Optimal Preconditioning for Mixed Finite Element Formulation of Second-Order Elliptic Problems

Powell, Catherine E.

2003

MIMS EPrint: **2006.83**

ISSN 1749-9097

Optimal Preconditioning for Mixed Finite Element Formulation of Second-Order Elliptic Problems

A thesis submitted to the University of Manchester Institute of Science and Technology
for the degree of Doctor of Philosophy

By

Catherine Elizabeth Powell

Department of Mathematics

September 2003

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institution of learning.

Abstract

The numerical solution of second-order elliptic partial differential equations (PDES), via so-called mixed finite element methods, gives rise to large, sparse linear systems. Such systems are usually ill-conditioned with respect to the discretisation parameter and the PDE coefficients, a feature that severely degrades the convergence of standard iterative solvers. However, improvements to convergence can often be achieved with preconditioners.

The main focus of this thesis is the design of fast and robust solution schemes for a symmetric and indefinite system arising in the mathematical modelling of flow in porous media. This is a standard variable diffusion problem, described by Darcy's law. The challenge is to treat problems with a wide range of coefficients in the same preconditioning framework.

In the modelling of groundwater flow, permeability coefficients commonly exhibit discontinuities and/or are anisotropic. Several previously suggested preconditioning schemes are not robust in such cases. Very few authors tackle the indefinite problem and previous work has not paid significant attention to the impact of the PDE coefficients. This thesis addresses these important issues.

For the model problem, we construct two block-diagonal preconditioners, consider practical parameter-free implementations of them, and evaluate their performance with respect to anisotropic and discontinuous diffusion coefficients. Generalising some of these ideas leads to a generic, black-box strategy for tackling mixed finite element formulations of a wide range of other elliptic PDES.

Acknowledgements

During the last three years, many people have influenced, directly and indirectly, the development of this thesis. I am indebted most especially to the following people.

First, I must thank my superviser David Silvester, for taking me on as a Ph.D student, for investing time in this project, for teaching me to work independently and for introducing me to the game that is academia, its politics and some of its players.

Money is undoubtedly one of the most important ingredients of a Ph.D thesis, and so I must thank the EPSRC for funding research students and UMIST mathematics department for awarding me one of their grants. Additional support for travel to conferences was also greatly appreciated. In particular, I am grateful to the organising committees of the Householder XV meeting and the 2003 Copper Mountain conference. I also owe thanks to the administrators of the UMIST Peter Allen travel fund for facilitating an academic visit to the University of Maryland.

Many thanks to Valeria Simoncini and Ricardo Nochetto for matrix data and to Doug Arnold for a multigrid code. I am also indebted to Howard Elman for his hospitality during a visit to Maryland.

This thesis could not have been written without the emotional support of my family. I thank my parents, Michael and Barbara, for never telling me to get a 'proper job' and for only ever wanting me to be happy.

Lastly, thanks to all the friends that have made the last three years enjoyable. I want to record some of your names here so that I can look back and remember people and places as well as mathematics. Thanks to Flavia and Eamonn for being such lovely housemates. Oak House wasn't the same without you. Thanks to Natasha for being

such a colourful officemate. Q-floor was a much better place for having you there. Gracias to Carme for Spanish philosophy at Fuel. Thanks to Zoe for being one of the nicest people I've ever met. Thanks to Helen for dinners and animating phone calls. Efharistw to Spyridoula for Greek coffee and gossip. Thanks to you all for keeping me sane with e-mail, dinners, movies and gallons of coffee ;o)

Dedication

For my parents, Michael and Barbara.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The numerical solution of partial differential equations (PDEs) is an extremely challenging task. From simple models of heat diffusion, to complex time-dependent non-linear models of fluid flow, mathematicians describe the physical processes that dominate our world using equations that express rates of change. Solving such equations allows us to predict planetary motion, weather patterns, climate change and pollution effects, understand the currents in our oceans, track orbits of satellites, predict evolution of stock prices, detect cancerous tumours in patients, deblurr satellite images and all manner of important things.

Many such mathematical models do not admit analytical solutions and so we must look to numerical methods to approximate them. This requires not only an understanding of the underlying physical laws governing the processes, but also knowledge of functional analysis, discretisation schemes such as finite element methods, linear algebra, and, crucially, consideration of physical limitations such as computing memory and time.

In the first six chapters of this thesis, we unite all of these considerations to analyse efficient preconditioning schemes for a particular finite element formulation of the following model variable diffusion problem.

1.1 A model diffusion problem

Let Ω be a bounded domain in $I\!R^d, d = 2, 3$. We consider scalar, second-order elliptic problems of the form,

find p satisfying,

$$\begin{aligned} -\nabla \cdot \mathcal{A}\nabla p &= f \quad \text{in } \Omega, \\ p &= g \quad \text{on } \partial\Omega_D, \\ \mathcal{A}\nabla p \cdot \vec{n} &= 0 \quad \text{on } \partial\Omega_N, \end{aligned} \tag{1.1}$$

where $\partial\Omega = \partial\Omega_D \cup \partial\Omega_N$ denotes the boundary of Ω, $\mathcal{A} = \mathcal{A}(\vec{x})$ is a $d \times d$ coefficient tensor and $\vec{n}$ denotes the unit outward normal vector. It is a classical result that a unique solution p exists provided that the tensor $\mathcal{A}$, the source term f, and the boundary data g, satisfy certain regularity requirements. (The problem is formulated rigorously in Chapter 2.)

The boundary-value problem (1.1) arises in mathematical models of fluid flow in porous media (see [30], [58], [40], [85], [49], [89]). The macroscopic flow of groundwater in a porous medium was first shown, by Henri Darcy in 1856, to satisfy the linear relation,

$$\vec{u} = -\frac{k}{\mu}\nabla P_R. \tag{1.2}$$

Here, $\vec{u}$ denotes fluid discharge or velocity, P_R is the 'residual pressure', k is the permeability coefficient of the medium and μ is the fluid viscosity. The true pressure is calculated via $P_R - \rho z g$ where ρ, z and g are, respectively, fluid density, height and the gravitational constant. Coupling (1.2) with the mass conservation law,

$$\nabla \cdot \vec{u} = 0, \tag{1.3}$$

yields precisely the first equation of (1.1) with $f = 0$, $\mathcal{A} = \frac{k}{\mu}\mathcal{I}$ and $p := P_R$. In the context of oil reservoir simulation (see, for example, [85]), (1.1) is referred to as the 'pressure equation'.

If the coefficient $\mathcal{A}$ has the same value at all points in the flow region considered, we describe the medium as homogeneous. In practice, flow domains are comprised of different media with varying porosity volumes, leading to heterogenous problems with discontinuous $\mathcal{A}$. If $\mathcal{A}$ depends on the direction of flow, then we say that the medium is anisotropic. This phenomenon occurs in stratified media e.g. soils with alternating layers. Although many typical values of $\mathcal{A}$ have been experimentally determined (see [58] and references therein), it is not realistic to obtain measurements for every point in space. This has led to probabilistic studies (see [40]) of the numerical solution of (1.1). In this thesis, we consider only the deterministic case. Further, if $\mathcal{A}$ is a given *variable* coefficient tensor, we assume that it can be locally approximated by a piecewise constant function. We do not consider variable diffusion tensors that are highly oscillatory at microscopic scales.

In typical applications, $\vec{u}$ is the variable of primary interest and so we look for a solution to the coupled first-order system,

find $\vec{u}$ and p satisfying,

$$\begin{aligned} \mathcal{A}^{-1}\vec{u} + \nabla p &= 0 & \\ \nabla \cdot \vec{u} &= f & \text{in } \Omega, \\ p &= g & \text{on } \partial\Omega_D, \\ \vec{u}\cdot\vec{n} &= 0 & \text{on } \partial\Omega_N. \end{aligned} \tag{1.4}$$

To fix ideas, we call p and $\vec{u} = -\mathcal{A}\nabla p$ the pressure and velocity solutions respectively. We solve (1.4) numerically, using a finite element method.

For an introduction to finite element methods see, for example, [17], [92] or [79]. These methods are derived from variational formulations. In standard or *primal* methods for discretising PDEs such as (1.1), the solution is regarded as a *minimiser* of an associated functional and Sobolev spaces are the natural choice for solution spaces. However, primal methods are unsuitable in fluid flow modelling. Instead, we shall apply one of a class of so-called *mixed* finite element methods (see Brezzi and Fortin, [26]) which facilitate the simultaneous approximation of p and $\vec{u}$. In this setting, the solution corresponds to a *saddle-point* of a functional. The use of mixed methods is essential in flow modelling because they ensure local mass conservation. Mixed velocity solutions are also known to be more robust with respect to variations in the coefficient term, (see [79, pp.240–241]). The simple-minded approach of post-processing primal pressure solutions to recover $\vec{u}$ can lead to highly inaccurate and nonphysical velocity solutions.

It will be shown in Chapter 2 that mixed finite element approximation of (1.4) yields a linear system of the form,

$$\underbrace{\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix}}_{C} \begin{pmatrix} \underline{u} \\ \underline{p} \end{pmatrix} = \begin{pmatrix} \underline{g} \\ \underline{f} \end{pmatrix}, \tag{1.5}$$

where $A \in \mathbb{R}^{n\times n}$ is symmetric and positive definite and $\underline{u}$ and $\underline{p}$ are the discrete velocity and pressure solutions. Many other applications give rise to systems with this structure; the fields of flow modelling, electromagnetics and constrained optimisation are rich in examples. If the chosen discretisation is 'stable' in a sense to be defined in Chapter 2, then $B \in \mathbb{R}^{m\times n}$ (with $m \leq n$) has rank m and the system is uniquely solvable. However, the question of *how* to solve (1.5) is not a trivial one.

We begin with a brief review of iterative methods applicable to systems of the form (1.5) and some concepts associated with preconditioning. Readers who are already familiar with these topics can skip to the end of this Chapter.

1.2 Iterative solution schemes

First, we require some notation and definitions. Let A be an arbitrary, symmetric matrix in $\mathbb{R}^{n\times n}$ with eigenvalues $\{\lambda_1, \ldots, \lambda_n\}$ and let $\underline{x}$, $\underline{z}$ be arbitrary vectors in $\mathbb{R}^n$. The Euclidean inner-product and induced vector norm are defined by,

$$\begin{aligned} (\underline{x}, \underline{z}) &= \underline{x}^T \underline{z}, \\ \| \underline{x} \|_2^2 &= (\underline{x}, \underline{x}) = \sum_{i=1}^{n} x_i^2, \end{aligned}$$

respectively. It will be our convention to underline discrete vectors $\underline{x}$ in $\mathbb{R}^n$ to distinguish from vector functions $\vec{v} : \Omega \to \mathbb{R}^d$.

Recall that if A is symmetric and positive definite, the values $\{\lambda_1, \ldots, \lambda_n\}$ are real and positive. In that case, we may also define the vector A-norm,

$$\| \underline{x} \|_A^2 = \underline{x}^T A \underline{x},$$

based on the inner-product,

$$(\,\underline{x},\,\underline{z}\,)_A = (\,A\underline{x},\,\underline{z}\,) = \underline{x}^T A\underline{z}.$$

We say that two vectors $\underline{x}$, $\underline{z}$ in $\mathbb{R}^n$ are *orthogonal* with respect to a given inner-product $(\cdot,\cdot)$ if $(\,\underline{x},\,\underline{z}\,) = 0$. Thus, $\underline{x}$ and $\underline{z}$ are *A-orthogonal* if,

$$(\underline{x},\,\underline{z})_A = \underline{x}^T A\underline{z} \;=\; 0.$$

Two important measures of a general $A \in \mathbb{R}^{n\times n}$ are the *spectral radius*,

$$\rho(A) \quad = \quad \max_{1\le i\le n} \mid \lambda_i \mid,$$

and the *condition number*,

$$\kappa \;=\; \kappa(A) \;=\; \frac{\max_i \mid \lambda_i \mid}{\min_i \mid \lambda_i \mid}.$$

More notation will be added later, as required.

Now, system (1.5) is sparse and in practical applications, will typically consist of millions of equations. Traditional direct solvers based on Gaussian elimination are too costly because they do not exploit the sparsity of C. Today, *sparse* direct solution methods are also available (see Duff [44]) and for problems in $\mathbb{R}^2$ they can match the efficiency of iterative methods. For problems in $\mathbb{R}^3$, however, their efficiency deteriorates (see, for example, [76].) To obtain practical schemes, iterative solution methods that exploit sparsity are essential. Our philosophy is that we should choose a method that:

1. Converges to a fixed tolerance in a number of iterations that is bounded from above independently of the discretisation parameter and the PDE coefficients.

2. Converges to a fixed tolerance in only $O(N)$ floating point operations (flops), where $N = n + m$ is the number of equations.

Possible choices of iterative methods for (1.5) are described in the next section.

For symmetric matrices, the convergence of *all* iterative solvers depends on the eigenvalue spectrum of the system matrix. By considering the congruence transformation,

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} = \begin{pmatrix} A & 0 \\ B & I \end{pmatrix} \begin{pmatrix} A^{-1} & 0 \\ 0 & -BA^{-1}B^T \end{pmatrix} \begin{pmatrix} A & B^T \\ 0 & I \end{pmatrix}, \qquad (1.6)$$

and applying Sylvester's law (see [56] p.416) of preservation of inertia, it is a standard result that if B is full rank then C in (1.5) has n positive and m negative eigenvalues. Hence, the problem is *indefinite.*

If an iterative method converges poorly for (1.5), we can look for matrices P that have the property that the eigenvalue spectrum of $P^{-1}C$ is more favourable to convergence and apply the iteration, instead, to the preconditioned system,

$$P^{-1} \begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} \underline{u} \\ \underline{p} \end{pmatrix} = P^{-1} \begin{pmatrix} \underline{g} \\ \underline{f} \end{pmatrix}. \qquad (1.7)$$

Crucially, computing the action of the inverse of the preconditioner P should require no more than $O(N)$ flops. If a preconditioner is implemented as in (1.7), we describe the process as *left preconditioning.* Alternatively, if P is symmetric and positive definite, we can perform *symmetric preconditioning* by writing $P = MM^T$ and solving,

$$M^{-1} \begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} M^{-T} \begin{pmatrix} \underline{y} \\ \underline{z} \end{pmatrix} = M^{-1} \begin{pmatrix} \underline{g} \\ \underline{f} \end{pmatrix}, \text{ with } \begin{pmatrix} \underline{y} \\ \underline{z} \end{pmatrix} = M^T \begin{pmatrix} \underline{u} \\ \underline{p} \end{pmatrix}.$$

In this thesis, an *h-optimal* preconditioner is a matrix operator that accelerates the convergence rate of the chosen iterative solver so that convergence to a fixed tolerance is independent of the discretisation parameter, h. For our model problem, we will also refer to the property of $\mathcal{A}$*-optimality*, which is analogously defined. Preconditioners should be parameter-free where possible and not require eigenvalue estimates or tuning with respect to PDE coefficients. We use the term *black-box* to describe such schemes.

1.2.1 Stationary iterative methods

So-called stationary iterative methods for solving general linear systems,

$$M\underline{x} = \underline{b}, \qquad (1.8)$$

with $M \in I\!R^{n \times n}$ an arbitrary, nonsingular matrix, were popularised in the 1950s. They are all based on a 'splitting' of the coefficient matrix into $M = P - N$ with

$P^{-1}M$ approximating, in some sense, the identity I. Starting from an initial guess $\underline{x}^{(0)}$, iterates are constructed via,

$$P\underline{x}^{(i+1)} = N\underline{x}^{(i)} + \underline{b}, \tag{1.9}$$

or, equivalently,

$$\underline{x}^{(i+1)} = G\underline{x}^{(i)} + \underline{c}, \tag{1.10}$$

where $G = P^{-1}N$ and $\underline{c} = P^{-1}\underline{b}$. Typical choices for the preconditioner P are the diagonal or a triangular part of M, or a linear combination of these. Popular examples include Jacobi and Gauss-Seidel iteration.

For any iterate $\underline{x}^{(i)}$, we define the *residual* $\underline{r}^{(i)} = \underline{b} - M\underline{x}^{(i)}$, and the *error* $\underline{e}^{(i)} = M^{-1}\underline{b} - \underline{x}^{(i)}$. Clearly, $M\underline{e}^{(i)} = \underline{r}^{(i)}$ and the error satisfies $\underline{e}^{(i)} = G^i\underline{e}^{(0)}$. It is not hard to show that the iteration converges if and only if $\rho(G) < 1$ (see [55, p.27]). The choice of the preconditioner is therefore crucial. For fast convergence, $\rho(P^{-1}N)$ must be small. However, convergence can be greatly improved with the use of dynamic parameters. Today, stationary iterative methods have been superceded as solvers although they still play an important role as smoothers in multigrid schemes (see Chapters 4 and 5.)

1.2.2 Uzawa methods

An iteration scheme that specifically tackles saddle-point systems of the form (1.5) is Uzawa's method. In its standard form, it is the algorithm shown in Fig. 1.1.

```
Given initial guess $\underline{p}^{(0)}$,
for $n = 1, 2, \ldots,$          until convergence:
    Solve:       $A\underline{u}^{(n+1)} = \underline{g} - B^T\underline{p}^{(n)}$
    Update:      $\underline{p}^{(n+1)} = \underline{p}^{(n)} + \alpha\left(B\underline{u}^{(n+1)} - \underline{f}\right)$
end
```

Figure 1.1: Standard Uzawa algorithm

The method requires the action of the inverse of the matrix A to be computed at each step. Since this is infeasible in most applications, modern variants use an approximation, leading to a nested iteration. The choice of the relaxation parameter α is crucial for convergence. Optimal values are known (see [52]) but require the solution of an eigenvalue problem. Another deficiency is the need to tune the parameter α and stopping tolerances for the inner-iteration. For a discussion of variants of this method and nested iteration schemes see [52], [11], [45], [86] and [10].

The augmented Lagrangian method, presented by Fortin and Glowinski in [52, Ch.1], refers to the application of Uzawa's algorithm to a modified saddle-point problem. Observe that for *any* r, the solution to (1.5) is also the solution to (1.11),

$$\begin{pmatrix} A + rB^TB & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} \underline{u} \\ \underline{p} \end{pmatrix} = \begin{pmatrix} \underline{g} + rB^T\underline{f} \\ \underline{f} \end{pmatrix}. \tag{1.11}$$

Applying Uzawa's method to this system yields the algorithm shown in Fig. 1.2.

```
Given initial guess $\underline{p}^{(0)}$,

for  $n = 1, 2, \ldots,$          until convergence:

    Solve:                $\left(A + rB^TB\right)\underline{u}^{(n+1)} = \underline{g} + rB^T\underline{f} - B^T\underline{p}^{(n)}$

    Update:               $\underline{p}^{(n+1)} = \underline{p}^{(n)} + \alpha\left(B\underline{u}^{(n+1)} - \underline{f}\right)$

end
```

Figure 1.2: Augmented Lagrangian algorithm

The advantage is that choosing the relaxation parameter $\alpha = r$, where r is large, yields arbitrarily fast convergence. However, the method has a serious flaw. $\kappa\left(A + rB^TB\right)$ increases with r (see Proposition 2.3 in [52]), crippling the inner-iteration. Preconditioners for tackling the inner-solve have been suggested (see for example Hiptmair, [60]) but optimal values of α and r have not been discussed. Parameter tuning is required.

1.2.3 Krylov subspace methods

Today, Krylov subspace methods are among the most powerful tools available for solving large, sparse linear systems. Starting from an initial guess $\underline{x}^{(0)}$, they generate a sequence of iterates $\underline{x}^{(1)}, \underline{x}^{(2)}, \ldots$, for (1.8) such that,

$$\underline{x}^{(i)} \in \underline{x}^{(0)} + \mathrm{span}\{\underline{r}^{(0)}, M\underline{r}^{(0)}, \ldots, M^{i-1}\underline{r}^{(0)}\} = \underline{x}^{(0)} + K_i\left(M, \underline{r}^{(0)}\right),$$

via an iteration of the form,

$$\underline{x}^{(i+1)} = \underline{x}^{(i)} + \alpha^{(i)}\underline{p}^{(i)}, \quad i = 1, 2, \ldots. \tag{1.12}$$

Here, $\alpha^{(i)}$ is a dynamic constant and $\underline{p}^{(i)}$ is a search direction. $K_i\left(M, \underline{r}^{(0)}\right)$ is the Krylov space. The attraction of such schemes is that they require only one matrix-vector multiplication and a few inner-products per iteration, which, for sparse matrices, can be performed in $O(N)$ flops. Two well known examples are the conjugate gradient method (CG) for symmetric positive definite systems and the minimum residual method (MINRES) for symmetric indefinite systems.

Both schemes can be motivated by simple linear algebra arguments. CG chooses,

$$\alpha^{(i)} = \frac{\left(\underline{r}^{(i)}, \underline{r}^{(i)}\right)}{\left(\underline{p}^{(i)}, M\underline{p}^{(i)}\right)},$$

and updates search directions via,

$$\underline{p}^{(i)} = \underline{r}^{(i)} + \beta^{(i-1)}\underline{p}^{(i-1)} \quad \text{with} \quad \beta^{(i-1)} = \frac{\left(\underline{r}^{(i)}, \underline{r}^{(i)}\right)}{\left(\underline{r}^{(i-1)}, \underline{r}^{(i-1)}\right)},$$

so that the error $\underline{e}^{(i+1)} = \underline{x} - \underline{x}^{(i+1)}$ is M-orthogonal to $\underline{p}^{(i)}$ and $\underline{p}^{(i-1)}$. By exploiting a certain three-term recurrence relation, it can be shown (see [55]) that the search directions are *all* M-orthogonal and form a basis for the Krylov space. Consequently, $\underline{e}^{(i+1)}$ is M-orthogonal to all previous search directions and CG minimises the M-norm of the error at each iteration over the space $K_i\left(M, \underline{r}^{(0)}\right)$. Templates can be found in [51] and [55].

When M is indefinite, as in (1.5), CG is unstable. Although M does not define a norm, we can minimise the Euclidean norm of the residuals by choosing, instead,

$$\alpha^{(i)} = \frac{\left(\underline{r}^{(i)}, M\underline{p}^{(i)}\right)}{\left(M\underline{p}^{(i)}, M\underline{p}^{(i)}\right)},$$

and then updating search directions via,

$$\underline{p}^{(i)} \;=\; \underline{r}^{(i)} - \beta^{(i-1)}\underline{p}^{(i-1)} \text{ with } \quad \beta^{(i-1)} = \frac{\left(M\underline{r}^{(i)}, M\underline{p}^{(i-1)}\right)}{\left(M\underline{p}^{(i-1)}, M\underline{p}^{(i-1)}\right)}.$$

MINRES is a scheme that has this minimisation property. However, it must be implemented with care. The iteration breaks down if any of the computed coefficients are zero. For this reason, CG and MINRES are best viewed as variants of the Lanczos method. This link was first established in the important work of Paige and Saunders in [74]. When M is symmetric, the Lanczos algorithm reduces to a three-term recurrence which can be expressed in matrix form as,

$$MQ_i \;=\; Q_{i+1}T_{i+1,i},$$

where $Q_i \in I\!R^{n\times i}$ is an orthonormal matrix whose columns are basis vectors for $K_i\left(M, \underline{r}^{(0)}\right)$ and $T_{i+1,i} \in I\!R^{i+1,i}$ is a symmetric, tridiagonal matrix of recursion coefficients. The coefficients $\alpha^{(i)}$, $\beta^{(i)}$ produced by CG correspond to an LU factorisation of this matrix. This is where break-down can occur for indefinite problems. The iterates produced are of the form,

$$\underline{x}^{(i)} = \underline{x}^{(0)} + Q_i\underline{y}^{(i)}.$$

Choosing $\underline{y}^{(i)}$ to be the vector of coefficients that minimises the Euclidean norm of the residual leads to a least squares problem that requires a QR decomposition of $T_{i+1,i}$. By exploiting recurrence formulae this decomposition can be performed cheaply using Givens rotations. Commercial MINRES codes such as the one in MATLAB (see [69]) use a stable implementation based on an algorithm outlined by Fischer in [51].

1.2.4 Preconditioned MINRES

Our approach to solving (1.5), like those of [6], [86], and [98], is to apply MINRES. Recall that MINRES minimises $\| \underline{r}^{(i)} \|_2$ over the space $\underline{r}^{(0)} + K_i\left(C, \underline{r}^{(0)}\right)$. It follows that there exists an optimal matrix polynomial $p_i^*(C)$, such that $\underline{r}^{(i)} = p_i^*(C)\underline{r}^{(0)}$ and,

$$\| \underline{r}^{(i)} \|_2 = \min_{p_i} \| p_i(C)\underline{r}^{(0)} \|_2,$$

where p_i denotes *any* polynomial of degree i or less satisfying $p_i(0) = 1$. Since C is symmetric and normal, we have the decomposition $C = Q\Lambda_C Q^T$ where Q is an orthogonal matrix and Λ_C denotes the diagonal matrix of eigenvalues of C. It then follows that,

$$\| \underline{r}^{(i)} \|_2 \;\; \leq \;\; \min_{p_i} \| p_i(\Lambda_C) \|_2 \, \| \underline{r}^{(0)} \|_2 \, .$$

Thus, an upper bound for the relative residual error is given by,

$$\frac{\| \underline{r}^{(i)} \|_2}{\| \underline{r}^{(0)} \|_2} \;\; \leq \;\; \min_{p_i} \max_{j=1:n+m} | \, p_i(\lambda_j) \, | , \tag{1.13}$$

where $\{\lambda_1, \ldots, \lambda_{n+m}\}$ denotes the set of eigenvalues of C. The relative residual error is reduced quickly only if it is possible to construct a polynomial of *low* degree, taking value one at the origin, that is close to zero at *all* of those values. Apart from the choice of initial vector, the convergence rate of MINRES is *completely* determined by the spread of the eigenvalues. For our model problem, however, C is ill-conditioned with respect to the discretisation parameter h and the coefficient tensor $\mathcal{A}$ (see Chapter 2.) For an optimal method, we require preconditioners P such that inclusion intervals for the eigenvalues of $P^{-1}C$ are independent of the problem parameters. This is the focus of this thesis.

1.3 Overview

The remainder of the thesis is organised as follows. Chapter 2 serves as an introduction to mixed finite element formulations and outlines technical preliminary results that are needed to understand the discussion in the sequel. We study the well-posedness, in a natural choice of norms, of the variational problem associated with conforming mixed finite element approximations of (1.4). We describe properties of the Raviart-Thomas finite element spaces and review some existing preconditioning schemes.

Chapters 3–5 constitute the main theoretical contributions of the thesis and contain further details and analysis of the preconditioners discussed by Powell and Silvester in [77] and [78]. In Chapter 3, a preconditioner is motivated using the standard stability theory. New eigenvalue analysis is supplied. In Chapter 4, a practical scheme based on the multigrid theory of Arnold et al., [6], is considered. Important new eigenvalue bounds are established.

In Chapter 5, we use an alternative stability theory to motivate a second class of preconditioners. A new practical method is proposed, the key building blocks for which are diagonal scaling for a weighted mass matrix and a fast solver for a generalised diffusion operator based on black-box algebraic multigrid (AMG).

In Chapter 6, we consider a commonly advocated positive definite reformulation of the model problem and make a numerical comparison to the black-box approach of Chapter 5. Finally, the discussion is extended to mixed finite element formulations of Stokes equations and Maxwell's equations in Chapter 7.

Chapter 2

Mixed finite element formulation

In this chapter, we give a rigorous statement of the model problem, and describe an appropriate mixed finite element formulation. Readers who are mainly interested in linear algebra must be patient until section 2.4. Readers who are already familiar with standard stability theory can skip section 2.3. It is necessary to review, first, some fundamental results and technical definitions from the field of functional analysis. For notation and style of presentation, we follow the conventions of Brezzi and Fortin [26, Ch.3] and Hackbush [57, Ch.6].

2.1 Notation and preliminary results

Let Ω be a bounded and connected subset of $I\!R^2$ or $I\!R^3$, with Lipschitz continuous boundary $\partial\Omega = \partial\Omega_D \cup \partial\Omega_N$. The symbols $\partial\Omega_D$ and $\partial\Omega_N$ indicate portions of the boundary where Dirichlet and Neumann conditions are prescribed, respectively. $C^k(\Omega)$ is the set of functions that are defined and have continuous derivatives of order k, or less, on Ω. As usual, $L^2(\Omega)$ denotes the space of scalar functions that are defined and square integrable over Ω in the sense of Lebesque,

$$L^2(\Omega) \quad = \quad \left\{ w \mid \int_\Omega w^2 d\Omega < \infty \right\}.$$

$L^2(\Omega)$ is a Hilbert space equipped with the inner-product,

$$(w, s) \;=\; \int_\Omega ws \; d\Omega,$$

and the induced norm $\| \, w \, \|_0^2 = \; (w, w)$. Analogously, for vector functions $\vec{v} = (v_1, \ldots, v_d)^T$, we define the Hilbert space,

$$L^2(\Omega)^d \;=\; \{\vec{v} \mid v_i \in L^2(\Omega), \, i = 1 : d\},$$

equipped with the inner-product,

$$(\vec{v}, \vec{u}) \;=\; \int_\Omega \vec{v} \cdot \vec{u} \; d\Omega, = \sum_{i=1}^{d} \int_\Omega v_i u_i \; d\Omega,$$

and the induced norm, $\| \; \vec{v} \; \|_0^2 = \; (\vec{v}, \vec{v})$. The overlap in notation should not cause confusion. The inner-product $(\cdot, \cdot)$ and norm $\| \; \cdot \; \|_0$ are understood to be defined componentwise for vectors.

A multi-index, $\vec{\alpha} = (\alpha_1, \cdots, \alpha_d)$ is a set of non-negative integers with

$$\mid \vec{\alpha} \mid = \textstyle\sum_{i=1}^d \alpha_i,$$

and is used to define the partial differential operator,

$$D^{\vec{\alpha}} = \tfrac{\partial^{|\vec{\alpha}|}}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}.$$

Now, given an integer $m \geq 0$, we define the Sobolev spaces,

$$H^m(\Omega) = \{ v \mid v \in L^2(\Omega) \text{ and } D^{\vec{\alpha}} v \in L^2(\Omega), \mid \vec{\alpha} \mid \leq m \}.$$

For a fixed $m \geq 0$, associated semi-norms and norms are given by,

$$\mid v \mid_m^2 = \sum_{|\vec{\alpha}|=m} \| \, D^{\vec{\alpha}} v \, \|_0^2, \qquad \| \, v \, \|_m^2 = \sum_{k \leq m} \mid v \mid_k^2,$$

respectively. We shall mainly be concerned with the Hilbert space,

$$H^1(\Omega) \;=\; \left\{ w \mid w \in L^2(\Omega) \text{ and } \frac{\partial w}{\partial x_1}, \ldots, \frac{\partial w}{\partial x_d} \in L^2(\Omega) \right\},$$

and the subspace,

$$H_0^1(\Omega) \;=\; \left\{ w \mid w \in H^1(\Omega) \text{ and } w = 0 \text{ on } \partial\Omega \right\}.$$

Recall that functions belonging to $H_0^1(\Omega)$ satisfy the following fundamental inequality.

Lemma 1 *'Poincaré-Friedrich's inequality'. For all* $w \in H_0^1(\Omega)$,

$$\| w \|_0 \leq C(\Omega) \mid w \mid_1,$$

where $C(\Omega)$ *is a constant that depends on* Ω.

Proof See Braess [17, p.30] for the case where Ω is contained in a d-dimensional cube.

For completeness, we also define the Sobolev dual space,

$$H^{-1}(\Omega) = \left\{ w \mid \int_{\Omega} ws < \infty \quad \forall s \in H_0^1(\Omega) \right\}.$$

In our mixed finite element approximation of the model diffusion problem, we will constantly refer to the space,

$$H(div;\Omega) = \left\{ \vec{v} \mid \vec{v} \in L^2(\Omega)^d \text{ and } \nabla \cdot \vec{v} \in L^2(\Omega) \right\},$$

which clearly satisfies the inclusion $\left(H^1(\Omega)\right)^d \subset H(div;\Omega) \subset L^2(\Omega)^d$. It possesses the natural inner-product,

$$(\vec{u}, \vec{v})_{div} = (\vec{u}, \vec{v}) + (\nabla \cdot \vec{u}, \nabla \cdot \vec{v}),$$

and the induced norm,

$$\| \vec{v} \|_{div}^2 = \| \vec{v} \|_0^2 + \| \nabla \cdot \vec{v} \|_0^2 .$$

To perform analysis with functions in $H(div;\Omega)$, the following version of Green's formula is often useful.

Lemma 2 *Let* $\vec{v} \in H\,(div;\Omega)$ *then,*

$$\int_{\Omega} \nabla \cdot \vec{v}\, w \, d\Omega = - \int_{\Omega} \vec{v} \cdot \nabla \, w \, d\Omega + \int_{\partial\Omega} \vec{v} \cdot \vec{n} \, w \, ds \quad \forall w \in H^1\,(\Omega)\,.$$

Proof See Brezzi and Fortin [26, p.91].

Notice that if the coefficient tensor $\mathcal{A}$ in (1.4) is *symmetric* and if there exist positive constants γ and Γ with $0 < \gamma \leq \Gamma$ such that,

$$\gamma(\vec{v}, \vec{v}) \leq (\mathcal{A}^{-1}\vec{v}, \vec{v}) \quad \leq \quad \Gamma(\vec{v}, \vec{v}), \tag{2.1}$$

for every $\vec{v} : \Omega \to I\!R^d$, then,

$$(\vec{u}, \vec{v})_{div,\mathcal{A}} = \left(\mathcal{A}^{-1}\vec{u}, \vec{v}\right) + (\nabla \cdot \vec{u}, \nabla \cdot \vec{v}),$$

also defines an inner-product on $H(div;\Omega)$ and induces a norm $\|\cdot\|_{div,\mathcal{A}}$ that is equivalent to $\|\cdot\|_{div}$. We shall make these assumptions in the sequel.

Next, since we have assumed that $\partial\Omega$ is smooth, we can define the *trace*, $w|_{\partial\Omega}$, of any $w \in H^1(\Omega)$. The set of all traces of such functions gives rise to the Hilbert space,

$$H^{\frac{1}{2}}(\partial\Omega) = \left\{ g \mid g = w|_{\partial\Omega} \text{ for some } w \in H^1(\Omega) \cap C^0(\overline{\Omega}) \right\}.$$

Similarly, for vector functions, $\vec{v} \in H(div;\Omega)$, the set of *normal traces*, $(\vec{v}\cdot\vec{n})\,|_{\partial\Omega}$, where $\vec{n}$ denotes the outward-pointing normal vector to $\partial\Omega$, gives rise to the dual space,

$$H^{-\frac{1}{2}}(\partial\Omega) = \left\{ q \mid q = (\vec{v}\cdot\vec{n})|_{\partial\Omega} \text{ for some } \vec{v} \in H(div;\Omega) \cap (C^0(\overline{\Omega}))^d \right\}.$$

Now, for any $g \in H^{\frac{1}{2}}(\partial\Omega)$ and $q \in H^{-\frac{1}{2}}(\partial\Omega)$, $\langle\cdot,\cdot\rangle$ represents the *duality pairing*,

$$\langle q, g\rangle = \int_{\partial\Omega} qg\, ds,$$

and we can define the subspace,

$$H_{0,N}(div;\Omega) = \left\{ \vec{v} \in H(div;\Omega) \mid \langle \vec{v}\cdot\vec{n}, w\rangle = 0 \quad \forall w \in H^1_{0,D}(\Omega) \right\}, \tag{2.2}$$

where,

$$H^1_{0,D}(\Omega) = \{ w \in H^1(\Omega) \mid w|_{\partial\Omega_D} = 0 \}. \tag{2.3}$$

We shall see that $H_{0,N}(div;\Omega)$ is an appropriate space in which to seek an approximation to the velocity solution of (1.4). The properties of this space, relative to a given partitioning of the domain, will play an important role in the finite element method.

Suppose, then, that Ω can be subdivided into a set T_h of subdomains, K, where h_K denotes the diameter of the escribed circle of K. We denote,

$$h = \max_{K\in T_h} h_K, \quad h_{min} = \min_{K\in T_h} h_K.$$

The subdomains are called *finite elements.* T_h is the finite element *mesh* and the *discretisation parameter*, h, describes the size of the elements in it. We shall make the following assumptions on T_h :

- $\overline{\Omega} = \cup_{K\in T_h}\overline{K}$
- $K_i \cap K_j = \emptyset \quad \forall i, j,\, i \neq j$

- There exists a constant, ξ, independent of h, such that

$$\frac{h_K}{\rho_K} \leq \xi \quad \forall K \in T_h,$$

where ρ_K denotes the diameter of the largest inscribed circle of K.

The last condition is known as *shape-regularity.* We will also refer to the property of *quasi-uniformity*:

- There exists a constant, τ, satisfying,

$$\frac{h}{\rho_K} \leq \tau \quad \forall K \in T_h,$$

which will be a requirement in some parts of our analysis.

For any given element K we can now define the space $H(div;K)$. The following lemma gives conditions on the type of functions, defined piecewise on the elements of T_h, that can be used to approximate $\vec{v} \in H_{0,N}(div;\Omega)$.

Lemma 3 *Let $\vec{v} \in L^2(\Omega)^d$. $\vec{v} \in H_{0,N}(div;\Omega)$ if and only if,*

$$\vec{v}|_K \in H(div;K) \quad \forall K \in T_h, \tag{2.4}$$

and,

$$\sum_{K \in T_h} \langle \vec{v}\cdot\vec{n},\, w\rangle_{\partial K} = 0 \quad \forall w \in H^1_{0,D}(\Omega). \tag{2.5}$$

Proof We give an outline of the proof, as suggested in Proposition 1.1 of [26, Ch.3]. First, let $\vec{v} \in H_{0,N}(div;\Omega)$. Clearly condition (2.4) holds. Using the definition (2.2) and the integration by parts rule in Lemma 2, we have, for any $w \in H^1_{0,D}(\Omega)$,

$$0 \;=\; \langle \vec{v}\cdot\vec{n},\, w\rangle \;=\; \int_{\partial\Omega} \vec{v}\cdot\vec{n}\, w\, ds \;=\; \int_\Omega \nabla\cdot\,\vec{v}\, w\, d\Omega + \int_\Omega \vec{v}\cdot\nabla\, w\, d\Omega.$$

Breaking the integral into pieces and using (2.4) yields,

$$0 \;=\; \sum_{K \in T_h} \left(\int_K \nabla\cdot\,\vec{v}\, w\, dK + \int_K \vec{v}\cdot\nabla\, w\, dK\right) \;=\; \sum_{K \in T_h} \langle \vec{v}\cdot\vec{n},\, w\rangle_{\partial K}\,.$$

Hence condition (2.5) holds.

Conversely, let $\vec{v} \in L^2(\Omega)^d$ and suppose that conditions (2.4) and (2.5) hold. Since $\vec{v}|_K \in H(div;K)$ on each K, we can apply Lemma 2 on each subdomain. Thus, $\forall w \in H^1_{0,D}(\Omega)$ we have,

$$\int_K \nabla\cdot\,\vec{v}|_K\, w|_K\, dK + \int_K \vec{v}|_K\cdot\nabla\, w|_K\, dK \quad = \quad \int_{\partial K} (\vec{v}\cdot\vec{n})\,|_K\; w|_K\, dK.$$

Summing over K and imposing condition (2.5) yields,

$$\int_\Omega \nabla\cdot\,\vec{v}\,w\,d\Omega = -\int_\Omega \vec{v}\cdot\nabla\,w\,d\Omega \quad \forall w\,\in\, H^1_{0,D}\,(\Omega)\,. \tag{2.6}$$

It follows that,

$$\left|\int_\Omega \nabla\cdot\,\vec{v}\,w\,d\Omega\right| = \left|\int_\Omega \vec{v}\cdot\nabla\,w\,d\Omega\right| \;\le\; \|\,\vec{v}\,\|_0|\,w\,|_1\,d\Omega \quad \forall w\,\in\, H^1_{0,D}\,(\Omega)\,.$$

Since $\vec{v}\,\in\,L^2(\Omega)^d$, and $w\,\in\,H^1(\Omega)$, the right-hand side is bounded and we must have $\nabla\cdot\vec{v}\,\in\,L^2(\Omega)$. In that case, $\vec{v}\,\in\,H(div;\Omega)$ and so we can now apply Lemma 2 to the whole domain. This yields,

$$\int_\Omega \nabla\cdot\,\vec{v}\,w\,d\Omega = -\int_\Omega \vec{v}\cdot\nabla\,w\,d\Omega\,+\,\langle\vec{v}\cdot\vec{n},\,w\rangle \quad \forall w\,\in\, H^1_{0,D}\,(\Omega)\,. \tag{2.7}$$

Equating (2.7) with (2.6) gives, $\langle\vec{v}\cdot\vec{n},\,w\rangle\;=\;0$ for all $w\;\in\;H^1_{0,D}\,(\Omega)\,$. Hence $\vec{v}\;\in\;H_{0,N}(div;\Omega)$ as required. □

Remark 1 *Any $\vec{v}$ that satisfies $(\vec{v}\cdot\vec{n})\,|_{\partial\Omega_N}=0$ and that has continuous normal components at the interior interelement boundaries of T_h satisfies (2.5).*

We are now ready to formally state the model problem.

2.2 Mixed problem definition

Let Ω be a bounded domain in $I\!R^d$, $d=2,3$. Given $f\,\in\,L^2(\Omega)$ and $g\in H^{\frac{1}{2}}(\partial\Omega_D)$, we look for a solution $(\vec{u},p)$ to the first-order PDE system,

$$\begin{aligned} \mathcal{A}^{-1}\vec{u}-\nabla p &= 0, \\ \nabla\cdot\vec{u} &= -f \quad \text{in } \Omega, \\ p &= g \quad \text{on } \partial\Omega_D, \\ \vec{u}\cdot\vec{n} &= 0 \quad \text{on } \partial\Omega_N, \end{aligned} \tag{2.8}$$

where $\partial\Omega_D\neq\emptyset$ and $\mathcal{A}=\mathcal{A}(\vec{x})$ is a $d\times d$ bounded, symmetric and positive definite coefficient tensor with smallest eigenvalue bounded away from zero uniformly with respect to $\vec{x} \in \Omega$ so that (2.1) holds. It can be shown (see [26, p.134]), that a unique solution exists and corresponds to the solution of the saddle-point problem,

$$\inf_{\vec{v} \in H_{0,N}(div;\Omega)} \sup_{w \in L^2(\Omega)} \frac{1}{2} \int_\Omega \mathcal{A}^{-1} \vec{v} \cdot \vec{v} \, d\Omega + \int_\Omega (\nabla \cdot \vec{v} + f) \, w \, d\Omega + \int_{\partial \Omega_D} g \, \vec{v} \cdot \vec{n} ds. \tag{2.9}$$

Note that imposing non-homogeneous Neumann boundary conditions in (2.8) is straightforward. We restrict attention to the homogeneous case because it corresponds to a 'no-flow' condition which is common in the context of flow models.

2.2.1 Continuous variational problem

Now, we designate test spaces $V = H_{0,N}(div;\Omega)$ and $W = L^2(\Omega)$. Multiplying by arbitrary test functions in (2.8), integrating over Ω and imposing boundary conditions yields the continuous, mixed variational problem,

find $(\vec{u}, p) \in V \times W$ satisfying,

$$\begin{aligned} a(\vec{u}, \vec{v}) + b(\vec{v}, p) &= \langle g, \vec{v} \cdot \vec{n} \rangle_{\partial \Omega_D} \quad && \forall \vec{v} \in V, \\ b(\vec{u}, w) &= -(f, w) && \forall w \in W, \end{aligned} \tag{2.10}$$

where $a(\cdot, \cdot) : V \times V \to I\!R$, and $b(\cdot, \cdot) : V \times W \to I\!R$ are the continuous bilinear forms,

$$a(\vec{u}, \vec{v}) \; = \; \left(\mathcal{A}^{-1} \vec{u}, \vec{v} \right), \quad b(\vec{v}, w) \; = \; (\nabla \cdot \vec{v}, w) \, .$$

Note that $a(\cdot, \cdot)$ and $b(\cdot, \cdot)$ are also bounded. Applying the Cauchy-Schwarz inequality, we obtain,

$$| \, b(\vec{v}, w) \, | \leq \| \, \nabla \cdot \vec{v} \, \|_0 \| \, w \, \|_0 \leq \| \, \vec{v} \, \|_{div} \| \, w \, \|_0,$$

and since the coefficients in $\mathcal{A}^{-1}$ are assumed to be bounded, there exists a positive constant $C_a < \infty$, depending on $\mathcal{A}$, such that,

$$| \, a(\vec{u}, \vec{v}) \, | \leq \, C_a \, \| \, \vec{u} \, \|_0 \| \, \vec{v} \, \|_0 \leq \, C_a \, \| \, \vec{u} \, \|_{div} \| \, \vec{v} \, \|_{div} \, . \tag{2.11}$$

2.2.2 Discrete variational problem

Given a partition T_h of Ω into finite elements $\{K\}$, a standard, *conforming* finite element approximation consists of choosing finite dimensional subspaces $V_h \subset V$ and $W_h \subset W$ and solving, find $(\vec{u_h}, p_h) \in V_h \times W_h$ satisfying,

$$\begin{aligned} a(\vec{u_h}, \vec{v_h}) + b(\vec{v_h}, p_h) &= \langle g, \vec{v_h} \cdot \vec{n} \rangle_{\partial\Omega_D} \quad \forall \vec{v_h} \in V_h, \\ b(\vec{u_h}, w_h) &= -(f, w_h) \qquad \forall w_h \in W_h. \end{aligned} \tag{2.12}$$

Clearly, functions in V_h must satisfy the criteria of Lemma 3. Remark 1 says that we can achieve this by choosing functions which, in particular, have continuous normal components at interelement boundaries. Discontinuous pressure functions are admissible. However, the compatibility of V_h and W_h is critical. The variational problems (2.10) and (2.12) are examples of the generic saddle-point problem studied by Brezzi and Fortin in [26, Ch.2] and as such can be analysed for stability in the framework described there.

2.3 Stability analysis

In this section we recall the abstract theory of Brezzi, leading to conditions on the finite element spaces V_h and W_h that guarantee existence and uniqueness of a solution to (2.12). We begin with the continuous problem (2.10).

2.3.1 Continuous problem

Theorem 1 *Given continuous, bounded bilinear forms, $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$, the following two conditions are sufficient for the existence and uniqueness of the solution $(\vec{u}, p) \in V \times W$ to (2.10),*

1. *Z-ellipticity: there exists a constant $\alpha > 0$ such that,*

$$a(\vec{v}, \vec{v}) \geq \alpha \parallel \vec{v} \parallel_V^2, \quad \forall \vec{v} \in Z, \tag{2.13}$$

where $Z = \{\vec{v} \in V \mid b(\vec{v}, w) = 0 \quad \forall w \in W\}$.

2. *Compatibility condition: there exists a constant $\beta > 0$ such that,*

$$\sup_{\vec{v} \in V} \frac{b(\vec{v}, w)}{\parallel \vec{v} \parallel_V} \geq \beta \parallel w \parallel_W \quad \forall v \in W. \tag{2.14}$$

Proof See Brezzi [23] for the original exposition or Roberts and Thomas [82, pp.568–569]. □

Condition (2.14) is referred to as the inf-sup inequality since we may restate it as,

$$\inf_{w \in W} \sup_{\vec{v} \in V} \frac{|b(\vec{v}, w)|}{\| w \|_W \| \vec{v} \|_V} \geq \beta. \tag{2.15}$$

The natural choice of norms is $\| \cdot \|_V = \| \cdot \|_{div}$ and $\| \cdot \|_W = \| \cdot \|_0$. However, a different choice will be considered in Chapter 5.

A crucial first observation is that since $\nabla \cdot V \subset W$, the constraint space Z contains only divergence-free vectors. Thus, for any $\vec{v} \in Z$, $\| \vec{v} \|_{div} = \| \vec{v} \|_0$. Now, by condition (2.1),

$$a(\vec{v}, \vec{v}) \geq \gamma \| \vec{v} \|_0^2 = \gamma \| \vec{v} \|_{div}^2 \quad \forall \vec{v} \in Z, \tag{2.16}$$

and so (2.13) holds with $\alpha = \gamma > 0$. To establish inf-sup stability, we require the following result.

Lemma 4 *For any $f \in L^2(\Omega)$, there exists a $\vec{h} \in H_{0,N}(div; \Omega)$ and a positive constant C satisfying,*

$$\| \vec{h} \|_{div} \leq C \| f \|_0 . \tag{2.17}$$

Proof Given any $f \in L^2(\Omega)$, it is a standard result that there exists a unique solution $s \in H^1(\Omega)$ to the homogeneous mixed boundary value problem,

$$\begin{aligned} -\nabla \cdot (\nabla s) &= f \quad \text{in } \Omega \\ s &= 0 \quad \text{on } \partial\Omega_D \\ \nabla s \cdot \vec{n} &= 0 \quad \text{on } \partial\Omega_N. \end{aligned}$$

Choosing $\vec{h} = -\nabla s$ gives $\nabla \cdot \vec{h} = f$. Since $\nabla s \in L^2(\Omega)$ and $f \in L^2(\Omega)$, we have $\vec{h} \in H(div; \Omega)$. Using $\vec{h} \cdot \vec{n}|_{\partial\Omega_N} = 0$ yields $\vec{h} \in H_{0,N}(div; \Omega)$. The bound (2.17) then follows by Remark 1.1 in Brezzi and Fortin [26, p.136] □

It follows that for any $w \in L^2(\Omega)$ we can choose $\vec{h}$ in $H_{0,N}(div; \Omega)$ as the uniquely determined vector $\vec{h} = -\nabla s$, satisfying (2.17) with $f = w$. Then,

$$\sup_{\vec{v} \in V} \frac{b(\vec{v}, w)}{\| \vec{v} \|_{div}} \geq \frac{b(\vec{h}, w)}{\| \vec{h} \|_{div}} = \frac{(\nabla \cdot \vec{h}, w)}{\| \vec{h} \|_{div}} = \frac{\| w \|_0^2}{\| \vec{h} \|_{div}} \geq \frac{1}{C} \frac{\| w \|_0^2}{\| w \|_0} = \frac{1}{C} \| w \|_0 . \tag{2.18}$$

Hence, by Theorem 1, a unique solution $(\vec{u}, p) \in H_{0,N}(div; \Omega) \times L^2(\Omega)$ to the continuous variational problem (2.10) exists. Establishing existence and uniqueness of a solution to the discrete problem (2.12) is, however, somewhat more complicated.

2.3.2 Discrete problem

Theorem 2 *If the following conditions hold,*

1. *Z_h-ellipticity: there exists a constant $\alpha_h > \alpha_* > 0$, with α_* independent of h, satisfying,*

$$a(\vec{v_h}, \vec{v_h}) \geq \alpha_h \parallel \vec{v_h} \parallel_{V_h}^2, \quad \forall \vec{v_h} \in Z_h, \tag{2.19}$$

where $Z_h = \{\vec{v_h} \in V_h \mid b(\vec{v_h}, w_h) = 0 \quad \forall w_h \in W_h\}$,

2. *Discrete inf-sup inequality: there exists a constant $\beta_h > \beta_* > 0$, with β_* independent of h, satisfying,*

$$\sup_{\vec{v_h} \in V_h} \frac{b(\vec{v_h}, w_h)}{\parallel \vec{v_h} \parallel_{V_h}} \geq \beta_h \parallel w_h \parallel_{W_h} \quad \forall w_h \in W_h, \tag{2.20}$$

then there is a unique $(\vec{u_h}, p_h) \in V_h \times W_h$ satisfying (2.12) for each h. Further, there exists a constant C, depending on α_*, β_*, and C_a in (2.11) such that

$$\parallel \vec{u} - \vec{u_h} \parallel_{V_h} + \parallel p - p_h \parallel_{W_h} \leq C(\inf_{\vec{v_h} \in V_h} \parallel \vec{u} - \vec{v_h} \parallel_{V_h} + \inf_{w_h \in W_h} \parallel p - w_h \parallel_{W_h}). \tag{2.21}$$

Proof See [26, Ch.2] or [82]. □

We look for spaces V_h and W_h that satisfy the conditions of Theorem 2 with norms $\parallel \cdot \parallel_{V_h} = \parallel \cdot \parallel_V = \parallel \cdot \parallel_{div}$ and $\parallel \cdot \parallel_{W_h} = \parallel \cdot \parallel_W = \parallel \cdot \parallel_0$. However, this is not a trivial task. Theorem 3, below, offers a constructive way to establish the discrete inf-sup inequality for chosen spaces V_h and W_h.

Theorem 3 *Let $M \subset V$ such that M is dense in V. Suppose the continuous inf-sup inequality holds in M and W_h. That is, there exists a constant β_M, satisfying,*

$$\sup_{\vec{v} \in M} \frac{b(\vec{v}, w_h)}{\parallel \vec{v} \parallel_V} \geq \beta_M \parallel w_h \parallel_{W_h} \quad \forall w_h \in W_h. \tag{2.22}$$

Suppose further that there exists a family of uniformly continuous operators $\Pi_h : M \to V_h$ satisfying, for every $\vec{v} \in M$ and $w_h \in W_h$,

$$b(\Pi_h \vec{v} - \vec{v}, w_h) = 0, \tag{2.23}$$

$$\parallel \Pi_h \vec{v} \parallel_V \leq C \parallel \vec{v} \parallel_V, \tag{2.24}$$

where C is a constant independent of h. Then, the inf-sup inequality (2.20) is satisfied.

Proof For any $w_h \in W_h$, conditions (2.23)–(2.24) imply,

$$\begin{aligned}\sup_{\vec{v}_h \in V_h} \frac{b(\vec{v}_h, w_h)}{\| \vec{v}_h \|_V} &\geq \sup_{\vec{v} \in M} \frac{b(\Pi_h \vec{v}, w_h)}{\| \Pi_h \vec{v} \|_V} = \sup_{\vec{v} \in M} \frac{b(\Pi_h \vec{v} - \vec{v}, w_h) + b(\vec{v}, w_h)}{\| \Pi_h \vec{v} \|_V} \\ &= \sup_{\vec{v} \in M} \frac{b(\vec{v}, w_h)}{\| \Pi_h \vec{v} \|_V} \geq \sup_{\vec{v} \in M} \frac{1}{C} \frac{b(\vec{v}, w_h)}{\| \vec{v} \|_V}.\end{aligned}$$

Note that we used the fact that M is dense in V in the first step. The result follows from (2.22) with $\beta_h = \frac{\beta_M}{C}$. □

2.4 Raviart-Thomas approximation

A family of local spaces that can be used to construct a suitable subspace $V_h \subset H_{0,N}(div; \Omega)$ are proposed by Raviart and Thomas in [80] for $I\!R^2$ and by Nedelec in [72] for $I\!R^3$. We denote, for any element K, the set of polynomials of degree $\leq k$ by $P_k(K)$. In $I\!R^2$, $Q_{i,j}(K)$ is the set of polynomials of degree $\leq i$ in the x-component and $\leq j$ in the y-component. Using this notation, the Raviart-Thomas-Nedelec element spaces of degree $k \geq 0$ are,

$$RT_k(K) \quad = \quad (P_k(K))^d + \vec{x}\, P_k(K), \qquad d = 2, 3, \tag{2.25}$$

if K is a triangle or a tetrahedron, and

$$RT_k(K) \quad = \quad Q_{k+1,k}(K) \times Q_{k,k+1}(K), \tag{2.26}$$

if K is a rectangle in $I\!R^2$. For brick elements in $I\!R^3$, we have,

$$RT_k(K) \quad = \quad Q_{k+1,k,k}(K) \times Q_{k,k+1,k}(K) \times Q_{k,k,k+1}(K). \tag{2.27}$$

We shall only consider the lowest-order elements ($k = 0$), since for discontinuous coefficient tensors, $\mathcal{A}$, we do not obtain high solution regularity. Our numerical experiments will be performed in $I\!R^2$. In that case, functions $\vec{v} \in RT_0(K)$ have the special forms,

$$\vec{v}|_{\triangle} = \begin{pmatrix} a + cx \\ b + cy \end{pmatrix}, \quad \vec{v}|_{\square} = \begin{pmatrix} a + cx \\ b + dy \end{pmatrix}, \tag{2.28}$$

for triangles and rectangles, respectively. These functions are uniquely defined by the values of their normal components at the element boundaries. A proof of this is given in [80] for $I\!R^2$ and in [72] for $I\!R^3$. It is also easy to see that $RT_0(K) \subset H(div;K)$. Hence, in view of Lemma 3 and Remark 1, the element spaces $RT_0(K)$ are a natural choice for constructing the space $V_h \subset H_{0,N}(div;\Omega)$ in (2.12).

We now describe the construction process of spaces V_h and W_h and operators Π_h that satisfy the criteria of Theorem 3, for triangular and tetrahedral elements. The arguments are technical but necessary to establish existence and uniqueness results. To begin, we define the space,

$$M(K) \;=\; \{\vec{v} \,\in\, (L^s(K))^d \;\mid\; \nabla\cdot\vec{v} \,\in\, L^2(K)\,\}, \quad s>2,$$

and a local interpolation operator $\pi_K : M(K) \to RT_0(K)$, via,

$$\int_{\partial K} \vec{v}\cdot\vec{n}\,p_0\,ds \;\;=\;\; \int_{\partial K} \pi_K\vec{v}\cdot\vec{n}\,p_0\,ds \qquad \forall\, p_0 \,\in\, R_0(\partial K), \tag{2.29}$$

where $R_0(\partial K) \;=\; \big\{p_0 \,\in\, L^2(\partial K), \quad p_0|_e \,\in\, P_0(e)\,\forall\, e \in K\big\}$. For technical reasons, we cannot choose $M(K) \;=\; H(div;K)$. We require $s\,>\,2$ so that the integrals in (2.29) are well defined. A crucial property of π_K is that it commutes with p_K, the local L^2-projection operator acting on the space $\nabla\cdot RT_0(K)$. To be specific, we have the following result.

Lemma 5 *For all $\vec{v}\in M(K)$, $\nabla\cdot(\pi_K\vec{v}) \;=\; p_K\nabla\cdot\vec{v}$.*

Proof For any $w \in \nabla\cdot RT_0(K)$, we obtain, using Green's formula on K,

$$\int_K \nabla\cdot(\pi_K\vec{v}-\vec{v})\,w\,dK \;\;=\;\; \int_K (\vec{v}-\pi_K\vec{v})\cdot\nabla w\,dK \;+\; \int_{\partial K} w(\vec{v}-\pi_k\vec{v})\cdot\vec{n}\,ds.$$

By definition of $RT_0(K)$, $w \,\in\, P_0(K)$. Hence, $w|_K \,\in\, R_0(\partial K)$ and $\nabla w|_K \;=\; 0$. The right-hand side vanishes by the definition of π_K in (2.29). Combining this with the definition of the L^2-projection operator, p_K, yields,

$$\int_K w\,\nabla\cdot\pi_K\vec{v}\,dK \;\;=\;\; \int_K w\,\nabla\cdot\vec{v}\,dK \;=\; \int_K w\,p_K\nabla\cdot\vec{v}\,dK,\;\; \forall\, w\in\nabla\cdot RT_0(K).$$

Since $\nabla\cdot\pi_K\vec{v}\in\nabla\cdot RT_0(K)$, the result follows. □

This commutativity property is usually illustrated by the diagram shown in Fig. 2.1.

$$\begin{tikzcd}
M(K) \arrow[r, "div"] \arrow[d, "\pi_K"'] & L^2(K) \arrow[d, "p_K"] \\
RT_0(K) \arrow[r, "div"'] & \nabla \cdot RT_0(K)
\end{tikzcd}$$

Figure 2.1: Local commutativity property

Next, we define the global spaces,

$$\begin{aligned}
RT_0(\Omega; T_h) &= \{\vec{v} \in H(div;\Omega) \mid \vec{v}\,|_K \in RT_0(K)\ \forall\ K\ \in T_h\,\}, \\
W_h &= \{\, w \in L^2(\Omega) \mid w\,|_K \in \nabla \cdot RT_0(K)\ \forall\ K\ \in T_h\}, && (2.30) \\
M &= \{\,\vec{v}\ \in H_{0,N}(div;\Omega) \mid \vec{v}\ \in\ (L^s(\Omega))^d\}, \quad s > 2, && (2.31)
\end{aligned}$$

and construct a global interpolation operator, $\Pi_h : M \to RT_0(\Omega; T_h)$, via

$$(\Pi_h \vec{v})|_K \;=\; \pi_K(\vec{v}|_K).$$

If we choose P_h to be the L^2-projection operator on W_h, then we can derive the global commutativity property $\nabla \cdot (\Pi_h \vec{v}) \,=\, P_h \nabla \cdot \vec{v}$.

Returning to the approximation problem (2.12), we now choose finite-dimensional subspaces,

$$V_h = \{\,\vec{v} \in\ RT_0(\Omega; T_h) \text{ and } \vec{v} \cdot \vec{n}|_{\partial\Omega_N} = 0\,\}, \tag{2.32}$$

and W_h as in (2.30) which is equivalent to,

$$W_h = \{\, w \in L^2(\Omega) \mid w\,|_K \in\ P_0(K) \quad \forall K \in T_h\}. \tag{2.33}$$

To establish a unique solution to (2.12), we require that the chosen V_h, W_h and Π_h satisfy condition (2.19) in Theorem 2 and the criteria of Theorem 3 with M defined in (2.31) and $\| \cdot \|_V = \| \cdot \|_{V_h} = \| \cdot \|_{div}$ and $\| \cdot \|_{W_h} = \| \cdot \|_0$. Since $\nabla \cdot V_h = W_h$, Z_h-ellipticity certainly holds. It can be shown that the vector $\vec{h}$ used to establish inf-sup stability for the continuous problem in Lemma 4 actually belongs to M for some fixed $s > 2$.

Hence (2.18) holds with V replaced by M. Since $W_h \subset W$, the inf-sup inequality (2.22) also holds. It remains only to verify the conditions (2.23)–(2.24) on Π_h. Using the definition of P_h, and the global commutativity property, for any $w_h \in W_h$ and $\vec{v} \in M$, we have,

$$\begin{aligned} b(\Pi_h \vec{v} - \vec{v}, w_h) &= \int_\Omega \nabla \cdot \Pi_h \vec{v}\, w_h \, d\Omega - \int_\Omega \nabla \cdot \vec{v}\, w_h \, d\Omega \\ &= \int_\Omega P_h \nabla \cdot \vec{v}\, w_h \, d\Omega - \int_\Omega \nabla \cdot \vec{v}\, w_h \, d\Omega \;=\; 0. \end{aligned}$$

By definition (2.29), the local operator π_K in (2.29) is bounded. Hence, by construction, Π_h is a also a bounded operator, from M to V_h. Property (2.24) follows. The reader is referred to [26, Ch.3] for full technical details.

Hence, for triangular and tetrahedral elements, with V_h and W_h defined in (2.32) and (2.33), respectively, there exists a unique solution $(\vec{u}, p) \in V_h \times W_h$ to (2.12). The same construction applies to rectangular elements, provided we choose V_h as in (2.32) with $RT_0(K)$ defined in (2.26) and

$$W_h \;=\; \{ w \in L^2(\Omega) \mid w \,|_K \in \nabla \cdot RT_0(K) \quad \forall K \in T_h \}.$$

The degrees of freedom for the lowest order elements in $I\!R^2$ are illustrated in Fig. 2.2. Normal components of velocities are sampled at edge midsides. The piecewise constant pressure approximation is sampled at element centroids.

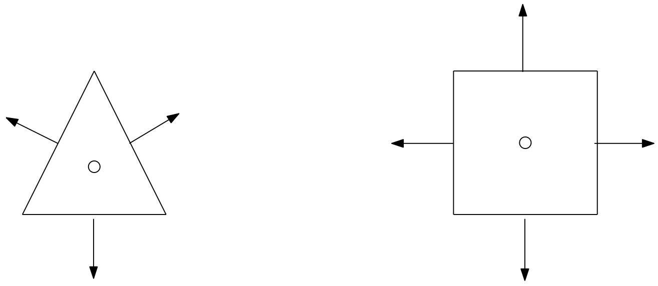

Figure 2.2: Degrees of freedom for $RT_0(K)$ in $I\!R^2$.

2.4.1 Error estimates

Error estimates for the resulting solutions $\vec{u}_h \in V_h$ and $p_h \in W_h$ to (2.12) can be derived from the following best approximation property.

Lemma 6 *Let $(\vec{u}, p)$ and $(\vec{u_h}, p_h)$ be the solutions to (2.10) and (2.12) respectively, with V_h, W_h chosen as in (2.32) and (2.33). There exists a constant C independent of h such that,*

$$\begin{aligned} \| \vec{u} - \vec{u_h} \|_{div} &\leq C \inf_{\vec{v_h} \in V_h} \| \vec{u} - \vec{v_h} \|_{div}, \\ \| p - p_h \|_0 &\leq C \left(\inf_{\vec{v_h} \in V_h} \| \vec{u} - \vec{v_h} \|_{div} + \inf_{w_h \in W_h} \| \ p - w_h \|_0 \right), \end{aligned}$$

Proof See Brezzi and Fortin [26, Ch.4].

Bounds can be derived in terms of $h = \max_K h_K$ by exploiting properties of the interpolation operators Π_h and P_h. Observe that,

$$\inf_{w_h \in W_h} \| \ p - w_h \|_0 \ \leq \ \| p - P_h p \|_0 \leq \left(\sum_K \| p - p_K p \|_{0,K} \right),$$

and if $\vec{u} \in M = H_{0,N}(div; \Omega) \cap L^s(\Omega)^d$ with $s > 2$,

$$\begin{aligned} \inf_{\vec{v_h} \in V_h} \| \vec{u} - \vec{v_h} \|_{div} &\leq \| \vec{u} - \Pi_h \vec{u} \|_{div} \\ &= \left(\sum_K \| \vec{u} - \pi_K \vec{u} \|_{0,K}^2 + \| \nabla \cdot \vec{u} - p_K \nabla \cdot \vec{u} \|_{0,K}^2 \right)^{\frac{1}{2}}. \end{aligned}$$

Provided we use an affine, shape regular mesh T_h, so that there exists an affine invertible mapping from each K onto a reference element K^*, we obtain,

$$\begin{aligned} \| \vec{u} - \pi_k \vec{u} \|_{0,K} &\leq C h_K \mid \vec{u} \mid_{1,K}, \\ \| p - p_k p \|_{0,K} &\leq C h_K \mid p \mid_{1,K}. \end{aligned}$$

Combining all of this information leads to the following standard result.

Lemma 7 *Let $(\vec{u}, p)$ be the solution to (2.10) and $(\vec{u_h}, p_h)$ be the solution to (2.12) with V_h and W_h defined in (2.32) and (2.33). Then,*

$$\begin{aligned} \| \vec{u} - \vec{u_h} \|_0 &\leq C h \left(\mid \vec{u} \mid_1 + \mid \nabla \cdot \vec{u} \mid_1 \right) \\ \| p - p_h \|_0 &\leq C h \left(\mid \vec{u} \mid_1 + \mid \nabla \cdot \vec{u} \mid_1 + \mid p \mid_1 \right), \end{aligned}$$

where C is a generic constant independent of h. □

The above estimates assume, then, that $\vec{u} \in H^1(\Omega)^d$, $\nabla \cdot \vec{u} \in H^1(\Omega)$ and $p \in H^1(\Omega)$. However, the regularity of these variables depends on the convexity of Ω and the continuity of the coefficient term (see [82, pp.582–583] for a discussion). If Ω is convex, and $\mathcal{A}$ is continuous, we have the following refined estimate, established by Falk and Osborn in [50].

Lemma 8 *Let $(\vec{u}, p)$ and $(\vec{u_h}, p_h)$ be the solutions to (2.10) and (2.12) with V_h and W_h defined in (2.32) and (2.33). Then,*

$$\| \vec{u} - \vec{u_h} \|_0 \leq Ch \| \vec{u} \|_1, \quad \| p - p_h \|_0 \leq Ch \| p \|_2,$$

where C is a generic constant independent of h.

Hence, for the lowest order schemes we obtain $O(h)$ estimates for both the velocity and pressure approximations.

2.4.2 Algebraic system

To see that (2.12) yields a linear algebra problem of the form (1.5), it is convenient to introduce linear operators $\boldsymbol{A} : V_h \to V_h$ and $\boldsymbol{B} : V_h \to W_h$, defined via,

$$\begin{aligned} (\boldsymbol{A}\vec{v_h}, \vec{q_h}) &= a(\vec{v_h}, \vec{q_h}) \qquad \forall \, \vec{v_h}, \vec{q_h} \in V_h, \\ (\boldsymbol{B}\vec{v_h}, w_h) &= b(\vec{v_h}, w_h) \qquad \forall \, \vec{v_h} \in V_h, \forall \, w_h \in W_h, \end{aligned}$$

and the adjoint operator $\boldsymbol{B^T} : W_h \to V_h$, satisfying,

$$\left(\vec{v_h}, \boldsymbol{B}^T w_h\right) = (\boldsymbol{B}\vec{v_h}, w_h) \quad \forall \, \vec{v_h} \in V_h, \forall \, w_h \in W_h.$$

Now (2.12) becomes,

$$\underbrace{\begin{pmatrix} \boldsymbol{A} & \boldsymbol{B^T} \\ \boldsymbol{B} & 0 \end{pmatrix}}_{\boldsymbol{C}} \begin{pmatrix} \vec{u_h} \\ p_h \end{pmatrix} = \begin{pmatrix} g_h \\ -f_h \end{pmatrix}, \tag{2.34}$$

where $f_h = P_h f$ and $g_h = \langle g, \vec{v_h} \cdot \vec{n} \rangle_{\partial\Omega_D}$. To realise (2.34) on a computer, we choose basis sets,

$$V_h = \ \mathtt{span}\{\vec{\varphi_i}\}_{i=1}^n, \quad W_h = \ \mathtt{span}\{\phi_j\}_{j=1}^m,$$

and construct the finite element matrices $A \in I\!\!R^{n \times n}$ and $B \in I\!\!R^{m \times n}$ via,

$$A_{ij} = \left(\mathcal{A}^{-1}\vec{\varphi_i}, \vec{\varphi_j}\right), \quad i, j = 1 : n, \tag{2.35}$$

$$B_{kj} = \left(\nabla \cdot \vec{\varphi_j}, \phi_k\right), \quad k = 1 : m, \, j = 1 : n. \tag{2.36}$$

Expanding the discrete solution variables in the chosen basis sets now yields,

$$(\vec{u_h}, p_h) \;=\; \left(\sum_{i=1}^{n} u_i \vec{\varphi_i} \,,\, \sum_{j=1}^{m} p_j \phi_j \right),$$

where $\underline{u} = [u_1, \ldots, u_n]^T$ and $\underline{p} = [p_1, \ldots, p_m]^T$ are the vectors of coefficients satisfying,

$$\underbrace{\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix}}_{C} \begin{pmatrix} \underline{u} \\ \underline{p} \end{pmatrix} = \begin{pmatrix} \underline{g} \\ \underline{f} \end{pmatrix}. \tag{2.37}$$

The right-hand side vectors $\underline{f} \in I\!\!R^m$ and $\underline{g} \in I\!\!R^n$ are constructed via,

$$\begin{aligned} g_i &= \langle g, \vec{\varphi_i} \cdot \vec{n} \rangle_{\partial\Omega_D} \qquad i = 1 : n, \\ f_k &= -(f, \phi_k) \qquad k = 1 : m. \end{aligned}$$

We call A the weighted velocity mass matrix and B is a discrete representation of the divergence operator. A is positive definite since (2.1) holds and the inf-sup inequality ensures that $rank(B) = m$ so that $null(B^T) = \{\underline{0}\}$. Combining these properties, it is easy to show that the coefficient matrix C in (2.37) is non-singular.

2.4.3 System assembly

We now give a brief description of the assembly procedure for the mixed finite element system (2.37). For any $w_h \in W_h$, defined in (2.33),

$$w_h(x, y) = \sum_{i=1}^{m} w_i \phi_i,$$

where $w_i = w_h(x_{c_i}, y_{c_i})$ is the value of w_h at the centroid of the ith element. The scalar global basis function ϕ_i is the characteristic function satisfying,

$$\phi_i \;=\; \begin{cases} 1 \text{ in element } K_i, & i = 1 : m = \#\text{elements} \\ 0 \text{ elsewhere.} \end{cases} \tag{2.38}$$

We fix a set of oriented normal vectors $\vec{\nu^i}$ to each edge (in $I\!\!R^2$, or face in $I\!\!R^3$) e_i of T_h. For uniform triangular meshes in $I\!\!R^2$ we shall use the orientation shown in Fig. 2.3. For any given domain, the orientation can be chosen arbitrarily. We denote by $\vec{n}^i_K$ the set of unit *outward* normal vectors at the edges, or faces, of element K and set

$$s^i_K = \begin{cases} +1 & \text{if } \vec{n}^i_K = \vec{\nu}^i_K, \\ -1 & \text{if } \vec{n}^i_K = -\vec{\nu}^i_K. \end{cases}$$

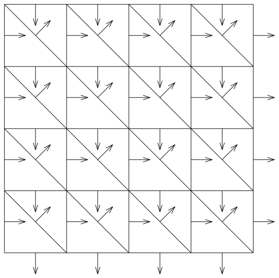

Figure 2.3: Possible configuration of global normal vectors.

In each K, $\vec{u}_h \in V_h$ has the local expansion,

$$\vec{u_h}(x,y)|_K = \sum_j \alpha_j \, \vec{\varphi}_j^K,$$

where $\alpha_j = \vec{u_h}|_K \cdot \vec{\nu}_K^j$, and the index j runs over the edges (faces) of K. Globally,

$$\vec{u}_h(x,y) = \sum_{i=1}^{n} u_i \, \vec{\varphi}_i,$$

where the index n runs over all the edges or faces of $T_h \backslash \partial\Omega_N$. We choose the basis functions to satisfy,

$$\vec{\varphi}_i \cdot \vec{\nu}^k = \begin{cases} 1 \text{ if } i = k, \\ 0 \text{ if } i \neq k \end{cases} \quad i, k = 1 : n, \tag{2.39}$$

so that the degrees of freedom are $u_i = \vec{u_h} \cdot \vec{\nu}^i$.

Remark 2 *Some authors write degrees of freedom for $\vec{u}_h$ in integral form. That is, the basis functions are chosen to satisfy,*

$$\int_{e_i} \vec{\varphi}_i \cdot \vec{\nu}^k \, ds \; = \begin{cases} 1 & \textit{if } k = i, \\ 0 & \textit{if } k \neq i, \end{cases} \tag{2.40}$$

so that $u_i = \int_{e_i} \vec{u_h} \cdot \vec{\nu}^i \, ds$, where e_i denotes the ith edge or face of T_h. All of the algebraic properties of the matrices A and B in (2.37) that we will derive correspond to the choice (2.39).

Now, using (2.36) we have,

$$B_{rj} \;=\; \int_{K_r} \nabla \cdot \vec{\varphi}_j \, dK_r \;=\; \int_{\partial K_r} \vec{\varphi}_j \cdot \vec{n}_K \, dK_r \;=\; \sum_e \int_e \vec{\varphi}_j \cdot \vec{n}_K^e \, ds,$$

where, in $I\!R^2$, $\vec{n}_K^e$ is the unit *outward* normal to edge e of K_r. Thus,

$$B_{rj} \;=\; \begin{cases} 0 & \text{if } e_j \notin K_r, \\ s^j_{K_r} \mid e_j \mid & \text{if } e_j \in K_r. \end{cases} \tag{2.41}$$

For the righthand side,

$$f_r \;=\; -\int_{K_r} f\, dK_r, \quad g_i \;=\; \begin{cases} 0 & \text{if } e_i \notin \partial\Omega_D, \\ \int_{e_i} g\, ds & \text{if } e_i \in \partial\Omega_D. \end{cases}$$

The matrix A is constructed from the element contributions,

$$A^K_{ij} \;=\; \int_K \mathcal{A}^{-1}|_K \vec{\varphi}^K_i \cdot \vec{\varphi}^K_j \, dK, \tag{2.42}$$

where the indices i, j run over the edges or faces of K. If the entries of the coefficient tensor $\mathcal{A}$ are variable, then we do not perform the integration exactly. Rather, we approximate the coefficients by piecewise constant functions. Thus, in $I\!R^2$, we write,

$$\mathcal{A}|_K \;=\; \begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{pmatrix}, \tag{2.43}$$

to denote $\mathcal{A}$ evaluated at the centroid of K. Making this assumption, the integrand in (2.42) involves only quadratic or lower order terms and can be performed exactly.

2.4.4 Structure of the weighted mass matrix

Next, we derive some algebraic properties of the element matrices A^K in $I\!R^2$, relative to the structure of the averaged coefficient tensor $\mathcal{A}$ in (2.43). Consider, first, the reference triangle K^* in Fig. 2.4, aligned with the co-ordinate axis (r, s).

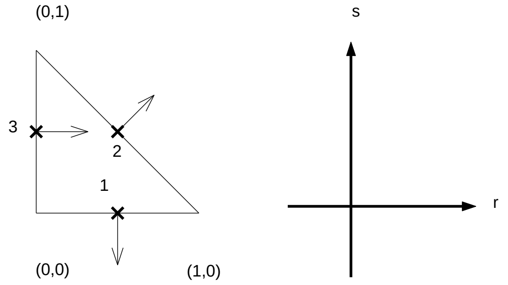

Figure 2.4: Reference triangle

If we fix oriented, unit, normal vectors,

$$\vec{\nu}^1_{K_*} = \begin{pmatrix} 0 \\ -1 \end{pmatrix}, \quad \vec{\nu}^2_{K_*} = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}, \quad \vec{\nu}^3_{K_*} = \begin{pmatrix} 1 \\ 0 \end{pmatrix},$$

at the edges, then the reference element basis functions are,

$$\vec{\varphi}^1_{K^*} = \begin{pmatrix} r \\ -1+s \end{pmatrix}, \quad \vec{\varphi}^2_{K^*} = \begin{pmatrix} \sqrt{2}r \\ \sqrt{2}s \end{pmatrix}, \quad \vec{\varphi}^3_{K^*} = \begin{pmatrix} 1-r \\ -s \end{pmatrix},$$

based on the orientation signs $s^1_{K^*} = +1, s^2_{K^*} = +1, s^3_{K^*} = -1$. On this element, integration yields,

$$A^{K^*} = \frac{1}{12det(\mathcal{A})} \begin{pmatrix} a_{22}+3a_{11}+3a_{12} & \sqrt{2}\left(a_{22}-a_{11}+a_{12}\right) & a_{22}+a_{11}+3a_{12} \\ \sqrt{2}\left(a_{22}-a_{11}+a_{12}\right) & 2\left(a_{22}+a_{11}-a_{12}\right) & \sqrt{2}\left(a_{22}-a_{11}-a_{12}\right) \\ a_{22}+a_{11}+3a_{12} & \sqrt{2}\left(a_{22}-a_{11}-a_{12}\right) & 3a_{22}+a_{11}+3a_{12} \end{pmatrix},$$

where $det(\mathcal{A}) = a_{11}a_{22} - a_{12}^2 > 0$. For diagonal coefficient tensors, this simplifies to,

$$A^{K^*} = \begin{pmatrix} \frac{1}{12a_{11}} + \frac{1}{4a_{22}} & \frac{\sqrt{2}}{12}\left(\frac{1}{a_{11}} - \frac{1}{a_{22}}\right) & \frac{1}{12a_{11}} + \frac{1}{12a_{22}} \\ \frac{\sqrt{2}}{12}\left(\frac{1}{a_{11}} - \frac{1}{a_{22}}\right) & \frac{1}{6a_{11}} + \frac{1}{6a_{22}} & \frac{\sqrt{2}}{12}\left(\frac{1}{a_{11}} - \frac{1}{a_{22}}\right) \\ \frac{1}{12a_{11}} + \frac{1}{12a_{22}} & \frac{\sqrt{2}}{12}\left(\frac{1}{a_{11}} - \frac{1}{a_{22}}\right) & \frac{1}{4a_{11}} + \frac{1}{12a_{22}} \end{pmatrix},$$

and if $\mathcal{A}$ is a constant or a scalar function, with $a_{11} = a_{22} = k$, we obtain,

$$A^{K^*} = \begin{pmatrix} \frac{1}{3k} & 0 & \frac{1}{6k} \\ 0 & \frac{1}{3k} & 0 \\ \frac{1}{6k} & 0 & \frac{1}{3k} \end{pmatrix}.$$

For general right-angled triangular elements of edge length h_K we obtain $A^K = h_K^2 A^{K^*}$. Stencils for A^K on equilateral triangles with $\mathcal{A} = 1$ are given in [76]. The important observation is that for triangles of all types, all of the coefficients influence all of the rows of A^K. This also occurs in $\mathbb{R}^3$.

Now consider the reference square K^* in Fig. 2.5. Fixing normal vectors,

$$\vec{\nu}^1_{K^*} = \vec{\nu}^2_{K^*} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \vec{\nu}^3_{K^*} = \vec{\nu}^4_{K^*} = \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

integration yields element basis functions,

$$\vec{\varphi}^1_{K^*} = \begin{pmatrix} \frac{1}{2} - \frac{1}{2}r \\ 0 \end{pmatrix}, \vec{\varphi}^2_{K^*} = \begin{pmatrix} \frac{1}{2} + \frac{1}{2}r \\ 0 \end{pmatrix} \vec{\varphi}^3_{K^*} = \begin{pmatrix} 0 \\ \frac{1}{2} - \frac{1}{2}s \end{pmatrix}, \vec{\varphi}^4_{K^*} = \begin{pmatrix} 0 \\ \frac{1}{2} + \frac{1}{2}s \end{pmatrix},$$

based on the orientation signs, $s^1_{K^*} = s^3_{K^*} = -1, s^2_{K^*} = s^4_{K^*} = +1$.

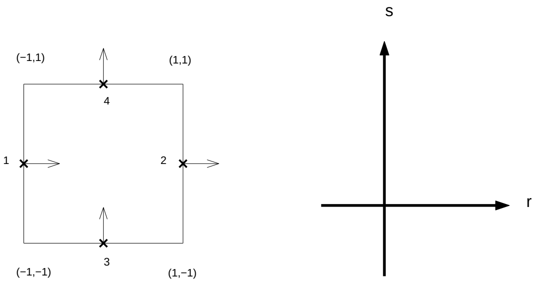

Figure 2.5: Reference square

The key point here is that we obtain two sets of mutually orthogonal basis functions $\{\vec{\varphi}^1_{K^*}, \vec{\varphi}^2_{K^*}\}$ and $\{\vec{\varphi}^3_{K^*}, \vec{\varphi}^4_{K^*}\}$. This will always be true for rectangles with edges aligned with the co-ordinate axes.

On the reference square, we find that,

$$A^{K^*} = \frac{1}{det(\mathcal{A})} \begin{pmatrix} \frac{4}{3}a_{22} & \frac{2}{3}a_{22} & -a_{12} & -a_{12} \\ \frac{2}{3}a_{22} & \frac{4}{3}a_{22} & -a_{12} & -a_{12} \\ -a_{12} & -a_{12} & \frac{4}{3}a_{11} & \frac{2}{3}a_{11} \\ -a_{12} & -a_{12} & \frac{2}{3}a_{11} & \frac{4}{3}a_{11} \end{pmatrix}$$

which simplifies, for diagonal coefficient tensors to,

$$A^{K^*} = \begin{pmatrix} \frac{4}{3}\frac{1}{a_{11}} & \frac{2}{3}\frac{1}{a_{11}} & 0 & 0 \\ \frac{2}{3}\frac{1}{a_{11}} & \frac{4}{3}\frac{1}{a_{11}} & 0 & 0 \\ 0 & 0 & \frac{4}{3}\frac{1}{a_{22}} & \frac{2}{3}\frac{1}{a_{22}} \\ 0 & 0 & \frac{2}{3}\frac{1}{a_{22}} & \frac{4}{3}\frac{1}{a_{22}} \end{pmatrix}. \tag{2.44}$$

Here, and for general rectangles, A^K has diagonal blocks. The crucial observation is that each one is scaled with respect to a *different* coefficient. Thus the impact of anisotropy on the structure and conditioning of A^K is entirely different for triangles and rectangles.

2.4.5 Eigenvalues of unpreconditioned system

To see that (2.37) requires a preconditioner, we now require bounds on the eigenspectra of C for our *specific* choice of V_h and W_h. Rusten and Winther established the following generic eigenvalue bound in [86].

Lemma 9 *Let $0 < \mu_1 \ldots \leq \mu_n$ be the eigenvalues of A and $0 < \sigma_1 \ldots \leq \sigma_m$ be the singular values of B. The eigenvalues of C in (2.37) lie in the intervals,*

$$\left[\frac{1}{2}\left(\mu_1 - \sqrt{\mu_1^2 + 4\sigma_m^2}\right), \frac{1}{2}\left(\mu_n - \sqrt{\mu_n^2 + 4\sigma_1^2}\right)\right] \cup \left[\mu_1, \frac{1}{2}\left(\mu_n + \sqrt{\mu_n^2 + 4\sigma_m^2}\right)\right]. \tag{2.45}$$

Proof See Lemma 2.1 in [86]. □

For a more explicit statement, we need bounds for μ_1, μ_n, σ_1 and σ_m that reflect the role of the discretisation parameter h and the coefficient tensor $\mathcal{A}$. The influence of h, on the condition number $\kappa(C)$, has been established by Scheichl in [89]. Below, we use the same arguments to establish the role of the constants γ and Γ from (2.1), as well as h, in (2.45). We reiterate that it is not $\kappa(C)$ that ultimately determines the success of the MINRES iteration but rather the *clustering* of the eigenvalues.

Remark 3 *In the sequel, c or c_i for some integer i, will always denote a positive generic constant, independent of h and the coefficients γ and Γ appearing in (2.1).*

The starting point is the following fundamental bounds for the L^2 norms of arbitrary pressure and velocity functions in the lowest-order spaces V_h and W_h.

Lemma 10 *Let $w_h \in W_h$ and $\vec{v}_h \in V_h$, with W_h and V_h defined in (2.33) and (2.32), respectively. Let $\underline{w}$ and $\underline{v}$ denote their vector expansions in the bases defined in (2.38) and (2.39). Then, for some constants c_1, c_2, c_3, c_4,*

$$c_1 h_{min}^d \underline{v}^T \underline{v} \quad \leq \quad \| \vec{v}_h \|_0^2 \leq c_2 h^d \underline{v}^T \underline{v}, \tag{2.46}$$

$$c_3 h_{min}^d \underline{w}^T \underline{w} \quad \leq \quad \| w_h \|_0^2 \leq c_4 h^d \underline{w}^T \underline{w}. \tag{2.47}$$

Proof The bound (2.46) is established by Scheichl in [89, Appendix A], and exploits an earlier result obtained by Hiptmair, [60], for a different choice of degrees of freedom.

The second bound follows immediately from,

$$\| w_h \|_0^2 = \sum_K \int_K \left(\sum_{i=1}^m w_i \phi_i \right)^2 dK,$$

and the fact that the basis functions ϕ_i are just characteristic functions satisfying $\phi_i = 1$ in K_i and $\phi_i = 0$ elsewhere. Thus,

$$\| w_h \|_0^2 = \sum_{j=1}^{m} \int_{K_j} w_j^2 \; dK_j, = \sum_{j=1}^{m} w_j^2 \mid K_j \mid,$$

where $\mid K_j \mid$ denotes the area or volume of K_j. Since we have assumed T_h is shape-regular, we easily obtain (2.47). □

By definition (2.1), we have,

$$\gamma \; \| \vec{v} \|_0^2 \leq \underline{v}^T A \underline{v} \; \leq \; \Gamma \; \| \vec{v} \|_0^2,$$

and so the following result is an obvious consequence of (2.46).

Lemma 11 *The eigenvalues μ_1 and μ_n of A satisfy,*

$$c_1 \gamma \, h_{min}^d \leq \mu_1 \; \leq \; \mu_n \leq c_2 \Gamma h^d. \tag{2.48}$$

Remark 4 *Lemma 11 indicates that $\kappa(A)$ is only independent of h if $h_{min} \approx h$. Moreover, $\kappa(A)$ depends on $\frac{\Gamma}{\gamma}$. Thus, A is not necessarily well-conditioned as is often assumed (cf. Rusten et al. [86], [87]) and cannot in general be approximated by a scaled identity matrix.*

Lemma 12 *The minimum and maximum singular values σ_1 and σ_m of B satisfy*

$$c_3 h_{min}^d \; \leq \; \sigma_1 \; \leq \; \sigma_m \; \leq \; c_4 h^{d-1}. \tag{2.49}$$

Proof See Scheichl [89, Proposition 2.26].

Combining Lemmas 11 and 12 with Rusten and Winther's standard result in Lemma 9 now yields the following eigenvalue bound.

Theorem 4 *If $\mu_n \; \leq \; 2\sigma_1$, the eigenvalues of the saddle-point system C in (2.37), arising in the lowest-order Raviart-Thomas approximation of (2.12), lie in the union of the intervals,*

$$\left[-c_4 h^{d-1} \, , \, \frac{1}{2} \left(1 - \sqrt{2} \right) c_1 \gamma h_{min}^d \right] \cup \left[c_1 \gamma h_{min}^d, \, \left(c_2 \Gamma h + c_4 \right) h^{d-1} \right].$$

Alternatively, if $\mu_n \geq 2\sigma_1$, we may re-state the eigenvalue bound as,

$$\left[-c_4 h^{d-1} \, , \, -\frac{\pi}{4} \frac{c_3^2 h_{min}^{2d}}{c_2 \Gamma \, h^d} \right] \cup \left[c_1 \gamma h_{min}^d, \, \left(c_2 \Gamma h + c_4 \right) h^{d-1} \right].$$

Proof The lower bound for the positive eigenvalues is obvious. The upper bound follows from $\frac{1}{2}\left(\mu_n + \sqrt{\mu_n^2 + 4\sigma_m^2}\right) \le \mu_n + \sigma_m$. For the negative eigenvalues it is clear that $\frac{1}{2}\left(\mu_1 - \sqrt{\mu_1^2 + 4\sigma_m^2}\right) \ge -\sigma_m \ge -c_4 h^{d-1}$ but for the upper bound we must now proceed as in [89, Theorem 2.27] and distinguish two cases.

If $\mu_n \le 2\sigma_1$, we obtain,

$$\frac{1}{2}\left(\mu_n - \sqrt{\mu_n^2 + 4\sigma_1^2}\right) \le \frac{1}{2}\left(\mu_n - \sqrt{2\mu_n^2}\right) = \frac{1}{2}\left(1 - \sqrt{2}\right)\mu_n \le \frac{1}{2}\left(1 - \sqrt{2}\right)\mu_1,$$

and the first result follows. On the other hand, suppose that $\mu_n \ge 2\sigma_1$. If we substitute $\tan\theta = \frac{2\sigma_1}{\mu_n}$, rearranging gives,

$$\frac{1}{2}\left(\mu_n - \sqrt{\mu_n^2 + 4\sigma_1^2}\right) = \frac{\mu_n}{2}\left(1 - \sqrt{1 + \tan^2\theta}\right).$$

By manipulating the identity,

$$\tan\theta = \frac{2\tan\frac{\theta}{2}}{1 - \tan^2\frac{\theta}{2}},$$

we obtain, $1 - \sqrt{1 + \tan^2\theta} = -\tan\theta\tan\frac{\theta}{2}$, and so,

$$\frac{1}{2}\left(\mu_n - \sqrt{\mu_n^2 + 4\sigma_1^2}\right) = -\sigma_1 \tan\frac{\theta}{2}.$$

Since $0 \le \tan\theta \le 1$ and $\tan\frac{\theta}{2} > 0$, we have $\tan\frac{\theta}{2} \le \tan\frac{\pi}{8}$. Hence,

$$\frac{\tan\frac{\theta}{2}}{\tan\theta} = \frac{1 - \tan^2\frac{\theta}{2}}{2} \ge \frac{1 - \tan^2\frac{\pi}{8}}{2} = \tan\frac{\pi}{8} > \frac{\pi}{8},$$

and so, finally,

$$\frac{1}{2}\left(\mu_n - \sqrt{\mu_n^2 + 4\sigma_1^2}\right) < -\sigma_1 \tan\theta\frac{\pi}{8} = -\frac{\pi}{4}\frac{\sigma_1^2}{\mu_n} \le -\frac{\pi}{4}\frac{c_3^2\, h_{min}^{2d}}{\mu_n} \le -\frac{\pi}{4}\frac{c_3^2 h_{min}^{2d}}{c_2 \Gamma\, h^d}. \quad \square$$

Remark 5 *Note that the condition $\mu_n \le 2\sigma_1$ is a condition on the magnitude of the* PDE *coefficients. On uniform meshes, the first bound in Theorem 4 is only likely to be tight if* $\Gamma << 1$.

The following Corollary is an obvious consequence of Theorem 4.

Corollary 1 *In $I\!R^2$, with uniform meshes, and unit coefficients, the eigenvalues of the saddle-point system C in (2.37), arising in the lowest-order Raviart-Thomas approximation of (2.12), lie in the union of the intervals,*

$$\left[ah\,,\,bh^2\right] \cup \left[ch^2,\,dh\right], \tag{2.50}$$

where a and b are negative constants and c and d are positive constants.

It is now clear why we require preconditioners for (2.37), even for trivial coefficients.

2.5 Preconditioning strategies

We conclude this chapter with a brief review of some existing preconditioning strategies for the saddle-point system (2.37). The fields of domain decomposition (see Chan [32]) and multigrid (see [95] or [29]) offer cheap, practical schemes but are mature for symmetric and positive definite problems only. Thus, many positive definite reformulations of the model problem, both at the PDE level and at the matrix level, have been suggested.

The most obvious way to achieve this is to eliminate the velocity variable to obtain the Schur complement problem,

$$BA^{-1}B^{T}\underline{p} = BA^{-1}\underline{g} - \underline{f}, \tag{2.51}$$

which can be solved with CG. This requires multiplication with $S = BA^{-1}B^{T}$ and thus computation of the action of A^{-1} at each iteration. In the special case of diagonal coefficients and rectangular finite elements, we have seen that A has a special block-diagonal structure. Solving for A^{-1} directly is advocated in [85], [48], [64] and [49]. However, this approach is infeasible for general meshes and general coefficients. In general, an inner iteration is required to approximate the action of A^{-1}. The success of the nested iteration depends on the selection of inner and outer stopping tolerances. In [86], Rusten and Winther demonstrate that convergence is highly sensitive to the choice of these parameters.

Other positive definite approaches include the penalty method (see Cai et al. [31] and Vassilevski and Wang [97]), the augmented Lagrangian method (see Hiptmair [60], [61] and [62]) and divergence-free basis methods (see Ewing and Wang [46], [47], Cliffe et al. [40] and Scheichl [89].) One of the most popular schemes is the so-called mixed-hybrid approach (see Arnold and Brezzi [8]), based on a non-conforming variational formulation of (2.8). This approach will be described in detail in Chapter 6. Note that none of these reformulations completely achieve both h-optimality *and* $\mathcal{A}$-optimality.

A few attempts have also been made at preconditioning the saddle-point system (2.37). Ewing et al., in [48], replace the leading matrix A with a lumped diagonal approximation and use the resulting saddle-point matrix to precondition the original one. The system is solved using a stationary iterative method that is h-optimal. However, the convergence rate deteriorates for anisotropic coefficients.

Algebraic approaches to solving (1.5) with Krylov methods were initiated by Rusten and Winther in [86]. The authors consider only well-conditioned coefficient tensors and propose a preconditioning operator of the form,

$$\mathcal{P} = \begin{pmatrix} \mathcal{I} & 0 \\ 0 & \mathcal{S} \end{pmatrix}, \tag{2.52}$$

where $\mathcal{S}$ is an approximation to the Laplacian operator $\nabla \cdot \nabla$ acting on W_h and $\mathcal{I}$ is the identity. For the Raviart-Thomas spaces, there is no *obvious* way to approximate the Laplacian operator on W_h. In [86], incomplete Cholesky factorisation of BB^T is recommended but it is known (see Greenbaum [55, Ch.11]) that such approximations are not h-optimal. Moreover, the method is not robust because the coefficient term is neglected. We will refer to operators such as (2.52) as 'H^1 preconditioners', since the main focus is on deriving approximations $\mathcal{S}$ for a global operator which is only well defined, in the classical sense, on subspaces of the Sobolev space $H^1(\Omega)$. Other preconditioning schemes which fall into the H^1 category are presented by Rusten et al. in [87] and [88]. These papers extend the discussion in [86] using domain decomposition ideas. The proposed methods are h-optimal but, again, neglect the coefficient term.

Arnold, Falk and Winther observe in [6] that for unit coefficients $\mathcal{A} = \mathcal{I}$, C in (1.5) also has the same mapping properties as the matrix operator,

$$\mathcal{P} = \begin{pmatrix} \mathcal{H} & 0 \\ 0 & \mathcal{I} \end{pmatrix}, \tag{2.53}$$

where $\mathcal{H} : H(div) \times H(div) \rightarrow I\!R$ is the $H(div)$ PDE operator defined, for vector functions $\vec{u}$ and $\vec{v}$ via,

$$(\mathcal{H}\vec{u}, \vec{v}) = (\vec{u}, \vec{v}) + (\nabla \cdot \vec{u}, \nabla \cdot \vec{v}).$$

We shall refer to this type of scheme as '$H(div)$ preconditioning'. $\mathcal{H}$ gives rise to a matrix similar to the one occurring in the penalty and augmented Lagrangian methods and thus can be approximated by any of the multilevel schemes suggested for those problems. A simpler and more practical method is presented in the form of a standard multigrid V-cycle, with special smoothing, by Arnold et al. in [6] and [7]. It is shown that the resulting approximation is optimal for $\mathcal{H}$ provided that the coefficient is of the form $\mathcal{A} = \rho\mathcal{I}$ where ρ is a constant. Discontinuous, anisotropic and variable ρ are not considered. We shall describe this scheme in more detail in Chapter 4.

Finally, in [98], Vassilevski and Lazarov consider preconditioners for the augmented indefinite system (1.11). A parameterised preconditioner of the form,

$$P = \begin{pmatrix} A + \alpha_1 B^T B & 0 \\ 0 & \alpha_2 I \end{pmatrix}, \tag{2.54}$$

is suggested, which calls on the multilevel schemes of [97] or [47] to approximate the leading block. Critically, convergence rates are highly sensitive to the artificial parameters, α_1 and α_2. Optimal values are not established.

2.6 Concluding remarks

In this introductory chapter, we reviewed stability theory for standard mixed finite element formulations of the model variable diffusion problem (1.1) and concluded that the discrete variational problem (2.12) is well-posed with an appropriate choice of finite element spaces. We introduced Raviart-Thomas approximation and derived the associated linear algebra problem (2.37). In addition, we made specific the dependence of the eigenvalue spectrum of the coefficient matrix on the discretisation parameter and the PDE coefficients. Algebraic properties of the element contributions to the weighted velocity mass matrix were derived.

Using all of this information, we are now ready to introduce and analyse new parameter-free preconditioning strategies for the saddle-point system (2.37).

Chapter 3

Ideal $H(div)$ preconditioning

In this chapter, we use the stability theory outlined in Chapter 2 to motivate an ideal preconditioner of the form (2.53).

The conditions of Z_h-ellipticity,

$$a(\vec{v_h}, \vec{v_h}) \geq \alpha_h \parallel \vec{v_h} \parallel_{V_h}^2, \quad \forall \vec{v_h} \in Z_h, \tag{3.1}$$

with $Z_h = \{\vec{v_h} \in V_h \mid b(\vec{v_h}, w_h) = 0 \quad \forall w_h \in W_h\}$, and discrete inf-sup stability,

$$\sup_{\vec{v_h} \in V_h} \frac{b(\vec{v_h}, w_h)}{\parallel \vec{v_h} \parallel_{V_h}} \geq \beta_h \parallel w_h \parallel_{W_h} \quad \forall w_h \in W_h, \tag{3.2}$$

essentially ensure (see Brezzi and Bathe, [24], and Proposition 2.1 in Brezzi, [23]), that the linear operator $\boldsymbol{C}$ in the saddle-point problem (2.34) defines an isomorphism from $H(div;\Omega) \times L^2(\Omega)$ onto it's dual space, and furthermore that $\parallel \boldsymbol{C}^{-1} \parallel$ is bounded in the natural norm defined on $H(div;\Omega) \times L^2(\Omega)$. Intuitively, choosing an operator $\boldsymbol{P}$ with the same mapping properties as $\boldsymbol{C}$, such that $\parallel \boldsymbol{P}^{-1} \parallel$ is bounded in the same norm, will give a good preconditioner for $\boldsymbol{C}$.

We thus consider a block-diagonal preconditioner P of the generic form,

$$P = \begin{pmatrix} P_1 & 0 \\ 0 & P_2 \end{pmatrix}, \tag{3.3}$$

where the symmetric and positive definite matrices $P_1 \in I\!R^{n \times n}$ and $P_2 \in I\!R^{m \times m}$ are chosen to represent norms on the lowest-order Raviart-Thomas spaces V_h and W_h, respectively. We investigate, first, the norms for which stability has been proved and choose P_1 and P_2 to satisfy,

$$\begin{aligned} \underline{v}^T P_1 \underline{v} &= \| \vec{v_h} \|_{div}^2 \quad \forall \vec{v}_h \in V_h, & (3.4) \\ \underline{w}^T P_2 \underline{w} &= \| w_h \|_0^2 \quad \forall w_h \in W_h, & (3.5) \end{aligned}$$

where $\underline{v}$ and $\underline{w}$ are the vectors of coefficients corresponding to the expansion of arbitrary $\vec{v_h} \in V_h$ and $w_h \in W_h$ in the chosen finite element basis sets.

To this end, we construct the *unweighted* velocity mass matrix $A_I \in I\!R^{n\times n}$ (equivalent to the weighted velocity mass matrix A in (2.35) with unit coefficients $\mathcal{A} = \mathcal{I}$) and $D \in I\!R^{n\times n}$ via,

$$\begin{aligned} A_{I\,ij} &= (\vec{\varphi_i}, \vec{\varphi_j}), & i,j = 1:n, \quad & (3.6) \\ D_{ij} &= (\nabla\cdot\vec{\varphi_i}, \nabla\cdot\vec{\varphi_j}), & i,j = 1:n. \quad & (3.7) \end{aligned}$$

This yields $P_1 = A_I + D$ since,

$$\| \vec{v}_h \|_{div}^2 = \| \vec{v}_h \|_0^2 + \| \nabla\cdot\vec{v}_h \|_0^2 = \underline{v}^T \left(A_{\mathcal{I}} + D\right) \underline{v}.$$

Note that since the matrix A_I is positive definite, so is $A_I + D$. Since we are dealing with the lowest-order Raviart-Thomas elements, $\nabla\cdot\vec{\varphi_i}$ is a constant for all velocity basis functions $\vec{\varphi_i}$. Integration in the construction of D is therefore trivial. Defining the pressure mass matrix N via,

$$N_{rs} = (\phi_r, \phi_s) \qquad r,s = 1:m, \tag{3.8}$$

yields $\| w_h \|_0^2 = \underline{w}^T N \underline{w}$ and thus $P_2 = N$. Again, the integration is trivial since,

$$\begin{aligned} N_{rs} &= \int_\Omega \phi_r\,\phi_s\,d\Omega \\ &= \sum_{i=1}^m \int_{K_i} \phi_r|_{K_i}\,\phi_s|_{K_i} dK_i \\ &= \begin{cases} |K_r| & \text{if } r = s, \\ 0 & \text{if } r \neq s. \end{cases} \end{aligned}$$

N is thus a positive definite *diagonal* matrix with entries corresponding to the areas or volumes of the finite elements in T_h.

To gauge the performance of the preconditioner,

$$P = \begin{pmatrix} A_I + D & 0 \\ 0 & N \end{pmatrix}, \tag{3.9}$$

our task is now to establish a theoretical bound for the eigenvalues $\{\lambda_i\}_{i=1}^{n+m}$ of,

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} \underline{u} \\ \underline{p} \end{pmatrix} = \lambda \begin{pmatrix} A_I + D & 0 \\ 0 & N \end{pmatrix} \begin{pmatrix} \underline{u} \\ \underline{p} \end{pmatrix}, \tag{3.10}$$

and determine the extent to which it depends on the discretisation parameter h and the coefficient tensor $\mathcal{A}$. To prove results, we begin by deriving the discrete matrix form of the inf-sup stability inequality.

3.1 Matrix form of inf-sup stability inequality

Since the discrete inf-sup condition (2.20) holds for $\|\cdot\|_{V_h} = \|\cdot\|_{div}$ and $\|\cdot\|_{W_h} = \|\cdot\|_0$, we see that,

$$\begin{aligned} \beta_h \left(\underline{w}^T N \underline{w}\right)^{\frac{1}{2}} &\leq \max_{\underline{v} \in I\!\!R^n} \frac{\underline{w}^T B \underline{v}}{\left(\underline{v}^T \left(A_I + D\right) \underline{v}\right)^{\frac{1}{2}}} \\ &= \max_{\underline{z} = (A_I + D)^{\frac{1}{2}} \underline{v}} \frac{\underline{w}^T B \left(A_I + D\right)^{-\frac{1}{2}} \underline{z}}{\left(\underline{z}^T \underline{z}\right)^{\frac{1}{2}}} \\ &= \frac{\underline{w}^T B \left(A_I + D\right)^{-1} B^T \underline{w}}{\left(\underline{w}^T B \left(A_I + D\right)^{-1} B^T \underline{w}\right)^{\frac{1}{2}}} \\ &= \left(\underline{w}^T B \left(A_I + D\right)^{-1} B^T \underline{w}\right)^{\frac{1}{2}}. \end{aligned}$$

Hence,

$$\beta_h^2 \leq \frac{\underline{w}^T B \left(A_I + D\right)^{-1} B^T \underline{w}}{\underline{w}^T N \underline{w}} \qquad \forall \underline{w} \in I\!\!R^m \backslash \{\underline{0}\}. \tag{3.11}$$

Further, the Cauchy-Schwarz inequality yields,

$$\begin{aligned} \sup_{\vec{v_h} \in V_h \backslash \{0\}} \frac{|\, b(\vec{v_h}, w_h) \,|^2}{\|\, \vec{v_h} \,\|_{div}^2 \|\, w_h \,\|_0^2} &\leq \sup_{\vec{v_h} \in V_h \backslash \{0\}} \frac{\|\, w_h \,\|_0^2 \|\, \nabla \cdot \vec{v_h} \,\|_0^2}{\|\, w_h \,\|_0^2 \left(\|\, \vec{v_h} \,\|_0^2 + \|\, \nabla \cdot \vec{v_h} \,\|_0^2\right)} \\ &= \sup_{\vec{v_h} \in V_h \backslash \{0\}} \frac{\|\, \nabla \cdot \vec{v_h} \,\|_0^2}{\left(\|\, \vec{v_h} \,\|_0^2 + \|\, \nabla \cdot \vec{v_h} \,\|_0^2\right)} \leq 1, \end{aligned}$$

and so we obtain the double-sided bound,

$$\beta_h^2 \leq \frac{\underline{w}^T B \left(A_I + D\right)^{-1} B^T \underline{w}}{\underline{w}^T N \underline{w}} \leq 1 \qquad \forall\, \underline{w} \in I\!R^m \backslash \{\underline{0}\}. \tag{3.12}$$

An alternative statement of (3.12) is the following eigenvalue bound.

Lemma 13 *The eigenvalues $\{\sigma_i\}_{i=1}^m$ of the generalised eigenvalue problem,*

$$B \left(A_I + D\right)^{-1} B^T \underline{w} \;\; = \;\; \sigma N \underline{w}, \tag{3.13}$$

arising in the Raviart-Thomas approximation to (2.12) are bounded by constants independent of h and lie in the interval $[\,\beta_h^2,\, 1\,]$.

Remark 6 *A computable upper bound for the discrete inf-sup constant is the square-root of the minimum of the eigenvalues of (3.13), i.e. $\beta_h \leq \sqrt{\sigma_{min}}$.*

3.2 Eigenvalue bounds

To establish eigenvalue bounds for (3.10), we require the following preliminary result.

Lemma 14 *If $\nabla \cdot V_h \subset W_h$, then,*

$$D \;\; = \;\; B^T N^{-1} B. \tag{3.14}$$

Proof Consider writing the matrices B, N and D in operator form, where D is the matrix in (3.7) and N and B are defined in (3.8) and (2.36), respectively. We have,

$$\begin{aligned}
(\boldsymbol{B}\vec{x_h}, z_h) &= (\nabla \cdot \vec{x_h}, z_h) = \left(\vec{x_h}, \boldsymbol{B}^T z_h\right), && \forall\, \vec{x_h} \,\in\, V_h,\, z_h \,\in\, W_h,\\
(\boldsymbol{D}\vec{x_h}, \vec{y_h}) &= (\nabla \cdot \vec{x_h}, \nabla \cdot \vec{y_h}), && \forall\, \vec{x_h},\, \vec{y_h}, \in\, V_h,\\
(\boldsymbol{N} z_h, w_h) &= (z_h, w_h), && \forall\, z_h,\, w_h \,\in\, W_h.
\end{aligned}$$

Here, $\boldsymbol{N}$ acts as the identity operator on W_h. For any $\vec{x}_h$ and $\vec{y}_h$ in V_h, if $\nabla \cdot V_h \subset W_h$,

$$\begin{aligned}
(\boldsymbol{D}\vec{x_h}, \vec{y_h}) = (\nabla \cdot \vec{x_h}, \nabla \cdot \vec{y_h}) &= (\boldsymbol{B}\vec{x_h},\, \nabla \cdot \vec{y_h})\\
&= \left(\boldsymbol{N}\boldsymbol{N}^{-1}\boldsymbol{B}\vec{x_h}, \nabla \cdot \vec{y_h}\right)\\
&= \left(\boldsymbol{N}^{-1}\boldsymbol{B}\vec{x_h},\, \nabla \cdot \vec{y_h}\right)\\
&= \left(\boldsymbol{B}^T\boldsymbol{N}^{-1}\boldsymbol{B}\vec{x_h},\, \vec{y_h}\right). \qquad \square
\end{aligned}$$

Remark 7 *For Raviart-Thomas approximation, $\nabla \cdot V_h = W_h$ by construction, so that $D = B^T N^{-1} B$.*

To proceed with the eigenvalue analysis, we distinguish two cases. Consider, first, the trivial case of unit coefficients, $\mathcal{A} = \mathcal{I}$. Here, the mass matrices A and A_I are identical and we obtain the following bound which is a new result.

Theorem 5 *The $n+m$ eigenvalues of the generalised eigenvalue problem,*

$$\begin{pmatrix} A_I & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} \underline{u} \\ \underline{p} \end{pmatrix} = \sigma \begin{pmatrix} A_I + D & 0 \\ 0 & N \end{pmatrix} \begin{pmatrix} \underline{u} \\ \underline{p} \end{pmatrix}, \tag{3.15}$$

arising in the Raviart-Thomas approximation of (2.12) are bounded by constants independent of h and lie in the intervals,

$$\left[-1, -\beta_h^2\right] \cup [1], \tag{3.16}$$

where β_h is the discrete inf-sup constant satisfying (2.20).

Proof The eigenvalues $\{\sigma_i\}_{i=1}^{m+n}$ satisfy,

$$\begin{aligned} A_I \underline{u} + B^T \underline{p} &= \sigma \left(A_I + D\right) \underline{u}, \\ B\underline{u} &= \sigma N \underline{p}. \end{aligned}$$

If $\sigma = 1$, then,

$$\left.\begin{aligned} B^T \underline{p} &= D\underline{u} \\ B\underline{u} &= N\underline{p} \end{aligned}\right. \Rightarrow \; B^T N^{-1} B \underline{u} \;=\; D\underline{u}. \tag{3.17}$$

Using Lemma 14 we deduce that this is true for any vector $\underline{u}$ in $I\!R^n$. Since $D \in I\!R^{n \times n}$ is symmetric, there are n linearly independent eigenvectors corresponding to $\sigma = 1$ and hence n distinct eigenvalues equal to unity.

Now suppose $\sigma \neq 1$. Then $\underline{u}, \underline{p}$ satisfy,

$$\begin{aligned} B^T \underline{p} &= (\sigma - 1)\left(A_I + D\right) \underline{u} + D\underline{u}, \\ B\underline{u} &= \sigma N \underline{p}. \end{aligned}$$

Using (3.14) we have that,

$$B\left(A_I + D\right)^{-1} B^T \underline{p} \;=\; (\sigma - 1) B\underline{u} + B\left(A_I + D\right)^{-1} D\underline{u},$$

$$\begin{aligned} &= \sigma(\sigma-1)N\underline{p} + B(A_I+D)^{-1}B^TN^{-1}B\underline{u} \\ &= \sigma(\sigma-1)N\underline{p} + \sigma B(A_I+D)^{-1}B^T\underline{p}. \end{aligned}$$

Thus, $B(A_I+D)^{-1}B^T\underline{p} = -\sigma N\underline{p}$ and the result follows from Lemma 13. $\square$

To summarise, when unit coefficients are present, choosing the preconditioner with diagonal blocks representing the norms for which stability holds, leads to an h-optimal eigenvalue bound. To efficiently precondition a problem with a general coefficient tensor, however, P *must* supply scaling with respect to the coefficient tensor $\mathcal{A}$. Hence, we now propose the preconditioner,

$$P = \begin{pmatrix} A+D & 0 \\ 0 & N \end{pmatrix}, \tag{3.18}$$

whose leading block represents the norm $\| \cdot \|_{div,\mathcal{A}}$ which is equivalent to $\| \cdot \|_{div}$.

To derive a corresponding eigenvalue bound, it is necessary, first of all, to establish the dependence on h of the minimum eigenvalue of the Schur complement matrix $BA^{-1}B^T$. Recall that the matrix A, here, is the *weighted* velocity mass matrix defined in (2.35).

Lemma 15 *Let μ_{min} denote the minimum eigenvalue of the matrix $BA^{-1}B^T$, arising in the Raviart-Thomas approximation of (2.12). There exists a constant c, independent of h and the coefficient tensor $\mathcal{A}$, such that,*

$$\frac{c\beta_h^2 h_{min}^d}{\Gamma} \leq \mu_{min}, \qquad d = 2, 3, \tag{3.19}$$

where $\beta_h > 0$ is the discrete inf-sup constant satisfying (2.20), $h_{min} = \min_K h_K$ and $\Gamma > 0$ is a constant satisfying (2.1).

Proof The elements of the proof are inf-sup stability, the bound (2.47), and assumption (2.1). First, recall from Lemma 10 that for any $w_h \in W_h$ there exist constants c_3 and c_4, independent of h, such that,

$$c_3 h_{min}^d \, \underline{w}^T\underline{w} \leq \| w_h \|_0^2 \leq c_4 h^d \underline{w}^T\underline{w}, \tag{3.20}$$

where $\underline{w}$ is the vector of coefficients corresponding to the expansion of w_h in the basis for W_h.

Combining this with inf-sup stability, and applying (2.1), we have,

$$\begin{aligned}\beta_h^2\, c_3\, h_{min}^d\, \underline{w}^T \underline{w} \;\le\; \beta_h^2 \parallel w_h \parallel_0^2 \;&\le\; \sup_{\vec{v}_h \in V_h \backslash \{\vec{0}\}} \frac{\mid (w_h, \nabla \cdot \vec{v}_h) \mid^2}{\parallel \vec{v}_h \parallel_{div}^2} \\ &\le\; \sup_{\vec{v}_h \in V_h \backslash \{\vec{0}\}} \frac{\mid (w_h, \nabla \cdot \vec{v}_h) \mid^2}{\parallel \vec{v}_h \parallel_0^2} \\ &\le\; \Gamma \sup_{\vec{v}_h \in V_h \backslash \{\vec{0}\}} \frac{\mid (w_h, \nabla \cdot \vec{v}_h) \mid^2}{(\mathcal{A}^{-1}\vec{v}_h, \vec{v}_h)}.\end{aligned}$$

Translating into matrix notation, we obtain,

$$\begin{aligned}\sup_{\vec{v}_h \in V_h \backslash \{\vec{0}\}} \frac{(w_h, \nabla \cdot \vec{v}_h)}{(\mathcal{A}^{-1}\vec{v}_h, \vec{v}_h)^{\frac{1}{2}}} \;&=\; \max_{\underline{v} \in I\!R^n \backslash \{\underline{0}\}} \frac{\underline{w}^T B \underline{v}}{(\underline{v}^T A \underline{v})^{\frac{1}{2}}} \;=\; \max_{\underline{z} = A^{\frac{1}{2}} \underline{v}} \frac{\underline{w}^T B A^{-\frac{1}{2}} \underline{z}}{(\underline{z}^T \underline{z})^{\frac{1}{2}}} \\ &=\; \frac{\underline{w}^T B A^{-1} B^T \underline{w}}{(\underline{w}^T B A^{-1} B^T \underline{w})^{\frac{1}{2}}} \;=\; \left(\underline{w}^T B A^{-1} B^T \underline{w}\right)^{\frac{1}{2}}.\end{aligned}$$

Hence, we see that,

$$\frac{c_3 \beta_h^2 h_{min}^d}{\Gamma} \;\le\; \frac{\underline{w}^T B A^{-1} B^T \underline{w}}{\underline{w}^T \underline{w}} \qquad \forall \underline{w} \in I\!R^m \backslash \{\underline{0}\}, \tag{3.21}$$

which proves the result. □

The following theorem extends the bound established by Vassilevski and Lazarov in [98]. (Our analysis is for a different matrix A and no artificial parameters.) The crucial difference is that the matrix D in the preconditioner (3.18) supplies scaling with respect to N. Recalling the bound (2.47), we obtain,

$$c_3 h_{min}^d\, \underline{w}^T \underline{w} \;\le\; \underline{w}^T N \underline{w} \;\le\; c_4 h^d \underline{w}^T \underline{w} \quad \forall \underline{w} \in I\!R^m \backslash \{\underline{0}\}, \tag{3.22}$$

and so N represents an h-dependent scaling. We begin with quasi-uniform meshes.

Theorem 6 *If T_h is quasi-uniform, the $n+m$ eigenvalues of the generalised eigenvalue problem,*

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} \underline{u} \\ \underline{p} \end{pmatrix} = \sigma \begin{pmatrix} A + D & 0 \\ 0 & N \end{pmatrix} \begin{pmatrix} \underline{u} \\ \underline{p} \end{pmatrix}, \tag{3.23}$$

arising in the Raviart-Thomas approximation of (2.12), lie in the union of the intervals,

$$\left(-1, -\left(\frac{c\mu_{min}}{|K|_{min} + \mu_{min}}\right)\right] \cup [1], \tag{3.24}$$

where μ_{min} is the minimum eigenvalue of the Schur complement matrix $BA^{-1}B^T$, $\mid K \mid_{min}$ is the volume of the smallest element in T_h and c is a constant independent of h and $\mathcal{A}$.

Proof The eigenvalues $\{\sigma_i\}_{i=1}^{m+n}$ satisfy,

$$\begin{aligned} A\underline{u} + B^T\underline{p} &= \sigma\left(A + D\right)\underline{u}, \\ B\underline{u} &= \sigma N\underline{p}. \end{aligned}$$

As in the proof of Theorem 5, there are n eigenvalues equal to unity and the remaining m eigenvalues $\{\sigma_i\}_{i=1}^{m}$ satisfy,

$$B\left(A + D\right)^{-1}B^T\underline{p} = -\sigma N\underline{p}. \tag{3.25}$$

Since $D = B^TN^{-1}B$, and N is diagonal, these are the same as the eigenvalues of the matrix,

$$-N^{-\frac{1}{2}}B\left(A + B^TN^{-1}B\right)^{-1}B^TN^{-\frac{1}{2}}.$$

Rearranging gives,

$$\begin{aligned} & N^{-\frac{1}{2}}B\left(A + B^TN^{-1}B\right)^{-1}B^TN^{-\frac{1}{2}} \\ = \; & N^{-\frac{1}{2}}BA^{-\frac{1}{2}}\left(I + A^{-\frac{1}{2}}B^TN^{-\frac{1}{2}}N^{-\frac{1}{2}}BA^{-\frac{1}{2}}\right)^{-1}A^{-\frac{1}{2}}B^TN^{-\frac{1}{2}} \qquad (3.26) \\ = \; & X\left(I + X^TX\right)^{-1}X^T, \end{aligned}$$

where $X = N^{-\frac{1}{2}}BA^{-\frac{1}{2}}$. Applying the Sherman-Morrison-Woodbury formula (see Golub and Van Loan [56, p.51]) yields,

$$\left(I + X^TX\right)^{-1} = I - X^T\left(I + XX^T\right)^{-1}X,$$

and so,

$$X\left(I + XX^T\right)^{-1}X^T = X\left(I - X^T\left(I + XX^T\right)^{-1}X\right)X^T. \tag{3.27}$$

Now we can apply Lemma 3.1 of [98] with $X = N^{-\frac{1}{2}}BA^{-\frac{1}{2}}$ to relate the eigenvalues of (3.25) to those of $BA^{-1}B^T$. For completeness, we reproduce this argument below.

Let $\underline{x}_i$ be an eigenvector of XX^T and λ_i denote the corresponding eigenvalue. Then, with (3.27), we obtain

$$\begin{aligned} X\left(I + XX^T\right)^{-1}X^T\underline{x}_i &= XX^T\underline{x}_i - XX^T\left(I + XX^T\right)^{-1}XX^T\underline{x}_i \\ &= \lambda_i\underline{x}_i - \left(\frac{\lambda_i^2}{1 + \lambda_i}\right)\underline{x}_i \\ &= \left(\frac{\lambda_i}{1 + \lambda_i}\right)\underline{x}_i. \end{aligned}$$

Hence, the eigenvalues of $X\left(I+XX^T\right)^{-1}X^T$ are the set of values $\{\frac{\lambda_i}{1+\lambda_i}\}_{i=1}^m$, where $\{\lambda_i\}_{i=1}^m$ are the eigenvalues of $XX^T = N^{-\frac{1}{2}}BA^{-1}B^TN^{-\frac{1}{2}}$.

Since $N^{-1}BA^{-1}B^T$ has the same eigenvalue spectrum as $N^{-\frac{1}{2}}BA^{-1}B^TN^{-\frac{1}{2}}$, the negative eigenvalues of *our* generalised eigenvalue problem (3.23) lie in the interval,

$$\left[-\max_i \frac{\lambda_i}{1+\lambda_i},\ -\min_i \frac{\lambda_i}{1+\lambda_i}\right].$$

Since A is positive definite and B is full rank, $\lambda_i > 0$ for all i, thus,

$$\{\sigma_i\}_{i=1}^m \quad \in \quad \left(-1,\ -\frac{\lambda_{min}}{1+\lambda_{min}}\right], \tag{3.28}$$

where λ_{min} is the minimum eigenvalue of $N^{-1}BA^{-1}B^T$. (This result was also stated in [60], for a different choice of N.)

Recall that *here* the eigenvalues of N are the volumes of the elements. Hence, an alternative statement of (3.22) is,

$$c_3 h_{min}^d \ \leq\ |K|_{min} \ \leq\ |K|_{max} \leq\ c_4 h^d, \tag{3.29}$$

where $|K|_{min}$ and $|K|_{max}$ denote the smallest and largest volumes of the finite elements in T_h. For quasi-uniform meshes, it can also be shown that there exist positive constants c_5 and c_6, independent of h, satisfying,

$$c_5 h_{min} \ \leq h\ \leq\ c_6 h_{min}. \tag{3.30}$$

Combining this with (3.29) yields,

$$|K|_{min} \ \leq\ |K|_{max} \ \leq\ \frac{c_4 c_6}{c_3}|K|_{min}. \tag{3.31}$$

If we denote $c = \frac{c_3}{c_4 c_6}$, it follows that λ_{min} satisfies,

$$\frac{c\mu_{min}}{|K|_{min}} \ \leq\ \frac{\mu_{min}}{|K|_{max}} \ \leq \lambda_{min} \ \leq\ \frac{\mu_{min}}{|K|_{min}}, \tag{3.32}$$

where μ_{min} is the minimum eigenvalue of $BA^{-1}B^T$. Hence,

$$-\frac{\lambda_{min}}{1+\lambda_{min}} \leq \frac{-\frac{\mu_{min}}{|K|_{max}}}{1+\frac{\mu_{min}}{|K|_{min}}} \leq \frac{-\frac{c\mu_{min}}{|K|_{min}}}{1+\frac{\mu_{min}}{|K|_{min}}},$$

and we obtain,

$$\{\sigma_i\}_{i=1}^m \quad \in \quad \left(-1,\ -\left(\frac{c\mu_{min}}{|K|_{min}\ +\ \mu_{min}}\right)\right]. \quad \Box$$

Remark 8 *For uniform meshes we obtain $c_5 = c_6 = 1$ in (3.30) and $c_3 = c_4$ in (3.29) so $c = 1$ in the eigenvalue bound (3.24).*

To show that (3.24) is an h-optimal eigenvalue bound, we refine the previous theorem to obtain the following two corollaries.

Corollary 2 *If T_h is quasi-uniform and if $|K|_{min} \leq \mu_{min}$, the $n + m$ eigenvalues of (3.23) lie in the union of the intervals,*

$$(-1, -\tilde{c}] \cup [1] \tag{3.33}$$

where $\tilde{c}$ is a constant independent of h and $\mathcal{A}$.

Proof If $|K|_{min} \leq \mu_{min}$ then clearly $|K|_{min} + \mu_{min} \leq 2\mu_{min}$, and so, in Theorem 6, we obtain,

$$-\left(\frac{c\mu_{min}}{|K|_{min} + \mu_{min}}\right) \quad \leq \quad -\frac{c\mu_{min}}{2\mu_{min}} = -\frac{c_3}{2c_4c_6} = -\tilde{c}. \quad \square$$

Corollary 3 *If T_h is quasi-uniform and $|K|_{min} > \mu_{min}$, the $n + m$ eigenvalues of (3.23) lie in the union of the intervals,*

$$\left(-1, -\frac{\tilde{c}\,\beta_h^2}{\Gamma}\right] \cup [1], \tag{3.34}$$

where $\tilde{c}$ is a constant independent of h and $\mathcal{A}$.

Proof If $|K|_{min} > \mu_{min}$, we have $|K|_{min} + \mu_{min} < 2|K|_{min}$, and so, in Theorem 6,

$$-\left(\frac{c\mu_{min}}{|K|_{min} + \mu_{min}}\right) \quad \leq \quad -\frac{c\mu_{min}}{2|K|_{min}} = -\frac{c_3\mu_{min}}{2c_4c_6|K|_{min}}.$$

Applying Lemma 15, (3.29) and (3.30) yields,

$$-\frac{c\mu_{min}}{2|K|_{min}} < -\frac{c\,c_3\,\beta_h^2}{2c_4\Gamma}\left(\frac{h_{min}}{h}\right)^d < -\frac{c\,c_3\,\beta_h^2}{2c_4c_6\Gamma} = -\frac{c^2\,\beta_h^2}{2\Gamma}.$$

The result follows with $\tilde{c} = \frac{1}{2}\left(\frac{c_3}{c_4c_6}\right)^2$. $\square$

Remark 9 *For uniform meshes we have $c_5 = c_6 = 1$ and $c_3 = c_4$, so $\tilde{c} = \frac{1}{2}$ in the eigenvalue bounds (3.33) and (3.34).*

For quasi-uniform meshes, Corollaries 2 and 3 tell us that the preconditioner (3.18) is h-optimal but may not be $\mathcal{A}$-optimal. Lemma 15 suggests that small coefficients, or, equivalently, large values of Γ, can cause μ_{min}, the minimum eigenvalue of $BA^{-1}B^T$, to be small. Theorem 6 suggests that MINRES convergence will not be efficient in such cases. However, the bound may be overly pessimistic in this respect. The dependence of μ_{min} on the coefficient tensor is not straightforward. We comment on two important classes of test problems in the next section.

For non quasi-uniform meshes, the bounds established in Theorems 6 and Corollaries 2 and 3 also hold. However, the constants defined in (3.29) and (3.30) depend on the ratio $\frac{h_{min}}{h}$. As a consequence, the constant c appearing in (3.24) tends to zero as h_{min} tends to zero. We do not include non quasi-uniform meshes in the statements of Theorem 6 and its corollaries because the results are derived from the bound (3.32) which is grossly pessimistic for highly non-uniform meshes. Hence, the eigenvalue bound stated in Theorems 6 is also too pessimistic for non-uniform meshes. Numerical evidence of this will be given later.

To illustrate that the above theory is tight, we present a simple numerical example.

Numerical example

Consider the model problem (1.4) discretised on $\Omega = [0,1] \times [0,1]$ with *uniform meshes* of right-angled triangles, unit coefficients $\mathcal{A} = \mathcal{I}$, and a homogeneous Dirichlet boundary condition, $\partial\Omega = \partial\Omega_D$. The observed eigenvalues of the preconditioned system $\{\sigma_1, \ldots, \sigma_{n+m}\}$ are listed in Table 3.1; they confirm that the bounds (3.16) and (3.24) in Theorems 5 and 6 are tight. Note that the constant c in (3.24) is one by Remark 8. The observed negative eigenvalues are plotted in Fig. 3.1.

h	$-\beta^2$	μ_{min}	$-\frac{\mu_{min}}{\lvert K\rvert_{min}+\mu_{min}}$	σ_1	σ_m	σ_{m+1}	σ_{m+n}
$\frac{1}{4}$	-0.9525	0.6268	-0.9525	-0.9983	-0.9525	1	1
$\frac{1}{8}$	-0.9519	0.1549	-0.9519	-0.9996	-0.9519	1	1
$\frac{1}{16}$	-0.9518	0.0386	-0.9518	-0.9999	-0.9518	1	1

Table 3.1: Bounds and observed eigenvalues of preconditioned system

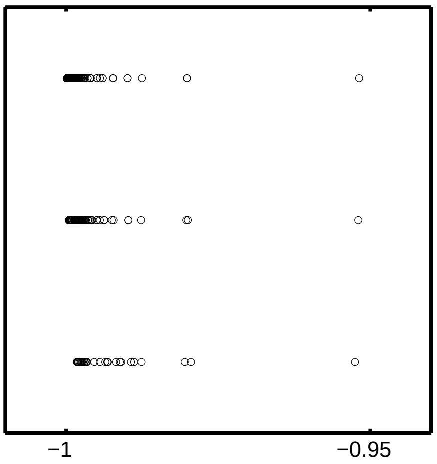

Figure 3.1: Negative eigenvalues of preconditioned system, $h = \frac{1}{4}, \frac{1}{8}, \frac{1}{16}$

3.2.1 Deterioration of the eigenvalue bound due to coefficients

Anisotropic coefficients

To illustrate the impact of anisotropic structure in $\mathcal{A}$ on the bound (3.24), we consider a class of test problems in $I\!R^2$. Choose,

$$\mathcal{A} = \begin{pmatrix} \epsilon & 0 \\ 0 & 1 \end{pmatrix} \quad \forall \vec{x} \in \Omega,$$

with anisotropy parameter $\epsilon \to \infty$ or $\epsilon \to 0$, so that $\mathcal{A}$ is ill-conditioned with,

$$\gamma = \min\{\frac{1}{\epsilon}, 1\}, \quad \Gamma = \max\{\frac{1}{\epsilon}, 1\}.$$

Although the bound (3.19) in Lemma 15 suggests that μ_{min} behaves like ϵ as $\epsilon \to 0$, this bad behaviour can be avoided by choosing a mesh aligned with the anisotropy. Here, the use of rectangular finite elements is crucial to the success of the preconditioning.

To see this, recall from Chapter 2 that using rectangular elements aligned with the coordinate axes produces element matrices A^K with the special block structure (2.44). Thus, with appropriate edge ordering, the global weighted mass matrix has the special structure,

$$A = \begin{pmatrix} \frac{1}{\epsilon}A_x & 0 \\ 0 & A_y \end{pmatrix},$$

with A_x and A_y defined via,

$$\begin{aligned} A_{x\,ij} &= \int_\Omega \vec{\varphi}_{i,x} \cdot \vec{\varphi}_{j,x} \, d\Omega, \qquad i,j = 1 : n_x, \\ A_{y\,rs} &= \int_\Omega \vec{\varphi}_{r,y} \cdot \vec{\varphi}_{s,y} \, d\Omega, \qquad r,s = n_x + 1 : n_x + n_y, \end{aligned}$$

where n_x and n_y denote the number of edges aligned with the x-axis and the y-axis respectively. Consequently, the Schur complement matrix $BA^{-1}B^T$ also has a special block structure. In fact this is true whenever diagonal coefficients and rectangular meshes are present. Moreover, this structure can be exploited to good effect by our preconditioner.

We can write,

$$\begin{aligned} BA^{-1}B^T &= B \begin{pmatrix} \epsilon A_x^{-1} & 0 \\ 0 & A_y^{-1} \end{pmatrix} B^T = \begin{pmatrix} B_x & B_y \end{pmatrix} \begin{pmatrix} \epsilon A_x^{-1} & 0 \\ 0 & A_y^{-1} \end{pmatrix} \begin{pmatrix} B_x^T \\ B_y^T \end{pmatrix} \\ &= \epsilon B_x A_x^{-1} B_x^T + B_y A_y^{-1} B_y^T. \end{aligned}$$

Thus, if we denote by $\lambda_{min}(\cdot)$ and $\lambda_{max}(\cdot)$ the minimum and maximum eigenvalues of a designated matrix, we obtain, as a consequence of the minimax theorem (see [56, p.411]),

$$\begin{aligned} \epsilon\lambda_{min}\left(B_x A_x^{-1} B_x^T\right) + \lambda_{min}\left(B_y A_y^{-1} B_y^T\right) &\leq \mu_{min} \\ \epsilon\lambda_{min}\left(B_x A_x^{-1} B_x^T\right) + \lambda_{max}\left(B_y A_y^{-1} B_y^T\right) &\geq \mu_{min}. \end{aligned}$$

Now we see that if $\epsilon << 1$, μ_{min} is bounded independently of the anisotropy parameter. Only when *all* the coefficients on the diagonal of $\mathcal{A}$ are small, does μ_{min} deteriorate. The same phenomenon occurs using brick elements aligned with the coordinate axes in $I\!R^3$.

Now, for triangular elements, the matrix A cannot be partitioned in the same way. Each row is scaled by *all* of the coefficients. Hence,

$$A_{ij} = \int_\Omega \frac{1}{\epsilon} \varphi_{\vec{i},x} \cdot \varphi_{\vec{j},x} \, d\Omega + \int_\Omega \varphi_{\vec{i},y} \cdot \varphi_{\vec{j},y} \, d\Omega \quad i,j = 1 : n, \tag{3.35}$$

and we can write, $A = \frac{1}{\epsilon} A_x + A_y$, where, here,

$$\begin{aligned} A_{x\,ij} &= \int_\Omega \vec{\varphi}_{i,x} \cdot \vec{\varphi}_{j,x} \, d\Omega, \qquad i,j = 1 : n, \\ A_{y\,rs} &= \int_\Omega \vec{\varphi}_{r,y} \cdot \vec{\varphi}_{s,y} \, d\Omega, \qquad r,s = 1 : n. \end{aligned}$$

Hence, in the limit $\epsilon \to 0$,

$$BA^{-1}B^T \to \epsilon \left(B{A_x}^{-1}B^T \right),$$

and so μ_{min} will also tend to zero. Unlike rectangular elements, if just one coefficient on the diagonal of the coefficient tensor $\mathcal{A}$ tends to zero then so does μ_{min}.

Consider again, the model problem discretised on $\Omega = [0,1] \times [0,1]$ with a homogeneous Dirichlet boundary condition. The minimum eigenvalues of the Schur complement matrix for uniform square meshes and varying ϵ are listed in Table 3.2. For $\epsilon \to 0$ and a fixed mesh, μ_{min} does not decay to zero and MINRES convergence is insensitive to ϵ. For the same problem solved using uniform triangular Raviart-Thomas elements, we observe that $\mu_{min} \to 0$ as $\epsilon \to 0$. To see this, compare the eigenvalues of the preconditioned systems in Fig. 3.2 for a fixed h and varying ϵ.

$h \quad \epsilon$	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}
$\frac{1}{8}$	0.1718	0.1578	0.1564	0.1562	0.1562	0.1562
$\frac{1}{16}$	0.0425	0.0391	0.0387	0.0387	0.0387	0.0387
$\frac{1}{132}$	0.0106	0.0097	0.0096	0.0096	0.0096	0.0096
$h \quad \epsilon$	10^{1}	10^{2}	10^{3}	10^{4}	10^{5}	10^{6}
$\frac{1}{8}$	1.7182	15.777	1.564e2	1.562e3	1.562e4	1.562e5
$\frac{1}{16}$	0.4254	3.9064	3.872e1	3.868e2	3.868e3	3.868e4
$\frac{1}{32}$	0.1061	0.9742	9.6557	9.647e1	9.646e2	9.646e3

Table 3.2: Observed values of μ_{min}, anisotropic coefficients, square elements

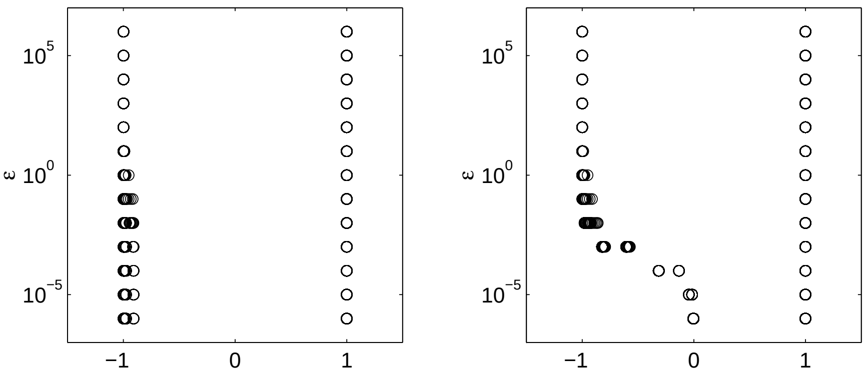

Figure 3.2: Eigenvalues of preconditioned system, $h = \frac{1}{16}$, $\epsilon \in [10^{-6}, 10^{6}]$, anisotropic coefficients, squares (left), triangles (right)

Discontinuous coefficients

In practical applications, both entries on the diagonal of $\mathcal{A}$ may be small. Consider then a class of discontinuous test problems. Take,

$$\mathcal{A} = \begin{cases} \epsilon I & \forall \vec{x} \in \Omega_*, \\ I & \forall \vec{x} \in \Omega \backslash \Omega_*, \end{cases} \tag{3.36}$$

where I is the identity matrix and $\Omega_* \subset \Omega = [0,1] \times [0,1]$. Setting $0 < \epsilon \ll 1$ in Ω_* describes a zone of low permeability, a feature common to groundwater flow problems.

To illustrate the numerical difficulties inherent in solving this problem, choose a jump zone $\Omega_* = [0.25, 0.5] \times [0.25, 0.75]$ and the permeability coefficient $\epsilon \in \left[10^{-6}, 10^{6}\right]$. The corresponding values of μ_{min}, for uniform meshes of square elements, are listed in Table 3.3. (Exactly the same behaviour occurs with uniform triangular meshes.)

$h \quad \epsilon$	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}
$\frac{1}{8}$	0.2074	0.0308	0.0036	3.648e-4	3.648e-5	3.649e-6
$\frac{1}{16}$	0.0506	0.0078	0.0008	8.051e-5	8.053e-6	8.054e-7
$\frac{1}{32}$	0.0156	0.0019	0.0002	1.894e-5	1.895e-6	1.895e-7
$h \quad \epsilon$	10^{1}	10^{2}	10^{3}	10^{4}	10^{5}	10^{6}
$\frac{1}{8}$	0.3409	0.3452	0.3456	0.3457	0.3547	0.3547
$\frac{1}{16}$	0.0846	0.0857	0.0858	0.0858	0.0858	0.0858
$\frac{1}{32}$	0.0211	0.0214	0.0214	0.0214	0.0214	0.0214

Table 3.3: Observed values of μ_{min}, discontinuous coefficient, square elements

If ϵ is large, the right-hand bound for the negative eigenvalues in Theorem 6 converges rapidly to -1 as $h \to 0$. MINRES convergence is fast. On the other hand, if $\epsilon \ll 1$ then μ_{min} is close to zero and MINRES convergence deteriorates.

Alternatively, take $\epsilon < 1$ and solve the same problem with coefficients,

$$\mathcal{A} = \begin{cases} I & \forall \vec{x} \in \Omega_* \\ \frac{1}{\epsilon} I & \forall \vec{x} \in \Omega \backslash \Omega_*. \end{cases} \tag{3.37}$$

The eigenvalues of the preconditioned systems associated with coefficients (3.36) and (3.37), for fixed h and $\epsilon \in \left[10^{-6}, 1\right]$ are shown in Fig. 3.3. The plot on the left illustrates the decay to zero of a subset of the negative eigenvalues of problem (3.36) as $\epsilon \to 0$. The plot on the right indicates that all the negative eigenvalues of problem (3.37) are bounded independently of the jump coefficient.

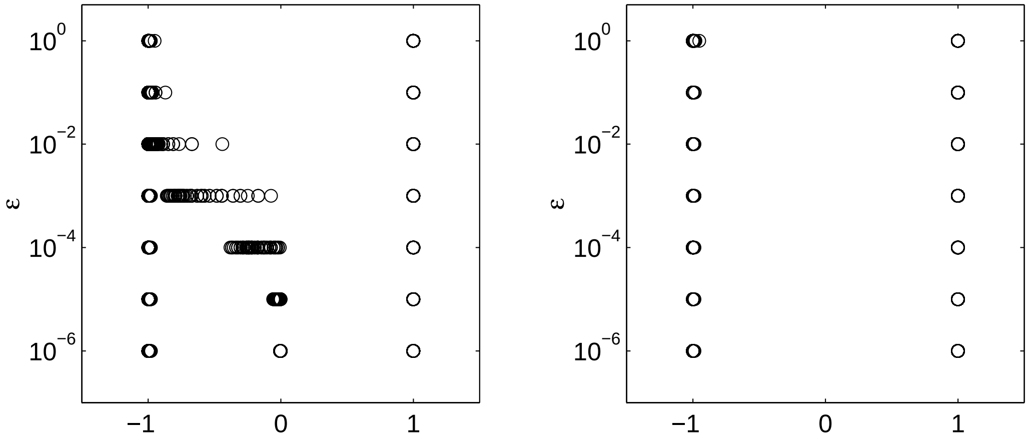

Figure 3.3: Eigenvalues of preconditioned system, discontinuous coefficients, $h = \frac{1}{16}$, $\epsilon \in [10^{-6}, 1]$, unscaled (left) and scaled (right)

Notice that if the source term f is rescaled, solving problem (3.37) instead of problem (3.36) corresponds to multiplying the underlying PDE by a constant and has the effect of multiplying all the eigenvalues of $BA^{-1}B^T$ by $\frac{1}{\epsilon}$. The result is that for $\epsilon \ll 1$, μ_{min} is large and by Theorem 6, large eigenvalues of $BA^{-1}B^T$ and uniform meshes produce a tight cluster of negative eigenvalues. For more general problems with discontinuous coefficients, scaling with respect to the smallest coefficient will have to be performed to ensure the iteration is efficient. The magnitude of the scaling parameter needed will be fixed by the problem at hand.

3.3 Preconditioned MINRES

To illustrate the practical implications of the above discussion, we now report on MINRES convergence for a range of coefficients. The iteration counts listed are for exact preconditioning. Thus,

$$P = \begin{pmatrix} A+D & 0 \\ 0 & N \end{pmatrix}, \qquad (3.38)$$

is factorised. Iteration counts for the *unpreconditioned* experiments are given in parentheses.

Consider problem (1.4) discretised on $\Omega = [0,1] \times [0,1]$. Uniform square meshes are used, unless stated otherwise. We apply MINRES to the assembled system with zero initial guess, and terminate the iteration when the residual error $\underline{r}^{(i)}$ satisfies,

$$\frac{\| \underline{r}^{(i)} \|_2}{\| \underline{r}^{(0)} \|_2} \leq 10^{-6}. \tag{3.39}$$

All experiments were performed with MATLAB 6.0 on a SUN ultraSPARC workstation. The symbol $*$ indicates that more than 500 iterations were required.

Example 1

We begin with a trivial case. Choose $\mathcal{A} = \mathcal{I}$, $f = 1$ and a homogenous Dirichlet boundary condition. Iteration counts are given in Table 7.1 and confirm the h-optimality of the preconditioner.

h	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$
	5 (25)	5 (75)	5 (165)	5 (311)

Table 3.4: MINRES iterations, Example 1, Dirichlet boundary condition

A simple flow problem is induced by setting $f = 0$ and introducing *mixed* boundary conditions, $p = 1$ on $\{0\} \times [0, 1]$, $p = 0$ on $\{1\} \times [0, 1]$ and $\vec{u} \cdot \vec{n} = 0$ on $(0, 1) \times \{0, 1\}$. The corresponding iteration counts are given in Table 3.5. Imposing the essential Neumann condition (see Chapter 2) has no impact on the performance of the preconditioner.

h	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$
	6 (18)	6 (34)	6 (66)	6 (130)

Table 3.5: MINRES iterations, Example 1, mixed boundary conditions

Example 2

For non-diagonal tensors, convergence is completely determined by the eigenvalues of $\mathcal{A}$. To demonstrate this, we solve the model problem with $f = 1$ and,

$$\mathcal{A} \;=\; \begin{pmatrix} 1 + 4\left(x^2 + y^2\right) & 3xy \\ 3xy & 1 + 11\left(x^2 + y^2\right) \end{pmatrix}. \tag{3.40}$$

Iteration counts are given in Table 3.6.

h	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$
	5	5	5	5
	(191)	(426)	(*)	(*)

Table 3.6: MINRES iterations, Example 2, full coefficient tensor

Choosing $\mathcal{A}$ to be the diagonal tensor $\mathcal{A} = \texttt{diag}(\lambda_1, \lambda_2)$, with entries corresponding to the eigenvalues of (3.40),

$$\lambda_1 = \frac{1}{2}\left(2 + 15\left(x^2+y^2\right) + \left(49\left(x^2+y^2\right)^2 + 36x^2y^2\right)^{\frac{1}{2}}\right),$$
$$\lambda_2 = \frac{1}{2}\left(2 + 15\left(x^2+y^2\right) - \left(49\left(x^2+y^2\right)^2 + 36x^2y^2\right)^{\frac{1}{2}}\right),$$

yields the iteration counts in Table 3.7. Observe that convergence does not deteriorate in the non-diagonal case.

h	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$
	5	5	5	5
	(146)	(343)	(*)	(*)

Table 3.7: MINRES iterations, Example 2, diagonal coefficient tensor

Example 3

Next, consider a problem with a variable diagonal coefficient tensor whose entries vary by three orders of magnitude across the domain. Choose,

$$\mathcal{A} \quad = \quad \begin{pmatrix} \frac{1}{1+1000(x^2+y^2)} & 0 \\ 0 & \frac{1}{1+1000(x^2+y^2)} \end{pmatrix}. \tag{3.41}$$

Iteration counts are listed in Table 3.8.

h	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$
	24	27	27	27
	(442)	(*)	(*)	(*)

Table 3.8: MINRES iterations, Example 3, variable coefficient tensor

Analysis of the negative eigenvalues reveals that this problem is more challenging than the previous examples due to the small magnitude of the entries of the coefficient tensor in some parts of the domain. The iteration count rises because μ_{min} is smaller here.

Example 4

Now consider the anisotropic test problem from section 3.2.1 with $\mathcal{A} = \texttt{diag}(\epsilon, 1)$. Using square uniform meshes we achieve both h-optimality and $\mathcal{A}$-optimality. The iteration counts listed in Table 3.9 below are perfectly consistent with the eigenvalue clusters shown in Fig. 3.2. Note that to resolve boundary layers in the solution when homogeneous Dirichlet boundary conditions are present, anisotropic meshes should be used. The experiment is simply to show the impact of ϵ with respect to the shape of the elements. For triangles, convergence completely breaks down if $\epsilon << 1$.

$\epsilon \quad h$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$
10^6	4	4	4	4
10^5	4	4	4	4
10^4	4	4	4	4
10^3	4	4	4	4
10^2	5	5	5	5
10^1	6	6	6	6
10^{-1}	7	7	7	7
10^{-2}	8	8	8	8
10^{-3}	8	8	8	8
10^{-4}	7	7	8	8
10^{-5}	7	7	7	7
10^{-6}	6	7	7	7

Table 3.9: MINRES iterations, Example 4, anisotropic coefficient tensor

If the anisotropy is not aligned with the coordinate axes, then there is no benefit in using rectangular elements. For example, choosing,

$$\mathcal{A} = \begin{pmatrix} 1 & \epsilon \\ \epsilon & 1 \end{pmatrix}, \tag{3.42}$$

with $\epsilon^2 \neq 1$, produces off-diagonal anisotropy as $\epsilon \to 1$. The problem is best tackled using meshes of uniform triangles with diagonal edges aligned to the direction of anisotropy. Iteration counts for triangular and square meshes are listed in Tables 3.10 and 3.11, respectively.

ϵ h	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$
0.9	6	6	6	6
0.99	6	6	6	6
0.999	7	6	6	6
0.9999	7	7	7	7

Table 3.10: MINRES iterations, Example 4, anisotropic coefficient tensor, triangles

ϵ h	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$
0.9	6	6	6	6
0.99	7	7	6	6
0.999	10	10	9	9
0.9999	13	16	17	17

Table 3.11: MINRES iterations, Example 4, anisotropic coefficient tensor, squares

Example 5

Next, we perform a discontinuous coefficient experiment. Choose $\mathcal{A}$ as in (3.36) with $\epsilon \in \left[10^{-6}, 10^{6}\right]$. Take $\Omega_* = [0.25, 0.75] \times [0.25, 1]$, $f = 0$ and mixed boundary conditions: $\vec{u} \cdot \vec{n} = 0$ on $\partial\Omega_N$ and $p = 1 - x$ on $\partial\Omega_D$ with $\partial\Omega_N = [0,1] \times 0 \cup \{0,1\} \times [0, 0.75]$ and $\partial\Omega_D = \Omega\backslash\Omega_N$. The pressure contours and velocity fields obtained for $\epsilon = 10^6$ and $\epsilon = 10^{-6}$ are shown in Figs. 3.4–3.5.

Without scaling, the iteration counts listed in Table 3.12 deteriorate as $\epsilon \to 0$. This behaviour is consistent with the eigenvalues shown in the right plot in Fig. 3.3. Again, this is due to the small magnitude of μ_{min}. However, if for $\epsilon < 1$, we solve the rescaled problem as discussed in section 3.2.1 by applying the coefficients (3.37), we obtain the iteration counts listed in Table 3.13. This behaviour is consistent with the eigenvalues shown in the left plot in Fig. 3.3.

The accuracy of the solution is not unaffected by the rescaling but we can compensate for this cheaply by iterating to a smaller tolerance. In this example, we apply a stopping tolerance of 10^{-9}. This is sufficient to ensure that the velocity solution to the rescaled problem is the same as that of the original problem to 8 decimal places for the smallest value of ϵ considered.

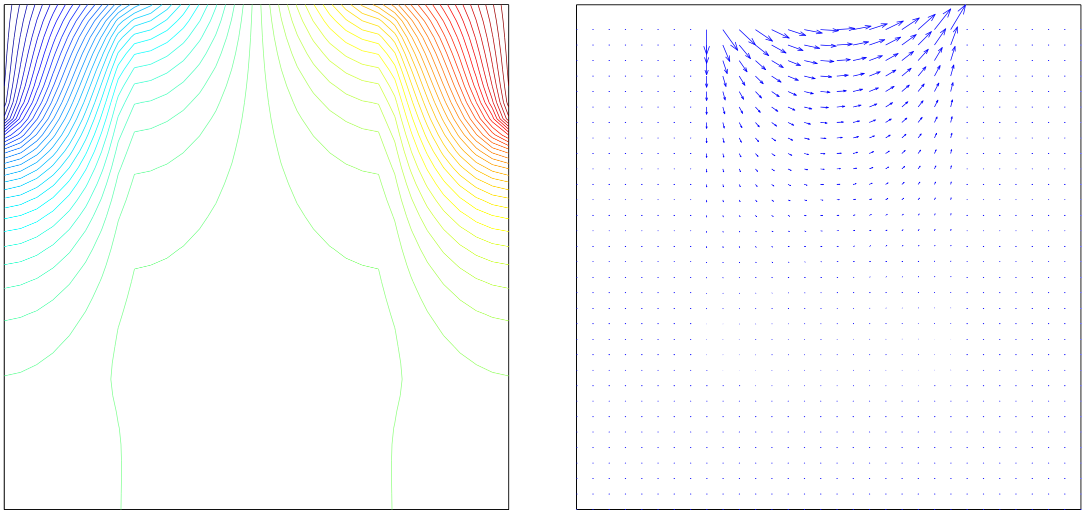

Figure 3.4: Pressure contours (left) and velocity field (right), $\epsilon = 10^6$

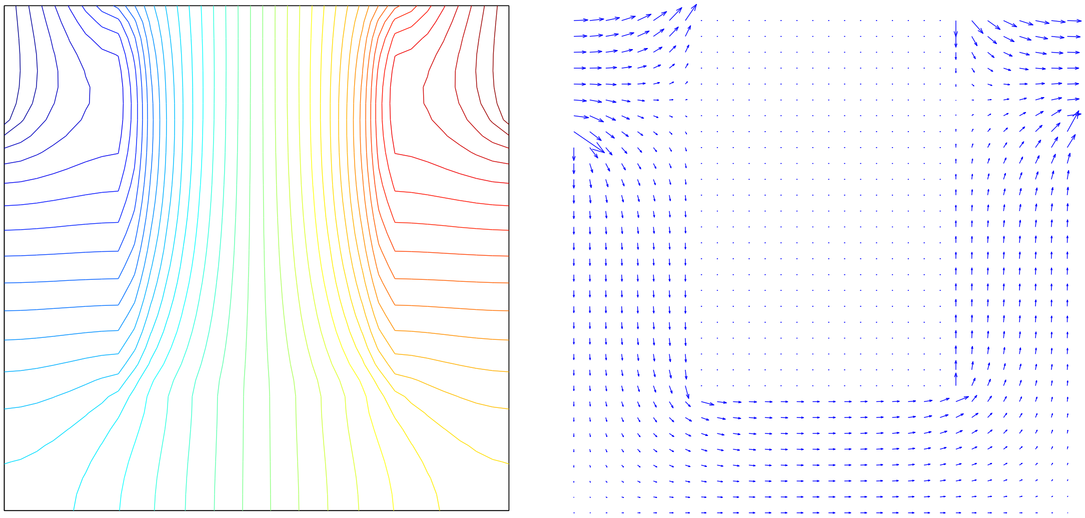

Figure 3.5: Pressure contours (left) and velocity field (right) , $\epsilon = 10^{-6}$

$\epsilon \quad h$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$
10^6	7	7	7
10^5	7	7	7
10^4	7	7	7
10^3	7	7	7
10^2	7	8	8
10^1	7	7	7
10^{-1}	9	9	9
10^{-2}	15	15	15
10^{-3}	30	30	32
10^{-4}	64	77	82
10^{-5}	89	143	191
10^{-6}	105	187	308

Table 3.12: MINRES iterations, Example 5, discontinuous coefficient tensor, unscaled

$\epsilon \quad h$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$
10^6	7	7	7
10^5	7	7	7
10^4	7	7	7
10^3	7	7	7
10^2	7	8	8
10^1	7	7	7
10^{-1}	7	7	7
10^{-2}	6	6	7
10^{-3}	6	7	7
10^{-4}	6	6	6
10^{-5}	6	6	6
10^{-6}	6	6	6

Table 3.13: MINRES iterations, Example 5, discontinuous coefficient tensor, scaled

Example 6

Finally, for a more challenging discontinuous coefficient example, we consider the so-called 'Kellogg problem' (see Kellogg [63]). The test problem we consider is one suggested by Morin et al. in [71]. Set $\Omega = [-1, 1] \times [-1, 1]$ and $f = 0$. $\mathcal{A}$ is chosen as $a_1\mathcal{I}$ in the first and third quadrants of Ω and $a_2\mathcal{I}$ in the second and fourth quadrants, so that the analytical pressure solution, in polar co-ordinates, is $p(r, \theta) = r^\gamma \mu(\theta)$, where,

$$\mu(\theta) \;=\; \begin{cases} \cos((\frac{\pi}{2}-\sigma)\gamma)\cdot\cos((\theta-\frac{\pi}{2}+\rho)\gamma) & 0\le\theta\le\frac{\pi}{2}, \\ \cos(\rho\gamma)\cdot\cos((\theta-\pi+\sigma)\gamma) & \frac{\pi}{2}\le\theta\le\pi, \\ \cos(\sigma\gamma)\cdot\cos((\theta-\pi-\rho)\gamma) & \pi\le\theta\le\frac{3\pi}{2}, \\ \cos((\frac{\pi}{2}-\rho)\gamma)\cdot\cos((\theta-\frac{3\pi}{2}-\sigma)\gamma) & \frac{3\pi}{2}\le\theta\le 2\pi, \end{cases}$$

and the constants σ, ρ and γ are chosen to satisfy the non-linear relations,

$$\begin{aligned} R = \frac{a_1}{a_2} \;&=\; -\tan((\frac{\pi}{2}-\sigma)\gamma)\cdot\cot(\rho\sigma), \\ \frac{1}{R} \;&=\; -\tan(\rho\sigma)\cdot\cot(\sigma\gamma), \\ R \;&=\; -\tan(\sigma\gamma)\cdot\cot((\frac{\pi}{2}-\rho)\gamma), \\ 0 \;&<\; \gamma \;<\; 2, \\ \max\{0,\pi\gamma-\pi\} \;&<\; 2\gamma\rho \;<\; \min\{\pi\gamma,\pi\} \\ \max\{0,\pi-\pi\gamma\} \;&<\; -2\gamma\sigma \;<\; \min\{\pi,2\pi-\pi\gamma\}. \end{aligned}$$

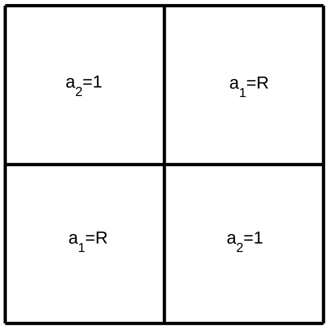

Figure 3.6: 2 × 2 checkerboard coefficient ordering

Following [71], we set $\gamma = 0.1$, producing a singular solution at the origin. Solving for the various constants using a Newton iteration, yields,

$$R = \frac{a_1}{a_2} \approx 161.4476387, \quad \rho = \frac{\pi}{4}, \quad \sigma \approx -14.922565105.$$

We then set $a_1 = R$ and $a_2 = 1$ (see Fig. 3.6.)

The interpolant of the exact solution on a uniform mesh with $h = \frac{1}{32}$ is shown (rotated though 90 degrees, clockwise, about the origin) in Fig. 3.7.

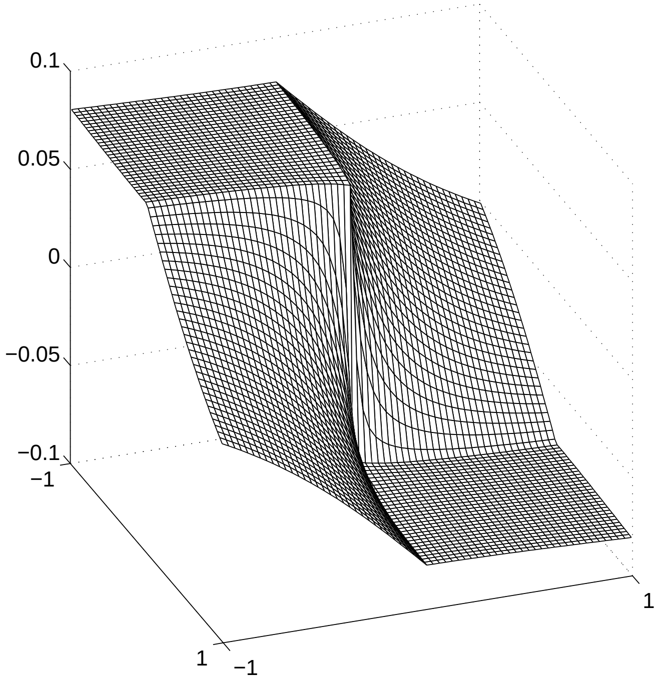

Figure 3.7: Interpolant of exact solution, Kellogg problem

To accurately capture the singularity at the origin, locally refined meshes should be used. For this preconditioning experiment, we use adaptive meshes generated by the ALBERT toolbox, [2]. Examples are shown in Fig. 3.8, below. Iteration counts corresponding to discretisations with varying levels of local refinement are listed in Table 3.14. Mesh-independent MINRES convergence is achieved.

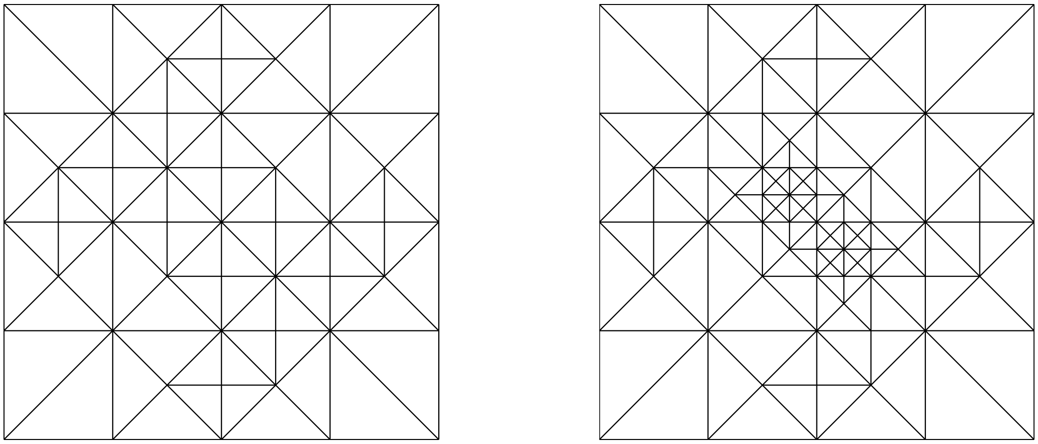

Figure 3.8: Locally adapted meshes

The observed eigenvalues of the preconditioned system are listed in Table 3.15. They

$\|K\|_{min}$	$\|K\|_{max}$	Iterations	
2^{-8}	2^{-2}	5	(65)
2^{-10}	2^{-2}	5	(91)
2^{-12}	2^{-4}	5	(128)
2^{-16}	2^{-4}	4	(196)
2^{-18}	2^{-8}	4	(269)
2^{-20}	2^{-8}	4	(339)

Table 3.14: MINRES iterations, Example 6, discontinuous coefficient tensor, adapted meshes

remain bounded independently of the mesh parameters, h_{min} and h. The right hand bound for the negative eigenvalues in Theorem 6 decays to zero with mesh refinement. This is clearly too pessimistic, as predicted.

$\|K\|_{min}$	$\|K\|_{max}$	Eigenvalues
2^{-8}	2^{-2}	$[-0.9999,\ -0.9244] \cup [1]$
2^{-10}	2^{-2}	$[-0.9999,\ -0.9284] \cup [1]$
2^{-12}	2^{-4}	$[-0.9999,\ -0.9238] \cup [1]$
2^{-16}	2^{-4}	$[-0.9999,\ -0.9280] \cup [1]$

Table 3.15: Observed eigenvalues, Example 6, Kellogg problem

3.4 Concluding remarks

In this chapter, we proposed the ideal, parameter-free preconditioner (3.38) with diagonal blocks representing the norms $\|\cdot\|_{div,\mathcal{A}}$ and $\|\cdot\|_0$ on the lowest order Raviart-Thomas spaces V_h and W_h. New inclusion intervals for the eigenvalues of the preconditioned system matrix were derived and we made specific the impact of the discretisation parameter h and general coefficient tensors $\mathcal{A}$ on those bounds. A range of numerical examples were performed in $I\!R^2$ to illustrate the theory.

For uniform and quasi-uniform meshes, our theoretical eigenvalue bounds are tight. The suggested preconditioner is h-optimal. The impact of the coefficient is not always trivial; anisotropic and discontinuous coefficients produce sub-optimal eigenvalue bounds. However, in some cases, the theoretical bounds are overly pessimistic and the difficulty can be overcome with the appropriate mesh or scaling with respect to PDE coefficients. In this way, $\mathcal{A}$-optimality can be also be achieved.

For non quasi-uniform meshes, the same inclusion intervals for the eigenvalues apply. However, the right-hand bound for the negative eigenvalues is too pessimistic. Numerical evidence suggests that mesh independent MINRES convergence is achieved with the suggested preconditioner even for highly non-uniform meshes.

Chapter 4

Practical $H(div)$ preconditioning

To obtain a practical scheme, the ideal preconditioner,

$$P \;=\; \begin{pmatrix} A+D & 0 \\ 0 & N \end{pmatrix} \tag{4.1}$$

must be implemented in a cost-effective way. Specifically, we require that the action of the inverse of P can be computed in only $O(n)$ flops, where n is the dimension of the system. Recall that the leading block, $A+D$, arises in the weighted $H(div)$ inner-product,

$$(\vec{u}_h,\, \vec{v}_h)_{div,\mathcal{A}} \;=\; \left(\mathcal{A}^{-1}\,\vec{u}_h,\, \vec{v}_h\right) + \left(\nabla\cdot\vec{u}_h, \nabla\cdot\vec{v}_h\right) \;=\; \underline{u}^T\,(A+D)\,\underline{v}. \tag{4.2}$$

In the sequel, we shall adopt the matrix notation $H \;=\; A+D$. We also define the associated PDE operator, $\mathcal{H}$, via,

$$(\vec{u}_h,\, \vec{v}_h)_{div,\mathcal{A}} \;=\; (\mathcal{H}\vec{u}_h,\, \vec{v}_h)\,. \tag{4.3}$$

We say that $\mathcal{H}$ is the 'pure' $H(div)$ operator if the coefficient tensor $\mathcal{A}$ is the identity matrix.

Implementing (4.1) cheaply is not easy since it is not a trivial task to compute the action of the inverse of H. There is no difficulty with the mass matrix, N, however, since it is a diagonal matrix. Now that we have a handle on the theoretical properties of the ideal preconditioner, we look to multigrid schemes to *approximately* invert the matrix H. This is not straightforward since the corresponding $\mathcal{H}$ is not a full elliptic operator, to which multigrid methods are best suited. However, there exists a special class of methods that have been suggested for the *pure* $H(div)$ operator. We will discuss a particular multigrid approximation introduced by Arnold, Falk and Winther in [6] and [7].

4.1 Multigrid

Multigrid methods were introduced in the 1960s as a means to solve finite difference equations arising in approximations of the Laplacian operator. They became widely used in the 1970s for solving finite element equations associated with other elliptic operators. This revolution is attributed to a work of Brandt, [16], who popularised the notion of 'multi-level adaptive solution techniques'. It is a powerful philosophy based on the observation that exploiting a hierarchy of coarser grids i.e. discretisations of the same PDE on different geometries, can lead to a solution scheme requiring only $O(n)$ flops, where n is the dimension of the problem to be solved. Today, 'multigrid' encompasses many families of different methods but all share two fundamental components: 'smoothing' and coarse-grid correction.

To understand these concepts, suppose that it is required to solve, in operator form, the symmetric and positive definite system,

$$\mathcal{M}_J x_J = b_J, \tag{4.4}$$

arising in the discretisation of a self-adjoint *elliptic* PDE, on a given mesh T_J. *Standard geometric* multigrid methods assume the existence of a nested sequence of $J+1$ solution spaces, $V_0 \subset V_1 \subset \ldots V_J = V_h$, associated with a set of uniformly refined grids, $T_0 \subset T_1 \subset \ldots T_J = T_h$. Nested solution spaces and uniform refinement are not necessary in general but in this chapter we shall only consider these cases.

Stationary iterative methods, (see section 1.2.1), perform efficiently on (4.4) for a *few* iterations but stall when oscillatory components of the error have been damped out. This phenomenon can be easily observed by applying Gauss-Seidel, say, to an elliptic problem with zero righthand side and initial guess comprising high and low frequency Fourier modes. The method is called a 'smoother' because high frequency modes are rapidly damped out. If we terminate the iteration once this is achieved, say after m iterations, the error $e_J^{(m)} = x_J - x_J^{(m)}$, where $x_J^{(m)}$ is the mth iterate, is geometrically smooth, and can be well represented on a coarser grid $T_{J-1} = T_H$, with mesh width $H > h$. A standard choice is $H = 2h$. If we construct a restriction operator $\mathcal{I}_J^{J-1} : V_J \to V_{J-1}$ (a matrix mapping vectors defined on T_J to vectors on T_{J-1}), then we can restrict the residual error $r_J^{(m)} = b_J - \mathcal{M}_J x_J^{(m)}$ to T_{J-1} and solve the error equations $M_{J-1} e_{J-1}^{(m)} = \mathcal{I}_J^{J-1} r_J^{(m)}$, on the coarser grid.

Smoothing on T_{J-1} is more efficient because $\mathcal{I}_J^{J-1} r_J^{(m)}$ appears more oscillatory than $r_J^{(m)}$. Moreover, iteration is cheaper because there are fewer unknowns. For a two grid method, the error equations are solved exactly on T_{J-1}, and an interpolated error is used to correct the initial approximation $x_J^{(m)}$ to x_J on T_J. If a full sequence of grids is available, we recursively combine smoothing, restriction and coarse grid correction on the whole set, solving the error equations exactly only on the coarsest grid T_0. A basic algorithm, employing ν_1 pre-smoothing and ν_2 post-smoothing steps, is described in Fig. 4.2.

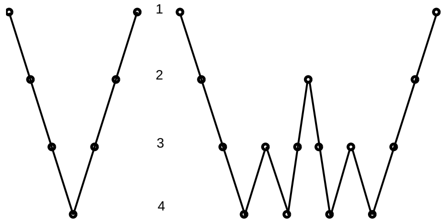

Figure 4.1: Standard V-cycle and W-cycle on 4 levels

The order in which the grids are visited gives rise to the terms 'V-cycle' and 'W-cycle' (see Fig. 4.1.) Crucially, it can be established (see, for example [95, p.74]), in $\mathbb{R}^2$ and $\mathbb{R}^3$, that each cycle can be performed in $O(N_J)$ flops, where N_J is the dimension of the system at the finest level. Computational work increases only linearly with respect to the problem size.

```
$q = J$, $x_q^{(0)}$ = zeros

function $v$ = mg_cycle($\mathcal{M}_q, b_q, x_q^{(0)}$)

    for $j = 1 : \nu_1$, smooth:                  $x_q^{(j)} = \mathcal{S} x_q^{(j-1)}$

    restrict:                                     $r_{q-1} = \mathcal{I}_q^{q-1} \left( b_q - \mathcal{M}_q x_q^{(\nu_1)} \right)$

    if $q = 1$ solve:                             $e_{q-1} = \left( \mathcal{M}_{q-1} \right)^{-1} r_{q-1}$

    elseif $q > 1$ initialise:                    $x_{q-1}^{(0)}$ = zeros

    call $p$ times:                               $v_{q-1} \approx e_{q-1}$ = mg_cycle ($\mathcal{M}_{q-1}, r_{q-1}, x_{q-1}^{(0)}$)

    correct                                       $x_q^{(\nu_1+1)} = x_q^{(\nu_1)} + \mathcal{I}_{q-1}^q v_{q-1}$

    for $j = \nu_1 + 1 : \nu_1 + \nu_2 + 1$, smooth:  $x_q^{(j)} = \mathcal{S} x_q^{(j-1)}$

    update                                        $v = x_q^{(\nu_1+\nu_2+1)}$

end function
```

Figure 4.2: Standard multigrid V-cycle ($p = 1$), W-cycle ($p = 2$)

The convergence of *any* multigrid algorithm is determined by the interaction of the chosen smoother and the coarse grid correction. Error not reduced by the smoother, $\mathcal{S}$, must lie in the range of the chosen interpolation operator $\mathcal{I}_{J-1}^J$. Error components that do not lie in the range of the interpolation operator, must be efficiently reduced by the smoother. A standard choice for interpolation is

$$\mathcal{I}_{J-1}^J = \left(\mathcal{I}_J^{J-1} \right)^T .$$

Choosing the coarse grid operator as,

$$M_{J-1} = \mathcal{I}_J^{J-1} M_J \mathcal{I}_{J-1}^J,$$

is known as *Galerkin approximation* and ensures that the coarse grid correction $e_{J-1}^{(m)}$ has minimal norm, (in the norm induced by M_{J-1}), over the space of all corrections v_{J-1} lying in the range of the interpolation operator.

Rigorous convergence theory is available for many schemes. For the nested approach, fundamental contributions were made in $I\!R^2$ by Bank and Dupont, [12], for the W-cycle and by Braess and Hackbush, [18], for the V-cycle. For a general discussion of multigrid methods see Briggs et al., [29], or Braess, [17]. For an introduction to convergence theory, and a review of the state of the art see Trottenberg et al., [95].

When solving problems in $I\!R^3$, the generation of large amounts of geometric information on nested grids is infeasible. As a remedy to this, *algebraic* multigrid methods (AMG), which generate coarse levels and transfer operators by exploiting only the stencil of the given matrix, were developed in the 1980s (see, for example, Ruge and Stüben [83], [84]). Today, they are becoming increasingly popular in many important applications (see, for example, [70], [81]). However, convergence theory for AMG is much less developed than for the geometric case and is limited to so-called M-matrices which are characterised by large positive diagonal entries and negative off-diagonal entries. We shall discuss AMG in more detail in Chapter 5.

4.2 Multigrid in $H(div;\Omega)$

Now, we require a multigrid approximation $\mathcal{V}$ to the operator $\mathcal{H} : V_h \to V_h$ arising in the inner-product (4.3), such that $\kappa\left(\mathcal{V}^{-1}\mathcal{H}\right)$ is bounded independently of the discretisation parameter h and the coefficient tensor $\mathcal{A}$. To begin, we note that $\mathcal{H}$ is a symmetric and positive definite operator.

Lemma 16 *$\mathcal{H}$ is a symmetric and positive definite operator with respect to the $L^2(\Omega)$ inner-product.*

Proof Since we have assumed that the coefficient tensor $\mathcal{A}$ is positive definite, it is a trivial consequence that $(\mathcal{H}\vec{u}_h, \vec{u}_h) = \left(\mathcal{A}^{-1}\vec{u}_h, \vec{u}_h\right) + (\nabla\cdot\vec{u}_h, \nabla\cdot\vec{u}_h) > 0$ for all $\vec{u}_h$ in $V_h\backslash\{\vec{0}\}$. Since we have also assumed that $\mathcal{A}$ is symmetric, we obtain for all $\vec{u}_h, \vec{v}_h \in V_h$,

$$\begin{aligned}(\mathcal{H}\vec{u}_h, \vec{v}_h) &= \left(\mathcal{A}^{-1}\vec{u}_h, \vec{v}_h\right) + (\nabla\cdot\vec{u}_h, \nabla\cdot\vec{v}_h) \\ &= \left(\vec{u}_h, \mathcal{A}^{-1}\vec{v}_h\right) + (\nabla\cdot\vec{u}_h, \nabla\cdot\vec{v}_h) = (\mathcal{H}\vec{v}_h, \vec{u}_h)\,. \quad \Box\end{aligned}$$

Unfortunately, standard multigrid methods described in the previous section are unsuitable (see Cai et al., [31], for numerical evidence of this) since $\mathcal{H}$ lacks some of the characteristics of elliptic operators that are necessary to obtain h-optimal convergence. A first suspicion of this can be gleaned from the stencil of the matrix corresponding to the so-called pure $H(div)$ operator.

For example, the element basis functions for $RT_0(K)$, on a square of edge length h, with oriented normal vectors as shown in Fig. 2.5, and edge mid-side co-ordinates (x_i, y_i), $i = 1:4$, are,

$$\vec{\varphi}_K^1 = \begin{pmatrix} 1 + \frac{x_1}{h} - \frac{x}{h} \\ 0 \end{pmatrix}, \vec{\varphi}_K^2 = \begin{pmatrix} -\frac{x_1}{h} + \frac{x}{h} \\ 0 \end{pmatrix}, \vec{\varphi}_K^3 = \begin{pmatrix} 0 \\ 1 + \frac{y_3}{h} - \frac{y}{h} \end{pmatrix}, \vec{\varphi}_K^4 = \begin{pmatrix} 0 \\ -\frac{y_3}{h} + \frac{y}{h} \end{pmatrix}.$$

Integrating, we find that the element contributions to the matrices A and D are,

$$A^K = \frac{h^2}{4} \begin{pmatrix} \frac{4}{3} & \frac{2}{3} & 0 & 0 \\ \frac{2}{3} & \frac{4}{3} & 0 & 0 \\ 0 & 0 & \frac{4}{3} & \frac{2}{3} \\ 0 & 0 & \frac{2}{3} & \frac{4}{3} \end{pmatrix}, \quad D^K = \begin{pmatrix} +1 & -1 & +1 & -1 \\ -1 & +1 & -1 & +1 \\ +1 & -1 & +1 & -1 \\ -1 & +1 & -1 & +1 \end{pmatrix}.$$

Adding them together produces the matrix stencils shown in Fig 4.3.

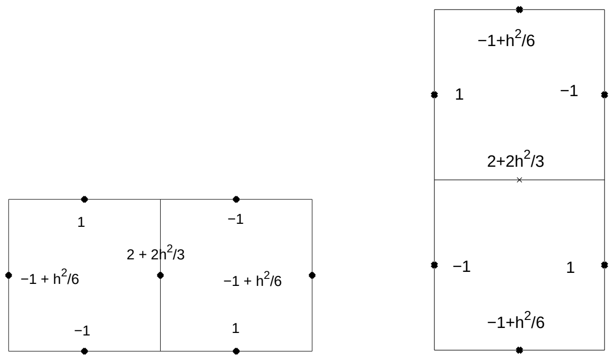

Figure 4.3: Stencil of H^K, squares

Similarly, for uniform right-angled triangles, with oriented normal vectors as shown in Fig. 2.4, and vertex co-ordinates (x_i, y_i), $i = 1:3$, we obtain element basis functions,

$$\vec{\varphi}_K^1 = \begin{pmatrix} -\frac{x_3}{h} + \frac{x}{h} \\ -1 - \frac{y_1}{h} + \frac{y}{h} \end{pmatrix}, \vec{\varphi}_K^2 = \begin{pmatrix} -\frac{x_3\sqrt{2}}{h} + \frac{x\sqrt{2}}{h} \\ -\frac{y_1\sqrt{2}}{h} + \frac{y\sqrt{2}}{h} \end{pmatrix}, \vec{\varphi}_K^3 = \begin{pmatrix} 1 + \frac{x_3}{h} - \frac{x}{h} \\ \frac{y_1}{h} - \frac{y}{h} \end{pmatrix},$$

yielding,

$$A^K = h^2 \begin{pmatrix} \frac{1}{3} & 0 & \frac{1}{6} \\ 0 & \frac{1}{3} & 0 \\ \frac{1}{6} & 0 & \frac{1}{3} \end{pmatrix}, \quad D^K = \begin{pmatrix} 2 & 2\sqrt{2} & -2 \\ 2\sqrt{2} & 4 & -2\sqrt{2} \\ -2 & -2\sqrt{2} & 2 \end{pmatrix}.$$

Matrix stencils for the sum H^K are given in Fig. 4.4.

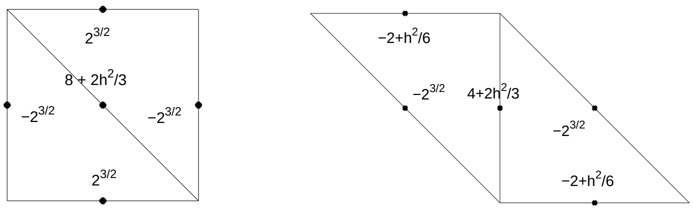

Figure 4.4: Stencil of H^K, triangles

An important observation is that we obtain positive and negative off-diagonal entries in each H^K. The global matrix H is certainly not an M-matrix, and, moreover, does not have the appearance of a standard discrete elliptic operator.

Now, error modes that cannot be damped out by standard smoothers such as Gauss-Seidel or Jacobi iteration are attributed to eigenfunctions associated with the eigenvalues closest to the origin. For elliptic problems, these eigenvalues correspond to geometrically smooth eigenfunctions which can be well represented on coarser grids. In contrast, standard smoothing for matrices with significant positive off-diagonal entries, such as H, leaves oscillatory error components (see Stüben, [93]) that do not lie in the range of standard interpolation operators. In such cases, multigrid with stationary iterative smoothers fails.

At a more abstract level, this 'non-elliptic' behaviour is a consequence of the non-trivial null space of the divergence operator. To gain insight into this, recall first that any $\vec{v} \in L^2(\Omega)^d$ satisfying $\nabla \cdot \vec{v} = 0$ can be represented as a curl field.

Theorem 7 *Let Ω be a bounded and simply-connected domain with Lipschitz continuous boundary. A vector-field $\vec{v} \in L^2(\Omega)^d$ satisfies $\nabla \cdot \vec{v} = 0$ in Ω if and only if there exists a $\vec{z} \in H^1(\Omega)^d$ satisfying $\vec{v} = \nabla \times \vec{z}$.*

Proof See Girault and Raviart, [54, Theorem 1.3.4].

Under the same assumptions on Ω we can also invoke the Helmholtz decomposition of an arbitrary vector $\vec{v} \in L^2(\Omega)^d$ into a gradient field and a curl field.

Theorem 8 *Every $\vec{v} \in L^2(\Omega)^d$, $d = 2, 3$, has a unique orthogonal decomposition,*

$$\vec{v} = \nabla w + \nabla \times \vec{z},$$

where $w \in H^1(\Omega) \backslash I\!R$ and $\vec{z} \in H^1(\Omega)^d$.

Proof See Theorem 1.3.2 and Corollary 1.3.4 in Girault and Raviart, [54].

For the lowest order triangular or tetrahedral Raviart-Thomas elements, we obtain the discrete orthogonal decomposition,

$$V_h = \nabla_h W_h \oplus \nabla \times S_h,$$

where $S_h = \{\, s \in H^1(\Omega) \mid s|_K \in P_1(K) \,\}$ is the set of continuous piecewise linear polynomials. Since the pressure space $W_h \not\subset H^1(\Omega)$, gradient fields are defined via,

$$(\nabla_h w_h, \vec{v}_h) \;=\; -\,(w_h, \nabla \cdot \vec{v}_h)\,.$$

Now, when restricted to the divergence-free subspace $\nabla \times S_h \subset V_h$, we obtain, for all $\vec{u}_h = \nabla \times s_1$ and $\vec{v}_h = \nabla \times s_2$,

$$\begin{aligned} (\mathcal{H}\,\vec{u}_h,\, \vec{v}_h) &= (\mathcal{H}\,\nabla \times s_1,\, \nabla \times s_2,) \\ &= \left(\mathcal{A}^{-1}\,\nabla \times s_1,\, \nabla \times s_2,\right) + (\nabla \cdot \nabla \times s_1, \nabla \cdot \nabla \times s_2) \\ &= \left(\mathcal{A}^{-1}\,\nabla \times s_1,\, \nabla \times s_2\right) \;=\; \left(\mathcal{A}^{-1}\vec{u}_h,\, \vec{v}_h\right). \end{aligned}$$

Hence, $\mathcal{H}$ behaves like the non-elliptic operator $\mathcal{A}^{-1}\mathcal{I}$.

Since $\mathcal{H}$ only fails to be elliptic on a subspace of V_h, it is no surprise that most existing multilevel approximations to operators of this form, exploit the decomposition of V_h into divergence-free and curl-free parts. In [31], Cai et al. propose a hierarchical basis preconditioner based on such a splitting. In [97], Vassilevski and Wang propose a domain decomposition method. The subject of that work is the penalty method. A domain decomposition approach is also recommended by Hiptmair in [60]. Divergence-free and curl-free subproblems are solved separately, at all levels. The same author, in [61], extends the method of [97] to the case of adaptively refined grids. However, the subject of that work is the augmented Lagrangian method in which parameters are present. None of these papers discuss general coefficient tensors $\mathcal{A}$.

The approach of Arnold, Falk and Winther, in [6] and [7], however, is set in the framework of the standard multigrid V-cycle. Only the smoother is modified. Their method offers a simpler implementation because the Helmholtz decomposition is employed only as a theoretical tool for obtaining error estimates. It is not physically performed. We give an outline of the method in the next section. Readers who are not interested in multigrid theory can skip to the next section to find details of the practical implementation.

4.3 Arnold-Falk-Winther multigrid

To investigate the performance of the method of Arnold et al., for our model problem (2.8), we restrict our attention, in the remainder of this chapter, to triangular and tetrahedral elements and shape-regular, quasi-uniform meshes. The domain Ω is assumed to be convex. Now, given a sequence of quasi-uniform meshes, $T_0 \subset T_1 \subset \dots T_J = T_h$, the corresponding velocity spaces V_j form a nested sequence of subspaces of $L^2(\Omega)$. That is,

$$V_0 \subset V_1 \subset \dots V_J = V_h \subset L^2(\Omega).$$

We define, for each level, $j = 0 : J$, the operator $\mathcal{H}_j : V_j \to V_j$ via,

$$(\mathcal{H}_j \vec{u}, \vec{v}) = (\vec{u}, \vec{v})_{div,\mathcal{A}} \quad \forall \vec{u}, \vec{v} \in V_j.$$

By Lemma 16, each $\mathcal{H}_j$ is a symmetric and positive definite operator. To obtain a multigrid algorithm, we require the L^2-projection operator $\mathcal{Q}_j : V_h \to V_j$ defined via,

$$(\mathcal{Q}_j \vec{u}, \vec{v}) = (\vec{u}, \vec{v}) \quad \forall \vec{u} \in V_h, \forall \vec{v} \in V_j,$$

and the $\mathcal{H}$-orthogonal projection operator $P_j : V_h \to V_j$, defined via,

$$(P_j \vec{u}, \vec{v})_{div,\mathcal{A}} = (\vec{u}, \vec{v})_{div,\mathcal{A}} \quad \forall \vec{u} \in V_h, \forall \vec{v} \in V_j.$$

We denote the generic smoothing operator by $\mathcal{S}_j : V_j \to V_j$.

Our aim is to solve a system of the form,

$$\mathcal{H}_J x_J = f_J,$$

for $x_J \in V_J$. Let $\mathcal{V}_J^{-1}$ denote the application of a V-cycle of multigrid to this system. If we employ the algorithm in Fig. 4.2, with m pre-smoothing and m post-smoothing steps, with Galerkin approximation and restriction given by $\mathcal{Q}_j$, then an h-optimal algorithm results, provided we can construct a set of smoothers $\{\mathcal{S}_j\}_{j=0}^J$ satisfying the conditions of the following theorem.

Theorem 9 *Suppose that, for $j = 0 : J$, the smoother $\mathcal{S}_j$ is L^2-symmetric and positive semidefinite and satisfies the conditions,*

$$\begin{aligned} ([\mathcal{I} - \mathcal{S}_j \mathcal{H}_j]\, \vec{v}, \vec{v})_{div,\mathcal{A}} &\geq 0 \quad \forall \vec{v} \in V_j, &(4.5)\\ \left(\mathcal{S}_j^{-1} \vec{v}, \vec{v}\right)_{div,\mathcal{A}} &\leq \alpha\, (\vec{v}, \vec{v})_{div,\mathcal{A}} \quad \forall \vec{v} \in (\mathcal{I} - \mathcal{P}_{j-1})\, V_j, &(4.6) \end{aligned}$$

where α is a positive constant, then

$$0 \leq \left([\mathcal{I} - \mathcal{V}_J^{-1} \mathcal{H}_J]\, \vec{v}, \vec{v}\right)_{div,\mathcal{A}} \leq \delta\, (\vec{v}, \vec{v})_{div,\mathcal{A}} \quad \forall \vec{v} \in V_J, \qquad (4.7)$$

where $\delta = \frac{\alpha}{(\alpha + 2m)}$ and m is the number of pre-smoothing steps.

Proof This result is established by Arnold et al. in [6, Appendix B] and is a modification of a more general theory due to Bramble [15, Theorem 3.6]. □

Remark 10 *If the conditions of Theorem 9 are satisfied, we see that the eigenvalues of $\mathcal{V}_J^{-1}\mathcal{H}$ lie in the interval $[\,1 - \delta, 1\,]$.*

The choice of the smoother is therefore critical. The class of smoothers recommended by Arnold et al. are the so-called additive and multiplicative Schwarz methods. These schemes are based on particular geometric decompositions of the triangulations T_j into ‘patches’ which induce a partition of the spaces V_j. To understand this, suppose that there exists a partition of T_j into overlapping subdomains $\{\Omega_j^k\}$ such that each V_j can be decomposed as a sum of closed subspaces,

$$V_j = \sum_k V_j^k,$$

where, $V_j^k = \{\vec{v} \in V_j \mid supp(\vec{v}) \subset \overline{\Omega_j^k}\}$. If we define, for each grid level j, and for each subdomain k, the $\mathcal{H}$-projection operator $\mathcal{P}_j^k : V_j \to V_j^k$ via,

$$\left(P_j^k \vec{u}, \vec{v}\right)_{div,\mathcal{A}} = (\vec{u}, \vec{v})_{div,\mathcal{A}} \quad \forall \vec{u} \in V_j, \forall \vec{v} \in V_j^k,$$

then the *additive* Schwarz smoother is defined as,

$$\mathcal{S}_j = \eta \sum_k \mathcal{P}_j^k \mathcal{H}_j^{-1}, \tag{4.8}$$

where $\eta > 0$ is a scaling parameter.

To be specific, admissible domain decompositions are those for which a bound on the $L^2(\Omega)$ norm, of the form,

$$\sum_k \| \vec{v}^k \|_0^2 \leq c \| \vec{v} \|_0^2,$$

holds with a constant c independent of h. For details, see [7] or Hiptmair, [60]. One possibility, for the Raviart-Thomas spaces, is a vertex-based decomposition. Hence, set Ω_j^k to be the patch of elements surrounding vertex k in T_j (see Fig. 4.5.)

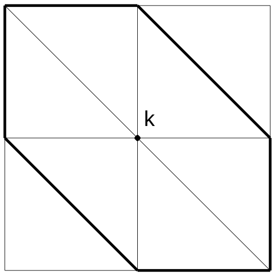

Figure 4.5: Vertex-centered patch

Then, in matrix form, each P_j^k takes the form,

$$P_j^k = R_k^T H_k^{-1} R_k H_j,$$

where H_k is the principal submatrix of H_j corresponding to the rows and columns associated with nodes lying on the edges or faces attached to vertex k. R_k is a restriction matrix with entries zero or one. Hence, applying the smoother,

$$S_j H_j = \eta \sum_k R_k^T H_k^{-1} R_k H_j = \eta \sum_k P_j^k,$$

corresponds to taking a scaled sum of solutions to local subproblems, and is equivalent to block-Jacobi iteration. Alternatively, a *multiplicative* smoother, akin to block Gauss-Seidel can be defined by solving the subproblems in sequence and updating the residuals after each local solve. We will apply the additive smoother (4.8).

In [6] and [7] it is demonstrated that the additive and multiplicative Schwarz smoothers satisfy the conditions of Theorem 9, if the bilinear form $(\cdot,\cdot)_{div,\mathcal{A}}$ is,

$$(\vec{u},\vec{v})_{div,\mathcal{A}} = \rho\,(\vec{u},\vec{v}) + (\nabla\cdot\vec{u}, \nabla\cdot\vec{v})\,,$$

for some constant $\rho > 0$. Moreover, it is established that δ is independent of ρ and the discretisation parameter h. For our model problem, this corresponds to choosing the coefficient tensor $\mathcal{A}(\vec{x}) = \rho\mathcal{I}$ for all points $\vec{x} \in \Omega$, and thus only corresponds to a trivial subset of the problems we would like to solve. However, we must not be discouraged.

Note, first, that the symmetry and positive definiteness of the smoother is completely unaffected by general, symmetric, coefficient tensors. To see that the additive smoother $\mathcal{S}_j$ is L^2-symmetric, observe that,

$$\begin{aligned}
(\mathcal{S}_j\vec{v},\vec{z}) = \eta\sum_k \left(\mathcal{P}_j^k\mathcal{H}_j^{-1}\vec{v},\vec{z}\right) &= \eta\sum_k \left(\mathcal{P}_j^k\mathcal{H}_j^{-1}\vec{v},\,\mathcal{H}_j^{-1}\vec{z}\right)_{div,\mathcal{A}} \\
&= \eta\sum_k \left(\mathcal{H}_j^{-1}\vec{v},\,\mathcal{P}_j^k\mathcal{H}_j^{-1}\vec{z}\right)_{div,\mathcal{A}} \\
&= \left(\mathcal{H}_j^{-1}\vec{v},\,\mathcal{S}_j\vec{z}\right)_{div,\mathcal{A}} \\
&= (\vec{v},\,\mathcal{S}_j\vec{z})\,.
\end{aligned}$$

It is also evident that (4.5), the first condition in Theorem 9, holds in the generic bilinear form (4.3). Indeed, the analysis of Arnold et al. can be applied without modifications. If we choose the additive smoother (4.8), we obtain, for any $\vec{v}$ in V_j,

$$\begin{aligned}
([\mathcal{I}-\mathcal{S}_j\mathcal{H}_j]\,\vec{v},\vec{v})_{div,\mathcal{A}} &= (\vec{v},\vec{v})_{div,\mathcal{A}} - (\mathcal{S}_j\mathcal{H}_j\vec{v},\vec{v})_{div,\mathcal{A}} \\
&= (\vec{v},\vec{v})_{div,\mathcal{A}} - \eta\sum_k \left(\mathcal{P}_j^k\vec{v},\vec{v}\right)_{div,\mathcal{A}} \\
&= (\vec{v},\vec{v})_{div,\mathcal{A}} - \eta\sum_k \left(\mathcal{P}_j^k\vec{v},\,\mathcal{P}_j^k\vec{v}\right)_{div,\mathcal{A}} \\
&= (\vec{v},\vec{v})_{div,\mathcal{A}} - \eta\sum_k \parallel \mathcal{P}_j^k\vec{v} \parallel_{div,\mathcal{A}}^2 .
\end{aligned}$$

Now, if we denote by $\|\cdot\|_{div,\mathcal{A},k}$ the weighted $H(div)$ norm with integration performed on the kth subdomain, we obtain,

$$\| \mathcal{P}_j^k \vec{v} \|_{div,\mathcal{A}}^2 \ = \ \left(\mathcal{P}_j^k \vec{v}, \mathcal{P}_j^k \vec{v}\right)_{div,\mathcal{A}} \ = \ \left(\mathcal{P}_j^k \vec{v}, \vec{v}\right)_{div,\mathcal{A}} \leq \| \mathcal{P}_j^k \vec{v} \|_{div,\mathcal{A}} \| \vec{v} \|_{div,\mathcal{A},k} \, .$$

Hence, $\| \mathcal{P}_j^k \vec{v} \|_{div,\mathcal{A}}^2 \leq \| \vec{v} \|_{div,\mathcal{A},k}^2$ and so,

$$\left([\mathcal{I} - \mathcal{S}_j \mathcal{H}_j]\, \vec{v}, \vec{v}\right)_{div,\mathcal{A}} \ \geq \ (\vec{v}, \vec{v})_{div,\mathcal{A}} - \eta \sum_k (\vec{v}, \vec{v})_{div,\mathcal{A},k} \ \geq \ (1 - \eta\omega)\, (\vec{v}, \vec{v})_{div,\mathcal{A}} \, ,$$

where ω is an overlap parameter denoting the maximum number of subdomains to which any point $\vec{x}$ in T_j belongs. The message is that choosing the scaling parameter η appropriately, guarantees (4.5) independently of $\mathcal{A}$.

Unfortunately, it is not possible to obtain α independent of $\mathcal{A}$ in (4.6) using the existing analysis. Intermediary results for (4.6), in [6] and [7], exploit the approximation properties of the Raviart-Thomas spaces V_j. The error bounds appearing in Lemmas 7 and 8, in Chapter 2, play a crucial role. Since the constants appearing in those bounds depend heavily on the regularity of the discrete solutions $\vec{u}_h$ and p_h, it is inevitable that the proofs contain constants that potentially blow-up for, say, highly anisotropic coefficients. As a consequence, the accuracy of the multigrid approximation described above is not currently understood for general coefficient tensors. However, this will not detract from the main message of this chapter. We will provide new eigenvalue analysis for the preconditioned saddle-point problem in the next section. For the benefit of the reader, we end this section with some hints for practical implementation in $I\!R^2$.

4.3.1 Implementation

The method is the V-cycle algorithm described in Fig 4.2, with Galerkin coarse-grid approximation. Applying the additive Schwarz smoother S_j in (4.8) to a vector, at level j, is straightforward. We require, for each vertex in the mesh T_j, the labels of the nodes lying on the edges that emanate from that vertex. Pseudo-code is given in Fig. 4.6. Here, H_j is the matrix H, corresponding to (4.3), assembled on the mesh T_j. Following [6], we choose the scaling parameter $\eta = \frac{1}{2}$.

It remains only to construct the interpolation operators $\mathcal{I}_{j-1}^j$. The restriction operators $\mathcal{I}_j^{j-1}$ are then defined as their transposes. We use so-called injection, but care has to be taken since the vectors we want to transfer between grids represent *normal*

```
function $\underline{y}$ = smooth($H_j, \underline{x}$)

   initialise:               $\underline{y}$ = zeros
   for $n = 1 : \#$ vertices
        get edge labels:     $\underline{L}$ = labels
        solve and update:    $\underline{y}(\underline{L}) = \underline{y}(\underline{L}) + H_j(\underline{L},\underline{L})^{-1}\underline{x}(\underline{L})$
   end
   scale:                    $\underline{y} = \eta\, \underline{y}$
```

Figure 4.6: Additive Schwarz smoothing

components of vector fields. To fix ideas, consider the meshes T_{j-1} and T_j shown in Fig. 4.7.

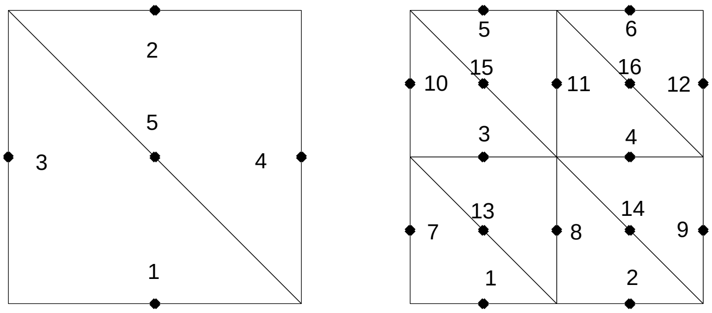

Figure 4.7: Coarse grid T_{j-1} (left), fine grid T_j (right)

Let $\underline{v}_{j-1}$ denote the discrete representation of a function $\vec{v}_{j-1}$ on T_{j-1} that we wish to transfer to T_j. Recall that $\underline{v}_{j-1}(i) = \vec{v}_{j-1} \cdot \vec{\nu}^i$, $i = 1 : 5$, where $\vec{\nu}^i$ is an oriented unit normal vector at edge i on T_{j-1}. Now, normal components of lowest-order Raviart-Thomas functions are constant along edges of triangles. Providing that the orientation of normal vectors is consistent, for nodes lying on edges of T_j that coincide with edges in T_{j-1} we can inject the vector values of $\underline{v}_{j-1}$ directly. Thus, for example, $\underline{v}_j(14) = \underline{v}_j(15) = \underline{v}_{j-1}(5)$.

Formally, we define a function $\vec{v}_j$ via the injection,

$$\vec{v}_j \cdot \vec{\nu}_j^i\,(x_i, y_i) \;=\; \mathcal{I}_{j-1}^{j}\left(\vec{v}_{j-1} \cdot \vec{\nu}_j^i\right)(x_i, y_i) = \left(\vec{v}_{j-1} \cdot \vec{\nu}_j^i\right)(x_i, y_i)\,,$$

where, here, $\vec{\nu}_j^i$ denotes the oriented unit normal vector to edge i in the *fine* level grid T_j and (x_i, y_i) denotes the mid-point of that edge. Thus, for a node i in T_j lying in the interior of a triangle K in T_{j-1}, we obtain, in discrete vector notation,

$$\underline{v}_j(i) = \vec{v}_j \cdot \vec{\nu}_j^{\,i}\,(x_i, y_i) \;=\; \left(\vec{v}_{j-1}|_K \cdot \vec{\nu}_j^{\,i}\right)(x_i, y_i)\,.$$

Expanding $\vec{v}_{j-1}$ in terms of local basis functions for K gives,

$$\vec{v}_{j-1}|_K = \sum_{n=1}^{3} \underline{v}_{j-1}(n)\,\vec{\varphi}_n^{\,K}\,.$$

Hence, if we denote the local edges of K at level $j-1$ by n_1, n_2, n_3, we obtain, in vector notation,

$$\underline{v}_j(i) = \underline{v}_{j-1}(n_1)c_1 + \underline{v}_{j-1}(n_2)c_2 + \underline{v}_{j-1}(n_3)c_3,$$

where,

$$c_k = \vec{\varphi}_{n_k}^{\,K}\,(x_i, y_i) \cdot \vec{\nu}_j^{\,i}, \quad k = 1:3.$$

Thus, in Fig. 4.7, $\underline{v}_j(3), \underline{v}_j(8)$ and $\underline{v}_j(13)$ are linear combinations of $\underline{v}_{j-1}(1), \underline{v}_{j-1}(5)$ and $\underline{v}_{j-1}(3)$. For the configuration of normal vectors shown in Fig. 2.3, we obtain the interpolation matrix,

$$\mathcal{I}_{j-1}^{j} \;=\; \begin{array}{c} \left(\begin{array}{cccccccccccccccc} 1 & 1 & \frac{1}{2} & 0 & 0 & 0 & 0 & -\frac{1}{2} & 0 & 0 & 0 & 0 & \frac{1}{2\sqrt{2}} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 1 & 1 & 0 & 0 & 0 & 0 & -\frac{1}{2} & 0 & 0 & 0 & 0 & \frac{1}{2\sqrt{2}} \\ 0 & 0 & -\frac{1}{2} & 0 & 0 & 0 & 1 & \frac{1}{2} & 0 & 0 & 0 & 0 & \frac{1}{2\sqrt{2}} & 0 & 0 & 0 \\ 0 & 0 & 0 & -\frac{1}{2} & 0 & 0 & 0 & 0 & 1 & 1 & \frac{1}{2} & 1 & 0 & 0 & 0 & \frac{1}{2\sqrt{2}} \\ 0 & 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & 0 & 0 & \frac{1}{\sqrt{2}} & 0 & 0 & \frac{1}{\sqrt{2}} & 0 & \frac{1}{2} & 1 & 1 & \frac{1}{2} \end{array}\right) \end{array} \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array}$$

We report on numerical experiments at the end of the chapter. First, we present our main result. This is a new eigenvalue analysis which establishes the impact on the eigenvalue bound of Theorem 6 of replacing H, in the ideal preconditioner (4.1), with an approximation.

4.4 Eigenvalue bounds

Let V be *any* symmetric and positive definite approximation to the matrix $H = A + D$, arising in (4.2), such that there exist positive constants θ and Θ, satisfying,

$$\theta \le \frac{\underline{u}^T H \underline{u}}{\underline{u}^T V \underline{u}} \le \Theta \le 1 \quad \forall \underline{u} \in I\!R^n \backslash \{\underline{0}\}. \tag{4.9}$$

We now consider preconditioning the lowest-order Raviart-Thomas saddle-point system (2.37) with the matrix,

$$P = \begin{pmatrix} V & 0 \\ 0 & N \end{pmatrix}. \tag{4.10}$$

To establish bounds for the eigenvalues of the preconditioned system, we require the following preliminary result.

Lemma 17 *The n eigenvalues $\{\sigma_i\}_{i=1}^{n}$ of $D\underline{u} = \sigma H\underline{u}$, lie in the interval $[0, 1)$.*

Proof Recall from (3.14) that $H = A + D = A + B^T N^{-1} B$. Since H is positive definite and D is semi-positive definite, it is easy to see that $\sigma \geq 0$. If $\underline{u} \in null(B)$, then $\sigma = 0$ by positive definiteness of A. The dimension of $null(B)$ is $n - m$, so there are $n - m$ zero eigenvalues. Now, if $\underline{u} \notin null(B)$, $\sigma > 0$ and we have,

$$(1 - \sigma)\, \underline{u}^T D\underline{u} \;=\; \sigma \underline{u}^T A\underline{u} > 0.$$

Since $\underline{u}^T D\underline{u} > 0$ in this case, we must have $1 - \sigma > 0$ and thus $\sigma < 1$. □

Our starting point is the eigenvalue bound (3.24) in Theorem 6, for the ideal preconditioned system (3.23). To simplify notation, let,

$$a = \left(\frac{c\mu_{min}}{\mid K \mid_{min} \, + \mu_{min}} \right), \tag{4.11}$$

so that the bound (3.24) is,

$$[-1, -a] \cup [1]. \tag{4.12}$$

We now obtain the following result.

Theorem 10 *The $n + m$ eigenvalues $\{\lambda_i\}_{i=1}^{n+m}$ of the generalised eigenvalue problem,*

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} \underline{u} \\ \underline{p} \end{pmatrix} = \lambda \begin{pmatrix} V & 0 \\ 0 & N \end{pmatrix} \begin{pmatrix} \underline{u} \\ \underline{p} \end{pmatrix}, \tag{4.13}$$

arising in the Raviart-Thomas approximation of (2.12), lie in the union of the intervals,

$$\left[-1, \frac{1}{2} \left(\theta \left(1 - a \right) - \sqrt{\theta^2 \left(a - 1 \right)^2 + 4a\theta} \right) \right] \cup [\theta, 1], \tag{4.14}$$

where θ is the positive constant satisfying (4.9) and a is the positive constant defined in (4.11).

Proof First, suppose that $\lambda > 0$. The eigenvalues $\{\lambda_i\}_{i=1}^{m+n}$ satisfy,

$$\begin{aligned} A\underline{u} + B^T\underline{p} &= \lambda V\underline{u}, \\ B\underline{u} &= \lambda N\underline{p}. \end{aligned}$$

Eliminating $\underline{p}$ yields,

$$\begin{aligned} \lambda A\underline{u} + B^T N^{-1} B\underline{u} &= \lambda^2 V\underline{u}, \\ \lambda A\underline{u} + D\underline{u} &= \lambda^2 V\underline{u}, \\ \lambda\left(A + D\right)\underline{u} + \left(1-\lambda\right)D\underline{u} &= \lambda^2 V\underline{u}, \\ \lambda H\underline{u} + \left(1-\lambda\right)D\underline{u} &= \lambda^2 V\underline{u}. \end{aligned}$$

Thus,

$$\left(1-\lambda\right)\underline{u}^T D\underline{u} = \lambda^2\underline{u}^T V\underline{u} - \lambda\underline{u}^T H\underline{u}, \tag{4.15}$$

and since, by assumption (4.9),

$$\underline{u}^T H\underline{u} \le \underline{u}^T V\underline{u} \le \frac{1}{\theta}\underline{u}^T H\underline{u}, \tag{4.16}$$

it follows that,

$$\left(1-\lambda\right)\underline{u}^T D\underline{u} \;\ge\; \left(\lambda^2 - \lambda\right)\underline{u}^T H\underline{u}.$$

Applying Lemma 17, and noting that H is positive definite, we obtain,

$$\begin{aligned} \left(1-\lambda\right)\underline{u}^T H\underline{u} &\ge \left(\lambda^2-\lambda\right)\underline{u}^T H\underline{u}, \\ \left(1-\lambda\right) &\ge \left(\lambda^2-\lambda\right), \\ 1 &\ge \lambda. \end{aligned}$$

From (4.15) and (4.16), it also follows that,

$$\left(1-\lambda\right)\underline{u}^T D\underline{u} \;\le\; \left(\frac{\lambda^2}{\theta} - \lambda\right)\underline{u}^T H\underline{u}.$$

Since $0 \le \underline{u}^T D\underline{u}$ and we have established $0 < \lambda \le 1$, it follows that,

$$0 \;\le\; \lambda\left(\frac{\lambda}{\theta} - 1\right)\underline{u}^T H\underline{u}.$$

Hence $\lambda \ge \theta$, and the bound for the positive eigenvalues is proved.

Now assume that $\lambda < 0$. Eliminating $\underline{u}$ yields,

$$B\left(A - \lambda V\right)^{-1} B^T \underline{p} \;=\; -\lambda N \underline{p}, \tag{4.17}$$

an implicit equation for λ. Note that since $\lambda < 0$, the matrix $(A - \lambda V)$ is positive definite. Now, the values of λ satisfying (4.17) are the eigenvalues of the matrix,

$$\begin{aligned}
& N^{-\frac{1}{2}} B\left(\lambda V - A\right)^{-1} B^T N^{-\frac{1}{2}} \\
= \;& N^{-\frac{1}{2}} B\left(\lambda V - H + D\right)^{-1} B^T N^{-\frac{1}{2}} \\
= \;& N^{-\frac{1}{2}} B\left(\lambda V - H + B^T N^{-1} B\right)^{-1} B^T N^{-\frac{1}{2}} \\
= \;& N^{-\frac{1}{2}} B Y^{-\frac{1}{2}} \left(I + Y^{-\frac{1}{2}} B^T N^{-1} B Y^{-\frac{1}{2}}\right)^{-1} Y^{-\frac{1}{2}} B^T N^{-\frac{1}{2}} \\
= \;& X\left(I + X^T X\right)^{-1} X^T,
\end{aligned}$$

where, here, $X = N^{-\frac{1}{2}} B Y^{-\frac{1}{2}}$ and $Y = \lambda V - H$. Applying the same arguments as in Theorem 6, and applying the Sherman-Morrison-Woodbury formula, the eigenvalues $\{\lambda_i\}$ we are seeking in (4.17) are the values:

$$\lambda_i \;=\; \frac{\sigma_i}{1 + \sigma_i}, \tag{4.18}$$

where each σ_i is an eigenvalue of,

$$B\left(\lambda_i V - H\right)^{-1} B^T \underline{p} \;=\; \sigma N \underline{p}. \tag{4.19}$$

We can obtain a bound for these values by exploiting the spectral equivalence of H and V defined in (4.9). Note that bounds for the eigenvalues of (4.17) cannot be obtained without the above manipulation, since we have no readily available information about the spectral equivalence of A and V.

Consider, first, the eigenvalues $\{\mu\}$ of,

$$\left(\lambda V - H\right)^{-1} \underline{u} \;=\; \mu H^{-1} \underline{u}. \tag{4.20}$$

Since $\lambda < 0$, the matrix $(\lambda V - H)$ is negative definite. Since H^{-1} is positive definite, the values of μ are negative. Rearranging gives,

$$\begin{aligned}
H\left(\lambda V - H\right)^{-1} \underline{u} \;&=\; \mu \underline{u} \\
\left(\lambda H^{-1} V - I\right)^{-1} \underline{u} \;&=\; \mu \underline{u}
\end{aligned}$$

$$
\begin{aligned}
\left(\lambda H^{-1}V - I\right)\underline{u} &= \frac{1}{\mu}\underline{x} \\
\lambda V\underline{u} &= \left(\frac{1}{\mu}+1\right)H\underline{u} \\
\frac{\lambda}{\left(\frac{1}{\mu}+1\right)} &= \frac{\underline{u}^T H\underline{u}}{\underline{u}^T V\underline{u}}.
\end{aligned}
$$

Combining this with (4.9) we obtain,

$$\theta \leq \frac{\lambda\mu}{(\mu+1)} \leq 1.$$

Recalling that $\theta > 0$, $\mu < 0$ and $\lambda < 0$, we find that,

$$\mu \in \left[\frac{1}{\lambda-1}, \frac{\theta}{\lambda-\theta}\right]. \tag{4.21}$$

Now, combining (4.20) and (4.21), we obtain $\forall \underline{p} \in I\!\!R^m \backslash \{\underline{0}\}$,

$$\frac{1}{\lambda-1} \leq \frac{\underline{p}^T B\left(\lambda V - H\right)^{-1} B^T \underline{p}}{\underline{p}^T B H^{-1} B^T \underline{p}} \leq \frac{\theta}{\lambda-\theta},$$

and so,

$$\left(\frac{1}{\lambda-1}\right)\frac{\underline{p}^T B H^{-1} B^T \underline{p}}{\underline{p}^T N \underline{p}} \leq \frac{\underline{p}^T B\left(\lambda V - H\right)^{-1} B^T \underline{p}}{\underline{p}^T N \underline{p}} \leq \left(\frac{\theta}{\lambda-\theta}\right)\frac{\underline{p}^T B H^{-1} B^T \underline{p}}{\underline{p}^T N \underline{p}}.$$

In Theorem 6, we obtained the bound,

$$a \leq \frac{\underline{p}^T B H^{-1} B^T \underline{p}}{\underline{p}^T N \underline{p}} \leq 1 \quad \forall \underline{p} \in I\!\!R^m \backslash \{\underline{0}\},$$

with $a > 0$. Since the bounds for the eigenvalues in (4.21) are negative, we obtain,

$$\left(\frac{1}{\lambda-1}\right) \leq \frac{\underline{p}^T B\left(\lambda V - H\right)^{-1} B^T \underline{p}}{\underline{p}^T N \underline{p}} \leq \left(\frac{a\theta}{\lambda-\theta}\right).$$

For any $\lambda < 0$, the eigenvalues $\{\sigma\}$ of (4.19) therefore satisfy,

$$\frac{1}{\lambda-1} \leq \sigma \leq \frac{a\theta}{\lambda-\theta} < 0, \tag{4.22}$$

and also,

$$\frac{1}{1+\frac{1}{\lambda-1}} \geq \frac{1}{1+\sigma} \geq \frac{1}{1+\frac{a\theta}{\lambda-\theta}} > 0.$$

Hence, we obtain,

$$\frac{\frac{1}{\lambda-1}}{1+\frac{1}{\lambda-1}} \leq \frac{\sigma}{1+\sigma} \leq \frac{\frac{a\theta}{\lambda-\theta}}{1+\frac{a\theta}{\lambda-\theta}},$$

and so, using (4.18), the eigenvalues λ of (4.17) satisfy,

$$\frac{1}{\lambda} \leq \lambda \leq \frac{a\theta}{\lambda + \theta\left(a-1\right)}.$$

Finally, solving for λ in,

$$1 \geq \lambda^2, \qquad \lambda^2 + \lambda\theta\left(a-1\right) - a\theta \geq 0,$$

yields, since $\lambda < 0$,

$$-1 \leq \lambda \leq \frac{1}{2}\left(\theta\left(1-a\right) - \sqrt{\theta^2\left(a-1\right)^2 + 4a\theta}\right). \quad \square$$

Remark 11 *Notice that when $\theta = 1$, we recover the eigenvalue bound (4.12) for the ideal preconditioner.*

We choose V to be the discrete form of the Arnold-Falk-Winther multigrid operator described in the last section. By Theorem 9, this V yields $\Theta = 1$ in (4.9).

Using the the result of Theorem 10, we can now deliver the key message of this chapter. The $\mathcal{A}$-optimality of the preconditioner (4.10) is completely determined by that of the ideal preconditioner (4.1). To see this, recall from Theorem 6 in Chapter 3 that,

$$a = \left(\frac{c\mu_{min}}{\mid K \mid_{min} + \mu_{min}}\right), \tag{4.23}$$

where μ_{min} is the minimum eigenvalue of $BA^{-1}B^T$ and, for quasi-uniform meshes, $c > 0$ is a constant independent of the discretisation parameter. If a decays to zero due to the small magnitude of the entries of the coefficient tensor $\mathcal{A}$, then, asymptotically,

$$\lim_{a\to 0}\frac{1}{2}\left(\theta\left(1-a\right) - \sqrt{\theta^2\left(a-1\right)^2 + 4a\theta}\right) = \frac{1}{2}\left(\theta - \sqrt{\theta^2}\right) = 0.$$

Thus, even if the chosen multigrid approximation to H is both h-optimal and $\mathcal{A}$-optimal, the eigenvalue bound (4.14) deteriorates. The deficiency in the ideal preconditioner carries over to the practical preconditioner and scaling will still be needed to obtain efficient MINRES iteration.

On the other hand, if the coefficient term is sufficiently large so that $a \to 1$, we obtain, asymptotically,

$$\lim_{a\to 1}\frac{1}{2}\left(\theta\left(1-a\right) - \sqrt{\theta^2\left(a-1\right)^2 + 4a\theta}\right) = -\frac{1}{2}\sqrt{4\theta} = -\sqrt{\theta}.$$

Hence, for large μ_{min}, the eigenvalue bound (4.14) takes the form,

$$\left(-1,\, -\sqrt{\theta}\right] \cup [\,\theta\,,\, 1]\,,$$

and the efficiency of the method is completely determined by the multigrid approximation.

4.5 Preconditioned MINRES

To illustrate that the above theory is tight, we now report on MINRES convergence for a range of coefficients, using the preconditioner (4.10). All of the experiments are performed with uniform meshes of right-angled triangles and the stopping tolerance (3.39).

Example 1

Consider the case $\Omega = [0,1] \times [0,1]$, $p = 0$ on $\partial\Omega$, and $f = 2(x^2 - x + y^2 - y)$, so that for $\mathcal{A} = \mathcal{I}$ we obtain the analytical pressure solution $p = x(x-1)y(y-1)$. The eigenvalues of $V^{-1}H$, using the method of Arnold et al., are listed in Table 4.1. They indicate that the multigrid approximation is h-optimal. In Table 4.2 we compare the observed eigenvalues of the preconditioned saddle-point system with the bounds in (4.14). Iteration counts are given in Table 4.3.

h	θ	Θ
$\frac{1}{4}$	0.5938	1
$\frac{1}{8}$	0.4595	1
$\frac{1}{16}$	0.4273	1

Table 4.1: Eigenvalues of $V^{-1}H$, unit coefficients

h	*bounds*	*observed*
$\frac{1}{4}$	$[-0.9983,\ -0.7381] \cup [0.5938, 1]$	$[-0.9879,\ -0.8507] \cup [0.5943, 1]$
$\frac{1}{8}$	$[-0.9996,\ -0.6504] \cup [0.4595, 1]$	$[-0.9972,\ -0.8438] \cup [0.4598, 1]$
$\frac{1}{16}$	$[-0.9999,\ -0.6503] \cup [0.4273, 1]$	$[-0.9994,\ -0.8481] \cup [0.4273, 1]$

Table 4.2: Theoretical bounds and observed eigenvalues, unit coefficients

h	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$	$\frac{1}{64}$
	17	18	18	18	18
	(81)	(183)	(367)	(*)	(*)

Table 4.3: MINRES iterations, Example 1

Example 2

Next, we introduce a jump in the coefficient and set $\mathcal{A} = \epsilon\mathcal{I}$ in one quadrant of Ω so that $\mu_{min} \to 0$ in (4.23), if $\epsilon << 1$. Values of θ are listed in Table 4.4. Although this case is not covered by the theory of Arnold et al., the approximation to $\mathcal{H}$ is $\mathcal{A}$-optimal and h-optimal; $\Theta = 1$ in all cases. The negative eigenvalues of the preconditioned saddle-point system, for $\epsilon = 10^{-3}$ and $\epsilon = 10^{-6}$ are listed in Tables 4.5–4.6.

ϵ h	$\frac{1}{8}$	$\frac{1}{16}$
10^{6}	0.4725	0.4326
10^{5}	0.4725	0.4326
10^{4}	0.4725	0.4326
10^{3}	0.4725	0.4326
10^{2}	0.4722	0.4325
10^{1}	0.4698	0.4316
10^{0}	0.4595	0.4273
10^{-1}	0.4099	0.3784
10^{-2}	0.3642	0.3423
10^{-3}	0.3513	0.3319
10^{-4}	0.3494	0.3302
10^{-5}	0.3491	0.3299
10^{-6}	0.3491	0.3299

Table 4.4: Values of θ, discontinuous $\mathcal{A}$, single jump

Observe that the righthand bound for the negative eigenvalues is tighter as the jump parameter decreases. This is illustrated in Fig. 4.8 where we compare the observed and predicted righthand negative eigenvalues for a fixed h and varying ϵ. The observed values are marked with crosses, the theoretical bounds are marked with circles. The scale on the y-axis corresponds to values of $\epsilon \in \left[10^{-6}, 10^{6}\right]$. For $\epsilon > 1$, we observe that the theoretical bound is a conservative estimate.

h	*bounds*	*observed*
$\frac{1}{4}$	$[-0.9982,\ -0.0695]$	$[-0.9864,\ -0.0766]$
$\frac{1}{8}$	$[-0.9996,\ -0.0665]$	$[-0.9969,\ -0.0741]$
$\frac{1}{16}$	$[-0.9999,\ -0.0653]$	$[-0.9993,\ -0.0733]$

Table 4.5: Theoretical bounds and observed negative eigenvalues, $\epsilon = 10^{-3}$

h	*bounds*	*observed*
$\frac{1}{4}$	$[-0.9982,\ -0.000083273]$	$[-0.9864,\ -0.000083284]$
$\frac{1}{8}$	$[-0.9996,\ -0.000080216]$	$[-0.9972,\ -0.000080227]$
$\frac{1}{16}$	$[-0.9999,\ -0.000079271]$	$[-0.9994,\ -0.000079284]$

Table 4.6: Theoretical bounds and observed negative eigenvalues, $\epsilon = 10^{-6}$

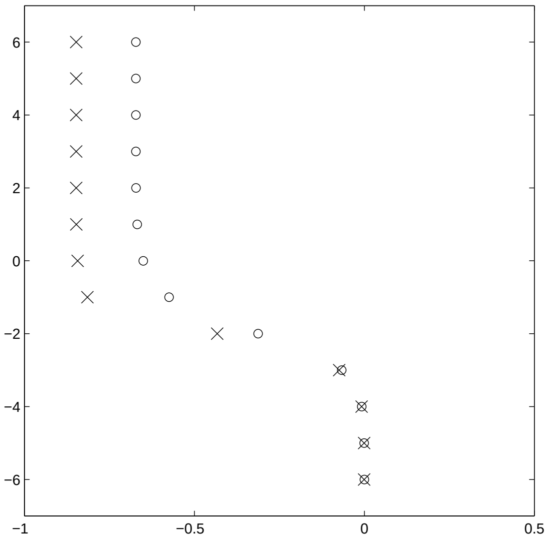

Figure 4.8: Observed negative eigenvalues (x) and theoretical bound (o), $\epsilon \in [10^{-6}, 10^{6}]$

The eigenvalues of the preconditioned system, for a fixed mesh and varying ϵ are plotted in Fig. 4.9. Again, the scale on the y-axis corresponds to values of $\epsilon \in \left[10^{-6}, 10^{6}\right]$. Despite the optimal multigrid performance, for values $\epsilon < 1$, it is clear that MINRES convergence will deteriorate. Iteration counts obtained with the ideal preconditioner and the multigrid preconditioner are listed in Table 4.7 and Table 4.8, respectively.

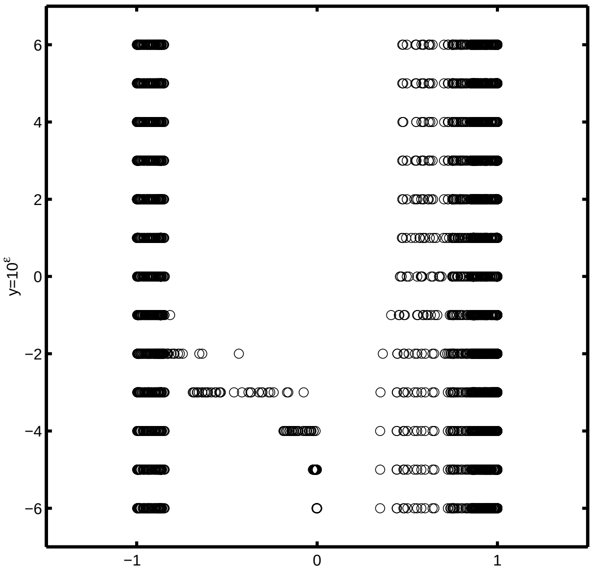

Figure 4.9: Eigenvalues of multigrid preconditioned system, $h = \frac{1}{16}, \epsilon \in \left[10^{-6}, 10^{6}\right]$

As our theory predicts, the multigrid preconditioner exhibits the same asymptotic behaviour as the exact version as $\epsilon \to 0$. The deterioration can be corrected, however, by rescaling the coefficients, as discussed in section 3.2.1.

ϵ \ h	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$
10^{6}	5	5	5
10^{5}	5	5	5
10^{4}	5	5	5
10^{3}	5	5	5
10^{2}	5	5	5
10^{1}	5	5	5
10^{0}	5	5	5
10^{-1}	6	6	6
10^{-2}	9	9	9
10^{-3}	17	19	19
10^{-4}	28	42	48
10^{-5}	35	70	107
10^{-6}	42	92	162

Table 4.7: MINRES iterations, exact preconditioner, Example 2

ϵ h	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$
10^{6}	17	18	18
10^{5}	17	18	18
10^{4}	17	18	18
10^{3}	17	18	18
10^{2}	17	18	18
10^{1}	18	18	18
10^{0}	17	18	18
10^{-1}	18	20	20
10^{-2}	22	24	25
10^{-3}	41	45	47
10^{-4}	71	119	130
10^{-5}	106	220	342
10^{-6}	131	328	574

Table 4.8: MINRES iterations, multigrid preconditioner, Example 2

Example 3

For a more challenging discontinuous coefficient example, we consider, again, the so-called 'Kellogg problem' described in Example 6 in Chapter 3. Recall that the coefficient is prescribed in a 2×2 checkerboard fashion on $\Omega = [-1, 1] \times [-1, 1]$. We choose $\mathcal{A} = a_1\mathcal{I} \approx 161.477 \times \mathcal{I}$ in two quadrants of Ω and $\mathcal{A} = \mathcal{I}$ elsewhere.

Unfortunately, the multigrid method of Arnold et al. is limited to quasi-uniform meshes, which are not desirable for this problem. We perform the experiment, however, to observe the behaviour of the multigrid approximation with respect to the complicated coefficient term. We will describe an alternative practical preconditioning scheme that is suited to locally refined meshes in Chapter 5.

Iteration counts are listed in Table 4.9. For $h = \frac{1}{8}$, we obtain $\theta \approx 0.1538$ and $\Theta = 1$, yielding the eigenvalue bound, $[-0.9999,\ -0.7568] \cup [0.1538,\ 1]$. The observed values lie in the interval $[-0.9998,\ -0.8431] \cup [0.1541,\ 1]$.

h	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$	$\frac{1}{128}$
	23	25	26	26	26
	(*)	(*)	(*)	(*)	(*)

Table 4.9: MINRES iterations, exact preconditioner, Example 3

Again, this type of coefficient tensor is not covered by the theory. However, we observe that the scheme is robust. To illustrate this, values of θ are listed in Table 4.10 for different values of a_1. $\Theta = 1$ in all cases. Although θ is smaller than in Example 2 for large coefficients, it remains bounded away from zero.

a_1 h	$\frac{1}{8}$	$\frac{1}{16}$
10^6	0.1470	0.1043
10^5	0.1470	0.1043
10^4	0.1471	0.1044
10^3	0.1481	0.1055
10^2	0.1578	0.1156
10^1	0.2358	0.1994
10^0	0.4340	0.4179
10^{-1}	0.4449	0.4211
10^{-2}	0.4457	0.4209
10^{-3}	0.4453	0.4208
10^{-4}	0.4451	0.4208
10^{-5}	0.4451	0.4208
10^{-6}	0.4451	0.4208

Table 4.10: Values of θ, discontinuous $\mathcal{A}$, 2×2 checkerboard jumps

Example 4

If we set the coefficient to be the full tensor (3.40) we obtain $\kappa(V^{-1}H) \approx 3.84$. Iteration counts are given in Table 4.11.

h	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$	$\frac{1}{128}$
	21 (322)	22 (*)	22 (*)	22 (*)	22 (*)

Table 4.11: MINRES iterations, Example 4

Example 5

Choosing the variable coefficient tensor (3.41) yields $\kappa(V^{-1}H) \approx 1.95$. Iteration counts are listed in Table 4.12. Once again, the multigrid approximation is h-optimal and so therefore is the preconditioner (4.10). However, the iteration count rises since the minimum eigenvalue of $BA^{-1}B^T$ is smaller here than in the other examples. (Recall

Example 3 in Chapter 3.)

h	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$	$\frac{1}{128}$
	45	59	67	69	69
	(*)	(*)	(*)	(*)	(*)

Table 4.12: MINRES iterations, Example 5

Example 6

Finally, consider an anisotropic test problem with $\mathcal{A} = \texttt{diag}(\epsilon, 1)$. Condition numbers for $V^{-1}H$ are listed in Table (4.13); they show that the suggested multigrid scheme is totally unsuitable as an approximation to the $H(div)$ operator with anisotropic weighting. $\Theta = 1$ in all cases but θ deteriorates with ϵ. MINRES does not converge, even for mild anisotropies.

ϵ	$h = \frac{1}{8}$	$h = \frac{1}{16}$
10^6	8.98e5	9.60e5
10^5	8.98e4	9.60e4
10^4	8.94e3	9.61e3
10^3	8.99e2	9.61e2
10^2	91.12	97.37
10^1	10.27	10.90
10^0	2.18	2.34
10^{-1}	9.71	10.75
10^{-2}	59.48	85.43
10^{-3}	1.98e2	4.30e2
10^{-4}	8.06e2	1.30e3
10^{-5}	6.42e3	7.25e3
10^{-6}	6.24e4	6.56e4

Table 4.13: Condition number of $V^{-1}H$, anisotropic coefficients

4.6 Concluding remarks

In this chapter, we developed the preconditioning scheme introduced in Chapter 3. To obtain a practical scheme, we replaced the weighted $H(div)$ operator with a multigrid approximation due to Arnold et al. We rigorously established the performance of the resulting preconditioner with new eigenvalue analysis. Further, we demonstrated the impact of general coefficient tensors on the performance of the multigrid approximation and on the theoretical eigenvalue bound.

Two key issues arise. First, the multigrid approximation failed in anisotropic test cases. This failure begs the question of whether general coefficient tensors can be handled efficiently in the V-cycle framework of Arnold, Falk and Winther. Our suspicion is that coefficient dependent transfer operators should be employed, rather than pure injection, between grids. However, we must defer this to future work. More importantly, it is clear that if the ideal preconditioner (4.1) is not $\mathcal{A}$-optimal, which is likely in practical simulations, then neither is the suggested practical preconditioner (4.10). Iteration may only be efficient if scaling with respect to the minimum coefficient, is applied. This is true independently of the approximation properties of the multigrid scheme.

Chapter 5

H^1 preconditioning

This chapter is motivated by the observation that the continuous variational problem (2.10) is well-posed in a *second* pair of function spaces. This leads to ideal preconditioners of the generic form (2.52). First, we outline an alternative stability theory which gives rise to the concept of a 'jump operator'. We discuss an ideal preconditioning scheme that incorporates this operator and consider some of the difficulties in constructing finite element matrices to represent it. Finally, we propose and analyse a novel practical scheme based on black-box algebraic multigrid (AMG).

5.1 Motivation

Our starting point is the mixed first-order PDE system (2.8) with homogeneous Dirichlet boundary condition, $p = 0$ on $\partial\Omega$. (The boundary condition will not be imposed in the sequel, it is simply to facilitate the initial discussion.) Now choose $V = L^2(\Omega)^d$ and $W = H_0^1(\Omega)$. Multiplying by arbitrary test functions, integrating and applying Green's formula to the *second* equation leads to the continuous, mixed variational problem,

find $(\vec{u}, p) \in V \times W$ satisfying,

$$\begin{aligned} a(\vec{u},\vec{v}) + b(\vec{v},p) &= 0 \quad \forall\, \vec{v} \in V, \\ b(\vec{u},w) &= -(f,w) \quad \forall\, w \in W, \end{aligned} \tag{5.1}$$

where $a(\cdot,\cdot) : V \times V \to I\!R$ and $b(\cdot,\cdot) : V \times W \to I\!R$ are now defined via,

$$a(\vec{u},\vec{v}) = \left(\mathcal{A}^{-1}\vec{u},\vec{v}\right), \quad b(\vec{u},w) = -\left(\nabla w,\vec{u}\right).$$

Z_h-ellipticity certainly holds in the norm $\| \cdot \|_0$ if (2.1) holds, and since $\nabla W \subset V$,

$$\sup_{\vec{v} \in V} \frac{b(\vec{v}, w)}{\| \vec{v} \|_0} = \sup_{\vec{v} \in V} - \frac{(\nabla w, \vec{v})}{\| \vec{v} \|_0} \geq \frac{\| \nabla w \|_0^2}{\| \nabla w \|_0} = | w |_1 \geq \frac{1}{\sqrt{1+c^2}} \| w \|_1, \tag{5.2}$$

where c is the constant arising in Friedrich's inequality (see Lemma 1). Hence, inf-sup stability holds in $\| \cdot \|_0$ and $\| \cdot \|_1$ with $\beta = \frac{1}{\sqrt{1+c^2}}$.

Following the discussion in Chapter 3, it seems feasible to consider block-diagonal preconditioners P of the generic form,

$$P = \begin{pmatrix} P_1 & 0 \\ 0 & P_2 \end{pmatrix} \tag{5.3}$$

with symmetric and positive definite blocks $P_1 \in I\!\!R^{n \times n}$ and $P_2 \in I\!\!R^{m \times m}$ chosen to represent the norms $\| \cdot \|_0$ and $\| \cdot \|_1$ on the subspaces V_h and W_h. However, formulating the discrete problem,

find $(\vec{u_h}, p_h) \in V_h \times W_h$ satisfying,

$$\begin{aligned} a(\vec{u}_h, \vec{v}_h) + b(\vec{v}_h, p_h) &= 0 \quad \forall \vec{v}_h \in V_h, \\ b(\vec{u}_h, w_h) &= -(f, w_h) \quad \forall w_h \in W_h, \end{aligned} \tag{5.4}$$

using the lowest-order Raviart-Thomas spaces now corresponds to a *non-conforming* approach because $W_h \not\subset H^1(\Omega)$. Recall that W_h is the space of piecewise constant functions and so the definition of ∇w does not hold in the classical sense. Hence, there is no *obvious* way to construct a suitable matrix P_2. Moreover, discrete inf-sup stability cannot be established in the *standard* Sobolev norm $\| \cdot \|_1$.

5.1.1 Alternative inf-sup inequality

As pointed out in [88], we *can* establish inf-sup stability in an alternative mesh-dependent norm, if T_h is quasi-uniform. The advantage is that the new norm can be defined locally on each element of T_h and has a simple algebraic representation. In essence, we require a 'jump operator'. The reader should note, however, that the details are specific to the choice (2.39) of the degrees of freedom for V_h. We outline an approach for the lowest-order schemes in $I\!\!R^2$. Similar concepts carry over to $I\!\!R^3$.

Assume, now, that T_h is a quasi-uniform partition of Ω and denote the set of *all* edges of T_h by $\mathcal{E}_h$. Then, for any given $w_h \in W_h$, we can define a norm via,

$$\| w_h \|_{1,h}^2 \quad = h^{-1} \textstyle\sum_{e \in \mathcal{E}_h} \int_e [w_h]_e^2 \, ds. \tag{5.5}$$

Here, h denotes the maximum characteristic edge length and e denotes a generic edge. $[w_h]_e$ denotes the jump in w_h across an edge between two elements K_1, K_2. Thus,

$$[w_h]_e = w_h \,|_{K_1} - w_h \,|_{K_2} \,.$$

By extending w_h by zero outside Ω, jumps across boundary edges are well defined. The norm (5.5) is a special case of a general norm, for higher order schemes, considered by Rusten et al. in [88]. We will use it to establish new eigenvalue bounds and to construct a matrix operator that is spectrally equivalent to the Schur complement matrix $BA^{-1}B^T$ for the choice (2.39). The authors of [88], on the other hand, derive an operator that is spectrally equivalent to BB^T. To begin, we have the following result.

Lemma 18 *Let V_h and W_h be the lowest-order Raviart-Thomas spaces defined in (2.32) and (2.33), respectively. Let the degrees of freedom for V_h be chosen according to (2.39). If T_h is quasi-uniform, there exists a constant $C > 0$, independent of h, satisfying,*

$$\sup_{\vec{v}_h \in V_h \setminus \{\vec{0}\}} \frac{(w_h, \nabla \cdot \vec{v}_h)}{\| \vec{v}_h \|_0} \leq C \, \| w_h \|_{1,h} \quad \forall w_h \in W_h. \tag{5.6}$$

Proof Let $w_h \in W_h$ be given. For any $\vec{v}_h \in V_h$ we obtain,

$$\begin{aligned}
(w_h, \nabla \cdot \vec{v}_h) &= \sum_K \int_K w_h|_K \, \nabla \cdot \vec{v}_h|_K \, dK \\
&= \sum_K w_h|_K \int_K \nabla \cdot \vec{v}_h|_K dK \\
&= \sum_K w_h|_K \int_{\partial K} \vec{v}_h \cdot \vec{n}_K \, ds \\
&= \sum_{e \in \mathcal{E}_h} [w_h]_e \int_e \vec{v}_h \cdot \vec{n}^e \, ds \\
&= \sum_{e \in \mathcal{E}_h} \int_e h^{-\frac{1}{2}} \, [w_h]_e \, h^{\frac{1}{2}} \, \vec{v}_h \cdot \vec{n}^e \, ds \\
&\leq \left(\sum_{e \in \mathcal{E}_h} h^{-1} \int_e [w_h]_e^2 \, ds \right)^{\frac{1}{2}} \left(\sum_{e \in \mathcal{E}_h} h \int_e (\vec{v}_h \cdot \vec{n}^e)^2 \right)^{\frac{1}{2}}
\end{aligned}$$

$$= \| w_h \|_{1,h} \left(\sum_{e \in \mathcal{E}_h} h\, h_e\, (\vec{v}_h \cdot \vec{n}^e)^2 \right)^{\frac{1}{2}} .$$

Here, h_e denotes the length of edge e. Recalling (2.39) and writing $\underline{v}$ as the vector corresponding to the expansion of $\vec{v}_h$ in the basis for V_h, we obtain, by Lemma 10,

$$\begin{aligned}
\sum_{e \in \mathcal{E}_h} h\, h_e\, (\vec{v}_h \cdot \vec{n}^e)^2 &\leq \sum_{e \in \mathcal{E}_h} h^2\, ((\pm 1)\ \vec{v}_h \cdot \vec{\nu}^e)^2 \\
&= \sum_{e \in \mathcal{E}_h} h^2\, (\vec{v}_h \cdot \vec{\nu}^e)^2 \\
&= h^2 \underline{v}^T \underline{v} \\
&\leq \frac{1}{c_1} \left(\frac{h}{h_{min}} \right)^2 \| \vec{v}_h \|_0^2 \\
&\leq C \| \vec{v}_h \|_0^2 .
\end{aligned}$$

Note we have applied quasi-uniformity in the last step. Combining the two bounds, we obtain,

$$(w_h, \nabla \cdot \vec{v}_h) \leq C \| w_h \|_{1,h} \| \vec{v}_h \|_0 \quad \forall \vec{v}_h \in V_h,$$

and the result immediately follows. □

Now, by making a particular choice of $\vec{v}_h = \vec{v}_h^*$ we can also deduce an alternative inf-sup inequality.

Lemma 19 *Let V_h and W_h be the lowest-order Raviart-Thomas spaces defined in (2.32) and (2.33), respectively. Let the degrees of freedom for V_h be chosen according to (2.39). If T_h is quasi-uniform, there exists a constant* $C > 0$, *independent of* h, *satisfying,*

$$\sup_{\vec{v}_h \in V_h \setminus \{\vec{0}\}} \frac{(w_h, \nabla \cdot \vec{v}_h)}{\| \vec{v}_h \|_0} \geq C \| w_h \|_{1,h} \quad \forall w_h \in W_h. \tag{5.7}$$

Proof Any $\vec{v}_h \in RT_0(K)$ is uniquely defined by the set of values of its normal components at the edges of K. Recall that $\vec{v}_h \cdot \vec{n}$ is a piecewise constant function and so, given any $w_h \in W_h$, we can define for all elements K, a unique $\vec{v}_h^* \in RT_0(K)$ via the set of jumps of w_h across the edges of that element. Hence we can construct a $\vec{v}_h^* \in V_h$ from the element contributions so that,

$$\vec{v}_h^* \cdot \vec{n}^e = [w_h]_e \quad \forall e \in \mathcal{E}_h.$$

As in the proof of Lemma 18, applying Lemma 10 and quasi-uniformity, yields,

$$\sum_{e\in\mathcal{E}_h} h\,h_e\ (\vec{v}_h^* \cdot \vec{n}^e)^2\ ds \ \geq\ h_{min}^2\,(\underline{v}^*)^T\,\underline{v}^* \geq \frac{1}{c_2}\left(\frac{h_{min}}{h}\right)^2 \parallel \vec{v}_h^* \parallel_0^2 \geq C \parallel \vec{v}_h^* \parallel_0^2,$$

where C is a constant independent of h. Now, for any $w_h \in W_h$,

$$\begin{aligned}
(w_h, \nabla\cdot\vec{v}_h^*) &= \sum_{e\in\mathcal{E}_h} [w_h]_e \int_e \vec{v}_h^* \cdot \vec{n}^e\, ds \\
&= \sum_{e\in\mathcal{E}_h} \int_e [w_h]_e^2\, ds \\
&= \left(h^{-1}\sum_{e\in\mathcal{E}_h}\int_e [w_h]_e^2\, ds\right)^{\frac{1}{2}} \left(h\sum_{e\in\mathcal{E}_h}\int_e [w_h]_e^2\, ds\right)^{\frac{1}{2}} \\
&= \parallel w_h \parallel_{1,h} \left(\sum_{e\in\mathcal{E}_h} h\,h_e\,[w_h]_e^2\right)^{\frac{1}{2}} \\
&= \parallel w_h \parallel_{1,h} \left(\sum_{e\in\mathcal{E}_h} h\,h_e\,(\vec{v}_h^* \cdot \vec{n}_e)^2\right)^{\frac{1}{2}} \\
&\geq \sqrt{C} \parallel w_h \parallel_{1,h} \parallel \vec{v}_h^* \parallel_0 .
\end{aligned}$$

Choosing $\vec{v}_h = \vec{v}_h^*$, we deduce that there exists a constant C, independent of h, satisfying,

$$\sup_{\vec{v}_h\in V_h\backslash\{\vec{0}\}} \frac{(w_h, \nabla\cdot\vec{v}_h)}{\parallel \vec{v}_h \parallel_0} \geq \frac{(w_h, \nabla\cdot\vec{v}_h^*)}{\parallel \vec{v}_h^* \parallel_0} \geq C \parallel w_h \parallel_{1,h} \quad \forall\, w_h \in W_h,$$

and the result is proved. □

5.1.2 Matrix form of alternative inf-sup inequality

Now suppose that we can construct a matrix X satisfying,

$$\underline{w}^T X \underline{w} = \parallel w_h \parallel_{1,h}^2 \quad \forall\, w_h \in W_h. \tag{5.8}$$

By writing the alternative inf-sup inequality (5.7) in matrix form, we see that, for quasi-uniform meshes, X is an h-optimal preconditioner for the matrix $BA^{-1}B^T$.

Lemma 20 *If T_h is quasi-uniform, there exist positive constants C_1 and C_2, independent of h and $\mathcal{A}$, satisfying,*

$$\frac{C_2}{\Gamma} \leq \frac{\underline{w}^T BA^{-1}B^T \underline{w}}{\underline{w}^T X \underline{w}} \leq \frac{C_1}{\gamma} \qquad \forall\, \underline{w} \in I\!R^m\backslash\{\underline{0}\}, \tag{5.9}$$

where X is defined in (5.8) and γ and Γ are positive constants satisfying (2.1).

Proof In matrix notation, we obtain,

$$\begin{aligned}
\sup_{\vec{v}_h \in V_h \backslash \{\vec{0}\}} \frac{(w_h, \nabla \cdot \vec{v}_h)}{\| \vec{v}_h \|_0} &= \max_{\underline{v} \in I\!\!R^n} \frac{\underline{w}^T B \underline{v}}{(\underline{v}^T A_{\mathcal{I}} \underline{v})^{\frac{1}{2}}} \\
&\geq \sqrt{\gamma} \max_{\underline{v} \in I\!\!R^n} \frac{\underline{w}^T B \underline{v}}{(\underline{v}^T A \underline{v})^{\frac{1}{2}}} \\
&= \sqrt{\gamma} \max_{\underline{z} = A^{\frac{1}{2}} \underline{v}} \frac{\underline{w}^T B A^{-\frac{1}{2}} \underline{z}}{(\underline{z}^T \underline{z})^{\frac{1}{2}}} \\
&= \sqrt{\gamma} \frac{\underline{w}^T B A^{-1} B^T \underline{w}}{(\underline{w}^T B A^{-1} B^T \underline{w})^{\frac{1}{2}}} \\
&= \sqrt{\gamma} \left(\underline{w}^T B A^{-1} B^T \underline{w}\right)^{\frac{1}{2}} \qquad \forall \underline{w} \in I\!\!R^m \backslash \{\underline{0}\}.
\end{aligned}$$

By Lemma 18 it now follows that there exists a constant C_1, independent of h and $\mathcal{A}$ satisfying,

$$\frac{\underline{w}^T B A^{-1} B^T \underline{w}}{\underline{w}^T X \underline{w}} \leq \frac{C_1}{\gamma} \qquad \forall \underline{w} \in I\!\!R^m \backslash \{\underline{0}\}.$$

Similarly,

$$\begin{aligned}
\sup_{\vec{v}_h \in V_h \backslash \{\vec{0}\}} \frac{(w_h, \nabla \cdot \vec{v}_h)}{\| \vec{v}_h \|_0} &\leq \sqrt{\Gamma} \max_{\underline{v} \in I\!\!R^n} \frac{\underline{w}^T B \underline{v}}{(\underline{v}^T A \underline{v})^{\frac{1}{2}}} \\
&= \sqrt{\Gamma} \left(\underline{w}^T B A^{-1} B^T \underline{w}\right)^{\frac{1}{2}} \qquad \forall \underline{w} \in I\!\!R^m \backslash \{\underline{0}\},
\end{aligned}$$

and so by Lemma 19, there exists a constant C_2, independent of h and $\mathcal{A}$ satisfying,

$$\frac{C_2}{\Gamma} \leq \frac{\underline{w}^T B A^{-1} B^T \underline{w}}{\underline{w}^T X \underline{w}} \qquad \forall \underline{w} \in I\!\!R^m \backslash \{\underline{0}\},$$

which proves the result. □

It will become evident why we require good preconditioners for the Schur complement matrix in the next section.

5.2 Ideal preconditioners

Block-elimination on the matrix associated with (2.34) yields a PDE operator of the form,

$$\begin{pmatrix} \mathcal{A}^{-1}\mathcal{I} & \nabla \\ 0 & \nabla \cdot (\mathcal{A}\nabla) \end{pmatrix}. \tag{5.10}$$

Hence, it is very natural to approximate (2.34) by a block-diagonal matrix (5.3) whose blocks are discrete representations of the operators $\mathcal{A}^{-1}\mathcal{I}$ and $\nabla \cdot (\mathcal{A}\nabla)$ acting on the spaces V_h and W_h, respectively. The Schur complement matrix $S = BA^{-1}B^T$ is a good choice to approximate the scalar diffusion operator $\nabla \cdot (\mathcal{A}\nabla)$ in the $2-2$ block. Indeed, it is a well-known result that the eigenvalues of,

$$\begin{pmatrix} A & 0 \\ 0 & S \end{pmatrix}^{-1} \begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix}, \tag{5.11}$$

lie in 3 clusters at $\frac{1}{2}\left(1-\sqrt{5}\right)$, 1, and $\frac{1}{2}\left(1+\sqrt{5}\right)$. To make the approach (5.11) feasible in practice, S must be replaced by a sparse matrix. In view of Lemma 20, the matrix X is an obvious choice when quasi-uniform meshes are used. Whilst some authors have used such approximations (see, for example, Rusten et al., [88]), we stress that it is *not good enough* for our purposes since it does not provide an $\mathcal{A}$-optimal approximation. This is illustrated in the following bound.

Lemma 21 *If T_h is quasi-uniform, the eigenvalues of the generalised eigenvalue problem,*

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} \underline{u} \\ \underline{p} \end{pmatrix} = \lambda \begin{pmatrix} A & 0 \\ 0 & X \end{pmatrix} \begin{pmatrix} \underline{u} \\ \underline{p} \end{pmatrix}, \tag{5.12}$$

arising in the lowest-order Raviart-Thomas approximation of (2.12), with X defined in (5.8), lie in the intervals $[-\hat{b}, -\hat{a}] \cup [1,1] \cup [1+\hat{a}, 1+\hat{b}]$ where,

$$\hat{a} = -\frac{1}{2} + \frac{1}{2}\sqrt{1+4C_2\Gamma^{-1}}, \quad \hat{b} = -\frac{1}{2} + \frac{1}{2}\sqrt{1+4C_1\gamma^{-1}}, \tag{5.13}$$

C_1, C_2 are constants independent of h and $\mathcal{A}$, and γ and Γ are positive constants satisfying (2.1).

Proof Given *any* approximation X to $S = BA^{-1}B^T$, it is a standard result (see, for example, [90, Theorem 2.3]) that the eigenvalues of (5.12) lie in the union of the intervals $[-\hat{b}, -\hat{a}] \cup [1,1] \cup [1+\hat{a}, 1+\hat{b}]$ where,

$$\hat{a} = -\frac{1}{2} + \frac{1}{2}\sqrt{1+4\sigma_{min}}, \quad \hat{b} = -\frac{1}{2} + \frac{1}{2}\sqrt{1+4\sigma_{max}},$$

and σ_{min} and σ_{max} are the minimum and maximum eigenvalues of,

$$S\underline{p} = \sigma X\underline{p}. \tag{5.14}$$

Applying Lemma 20 for our particular choice of X yields the stated result. □

In view of (5.13), the ideal preconditioner,

$$P = \begin{pmatrix} A & 0 \\ 0 & X \end{pmatrix}, \tag{5.15}$$

is h-optimal for quasi-uniform meshes. However, since the constants γ and Γ appear in the eigenvalue bound, it is clear that the jump operator in (5.8) does not provide the right kind of scaling with respect to the coefficients. Forsaking, temporarily, the notion of 'jump operators', we now motivate a different ideal preconditioner (introduced in [78]), using purely algebraic arguments.

Suppose that we have an approximation P_A to the weighted velocity mass matrix A, satisfying,

$$\tilde{\mu}_1 \leq \frac{\underline{u}^T A \underline{u}}{\underline{u}^T P_A \underline{u}} \leq \tilde{\mu}_n \quad \forall \underline{u} \in I\!R^n \backslash \{\underline{0}\}, \tag{5.16}$$

with positive constants $\tilde{\mu}_1$ and $\tilde{\mu}_n$, and consider the preconditioner,

$$P = \begin{pmatrix} P_A & 0 \\ 0 & BP_A^{-1}B^T \end{pmatrix}. \tag{5.17}$$

For efficiency, we choose P_A to be a diagonal matrix. If we consider symmetric preconditioning, the following result is a simple consequence of Rusten and Winther's standard eigenvalue bound in Lemma 9.

Lemma 22 *Let $0 < \tilde{\mu}_1 \ldots \leq \tilde{\mu}_n$ be the eigenvalues of $P_A^{-1}A$, then the eigenvalues of the generalised eigenvalue problem,*

$$\underbrace{\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix}}_{C} \begin{pmatrix} \underline{u} \\ \underline{p} \end{pmatrix} = \lambda \underbrace{\begin{pmatrix} P_A & 0 \\ 0 & BP_A^{-1}B^T \end{pmatrix}}_{P} \begin{pmatrix} \underline{u} \\ \underline{p} \end{pmatrix}, \tag{5.18}$$

lie in the union of the intervals,

$$\left[\frac{1}{2}\left(\tilde{\mu}_1 - \sqrt{\tilde{\mu}_1^2 + 4}\right), \frac{1}{2}\left(\tilde{\mu}_n - \sqrt{\tilde{\mu}_n^2 + 4}\right)\right] \cup \left[\tilde{\mu}_1, \frac{1}{2}\left(\tilde{\mu}_n + \sqrt{\tilde{\mu}_n^2 + 4}\right)\right]. \tag{5.19}$$

Proof Observe that,

$$P^{-\frac{1}{2}}CP^{-\frac{1}{2}} = \begin{pmatrix} P_A^{-\frac{1}{2}}AP_A^{-\frac{1}{2}} & P_A^{-\frac{1}{2}}B^T\left(BP_A^{-1}B^T\right)^{-\frac{1}{2}} \\ \left(BP_A^{-1}B^T\right)^{-\frac{1}{2}}BP_A^{-\frac{1}{2}} & 0 \end{pmatrix} = \begin{pmatrix} \tilde{A} & \tilde{B}^T \\ \tilde{B} & 0 \end{pmatrix},$$

is a saddle-point matrix. Further,

$$\tilde{B}\tilde{B}^T = \left(BP_A^{-1}B^T\right)^{-\frac{1}{2}} BP_A^{-\frac{1}{2}}P_A^{-\frac{1}{2}}B^T \left(BP_A^{-1}B^T\right)^{-\frac{1}{2}} = I,$$

where I is the identity matrix. The result follows immediately from Lemma 9 since the singular values of $\tilde{B}$ are all equal to one. □

The success of the preconditioner (5.17) is completely determined by the choice of P_A. An h-optimal and $\mathcal{A}$-optimal eigenvalue bound results if an h-optimal and $\mathcal{A}$-optimal approximation for the weighted mass matrix A can be found. We now consider the simple choice $P_A = \mathtt{diag}\,(A) = A_{diag}$.

5.2.1 Diagonal scaling for the weighted mass matrix

The advantage here is that it is sufficient to consider the *element* matrices A^K.

Lemma 23 *Let λ_{min}^K and λ_{max}^K denote the minimum and maximum eigenvalues of the diagonally scaled element matrix* $(\mathtt{diag}\ (A^K))^{-1}A^K$. *Then,*

$$\min_K \left\{\lambda_{min}^K\right\} \leq \tilde{\mu}_1, \quad \tilde{\mu}_n \leq \max_K \left\{\lambda_{max}^K\right\}.$$

Proof See Wathen [101]. □

Using this result, we now consider the efficiency of diagonal scaling for A using triangle and square elements in turn.

Triangles

Let K be a right-angled triangle of edge length h_K with oriented normal vectors as shown in Fig. 2.4. Recall that,

$$\mathcal{A}|_K = \begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{pmatrix}$$

denotes the coefficient tensor $\mathcal{A}$ evaluated at the centroid of that element. We distinguish three cases. For full $\mathcal{A}|_K$, recall from section 2.3.4 that integration yields,

$$A^K = \frac{h_K^2}{12det(\mathcal{A}|_K)} \begin{pmatrix} a_{22}+3a_{11}+3a_{12} & \sqrt{2}\,(a_{22}-a_{11}+a_{12}) & a_{22}+a_{11}+3a_{12} \\ \sqrt{2}\,(a_{22}-a_{11}+a_{12}) & 2\,(a_{22}+a_{11}-a_{12}) & \sqrt{2}\,(a_{22}-a_{11}-a_{12}) \\ a_{22}+a_{11}+3a_{12} & \sqrt{2}\,(a_{22}-a_{11}-a_{12}) & 3a_{22}+a_{11}+3a_{12} \end{pmatrix},$$

which simplifies to,

$$A^K = h_K^2 \begin{pmatrix} \frac{1}{12a_{11}} + \frac{1}{4a_{22}} & \frac{\sqrt{2}}{12}\left(\frac{1}{a_{11}} - \frac{1}{a_{22}}\right) & \frac{1}{12a_{11}} + \frac{1}{12a_{22}} \\ \frac{\sqrt{2}}{12}\left(\frac{1}{a_{11}} - \frac{1}{a_{22}}\right) & \frac{1}{6a_{11}} + \frac{1}{6a_{22}} & \frac{\sqrt{2}}{12}\left(\frac{1}{a_{11}} - \frac{1}{a_{22}}\right) \\ \frac{1}{12a_{11}} + \frac{1}{12a_{22}} & \frac{\sqrt{2}}{12}\left(\frac{1}{a_{11}} - \frac{1}{a_{22}}\right) & \frac{1}{4a_{11}} + \frac{1}{12a_{22}} \end{pmatrix},$$

for diagonal $\mathcal{A}|_K$, and further to,

$$A^K = h_K^2 \begin{pmatrix} \frac{1}{3a_{11}} & 0 & \frac{1}{6a_{11}} \\ 0 & \frac{1}{3a_{11}} & 0 \\ \frac{1}{6a_{11}} & 0 & \frac{1}{3a_{11}} \end{pmatrix},$$

if $a_{11} = a_{22}$.

Applying diagonal scaling, yields,

$$(\mathtt{diag}(A^K))^{-1}A^K = \begin{pmatrix} 1 & \frac{\sqrt{2}(a_{22}-a_{11}+a_{12})}{a_{22}+3a_{11}+3a_{12}} & \frac{a_{22}+a_{11}+3a_{12}}{a_{22}+3a_{11}+3a_{12}} \\ \frac{(a_{22}-a_{11}+a_{12})}{\sqrt{2}(a_{22}+a_{11}-a_{12})} & 1 & \frac{(a_{22}-a_{11}-a_{12})}{\sqrt{2}(a_{22}+a_{11}-a_{12})} \\ \frac{a_{22}+a_{11}+3a_{12}}{3a_{22}+a_{11}+3a_{12}} & \frac{\sqrt{2}(a_{22}-a_{11}-a_{12})}{3a_{22}+a_{11}+3a_{12}} & 1 \end{pmatrix},$$

in the first case,

$$(\mathtt{diag}(A^K))^{-1}A^K = \begin{pmatrix} 1 & \frac{\sqrt{2}(a_{22}-a_{11})}{a_{22}+3a_{11}} & \frac{a_{22}+a_{11}}{a_{22}+3a_{11}} \\ \frac{(a_{22}-a_{11})}{\sqrt{2}(a_{22}+a_{11})} & 1 & \frac{(a_{22}-a_{11})}{\sqrt{2}(a_{22}+a_{11})} \\ \frac{a_{22}+a_{11}}{3a_{22}+a_{11}} & \frac{\sqrt{2}(a_{22}-a_{11})}{3a_{22}+a_{11}} & 1 \end{pmatrix},$$

in the second and,

$$(\mathtt{diag}(A^K))^{-1}A^K = \begin{pmatrix} 1 & 0 & \frac{1}{2} \\ 0 & 1 & 0 \\ \frac{1}{2} & 0 & 1 \end{pmatrix}, \tag{5.20}$$

for $\mathcal{A}|_K = a_{11}\mathcal{I}$.

In all cases it is clear that the eigenvalues of the scaled matrices are independent of h_K. For piecewise constant scalar coefficients, we trivially obtain the following result.

Lemma 24 *For meshes of right-angled triangles, and piecewise constant scalar coefficients of the form $\mathcal{A}|_K = \alpha_K\mathcal{I}$ with $\alpha_K \in \mathbb{R}^+$, $\forall\, K \in T_h$, we obtain,*

$$\frac{1}{2} \leq \frac{\underline{u}^T A \underline{u}}{\underline{u}^T A_{diag} \underline{u}} \leq \frac{3}{2} \quad \forall \underline{u} \in \mathbb{R}^n \backslash \{\underline{0}\}, \tag{5.21}$$

where A is the weighted velocity mass matrix arising in the lowest-order Raviart-Thomas approximation of (2.12).

Proof The result follows immediately from Lemma 23 and the fact that the eigenvalues of (5.20), for each $K \in T_h$, are,

$$\lambda_1^K = \frac{1}{2}, \quad \lambda_2^K = 1, \quad \lambda_3^K = \frac{3}{2}. \quad \square$$

Hence, for discontinuous $\mathcal{A}$, diagonal scaling is $\mathcal{A}$-optimal, independently of the size of the jumps, provided that the coefficient is constant in each element. Notice that the mesh does not have to be uniform or quasi-uniform. The bound (5.21) also holds for locally refined meshes if each element is a right-angled triangle. A similar bound can be derived for equilateral triangles (see [76]).

For general diagonal and full tensors, characteristic polynomials can be derived and studied. We distinguish one important case. Suppose we have a diagonal coefficient tensor $\mathcal{A}|_K$ with strong anisotropy, say with a_{11} fixed and $a_{22} \to 0$ so that $a_{11} >> a_{22}$. In the limit $a_{22} \to 0$, we obtain,

$$(\mathtt{diag}(A^K))^{-1}A^K \to \begin{pmatrix} 1 & -\frac{\sqrt{2}}{3} & \frac{1}{3} \\ -\frac{1}{\sqrt{2}} & 1 & -\frac{1}{\sqrt{2}} \\ 1 & -\sqrt{2} & 1 \end{pmatrix}.$$

The characteristic polynomial is,

$$-\lambda^3 + 3\lambda^2 - \frac{4}{3}\lambda,$$

with roots,

$$\lambda_1 = 0, \quad \lambda_2 = \frac{3}{2} - \frac{1}{2}\sqrt{\frac{11}{3}}, \quad \lambda_3 = \frac{3}{2} + \frac{1}{2}\sqrt{\frac{11}{3}}.$$

In that case, we obtain,

$$0 \le \tilde{\mu}_1, \quad \tilde{\mu}_n \le \frac{3}{2} + \frac{1}{2}\sqrt{\frac{11}{3}}.$$

In practice, the minimum eigenvalue $\tilde{\mu}_1$ is close to zero. $\mathcal{A}$-optimal results cannot be achieved using triangular elements in such anisotropic cases but this can be remedied by using rectangular elements instead.

Squares

Let K be a square of edge length h_K with oriented normal vectors as shown in Fig.2.5. For full coefficient tensors we obtain,

$$A^K = \frac{h_K^2}{4\,det(\mathcal{A})} \begin{pmatrix} \frac{4a_{22}}{3} & \frac{2a_{22}}{3} & -a_{12} & -a_{12} \\ \frac{2a_{22}}{3} & \frac{4a_{22}}{3} & -a_{12} & -a_{12} \\ -a_{12} & -a_{12} & \frac{4a_{11}}{3} & \frac{2a_{11}}{3} \\ -a_{12} & -a_{12} & \frac{4a_{11}}{3} & \frac{4a_{11}}{3} \end{pmatrix},$$

and diagonal scaling yields,

$$(\mathtt{diag}(A^K))^{-1}A^K = \begin{pmatrix} 1 & \frac{1}{2} & -\frac{3a_{12}}{4a_{22}} & -\frac{3a_{12}}{4a_{22}} \\ \frac{1}{2} & 1 & -\frac{3a_{12}}{4a_{22}} & -\frac{3a_{12}}{4a_{22}} \\ -\frac{3a_{12}}{4a_{11}} & -\frac{3a_{12}}{4a_{11}} & 1 & \frac{1}{2} \\ -\frac{3a_{12}}{4a_{11}} & -\frac{3a_{12}}{4a_{11}} & \frac{1}{2} & 1 \end{pmatrix}.$$

The associated characteristic polynomial is,

$$\lambda^4 - 4\lambda^3 + \left(\frac{11}{2} - \frac{9a_{12}^2}{4a_{11}a_{22}}\right)\lambda^2 + \left(\frac{9a_{12}^2}{4a_{11}a_{22}} - 3\right)\lambda + \left(\frac{9}{16} - \frac{9a_{12}^2}{16a_{11}a_{22}}\right),$$

yielding eigenvalues,

$$\lambda_1^K = \frac{1}{2}, \quad \lambda_2^K = \frac{1}{2}, \quad \lambda_3^K = \frac{3}{2}\left(1 + \frac{a_{12}}{\sqrt{a_{11}a_{22}}}\right), \quad \lambda_4^K = \frac{3}{2}\left(1 - \frac{a_{12}}{\sqrt{a_{11}a_{22}}}\right).$$

For diagonal $\mathcal{A}|_K$, we obtain,

$$(\mathtt{diag}(A^K))^{-1}A^K = \begin{pmatrix} 1 & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 1 & 0 & 0 \\ 0 & 0 & 1 & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & 1 \end{pmatrix}, \tag{5.22}$$

and eigenvalues,

$$\lambda_1^K = \frac{1}{2}, \quad \lambda_2^K = \frac{1}{2}, \quad \lambda_3^K = \frac{3}{2}, \quad \lambda_4^K = \frac{3}{2}, \tag{5.23}$$

independently of a_{11} and a_{22}. An h-optimal and $\mathcal{A}$-optimal approximation results for all diagonal coefficient tensors including anisotropic cases.

Lemma 25 *For meshes of squares, and any piecewise constant diagonal coefficient tensor, we obtain,*

$$\frac{1}{2} \leq \frac{\underline{u}^T A \underline{u}}{\underline{u}^T A_{diag}\underline{u}} \leq \frac{3}{2} \quad \forall \underline{u} \in I\!R^n \backslash \{\underline{0}\}, \tag{5.24}$$

where A is the weighted velocity mass matrix arising in the lowest-order Raviart-Thomas approximation of (2.12).

Proof The result follows immediately from Lemma 23, (5.22) and (5.23). □

Note that square, uniform meshes are not essential. The general requirement for h-optimality is that the elements should not be too stretched.

In $I\!R^3$, the same analysis can be applied. For any given coefficient tensor in $I\!R^2$ or $I\!R^3$, Lemma 23 offers a quick and cheap criteria for assessing optimality of the suggested preconditioner,

$$P = \begin{pmatrix} A_{diag} & 0 \\ 0 & BA_{diag}^{-1}B^T \end{pmatrix}, \tag{5.25}$$

via the bound (5.19). Note that the optimality of diagonal scaling for A must be verified before implementing (5.25). For full coefficient tensors, diagonal scaling for A may be less efficient.

5.2.2 Jump operator

Returning to more abstract considerations, it is clear that the matrix X, representing the 'pure' jump operator in (5.8) does not provide an adequate preconditioner for the Schur complement matrix. Moreover, h-optimality is only achieved with quasi-uniform meshes. However, it is now easy to see that the matrix, $BA_{diag}^{-1}B^T$ in (5.25) provides exactly the right kind of scaling. Indeed, given $\tilde{\mu_1}$ and $\tilde{\mu_n}$ satisfying (5.16), with $P_A = A_{diag}$, we obtain,

$$\frac{1}{\tilde{\mu}_n} \leq \frac{\underline{p}^T BA^{-1}B^T\underline{p}}{\underline{p}^T BA_{diag}^{-1}B^T\underline{p}} \leq \frac{1}{\tilde{\mu}_1} \quad \forall \underline{p} \in I\!R^m \backslash \{\underline{0}\}, \tag{5.26}$$

which is now an $\mathcal{A}$-optimal and h-optimal approximation if diagonal scaling for A is $\mathcal{A}$-optimal and h-optimal. We shall see that this can be achieved without quasi-unform meshes.

At first glance, it appears that the preconditioner (5.25) has very little to do with the preconditioner (5.15) and the norms in which the alternative inf-sup inequality property (5.7) were established. However, with a little linear algebra, we can show that $BA_{diag}^{-1}B^T$ is actually a weighted jump operator, representing a weighted mesh-dependent norm of the form (5.5). To see this in $I\!R^2$, observe that,

$$\begin{aligned}
\| p \|_{1,h}^2 = \underline{p}^T X \underline{p} &= h^{-1} \sum_e \int_e [p]_e^2 \\
&= h^{-1} \sum_e h_e \left(p\,|_{K_i} - p\,|_{K_j} \right)^2 \quad e \subset K_i \cap K_j \\
&= h^{-1} \sum_e h_e \left(\underline{p}_i^2 + \underline{p}_j^2 - 2\underline{p}_i \underline{p}_j \right) \quad e \subset K_i \cap K_j \\
&= h^{-1} \sum_{K_i} \underline{p}_i^2 \left(\sum_{e \subset K_i} h_e \right) - h^{-1} \sum_{K_i} \sum_{K_j \neq K_i} \underline{p}_i \underline{p}_j \left(\sum_{e \subset K_i \cap K_j} h_e \right), \\
&= \sum_{K_i} \underline{p}_i^2 \left(\sum_{e \subset K_i} \frac{h_e}{h} \right) - \sum_{K_i} \sum_{K_j \neq K_i} \underline{p}_i \underline{p}_j \left(\sum_{e \subset K_i \cap K_j} \frac{h_e}{h} \right). \qquad (5.27)
\end{aligned}$$

That is, the jump operator (5.5) has the algebraic representation (5.27).

Now consider the matrix $BA_{diag}^{-1}B^T$. We have,

$$\left(BA_{diag}^{-1} B^T \right)_{ij} = \sum_e \frac{B_{ie} B_{je}}{A_{ee}}, \quad i, j = 1 : m.$$

Recalling (2.41), the diagonal entries are,

$$\left(BA_{diag}^{-1} B^T \right)_{ii} = \sum_e \frac{B_{ie}^2}{A_{ee}} = \sum_{e \subset K_i} \frac{B_{ie}^2}{A_{ee}} = \sum_{e \subset K_i} \frac{h_e^2}{A_{ee}}, \qquad i = 1 : m. \qquad (5.28)$$

For the off-diagonal entries $(i \neq j)$, recall that $\left(BA_{diag}^{-1} B^T \right)_{ij}$ is zero unless elements K_i and K_j share a common edge e. If K_i and K_j share an edge,

$$B_{ie} = \int_e \vec{\varphi}_e \cdot \vec{n}_e^{K_i}\, ds, \quad B_{je} = \int_e \vec{\varphi}_e \cdot \vec{n}_e^{K_j}\, ds.$$

By construction we have,

$$\vec{\varphi}_e \cdot \vec{\nu}^k = \begin{cases} 1 & \text{if } k = e, \\ 0 & \text{if } k \neq e, \end{cases}$$

where $\vec{\nu}^k$ is a fixed oriented normal vector at edge k. Since it is essential to impose continuity of normal components, one of $\left\{ \vec{n}_e^{K_i}, \vec{n}_e^{K_j} \right\}$ coincides with $\vec{\nu}^k$; the other carries the opposite sign. Hence,

$$\left(BA_{diag}^{-1} B \right)_{ij} = -\frac{B_{je}^2}{A_{ee}} = -\frac{h_e^2}{A_{ee}}. \qquad (5.29)$$

Using (5.28) and (5.29), we obtain,

$$\underline{p}^T BA_{diag}^{-1} B^T \underline{p} = \sum_{K_i} \sum_{K_j} \underline{p}_i \left(BA_{diag}^{-1} B^T \right)_{ij} \underline{p}_j$$

$$
\begin{aligned}
&= \sum_{K_i} \underline{p}_i^2 \left(BA_{diag}^{-1}B^T\right)_{ii} + \sum_{K_i} \sum_{K_j \neq K_i} \underline{p}_i \left(BA_{diag}^{-1}B^T\right)_{ij} \underline{p}_j \\
&= \sum_{K_i} \underline{p}_i^2 \left(\sum_{e \subset K_i} \frac{B_{ie}^2}{A_{ee}} \right) + \sum_{i} \sum_{K_j \neq K_i} \underline{p}_i \left(\sum_{e \subset K_i \cap K_j} \frac{B_{ie}B_{je}}{A_{ee}} \right) \underline{p}_j \\
&= \sum_{K_i} \underline{p}_i^2 \left(\sum_{e \subset K_i} \frac{h_e^2}{A_{ee}} \right) - \sum_{K_i} \sum_{K_j \neq K_i} \underline{p}_i \underline{p}_j \left(\sum_{e \subset K_i \cap K_j} \frac{h_e^2}{A_{ee}} \right).
\end{aligned}
$$

Recalling, now, that the diagonal entries of the weighted velocity mass matrix are,

$$
A_{ee} = \int_{\Omega} \mathcal{A}^{-1} \vec{\varphi}_e \cdot \vec{\varphi}_e \, d\Omega,
$$

we obtain, by Lemma 10 and condition (2.1),

$$
c_1 \gamma \, h_{min}^2 \leq A_{ee} \leq c_2 \Gamma h^2,
$$

for all edges e. Hence, for quasi-uniform meshes we obtain,

$$
\underline{p}^T BA_{diag}^{-1}B^T \underline{p} = \sum_{K_i} \underline{p}_i^2 \left(\sum_{e \subset K_i} \frac{h_e^2}{C_e(\mathcal{A})h^2} \right) - \sum_{K_i} \sum_{K_j \neq K_i} \underline{p}_i \underline{p}_j \left(\sum_{e \subset K_i \cap K_j} \frac{h_e^2}{C_e(\mathcal{A})h^2} \right), \tag{5.30}
$$

where $C_e(\mathcal{A})$ is a constant depending on $\mathcal{A}$. Comparing (5.30) with (5.27), we see that $BA_{diag}^{-1}B^T$ is also an algebraic representation of a jump operator. The difference is that in (5.30), each term in the sum is weighted with respect to $\mathcal{A}$.

To summarise, we have achieved, via algebraic arguments, an ideal preconditioner (5.25), whose diagonal blocks roughly represent coefficient and mesh-dependent versions of the $L^2(\Omega)$ and $H^1(\Omega)$ norms,

$$
\underline{u}^T A_{diag} \underline{u} = \| \vec{u} \|_{0,\mathcal{A}}^2, \quad \underline{p}^T BA_{diag}^{-1}B^T \underline{p} = \| p \|_{1,h,\mathcal{A}}^2 .
$$

5.3 Practical preconditioning

To obtain a practical scheme, we again look to multigrid methods to approximately invert the sparse matrix $BA_{diag}^{-1}B^T$ in $O(m)$ flops. (Recall that, here, the dimension of the system, m, corresponds to the number of elements in the mesh.)

First, observe that for *any* given symmetric positive definite matrix V satisfying,

$$
\theta \leq \frac{\underline{p}^T BA_{diag}^{-1}B^T \underline{p}}{\underline{p}^T V \underline{p}} \leq \Theta \quad \forall \underline{p} \in I\!R^m \setminus \{\underline{0}\}, \tag{5.31}
$$

for some positive constants θ and Θ, Lemma 22 can be extended to obtain the following theoretical eigenvalue bound.

Lemma 26 *Let $0 < \tilde{\mu}_1 \ldots \leq \tilde{\mu}_n$ be the eigenvalues of $A_{diag}^{-1}A$, then the eigenvalues of the generalised eigenvalue problem,*

$$\underbrace{\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix}}_{C} \begin{pmatrix} \underline{u} \\ \underline{p} \end{pmatrix} = \lambda \underbrace{\begin{pmatrix} A_{diag} & 0 \\ 0 & V \end{pmatrix}}_{P} \begin{pmatrix} \underline{u} \\ \underline{p} \end{pmatrix}, \tag{5.32}$$

lie in the union of the intervals,

$$\left[\frac{1}{2}\left(\tilde{\mu}_1 - \sqrt{\tilde{\mu}_1^2 + 4\Theta}\right), \frac{1}{2}\left(\tilde{\mu}_n - \sqrt{\tilde{\mu}_n^2 + 4\theta}\right)\right] \cup \left[\tilde{\mu}_1, \frac{1}{2}\left(\tilde{\mu}_n + \sqrt{\tilde{\mu}_n^2 + 4\Theta}\right)\right], \tag{5.33}$$

where θ and Θ are positive constants satisfying (5.31).

Proof The result follows directly from Lemma 9 and Lemma 22 with $\tilde{A} = A_{diag}^{-\frac{1}{2}} A A_{diag}^{-\frac{1}{2}}$ and $\tilde{B} = V^{-\frac{1}{2}} B A_{diag}^{-\frac{1}{2}}$. Applying (5.31), we obtain,

$$\begin{aligned} \underline{p}^T \tilde{B}\tilde{B}^T \underline{p} &= \underline{p}^T V^{-\frac{1}{2}} \left(B A_{diag}^{-1} B^T\right) V^{-\frac{1}{2}} \underline{p} \\ &\leq \Theta \underline{p}^T V^{-\frac{1}{2}} V V^{-\frac{1}{2}} \underline{p} \\ &= \Theta \underline{p}^T \underline{p}. \end{aligned}$$

That is, the maximum singular value $\tilde{\sigma}_m$ of $\tilde{B}$ satisfies, $\tilde{\sigma}_m^2 \leq \Theta$. Similarly, the minimum singular value satisfies $\tilde{\sigma}_1^2 \geq \theta$. □

Observe that we are dealing with a sparse representation of the scalar diffusion operator, $\nabla \cdot \mathcal{A} \nabla$—the most widely studied of standard elliptic operators—and so many possibilities for V arise. Suitable codes in the public domain include multigrid, domain decomposition, algebraic multilevel iteration, AMLI, (see Axelsson and Vassilevski, [9]) and algebraic multigrid, AMG, (see Ruge and Stüben, [83], [84]). However, our goal is to have a truly black-box method which requires no tuning for anisotropic and discontinuous coefficients. In addition, we would like to solve problems with non quasi-uniform meshes without having to generate large amounts of geometric information. Only the second two types of method are suited to these requirements.

Algebraic methods have recently undergone a resurgence in popularity due to increasing geometric complexity of physical models. AMLI and AMG offer the possibility of creating a hierarchy of 'levels' using only information in the coefficient matrix. No geometric information is required. The former is based on block incomplete factorisation and recursively solves subproblems using matrix polynomials defined on the intervals of eigenvalues of the created sub-matrices. Public domain code is currently limited to $\mathbb{R}^2$ (see Neytcheva, [73].)

The classical AMG method, introduced and analysed by Ruge and Stüben in the 1980s, is still freely available as the black-box code[1] `amg1r5`. Today, many variants exist (see, for example, [67], [22], [34], [35], [96]) and 'AMG' now refers to a philosophy rather than a single algorithm. From an algorithmic point of view, it fits into the V-cycle or W-cycle framework outlined in Fig. 4.2 in Chapter 4. However, coarsening, interpolation and restriction operations are all performed using the simple algebraic notion of 'strength of dependence' of the matrix entries. Consequently, all multigrid components are tailored to the underlying PDE operator.

The theoretical foundations of AMG, are, however, somewhat sketchy compared to geometric multigrid methods. AMG is heuristically motivated and full V-cycle convergence theory has not been achieved to date using *purely* algebraic arguments. Recall that our aim is to satisfy (5.31) with constants θ and Θ that are independent of h and $\mathcal{A}$. However, we must not be discouraged. Two-level AMG V-cycle convergence *is* proved to be h-optimal (see Ruge and Stüben, [84], [93]) for a particular class of matrices, namely, *symmetric M-matrices.* These are diagonally dominant matrices with positive diagonal entries and negative off-diagonal entries. Further, a wealth of numerical evidence (see [83], [84], [39], [95]) demonstrates that *full* V-cycle convergence is likely to be h-optimal for matrices belonging to this class. Lemma 27, below, tells us that AMG is perfectly suited to the problem at hand.

Lemma 27 *The matrix $BA_{diag}^{-1}B^T$, arising in the lowest-order Raviart-Thomas approximation of (2.12), with the choice of degrees of freedom (2.39), is symmetric, positive definite and diagonally dominant with positive diagonal entries and negative*

[1] The code can be downloaded from `www.mgnet.org`

off-diagonal entries.

Proof Symmetry is obvious. By assumption,

$$\left(\mathcal{A}^{-1}\vec{v},\,\vec{v}\right) \geq \gamma\,(\vec{v},\,\vec{v})\,, \quad \forall\,\vec{v}:\Omega \to I\!R^d,$$

with $\gamma > 0$. The matrix A therefore has positive diagonal entries since,

$$A_{ii} \;=\; \left(\mathcal{A}^{-1}\vec{\varphi}_i,\vec{\varphi}_i\right) \;\geq\; \gamma\,(\vec{\varphi}_i,\vec{\varphi}_i) > 0, \quad i = 1:n.$$

Positive definiteness of $BA_{diag}^{-1}B^T$ follows from that of A_{diag}. Now,

$$\left(BA_{diag}^{-1}B^T\right)_{ij} = \sum_e \frac{B_{ie}B_{je}}{A_{ee}}, \quad i,j = 1:m,$$

where e denotes an edge (or face) of T_h. The diagonal entries are positive by (5.28) since the diagonal entries of A are positive. The off-diagonal entries ($i \neq j$) are negative by (5.29). Diagonal dominance follows immediately from (5.28) and (5.29). □

Remark 12 *For the alternative choice (2.40) for the degrees of freedom for* V_h*, we obtain,* $B_{ij} = \pm 1$ *with,*

$$\left(BA_{diag}^{-1}B^T\right)_{ii} = \sum_{e\subset K_i} \frac{1}{A_{ee}}, \quad \left(BA_{diag}^{-1}B^T\right)_{i,j} = -\frac{1}{A_{ee}}, \quad i \neq j,$$

and Lemma 27 still holds.

To summarise, $BA_{diag}^{-1}B^T$ is a symmetric M-matrix in $I\!R^2$ and $I\!R^3$. When applying black-box methods, it is crucial to give consideration to the structure and stencil of the coefficient matrix of the system to be solved. Observe that the saddle-point matrix C and the Schur complement matrix S cannot be tackled directly with AMG. Here, we are applying the method to a matrix which has *all* the necessary characteristics to achieve h-optimal convergence.

5.3.1 Algebraic multigrid

We now choose V in (5.31) to be a single V-cycle of the AMG code `amg1r5`. Since our argument is that we can apply it as a black-box method—without geometric information and without having to tune any parameters to deal with different coefficients—we do not go into the fine details of the algorithm. Rather we give a brief outline of the main concepts and make some comments on how the method *must* be tailored for use as a preconditioner with MINRES.

Full details of the original code can be found in [83]. Two-level convergence theory is discussed in [84]. A review of AMG philosophy and recent developments is given in [94] by Stüben. A range of numerical examples is given by Cleary et al. in [39]. The reader should bear in mind that many variants of each of the AMG components to be described exist. For symmetric M-matrices, however, the basic choices perform efficiently.

Henceforth in this chapter, we are concerned with generic linear systems, of dimension m,

$$\sum_j M_{ij}\,\underline{x}_i \;=\; \underline{b}_i \qquad i = 1 : m,$$

where M is a symmetric M-matrix. The reader should now recall the terminology of the standard geometric multigrid V-cycle in Fig. 4.2. The components of basic two-level AMG, which is based on the Galerkin principle (see Chapter 4), are summarised in Table 5.1.

Fine grid:	$\Omega_J = 1 : m$
Splitting routine:	$\Omega_J \to C \cup F$
Coarse grid:	$\Omega_{J-1} = C \subset \Omega_J$
Smoother:	$\mathcal{S}$ = stationary iterative method
Interpolation:	$\left(I^J_{J-1}\underline{e}^{J-1}\right)_i = \begin{cases} \underline{e}^{J-1}_i & \text{if } i \in C \\ \sum_{j\in P_i} \omega_{ij}\underline{e}^{J-1}_j & \text{if } i \in F \end{cases}$
Restriction:	$I^{J-1}_J = \left(I^J_{J-1}\right)^T$
Coarse grid operator:	$M^{J-1} = I^{J-1}_J M^J I^J_{J-1}$

Table 5.1: Basic components of AMG

To define a concrete algorithm we must specify a coarsening algorithm and choose interpolation variables P_i and weights $\{\omega_{ij}\}$. In contrast to standard multigrid, the AMG philosophy is to fix the smoother to be a stationary iterative method such as Jacobi or Gauss-Seidel and then design interpolation to eliminate the remaining error components. For discontinuous coefficients, error remaining after this type of smoothing will be geometrically oscillatory (see Alcouffe et al., [3], or Wan, [100]) so 'smooth error' is now defined to be any error components not reduced by the initial iteration. It has no geometric connotation.

The smoother is fixed in this way because it is easy to write down algebraic equations for smooth error for the class of symmetric M-matrices. Consider, for example, weighted-Jacobi iteration, with scaling parameter σ. Splitting $M = D - L - U$, into diagonal and triangular parts, recall that the associated iteration matrix is $G = \left(I - \sigma D^{-1} M\right)$. The iteration stalls after i steps if and only if, $\underline{e}^{(i+1)} \approx \underline{e}^{(i)} = G^i \underline{e}^{(0)}$. Thus, smooth error, satisfies $G\underline{e} \approx \underline{e}$, or alternatively, $(G\underline{e}, M\underline{e}) \approx (\underline{e}, M\underline{e})$. As a consequence, we obtain,

$$\left(D^{-1} M\underline{e}, M\underline{e}\right) << (M\underline{e}, \underline{e}),$$

or equivalently,

$$\sum_{i=1}^{m} \frac{\underline{r}_i^2}{M_{ii}} << \sum_{i=1}^{m} \underline{r}_i \underline{e}_i,$$

and so, on average, for each i, $\mid \underline{r}_i \mid << M_{ii} \mid \underline{e}_i \mid$. The same result holds for Gauss-Seidel relaxation. Now we can say 'smooth error is characterised by small residuals' and approximately satisfies $M\underline{e} \approx 0$. Under this assumption, solving the error equations for the ith component, or node, yields,

$$M_{ii}\, \underline{e}_i = \left(\underline{r}_i - \sum_{j \neq i} M_{ij}\, \underline{e}_j\right) \approx -\left(\sum_{j \neq i} M_{ij}\, \underline{e}_j\right) = -\left(\sum_{j \in N_i} M_{ij}\, \underline{e}_j\right),$$

where $N_i = \{j \mid M_{ij} \neq 0\}$ denotes the set of nodes that have non-zero connection to i. This is the error that interpolation must capture efficiently. To derive a concrete scheme, we approximate the sum using only the *strongest* connections.

For M-matrices, a node j is defined to be *strongly connected* to i, relative to some given parameter α_S with $0 < \alpha_S < 1$, if

$$-M_{ij} \geq \alpha_S \max_{k \neq i}\{-M_{ik}\}. \tag{5.34}$$

(Here, off-diagonal connections are implicitly assumed to be negative.) For any node i, we can then identify a set of strongly connected neighbours S_i, by examining the magnitude of the entries in the ith row of M,

$$S_i = \{j \in N_i \mid -M_{ij} \geq \alpha_S \max_{k \neq i}\{-M_{ik}\}\}.$$

Now, at level J, given a splitting of the nodes into coarse level (C) and fine level (F) subsets, we can distinguish, for each i, three types of connections. We have $N_i = C_i^s \cup D_i^s \cup D_i^w$, where $C_i^s = S_i \cap C$, is the set of strong C-connections, $D_i^s = S_i \cap F$, denotes strong F-connections and $D_i^w = \{j \in N_i \mid j \notin S_i\}$, is the set of all connections that are weak.

Hence, for smooth error we have,

$$M_{ii}\underline{e}_i \quad \approx \quad -\sum_{j\in C_i^s} M_{ij}\,\underline{e}_j - \sum_{j\in D_i^s} M_{ij}\,\underline{e}_j - \sum_{j\in D_i^w} M_{ij}\,\underline{e}_j. \tag{5.35}$$

A simple, direct interpolation scheme to determine $\underline{e}_i$ as a linear combination of errors at neighbouring points is derived from (5.35) by choosing the set of interpolation points $P_i = C_i^S$. First, we must express $\underline{e}_j$ for $j \in D_i^s \cup D_i^w$ in terms of $\underline{e}_i$. For $j \in D_i^w$ we can simply write $\underline{e}_j \approx \underline{e}_i$, and lump those connections onto the diagonal,

$$\left(M_{ii} + \sum_{j\in D_i^w} M_{ij}\right)\underline{e}_i \quad \approx \quad -\sum_{j\in C_i^s} M_{ij}\,\underline{e}_j - \sum_{j\in D_i^s} M_{ij}\,\underline{e}_j.$$

Note that if D_i^w contains any connections of significant magnitude, it can be shown (see [93, pp.439–443]) that error varies slowly in the direction of strong dependence, provided that the M-matrix property is satisfied. Hence, this is always a valid approximation. Now, for each $j \in D_i^s$, we approximate $\underline{e}_j$ by a weighted linear combination of errors at points in C_i^S. That is,

$$\underline{e}_j \approx \frac{\sum_{k\in C_i^s} M_{jk}\,\underline{e}_k}{\sum_{k\in C_i^s} M_{jk}}, \qquad \forall\, j\ \in D_i^s.$$

The motivation for this is that coarse grids are constructed so that any F node that is strongly connected to i, also strongly depends on the set C_i^s. Finally, then, we obtain,

$$\begin{aligned}\left(M_{ii} + \sum_{n\in D_i^w} M_{in}\right)\underline{e}_i \quad &\approx \quad -\sum_{j\in C_i^s} M_{ij}\underline{e}_j - \sum_{m\in D_i^s} M_{im}\left(\frac{\sum_{k\in C_i^s} M_{mk}\underline{e}_k}{\sum_{k\in C_i^s} M_{mk}}\right)\\ &= \quad -\sum_{j\in C_i^s} M_{ij}\underline{e}_j - \sum_{j\in C_i^s}\sum_{m\in D_i^s}\left(\frac{M_{im}M_{mj}}{\sum_{k\in C_i} M_{mk}}\right)\underline{e}_j\\ &= \quad -\sum_{j\in C_i^s}\left(M_{ij} + \sum_{m\in D_i^s}\left(\frac{M_{im}\,M_{mj}}{\sum_{k\in C_i^s} M_{mk}}\right)\right)\underline{e}_j.\end{aligned}$$

Hence we arrive at an interpolation as outlined in Table 5.1 with weights,

$$w_{ij} = \frac{M_{ij} + \sum_{m\in D_i^s}\left(\frac{M_{im}M_{mj}}{\sum_{k\in C_i^s} M_{mk}}\right)}{M_{ii} + \sum_{n\in D_i^w} M_{in}}.$$

If i is designated a C node, then simple injection is used.

To fully define interpolation, a splitting algorithm must also be chosen. Standard coarsening has two stages and aims to satisfy the following heuristics:

I No C-point should strongly depend on another C-point.

II For each F-point i, each $j \in S_i$ should either belong to the interpolation set C_i^s or should strongly depend on at least one point in C_i^s.

Condition I is imposed to regulate grid sizes and hence the amount of computational work per cycle. Condition II loosely says that there should be no strong F-F connections unless they have an interpolation variable in common. This is crucial for efficient interpolation. If an F node is strongly connected to i but does not depend on any of the points $j \in C_i^s$, error at that point is not represented in the interpolation.

For a given set of points $\{1, \ldots, m\}$, an initial C-F splitting, designed to satisfy I, is achieved as follows:

1. set $C = F = \emptyset$ and $U = 1 : m$
2. for all undecided points $i \in U$, calculate $S_i^T = \{j \mid i \in S_j\}$ and define the rank of i via $R_i =\mid S_i^T \mid$
3. choose an i with maximum rank R_i and set $C = C \cup \{i\}$, $U = U - \{i\}$,
4. $\forall j \in S_i^T \cap U$, set $F = F \cup \{j\}$ and $U = U - \{j\}$
5. $\forall k \in S_j^T \cap U$, set $R_k = R_k + 1$
6. $\forall j \in S_i^T \cap U$, set $R_j = R_j - 1$
7. Return to 3. and continue until $U = \emptyset$

We sweep through the undecided points and calculate, for each one, how many other nodes strongly depend on it. A rank is assigned to each node by counting how many other nodes strongly depend on it. It makes sense to designate the points with the largest ranks C nodes because they are the most desirable interpolation variables.

For example, applying this to $BA_{diag}^{-1}B^T$, with the default value $\alpha_S = 0.25$, on a geometrically uniform square grid with unit coefficients, produces the splitting depicted in Fig. 5.1. In this case, condition II is already satisfied. In general, however, more steps will be needed. Strong F-F connections are identified and examined. If there are strong F-F connections that do not share an interpolation C-point, some of the F-points are changed to C-points.

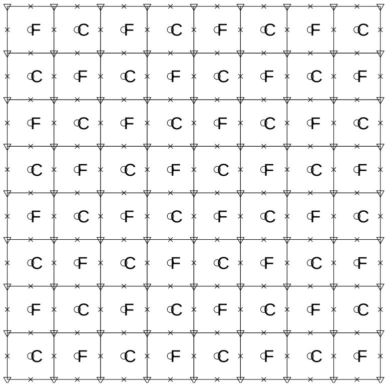

Figure 5.1: Initial coarsening, $h = \frac{1}{8}$, $\alpha_S = 0.25$, $\mathcal{A} = \mathcal{I}$

By constructing coarse grids (sets of C-points) in this way, the *strongest* connections tend to be chosen as C-points. This means that coarsening occurs in the direction of strong dependence. For example, applying `amg1r5` to $BA_{diag}^{-1}B^T$, with $\alpha_S = 0.25$, on a uniform grid with anisotropic coefficient tensor $\mathcal{A} = \mathtt{diag}(10^3, 1)$, produces the splitting depicted in Fig. 5.2. Geometrically, this makes sense since the solution only varies in the y direction.

Clearly, before multilevel cycling can begin, all of the coarse grids, interpolation weights, transfer operators and coarse grid operators have to defined. This is known as the *set-up phase*. The expense of this part of the algorithm varies from problem to problem, and may be considerably higher in $I\!R^3$ than in $I\!R^2$. However, the time cost is typically the same as that of a few V-cycles. When applying AMG as a *preconditioner*, set-up only has to be performed once, outside the iteration loop.

We will comment, in more detail, on the computational work involved in applying AMG in Chapter 6 where we compare the current approach with another method. Here,

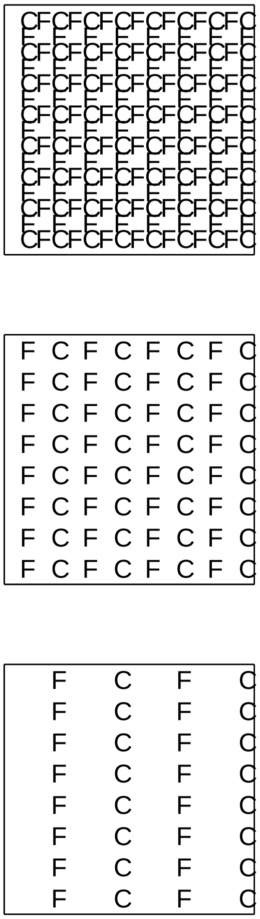

Figure 5.2: Full coarsening, $\alpha_S = 0.25$, anisotropic $\mathcal{A}$, triangles

we simply make a few remarks. Two important measures that directly influence the computational cost of applying AMG are the grid complexity, C_Ω, and the operator complexity, C_A, defined via,

$$C_\Omega : \quad = \quad \frac{\textit{Total number of nodes on all levels}}{\textit{number of nodes at finest level}},$$

$$C_A : \quad = \quad \frac{\textit{Total number of non-zeros in } M_j \textit{ on all levels}}{\textit{number of non-zeros in } M_J \textit{ at finest level}}.$$

In Tables 5.2–5.3, we summarise details of standard AMG coarsening for $BA_{diag}^{-1}B^T$ for a couple of test problems in $I\!\!R^2$. The results in the first table correspond to Poisson's equation, with unit coefficients, as described in Example 1 of Chapter 3. The second table gives information for the flow problem considered in Example 5 of Chapter 3 with jump parameter $\epsilon = 10^{-3}$.

Level	$m = 512$	$m = 2,048$	$m = 8,192$
1	512	2,048	8,192
2	256	1,024	4,096
3	86	324	1,366
4	26	119	460
5	-	29	126
6	-	11	34
7	-	-	11
C_A	2.45	2.66	2.72
C_Ω	1.72	1.74	1.74
Set-up time	0.015	0.059	0.262

Table 5.2: AMG coarsening, unit coefficients

Level	$m = 512$	$m = 2,048$	$m = 8,192$
1	512	2,048	8,192
2	257	1,025	4,097
3	89	345	1,369
4	33	117	471
5	14	39	153
6	-	11	58
7	-	-	23
C_A	2.57	2.64	2.75
C_Ω	1.77	1.75	1.75
Set-up time	0.015	0.059	0.262

Table 5.3: AMG coarsening, discontinuous coefficients

Observe that there is essentially no difference in the way set-up is performed for these two examples. The time cost grows only linearly with respect to problem size, as desired. The grid complexities are entirely acceptable, although the operator complexities are a little high by conventional standards. The value $C_A < 2$ is desirable. This is something that we would try to improve, with alternative coarsening strategies, if we intended to use AMG as a solver.

In fact, AMG has many parameters, controlling, for instance, the coarsest grid size and the strength of dependence of nodes. We would be tempted to try to tune these if we intended to apply AMG as a solver. However, since we are applying it as a preconditioner, there is little to gain. Convergence, for us, is determined by the minimisation properties of the Krylov subspace solver MINRES and the eigenvalue bound (5.33), *provided* that the V-cycle operator corresponding to the matrix V in (5.31) is *symmetric.* To achieve this, for our model problem, we must modify the smoother.

5.3.2 Symmetric smoothing

Freely available and commercial AMG codes apply point Gauss-Seidel smoothing. Recall that for a symmetric matrix M, we can write $M = D + L + U = D + L + L^T$. The standard Gauss-Seidel smoothing operator $\mathcal{S}_{GS}$ is defined as $\mathcal{S}_{GS} = (D + L)$, since, in solving the linear system $M\underline{x} = \underline{b}$ we have $\mathcal{S}_{GS}\underline{x} = \underline{b} - U\underline{x}$, from which we can derive the iteration,

$$\underline{x}^{(i+1)} = \mathcal{S}_{GS}^{-1} \left(\underline{b} - U\underline{x}^{(i)} \right).$$

This can be solved, pointwise, via,

$$\underline{x}^{(i+1)} = D^{-1} \left(\underline{b} - L\underline{x}^{(i+1)} - U\underline{x}^{(i)} \right),$$

sweeping though the indices 1 to m, updating each $\underline{x}^{(i)}$ as it becomes available. This is easy to implement but $\mathcal{S}_{GS} = (D + L)$ is not a symmetric operator. The resulting preconditioner,

$$P = \begin{pmatrix} A_{diag} & 0 \\ 0 & V \end{pmatrix}, \tag{5.36}$$

is not symmetric, the values of θ and Θ in (5.31) are complex and, in that case, MINRES convergence is totally unpredictable.

Instead, we choose the symmetric Gauss-Seidel smoothing operator,

$$\mathcal{S}_{SGS} = (D + L)\, D^{-1}\, (D + L)^T = \mathcal{S}_{GS} D^{-1} \mathcal{S}_{GS}^T,$$

which can be implemented by applying one iteration of standard Gauss-Seidel to obtain a vector $\hat{\underline{x}}$ and then solving,

$$\underline{x}^{(i+1)} = D^{-1} \left(\underline{b} - L\hat{\underline{x}} - U\underline{x}^{(i+1)} \right),$$

by performing a second Gauss-Seidel iteration, sweeping through the points in reverse order.

Numerical tests revealed that by performing one pre-smoothing step and one post-smoothing step with $\mathcal{S}_{SGS}$, instead of $\mathcal{S}_{GS}$, the resulting AMG V-cycle operator as defined by the algorithm in Fig 4.2 always yields real values θ and Θ. This means that the performance of the preconditioner (5.36) is completely determined by the eigenvalue bound (5.33) in Lemma 26.

Typical values of θ and Θ for discontinuous, variable and anisotropic coefficient test problems on uniform grids in $I\!R^2$ are listed in Table 5.4.

	Discontinuous		Variable		Anisotropic	
h	θ	Θ	θ	Θ	θ	Θ
$\frac{1}{8}$	0.918	1	0.957	1	0.949	1
$\frac{1}{16}$	0.845	1	0.955	1	0.948	1
$\frac{1}{32}$	0.836	1	0.950	1	0.948	1

Table 5.4: Values of θ and Θ, assorted coefficients

Observe that these values do not depend on the discretisation parameter and are insensitive to the coefficient tensors in each case. More details are given in the next section.

5.4 Preconditioned MINRES

We now report on MINRES convergence for a range of coefficients. We apply one V-cycle of `amg1r5`, with symmetric Gauss-Seidel smoothing, to the sparse matrix $BA_{diag}^{-1}B^T$. The code is implemented as a black-box; no parameters are estimated a priori. Iteration counts are reported for no preconditioning ($P = I$), ideal preconditioning (P_{ideal}) corresponding to P in (5.25), and AMG preconditioning (P_{amg}) corresponding to P in (5.36). The time units reported in parentheses are elapsed time in seconds for the total solve with a stopping tolerance of 10^{-6} on the relative residual error, using a `mex` fortran interface in MATLAB 6.0. The symbol '*' indicates that more than 500 iterations were required. Unless otherwise indicated, uniform meshes of triangles are employed.

Example 1

First, consider $\Omega = [0,1] \times [0,1]$ with unit coefficients, $f = 1$ and $g = 0$. MINRES iteration counts for the assembled linear system are reported in Table 5.5.

h	$P = I$	P_{ideal}	P_{amg}	
$\frac{1}{16}$	186	26	26	(0.18)
$\frac{1}{32}$	375	26	26	(0.48)
$\frac{1}{64}$	*	26	26	(1.90)
$\frac{1}{128}$	*	26	26	(9.06)

Table 5.5: MINRES iterations (and time), Example 1

Solve times grow linearly with respect to problem size, as desired. The eigenvalues of the multigrid preconditioned system are shown in Table 5.6. The spectral equivalence of the ideal and inexact version of the preconditioner is illustrated in Fig. 5.3. We obtain $\theta \approx 0.954$ and $\Theta = 1$, yielding the theoretical bound $[-0.781, -0.477] \cup [0.5, 2]$ in (5.33).

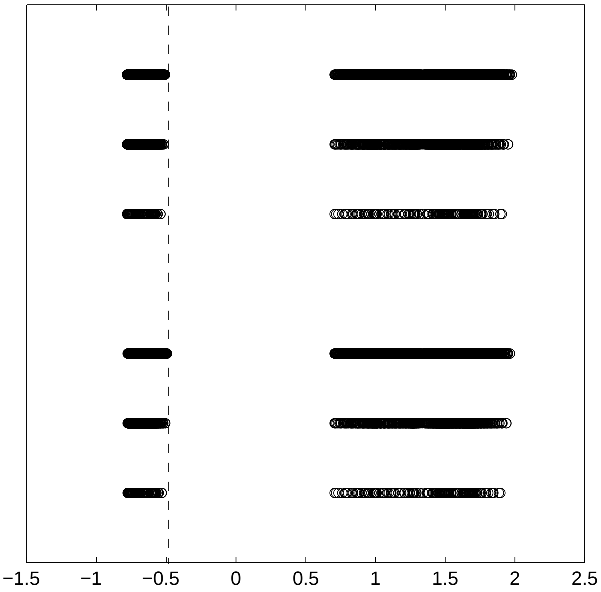

Figure 5.3: Eigenvalues of preconditioned system; P_{ideal} (top), P_{amg} (bottom), $h = \frac{1}{8}, \frac{1}{16}, \frac{1}{32}$

h	$\tilde{\mu}_1$	$\tilde{\mu}_n$	θ	Θ	Observed eigenvalues
$\frac{1}{8}$	0.5	1.5	0.957	1	$[-0.777,\ -0.532] \cup [0.707,\ 1.896]$
$\frac{1}{16}$	0.5	1.5	0.955	1	$[-0.777,\ -0.508] \cup [0.707,\ 1.939]$
$\frac{1}{32}$	0.5	1.5	0.954	1	$[-0.777,\ -0.496] \cup [0.707,\ 1.964]$

Table 5.6: Eigenvalues of preconditioned system, Example 1

Example 2

Choosing $\mathcal{A}$ to be the variable coefficient tensor (3.41) yields the iteration counts listed in Table 5.7.

h	$P = I$	P_{ideal}	P_{amg}	
$\frac{1}{16}$	$*$	26	26	(0.20)
$\frac{1}{32}$	$*$	26	26	(0.65)
$\frac{1}{64}$	$*$	26	26	(2.57)
$\frac{1}{128}$	$*$	26	26	(11.57)

Table 5.7: MINRES iterations (and time), Example 2

h	$\tilde{\mu}_1$	$\tilde{\mu}_n$	θ	Θ	Observed eigenvalues
$\frac{1}{8}$	0.5	1.5	0.957	1	$[-0.777,\ -0.539] \cup [0.707,\ 1.879]$
$\frac{1}{16}$	0.5	1.5	0.955	1	$[-0.777,\ -0.512] \cup [0.707,\ 1.931]$
$\frac{1}{32}$	0.5	1.5	0.950	1	$[-0.777,\ -0.498] \cup [0.707,\ 1.960]$

Table 5.8: Eigenvalues of preconditioned system, Example 2

Eigenvalues of the preconditioned system are given in Table 5.8. We obtain $\tilde{\mu_1} = \frac{1}{2}$, $\tilde{\mu_n} = \frac{3}{2}$, $\theta \approx 0.950$, and $\Theta = 1$, leading to the theoretical bound $[-0.7808, -0.5] \cup [0.5, 2]$ in (5.33).

Example 3

Next, we apply anisotropic coefficients $\mathcal{A} = \texttt{diag}(10^{-4}, 1)$. Since diagonal scaling for the weighted mass matrix is now not efficient on triangles, we use square elements. Iteration counts are listed in Table 5.9. We obtain $\tilde{\mu_1} = \frac{1}{2}$, $\tilde{\mu_n} = \frac{3}{2}$, $\theta \approx 0.9477$, and $\Theta = 1$, leading to the theoretical bound $[-0.781, -0.479] \cup [0.5, 2]$ in (5.33) which is observed to be tight.

h	$P=I$	P_{ideal}	P_{amg}
$\frac{1}{16}$	*	27	27 (0.10)
$\frac{1}{32}$	*	27	27 (0.31)
$\frac{1}{64}$	*	25	27 (1.29)
$\frac{1}{128}$	*	24	26 (8.68)

Table 5.9: MINRES iterations (and time), Example 3

Unlike the $H(div)$ preconditioner in Chapter 4, the performance of the AMG scheme is unaffected by the anisotropy.

Example 4

Next consider the flow problem, with discontinuous coefficients, described in Example 5 in Chapter 3. Iteration counts are given in Table 5.10.

h	$P=I$	P_{ideal}	P_{amg}
$\frac{1}{16}$	*	25	25 (0.17)
$\frac{1}{32}$	*	25	26 (0.46)
$\frac{1}{64}$	*	25	27 (2.67)
$\frac{1}{128}$	*	25	27 (10.74)

Table 5.10: MINRES iterations (and time), Example 4

Eigenvalues are reported in Table 5.11. Fig. 5.4 illustrates the spectral equivalence of the ideal and inexact versions of the preconditioner. We observe that $\theta \approx 0.835$ and $\Theta = 1$, yielding the theoretical bound $[-0.781, -0.433] \cup [0.5, 2]$ in (5.33).

h	$\tilde{\mu}_1$	$\tilde{\mu}_n$	θ	Θ	Observed eigenvalues
$\frac{1}{8}$	0.5	1.5	0.918	1	$[-0.774, -0.480] \cup [0.822, 1.965]$
$\frac{1}{16}$	0.5	1.5	0.845	1	$[-0.775, -0.448] \cup [0.822, 1.965]$
$\frac{1}{32}$	0.5	1.5	0.836	1	$[-0.775, -0.452] \cup [0.822, 1.974]$

Table 5.11: Eigenvalues of indefinite preconditioned system; Example 4

Now, if we increase the magnitude of the jump, θ remains bounded (see Table 5.12). The preconditioner is completely insensitive to the magnitude of the coefficient.

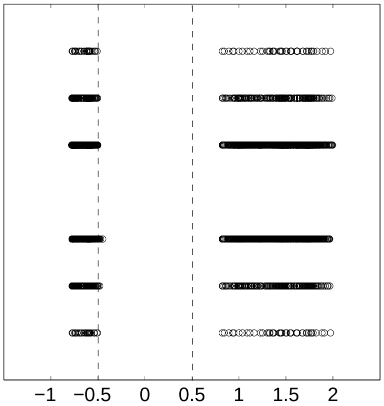

Figure 5.4: Eigenvalues of indefinite preconditioned matrix; P_{ideal} (top), P_{amg} (bottom), $h = \frac{1}{8}, \frac{1}{16}, \frac{1}{32}$

	10^0	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}
θ	0.929	0.868	0.850	0.845	0.845	0.845	0.845
Θ	1	1	1	1	1	1	1

Table 5.12: Values of θ and Θ for varying jump coefficient, $h = \frac{1}{16}$

Example 5

For the final example, we return to the Kellogg problem (see Fig. 3.7) described in Example 6 in Chapter 3. We assemble and solve the test problem on three different types of meshes, generated using the ALBERT toolbox (see [2]). The advantage here is that no modifications need to be made to the AMG algorithm for locally refined meshes. In contrast to the multigrid method described in Chapter 4, complex geometries require no special treatment.

Iteration counts corresponding to discretisations on uniform, graded, as well as locally adapted meshes of triangular elements, (illustrated in Fig. 5.5) are listed in Tables 5.13–5.14. Note that, here, N corresponds to the dimension of the saddle-point system. The minimum and maximum eigenvalues of the multigrid preconditioned operator $V^{-1}BA_{diag}^{-1}B^T$ are listed in Tables 5.16–5.17. We obtain $\tilde{\mu}_1 = \frac{1}{2}$, $\tilde{\mu}_n = \frac{3}{2}$, in all cases. We observe that AMG performance is not influenced by mesh type.

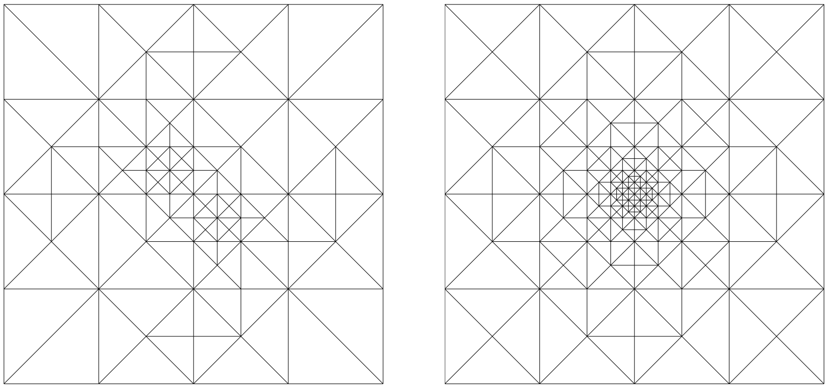

Figure 5.5: Locally adapted (left) and graded (right) meshes

N	$P = I$	$P = P_{ideal}$	$P = P_{amg}$
168	127	17	22 (0.08)
656	268	18	22 (0.09)
2,952	519	18	22 (0.25)
10,304	954	18	22 (1.27)
41,088	1,675	18	23 (6.06)

Table 5.13: MINRES iterations (and time), uniform meshes, Example 5

N	$P = I$	$P = P_{ideal}$	$P = P_{amg}$
184	147	23	25 (0.07)
728	312	24	24 (0.08)
2,896	602	24	25 (0.36)
11,552	1,126	24	25 (2.50)
46,144	2,094	23	25 (8.34)

Table 5.14: MINRES iterations (and time), graded meshes, Example 5

N	$P=I$	$P=P_{ideal}$	$P=P_{amg}$
680	365	23	25 (0.08)
1,910	592	24	25 (0.14)
2,290	673	24	26 (0.26)
3,110	799	26	26 (0.36)
4,844	919	25	26 (0.40)
13,010	1,368	24	26 (1.85)
18,780	1,407	24	26 (3.05)
30,526	1,795	24	26 (6.35)

Table 5.15: MINRES iterations (and time), adapted meshes, Example 5

m	θ	Θ
64	0.8626	1
256	0.8516	1
1024	0.8495	1

Table 5.16: Eigenvalues of $V^{-1}BA_{diag}^{-1}B^T$, uniform meshes, Example 5

m	θ	Θ
72	0.8146	1
288	0.8521	1
1,152	0.8061	1

Table 5.17: Eigenvalues of $V^{-1}BA_{diag}^{-1}B^T$, graded meshes, Example 5

5.5 Concluding remarks

In this chapter, we introduced a second generic approach to preconditioning the model variable diffusion problem (1.4). We used simple algebraic arguments to derive an h-optimal, ideal preconditioner and demonstrated that the diagonal blocks of the matrix actually represent weighted and mesh-dependent versions of a second pair of norms in which the underlying variational problem is stable. To obtain a practical scheme,

m	θ	Θ
128	0.8722	1
268	0.8645	1
756	0.8579	1

Table 5.18: Eigenvalues of $V^{-1}BA_{diag}^{-1}B^T$, adapted meshes, Example 5

we applied a black-box algebraic multigrid method, taking care to implement it symmetrically so as to preserve the minimisation properties of the chosen Krylov subspace solver. In $I\!R^2$, h-optimal results were obtained for uniform and locally-refined meshes. For diagonal coefficient tensors, $\mathcal{A}$-optimality was also achieved.

In Chapters 3–5, we have studied two different h-optimal, parameter-free preconditioning schemes. Both have advantages and disadvantages. On the one hand, the $H(div)$ preconditioner is sensitive to the magnitude of the entries of the coefficient tensor, but deals with diagonal and full tensors equally efficiently. The H^1 preconditioner, on the other hand, is not influenced by the magnitude of the coefficients as long as they are diagonal. However, structure is important because this determines how efficient diagonal scaling will be for the weighted mass matrix. In $I\!R^2$, $\mathcal{A}$-optimality can always be achieved, however, provided that due consideration is paid to the remarks we have made concerning mesh limitations and scaling requirements.

The black-box preconditioning strategy developed in this chapter, is, however, more generic and more widely applicable. Not only can we solve variable diffusion problems on diverse geometries, with no extra computational effort, but we now also have a device for preconditioning saddle-point systems that arise in mixed finite element formulations of other second-order, self-adjoint, elliptic PDES. In fact, whenever mixed formulations give rise to subproblems in which variable diffusion operators or Poisson-like operators are present, AMG is always a good choice for a plug-in solver or preconditioner. We shall explore this in Chapter 7.

The ideal versions of the preconditioners we have suggested are derived from the stability properties of the underlying discrete variational problem. In our philosophy, this is the most natural approach of all to preconditioning the model problem. However, many other authors, in an attempt to avoid solving saddle-point systems, have reformulated the problem as a positive-definite one and have derived preconditioners for those systems, instead. We now consider one of the most popular of these schemes and make a performance comparison with the black-box method we have developed for the indefinite problem in this chapter.

Chapter 6

Lagrange multiplier problem

Recall that the standard mixed variational formulation of (2.8) is,

find $\vec{u}_h \in V_h$, $p_h \in W_h$ satisfying,

$$\begin{aligned} \left(\mathcal{A}^{-1}\vec{u}_h, \vec{v}_h\right) + (p_h, \nabla \cdot \vec{v}_h) &= \langle g, \vec{v}_h \cdot \vec{n}\rangle_{\partial\Omega_D} && \forall\, \vec{v}_h \in V_h, \\ (\nabla \cdot \vec{u}_h, w_h) &= -\,(f, w_h) && \forall\, w_h \in W_h. \end{aligned} \tag{6.1}$$

For a conforming approximation, we require $V_h \subset H_{0,N}(div;\Omega)$. To achieve this, V_h is chosen to contain functions that vanish on the Neumann boundary and whose normal components are continuous across interelement boundaries. Hence, for the lowest order Raviart-Thomas scheme, we construct,

$$V_h = \{\, \vec{v}_h \in\, RT_0(\Omega; T_h) \text{ and } \vec{v}_h \cdot \vec{n} = 0 \ \text{ on } \partial\Omega_N \,\}, \tag{6.2}$$

where,

$$RT_0(\Omega; T_h) := \left\{ \vec{v}_h \in L^2(\Omega)^d \ \mid \ \vec{v}_h \mid_K \in RT_0(K) \ \forall \ K \ \in \ T_h \text{ and } [\vec{v}_h \cdot \vec{n}]_e = 0 \ \forall \, e \, \in \mathcal{E}_I \right\}.$$

Here, $\mathcal{E}_I$ denotes the set of interior interelement boundaries, e, of the triangulation, T_h, and $[\vec{v}_h \cdot \vec{n}]_e$ denotes the jump in the normal component of $\vec{v}_h$ across e with respect to the unit normal vector. In section 2.4.3, we demonstrate that this is simple to achieve. We fix oriented normal vectors $\vec{\nu}_e^{\,i}$ at each boundary e (see Fig. 2.3) and use them to define element basis functions for V_h before assembling the corresponding finite element matrices. As we have seen, the resulting linear system (2.37) is indefinite.

Most numerical analysts avoid solving the indefinite system (2.37). Indeed, Brezzi and Fortin, [20], describe it as a 'considerable source of trouble'. They advocate, instead, an idea attributed to Fraeijs de Veubeke (see [53].) In that approach, the continuity constraint on V_h is relaxed and enforced, implicitly, by imposing extra constraints involving a Lagrange multiplier. This leads to a so-called mixed-hybrid variational problem which can be condensed to a positive definite linear system and solved iteratively using the conjugate gradient method (CG.)

In [79, Ch.7], Quarteroni and Valli compare the computational cost of solving this so-called 'Lagrange multiplier problem', using preconditioned CG, with that of solving the positive-definite Schur complement system (2.51) of the original saddle-point matrix (2.37). Their study suggests that solving the latter problem is cheaper. However, optimal preconditioners for the Lagrange multiplier problem are not considered and the results of that study are, in our opinion, misleading.

In this chapter, we review the main concepts of the mixed-hybrid scheme. For triangular elements, and diagonal coefficient tensors, the system matrix is known to be an M-matrix. We consider black-box AMG preconditioning for the Lagrange multiplier system, in $\mathbb{R}^2$, with triangular *and* rectangular elements. We compare the computational cost of applying AMG to this system with that of applying AMG to the matrix $BA_{diag}^{-1}B^T$, using the black-box preconditioning approach described in Chapter 5.

6.1 A mixed-hybrid method

First, we denote by $\mathcal{E}_I$, $\mathcal{E}_D$ and $\mathcal{E}_N$, the sets of edges lying in the interior of T_h and on the boundaries $\partial\Omega_D$ and $\partial\Omega_N$, respectively. Now, for any $g \in L^2\left(\partial\Omega_D\right)$, we define the space of Lagrange multipliers, $L_{g,D} \subset L^2\left(\mathcal{E}_h\right)$ via,

$$L_{g,D} = \left\{ \mu_h \in P_0(\mathcal{E}_h) \mid \mu_h = \frac{1}{|e|}\int_e g\, ds \quad \forall e \in \mathcal{E}_D \right\}.$$

Recall that $P_0(\mathcal{E}_h)$ denotes the set of piecewise constant functions on the set $\mathcal{E}_h$. To define a mixed-hybrid method, our starting point is the variational problem (6.1). We now seek a velocity approximation, $\vec{u}_h$, in the *discontinuous* Raviart-Thomas space,

$$\tilde{V}_h = \left\{ \vec{v}_h \in L^2\left(\Omega\right)^d \mid \vec{v}_h\,|_K \in RT_0(K) \quad \forall K \in T_h \right\}.$$

Since $\tilde{V}_h \not\subset H(div;\Omega)$, we replace the inner product $(p_h, \nabla \cdot \vec{v}_h)$, for $\vec{v}_h \in \tilde{V}_h$, with,

$$\sum_{K \in T_h} \int_K p_h \nabla \cdot \vec{v}_h \, dK = \sum_{K \in T_h} (p_h, \nabla \cdot \vec{v}_h)_K ,$$

and similarly for the inner-product $(\nabla \cdot \vec{u}_h, w_h)$, with $\vec{u}_h \in \tilde{V}_h$.

An important observation is that any $\vec{v}_h \in \tilde{V}_h$ can be forced to belong to the continuous space V_h, defined in (6.2), by imposing condition (6.3), below.

Lemma 28 *Let $\vec{v}_h \in \tilde{V}_h$, then $\vec{v}_h \in V_h$ if and only if,*

$$\sum_{K \in T_h} \int_{\partial K} \mu_h \vec{v}_h \cdot \vec{n}_K \, ds = 0 \quad \forall \mu_h \in L_{0,D}. \tag{6.3}$$

Proof This is easy to see from the definitions of V_h, $L_{0,D}$, and the fact that,

$$\begin{aligned} \sum_{K \in T_h} \int_{\partial K} \mu_h \vec{v}_h \cdot \vec{n}_K \, ds &= \sum_{e \in \mathcal{E}_D} \int_e \mu_h \vec{v}_h \cdot \vec{n}_e \, ds + \sum_{e \in \mathcal{E}_N} \int_e \mu_h \vec{v}_h \cdot \vec{n}_e \, ds \\ &+ \sum_{e \in \mathcal{E}_I} \int_e \mu_h [\vec{v}_h \cdot \vec{n}_e]_e \, ds. \quad \square \end{aligned}$$

Now it follows that a mixed-hybrid variational formulation of the model variable diffusion problem is,

find $\vec{u}_h \in \tilde{V}_h$, $p_h \in W_h$ and $\lambda_h \in L_{g,D}$ satisfying,

$$\begin{aligned} \left(\mathcal{A}^{-1}\vec{u}_h, \vec{v}_h\right) + \sum_K (p_h, \nabla \cdot \vec{v}_h)_K - \sum_K \langle \lambda_h, \vec{v}_h \cdot \vec{n}_K \rangle_{\partial K} &= 0 \qquad \forall \vec{v}_h \in \tilde{V}_h, \\ \sum_K (\nabla \cdot \vec{u}_h, w_h) &= -(f, w_h) \ \forall w_h \in W_h, \\ \sum_K \langle \vec{u}_h \cdot \vec{n}_K, \mu_h \rangle_{\partial K} &= 0 \qquad \forall \mu_h \in L_{0,D}. \end{aligned} \tag{6.4}$$

The third equation of (6.4) imposes condition (6.3), so we actually obtain a velocity solution $\vec{u}_h \in V_h$. Further, it can be shown (see Brezzi and Fortin, [26, pp.178–180]) that the unique solutions $\vec{u}_h$ and p_h to (6.4) coincide with the unique solutions to (6.1). The form of the first equation suggests that the Lagrange multiplier solution, λ_h, is an approximation to the trace of the pressure solution, p_h, at interelement boundaries. Arnold and Brezzi studied this in [8] and demonstrated that an improved approximation to the pressure unknown can be achieved by combining λ_h and p_h from (6.4) in a post-processing step.

The resulting linear system has the block form,

$$\begin{pmatrix} A & B^T & C^T \\ B & 0 & 0 \\ C & 0 & 0 \end{pmatrix} \begin{pmatrix} \underline{u} \\ \underline{p} \\ \underline{\lambda} \end{pmatrix} = \begin{pmatrix} \underline{g} \\ \underline{f} \\ \underline{0} \end{pmatrix}, \tag{6.5}$$

and, like the original system (2.37), is symmetric and indefinite. Here,

$$\begin{aligned} A_{ij} &= \left(\mathcal{A}^{-1}\vec{\chi_i}, \vec{\chi_j}\right), & i, j = 1 : n^*, \quad (6.6)\\ B_{kj} &= \sum_K \left(\nabla \cdot \vec{\chi_j}, \phi_k\right)_K, & k = 1 : m,\ j = 1 : n^*, \quad (6.7)\\ C_{kj} &= -\sum_K \left\langle \vec{\chi_j} \cdot \vec{n}_K, \psi_k\right\rangle_{\partial K}, & k = 1 : l,\ j = 1 : n^*, \quad (6.8) \end{aligned}$$

where $\{\vec{\chi_i}\}_{i=1}^{n^*}$, $\{\phi_k\}_{k=1}^{m}$, and $\{\psi_k\}_{k=1}^{l}$ denote standard basis sets for $\tilde{V}_h$, W_h and $L_{0,D}$, respectively. Recall that m is the number of elements, and the row dimension, l, of C, is the number of interior edges. Since no continuity is imposed on $\tilde{V}_h$, the leading matrix, A, is now block-diagonal with m blocks of dimension 3 × 3, for triangles, or 4 × 4, for rectangles.

To arrive at a positive definite problem for the discrete Lagrange multiplier solution, $\underline{\lambda}$, we may proceed in two ways. The naive approach is to construct (6.5). Then, since A is block-diagonal, and can be inverted relatively cheaply, the discrete velocity solution $\underline{u}$, can be eliminated, yielding,

$$\begin{aligned} -BA^{-1}B^T\underline{p} - BA^{-1}C^T\underline{\lambda} &= \underline{f} - BA^{-1}\underline{g}, \\ -CA^{-1}B^T\underline{p} - CA^{-1}C^T\underline{\lambda} &= -CA^{-1}\underline{g}. \end{aligned}$$

For the lowest-order Raviart-Thomas schemes in $\mathbb{R}^2$, (and in $\mathbb{R}^3$), $BA^{-1}B^T$ is a diagonal matrix here (see Brezzi and Fortin [26, pp.184-185].) Hence $\underline{p}$ can also be eliminated cheaply, yielding, finally, a symmetric, positive definite system,

$$L\underline{\lambda} = \underline{F}, \tag{6.9}$$

of dimension l, where,

$$\begin{aligned} L &= CA^{-1}B^T\left(BA^{-1}B^T\right)^{-1}BA^{-1}C^T - CA^{-1}C^T, \\ \underline{F} &= CA^{-1}B^T\left(BA^{-1}B^T\right)^{-1}\left(BA^{-1}\underline{g} - \underline{f}\right) - CA^{-1}\underline{g}. \end{aligned}$$

Alternatively, we can construct L and $\underline{F}$ in (6.9) directly. This can be achieved with element by element calculations, by making special choices of test functions in (6.4). Such a process is described by Brezzi et al., [28], [27], for triangular elements and scalar coefficient functions in $I\!R^2$. In [36], Chen demonstrates that this so-called algebraic condensation can be extended to tetrahedral and brick elements. Further, general coefficient tensors can be treated in the same framework if we approximate $\mathcal{A}$ in (6.4) with a piecewise constant function. More details of this will be given later. On the other hand, (6.9) can also be constructed directly by observing that it is equivalent to the algebraic systems arising in certain other variational formulations of our model problem.

6.2 Equivalence results

Analysis of the algebraic system (6.9), reveals that the Lagrange multiplier solution, λ_h, generated by the mixed-hybrid method, is equivalent to certain non-conforming Galerkin approximations of the primal problem,

find p satisfying,

$$\begin{aligned} -\nabla \cdot \mathcal{A} \nabla p &= f \quad \text{in } \Omega, \\ p &= g \quad \text{on } \partial\Omega_D, \\ \mathcal{A}\nabla p \cdot \vec{n} &= 0 \quad \text{on } \partial\Omega_N. \end{aligned} \tag{6.10}$$

Such equivalence results are the starting point for most solution methods for (6.9).

Arnold and Brezzi, [8], establish a variational equivalence for (6.4) in $I\!R^2$, using triangles and a homogeneous Dirichlet boundary condition. To describe it, we must define the spaces,

$$\begin{aligned} S_h^k &= \left\{ w \in L^2(\Omega) \mid w\,|_K \in P_k(K) \quad \forall K \in T_h \right\}, \\ X_h &= \left\{ w \in S_h^1 \mid \; w \text{ is continuous at the midpoints of } e \in \mathcal{E}_I, w\,|_e = 0 \quad \forall e \in \mathcal{E}_D \right\}, \\ B_h &= \left\{ w \in S_h^3 \cap H_0^1(\Omega) \mid w\,|_e = 0 \, \forall e \in \mathcal{E}_h \right\}. \end{aligned}$$

Thus, $N_h = X_h + B_h$ is the space of midside continuous, piecewise linear functions, augmented by cubic bubble functions. In addition, we require the L^2-projection operators,

$P_h : N_h \to W_h$ and $\Pi_h : N_h \to L_{0,D}$, defined via,

$$(P_h \varphi_h, w_h) \; = \; (\varphi_h, w_h) \quad \forall\, \varphi_h \in N_h, \forall\, w_h \in W_h, \tag{6.11}$$

$$(\Pi_h \varphi_h, \mu_h) \; = \; (\varphi_h, \mu_h) \quad \forall\, \varphi_h \in N_h, \forall\, \mu_h \in L_{0,D}, \tag{6.12}$$

and the weighted L^2-projection operator $P_{RT,\mathcal{A}^{-1}} : \nabla N_h \to \tilde{V}_h$, defined via,

$$\left(\mathcal{A}^{-1} \left(P_{RT,\mathcal{A}^{-1}} \vec{\tau}_h\right), \vec{v}_h\right) \; = \; \left(\mathcal{A}^{-1} \vec{\tau}_h, \vec{v}_h\right) \quad \forall\, \vec{\tau}_h \in \nabla N_h, \forall\, \vec{v}_h \in \tilde{V}_h. \tag{6.13}$$

Arnold and Brezzi's classical result is summarised in the following lemma.

Lemma 29 *Let $(\vec{u_h}, p_h, \lambda_h) \in \tilde{V}_h \times W_h \times L_{0,D}$ be the solution of the mixed-hybrid variational problem (6.4), discretised using lowest-order, triangular, Raviart-Thomas element. Let $\varphi_h \in N_h$ be defined via,*

$$P_h \varphi_h = p_h, \quad \Pi_h \varphi_h = \lambda_h. \tag{6.14}$$

Then, φ_h is the unique solution to: find $\varphi_h \in N_h$ such that,

$$\sum_K \int_K \left(P_{RT,\mathcal{A}^{-1}} \left(\mathcal{A} \nabla \varphi_h\right), \nabla \chi_h\right) \; = \; (P_h f, \chi_h) \quad \forall\, \chi_h \in N_h. \tag{6.15}$$

Proof See [8, pp.25-31]. □

The reduced problem (6.15) is symmetric and positive definite and can therefore be solved with CG. However, the condition number of the corresponding system matrix grows like h^{-2} and preconditioners are required. A suitable h-optimal multigrid method is described by Brenner in [21] but the impact of the coefficient tensor is not discussed.

In [36], [37] and [38], Chen et al. study a slightly modified version of the mixed-hybrid problem (6.4). Again, a homogeneous Dirichlet boundary condition is assumed. By replacing the inverse of the coefficient tensor, $\mathcal{A}^{-1}$, by its L^2-projection, $P_h \mathcal{A}^{-1}$, onto W_h, the space of piecewise constant functions, algebraic equivalence results for the linear system (6.9) can be established without employing bubble functions. The corresponding result for triangles, in $I\!R^2$, is summarised below.

Lemma 30 *For lowest-order, triangular, Raviart-Thomas elements, the linear system (6.9) associated with (6.4), with $\mathcal{A}^{-1}$ approximated by $P_h \mathcal{A}^{-1}$, is equivalent to the linear system in the problem: find $\varphi_h \in X_h$ such that,*

$$\sum_K \int_K \left(\left(P_h \mathcal{A}^{-1}\right)^{-1} \nabla \varphi_h, \nabla \chi_h\right) \; = \; (P_h f, \chi_h) \quad \forall\, \chi_h \in X_h. \tag{6.16}$$

Proof See Chen, [36]. □

Hence, the advantage is that (6.16) is more easily tackled with multigrid. For triangles, convergence of the W-cycle is proved by Brenner, [20], and Braess and Verfürth, [19]. Alternative V-cycle and W-cycle multigrid schemes are proposed in [36] and [65]. Domain decomposition methods are suggested in [37]. Although mesh independent convergence is achieved in all of these methods, it is not clear how the coefficient tensor influences their efficiency. In fact, the existence of $\mathcal{A}$-optimal preconditioners for (6.9) appears to be an open question.

An attempt was made to address this important issue, for diagonal, anisotropic coefficient tensors, in $I\!R^3$, by Maliassov, [66], and Chen et al., [38]. In those studies, an ideal preconditioner for L in (6.9) is constructed by assembling parameterised versions of the local stiffness matrices defined on prism-shaped and tetrahedral subdomains. Although the condition number of the preconditioned system is shown to be h-optimal, it deteriorates severely if the direction of the anisotropy is not aligned with the subdomains.

6.3 System assembly

Before demonstrating the efficiency of black-box AMG preconditioning for (6.9), in $I\!R^2$, we outline the algebraic condensation method of Brezzi et al., [28], and Chen, [36], for assembling (6.9) and recovering discrete approximations to the original unknowns $\vec{u}$ and p. We stick to the notation conventions used in [36]. Hence, we restate the mixed-hybrid problem as,

find $\vec{u}_h \in \tilde{V}_h$, $p_h \in W_h$ and $\lambda_h \in L_{g,D}$ satisfying,

$$\begin{aligned}
\left(\mathcal{A}^{-1}\vec{u}_h, \vec{v}_h\right) - \sum_K (p_h, \nabla\cdot\vec{v}_h)_K \quad + \quad \sum_K \langle \lambda_h, \vec{v}_h\cdot\vec{n}_K\rangle_{\partial K} &= 0 \qquad \forall\, \vec{v}_h \in \tilde{V}_h, \\
\sum_K (\nabla\cdot\vec{u}_h, w_h) &= (f, w_h)\ \forall\, w_h \in W_h, \quad (6.17) \\
\sum_K \langle \vec{u}_h\cdot\vec{n}_K, \mu_h\rangle_{\partial K} &= 0 \qquad \forall\, \mu_h \in L_{0,D}.
\end{aligned}$$

Note this is the same as (6.4) with $\vec{u}_h$ replaced by $-\vec{u}_h$.

The derivation of (6.9) for triangles and general coefficient tensors can be found in [36], [38] or [66]. For rectangles, the derivation is performed by Chen in [36], for scalar coefficient functions only. Below, we give full details for rectangles and diagonal coefficient tensors. This includes discontinuous and anisotropic coefficients. For non-diagonal coefficient tensors, the method is the same but the algebra is 'messy'.

6.3.1 Rectangles

Recall that, for rectangular elements, $\vec{u}_h \in \tilde{V}_h$ and $p_h \in W_h$ have the local definitions,

$$\vec{u}_h \mid_K = \begin{pmatrix} a^K + b^K x \\ c^K + d^K y \end{pmatrix}, \quad p_h \mid_K = p^K.$$

Hence the unknowns, $\vec{u}_h$ and p_h, in (6.4), are fully defined by the set of constants a^K, b^K, c^K, d^K and p^K in each rectangle, K.

Let $\mathcal{A} = \mathtt{diag}(a_{11}(x,y), a_{22}(x,y))$ be any given diagonal coefficient tensor and let $P_h\mathcal{A}^{-1}$ denote the L^2-projection of the inverse onto the set of piecewise constant functions. In each rectangle K, we write,

$$P_h\mathcal{A}^{-1}|_K = \begin{pmatrix} \alpha_{11}^K & 0 \\ 0 & \alpha_{22}^K \end{pmatrix}.$$

Computationally, this means that each entry of $\mathcal{A}^{-1}$ is approximated by a constant.

Our starting point is the equations (6.17). Setting $w_h = 1$ in rectangle K and zero elsewhere, in the second equation, yields,

$$b^K + d^K = \frac{1}{\mid K \mid}\int_\Omega f \, d\Omega = f^K, \tag{6.18}$$

and hence, $b^K = f^K - d^K$. Choosing the test function $\vec{v}_h$ in the first equation of (6.4) to be $\vec{v}_h = (1,0)^T$ or $\vec{v}_h = (0,1)^T$ in rectangle K, and zero elsewhere yields,

$$a^K \int_K \alpha_{11}^K \, dK = -b^K \int_K \alpha_{11}^K x \, dK - \sum_{i=1}^4 \lambda_h \mid_{e_i^K} \vec{n}_K^{i(x)} \mid e_K^i \mid, \tag{6.19}$$

$$c^K \int_K \alpha_{22}^K \, dK = -d^K \int_K \alpha_{22}^K y \, dK - \sum_{i=1}^4 \lambda_h \mid_{e_i^K} \vec{n}_K^{i(y)} \mid e_K^i \mid. \tag{6.20}$$

Here, $\lambda_h \mid_{e_i^K}$ denotes the value of λ_h at the ith edge of rectangle K. $\vec{n}_K^i$ is the unit outward normal vector to that edge, and $\mid e_K^i \mid$ is the edge length.

To evaluate the integrals in (6.19) and (6.20), a simple calculation shows that, for a rectangle of dimension $h_x \times h_y$, with edges aligned with the co-ordinate axes, we obtain,

$$\int_K x \, dK = x_c h_x h_y = x_c \mid K \mid, \qquad \int_K y \, dK = y_c h_x h_y = y_c \mid K \mid,$$

where (x_c, y_c) denotes the co-ordinates of the centroid of the element. Substituting for b^K from (6.18) in (6.19) and (6.20) and writing,

$$C^K = \begin{pmatrix} \int_K \alpha_{11}^K \, dK & 0 \\ 0 & \int_K \alpha_{22}^K \, dK \end{pmatrix}^{-1} = \begin{pmatrix} C_{11}^K & 0 \\ 0 & C_{22}^K \end{pmatrix}, \tag{6.21}$$

yields,

$$a^K \;=\; x_c d^K - x_c f^K - C_{11}^K \sum_{i=1}^{4} \lambda_h \,|_{e_i^K} \, \vec{n}_K^{i(x)} \mid e_K^i \mid, \tag{6.22}$$

$$c^K \;=\; -y_c d^K - C_{22}^K \sum_{i=1}^{4} \lambda_h \,|_{e_i^K} \, \vec{n}_K^{i(y)} \mid e_K^i \mid. \tag{6.23}$$

In the same way, choosing $\vec{v}_h = (x,0)^T$ and $\vec{v}_h = (0,y)^T$ in element K and zero elsewhere, yields the pair of equations,

$$a^K \int_K \alpha_{11}^K x \, dK \;=\; -b^K \int_K \alpha_{11}^K x^2 \, dK - p^K \mid K \mid - \sum_{i=1}^{4} \lambda_h \,|_{e_i^K} \, \vec{n}_K^{i(x)} \int_{\partial e_i^K} x \, ds,$$

$$c^K \int_K \alpha_{22}^K y \, dK \;=\; -d^K \int_K \alpha_{22}^K y^2 \, dK - p^K \mid K \mid - \sum_{i=1}^{4} \lambda_h \,|_{e_i^K} \, \vec{n}_K^{i(y)} \int_{\partial e_i^K} y \, ds.$$

To evaluate the integrals, we calculate that,

$$\int_K x^2 \, dK = \left(x_c^2 h_x + \frac{1}{12} h_x^3 \right) h_y, \quad \int_K y^2 \, dK = \left(y_c^2 h_y + \frac{1}{12} h_y^3 \right) h_x.$$

Using these expressions and substituting for b^K from (6.18), yields,

$$a^K \;=\; -\frac{d^K - f^K}{x_c} \left(x_c^2 + \frac{h_x^2}{12} \right) - \frac{p^K C_{11}^K \mid K \mid}{x_c} - \frac{C_{11}^K}{x_c} \sum_{i=1}^{4} \lambda_h \,|_{e_i^K} \, \vec{n}_K^{i(x)} \int_{\partial e_i^K} x \, ds, \tag{6.24}$$

$$c^K \;=\; -\frac{d^K}{y_c} \left(y_c^2 + \frac{h_y^2}{12} \right) - \frac{p^K C_{22}^K \mid K \mid}{y_c} - \frac{C_{22}^K}{y_c} \sum_{i=1}^{4} \lambda_h \,|_{e_i^K} \, \vec{n}_K^{i(y)} \int_{\partial e_i^K} y \, ds. \tag{6.25}$$

Thus, combining (6.18), (6.19)–(6.20) and (6.24)–(6.25), we have five equations in five unknowns. Using the last two, we can eliminate p^K and substitute for a^K and c^K from (6.19)–(6.20) to obtain an explicit expression for d_K. We omit the details. Defining,

$$R_{\mathcal{A}} = h_x^2 \int_K \alpha_{11}^K \, dK + h_y^2 \int_K \alpha_{22}^K \, dK,$$

we obtain, after a lot of simplification (due to the particularly simple form of the vectors $\vec{n}_K^1, \vec{n}_K^2, \vec{n}_K^3, \vec{n}_K^4$ for rectangles),

$$d^K \;=\; \frac{h_x^2 f^K \int_K \alpha_{11}^K}{R_{\mathcal{A}}} + \frac{6 \mid K \mid}{R_{\mathcal{A}}} \sum_{i=1}^{4} \left(\mid \vec{n}_K^{i(x)} \mid - \mid \vec{n}_K^{i(y)} \mid \right) \lambda_h \,|_{e_i^K} \,.$$

Then, substituting for d^K in the remaining equations, we obtain the following explicit expressions for the other unknowns,

$$\begin{aligned}
a^K &= -\frac{x_c h_y^2 f^K \int_K \alpha_{22}^K}{R_{\mathcal{A}}} + \mid K \mid \sum_{i=1}^{4} \left(\frac{6x_c}{R_{\mathcal{A}}} \left(\mid \vec{n}_K^{i(x)} \mid - \mid \vec{n}_K^{i(y)} \mid \right) - \frac{\vec{n}_K^{i(x)}}{h_x \int_K \alpha_{11}^K} \right) \lambda_h \mid_{e_i^K}, \\
b^K &= \frac{h_y^2 f^K \int_K \alpha_{22}^K}{R_{\mathcal{A}}} + \frac{6 \mid K \mid}{R_{\mathcal{A}}} \sum_{i=1}^{4} \left(- \mid \vec{n}_K^{i(x)} \mid + \mid \vec{n}_K^{i(y)} \mid \right) \lambda_h \mid_{e_i^K}, \\
c^K &= -\frac{y_c h_x^2 f^K \int_K \alpha_{11}^K}{R_{\mathcal{A}}} + \mid K \mid \sum_{i=1}^{4} \left(\frac{6y_c}{R_{\mathcal{A}}} \left(- \mid \vec{n}_K^{i(x)} \mid + \mid \vec{n}_K^{i(y)} \mid \right) - \frac{\vec{n}_K^{i(y)}}{h_y \int_K \alpha_{22}^K} \right) \lambda_h \mid_{e_i^K}, \\
p^K &= \frac{\mid K \mid f^K \int_K \alpha_{11}^K \int_K \alpha_{22}^K}{12 R_{\mathcal{A}}} + \frac{1}{2R_{\mathcal{A}}} \sum_{i=1}^{4} \left(h_y^2 \mid \vec{n}_K^{i(x)} \mid \int_K \alpha_{22}^K + h_x^2 \mid \vec{n}_K^{i(y)} \mid \int_K \alpha_{11}^K \right) \lambda_h \mid_{e_i^K}.
\end{aligned}$$

For the case $\alpha_{11}^K = \alpha_{22}^K$, these equations reduce to the expressions given in [36].

Once a discrete approximation to λ_h is obtained, the original pressure and velocity approximations can be recovered by computing a^K, b^K, c^K, d^K and p^K in each element. To obtain (6.9), we study the third equation of (6.17). For a specific edge of a given rectangle K, we choose the test function μ_h that takes value one at that edge and zero at all other edges. Repeating this for all four edges of the chosen rectangle yields the system of equations,

$$\mid e_K^i \mid \vec{n}_K^{i(x)} a^K + \mid e_K^i \mid \vec{n}_K^{i(y)} c^K + b^K \vec{n}_K^{i(x)} \left(\int_{\partial e_i^K} x \, ds \right) + d^K \vec{n}_K^{i(y)} \left(\int_{\partial e_i^K} y \, ds \right) = 0, \quad i = 1:4.$$

Substituting for a^K, b^K, c^K and d^K yields, for $i = 1:4$,

$$\begin{aligned}
& \mid e_K^i \mid \vec{n}_K^{i(x)} \mid K \mid \sum_{j=1}^{4} \left\{ \frac{6x_c}{R_{\mathcal{A}}} \left(\mid \vec{n}_K^{j(x)} \mid - \mid \vec{n}_K^{j(y)} \mid \right) - \frac{\vec{n}_K^{j(x)}}{h_x \int_K \alpha_{11}^K} \right\} \lambda_h \mid_{e_j^K} \\
+ \; & \mid e_K^i \mid \vec{n}_K^{i(y)} \mid K \mid \sum_{j=1}^{4} \left\{ \frac{6y_c}{R_{\mathcal{A}}} \left(\mid \vec{n}_K^{j(x)} \mid - \mid \vec{n}_K^{j(y)} \mid \right) - \frac{\vec{n}_K^{j(y)}}{h_y \int_K \alpha_{22}^K} \right\} \lambda_h \mid_{e_j^K} \\
- \; & \vec{n}_K^{i(x)} \left(\int_{\partial e_i^K} x \, ds \right) \frac{6 \mid K \mid}{R_{\mathcal{A}}} \sum_{j=1}^{4} \left(\mid \vec{n}_K^{j(x)} \mid - \mid \vec{n}_K^{j(y)} \mid \right) \lambda_h \mid_{e_j^K} \\
+ \; & \vec{n}_K^{i(y)} \left(\int_{\partial e_i^K} y \, ds \right) \frac{6 \mid K \mid}{R_{\mathcal{A}}} \sum_{j=1}^{4} \left(\mid \vec{n}_K^{j(x)} \mid - \mid \vec{n}_K^{j(y)} \mid \right) \lambda_h \mid_{e_j^K} \\
= \; & \frac{f^K \mid e_K^i \mid}{R_{\mathcal{A}}} \left\{ x_c h_y^2 \vec{n}_K^{i(x)} \left(\int_K \alpha_{22}^K \, dK \right) + y_c h_x^2 \vec{n}_K^{i(y)} \left(\int_K \alpha_{11}^K \, dK \right) \right\} \\
- \; & \frac{f^K}{R_{\mathcal{A}}} \left\{ h_y^2 \vec{n}_K^{i(x)} \left(\int_K \alpha_{22}^K \, dK \right) \left(\int_{\partial e_i^K} x \, ds \right) + h_x^2 \vec{n}_K^{i(y)} \left(\int_K \alpha_{11}^K \, dK \right) \left(\int_{\partial e_i^K} y \, ds \right) \right\}.
\end{aligned}$$

Using the mid-point rule to evaluate the edge integrals and labelling the edges aligned with the x-axis e_1^K and e_2^K, and the edges aligned with the y-axis, e_3^K and e_4^K, we obtain,

$$\begin{aligned}
\int_{\partial e_1^K} x\,ds &= x_c h_x, & \int_{\partial e_1^K} y\,ds &= \left(y_c - \frac{h_y}{2}\right) h_x,\\
\int_{\partial e_2^K} x\,ds &= x_c h_x, & \int_{\partial e_2^K} y\,ds &= \left(y_c + \frac{h_y}{2}\right) h_x,\\
\int_{\partial e_3^K} x\,ds &= \left(x_c - \frac{h_x}{2}\right) h_y, & \int_{\partial e_3^K} y\,ds &= y_c h_y,\\
\int_{\partial e_4^K} x\,ds &= \left(x_c + \frac{h_x}{2}\right) h_y, & \int_{\partial e_4^K} y\,ds &= y_c h_y.
\end{aligned}$$

Making these substitutions, on the left-hand side, we finally obtain a 4×4 local problem on the chosen K, of the form,

$$L^K \underline{\lambda}^K = \underline{F}^K, \tag{6.26}$$

with $\underline{\lambda}_i^K = \lambda_h \mid_{e_i^K}$, for $i = 1:4$, and,

$$\begin{aligned}
L_{ij}^K &= \mid e_i^K \mid\mid e_j^K \mid \left(C^K \vec{n}_K^i\right) \cdot \vec{n}_K^j + \frac{3 \mid K \mid^2}{R_{\mathcal{A}}} \left(\mid \vec{n}_K^{i(x)} \mid - \mid \vec{n}_K^{i(y)} \mid\right) \left(\mid \vec{n}_K^{j(x)} \mid - \mid \vec{n}_K^{j(y)} \mid\right),\\
\underline{F}_i^K &= \frac{f^K \mid e_K^i \mid}{R_{\mathcal{A}}} \left(x_c h_y^2 \vec{n}_K^{i(x)} \int_K \alpha_{22}^K\,dK + y_c h_x^2 \vec{n}_K^{i(y)} \int_K \alpha_{11}^K\,dK\right)\\
&- \frac{f^K}{R_{\mathcal{A}}} \left\{ h_y^2 \vec{n}_K^{i(x)} \left(\int_K \alpha_{22}^K\,dK\right) \left(\int_{\partial e_i^K} x\,ds\right) + h_x^2 \vec{n}_K^{i(y)} \left(\int_K \alpha_{11}^K\,dK\right) \left(\int_{\partial e_i^K} y\,ds\right) \right\}.
\end{aligned}$$

The global system (6.9) is assembled from these element contributions.

6.3.2 Triangles

For triangles, the derivation of (6.9) is simpler because the velocity solution $\vec{u}_h \in \tilde{V}_h$ has only three degrees of freedom per element. That is, for any triangle K,

$$\vec{u}_h \mid_K = \begin{pmatrix} a^K + b^K x \\ c^K + b^K y \end{pmatrix}.$$

Given a general, symmetric, coefficient tensor,

$$\mathcal{A} = \begin{pmatrix} a_{11}(x,y) & a_{12}(x,y) \\ a_{12}(x,y) & a_{22}(x,y) \end{pmatrix},$$

we denote,

$$P_h \mathcal{A}^{-1} \mid_K = \begin{pmatrix} \alpha_{11}^K & \alpha_{12}^K \\ \alpha_{12}^K & \alpha_{22}^K \end{pmatrix}, \quad C^K = \begin{pmatrix} \int_K \alpha_{11}^K\,dK & \int_K \alpha_{12}^K\,dK \\ \int_K \alpha_{21}^K\,dK & \int_K \alpha_{22}^K\,dK \end{pmatrix}^{-1}. \tag{6.27}$$

Since we have already illustrated the method of derivation, we simply state the result of Chen, [36]. We obtain 3×3 local problems of the form,

$$L^K \underline{\lambda}^K = \underline{F}^K, \tag{6.28}$$

with $\underline{\lambda}_i^K = \lambda_h \mid_{e_i^K}$, for $i = 1:3$, and,

$$\begin{aligned} L_{i,j}^K &= \mid e_i^K \mid\mid e_j^K \mid \left(\vec{n}_K^i\right)^T C^K \vec{n}_K^j, \qquad i,j = 1:3, \\ \underline{F}_i^K &= -\frac{f^K}{2 \mid K \mid} \int_K x \mid e_K^i \mid \left(\vec{n}_K^{i(x)} + y\, \vec{n}_K^{i(y)}\right) dK + \frac{f^K}{2} \int_{\partial e_i^K} \left(x\, \vec{n}_K^{i(x)} + y\, \vec{n}_K^{i(y)}\right) ds. \end{aligned}$$

For both rectangles and triangles, recovering the original velocity and pressure unknowns requires the inversion of the 2×2 matrix, C^K, in (6.27), in each of the m elements. Thus, in addition to the cost of solving the Lagrange multiplier system (6.9) for $\underline{\lambda}$, this approach incurs an extra cost of inverting a block-diagonal matrix with m diagonal blocks of dimension two.

6.4 Structure of element matrices

In [36], Chen claims that the Lagrange multiplier system matrix, L, is always an M-matrix for triangles. Recall, then, that M-matrices are diagonally dominant matrices, characterised by positive diagonal entries and negative off-diagonal entries. For scalar coefficient functions, and diagonal coefficient tensors with positive diagonal entries, the sign of the off-diagonal entries of L^K in (6.28) depends on the cosine of the angle, θ, between two normal vectors to two edges of the triangle. The sign is negative if $\theta \leq \frac{\pi}{2}$. However, for general coefficient tensors, the same argument does *not* apply. Unfortunately, we have observed that the M-matrix property does not hold, in general, for non-diagonal coefficient tensors.

To illustrate this, consider right-angled triangles with $\vec{n}_K^1 = (0,-1)^T$, $\vec{n}_K^2 = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})^T$, $\vec{n}_K^3 = (-1,0)^T$, and $\mid e_1^K \mid = h$, $\mid e_2^K \mid = \sqrt{2}h$, $\mid e_3^K \mid = h$. The element matrix L^K in (6.28) is then,

$$L^K = \frac{2}{det} \begin{pmatrix} \alpha_{11}^K & \alpha_{12}^K - \alpha_{11}^K & -\alpha_{12}^K \\ \alpha_{12}^K - \alpha_{11}^K & \alpha_{11}^K + \alpha_{22}^K - 2\alpha_{12}^K & \alpha_{12}^K - \alpha_{22}^K \\ -\alpha_{12}^K & \alpha_{12}^K - \alpha_{22}^K & \alpha_{22}^K \end{pmatrix}, \tag{6.29}$$

where $det = \alpha_{11}^K \alpha_{22}^K - \left(\alpha_{12}^K\right)^2$. Suppose the coefficient tensor is non-diagonal with constant entries,

$$\mathcal{A}\mid_K = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix},$$

Rewriting L^K in terms of the original coefficients, we obtain,

$$L^K = 2\begin{pmatrix} a_{22} & -a_{12}-a_{22} & a_{12} \\ -a_{12}-a_{22} & a_{11}+a_{22}+2a_{12} & -a_{12}-a_{11} \\ a_{12} & -a_{12}-a_{11} & a_{11} \end{pmatrix}, \tag{6.30}$$

which has positive off-diagonal entries whenever $a_{12} > 0$. L^K is not an M-matrix and neither is the global matrix L. If, on the other hand, $a_{12} = 0$, we obtain,

$$L^K = 2\begin{pmatrix} a_{22} & -a_{22} & 0 \\ -a_{22} & a_{11}+a_{22} & -a_{11} \\ 0 & -a_{11} & a_{11} \end{pmatrix}, \tag{6.31}$$

which is always an M-matrix for positive a_{11} and a_{22}. In view of our discussion in Chapter 5, these observations have consequences for black-box AMG preconditioning.

For rectangles, the M-matrix property almost never holds, even for diagonal coefficient tensors. Consider rectangular elements with edges aligned with the co-ordinate axes, as shown in Fig. 6.1, with $\mid e_1^K \mid = \mid e_2^K \mid = h_x$, and $\mid e_3^K \mid = \mid e_4^K \mid = h_y$.

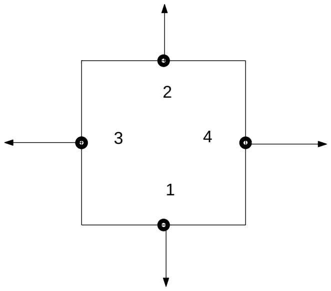

Figure 6.1: Rectangular element

Since the four unit outward normal vectors are,

$$\vec{n}_K^1 = \begin{pmatrix} 0 \\ -1 \end{pmatrix}, \quad \vec{n}_K^2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \vec{n}_K^3 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \vec{n}_K^4 = \begin{pmatrix} 1 \\ 0 \end{pmatrix},$$

the element matrix, L^K, in (6.26), for diagonal coefficient tensors, is,

$$L^K = \begin{pmatrix} \frac{h_x}{\alpha_{22}^K h_y} + \frac{3h_x h_y}{R_\mathcal{A}} & -\frac{h_x}{\alpha_{22}^K h_y} + \frac{3h_x h_y}{R_\mathcal{A}} & -\frac{3h_x h_y}{R_\mathcal{A}} & -\frac{3h_x h_y}{R_\mathcal{A}} \\ -\frac{h_x}{\alpha_{22}^K h_y} + \frac{3h_x h_y}{R_\mathcal{A}} & \frac{h_x}{\alpha_{22}^K h_y} + \frac{3h_x h_y}{R_\mathcal{A}} & -\frac{3h_x h_y}{R_\mathcal{A}} & -\frac{3h_x h_y}{R_\mathcal{A}} \\ -\frac{3h_x h_y}{R_\mathcal{A}} & -\frac{3h_x h_y}{R_\mathcal{A}} & \frac{h_y}{\alpha_{11}^K h_x} + \frac{3h_x h_y}{R_\mathcal{A}} & -\frac{h_y}{\alpha_{11}^K h_x} + \frac{3h_x h_y}{R_\mathcal{A}} \\ -\frac{3h_x h_y}{R_\mathcal{A}} & -\frac{3h_x h_y}{R_\mathcal{A}} & -\frac{h_y}{\alpha_{11}^K h_x} + \frac{3h_x h_y}{R_\mathcal{A}} & \frac{h_y}{\alpha_{11}^K h_x} + \frac{3h_x h_y}{R_\mathcal{A}} \end{pmatrix},$$

where, here, we redefine, $R_\mathcal{A} = \alpha_{11}^K h_x^2 + \alpha_{22}^K h_y^2$. Now, L^K has a 2×2 block structure. The entries in the off-diagonal blocks always have negative signs, since we assume that α_{11}^K and α_{22}^K are positive. This property is due to the difference in the signs of the normal vectors and is independent of the coefficients. In the trivial case $h_x = h_y$ and $\mathcal{A} = \mathcal{I}$, we obtain,

$$\begin{pmatrix} \frac{5}{2} & \frac{1}{2} & -\frac{3}{2} & -\frac{3}{2} \\ \frac{1}{2} & \frac{5}{2} & -\frac{3}{2} & -\frac{3}{2} \\ -\frac{3}{2} & -\frac{3}{2} & \frac{5}{2} & \frac{1}{2} \\ -\frac{3}{2} & -\frac{3}{2} & \frac{1}{2} & \frac{5}{2} \end{pmatrix}. \qquad (6.32)$$

Even here, the global matrix L is not an M-matrix.

6.5 AMG preconditioned CG

In Chapter 5, we introduced algebraic multigrid and explained how its standard components are tailored to M-matrices. For instance, in (5.34), in the definition of strong connections, it is implicitly assumed that off-diagonal entries of the matrix are negative. Positive connections are ignored. Thus, if large positive entries exist, the interpolation scheme we described can be very inefficient. Further, the interpolation weights may be negative or undefined. Of course, all of the components of AMG can be modified to cater for other types of matrix but this requires tuning. Recall that the matrix, $BA_{diag}^{-1}B^T$, to which we applied AMG in Chapter 5, is always an M-matrix. Standard black-box codes will always be robust for that problem.

We are now interested in the efficiency of standard black-box AMG, for solving the Lagrange multiplier problem, $L\underline{\lambda} = \underline{F}$. For triangular elements and diagonal coefficients, at least, we are assured that L is an M-matrix. For rectangles, if positive off-diagonal entries occur, but are small in magnitude, relative to the other entries, then we can still expect standard AMG to work well, particularly if we apply it as a preconditioner in conjunction with a Krylov subspace solver.

We use CG to solve (6.9), applying one V-cycle of the code `amg1r5`, with symmetric smoothing, as a preconditioner. The test problem we consider is (1.4), discretised on $\Omega = [0,1] \times [0,1]$, with $g = 0$. Iteration counts are reported below for a range of diagonal coefficients with uniform meshes of triangles and squares. The time units reported in parentheses correspond to the average elapsed time in seconds for the total solve with a stopping tolerance of 10^{-6} on the relative residual error.

Example 1

We begin with unit coefficients. Iteration counts for triangular and square elements are reported in Tables 6.1 and 6.2. Here, l refers to the dimension of the system matrix. Eigenvalues of the multigrid preconditioned operator $V^{-1}L$ are listed in Table 6.3. The minimum eigenvalue, λ_{min}, is slightly smaller for square elements, and thus accounts for the slightly higher iteration count.

h	l	CG	CG-AMG	time
$\frac{1}{16}$	736	50	4	(0.058)
$\frac{1}{32}$	3,008	102	4	(0.177)
$\frac{1}{64}$	12,160	202	4	(0.859)
$\frac{1}{128}$	48,896	408	5	(4.543)

Table 6.1: CG iterations, Example 1, unit coefficients, triangles

h	l	CG	CG-AMG	time
$\frac{1}{16}$	480	40	6	(0.055)
$\frac{1}{32}$	1,984	86	7	(0.118)
$\frac{1}{64}$	8,064	175	7	(0.452)
$\frac{1}{128}$	32,512	354	7	(1.813)

Table 6.2: CG iterations, Example 1, unit coefficients, squares

	triangles		squares	
h	λ_{min}	λ_{max}	λ_{min}	λ_{max}
$\frac{1}{8}$	0.9656	1	0.7389	1
$\frac{1}{16}$	0.9532	1	0.7307	1
$\frac{1}{32}$	0.9473	1	0.7247	1

Table 6.3: Eigenvalues of $V^{-1}L$, Example 1, unit coefficients

Note that L is not an M-matrix in the second case but black-box AMG yields an h-optimal preconditioner, for both elements. However, if we perform the same experiment with rectangles, say with $h_x = \frac{1}{2}h_y$, then the black-box AMG scheme fails.

Example 2

Next, consider anisotropic coefficients, $\mathcal{A} = \mathtt{diag}(\epsilon, 1)$. CG-AMG iteration counts for triangles are reported in Table 6.4. For square elements, convergence is observed to be highly erratic. Iteration counts for the case $\epsilon = 10^3$ are given in Table 6.5. Again, the M-matrix property is violated too strongly, causing the convergence rate to deteriorate with mesh refinement.

h	$\epsilon = 10^{-6}$	$\epsilon = 10^{-3}$	$\epsilon = 10^{3}$	$\epsilon = 10^{6}$	time
$\frac{1}{8}$	4	4	4	4	(0.0172)
$\frac{1}{16}$	4	4	4	4	(0.0294)
$\frac{1}{32}$	4	4	4	4	(0.0743)
$\frac{1}{64}$	4	4	4	4	(0.2540)

Table 6.4: CG iterations, Example 2, anisotropic coefficients, triangles

h	CG	CG-AMG
$\frac{1}{8}$	23	5
$\frac{1}{16}$	69	9
$\frac{1}{32}$	101	20
$\frac{1}{64}$	307	42

Table 6.5: CG iterations, Example 2, anisotropic coefficients, squares

h	$\epsilon = 10^{-6}$	$\epsilon = 10^{-3}$	$\epsilon = 10^{3}$	$\epsilon = 10^{6}$	time
$\frac{1}{8}$	4	4	4	5	(0.0354)
$\frac{1}{16}$	4	4	4	4	(0.0588)
$\frac{1}{32}$	4	4	4	4	(0.1775)
$\frac{1}{64}$	4	4	5	5	(0.8705)

Table 6.6: CG iterations, Example 3, discontinuous coefficients, triangles

Example 3

Now consider discontinuous coefficients. Choose $\mathcal{A} = \epsilon\mathcal{I}$ in $\Omega^* = [0.5, 1] \times [0, 0.5]$, and $\mathcal{A} = \mathcal{I}$ in $\Omega \backslash \Omega*$. Iteration counts for triangles and squares are listed in Tables 6.6 and 6.7, respectively. Black-box AMG yields an optimal preconditioner in *both* cases.

h	$\epsilon = 10^{-6}$	$\epsilon = 10^{-3}$	$\epsilon = 10^{3}$	$\epsilon = 10^{6}$	time
$\frac{1}{8}$	7	7	7	7	(0.0195)
$\frac{1}{16}$	7	7	7	7	(0.0375)
$\frac{1}{32}$	7	7	7	7	(0.1104)
$\frac{1}{64}$	7	7	7	7	(0.4003)

Table 6.7: CG iterations, Example 3, discontinuous coefficients, squares

Example 4

Finally, we choose a variable diagonal tensor, $\mathcal{A} = \left(1 + 100(x^2 + y^2)\right)^{-1}\mathcal{I}$. Iteration counts for triangles and squares are listed in Tables 6.8 and 6.9, respectively. Again, black-box AMG yields an optimal preconditioner in *both* cases.

h	CG	CG-AMG	time
$\frac{1}{8}$	187	4	(0.0313)
$\frac{1}{16}$	296	4	(0.0591)
$\frac{1}{32}$	877	4	(0.1743)
$\frac{1}{64}$	2,292	4	(1.1950)

Table 6.8: CG iterations, Example 4, variable coefficients, triangles

h	CG	CG-AMG	time
$\frac{1}{8}$	57	6	(0.0209)
$\frac{1}{16}$	193	7	(0.0417)
$\frac{1}{32}$	600	7	(0.1189)
$\frac{1}{64}$	1,636	7	(0.4182)

Table 6.9: CG iterations, Example 4, variable coefficients, squares

6.6 Computational work

For triangles, and diagonal coefficients, black-box AMG always provides a robust preconditioner for the Lagrange multiplier system because the M-matrix property is satisfied. For rectangles, the numerical results above show that AMG is *not* a reliable preconditioner. However, in cases where the M-matrix property is not violated too strongly, AMG *does* work and provides an h-optimal and $\mathcal{A}$-optimal preconditioner. CG convergence, in those cases, is fast. Hence, we are faced with the question, 'is solving the Lagrange multiplier problem a more efficient approach to solving the model variable diffusion problem?' A comparison with the approach of Chapter 5 cannot be made solely on the basis of iteration counts. More information is needed.

In the standard indefinite problem, recall that we applied AMG to the positive definite matrix, $BA_{diag}^{-1}B^T$. For triangles and rectangles, it has 4 and 5 non-zero entries per row, respectively (see Fig. 6.2.)

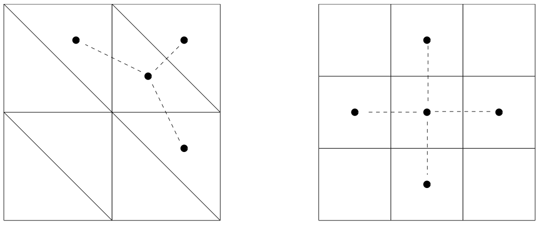

Figure 6.2: Connectivity of $BA_{diag}^{-1}B^T$

The dimension of the system corresponds to the number of finite elements in the mesh. In the Lagrange multiplier problem, the coefficient matrix L is denser, with a maximum of 5 or 7 non-zero entries per row (see Fig. 6.3.) It is also larger since its dimension is equal to the number of edges in the mesh.

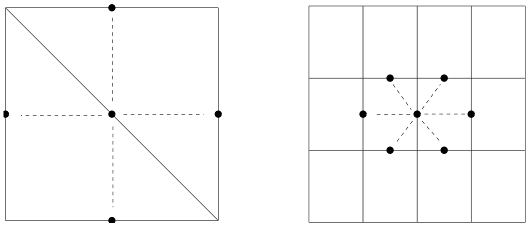

Figure 6.3: Connectivity of L

The computational cost of applying V-cycles of AMG to a given matrix depends on many factors, including the size and density of the coefficient matrices and the number of chosen interpolation points, at all levels. To evaluate this cost, for the two preconditioning schemes we have described, we use the standard theoretical estimates of Ruge and Stüben. Full details of the calculations can be found in [83] or [59].

Suppose that we are solving a linear system, $M\underline{x} = \underline{b}$, of dimension N. The major tasks in the AMG set-up phase are the construction of the coarse grid operators and the interpolation weights. We denote, by F_C and F_W, respectively, theoretical estimates of flop counts for these processes. Let s denote the total number of levels created by the multigrid algorithm. $M^s = M$ is the matrix associated with the finest level. N_M^j is the number of non-zeros in the coarse-grid matrix, M^j, and N_Ω^j is the number of points used at level j. With this notation, the operator complexity and the grid complexity, are, respectively,

$$C_A = \frac{\sum_{j=1}^s N_M^j}{N_M^s}, \quad C_\Omega = \frac{\sum_{j=1}^s N_\Omega^j}{N_\Omega^s}.$$

Next, we denote, by a, the average number of non-zero entries per row in M^s, and define,

$$\tilde{a} = a\,\frac{C_A}{C_\Omega},$$

the average number of non-zeros entries per row, of the matrices M^j, over all levels. Finally, let p denote the average, over all levels, of the number of interpolation points per F-point. Then, according to [83], we obtain,

$$\mathrm{F}_C = Np(2p(\tilde{a}-p)+3p+\tilde{a}), \tag{6.33}$$

$$\mathrm{F}_W = N3(\tilde{a}-(p+1))(p+1)+p. \tag{6.34}$$

Once set-up has been performed, a V-cycle is composed of smoothing, restriction and interpolation operations. If we perform a total of ν smoothing steps, at each level, the estimated total cost (again, see [83]) is,

$$\mathrm{F}_V = N\left(2(1+\nu)aC_A + 4p + C_\Omega - 1\right). \tag{6.35}$$

Combining (6.33)–(6.35), an estimate of the total flop count associated with applying AMG as a preconditioner is,

$$(F_C + F_W) + iF_V. \tag{6.36}$$

Here, i is the number of iterations required by the Krylov subspace solver to reduce the residual error to 10^{-6}. Note that set-up only has to be performed once. Below, we compare time costs (seconds) and estimated work costs (flops) for applying AMG as a preconditioner for L, in the Lagrange multiplier problem, and for $BA_{diag}^{-1}B^T$, in the indefinite problem described in Chapter 5. We consider the test problem described in the last section with fixed mesh size, $h = \frac{1}{64}$, and one pre- and one post-smoothing step.

6.6.1 Examples

We begin with uniform meshes of right-angled triangles and unit coefficients. Costs for the associated Lagrange multiplier problem are given in Table 6.10. Costs for the indefinite problem are given in Table 6.11.

Grids	Set-up time	V-cycle time	i	Total time
6	0.1419	0.0225	4	0.2319
a	C_A	C_Ω	$\tilde{a}$	p
3.65	1.99	1.56	4.67	1.91
F_C	F_W	F_V	N	Total MFlops
39.91	17.28	51.77	12,160	3.21

Table 6.10: Lagrange multiplier problem, triangles, unit coefficients

Consider, also, the discontinuous coefficient problem described in Example 3, above, with $\epsilon = 10^{-3}$. Costs for the associated Lagrange multiplier problem are given in Table

Grids	Set-up time	V-cycle time	i	Total time
6	0.1380	0.0191	26	0.6346
a	C_A	C_Ω	$\tilde{a}$	p
3.97	2.63	1.73	6.04	2.13
F_C	F_W	F_V	N	Total MFlops
63.05	30.38	73.34	8,192	16.39

Table 6.11: Indefinite problem, triangles, unit coefficients

6.12. Costs for the indefinite problem are given in Table 6.13. In both examples, applying AMG to $BA_{diag}^{-1}B^T$ is approximately 5 times more expensive than applying it to the Lagrange multiplier system.

Grids	Set-up time	V-cycle time	i	Total time
6	0.1455	0.0237	4	0.2403
a	C_A	C_Ω	$\tilde{a}$	p
3.65	1.99	1.56	4.67	1.90
F_C	F_W	F_V	N	Total MFlops
39.70	17.0	51.74	12,160	3.21

Table 6.12: Lagrange multiplier problem, triangles, discontinuous coefficients

This is easily explained. The combination of right-angled triangles and *any* diagonal coefficient tensor, $\mathcal{A}$, produces a matrix, L, that is actually sparser than $BA_{diag}^{-1}B^T$ (compare the values of a and $\tilde{a}$.) The operator complexity, C_A, is also slightly higher for $BA_{diag}^{-1}B^T$. Combining these features with the increased iteration count, for the indefinite problem, results in a more expensive preconditioner. For test problems of this size, in $\mathbb{R}^2$, however, note that the difference in time costs is negligible.

Now consider square elements. Again, we begin with unit coefficients. Costs for the associated Lagrange multiplier problem are given in Table 6.14. Costs for the indefinite problem are given in Table 6.15. Although more iterations are required for the indefinite problem, the total AMG cost is now sightly lower than for the Lagrange multiplier problem. This is due to the superior sparsity of $BA_{diag}^{-1}B^T$ and its relatively

Grids	Set-up time	V-cycle time	i	Total time
6	0.1469	0.0202	27	0.6923
a	C_A	C_Ω	$\tilde{a}$	p
3.97	2.58	1.72	5.96	2.10
F_C	F_W	F_V	N	Total MFlops
59.79	28.70	70.58	8,192	16.34

Table 6.13: Indefinite problem, triangles, discontinuous coefficients

small dimension.

We observe the same phenomenon for squares and other diagonal coefficient tensors. Consider, finally, the variable coefficient tensor in Example 4. Costs for the associated Lagrange multiplier problem are given in Table 6.16. Costs for the indefinite problem are given in Table 6.17. Once again, performing AMG on $BA_{diag}^{-1}B^T$, in the indefinite problem is slightly cheaper.

Grids	Set-up time	V-cycle time	i	Total time
6	0.1929	0.0253	7	0.37
a	C_A	C_Ω	$\tilde{a}$	p
6.91	2.40	1.67	9.93	2.35
F_C	F_W	F_V	N	Total MFlops
123.62	68.48	109.57	8,064	7.73

Table 6.14: Lagrange multiplier problem, squares, unit coefficients

Grids	Set-up time	V-cycle time	i	Total time
5	0.0724	0.0090	22	0.27
a	C_A	C_Ω	$\tilde{a}$	p
4.94	2.26	1.68	6.65	2.25
F_C	F_W	F_V	N	Total MFlops
74.70	35.40	76.67	4,096	7.36

Table 6.15: Indefinite problem, squares, unit coefficients

Grids	Set-up time	V-cycle time	i	Total time
6	0.1960	0.0255	7	0.3745
a	C_A	C_Ω	$\tilde{a}$	p
6.91	2.40	1.67	9.93	2.35
F_C	F_W	F_V	N	Total MFlops
100.29	68.48	109.58	8,064	7.55

Table 6.16: Lagrange multiplier problem, squares, variable coefficients

Grids	Set-up time	V-cycle time	i	Total time
5	0.0714	0.0089	22	0.2672
a	C_A	C_Ω	$\tilde{a}$	p
4.94	2.18	1.67	6.84	2.34
F_C	F_W	F_V	N	Total MFlops
76.53	33.50	74.64	4,096	7.18

Table 6.17: Indefinite problem, squares, variable coefficients

The cost of applying AMG, in the preconditioning schemes outlined in this chapter and in Chapter 5, is the major expense, but not the total cost of solving (1.4). The indefinite system (2.37) is larger than L in (6.9). Matrix-vector multiplications in the Krylov subspace iteration may be more expensive in the former case. The total time cost for solving the indefinite problem is usually slightly higher. For the Lagrange multiplier problem, we must also factor in the cost of recovering the velocity solution, which, after all, is what we seek. For non-diagonal coefficient tensors, this amounts to inverting a matrix with 2×2 diagonal blocks.

6.7 Concluding remarks

In this chapter, we reviewed properties of the mixed-hybrid formulation of the model variable diffusion problem (1.4) and explicitly derived the associated Lagrange multiplier system (6.9) for rectangular elements and diagonal coefficients. We established that the positive definite system matrix is not strictly an M-matrix, except in the special case of triangular elements and diagonal coefficients. Numerical evidence suggests, however, that black-box AMG is an h-optimal and $\mathcal{A}$-optimal preconditioner, for triangular and square elements, whenever the M-matrix property is not violated too strongly. Whether or not the same observation applies in $I\!R^3$, is a subject for future research.

In Chapter 5, we outlined a preconditioning strategy for the standard indefinite problem (2.37), which requires the application of AMG to the positive definite matrix $BA_{diag}^{-1}B^T$. For right-angled triangles, and diagonal coefficient tensors, we established that applying AMG to the Lagrange multiplier system is cheaper than applying it to the matrix $BA_{diag}^{-1}B^T$. We conclude that solving the Lagrange multiplier system is the cheapest approach to solving the model variable diffusion problem in this special case.

For rectangles, however, AMG is *not* a robust preconditioner for the Lagrange multiplier system and may fail, even for simple diagonal coefficients. The search for h-optimal and $\mathcal{A}$-optimal preconditioners for that system continues. When the M-matrix property is not violated too strongly, and AMG does work, it is *not* cheaper to apply it to the Lagrange multiplier system than to $BA_{diag}^{-1}B^T$.

Since $BA_{diag}^{-1}B^T$ is always an M-matrix, the indefinite problem (2.37) is more amenable to solution by black-box multigrid methods than the positive definite system (6.9). The preconditioning scheme described in Chapter 5 will never breakdown and is applicable to a broader range of geometries and coefficient tensors. We conclude that solving the indefinite system, using the method described in Chapter 5, is the only reliable *black-box* approach to solving the model variable diffusion problem.

Chapter 7

Black-box preconditioning for other saddle-point systems

Our analysis of solution schemes for the *particular* saddle-point system that arises in the lowest-order Raviart-Thomas approximation of the model variable diffusion problem, is now complete. We have described two distinct preconditioning schemes, so-called '$H(div)$ preconditioning' and 'H^1 preconditioning', each having links to stability properties of the underlying variational problem. To conclude our discussion, we now want to point out that the black-box scheme outlined in Chapter 5, provides a generic framework for tackling saddle-point systems that arise in *other* applications.

Many important physical processes are modelled by second-order elliptic PDES and are discretised with mixed finite element methods to preserve physical properties of the unknowns. Consequently, saddle-point systems of the generic form,

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} \underline{u} \\ \underline{p} \end{pmatrix} = \begin{pmatrix} \underline{f}_1 \\ \underline{f}_2 \end{pmatrix}, \tag{7.1}$$

arise in many fields in mathematics and engineering. If A is symmetric, we can always tackle (7.1) with MINRES. If A is also positive definite and B has full rank, then it is easy to show that an ideal, generic preconditioner is,

$$P = \begin{pmatrix} A & 0 \\ 0 & S \end{pmatrix}, \tag{7.2}$$

where $S = BA^{-1}B^T$. To obtain a practical scheme, we usually require approximations of A and/or S. Clearly, this requires knowledge of the properties of the underlying PDE operators. However, if the diagonal blocks of (7.2) represent Laplacian or diffusion operators, and are close to being M-matrices, then we now know that applying a V-cycle of black-box algebraic multigrid (AMG) is a good choice.

In this final chapter, we briefly outline black-box preconditioning schemes for saddle-point problems arising in two other applications. Numerical results are presented for the Stokes equations arising in incompressible flow modelling and Maxwell's equations, describing the classical magnetostatic problem. In analogy to the discussion in Chapter 5, we demonstrate that an optimal tool for preconditioning the associated saddle-point systems, is a V-cycle of black-box AMG.

7.1 Stokes equations

Let $\Omega \subset \mathbb{R}^d$, $(d = 2, 3)$, be a specified flow domain, with piecewise smooth boundary $\partial\Omega$. The classical Stokes problem is,

find a velocity $\vec{u}$ and a pressure p satisfying,

$$\begin{aligned} -\nu\nabla^2\vec{u} + \nabla p &= \vec{f}, \\ \nabla \cdot \vec{u} &= 0 \quad \text{in } \Omega, \\ \vec{u} &= \vec{0} \quad \text{on } \partial\Omega. \end{aligned} \tag{7.3}$$

For simplicity in our description, we consider a 'no-flow' boundary condition. The parameter ν represents viscosity of the fluid and $\vec{f}$ is a given forcing term. The equations (7.3) arise in the modelling of low-speed, viscous flows, and play an important role in steady-state approximations of the Navier-Stokes equations.

A standard, conforming mixed finite element approximation is obtained by constructing finite-dimensional subspaces $V_h \subset \left(H_0^1(\Omega)\right)^d$ and $W_h \subset L^2(\Omega)$. The associated discrete variational formulation is, find $\vec{u}_h \in V_h$, $p_h \in W_h$ satisfying,

$$\begin{aligned} \nu\left(\nabla \vec{u}_h, \nabla \vec{v}_h\right) - \left(p_h, \nabla \cdot \vec{v}_h\right) &= \left(\vec{f}, \vec{v}_h\right) \qquad \forall \vec{v}_h \in V_h & (7.4)\\ \left(\nabla \cdot \vec{u}_h, w_h\right) &= 0 \qquad\qquad \forall w_h \in W_h. & (7.5) \end{aligned}$$

Choosing basis sets,

$$V_h = \mathtt{span}\left\{\vec{\psi}_i\right\}_{i=1}^{n}, \quad W_h = \mathtt{span}\left\{\phi_j\right\}_{j=1}^{m}, \tag{7.6}$$

leads to a linear algebra problem of the form (7.1). We obtain,

$$\begin{pmatrix} \nu A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} \underline{u} \\ \underline{p} \end{pmatrix} = \begin{pmatrix} \underline{f} \\ \underline{0} \end{pmatrix}, \tag{7.7}$$

where $\vec{u}_h = \sum_{i=1}^{n} u_i \vec{\psi}_i$ and $p_h = \sum_{j=1}^{m} p_j \phi_j$ with,

$$\begin{aligned} A_{ij} &= \int_\Omega \nabla \vec{\psi}_i : \nabla \vec{\psi}_j \, d\Omega & & i,j = 1:n, \\ B_{kj} &= -\left(\nabla \cdot \vec{\psi}_j, \phi_k\right) & & k = 1:m,\ j = 1:n, \\ \underline{f}_i &= \left(\vec{f}, \vec{\psi}_i\right) & & i = 1:n. \end{aligned}$$

Note that, here, A is a *vector* Laplacian matrix with d diagonal blocks corresponding to scalar Laplacian matrices.

To ensure existence and uniqueness of a solution to the problem (7.4)–(7.5), we choose V_h and W_h to satisfy the discrete inf-sup condition,

$$\beta \parallel w_h \parallel_0 \le \sup_{\vec{v}_h \in V_h} \frac{\left(w_h, \nabla \cdot \vec{v}_h\right)}{\parallel \nabla \vec{v}_h \parallel_0} \quad \forall w_h \in W_h,\ w_h \neq \text{ constant}, \tag{7.8}$$

where $\beta \geq \beta_0 > 0$ and β_0 is a constant independent of h. Note that since we have specified the velocity everywhere on the boundary, the pressure solution is only defined up to a constant. Also, since $V_h \subset \left(H_0^1(\Omega)\right)^d$, the norms, $\parallel \nabla \vec{v}_h \parallel_0$ and $\parallel \vec{v}_h \parallel_1$ are equivalent on V_h.

Now, by defining the pressure mass matrix,

$$Q_{kl} = \left(\phi_k, \phi_l\right), \qquad k, l = 1:m,$$

the inf-sup condition (7.8) can be expressed in matrix form as,

$$\beta^2 \le \frac{\underline{p}^t B A^{-1} B^T \underline{p}}{\underline{p}^t Q \underline{p}} \quad \forall \underline{p} \in I\!R^m \setminus \left\{\underline{1}\right\}. \tag{7.9}$$

(The reader is referred to Chapter 3 for the derivation of such results.) It can also be shown that for a conforming finite element space,

$$\frac{\underline{p}^t B A^{-1} B^T \underline{p}}{\underline{p}^t Q \underline{p}} \leq 1 \quad \forall \underline{p} \in I\!R^m \backslash \{\underline{0}\}. \tag{7.10}$$

Thus, if we choose finite element spaces V_h and W_h that satisfy (7.8), it follows that the associated pressure mass matrix Q provides an h-optimal preconditioner for the Schur complement matrix $BA^{-1}B^T$.

7.1.1 Preconditioning strategy

In [90] and [91], Silvester and Wathen prove that for any stable finite element method, the eigenvalues, $\{\sigma_i\}_{i=1}^{n+m}$, of the generalised eigenvalue problem,

$$\begin{pmatrix} \nu A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} \underline{u} \\ \underline{p} \end{pmatrix} = \sigma \begin{pmatrix} \nu P_A & 0 \\ 0 & \frac{1}{\nu} P_Q \end{pmatrix} \begin{pmatrix} \underline{u} \\ \underline{p} \end{pmatrix}, \tag{7.11}$$

are bounded by constants independent of the discretisation parameter, h, if P_A and P_Q are chosen to satisfy,

$$\lambda \leq \frac{\underline{u}^T A \underline{u}}{\underline{u}^T P_A \underline{u}} \leq \Lambda \quad \forall \underline{u} \in I\!R^n \backslash \{\underline{0}\}, \tag{7.12}$$

$$\theta \leq \frac{\underline{p}^T Q \underline{p}}{\underline{p}^T P_Q \underline{p}} \leq \Theta \quad \forall \underline{p} \in I\!R^m \backslash \{\underline{0}\}, \tag{7.13}$$

with positive constants λ, Λ, θ and Θ, independent of h. Specifically, they prove the following Lemma.

Lemma 31 *Assume a stable discretisation satisfying (7.9) and (7.10) and choose P_A and P_Q to satisfy (7.12) and (7.13), respectively. The eigenvalues of the generalised eigenvalue problem (7.11) lie in the union of the intervals,*

$$\begin{gathered} \left[\tfrac{1}{2}\left(\lambda - \sqrt{\lambda^2 + 4\Theta\Lambda}\right), \tfrac{1}{2}\left(\lambda - \sqrt{\lambda^2 + 4\beta^2\theta\lambda}\right)\right], \\ \cup \left[\lambda, \tfrac{1}{2}\left(\Lambda + \sqrt{\Lambda^2 + 4\Theta\Lambda}\right)\right]. \end{gathered} \tag{7.14}$$

Proof See [91]. □

By applying, once again, the analysis of Wathen, [101], it is possible to show that (7.13) holds with $P_Q = \mathtt{diag}\,(Q)$. In other words, diagonal scaling is always an h-optimal preconditioner for the Stokes pressure mass matrix. One of the best Stokes

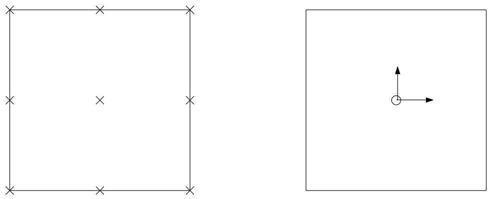

Figure 7.1: $Q_2 - P_{-1}$ element; velocity components (left), pressure $\left(p, \frac{\partial p}{\partial x}, \frac{\partial p}{\partial y}\right)$ components (right)

approximations in $I\!\!R^2$ is the $Q_2 - P_{-1}$ finite element (see Fig. 7.1), consisting of piecewise biquadratic velocity and discontinuous piecewise linear pressure. Not only is (7.8) satisfied but Q is also a diagonal matrix. Thus, $P_Q = Q$ is an optimal choice with $\theta = \Theta = 1$.

The task of defining an efficient preconditioner for (7.7) is thus reduced to finding a P_A satisfying (7.12). Since A is a discrete representation of the vector Laplacian operator ∇^2, any black-box solver for Poisson problems is a potential candidate. Silvester and Wathen demonstrate numerically in [91] that one V-cycle of standard *geometric* multigrid is an optimal choice, for uniform grids in $I\!\!R^2$. Instead, we choose P_A in (7.11) to be one V-cycle of the *algebraic* multigrid algorithm `amg1r5`, described in Chapter 5. Thus, in $I\!\!R^2$, the preconditioner we propose is,

$$P_{amg} = \begin{pmatrix} \nu V(A_x) & 0 & 0 \\ 0 & \nu V(A_y) & 0 \\ 0 & 0 & \frac{1}{\nu} Q \end{pmatrix}, \tag{7.15}$$

where $V(A_x)$ and $V(A_y)$ denote the application of a single V-cycle of AMG to the diagonal blocks A_x and A_y of A. Again, we implement AMG as a black-box, with symmetric Gauss-Seidel smoothing. No parameters are estimated a-priori.

Remark 13 *When biquadratic approximation is used for the velocity, the matrices A_x and A_y have positive as well as negative off-diagonal entries. However, the positive entries are small in magnitude, relative to the other entries, and do not adversely affect the multigrid approximation.*

7.1.2 Numerical example

To illustrate the h-optimality of (7.15), we perform a numerical example. Consider the Stokes equations (7.3), discretised on $[-1,1] \times [-1,1]$ using stable $Q_2 - P_{-1}$ approximation. Choose $\nu = 1$, $\vec{f} = \vec{0}$, and the velocity boundary condition,

$$(u_x, u_y) \;=\; \big\{ \big(1 - x^4, 0\big) \;\mid\; y \;=\; 1,\, -1 \leq \, x \, \leq 1 \big\}\,. \tag{7.16}$$

This corresponds to the benchmark, 'regularised lid-driven cavity' flow problem (see Fig. 7.2).

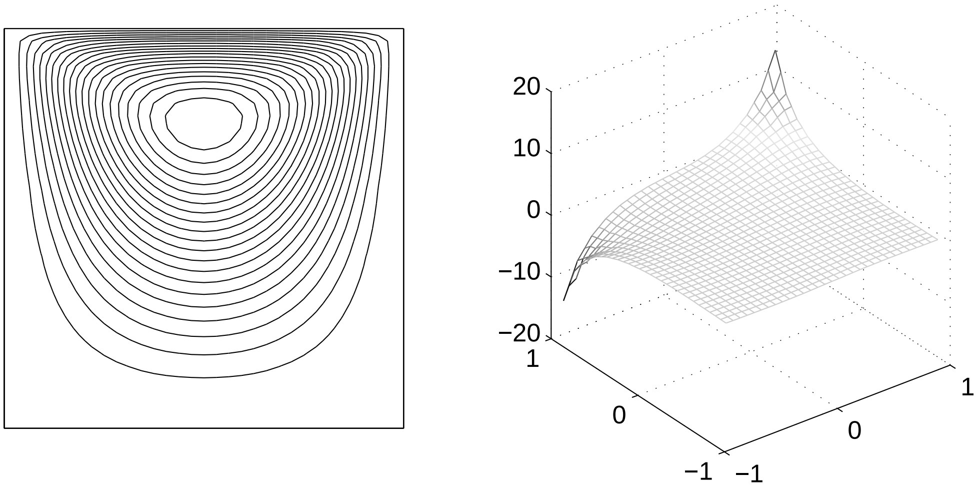

Figure 7.2: Contour plot of constant streamlines (left) and pressure solution (right)

We apply preconditioned MINRES to the assembled system (7.7), with a stopping tolerance of 10^{-6} on the relative residual error. Iteration counts for the preconditioned system are reported in Table 7.1. The second column contains unpreconditioned counts; the third column lists counts for the ideal (without AMG) version of the preconditioner. The time units reported in parentheses are elapsed time in seconds for the total solve, including AMG set-up time.

The eigenvalues of the preconditioned system are reported in Table 7.2. Substituting $\theta = 1$, $\Theta = 1$, $\lambda = 0.884$, $\Lambda = 1$ and $\beta^2 = 0.222$ in (7.14) yields the theoretical bound, $[-0.651, -0.184] \,\cup\, [0.884, 1.618]$. Fig. 7.3 illustrates the spectral equivalence of the AMG version (P_{amg}) and the ideal version (P_{ideal}) of the preconditioner. Note that in this example the preconditioned system is singular. The nullspace is spanned by the hydrostatic constant pressure solution. This accounts for the zero eigenvalue.

h	$P = I$	P_{ideal}		P_{amg}	
$\frac{1}{8}$	64	22	(0.2)	25	(0.2)
$\frac{1}{16}$	147	24	(1.3)	26	(0.5)
$\frac{1}{32}$	274	24	(11.5)	28	(2.2)
$\frac{1}{64}$	513	24	(96.8)	29	(9.2)

Table 7.1: MINRES iterations, Stokes problem

h	λ	Λ	β^2	Observed eigenvalues
$\frac{1}{8}$	0.951	1	0.256	$[-0.615,\ -0.211] \cup [0] \cup [0.960,\ 1.613]$
$\frac{1}{16}$	0.917	1	0.235	$[-0.618,\ -0.196] \cup [0] \cup [0.939,\ 1.617]$
$\frac{1}{32}$	0.884	1	0.222	$[-0.618,\ -0.186] \cup [0] \cup [0.924,\ 1.617]$

Table 7.2: Eigenvalues of preconditioned system

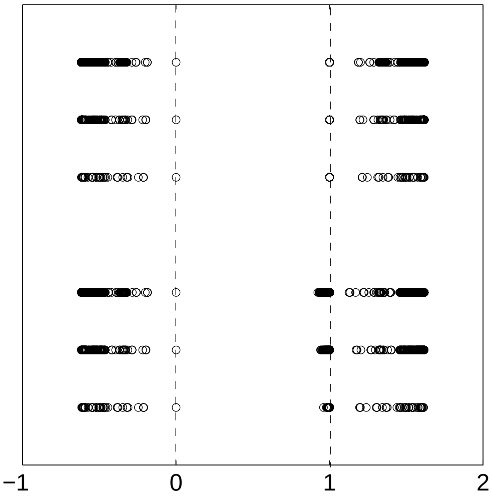

Figure 7.3: Eigenvalues of indefinite preconditioned matrix; P_{ideal} (top), P_{amg} (bottom), $h = \frac{1}{8}, \frac{1}{16}, \frac{1}{32}$

The experiment shows that we have a simple, h-optimal preconditioner for Stokes problems. Next, we consider the so-called ‘magnetostatic problem’, in $\mathbb{R}^2$.

7.2 Maxwell's equations

Let $\Omega \subset I\!R^2$ be a convex polygon, with piecewise smooth boundary $\partial\Omega = \partial\Omega_H \cup \partial\Omega_B$. The classical magnetostatic problem is,

find a magnetic field, $\vec{H}$, and a magnetic displacement field, $\vec{B}$, satisfying,

$$\begin{aligned} \nabla \cdot \vec{B} &= 0 \quad \text{in } \Omega, & (7.17)\\ \nabla \times \vec{H} &= j \quad \text{in } \Omega, & (7.18)\\ \vec{B} &= \mu\vec{H} \quad \text{in } \Omega, & (7.19)\\ \vec{B} \cdot \vec{n} &= 0 \quad \text{on } \partial\Omega_B, \\ \vec{H} \cdot \vec{t} &= 0 \quad \text{on } \partial\Omega_H. \end{aligned}$$

This corresponds to a steady-state subset of Maxwell's equations, and models the properties of the magnetic field obtained when electrical currents are passed through magnetic media. Again, for simplicity in our description, we consider homogeneous boundary conditions. Note that in $I\!R^2$, we define the curl operator via,

$$\nabla \times \vec{H} = -\frac{\partial H_x}{\partial y} + \frac{\partial H_y}{\partial x}.$$

The parameter μ is the magnetic permeability coefficient and is assumed to be piecewise constant. The vectors $\vec{n}$ and $\vec{t}$ represent the unit normal and tangential vectors to $\partial\Omega_B$ and $\partial\Omega_H$, respectively, and j denotes an imposed current density.

There are many different ways to formulate (7.17)–(7.19) as a mixed variational problem. The scheme we have in mind is one described by Perugia and Simoncini, [76], and Perugia et al., [75], which aims to preserve the physical properties of $\vec{B}$ *and* $\vec{H}$ by solving (7.17)–(7.18) exactly, whilst minimising the residual of (7.19). Specifically, we solve the constrained optimization problem,

$$\min_{\nabla\cdot\vec{B}=0, \nabla\times\vec{H}=j} \frac{1}{2} \parallel \vec{B} - \mu\vec{H} \parallel^2 \ \text{ in } \Omega,$$

by introducing two Lagrange multipliers, λ_1, λ_2, for the constraints (7.17) and (7.18). The natural norm, $\parallel \cdot \parallel$, for the minimisation is the weighted L^2-norm, defined via, $\parallel \vec{v} \parallel^2 = \left(\mu^{-1}\vec{v}, \vec{v}\right)$. Full details can be found in [75] and [76] and the references therein.

For a conforming formulation, we require subspaces $W_h^1 \subset L^2(\Omega)$, $W_h^2 \subset L^2(\Omega)$, $V_h^1 \subset H_{0,B}(div;\Omega)$, and $V_h^2 \subset H_{0,H}(curl;\Omega)$, to approximate λ_1, λ_2, $\vec{B}$ and $\vec{H}$, respectively. Here,

$$\begin{aligned} H_{0,B}(div;\Omega) &= \{\vec{v} \in L^2(\Omega)^2 \mid \nabla\cdot\vec{v} \in L^2(\Omega) \text{ and } \vec{v}\cdot\vec{n}|_{\partial\Omega_B} = 0\}. \\ H_{0,H}(curl;\Omega) &= \{\vec{h} \in L^2(\Omega)^2 \mid \nabla\times\vec{h} \in L^2(\Omega) \text{ and } \vec{h}\cdot\vec{t}|_{\partial\Omega_H} = 0\}. \end{aligned}$$

Now, given a mesh T_h, an inf-sup stable approximation is obtained by choosing $W_h^1 = W_h^2 = W_h$, the set of piecewise constant functions, $V_h^1 = V_h$, the continuous Raviart-Thomas space defined in (6.2), and $V_h^2 = E_h \cap H_{0,H}(curl;\Omega)$, where E_h is the so-called 'edge-element' space, defined via,

$$E_h = \{\vec{h} \mid \vec{h} \in E(K) \quad \forall K \in T_h\}, \quad E(K) = (P_0(K))^2 + (-y,x)^T\, P_0(K).$$

Functions in V_h^2 have continuous *tangential* components at interelement boundaries.

The corresponding discrete mixed variational problem, is,

find $\vec{B}_h \in V_h^1$, $\vec{H}_h \in V_h^2, \lambda_{1,h} \in W_h$ and $\lambda_{2,h} \in W_h$, satisfying,

$$\begin{aligned} \left(\frac{1}{\mu}\vec{B}_h,\vec{v}_h\right) - \left(\vec{H}_h,\vec{v}_h\right) + \left(\frac{1}{\mu}\lambda_{1,h},\nabla\cdot\vec{v}_h\right) &= 0 & \forall\,\vec{v}_h \in V_h^1, \quad (7.20)\\ -\left(\vec{B}_h,\vec{w}_h\right) + \left(\mu\vec{H}_h,\vec{w}_h\right) + (\mu\lambda_{2,h},\nabla\times\vec{w}_h) &= 0 & \forall\,\vec{w}_h \in V_h^2, \quad (7.21)\\ \left(\frac{1}{\mu}\nabla\cdot\vec{B}_h, x_h\right) &= 0 & \forall\,x_h \in W_h, \quad (7.22)\\ \left(\mu\nabla\times\vec{H}_h, y_h\right) &= (\mu j, y_h) & \forall\,y_h \in W_h. \quad (7.23) \end{aligned}$$

Choosing basis sets,

$$V_h^1 = \mathtt{span}\left\{\vec{\psi}_i\right\}_{i=1}^n, \quad V_h^2 = \mathtt{span}\left\{\vec{\chi}_j\right\}_{j=1}^n, \quad W_h = \mathtt{span}\{\phi_k\}_{k=1}^m,$$

we again obtain a saddle-point system, with the block form,

$$\begin{pmatrix} A_{11} & A_{12} & B_1^T & 0 \\ A_{12}^T & A_{22} & 0 & B_2^T \\ B_1 & 0 & 0 & 0 \\ 0 & B_2 & 0 & 0 \end{pmatrix} \begin{pmatrix} \underline{B} \\ \underline{H} \\ \underline{\lambda_1} \\ \underline{\lambda_2} \end{pmatrix} = \begin{pmatrix} \underline{0} \\ \underline{0} \\ \underline{0} \\ \underline{f} \end{pmatrix}, \tag{7.24}$$

where, for $i,j = 1:n$, and $k = 1:m$,

$$\begin{aligned} A_{11,\,ij} &= \left(\tfrac{1}{\mu}\vec{\psi}_i,\,\vec{\psi}_j\right), & A_{12,\,ij} &= -\left(\vec{\psi}_i,\,\vec{\chi}_j\right), & A_{22,\,ij} &= \left(\mu\vec{\chi}_i,\,\vec{\chi}_j\right), \\ B_{1,\,kj} &= \left(\phi_k,\,\tfrac{1}{\mu}\nabla\cdot\vec{\psi}_j\right), & B_{2,\,kj} &= \left(\phi_k,\,\mu\nabla\times\vec{\chi}_j\right), & \underline{f}_k &= \left(\mu\, j,\,\phi_k\right). \end{aligned} \tag{7.25}$$

7.2.1 Preconditioning strategy

Sparse direct and iterative solution schemes for (7.24), based on factorisation methods, are discussed and evaluated by Perugia et al. in [75] and [76]. In particular, following Rusten and Winther, [86], the authors consider the performance of the ideal block-diagonal preconditioner,

$$P = \begin{pmatrix} I & 0 & 0 & 0 \\ 0 & I & 0 & 0 \\ 0 & 0 & B_1B_1^T & 0 \\ 0 & 0 & 0 & B_2B_2^T \end{pmatrix}. \tag{7.26}$$

As we have already explained in Chapter 2, and Chapter 5, the major deficiency of this approach is that it assumes that the leading block of (7.24),

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{12}^T & A_{22} \end{pmatrix},$$

can be well approximated by the identity matrix. To see this, consider (7.14) in Lemma 31 with $P_A = I$, $\nu = 1$ and $P_Q = Q$ replaced by $BB^T = \mathtt{diag}\left(B_1B_1^T, B_2B_2^T\right)$.

The preconditioner P in (7.26) is not h-optimal. To improve on this, Perugia et al. perform diagonal scaling on the whole of the system matrix (7.24), before applying (7.26). This has the effect of damping out some of the ill-conditioning due to the permeability coefficient μ, which can be highly discontinuous in practical simulations. A-priori scaling is essential to obtain h-optimal MINRES convergence. It is not clear, however, that μ-optimal convergence can be achieved, even with diagonal scaling.

To obtain a practical scheme, Perugia et al. propose incomplete Cholesky factorisation for the diagonal blocks $B_1B_1^T$ and $B_2B_2^T$. We denote the resulting preconditioner $P_{ic,\epsilon}$, where ϵ is the chosen drop tolerance parameter for the factorisation. Instead, we propose the preconditioner,

$$P_{amg} = \begin{pmatrix} I & 0 & 0 & 0 \\ 0 & I & 0 & 0 \\ 0 & 0 & V\left(B_1B_1^T\right) & 0 \\ 0 & 0 & 0 & V\left(B_2B_2^T\right) \end{pmatrix}, \tag{7.27}$$

where V denotes the application of a single V-cycle of black-box AMG. Our choice is appropriate because the matrix $B_1 B_1^T$ is, by definition, a discrete representation of a scalar diffusion operator. Any fast solver for Poisson problems is, again, a potential candidate but with AMG we can handle unstructured meshes with ease. The second matrix, $B_2 B_2^T$, is a discrete representation of the curl operator acting on the space V_h^2. In $I\!R^2$, it has the appearance of a diffusion operator and hence can also be preconditioned effectively with AMG.

7.2.2 Numerical example

To illustrate the h-optimality of the preconditioner P_{amg} in (7.27), we perform a numerical example. The test case we consider is the one proposed in [75] and [76]. Consider a hollow iron cylinder, placed in a prescribed induction field that is uniform in the direction orthogonal to the cylinder axis. To fix ideas, suppose that the cylinder axis is aligned with the z-axis, so that it is sufficient to consider a cross-section in the x-y plane. For simplicity, assume that the cross-section is centered at the origin of that plane. Due to symmetry, only a quadrant of the cross-section then needs to be studied.

Thus, we consider the equations (7.17)–(7.19), discretised on the square domain $[0,1] \times [0,1]$. Let the internal and external radii of the cylinder be 0.2 and 0.1 units respectively. The jump in the magnetic permeability μ at the iron-air boundary is of three orders of magnitude. Boundary conditions are homogeneous, except at the side of the square that coincides with the inflow direction of the induction field, where $\vec{B} \cdot \vec{n}$ takes a prescribed value. Unstructured, non-uniform meshes are employed.

N	P_{ideal}	$P_{ic,0}$		$P_{ic,10^{-2}}$		$P_{ic,10^{-4}}$	
2,088	49	180	(1.85)	72	(0.91)	49	(0.57)
3,810	49	233	(3.88)	85	(1.69)	49	(1.17)
9,102	52	365	(17.53)	123	(5.69)	53	(3.78)
14,808	49	462	(37.11)	153	(11.98)	50	(6.45)

Table 7.3: MINRES iterations, magnetostatic problem

We apply MINRES to the diagonally scaled version of the assembled system (7.24) with the preconditioners $P_{ic,\epsilon}$ and P_{amg}. For comparison purposes, we use a stopping tolerance of 10^{-8} on the relative residual error. Iteration counts for the preconditioned system with $P_{ic,\epsilon}$ are reported in Table 7.3. In the first column, N corresponds to the dimension of the system. The second column contains iteration counts for the ideal version of the preconditioner (7.26). The remaining columns list iteration counts for the preconditioner with incomplete Cholesky factorisation, with various values of ϵ. The time units in parentheses are elapsed time in seconds for the total solve, including factorisation time.

The key observation is that the drop tolerance parameter needs to be tuned to obtain h-optimal convergence. Since we do not know, a-priori, how much fill-in is required, this is a serious drawback. Iteration counts for the preconditioned system with P_{amg} are reported in Table 7.4. We apply AMG as a black-box with symmetric smoothing. Here, no parameters are tuned at all. Observe that h-optimal convergence is immediately achieved.

N	P_{ideal}	P_{amg}	
2,088	49	51	(0.82)
3,810	49	51	(1.33)
9,102	52	55	(3.32)
14,808	49	53	(4.73)

Table 7.4: MINRES iterations, magnetostatic problem

Remark 14 *In simulations in $I\!R^2$, it is clear that P_{amg} is a more desirable preconditioner than $P_{ic,\epsilon}$. Unfortunately, we cannot apply black-box* AMG *in $I\!R^3$ because the matrix $B_2 B_2^T$ has very different properties.*

7.3 Concluding remarks

The message that we want to convey in this final chapter is that the saddle-point systems arising in mixed finite element formulations of variable diffusion problems, Stokes flows problems and the magnetostatic problem in $I\!R^2$, can all be solved efficiently, with a variety of coefficients, and on a range of finite element meshes, using the same preconditioning strategy. The ingredients are MINRES, a V-cycle of black-box AMG and possibly diagonal scaling, applied to one or more diagonal blocks of the matrix (7.2). Any other saddle-point system of the form (7.1) that gives rise to matrices A and $S = BA^{-1}B^T$, that are close to being M-matrices, can be treated in the same way.

Appendix A

Mathematical Notation

Symbol	**Definition**
	Function spaces and norms
Ω	:= bounded and connected domain in $\mathbb{R}^d$
$\partial\Omega$	:= boundary of Ω
$\overline{\Omega}$	:= $\Omega \cup \partial\Omega$
$\partial\Omega_N$	:= Dirichlet boundary of Ω
$\partial\Omega_D$	:= Neumann boundary of Ω
$\vec{v}$	:= $(v_1, \ldots, v_d)^T$, a vector in $\mathbb{R}^d$
∇w	:= $\left(\frac{\partial w}{\partial x_1}, \ldots, \frac{\partial w}{\partial x_d} \right)^T$, the gradient operator
$\nabla \cdot \vec{v}$	:= $\sum_{i=1}^{d} \frac{\partial v_i}{\partial x_i}$, the divergence operator
$\vec{\alpha}$	:= $(\alpha_1, \alpha_2, \ldots, \alpha_d)$, a multi-index
$\lvert \vec{\alpha} \rvert$	:= $\sum_{i=1}^{d} \alpha_i$
$D^{\vec{\alpha}} w$	:= $\frac{\partial^{\lvert\vec{\alpha}\rvert} w}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}$
$C^k(\Omega)$	:= set of functions with continuous derivatives of up to degree k
$L^2(\Omega)$	:= $\{ w \mid \int_\Omega w^2 \, d\Omega < \infty \}$
$L^2(\Omega)^d$	:= $\{ \vec{v} \mid v_i \in L^2(\Omega), \, i = 1 : d \}$
(w, s)	:=$\int_\Omega w \, s \, d\Omega$
$(\vec{u}, \vec{v})$	:= $\int_\Omega \vec{u} \cdot \vec{v} \, \partial\Omega$
$\lVert \vec{u} \rVert_0$	:= $(\vec{u}, \vec{u})^{\frac{1}{2}}$
$H^k(\Omega)$	:= $\{ w \in L^2(\Omega) \mid D^{\vec{\alpha}} w \in L^2(\Omega), \lvert \vec{\alpha} \rvert \leq k \}$

$| \, w|_k \quad := \left(\sum_{|\vec{\alpha}|=k} \| \, D^{|\vec{\alpha}|} \, \|_0^2 \right)^{\frac{1}{2}}$

$\| \, w \, \|_k \quad := \left(\| \, w \, \|_0 \, + \sum_{i=1}^{k} \, | \, w \, |_i \right)^{\frac{1}{2}}$

$H^1(\Omega) \quad := \{ \, w \in L^2(\Omega) \, \mid \, \nabla w \in L^2(\Omega)^d \, \}$

$H^1_{g,D}(\Omega) \quad := \{ \, w \in H^1(\Omega) \, \mid \, w = g \text{ on } \partial\Omega_D \, \}$

$H^{-1}(\Omega) \quad := \{ \, w \, \mid \, \int_\Omega s \, w \, d\Omega \, < \infty \quad \forall \, s \in H^1_0(\Omega) \}$

$H(div; \Omega) \quad := \left\{ \, \vec{v} \in L^2(\Omega)^d \mid \nabla \cdot \vec{v} \in L^2(\Omega) \, \right\}$

$H_{0,N}(div; \Omega) \quad := \left\{ \, \vec{v} \in H(div; \Omega) \mid \langle \vec{v} \cdot \vec{n}, w \rangle = 0 \, \forall \, w \, \in H^1_{0,D}(\Omega) \, \right\}$

$\langle \vec{v} \cdot \vec{n}, \, w \rangle \quad := \int_{\partial\Omega} w \, \vec{v} \cdot \vec{n} \, ds$

$(\vec{u}, \vec{v})_{div} \quad := (\vec{u}, \vec{v}) + (\nabla \cdot \vec{u}, \nabla \cdot \vec{v})$

$\| \, \vec{u} \, \|_{div} \quad := (\vec{u}, \vec{u})^{\frac{1}{2}}_{div}$

$(\vec{u}, \vec{v})_{div,\mathcal{A}} \quad := \left(\mathcal{A}^{-1} \vec{u}, \vec{v} \right) + (\nabla \cdot \vec{u}, \nabla \cdot \vec{v})$

$\| \, \vec{u} \, \|_{div,\mathcal{A}} \quad := (\vec{u}, \vec{u})^{\frac{1}{2}}_{div,\mathcal{A}}$

$H^{\frac{1}{2}}(\partial\Omega) \quad := \left\{ \, g \, \mid \, g = w|_{\partial\Omega} \, , w \in H^1(\Omega) \cap C^0(\overline{\Omega}) \right\}$

$H^{-\frac{1}{2}}(\partial\Omega) \quad := \left\{ \, \gamma \, \mid \, \gamma = (\vec{v} \cdot \vec{n})|_{\partial\Omega} \, , \vec{v} \in H(div) \cap (C^0(\overline{\Omega}))^2 \right\}$

PDE parameters

$\mathcal{A} \quad := \mathcal{A}(\vec{x}) \in I\!R^{d \times d}$, permeability tensor

$\gamma, \, \Gamma \quad := \gamma(\vec{v}, \vec{v}) \leq \, (\mathcal{A}^{-1} \vec{v}, \vec{v}) \, \leq \, \Gamma(\vec{v}, \vec{v}) \quad \forall \, \vec{v} : \Omega \to I\!R^d$

Finite element meshes and spaces

$T_h \quad :=$ partition of Ω

$K \quad :=$ finite element of T_h

$h_K \quad :=$ diameter of K

$h \quad := \max_{K \in T_h} h_K$

$h_{min} \quad := \min_{K \in T_h} h_K$

$P_k(K) \quad :=$ set of polynomials of degree $\leq k$ on K

$Q_{r,s}(K) \quad :=$ set of polynomials of degree $\leq k$ in x
and degree $\leq s$ in y on K

$w|_K \quad :=$ restriction of a function $w(\vec{x})$ to element K

$\mathcal{E}_h$: = set of all edges of T_h

$\mathcal{E}_\mathcal{I}$: = set of interior edges of T_h

$\mathcal{E}_\mathcal{D}$: = set of edges of T_h on $\partial\Omega_D$

$\mathcal{E}_\mathcal{N}$: = set of edges of T_h on $\partial\Omega_N$

e_i := an edge belonging to $\mathcal{E}_h$ with global label i

e_j^K := an edge of element K with local label j

$\vec{n}^i$:= unit normal vector to global edge e_i

$\vec{n}_K^j$:= unit *outward* normal vector to local edge j of K

$\vec{\nu}^i$:= *orientated* unit normal vector of global edge e_i

$\vec{\nu}_K^j$:= *orientated* unit normal vector of local edge j of K

$$s_K^i := \begin{cases} +1 & \text{if } \vec{n}_K^i = \vec{\nu}_K^i \\ -1 & \text{if } \vec{n}_K^i = -\vec{\nu}_K^i \end{cases}$$

Raviart-Thomas approximation

$RT_k(K)$:= $(P_k(K))^d + \vec{x}\, P_k(K)$ for triangular K

$RT_k(K)$:= $Q_{k+1,k}(K) \times Q_{k,k+1}(K)$ for rectangular K

$RT_k(\Omega; T_h)$:= $\{\vec{v} \in H(div;\Omega) \mid \vec{v}|_K \in RT_k(K)\ \forall\ K \in T_h\}$

$R_k(\partial K)$:= $\{s \mid s \in L^2(\partial K),\ s|_{e_i^K} \in P_k(e_i^K)\, \forall\, e_i^K\}$

$M(K)$:= $\{\vec{v} \in (L^s(K))^d \mid \nabla\cdot\vec{v} \in L^2(\Omega)\}$, $s > 2$

M := $\{\vec{v} \in H_{0,N}(div;\Omega) \mid \vec{v} \in (L^s(\Omega))^d\}$, $s > 2$,

V_h := $\{\vec{v} \in RT_k(\Omega;T_h)$ and $\vec{v}\cdot\vec{n} = 0$ on $\partial\Omega_N\}$

W_h := $\{w \in L^2(\Omega) \mid w\,|_K \in P_k(K)\,\forall K \in T_h\}$

$\tilde{V}_h$:= $\{\vec{v} \in L^2(\Omega)^d \mid \vec{v} \in RT_k(K)\,\forall K \in T_h\}$

$L_{g,D}$:= $\{w \in P_0(\mathcal{E}_h) \mid w|_e = \frac{1}{|e|}\int_e g\,ds = 0, \forall\, e \in \mathcal{E}_h \cap \partial\Omega_D\}$

n := dimension of V_h, the number of edges (faces) in $T_h \backslash \partial\Omega_N$

m := dimension of W_h, the number of elements in T_h

$\{\vec{\varphi}_i\}_{i=1}^n$:= vector basis functions for V_h

$\{\phi_j\}_{i=1}^m$:= scalar basis functions for W_h

Finite element matrices

$A_{ij} := \left(\mathcal{A}^{-1}\vec{\varphi}_i, \vec{\varphi}_j\right), \quad i,\, j = 1:n$

$B_{rj} := (\phi_r, \nabla \cdot \vec{\varphi}_j), \quad r = 1:m, j = 1:n$

$A_{I\,ij} := (\vec{\varphi}_i, \vec{\varphi}_j), \quad i,\, j = 1:n$

$D_{ij} := (\nabla \cdot \vec{\varphi}_i, \nabla \cdot \vec{\varphi}_j), \quad i,\, j = 1:n$

$N_{rs} := (\phi_r, \phi_s), \quad r,\, s = 1:m$

$S := BA^{-1}B^T$

$H := A + D$

$C := \begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix}$

Standard stability analysis

$a(\vec{u}, \vec{v}) := \left(\mathcal{A}^{-1}\vec{u}, \vec{v}\right)$

$b(\vec{v}, w) := (\nabla \cdot \vec{v}, w)$

α := continuous coercivity constant

β := continuous inf-sup stability constant

α_h := discrete coervicity constant

β_h := discrete inf-sup stability constant

$Z := \{\vec{v} \in V \mid b(\vec{v}, w) = 0 \quad \forall\, w \in W\}$

$Z_h := \{\vec{v} \in V_h \mid b(\vec{v}, w) = 0 \quad \forall\, w \in W_h\}$

Multigrid

ν := no. of smoothing steps

$\mathcal{S}$:= generic smoothing operator

$\mathcal{S}_{SGS}$:= Symmetric Gauss-Seidel smoothing operator

T_J : = finite element mesh, at level J

$\mathcal{M}_J$:= representation of operator $\mathcal{M}$ on level J

$x_J^{(m)}$:= mth approximation to vector defined on level J

$\mathcal{M}_{J-1}$:= coarse grid representation of operator $\mathcal{A}$

$x_{J-1}^{(m)}$:= mth approximation to vector defined on level $J-1$

$\mathcal{I}_J^{J-1}$:= restriction operator

$\mathcal{I}_{J-1}^J$:= prolongation operator

$\underline{e}_J^{(m)}$:= algebraic error in mth iterate at level J

$\underline{r}_J^{(m)}$:= residual error in mth iterate at level J

Algebraic Multigrid

α_S := parameter measuring strength of dependence

P_i := set of interpolation variables for node i

C := set of coarse grid variables

F := set of fine grid variables

N_i := set of non-zero connections to node i

D_i^w := set of all weak connections to node i, relative to α_S

C_i^s := set of strong C-connections to node i, relative to α_S

D_i^s := set of strong F-connections to node i, relative to α_S

References

[1] R.A. Adams, Sobolev Spaces, Academic Press, 1975.

[2] ALBERT toolbox documentation, `http://www.mathematik.uni-frieberg.de/IAM/Research/projectsdz/albert/doc.html`.

[3] R.E. Alcouffe, A. Brandt, J.E. Dendy Jr. and J.W. Painter, The multi-grid method for the diffusion equation with strongly discontinuous coefficients, *SIAM J. Sci. Comput.,* 2 (4), pp.430–454, 1981.

[4] T. Arbogast and Z. Chen, On the implementation of mixed methods as nonconforming methods for second-order elliptic problems, *Math. Comp.,* 64(211), pp.943–972, July 1995.

[5] D.N. Arnold, Mixed finite element methods for elliptic problems, *Meth. Appl. in Math. Eng,* 82, pp.281–300, 1990.

[6] D.N. Arnold, R.S. Falk and R. Winther, Preconditioning in H(div) and applications, *Math. Comp.,* 66(219), pp.957–984, 1997.

[7] D.N. Arnold, R.S. Falk and R. Winther, Multigrid in H(div) and H(curl), *Numer. Math.,* 85, pp.197–218, 2000.

[8] D.N. Arnold and F. Brezzi, Mixed and non-conforming finite element methods: implementation, postprocessing and error estimates, *M^2AN Math. Model. Numer. Anal.,* 19(1), pp.7–32, 1985.

[9] O. Axelsson and P.S. Vassilevski, Algebraic multilevel preconditioning methods I, *Numer. Math.,* 56, pp.157–177, 1989.

[10] O. Axelsson and M. Neytcheva, Preconditioning methods for linear systems arising in constrained optimisation problems, *Numer. Linear Algebra Appl.*, 10(3), pp.3–31, 2003.

[11] R. Bank, B. Welfert and H. Yserentant, A class of iterative methods for solving saddle-point problems, *Numer. Math.,* 56, pp.645–666, 1990.

[12] R. Bank and T. Dupont, An optimal order process for solving finite element equations, *Math. Comp.,* 36(153), pp.35–51, 1981.

[13] R. Barrett, M. Berry, T.F. Chan, J. Demmel, J. Donato, J.Dongarra, V. Eijkhout, R. Pozo, C. Romine and H. Van der Vorst, Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods, SIAM, Philadelphia, 1994.

[14] M. Bercovier and O. Pironneau, Error estimates for finite element method solution of the Stokes problem in the primitive variables, *Numer. Math.,* 33, pp.211–224, 1979.

[15] J.H. Bramble, Multigrid Methods. Pitman Research Notes in Mathematics Series, 294, Longman 1993.

[16] A. Brandt, Multi-level adaptive solutions to boundary-value problems, *Math. Comp.,* 31(138), pp.333–390, 1977.

[17] D. Braess, Finite Elements: Theory, Fast Solvers and Applications in Solid Mechanics, Cambridge University Press, 1992.

[18] D. Braess and W. Hackbusch, A new convergence proof for the multigrid method including the V-cycle, *SIAM J. Numer. Anal.,* 20(5), pp.967–975, 1990.

[19] D. Braess and R. Verfürth, Multigrid methods for nonconforming finite element methods, *SIAM J. Numer. Anal.,* 4, pp.979–986, 1990.

[20] S.C. Brenner, An optimal-order multigrid method for P1 nonconforming finite elements, *Math. Comp.,* 52(185), pp.1–15, 1989.

[21] S.C. Brenner, A multigrid algorithm for the lowest-order Raviart-Thomas mixed triangular finite element method, *SIAM J. Numer. Anal.,* 29(3), pp. 647–678, 1992.

[22] M. Brezina, A.J. Cleary, R.D. Falgout, V.E. Henson, J.E. Jones, T.A. Manteuffel, S.F. Mccormick and J.W. Ruge, Algebraic multigrid based on element interpolation (AMGe), *SIAM J. Sci. Comput.,* 22 (5), pp.1570–1592, 2001.

[23] F. Brezzi, On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers, *RAIRO Anal. Numér.,* 8, pp.129–151, 1974.

[24] F. Brezzi and K.J. Bathe, A discourse on the stability conditions for mixed finite element formulations, *Comput. Methods Appl. Mech. Engrg.,* 82, pp.27–57, North-Holland, 1990.

[25] F. Brezzi and J. Douglas Jr., Stabilised mixed methods for the Stokes problem, *Numer. Math.,* 53, pp.225–235, 1988.

[26] F. Brezzi and M. Fortin, Mixed and Hybrid Finite Element Methods, Springer-Verlag, New York, 1991.

[27] F. Brezzi, L.D. Marini and P. Pietra, Numerical simulation of semiconductor devices, *Comput. Methods Appl. Mech. Engrg.,* 75, pp.493–514, 1989.

[28] F. Brezzi, L.D. Marini and P. Pietra, On some numerical problems in semiconductor device simulation, *Lecture Notes in Math.,* 1460, pp.31–42, 1991.

[29] W.L. Briggs, Van E. Henson and S.F. McCormick, A Multigrid Tutorial, SIAM, 2000.

[30] X. Cai, B.F. Nielsen and A. Tveito, An analysis of a preconditioner for the discretised pressure equation arising in reservoir simulation, *IMA J. Numer. Anal.,* 19, pp.291–316, 1999.

[31] Z. Cai, C.I. Goldstein and J.E. Pasciak, Multilevel iteration for mixed finite element systems with penalty, *SIAM J. Sci. Comput.,* 14(5), pp.1072–1088, 1993.

[32] T.F. Chan, Domain decomposition algorithms, *Acta Numer.,*1994, pp.61–143.

[33] T.F. Chan and W.L. Wan, Robust multigrid methods for elliptic linear systems, *J. Comput. Appl. Math.*, 123, pp.323–352, 2000.

[34] Q. Chang, Y.S. Wong and H.Fu, On the algebraic multigrid method, *J. Comput. Phys.,* 125, pp.279–292, 1996.

[35] Q. Chang, Z. Huang, Efficient algebraic multigrid algorithms and their convergence, *SIAM J. Sci. Comput.,* 24(2), pp.597–618, 2002.

[36] Z. Chen, Equivalence between and multigrid algorithms for nonconforming and mixed methods for second-order elliptic problems, *East-West J. Numer. Math.,* 4(1), pp.1–33, April 1996.

[37] Z. Chen, R.E. Ewing and R.D. Lazarov, Domain decomposition algorithms for mixed methods for second-order elliptic problems, *Math. Comp.,* 65(214), pp.467–490, April 1996.

[38] Z. Chen, R.E. Ewing, Y.A. Kuznetsov, R.D. Lazarov and S. Maliassov, Multilevel preconditioners for mixed methods for second-order elliptic problems, *Numer. Linear Algebra Appl.*, 1(1), pp.1–27, 1996.

[39] A.J. Cleary, R.D. Falgout, V.E. Henson, J.E. Jones, T.A. Manteuffel, S.F. Mccormick, G.N. Miranda, and J.W. Ruge, Robustness and scalability of algebraic multigrid, *SIAM J. Sci. Comput.,* 21 (5), pp.1886–1908, 2000.

[40] K.A. Cliffe, I.G. Graham, R. Scheichl and L. Stals, Parallel computation of flow in heterogeneous media modelled by mixed finite elements, *J. Comput. Phys.,* 164(2), 2000.

[41] J. Douglas, R.E. Ewing and M.F. Wheeler, The approximation of the pressure by a mixed method in the simulation of miscible displacement, *RAIRO*, 17(1), pp.17–33, 1983.

[42] J. Douglas and J.E. Roberts, Mixed finite element methods for second-order elliptic problems, *Mat. Aplic. Comp.,* 1(1), pp. 91–103, 1982.

[43] J. Douglas and J.E. Roberts, Global estimates for mixed methods for second-order elliptic problems, *Math. Comp.,* 44(169), pp. 39–52, 1985.

[44] I.S. Duff, The solution of augmented systems. In: Numerical Analysis 1993. Pitman Research in Mathematics Series, D.F. Griffiths and G.A. Watson (Eds.), 303, Longman 1994.

[45] H.C. Elman and G.H. Golub, Inexact and preconditioned Uzawa algorithms for saddle point problems, *SIAM J. Numer. Anal.,* 31(6), pp.1645–1661, 1994.

[46] R.E. Ewing and J. Wang, Analysis of the Schwarz algorithm for mixed finite elements methods, *M^2AN, Mathematical Modelling and Numerical Analysis,* 26(6), pp.739–756, 1992.

[47] R.E. Ewing and J. Wang, Analysis of multilevel decomposition iterative methods for mixed finite element methods, *M^2AN, Math. Model. Numer. Anal.,* 28(4), pp.377–398, 1994.

[48] R.E. Ewing, R.D. Lazarov, P. Lu and P.S. Vassilevski, Preconditioning indefinite systems arising from the mixed finite element discretization of second-order elliptic systems. In: Preconditioned Conjugate Gradient Methods, O. Axelsson and L. Kolotilina (Eds.) *Lecture Notes in Math.* Springer-Verlag, Berlin, pp.28–43, 1990.

[49] R.E. Ewing and M.F. Wheeler, Computational aspects of mixed finite element methods. In: Scientific Computing, R. Stepleman (Ed), *IMACS*, North-Holland, pp.163–172, 1983.

[50] R.S. Falk and J.E. Osborn, Error estimates for mixed methods, *RAIRO,* 14(3) pp. 249–277, 1980.

[51] B. Fischer, Orthogonal polynomials and polynomial based iteration methods for indefinite linear systems, *Habilitation Thesis,* University of Hamburg, 1994.

[52] M. Fortin and R. Glowinski, Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-Value Problems, *Stud. Math. Appl.,* 15, North-Holland, Amsterdam, 1983.

[53] B. Fraeijs de Veubeke, Displacement and equilibrium models in the finite element method. In: Stress Analysis, O.C. Zienkiewicz and G. Holister, (Eds.), Wiley, New York, 1965.

[54] V. Girault and P.A. Raviart, Finite Element Methods for the Navier-Stokes Equations, *Springer Ser. Comput. Math.* 5, Springer-Verlag, New York, 1986.

[55] A. Greenbaum, Iterative Methods for Solving Linear Systems, SIAM, Philadelphia, 1997.

[56] G.H. Golub and C.F. Van Loan, Matrix Computations, John Hopkins University Press, 1989.

[57] W. Hackbusch, Elliptic Differential Equations: Theory and Numerical Treatment, Springer, Berlin, 1982.

[58] M.E. Harr, Groundwater and Seepage, Dover publications, New York, 1990.

[59] V.E. Henson, An algebraic multigrid tutorial, `http://www.mgnet.org/mgnet/Conferences/CopperMtn99/Tutorials/amgtut_files/v3_document.htm`, 1999.

[60] R. Hiptmair, Multilevel preconditioning for mixed problems in three dimensions, *PhD. Thesis,* Augsburg University, 1996.

[61] R. Hiptmair, Multigrid method for H(div) in three dimensions, *ETNA*, 6, pp.133–152, 1997.

[62] R. Hiptmair, B. Wohlmuth, T. Scheikofer, Multilevel preconditioned augmented Lagrangian techniques for 2nd order mixed problems, *Computing,* 57, pp.25–48, 1996.

[63] R.B. Kellogg, On the Poisson equation with intersecting interfaces, *Appl., Anal.,* 4, pp.101–129, 1975.

[64] C. Kim, On iteration and approximation methods for anisotropic problems, *Ph.D Thesis,* Texas A&M University, 2001.

[65] C.O. Lee, A nonconforming multigrid method using conforming subspaces, NASA conference publication, 3224, pp.317-330, 1993.

[66] S. Maliassov, Optimal Order Preconditioners for Mixed and Non-conforming Finite Element Approximations of Elliptic Problems with Anisotropy, *PhD. Thesis,* Texas A&M University, 1996.

[67] J. Mandel and P. Vaněk, Energy optimization of algebraic multigrid bases, *Computing, to appear.*

[68] K.A. Mardal, X.C. Tai and R. Winther, A robust finite element method for Darcy-Stokes flow, *SIAM J. Numer. Anal., to appear.*

[69] MATLAB Version 5, User's Guide, The Mathworks Inc, Prentice Hall, New Jersey, 1997.

[70] M. Mohr, B. Vanrumste, Comparing iterative solvers for linear systems associated with the finite difference discretisation of the forward problem in electro-encephalographic source analysis, *Medical and biological engineering and computing,* 41, pp.75–84, 2003.

[71] P. Morin, R.H. Nochetto and K.G. Siebert, Convergence of adaptive finite element methods, *SIAM rev.,* pp.1–28, 2002.

[72] J.C. Nedelec, Mixed finite elements in $I\!R^3$, *Numer. Math.,* 35, pp.315–341, 1980.

[73] M.G. Neytcheva, Algebraic multilevel iteration preconditioning technique: a matlab implementation, `http://www.netlib.org.`

[74] C.C. Paige and M.A. Saunders, Solution of sparse indefinite systems of linear equations, *SIAM. J. Numer. Anal.,* 12, pp.617–629, 1975.

[75] I. Perugia, V. Simoncini and M. Arioli, Linear algebra methods in a mixed approximation of magnetostatic problems, *SIAM J. Sci. Comput.,* 21(3), pp.1085–1101, 1999.

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

[77] C.E. Powell and D. Silvester, Optimal preconditioning for Raviart-Thomas mixed formulation of second-order elliptic problems, *SIAM J. Matrix Anal. Appl., to appear.*

[78] C.E. Powell and D. Silvester, Black-box preconditioning for mixed formulation of self-adjoint elliptic pdes, *Lect. Notes Comput. Sci. Eng., to appear.*

[79] A. Quarteroni and A. Valli, Numerical Approximation of Partial Differential Equations, Springer-Verlag, 1994.

[80] P.A. Raviart and J.M. Thomas, A mixed finite element method for second-order elliptic problems. In: Mathematical Aspects of the Finite Element Method, *Lect. Notes in Math.,* 606, Springer-Verlag, New York, 1977.

[81] S. Reitzinger, U. Schreiber, U. van Rienen, Electro-quasistatic calculation of electric field stength on high voltage insulators with an algebraic multigrid algorithm, *IEEE transactions on magnetics,* 39(4), pp.2129–2132, 2003.

[82] J.E. Roberts and J.M. Thomas, Mixed and hybrid methods. In: Handbook of Numerical Analysis, Vol II: Finite Element Methods (Part 1), P.G. Ciarlet and J.L. Lions, (Eds), North-Holland, pp.523–633, 1991.

[83] J.W. Ruge and K. Stüben, *Efficient solution of finite difference and finite element equations by algebraic multigrid (AMG).* In: Multigrid Methods for Integral and Differential Equations, The Institute of Mathematics and its Applications Conference Series, D.J. Paddon, H. Holstein (Eds.), New Series 3, Clarendon Press, Oxford, 1985, pp.169–212.

[84] J.W. Ruge and K. Stüben, Algebraic Multigrid. In: Multigrid Methods, S.F. McCormick (Ed.), SIAM, Philidelphia, 1987, pp.73–130.

[85] T.F. Russell and M.F. Wheeler, Finite element and finite difference methods for continuous flows in porous media. In: Mathematics of Reservoir Simulation, R.E. Ewing (Ed.), pp.35–106, 1983.

[86] T. Rusten and R. Winther, A preconditioned iterative method for saddlepoint problems, *SIAM J. Matrix Anal. Appl.,* 13(3), pp.887–904, 1992.

[87] T. Rusten and R. Winther, Substructure preconditioners for elliptic saddle point problems,*Math. Comp.,* 60(201), pp.23–48, 1993.

[88] T. Rusten, P.S. Vassilevski and R. Winther, Interior preconditioners for mixed finite element approximations of elliptic problems, *Math. Comp.,* 65(214), pp.447–466, 1996.

[89] R. Scheichl, Iterative Solution of Saddle-Point Problems Using Divergence-free Finite Elements with Applications to Groundwater Flow, *PhD. Thesis,* University of Bath, 2000.

[90] D. Silvester and A. Wathen, Fast iterative solution of stabilised Stokes systems part II: using general block preconditioners,*SIAM J. Numer. Anal.,* 31(5), pp.1352–1367, 1994.

[91] D. Silvester and A. Wathen, Fast and robust solvers for time-discretised incompressible Navier-Stokes equations. In: Numerical Analysis 1995. Pitman Research in Mathematics Series. D.F. Griffiths and G.A. Watson (Eds.), 344, Longman, 1996.

[92] G. Strang and G.J. Fix, An Analysis of the Finite Element Method, Prentice Hall, 1973.

[93] K. Stüben, *Algebraic multigrid (AMG): an introduction with applications.* In: Multigrid, U. Trotenberg, C.W. Oosterlee, A. Schüller (Eds.), Academic Press, New York, 2000.

[94] K. Stüben, A review of algebraic multigrid, *J. Comput. Appl. Math.* 128, pp.281–309, 2001.

[95] U. Trottenberg, C. Oosterlee and A. Schüller, Multigrid, Academic Press, 2001.

[96] P. Vaněk, M. Brezina and J. Mandel, Convergence of algebraic multigrid based on smoothed aggregation, *Numer. Math.*, 88, pp.559–579, 2001.

[97] P.S. Vassilevski and J. Wang, Multilevel iterative methods for mixed finite element discretizations of elliptic problems,*Numer. Math.*, 63, pp.503–520, 1992.

[98] P.S. Vassilevski and R.D. Lazarov, Preconditioning mixed finite element saddle-point elliptic problems, *Numer. Linear Algebra Appl.*, 3(1), pp.1–20, 1996.

[99] W.L. Wan, T.F. Chan, and B. Smith. An energy-minimizing interpolation for robust multigrid, *Technical report 98-6*, Dept. of Mathematics, UCLA, 1998.

[100] W.L. Wan, Interface preserving coarsening multigrid for elliptic problems with highly discontinuous coefficients, *Numer. Linear Algebra Appl.*, 7, pp.727–741, 2000.

[101] A.J. Wathen, Realistic eigenvalue bounds for the galerkin mass matrix, *IMA J. Numer. Anal.*, 7(1), pp. 449–457, 1987.

[102] A. Wathen and D. Silvester, Fast iterative solution of stabilised Stokes systems part I: using simple diagonal preconditioners, *SIAM J. Numer. Anal.,* 30(3), pp.630–649, 1993.