22] Howard C. Elman. *Iterative methods for large, sparse, nonsymmetric systems of linear equations*. PhD thesis, Department of Computer Science, Yale University New Haven, Connecticut, 1982.
- [23] Howard C. Elman, David J. Silvester, and Andrew J. Wathen. *Finite Elements and Fast Iterative Solvers: with applications in incompressible fluid dynamics*. Oxford University Press, New York, 2005.
- [24] Mark Embree. How descriptive are GMRES convergence bounds? Technical Report 08, Oxford University Computing Laboratory, 1999.
- [25] Lawrence C. Evans. *Partial Differential Equations*, volume 19 of *Graduate studies in mathematics*. American Mathematical Society, Providence, RI, 1998.
- [26] Robert D. Falgout and Panayot S. Vassilevski. On generalizing the algebraic multigrid framework. *SIAM Journal on Numerical Analysis*, 42:1669–1693, 2004.
- [27] Mark S. Gockenbach. *Understanding and Implementing the Finite Element Method*. SIAM, Philadelphia, 1987.
- [28] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 4th edition, 2013.

- [29] Anne Greenbaum. *Iterative Methods for Solving Linear Systems*. SIAM, Philadelphia, 1997.
- [30] Wolfgang Hackbusch. *Multi-Grid Methods and Applications*. Springer, Berlin, Heidelberg, 1985.
- [31] Navraj S. Hanspal, Vahid Nassehi, and Abhijit Kulkarni. Three-dimensional finite element modelling of coupled free/porous flows: Applications to industrial and environmental flows. *International Journal for Numerical Methods in Fluids*, 71:1382–1421, 2013.
- [32] Magnus R. Hestenes and Eduard Stiefel. Methods of conjugate gradients for solving linear systems. *Journal of Research of the National Bureau of Standards*, 49:409–436, 1952.
- [33] Nicholas J. Higham. The Matrix Computation Toolbox. <http://www.ma.man.ac.uk/~higham/mctoolbox>.
- [34] Michael Hinze, Rene Pinnau, Michael Ulbrich, and Stefan Ulbrich. *Optimization with PDE constraints*, volume 23 of *Mathematical Modelling: Theory and Applications*. Springer, New York, 2009.
- [35] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, UK, 2012.
- [36] Carsten Keller, Nicholas I.M. Gould, and Andrew J. Wathen. Constraint preconditioning for indefinite linear systems. *SIAM Journal on Matrix Analysis and Applications*, 21:1300–1317, 2000.
- [37] Axel Klawonn and Gerhard Starke. Block triangular preconditioners for nonsymmetric saddle point problems: field-of-values analysis. *Numerische Mathematik*, 81:577–594, 1999.
- [38] Jörg Liesen and Zdenek Strakos. *Krylov subspace methods: principles and analysis*. Oxford University Press, Oxford, 2012.

- [39] Daniel Loghin and Andrew J. Wathen. Analysis of preconditioners for saddle-point problems. *SIAM Journal on Scientific Computing*, 25:2029–2049, 2004.
- [40] Ladislav Lukšan and Jan Vlček. Indefinitely preconditioned inexact Newton method for large sparse equality constrained nonlinear programming problems. *Numerical Linear Algebra with Applications*, 5:219–247, 1998.
- [41] Malcolm F. Murphy, Gene H. Golub, and Andrew J. Wathen. A note on preconditioning for indefinite linear systems. *SIAM Journal on Scientific Computing*, 21:1969–1972, 2000.
- [42] Jorge Nocedal and Stephen J. Wright. *Numerical optimization*. Springer, New York, 2nd edition, 2006.
- [43] Christopher C. Paige and Michael A. Saunders. Solution of sparse indefinite systems of linear equations. *SIAM Journal on Numerical Analysis*, 12:617–629, 1975.
- [44] Ilaria Perugia and Valeria Simoncini. Block-diagonal and indefinite symmetric preconditioners for mixed finite element formulations. *Numerical Linear Algebra with Applications*, 7:585–616, 2000.
- [45] Béatrice Rivière. Analysis of a discontinuous finite element method for the coupled Stokes and Darcy problems. *Journal of Scientific Computing*, 22:479–500, 2005.
- [46] John W. Ruge and Klaus Stüben. Algebraic multigrid. In Steve. F. McCormick, editor, *Multigrid Methods*, volume 3 of *Frontiers in Applied Mathematics*, pages 73–130. SIAM, Philadelphia, 1987.
- [47] Yousef Saad. A flexible inner-outer preconditioned GMRES algorithm. *SIAM Journal on Scientific Computing*, 14:461–469, 1993.
- [48] Yousef Saad. *Iterative methods for sparse linear systems*. SIAM, Philadelphia, 2003.

- [49] Yousef Saad and Martin H. Schultz. GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM Journal on Scientific and Statistical Computing*, 7:856–869, 1986.
- [50] Philip Geoffrey Saffman. On the boundary condition at the surface of a porous medium. *Studies in Applied Mathematics*, 50:93–101, 1971.
- [51] Debora Sesana and Valeria Simoncini. Spectral analysis of inexact constraint preconditioning for symmetric saddle point matrices. *Linear Algebra and its Applications*, 438:2683–2700, 2013.
- [52] Chris Siefert and Eric de Sturler. Preconditioners for generalized saddle-point problems. *SIAM Journal on Numerical Analysis*, 44:1275–1296, 2006.
- [53] Valeria Simoncini and Daniel B. Szyld. On the occurrence of superlinear convergence of exact and inexact Krylov subspace methods. *SIAM review*, 47:247–272, 2005.
- [54] Valeria Simoncini and Daniel B. Szyld. Recent computational developments in Krylov subspace methods for linear systems. *Numerical Linear Algebra with Applications*, 14:1–59, 2007.
- [55] Gilbert Strang. A framework for equilibrium equations. *SIAM Review*, 30:283–297, 1988.
- [56] Cedric Taylor and Paul Hood. A numerical solution of the Navier-Stokes equations using the finite element technique. *Computers & Fluids*, 1:73–100, 1973.
- [57] Lloyd Nicholas Trefethen and David Bau III. *Numerical Linear Algebra*. SIAM, Philadelphia, 1997.
- [58] Henk A Van der Vorst and Kees Vuik. The superlinear convergence behaviour of GMRES. *Journal of Computational and Applied Mathematics*, 48:327–341, 1993.

- [59] Panayot S. Vassilevski. *Multilevel Block Factorization Preconditioners*. Springer, New York, 2008.
- [60] Walter Zulehner. Analysis of iterative methods for saddle point problems: a unified approach. *Mathematics of computation*, 71:479–506, 2002.